COMPARATIVE TECHNIQUES FOR THE EVALUATION OF CLUSTERING METHODS

By

JANICE LYNN DUBIEN

Bachelor of Science
Illinois State University
Normal, Illinois
1969

Master of Science
Oklahoma State University
Stillwater, Oklahoma
1973

Submitted to the Faculty of the Graduate College
of the Oklahoma State University
in partial fulfillment of the requirements
for the Degree of
DOCTOR OF PHILOSOPHY
July, 1976

COMPARATIVE TECHNIQUES FOR THE EVALUATION

OF CLUSTERING METHODS

Thesis Approved:

Thesis Adviser

Dean of the Graduate College

ACKNOWLEDGEMENTS

I would like to extend my sincere thanks and appreciation to Dr. William D. Warde for suggesting the problem and for his invaluable guidance and assistance during the course of this research. I would also like to thank Dr. J. Leroy Folks, Dr. Lyle D. Broemeling, and Dr. Joe V. Whiteman for serving on my advisory committee.

I would like to express my sincere appreciation to my parents, Harold and Eleanor DuBien, for their interest and encouragement throughout my life.

This manuscript was made possible through the invaluable technical assistance of my friends, Jan Jones, Ginny Gann, and Ceal Holbert. I am indebted to Jan Jones for the professional typing of this manuscript and to Ginny Gann for the graphical portrayals of the comparative study. I would also like to thank Ceal Holbert for assisting me with the proofreading of this manuscript.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

CHAPTER I

INTRODUCTION

Perspectives

The problem of finding the "best" procedure for classifying m individuals (generic) into k homogeneous populations on the basis of n observable characteristics has perplexed man through the ages. If the classification categories are known *a priori*, then discriminant analysis provides a solution to the general classification problem. However, if the classification categories must be generated from the data, then cluster analysis is the multivariate, descriptive method necessary to make sense out of the data.

The general classification problem has a very long and rich history, being dated at least to the time of Aristotle for its philosophical foundations. In essence, there is a "need to classify" in man (generic) which pervades his perpetual compulsion to organize and reorganize his world in search of a "perfect" organizational structure for each segment of his world and, ultimately, the universe. Man feels compelled to organize everything around him, and most conflicts among men are derived from different perceptions of what constitutes the "best" organizational structure for some segment of the world. The concepts of "necessary property," "natural grouping," and "natural kind" are all attributable to Aristotle, and they symbolize the origin

of man's belief in the existence of "natural" structure in the universe and in the existence of a "best" classification for any set of objects. Ideally, everything in the universe has a unique position in the "natural" grouping.

On the other hand, cluster analysis is still in a relatively embryonic state being dated in a sense with the publication in 1963 of The Principles of Numerical Taxonomy by Sokal and Sneath; for initially, it was in the context of applying quantitative methods to taxonomical data that clustering methods evolved to provide solutions for the general classification problem. Cluster analysis has developed in a relatively isolated state in many diverse fields of application including biology, psychiatry, criminology, ecology, psychology, sociology, engineering, soil science, economics, and marketing research to mention only a few. A more complete and organized listing with discussion appears in Anderberg (1973). In addition, some of the relevant cluster analysis research is being published in computer science and statistical journals.

The result of all of this diversity in the evolution of cluster analysis is a lack of any standard notation or terminology for the concepts of cluster analysis, a duplication of research, and the development of fringe areas to satisfy a more manageable and well-defined set of objectives. Anderberg (1973, p. 7) offers some causes for and some criticisms of the diversity in cluster analysis.

> The cause [of diversity] is probably a mixture of professional jealousy, a relative isolation among the fields, and genuine differences of viewpoint. For the novice, the disarray is bewildering and confusing; ultimately it is highly duplicative since the same idea is discovered repeatedly and published in a variety of journals.

On the fringes of cluster analysis are such diverse fields as pattern recognition, information theory, mixtures of probability distributions, graph theory, multidimensional scaling, and artificial intelligence.

In spite of the shades of gray and the diversity of evolution in cluster analysis, a unifying framework for the development of the theoretical aspects of cluster analysis can be found among the statistical methods. Since statistics is a body of methods purporting to aid in making sense out of data, cluster analysis belongs among the descriptive, statistical methods; and as a descriptive method, cluster analysis possesses the following noteworthy characteristics:

1. It is an exploratory technique to be used in the initial stages of research which, hopefully, will precipitate hypotheses for further research;
2. It has as its goal simplification through organization.

Within the body of statistical methods presently available for data analysis, there exists a hierarchy of descriptive methods based on the dimensionality of the data to be analyzed. This hierarchy of descriptive methods is briefly outlined below:

1. Ordering (ranking) -- univariate,
2. Graphing (scatter plots) -- bivariate,
3. Response surfaces (models) -- trivariate,
4. Factor analysis (Principal Components) -- multivariate,
5. Cluster analysis (Numerical Taxonomy) -- multivariate.

As Warde (1975) indicates, cluster analysis may also be viewed as the multivariate analogue to multiple comparisons.

In viewing cluster analysis from its philosophical, historical, and statistical perspectives, the inherent difficulties of research within this area imposed by its voluminous and diverse literature have become apparent. Consequently, any meaningful research within the realm of cluster analysis must be limited to a well-defined facet of cluster analysis, and a consistent set of terms, definitions, and symbols must be imposed for the exposition of this research. Thus, before defining the limits of this study, some definitions will be tendered.

A Discussion of Fundamental Concepts with Some General Definitions

The central concept in cluster analysis is that of cluster, but its definition is as diverse as the many applications of cluster analysis. In fact, as Kendall (1973, p. 181) states, "The fundamental problem in cluster analysis is to define what we mean by 'cluster'." Intuitively, the concept of cluster encompasses the duality of homogeneity within clusters and heterogeneity between clusters. Thus, there must also be some concept of "closeness." However, Rand (1971, p. 846) believes, "that every definition of 'closer' is natural for some situation." In the following passage, Kendall (1973, p. 181) further exemplifies the contextual variation which occurs in the concept of a cluster:

> But what are we to say of the particles which compose one of Saturn's rings, which are certainly a grouping, but a hollow one; or the tracks of a particle in a Wilson cloud chamber, which is an organized series of droplets but a linear one? And if we allow a scatter of points inside an ellipse to constitute a cluster, what are we to say of two such shapes with common centre and major axes at right-angles — are they one cluster or two overlapping clusters?

Hence, the difficulty ascribed to defining a cluster is one of specificity rather than generality. Ideally, the definition of a cluster defines very special clusters for each specific application of cluster analysis; and at the same time, it must be completely general, defining a cluster for every possible application of cluster analysis. This ideal, of course, is a logical impossibility. With an example, Norton (1975) also cites the inherent difficulty involved in any attempt to find a single all purpose definition of cluster. Through the literature, there are a multitude of different, idealistic definitions of a cluster. Practically, however, most definitions of cluster are operational in the sense that a clustering method is chosen which then determines the kind of cluster generated. Unfortunately, very little information is available concerning the association between clustering method and type of cluster generated.

The definitional problems associated with cluster analysis can be at least partially resolved by a mathematical approach to the problem. Using some of Rand's (1969) notation to formalize the presentation, a general, set theoretic framework will be established for cluster analysis.

Noting that the primitive components of set theory are element and set, parallel concepts in cluster analysis are the elements to be clustered and the set consisting of these elements. In general terms, the elements to be clustered have been called objects, individuals, patterns, and by Sokal and Sneath (1963) operational taxonomic units (OTUs). The elements to be clustered shall be referred to as data points in this paper, and each data point shall be represented by a $p \times 1$ vector, X_i, where

$$X_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ . \\ . \\ . \\ x_{ip} \end{bmatrix} .$$

The components, x_{ij}, of X_i will be termed variables. The set of all elements to be clustered shall be called the object space and symbolized by X. Letting N be the number of data points, then

$$X = \{X_1, X_2, \ldots, X_N\} .$$

Obviously, the object space is embedded in Euclidean p-space. Thus, if E_p represents Euclidean p-space, then $X \subseteq E_p$.

A popular conceptualization of the object space is the data matrix which is formed by stacking the data points as rows of a matrix. Letting $X_{N,p}$ represent the data matrix, where N is the number of data points and p is the number of variables, then

$$X_{N,p} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ . & . & & . \\ . & . & \cdots & . \\ . & . & & . \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{bmatrix} .$$

Having laid a set-theoretic foundation for discussing cluster analysis concepts, mathematical definitions for cluster and clustering can be given.

Definition 1. A cluster, Y_k, is any nonempty subset of the object space. Symbolically, $Y_k \subseteq X$ which means that if $X_i \in Y_k$, then

$X_i \in X$.

Thus, a cluster is simply a collection of data points.

Definition 2. A clustering, Y, is any partition of the object space. Symbolically, $Y = \{Y_1, Y_2, \ldots, Y_K\}$ is a partition of X, if the following three conditions hold:

(i) For every $Y_k \in Y$, $Y_k \neq \phi$.

(ii) If $Y_k \in Y$, $Y_m \in Y$, and $Y_k \neq Y_m$, then $Y_k \cap Y_m = \phi$.

(iii) $\bigcup_{k=1}^{K} Y_k = X$.

Hence, a clustering is simply a special kind of collection of clusters.

A clustering of N data points can consist of $K = 1, 2, \ldots, N$ clusters. The number of clusters contained in a clustering shall be termed the size of the clustering, and this designation will be incorporated into the general notation for a clustering by the use of a superscript. For example, if clustering Y contains K clusters, then Y^K denotes a clustering of size K. The set of all possible clusterings of the object space shall be denoted by $\mathcal{Y}$. The fact that even for small values of N, the cardinality of $\mathcal{Y}$ is quite large has motivated the development of a multitude of clustering methods, not all of which are distinct.

In very general terms, a clustering method consists of a criterion and a technique in which case the criterion assigns a numerical value to each clustering and the technique selects a subset of the set of all possible clusterings over which the criterion is optimized (providing only a local optimum). A problem of major proportions is to classify the many clustering methods into a small number of different types. Noteworthy attempts at classifying and reviewing clustering methods appear

in Sneath and Sokal (1973), Norton (1975), Cormack (1971), Anderberg (1973), and Everitt (1974). However, no standard terminology has emerged to clarify the confused nomenclature that exists for designating an entire family of similar clustering methods. Apparently, "agglomerative hierarchical" given by Anderberg (1973) and Everitt (1974), "sequential, agglomerative, hierarchal" given by Norton (1975), and "sequential, agglomerative, hierarchic, nonoverlapping (SAHN)" given by Sneath and Sokal (1973) are all descriptors for the same class of clustering methods which was also defined as a "hierarchical clustering scheme (HCS)" by Johnson (1967). The previously described class of clustering methods will be of primary importance in this paper, and these clustering methods shall be referred to simply as agglomerative clustering methods.

Agglomerative clustering methods are some of the oldest and most frequently used clustering methods. An agglomerative clustering method may be characterized as proceeding sequentially by joining pairs of clusters from the partition which consists of each data point grouped as a single cluster to the partition which consists of all data points grouped together in a single cluster (if no stopping rule is provided). An important concept in the definition of an agglomerative clustering method is an hierarchy.

Assuming that there are N data points, formal definitions for hierarchy and agglomerative clustering method are given as Definitions 3 and 4, respectively.

Definition 3. A hierarchy, H, on the object space is an ordered sequence of nested clusterings. Symbolically,

$$H: \; Y^N, Y^{N-1}, \ldots, Y^2, Y^1,$$

$$\text{where } Y^N \subset Y^{N-1} \subset \ldots \subset Y^2 \subset Y^1 \; .$$

One useful visualization of a hierarchy is a tree diagram which is often called a dendrogram in cluster analysis applications. Summarizing, a hierarchy on the object space is a nested collection of clusterings (each consisting of a set of clusters) which may be aptly depicted by a dendrogram.

Definition 4. An agglomerative clustering method is any clustering method, m, which produces a hierarchy on the object space subject to the following constraints:

(i) Y^N is the initial clustering;

(ii) Clustering Y^{K-1}, $K \leq N$, is obtained from clustering Y^K by joining the two "closest" clusters in clustering Y^K; i.e., if $Y_i, Y_j \in Y^K$ and they are deemed "closest", then $Y_i \cup Y_j \in Y^{K-1}$.

Thus, the application of an agglomerative clustering method to the N data points results in a special kind of hierarchy, thereby imposing an hierarchical structure on the object space.

The resolution of a clustering problem by the application of an agglomerative clustering method to a data set can be described by the triple (X, H, m); for future reference, the components of this triple have been carefully defined in this section. Recalling that, in general, a clustering method consists of a criterion and a technique, an agglomerative clustering method may be more specifically viewed as consisting of a measure of similarity or dissimilarity (usually a metric) and an algorithm (usually a form of linkage). The measure of similarity or

dissimilarity explicates "close", initially; and the algorithm reevaluates the "closeness" of clusters after each join. As a further limitation, the agglomerative clustering methods of particular interest in this paper may be denoted by the pair (metric, algorithm).

Further delineation of the particular agglomerative clustering methods of interest will be given in Chapter III. However, sufficient terminology and notation have been developed to define the scope of the study being presented in this paper.

The Rationale and Scope of This Study

Having placed cluster analysis among the descriptive, statistical methods, the problem of actually implementing a clustering method, given a "real" set of data, is a bewildering one. The data analyst must make many choices before a data set can be cluster analyzed such as the following questions exemplify:

1. Should he standardize the variables?
2. Should he factor analyze the variables before clustering the data points?
3. What value of K, the number of clusters to be found in the data set, should he specify?
4. What clustering method should he use?

Although this study primarily addresses itself to the fourth question, a brief discussion of the first three questions is relevant.

The first two questions make reference to often advocated solutions for frequently encountered problems concerning the variables observed on each data point. Typically, a data point consists of measurements on a myriad of related variables with divergent ranges, and often these

measurements are made in many different incompatible units. Since, inevitably, the variables are combined in a measure of similarity or dissimilarity or in a criterion, the incompatibility of units problem cannot be entirely ignored, and standardization of the variables does at least result in unitless quantities (making at least the mathematicians happy). However, from a statistical point of view, standardization is not the panacea its advocates would lead one to believe, especially since only sample moments are available for use in this process. It is worth noting that Kendall (1973) favors standardization of the variables as the lesser of several evils, but for the most part, standardization is opposed by Anderberg (1973). Within the numerical taxonomy literature, there exist many philosophical discussions concerning the importance of weighting certain characteristics and the hazards of forcing all characteristics to have the same relative weights. Sneath and Sokal (1973) provide a good reference to the numerical taxonomy literature and to the biological viewpoint on philosophical questions. Anderberg (1973) provides an extensive discussion of alternatives to standardization based on the scale of measurement of the variables. Applying factor analysis or even principle components to a set of variables before cluster analyzing the data points may reduce the number of variables, but research on the invariance of clustering methods to these transformations is lacking. It should also be noted that there is no reason to believe that simple correlation is the only relationship between pairs of variables. In this study, problems concerning variables will be ignored. However, additional research on this subject would be valuable.

The third question is relatively unimportant when agglomerative clustering methods are being used. If feasible, the complete hierarchy should be examined as the output from the application of an agglomerative clustering method to the data set. Often valuable additional information about the data points can be gained from the sequence of clusterings, which would be totally lost if only one clustering was examined.

The purpose of this study is to provide a "dynamic" comparison of agglomerative clustering methods, which will guide the matching of clustering method with type of cluster generated. Ideally, the comparative study would follow the suggestions made by Anderberg (1973, p. 201) in the following passage:

> What seems to be needed is an approach to evaluation which systematically can relate the key characteristics of cluster analysis problems to the capacities of various cluster analysis methods; in other words, find the elements which make problems difficult and match them with the strengths of powerful methods. If there could be found a set of significant concept dimensions which describes problems and another such set which describes methods, then a variety of important capabilities might be within reach.

Through the literature, there have been both analytical and empirical attempts to compare some clustering methods, but because of the large number of clustering methods now in existence and because of the number of factors requiring controlled change to make the comparisons relevant, a useable comparative summary of clustering methods is non-existent.

Consequently, the comparative study presented in this paper is limited to agglomerative clustering methods of the form (metric, algorithm), but a comprehensive study of these clustering methods is attempted in this paper. Chapter III contains an algebraic analysis of agglomerative

clustering method algorithms, which results in a graphic portrayal of these algorithms and a classification scheme for these algorithms based on the degree of distortion perpetrated on the object space by the algorithms in each group. Chapter IV presents a statistical analysis of the comparative statistic employed in Chapter V, which provides a distribution for the statistic under the specific model assumptions considered. Chapter V delineates the important considerations in any extensive, systematic comparison of clustering methods, and then it presents an empirical investigation of the effect of correlated variables on the "retrieval" ability of agglomerative clustering methods. First, however, a review of cluster analysis literature will be given for perspective.

CHAPTER II

A REVIEW OF CLUSTER ANALYSIS LITERATURE

A Classification of Cluster Analysis Literature

The voluminous and diversified nature of the cluster analysis literature has already been alluded to as a major impasse to research in cluster analysis. Considering the present state of knowledge in the realm of cluster analysis, making sense out of the cluster analysis literature would represent a major advance in cluster analysis research. Initially, a classification of the cluster analysis literature into representative categories would be a valuable implement.

In the preparation of this thesis, a sizeable sample of the cluster analysis literature was perused. Thus, the problem at hand is how to efficiently summarize a set of publications all purported to discuss subject matter related to cluster analysis. Rhetorically, the solution would be to write a "comparison and contrast" of the publications. Essentially, this means to extract those things which are similar and those things which make each publication unique, which in essence is the goal in the general classification problem. Thus, a particular instance of the general classification problem is to be solved as an efficient means to summarizing a sample of the cluster analysis literature.

In this chapter, a subjective classification of the publications into representative categories based on what is perceived to be their primary purpose is tendered. First, however, it should be noted that

most journal articles in the realm of cluster analysis either propose a new clustering method or as Cormack (1971, p. 323) comments:

> Unfortunately the current swell of classificatory publications (estimated at more than 1,000 a year) is mainly devoted to 'testing' published techniques on data for which 'standard' classifications exist. When the technique fails the author's response is to modify the technique instead of thinking about the 'standard' classification or questioning the value of the whole process.

With this in mind, the four primary purposes discerned from the cluster analysis publications sampled are as follows:

1. To survey classification procedures;
2. To propose or modify a clustering method;
3. To present statistical aspects of cluster analysis;
4. To compare aspects of clustering methods.

Since this classification is monothetic, the four primary purposes define a partition of the sample of cluster analysis publications into four clusters. However, the unavoidable overlapping of related publications becomes apparent when their secondary purposes are examined. Although significant secondary purposes could be used to refine the classification by defining sub-clusters, in the present review of cluster analysis literature, the initial four clusters are deemed adequate, and all relevant secondary purposes are revealed within the defined clusters as significant contributions to cluster analysis research.

Since to compare clustering methods is of principal interest in this thesis, an extensive critical review of publications falling in the cluster defined by a primary purpose "to compare" will be given. First, however, some of the publications falling in the other three clusters will be briefly discussed with particular emphasis being given to their significant contributions within the realm of cluster analysis.

Publications Having the Primary Purpose to Survey Classification Procedures

The first cluster of publications defined by a primary purpose "to survey classification procedures" or simply "to survey" contains several journal articles, two monographs, and two books. Since cluster analysis has been developing separately in a multitude of different applied fields, an interesting overview of the publications falling in this cluster is obtained by considering the viewpoint of the author. The important question is: For whom is the publication being written? The following listing of publication by perspective is enlightening:

1. From a biological sciences perspective -- Sneath and Sokal (1973)
2. From a social sciences perspective -- Everitt (1974), Ball (1965), and Fleiss and Zubin (1969)
3. From the viewpoint of the data analyst -- Anderberg (1973)
4. From the viewpoint of the econometrician -- Duran and Odell (1974)
5. From a statistical perspective -- Cormack (1971)
6. From a philosophical perspective -- Sneath (1969), Sokal (1974), and Kendall (1973).

The book by Sneath and Sokal (1973) is certainly a landmark in numerical taxonomy, but the biological nomenclature and the extensive discussion of special problems associated with taxonomy make it less valuable as a general reference in the realm of cluster analysis than the book by Anderberg (1973) or the monographs by Everitt (1974) and by Duran and Odell (1974). These other three publications are presented in an essentially context free manner, and each of these publications

provides a comprehensive general review of clustering methods, including a classification of the clustering methods into broad general categories and discussions of measures of similarity, measures of dissimilarity, measures of association, clustering algorithms, clustering criteria, and clustering techniques. Of special significance, however, are the noteworthy original contributions to cluster analysis research that each of these three publications makes.

Anderberg (1973) provides a self-contained presentation of cluster analysis which is organized to guide the data analyst sequentially from the raw data to the finished cluster analysis, including an extensive collection of well-documented computer programs to implement the complete sequence from raw data to finished analysis. His comprehensive analysis of problems pertaining to variables, scales of measurement, and measures of association includes commentary on strategies for mixed variable data sets, conversion of variables from scale to scale, compatibility of measures of association across variables, and weighting of variables, both explicitly and implicitly. The chapter entitled "Comparative Evaluation of Cluster Analysis Methods" provides the framework for a "dynamic" comparison of clustering methods, which includes a suggestion for making sense out of the resultant clusterings, namely, cluster the clustering methods. Anderberg (1973, p. 201) states:

> A possible approach for discovering these concept dimensions is to turn cluster analysis on itself and cluster the results obtained by applying available methods to specially constructed data sets. The similarities and differences among various clustering methods may be identified through comparison of the results obtained by clustering data sets of known characteristics, and the characteristics of various data sets may be discovered through clustering them with methods having known properties.

Thus, Anderberg (1973) gives some philosophical perspectives on

comparative studies, which should be considered in any attempt to compare clustering methods.

The monograph by Everitt (1974) presents an incisive discussion of the problems encountered when applying cluster analysis to "real" data, which includes enlightening commentary on defining a cluster, choosing the variables, choosing a measure of similarity or distance, choosing the number of clusters present in the data, and special problems associated with each type of clustering method. Everitt (1974) then aptly demonstrates the problems associated with various clustering methods by applying representatives from different types of clustering methods to data sets generated from bivariate normal distributions, having various degrees and kinds of structure. He also includes scatter plots for each generated data set to give an elucidative illustration of the structure and irregularities within the data sets which lead to the anomalous clusterings. The main purpose of the empirical investigation of different classes of clustering methods is not to compare the clustering methods, but to discover how a wide variety of supposedly different clustering methods perform on a few well-defined types of data structure. In fact, Everitt (1974) deliberately constructs his empirical investigation to test the strength (without a quantitative measure of it) of the underlying assumptions of various clustering methods to impose a structure on the data rather than find the structure existing in the data set. Everitt (1974, p. 87) concludes:

> All the methods make implicit assumptions about the type of structure present: when these assumptions fail to be met spurious solutions are likely to be obtained.

Duran and Odell (1974) attempt to unify the various results of research in the realm of cluster analysis and present them in a coherent

fashion, establishing mathematical notation for many of the concepts of cluster analysis. The resultant monograph consists primarily of a classification of clustering methods into broad, general categories with in depth and mathematically rigorous (extensively employing graph theory in the case of agglomerative clustering methods) discussions of the clustering methods contained in each group, emphasizing their common characteristics. A valuable contribution of this monograph is the chapter on clustering by complete enumeration and the subsequent chapter on dynamic programming techniques as "good" approximations to clustering by complete enumeration.

The journal articles by Ball (1965) and by Fleiss and Zubin (1969) are both written for the social scientist. Ball (1965) gives a comprehensive discussion of the seven major classifications of cluster seeking techniques with summaries of known measures of similarity, criteria for clustering, and techniques for clustering. He essentially provides a case against the normal assumption and a case for iterative clustering methods. On the other hand, Fleiss and Zubin (1969) present a brief critical review of factor analysis, cluster analysis, and mixtures of distributions as procedures for clustering individuals into homogeneous groups with specific emphasis on the logical and technical problems which arise in cluster analysis.

Each of the last four journal articles offers a measure of philosophical insight into the concept dimension of cluster analysis. The article by Cormack (1971) represents an in depth survey of all aspects of the general classification problem along with many amusing philosophical comments to amplify his scintillating style. In contrast, the article by Sneath (1969) represents a more limited survey of some aspects

of cluster analysis with particular emphasis on the unsolved problems in this relatively new branch of multivariate statistical analysis. Sokal (1974) presents an enlightening discussion of the purposes, principles, progress, prospects, and problems of classification from a philosophical perspective. Finally, Kendall (1973) discusses from a non-technical, but philosophical, perspective the nature of the problems of cluster analysis.

Publications Having the Primary Purpose to Propose or Modify a Clustering Method

The second cluster of publications defined by a primary purpose "to propose or modify a clustering method" consists of numerous journal articles. However, some of these journal articles also provide valuable, theoretical and practical discussions.

The journal articles by Fisher (1958), Edwards and Cavalli-Sforza (1965), Mayer (1971), and Scott and Knott (1974) present clustering methods which are essentially univariate. The divisive clustering method devised by Edwards and Cavalli-Sforza (1965) is used by Scott and Knott (1974) to group treatment means. The clustering method proposed by Mayer (1971) involves the choice of a primary variable to make the initial monothetic clustering, and then the secondary variables are used to refine the initial clustering.

Some specialized clustering methods are given by Fortier and Solomon (1966), King (1967), and Hartigan (1970). King (1967) proposes a step-wise, "quick and dirty" clustering method for separating a large number of variables into a group of clusters so that the variables within a cluster are highly intercorrelated and variables from different

clusters are not so highly intercorrelated. Hartigan (1970) presents an extensive review of cluster analysis by emphasizing six problem areas of cluster analysis; however, his primary purpose is to present a new clustering technique which simultaneously clusters variables and cases of a data matrix. He gives the following two justifications for this "better" clustering method:

> The principle justification for this technique is that the clusters obtained may be interpreted directly on the data matrix, rather than on the distance function usually necessary in other techniques. A second justification is that this direct clustering technique seems more in accord with the practice of biological taxonomists, who associate with each cluster (taxon) of animals, the cluster properties the animals have in common (Hartigan, 1970, p. 1.2).

Two of the journal articles in this cluster tender generalizations of the single linkage clustering method. Jardine and Sibson (1968) propose a sequence of overlapping clustering methods as an extension of the single-link method to reduce chaining after claiming that the single-link method is the "best" of the well-known agglomerative clustering methods with respect to their seven properties of a hierarchic classificatory scheme. Wishart (1969b) devises mode analysis to reduce the chaining effect associated with the single linkage clustering method.

Two journal articles by Lance and Williams (1966, 1967) form the basis for Chapter III of this paper. Lance and Williams (1966) tender a general linear combinatorial strategy based on four parameters, which yields an agglomerative clustering method algorithm for each choice of parameter values. The parameter values for five of the well-known agglomerative clustering methods are also given. The parameter values of this general linear combinatorial strategy for Ward's (1963) sum of squares clustering method are derived by Wishart (1969a). The second

journal article by Lance and Williams (1967) presents some properties associated with the general linear combinatorial strategy and a new agglomerative clustering method called the flexible strategy.

The journal articles by Hartigan (1967) and by Gower and Ross (1969) provide graph-theoretic approaches to clustering. Hartigan (1967) creates a measure of distance between a similarity matrix and a tree. Gower and Ross (1969) introduce the minimum spanning tree as a useful ancillary technique.

In addition to their primary purpose "to propose or modify a clustering method," three of the journal articles make noteworthy mathematical and statistical contributions to cluster analysis research. Johnson (1967) introduces the ultrametric inequality to define a hierarchical clustering scheme. Rubin (1967) presents a general framework for cluster analysis through mathematical definitions, properties, and proofs; he also creates a measure of object stability. Besides a local optimization program with single point reassignment and amalgamation of clusters criteria, Beale (1969) gives a reasonable criterion for the number of clusters based on a one-way classification MANOVA and an F-test.

Publications Having the Primary Purpose to Present Statistical Aspects of Cluster Analysis

The third cluster of publications defined by a primary purpose "to present statistical aspects of cluster analysis" contains two theses and five journal articles of a thoeretical nature. The journal articles by Marriott (1971) and by Scott and Symons (1971) are grouped in this

cluster because they contain numerous applications of statistical tools to cluster analysis problems. Marriott (1971) uses MANOVA criteria and the distribution theory associated with a multivariate analysis of variance. Scott and Symons (1971) employ likelihood ratio criteria in their investigation of cluster analysis.

The journal articles by Goodall (1967), Engelman and Hartigan (1969), and Bolshev (1969) represent attempts to develop theoretical aspects of cluster analysis. Goodall (1967) gives a distribution for the matching coefficient under certain sets of assumptions. Engelman and Hartigan (1969) empirically derive a table of percentage points of a test for the presence of clusters in data, but their test for the presence of structure is limited to the univariate case. Bolshev (1969) makes an initial attempt at constructing a general probabilistic theory of cluster analysis.

The thesis by Mrachek (1972) and the thesis by Norton (1975) necessarily make valuable contributions to the theoretical development of cluster analysis, and both of these theses are at least partially concerned with the problem of testing for the presence of structure in data. Mrachek (1972) develops a distribution theory for his metric of Euclidean distance so that he can apply inferential theory to the two approximate tests for structure which he suggests. He also considers the effect of uninformative variables on the ability of the single linkage and the complete linkage clustering algorithms to provide the correct clustering of a structured data set. Norton (1975) discerns two types of cluster analysis, which he refers to as mathematical clustering and inferential clustering, based on the type of "evidence" provided by the cluster analysis with respect to the data. Norton (1975) demonstrates the

difficulties encountered in attempts to construct "good" tests for the presence of structure based on closed form sampling distributions, and then he proposes several approximate tests for the presence of clusters based on agglomerative clustering methods. Specifically, he presents tests to detect the presence of more than one univariate normal population along with tabulated percentage points of their null distributions for selected sample sizes.

A Critical Review of Publications Having a Primary Purpose to Compare Aspects of Clustering Methods

The fourth cluster of publications defined by a primary purpose "to compare aspects of clustering methods" or simply "to compare" is of principal importance to the research being reported in this paper. The comparative studies of this cluster are either primarily theoretical, both analytical and empirical, or primarily empirical in nature. It should be noted that most of the development within this cluster is of fairly recent vintage.

Until recently, the cophenetic correlation coefficient, originated by Sokal and Rohlf (1962), was the only comparative statistic available for use in cluster analysis. Essentially, the cophenetic correlation coefficient is the ordinary product moment correlation coefficient computed from the corresponding elements of the original similarity (dissimilarity) matrix and the elements of a similarity (dissimilarity) matrix derived from a dendrogram; it may be computed on any two similarity (dissimilarity) matrices derived from dendrograms representing the same set of data (the matrices, of course, must have the same

dimensions). By the method of cophenetic correlation, different hierarchical clustering methods can be indirectly compared with each other through their derived dendrograms, and their derived dendrograms can be compared with the original similarity (dissimilarity) matrix to provide a measure of distortion for each clustering method with respect to the data set.

The method of cophenetic correlation has come under heavy criticism since its inception with impetus for this criticism being provided in a journal article by Farris (1969). Farris (1969) derives some algebraic properties of the cophenetic correlation coefficient, and he discovers the conditions under which the cophenetic correlation coefficient is maximized for a dendrogram. His analysis implies that agglomerative clustering methods based on an average linkage clustering algorithm should produce the highest cophenetic correlation coefficients among existing agglomerative clustering methods, when these clustering methods are compared against the original similarity (dissimilarity) matrix by the method of cophenetic correlation; and this implication is not tied to any underlying data structure. In theory, at least, a "best" clustering method with respect to the cophenetic correlation coefficient can be constructed.

The journal articles by Gower (1967) and by Fisher and Van Ness (1971) present comparative studies which are primarily theoretical in nature. Gower (1967) compares three well-known clustering methods from a geometrical point of view in order to expose the underlying cluster structure being assumed by these clustering methods. Fisher and Van Ness (1971), along with the extension of their work by Van Ness (1973), list eleven admissibility criteria which any "good" clustering method

should possess. They then compare nine different clustering methods with respect to these admissibility criteria, but their comparison is entirely theoretical employing mathematical proof to construct an admissibility table.

The journal articles by Friedman and Rubin (1967), Chaddha and Marcus (1968), and Maronna and Jacovkis (1974) represent extensive comparative studies containing both analytical and empirical comparisons. The journal article by Friedman and Rubin (1967) contains both an analytical and an empirical comparison of three generalized variance criteria along with many other theoretical and practical considerations relevant to clustering methods. Chaddha and Marcus (1968) compare three generalized distance statistics both analytically and empirically. Maronna and Jacovkis (1974) compare five diverse metrics with only Euclidean distance coming from the family of Minkowski metrics. Initially, their comparison of these metrics is analytical exhibiting the relationships between the five metrics and generalized variance criteria. Then the three "best" metrics based on the theoretical analysis are combined with an iterative technique and compared empirically on both "real" data and data generated from bivariate normal populations.

Several of the publications in this cluster present comparative studies which are primarily empirical in nature. Two of the earlier empirical, comparative studies are given by Williams, Lambert, and Lance (1965) and by Boyce (1969). Williams, Lambert, and Lance (1965) provide an empirical comparison of ten different clustering methods formed by using the single linkage and the centroid clustering algorithms in combination with each of five different measures of similarity or dissimilarity; these clustering methods were compared using "real"

data sets from ecology and with respect to the amount of chaining observed as measured by a coefficient of chaining which was also developed in the article. One interesting conclusion drawn from this comparative study by the authors is that there exists interaction between measures of similarity or dissimilarity and clustering algorithms. The journal article by Boyce (1969) represents an extensive empirical, comparative study, using cophenetic correlation techniques and graphic techniques to compare three agglomerative, pair-group clustering methods amongst themselves and against a principal components analysis of the data. This journal article also includes a comparison of five measures of similarity or dissimilarity from a theoretical point of view and from an empirical study using the unweighted pair-group algorithm based on averages. For the anthropological data employed in this study, the overall pattern of relationships was unaffected by the measure of similarity or dissimilarity used.

The recent journal article by Kuiper and Fisher (1975) is a prime example of a very poorly reported empirical, comparative study. The journal article by Kuiper and Fisher (1975) suffers more from what they did not say than it benefits from what they did say. Just to exemplify the absurdity of their style of reporting, the following quote is offered as evidence of their attempt to conceal any potentially enlightening details of their empirical study:

> It is neither feasible nor desirable to present most of the output. The percentages given below are averages of average values across various configurations (or probability distributions) (Kuiper and Fisher, 1975, p. 778).

The journal article by Kuiper and Fisher (1975) suffers from the following major defects and omissions:

1. It does not indicate which measure of similarity or dissimilarity was used with the six agglomerative clustering algorithms employed in the study;
2. Although the authors indicate that the Monte Carlo runs were made on a CDC 6400 computer, they give absolutely no indication of the procedure or computer package used to generate the multivariate normal data sets;
3. Even the configuration of the mean vectors is omitted for the cases where there are more than two multivariate normal populations being generated;
4. For the case of two bivariate normal populations, the configuration of mean vectors implies that one variable is completely uninformative, and thus the supposedly bivariate clustering problem is really reduced to a univariate clustering problem with "noise";
5. Averaging all results over configurations as well as the small number of replications (30) of each configuration makes the reported results totally uninterpretable.

In all fairness, the journal article by Kuiper and Fisher (1975) is a relatively short article that might have been substantially chopped before publication. Unfortunately, however, the conclusions and comments (based on all of the research done, not just the reported results) made in this journal article could have been completely anticipated based on previous comparative studies and theoretical knowledge of the clustering algorithms used.

In contrast, the technical report by Dubes and Jain (1975) is an outstanding example of a well reported and well conducted empirical,

comparative study with many new insights to offer the potential cluster analysis user. Dubes and Jain (1975) produce a comprehensive data analysis of a 192 X 8 dimensional subset of the Munson handprinted Fortran character set referred to as IMOX, which does not cluster in a trivial manner. Their objective is not to find a "best" clustering method, but to explore the strengths and peculiarities of several diverse clustering methods on a challenging data set for which a "natural" classification exists.

Comparisons of clustering methods which are from different classes such as the hierarchical and non-hierarchical classes of clustering methods are practically nonexistent because the outputs from clustering methods which are from different classes are, in general, noncomparable. However, Dubes and Jain (1975) successfully compare the performance of eight clustering methods representing three diverse classes (squared-error, hierarchical, and graph-theoretic) of clustering methods on the IMOX data set by utilizing the suggestion of Anderberg (1973) to cluster the clustering methods. Noteworthy features of their comparative study are delineated below:

1. Various types of evidence concerning the nature of the IMOX data set are presented, such as selected scatter plots;
2. A complete description of each clustering method employed in the empirical study is given, including practical considerations relevant to its computer implementation;
3. A complete summary of all results from the application of each clustering method to the IMOX data set is given, including the CPU time used, the number of clusters found, the number of patterns misclassified, and a cluster by category table;

4. Using Rand's (1971) statistic as a measure of similarity between clustering methods, a similarity matrix is derived to summarize the degree of similarity among the eight clustering methods with respect to the IMOX data set;
5. Two dendrograms are derived from the similarity (between clustering methods) matrix to determine which clustering methods really produced different results when applied to the IMOX data set;
6. Using one of the multidimensional scaling techniques, a one-dimensional comparison of the eight clustering methods is also provided.

The conclusions drawn by Dubes and Jain (1975) from their comparison of eight clustering methods are enlightening. For the IMOX data set, the complete linkage clustering represented the average of four different squared-error clusterings. The two clustering methods which are most dissimilar are both from the graph-theoretic class of clustering methods. Choosing a single clustering method from each of the three classes of clustering methods would not cover the gamut of possible clusterings for the IMOX data set. Finally, the two graph-theoretic clustering methods plus the complete linkage clustering method are sufficient to provide several alternative hypotheses about the structure of the IMOX data base.

Unfortunately, one recent trend in empirical, comparative studies involves the revival of the method of cophenetic correlation with non-parametric measures of correlation being substituted for the ordinary product moment correlation coefficient. The proponents of this "new" comparative method are, apparently, aware of the criticisms of the

cophenetic correlation coefficient as a measure of similarity between dendrograms given by Farris (1969). However, they also, apparently, missed, or at least ignored, Farris's (1969) overall skepticism concerning the method of cophenetic correlation itself. Some of the deficiencies attributable to the method of cophenetic correlation are functions of the methodology itself, which cannot be completely overcome by merely changing the measure of correlation. The method of cophenetic correlation is applicable only to hierarchical clustering methods; and more specifically, this method is used to compare agglomerative clustering method algorithms amongst themselves and with respect to the original similarity or dissimilarity matrix.

It should be recalled that for the purposes of this thesis, a clustering method was very carefully defined as consisting of two parts; and specifically, an agglomerative clustering method was characterized as consisting of some measure of distance, determining the original dissimilarity matrix, and an algorithm for recomputing distances after each join. The application of an agglomerative clustering method algorithm to a distance matrix imposes a hierarchy on the data set which may be conveniently visualized by means of a dendrogram. Typically, a dendrogram consists of a tree and a vertical scale of measurement which affords information on the distance at which the two clusters in clustering Y^K joined to form clustering Y^{K-1}; this distance will be called the joining distance for clustering Y^{K-1}. Initially, there are $N(N-1)/2$ distances associated with N data points, and these are reduced to $N - 1$ joining distances by the application of an agglomerative clustering method algorithm to the original distance matrix. Thus, summarizing a distance matrix by means of a dendrogram necessitates a loss

of information with respect to distances, but the purpose of cluster analyzing a set of data is to provide a summary of the data set which substantially reduces its proportions. A distance matrix is itself a summary of the data set; but even for small values of N , a distance matrix is difficult to assimilate. An agglomerative clustering method algorithm provides an interpretation for the distance matrix, which can be more easily assimilated.

From a philosophical point of view, it is important to consider the primary purpose for cluster analyzing a data set. The relevant question appears to be: Is the primary purpose of cluster analysis to describe the data points or to describe the distance matrix, which is assumed to be a "good" representation of the relationship between data points. The method of cophenetic correlation implicitly assumes that the initial distance matrix is the "best" summary of the relationships which exist among the data points. As a consequence, the comparison of clustering algorithms by means of the method of cophenetic correlation is not directly related to the data points or the sequence of clusterings; this comparative technique only considers how well a clustering algorithm represents the original distance matrix as depicted by the set of joining distances. For example, the cophenetic correlation coefficient for comparing a dendrogram resulting from the application of the single linkage algorithm with a dendrogram resulting from the application of the complete linkage algorithm can not be equal to one, (except in specially contrived cases) even when all clusterings in the hierarchy are exactly the same. Farris (1969, p. 284) comments on the cophenetic correlation coefficient (CPCC) as an optimality criterion as follows:

> The CPCC is a true measure of optimality of a classification only for a particular definition of taxonomic 'information.' Under the usual criterion that similar OTUs should be clustered together in a 'good' classification, the CPCC is not a direct measure of optimality of classifications. Further, the problem of finding the most appropriate optimality criterion for classifications will have to be considered jointly with the question of what is the most appropriate measure of 'similarity' between OTUs.

Thus, the practice of beginning a comparison of agglomerative clustering methods a step beyond the choice of a measure of similarity or dissimilarity is at best questionable.

Apparently, Cunningham and Ogilvie (1972) initiated the trend of comparing agglomerative clustering method algorithms by means of the method of cophenetic correlation in conjunction with a measure of rank correlation; for simplicity, this method will be referred to as the rank method of comparison. Theoretically, substituting a measure of rank correlation for the ordinary product moment correlation coefficient in the method of cophenetic correlation will alleviate the problem of the coefficient not accurately portraying the similarity in the sequence of clusterings. Now, supposedly, when the sequence of clusterings are the same in two different dendrograms (joining distances differ), the rank method will yield a coefficient of one. However, the reduction of the initial distance matrix to a set of joining distances gives rise to the mechanical problem of tied ranks, which represents a serious encumbrance to the rank method of comparison regardless of the rank correlation coefficient chosen.

As a justification for their methodology, Cunningham and Ogilvie (1972) define a perfect grouping as one which retains the information contained in the initial distance matrix, but this definition implicitly assumes that the initial distance matrix is a "correct" representation

of the structure present in the data set. They choose two goodness of fit measures, Kendall's (1948) tau (τ) which measures concordance in order relationship and a stress type measure which assesses agreement in absolute value, to quantify the amount of distortion imposed on the initial distance matrix by each of seven well-known agglomerative clustering method algorithms. Unfortunately, Cunningham and Ogilvie (1972) give no indication of the formula being used to compute τ, nor do they indicate that a correction has been made in the usual expression for τ to handle the mechanical problems associated with tied ranks. In fact, they make no reference to the existence of tied ranks. Both Baker (1974) and Hubert (1974) indicate that τ is not an appropriate measure of rank correlation in the presence of tied ranks because it does not have a probabilistic interpretation when tied ranks occur. It should be noted that if Cunningham and Ogilvie (1972) used Kendall's (1938) tau as originally defined with no correction for tied ranks to compare the clustering algorithms to the initial distance matrix, then many of the values of τ appearing in their tabled results can be shown to be unattainable. Further, Cunningham and Ogilvie (1972, p. 213) allude to a possible deficiency in the rank method of comparison when their measure of stress is chosen as the goodness of fit criterion in the following statement:

> Computed distances, unlike average distances, are not necessarily in the same range as the input distances, and therefore can inflate the value of stress.

Cunningham and Ogilvie (1972) may also be credited with initiating another trend in recent empirical, comparative studies. The construction of test data sets that are artificially contrived to represent

certain types of ideal structure in an attempt to reveal fundamental differences between clustering methods appears to be a new approach to comparing clustering methods. Cunningham and Ogilvie (1972, p. 210) give the following rationale for basing a comparative study on artificially contrived distance matrices:

> Several sets of data were tried out in an attempt to find if there are distinguishable 'types' of data which fit into a hierarchical structure in a characteristic way.

The ideal data set concept provides an interesting approach to comparing clustering methods, which is continued by Baker (1974) and by Hubert (1974). However, artificially contrived data sets necessitate a comparative study of a more limited scope than the usual Monte Carlo approach to generating data sets would permit. There are no replications in the empirical, comparative study reported by Cunningham and Ogilvie (1972). Finally, they also used their overall framework (ideal data sets and rank method of comparison) to explore robustness against random permutation and robustness against random perturbation of the chosen agglomerative clustering method algorithms.

Baker (1974) presents an "improved" version of the "robustness against random perturbation" investigation originated by Cunningham and Ogilvie (1972). Baker's (1974) empirical, comparative study suffers from an artificial quality which makes it difficult to relate his results to the data analyst's problem of choosing a clustering method. For example, there is no data in his comparative study, only basal taxonomies representing ideal data structures (such as a completely chained structure). An "error-free" matrix of ranks, the initial rank matrix, is derived from each of three basal taxonomies such that the application of either the single linkage or the complete linkage clustering

algorithm to the initial rank matrix recreates the original basal taxonomy. It should be noted that both the single linkage and the complete linkage clustering algorithms require only an ordinal scale of measurement for their application. However, since the ordinal scale of measurement is fundamental to Baker's (1974) comparative study, it is not generalizable to other agglomerative clustering method algorithms.

Baker's (1974) objective is to compare the single linkage and the complete linkage clustering algoirthms with respect to their sensitivity to random perturbation of the data. However, there is no data to which random error may be added. Instead, Baker (1974) adds random perturbations (by a seemingly complex scheme) to each entry of the initial rank matrix. Although he has three different levels of random error, it is very difficult to visualize the different levels of perturbation of the ranks as relating to different degrees of perturbation at the variable level. Instead, a higher level of perturbation of ranks may be merely an indication of additional variables being used to describe each data point.

In Baker's (1974) empirical, comparative investigation, each of the perturbed rank matrices is clustered by the single linkage and the complete linkage clustering algorithms. The resultant hierarchies are compared to the basal taxonomy by means of the rank method of comparison in conjunction with the Goodman and Kruskal (1954) gamma coefficient as an alternative goodness of fit measure to Kendall's (1938) tau coefficient. Although, the gamma coefficient retains a probabilistic interpretation even in the presence of tied ranks, there is still a considerable loss of information resulting from the tied ranks. Paradoxically, the gamma coefficient probably attains its highest values, when the greatest

amount of information is lost due to tied ranks. This observation might partially account for the following conclusions alluded to by Baker (1974):

1. The single linkage clustering algorithm is more sensitive to random perturbation of the ranks than is the complete linkage clustering algorithm;
2. A completely chained data structure is more easily obscured by a fixed level of random perturbation of the ranks than are the other two types of data structure employed in this comparative study.

Hubert (1974), like Baker (1974), is concerned with the single linkage and the complete linkage clustering algorithms and the concept of "noise." Hubert (1974) also employs the basic framework developed by Baker (1974), i.e., initial rank matrix and gamma coefficient as a measure of goodness of fit. However, Hubert (1974) explicitly bases his empirical comparative study on Ling's (1973) assumption that every permutation of the object pairs has an equal chance of occurring; and thus, he proceeds to randomly select with replacement from the set of all possible permutations of the object pairs from an initial rank matrix. This assumption appears to be a very poor basis for an empirical study because for a fixed p-dimensional Euclidean space, a large proportion of the set of all possible permutations may be geometrically impossible. It is analogous to assuming that the data points come from an infinite dimensional space.

A simple example will aptly depict the inappropriateness of assuming that every possible permutation of the object pairs is equally likely to occur at least from a geometric point of view. For $N = 4$

(distinct) data points, there are $N(N-1)/2 = 6$ ranks in the initial rank matrix. For these six ranks, there exist 720 possible permutations of the ranks. It can be easily shown that in one-dimensional Euclidean space (i.e., on a line) 5/6 or 600 of the 720 possible permutations are geometrically impossible. Let it suffice to pose the question: Would these 600 impossible cases produce high values of the gamma coefficient? The main difficulty, however, lies in trying to interpret Hubert's (1974) comparative study in an applied sense without a "real" world context.

For the purposes of this thesis, the empirical, comparative study reported by Rand (1969, 1971) is of primary importance. Chapter V of this thesis represents an extension of one aspect of the empirical studies reported in a thesis by Rand (1969) and in a subsequent journal article by Rand (1971), which summarized and supplemented the original thesis. Consequently, an extensive critical review of Rand's (1969, 1971) comparative studies will be given with additions and possible extensions being noted. Rand's major contribution to the problem of comparing clustering methods is a statistic, c, which measures the similarity between pairs of clusterings; the c statistic is the subject of Chapter IV of this thesis.

Rand (1969, 1971) uses the measure of similarity between clusterings, c, to investigate four relevant questions in a series of Monte Carlo studies, reporting the sample mean of c, the sample standard deviation of c, and the percentage of complete agreement for each case considered. The four fundamental aspects of clustering methods proposed by Rand (1971, p. 848) are exemplified by the following questions:

1. "How well does a method retrieve 'natural' clusters?"
2. "How sensitive is a method to perturbation of the data?"

3. "How sensitive is a method to missing individuals?"
4. "Given two methods, do they produce different results when applied to the same data?"

Chapter V of this thesis is primarily concerned with the "retrieval" ability of agglomerative clustering methods for particular types of structure.

Without intending to be critical of Rand's empirical studies, the following criticisms and comments should be noted as indications of possible extensions and as indications of factors not considered, which could make a comparative study of clustering methods more "dynamic" and more meaningful to the data analyst:

1. The clustering methods compared by Rand are not well-known clustering methods, and they appear to be poor for the purpose of "retrieval" and computationally inefficient.
2. For all of the Monte Carlo studies except that of "retrieval," he generated all of the data points from a single distribution.
3. For the "retrieval" study, he generated the same number of points from each population.
4. Rand did not attempt to relate the distance between populations to the "retrieval" ability of the clustering methods.
5. The only measures of similarity or dissimilarity considered by Rand were forms of Euclidean distance.
6. All of the multivariate normal data was generated from populations having an identity variance-covariance matrix.
7. More use could be made of the fact that c is a valid statistic for comparing clusterings even when the clusterings contain different numbers of clusters.

The main point of the observations given above is that Rand's empirical, comparative studies could be naturally extended by the controlled change of a wider range of contextual variables. However, the concept of comparing clustering methods based on the clusterings produced rather than the joining distances seems more relevant to the objectives of cluster analysis from a practical point of view.

Some Reflections

The literature of cluster analysis, obviously, suffers from fragmentation due to its diverse evolution. Consequently, the lack of a standard nomenclature for cluster analysis concepts, even, resists attempts to edit the discussion of cluster analysis research to provide a consistent exposition of the literature. Very simply, with respect to the same concept, subtle differences of meaning, as reflected by the diverse terminology, exist across fields of application. In summary, since the primary purpose of this thesis is "to compare," some reflections on the philosophical basis for comparing clustering methods appear to be necessary before proceeding to a discussion of the present research effort.

The conclusions from an empirical study are necessarily embedded in some context (initial specifications and underlying assumptions) or parameter space, whether this fact is explicitly acknowledged or not. The infelicitous aspect of empirical, comparative studies which begin with an initial distance or rank matrix rather than an initial set of data points is that the aforementioned procedure effectively causes the context to be unknown; i.e., certain, important control parameters are inestimable. Regardless of the level at which an empirical, comparative

investigation is begun, it is not independent of contextual variables or control parameters as they will be referred to in this discussion. Instead, failure to specify the necessary control parameters renders the results uninterpretable in an applied sense. The consequences of this general discussion for the comparison of agglomerative clustering methods is worth considering.

From Chapter I, it should be recalled that the resolution of a clustering problem by the application of an agglomerative clustering method to a data set can be described by the triple (X, H, m). The object space, X, and the clustering method, m, are elements of the parameter space which require specification, initially, and the hierarchy, H, is the resultant sequence of clusterings for the specified pair (X, m). Since X is essentially specified by N, the number of data points, and p, the dimension of the Euclidean space in which the object space is embedded, and since m is specified by the pair (measure of distance, clustering algorithm), the parameter space may be completely specified by the quadruple (N, p, measure of distance, clustering algorithm). The specification of all four of these parameters is required for the application of an agglomerative clustering method to a set of data points, and all conclusions concerning the resultant hierarchy are dependent on these initial specifications.

When agglomerative clustering algorithms are compared based only on an initial distance or rank matrix being generated without the existence of data points <u>per se</u>, then only the pair (N, clustering algorithm) is specified to obtain the sequence of clusterings. The parameter pair (p, measure of distance) is left undefined, and these control parameters are, essentially, inestimable or unrecoverable. However, conclusions

concerning H are not independent of the parameters p and measure of distance. Instead, conclusions concerning H are based on one pair of unknown control parameters and one pair of known control parameters. H exists only for some unknown subset of the set of all possible choices of the pair (p, measure of distance), and the possibility of this subset being empty cannot be theoretically eliminated. If this subset is nonempty, recovery of a parameter pair (p, measure of distance) may be accomplished by showing that the initial distance or rank matrix is obtainable from the parameter triple (N, p, measure of distance). Thus, the validity of any conclusions concerning the relative merits of the agglomerative clustering algorithms being compared is difficult to assess when the empirical, comparative study is based on an initial distance or rank matrix without reference to a set of data points.

The necessity to specify all four members of the quadruple (N, p, measure of distance, clustering algorithm) places a serious restriction on the generalizations which may be made from an empirical, comparative investigation of agglomerative clustering methods. It should be noted that generalizations of empirical, comparative studies conducted in p-space, are not necessarily valid for any other choice of p; i.e., generalization to either a higher or a lower dimensional Euclidean space is usually not possible. It is also quite possible that there is an interaction between the measure of distance and the clustering algorithm. At least, both members of the pair defining the agglomerative clustering method contribute to the process which produces the dendrogram, and varying either member of this pair may produce a different sequence of clusterings for a particular data set. In conclusion, the further removed an empirical study, within the realm of cluster analysis, is

from the data analyst's problems alluded to in Chapter I, the less viable is the research.

CHAPTER III

THE PROPERTIES OF AN INFINITE SET OF AGGLOMERATIVE CLUSTERING ALGORITHMS

A General Formulation for Agglomerative Clustering Algorithms

For the purposes of this chapter, the application of an agglomerative clustering method to a set of data requires that a measure of distance, d, be imposed on the object space, X. Thus, the properties and some examples of distance measures will be established before giving a general formulation for agglomerative clustering algorithms.

In very general terms, a measure of distance, d, on some arbitrary set, S, is a real-valued function on $S \times S$. In particular, some of the relevant properties which a measure of distance may possess will be given with respect to the object space, X. However, these properties may apply to an arbitrarily defined measure of distance on any set.

Letting d_{ij} denote the distance between data point X_i and data point X_j, the hierarchy of properties for a measure of distance is aptly depicted in Definitions 5, 6, and 7.

Definition 5. A semi-metric on the object space, X, is a function,

$$d: X \times X \longrightarrow R ,$$

such that the following two properties hold for every pair of data points, X_i and X_j, in X:

(i) d is a strictly positive function, i.e.,

$$\forall\ X_i, X_j \in X, \qquad d_{ij} \geq 0$$

$$\text{and} \qquad d_{ij} = 0 \ \text{ iff } \ X_i = X_j \ ;$$

(ii) d is a symmetric function, i.e.,

$$\forall\ X_i, X_j \in X, \qquad d_{ij} = d_{ji} \ .$$

Definition 6. A metric on the object space, X, is a semi-metric d such that the following third property also holds for every X_i, X_j, and X_k in X:

(iii) d satisfies the triangle inequality, i.e.,

$$\forall\ X_i, X_j, X_k \in X,$$

$$d_{ik} \leq d_{ij} + d_{jk} \ .$$

Definition 7. An ultrametric (Johnson, 1967) on the object space, X, is a metric d such that the following fourth property also holds for every X_i, X_j, and X_k in X:

(iv) d satisfies the ultrametric inequality, i.e.,

$$\forall\ X_i, X_j, X_k \in X,$$

$$d_{ik} \leq \max \{d_{ij}, d_{jk}\} \ .$$

The ultrametric inequality is a stronger property than the triangle inequality. Thus, if the ultrametric inequality holds for a measure of distance on X, then the triangle inequality necessarily holds for that measure of distance on X. It is also worth noting that an ultrametric measure of distance is invariant to all monotonic transformations of d. A metric measure of distance, however, is not, in general, invariant to monotonic transformations of the measure of distance because the triangle inequality is not preserved under all monotonic transformations of d.

It should be noted that for the derivations presented in this chapter, only a semi-metric measure of distance is required as a basis for the initial distance matrix.

A well-known family of distance measures for which the metric properties hold is the family of Minkowski metrics. The $m^{\underline{th}}$ member of the family of Minkowski metrics will be designated by ℓ_m. Recalling that X_i is a p-component vector, if x_{iv} denotes the $v^{\underline{th}}$ component of data point X_i and x_{jv} denotes the $v^{\underline{th}}$ component of data point X_j, then the $m^{\underline{th}}$ Minkowski metric between data points X_i and X_j is computed by the following formula:

$$\ell_m(X_i, X_j) = \left[\sum_{v=1}^{p} |x_{iv} - x_{jv}|^m \right]^{1/m} ,$$

where $m \geq 1$.

Euclidean distance is a member of the family of Minkowski metrics, namely, ℓ_2. However, squared Euclidean distance (in common use with some agglomerative clustering algoirthms) is only a semi-metric measure of distance, since the triangle inequality is not preserved under the operation of squaring distances.

From this brief background on measures of distance, the general formulation for agglomerative clustering algorithms given by Lance and Williams (1966) can be presented in a notation consistent with the present development. First, however, with respect to an agglomerative clustering method, some subtle distinctions, concerning the set on which d is a measure of distance, are necessitated.

In the application of an agglomerative clustering method to a set of data, initially, the distance between each pair of data points, X_i and X_j, is computed using some measure of distance, d, which is at

least semi-metric. Since d is at least semi-metric, the resultant set of distances may be denoted by

$$D = \{d_{ij} \mid i < j,\ i = 1,2,\ldots,N-1,\ j = 2,3,\ldots,N\} \quad .$$

A convenient device for displaying D is the distance matrix $D_{N,N}$, where only the $N(N-1)/2$ upper triangular elements of $D_{N,N}$ are necessary.

Therefore, d is a measure of distance on X. However, the set of single-point clusters, Y^N, corresponds to X. Consequently, d is also a measure of distance on Y^N, where an element of Y^N is a cluster, Y_i, corresponding to data point X_i. Hence, the process of clustering a set of data by means of an agglomerative clustering method is initiated by viewing the measure of distance on X as a measure of distance on Y^N; and thereby, D becomes the set of all distances between pairs of clusters in Y^N.

The role of the agglomerative clustering algorithm is to sequentially impose a measure of distance on each clustering, Y^K, $K = 1, 2, \ldots, N-1$, in the hierarchy such that the measure of distance imposed on Y^K is functionally related to the measure of distance imposed on Y^{K+1}. In a sense, d is not the same measure of distance on Y^K and on Y^{K+1} (i.e., on two clusterings of different sizes). In fact, even when d is initially a metric, for some clustering in the hierarchy, d may not even be semi-metric, and this anomalous situation will be illustrated in the next section.

To clarify the notation, since Y^K, $K = 1, 2, \ldots, N$, is a set of clusters, a measure of distance may be imposed on Y^K, and d_{ij} shall now be used to denote the distance between cluster Y_i and cluster Y_j,

where $Y_i, Y_j \in Y^K$, $K = 1, 2, \ldots, N$. This is not inconsistent since in the case of Y^N, X_i and Y_i correspond. Thus, the distance between data points is a special case of the distance between clusters, and this distance between data points will be used to initiate a recursive algorithm for the recomputation of distance between clusters after each joining of two clusters. As a further simplification of the notation, if $Y_i, Y_j \in Y^K$ join at distance d_{ij} to form Y^{K-1}, then $Y_{(ij)}$ will denote the new cluster, i.e.,

$$Y_{(ij)} = Y_i \cup Y_j \quad ,$$

and d_{ij} shall be termed the joining distance for clustering Y^{K-1}.

Using the notation of this section, the general linear combinatorial strategy originally presented by Lance and Williams (1966) is given as Equation (3.1), and it represents a family of agglomerative clustering algorithms. For any clustering Y^K in the hierarchy, if the distances d_{ij}, d_{ik}, and d_{jk} between pairs of clusters are obtained from some source (recursively from clustering Y^{K+1}, $K \neq N$), then the distance between the new cluster $Y_{(ij)}$ and any other cluster $Y_k \in Y^K$ can be computed from the following formula:

$$d_{(ij)k} = \alpha_i d_{ik} + \alpha_j d_{jk} + \beta d_{ij} + \gamma |d_{ik} - d_{jk}| \quad , \qquad (3.1)$$

where: α_i, α_j, β, and γ are specified parameters, defining the particular member of the family of agglomerative clustering algorithms.

Beginning with the initial distance matrix obtained by imposing d on X, Equation (3.1) is applied recursively to obtain each clustering in the hierarchy.

The objective of this chapter is to explore the properties of $d_{(ij)k}$ under a particular set of constraints. To motivate the choice of "interesting" properties, a brief discussion of the consequences of particular choices of the parameter values in Equation (3.1) will be given in the next section.

Some Examples of the Consequences of Arbitrary Parameter Choices

Equation (3.1) characterizes a family of agglomerative clustering algorithms so that for each choice of the parameter quadruple $(\alpha_i, \alpha_j, \beta, \gamma)$, a particular member of this family of agglomerative clustering algorithms is specified. In this section, two parameter quadruples will be specified, and the resultant algorithms will be applied to a set of Euclidean distances, D, derived from a small set of generated data points. Since, initially, the measure of distance being used is Euclidean distance, d is a metric on X. However, the triangle inequality is not necessarily preserved under the application of an agglomerative clustering algorithm to D.

Figure 1 gives the six bivariate normal data points and the Euclidean distance between each pair of data points, conveniently displayed in a two-way table. The first three data points, X_1, X_2, and X_3, were generated to simulate a random sample from a bivariate normal population with a mean vector given by $\mu' = (0, 5)$ and a variance-covariance matrix given by the identity matrix. The last three data points, X_4, X_5, and X_6, were generated to simulate a random sample from a bivariate normal population with a mean vector given by $\mu' = (0, 0)$ and a variance-covariance matrix given by the identity matrix. It should be

$X_1' = (-.333, 4.634)$

$X_2' = (-.728, 3.929)$

$X_3' = (.664, 5.800)$

$X_4' = (-.342, -.985)$

$X_5' = (1.491, 1.078)$

$X_6' = (.222, .453)$

a) Six Bivariate Normal Data Points

	1	2	3	4	5	6
1	0.0	.808	1.535	5.619	3.997	4.217
2		0.0	2.332	4.929	3.613	3.603
3			0.0	6.850	4.794	5.365
4				0.0	2.759	1.545
5					0.0	1.415
6						0.0

b) The Euclidean Distance Between Each Pair of Single-point Clusters or between Each Pair of Data Points

Figure 1. The Generated Data and an Initial Distance Matrix for the Examples

noted from Figure 1b that data points X_1 and X_2 are "closest" since $d_{1,2} = .808$ is the smallest distance in D. As a consequence, clusters Y_1 and Y_2 will join first, regardless of the choice of algorithm, and

their joining distance will be $d_{1,2} = .808$ (it is circled in Figure 1b because it is the first joining distance). Hence, clustering Y^5 is obtained from clustering Y^6 by replacing clusters Y_1 and Y_2 by cluster $Y_{(12)} = Y_1 \cup Y_2$. Before proceeding to clustering Y^4, distances from cluster $Y_{(12)}$ to each of the other clusters must be obtained, but this requires the specification of a particular member of the family of agglomerative clustering algorithms.

Choosing $\alpha_i = 1/2$, $\alpha_j = 1/2$, $\beta = -1/2$, and $\gamma = -1$, then

$$\begin{aligned} d_{(ij)k} &= (1/2)d_{ik} + (1/2)d_{jk} - (1/2)d_{ij} - |d_{ik} - d_{jk}| \\ &= (1/2)d_{ik} + (1/2)d_{jk} - (1/2)d_{ij} - \max\{d_{ik}, d_{jk}\} \\ &\qquad + \min\{d_{ik}, d_{jk}\} \\ &= (3/2)\min\{d_{ik}, d_{jk}\} - (1/2)\max\{d_{ik}, d_{jk}\} - (1/2)d_{ij} \end{aligned} \tag{3.2}$$

Derived from the recursive application of Equation (3.2) to the sets of distances, Figure 2 depicts the sequence of clusterings and their associated sets of distances, conveniently displayed in two-way tables. The joining distance for each successive clustering is circled in each set of distances. It should be observed that the sequence of joining distances is not monotone increasing, which is a somewhat undesirable situation, especially when a dendrogram is to be used to portray the hierarchy. It is also interesting to observe that for the set of distances obtained after the first join (in Figure 2a), the triangle inequality no longer holds for all choices of three clusters. For example,

$$d_{(ij)k} = (3/2)\min\{d_{ik}, d_{jk}\} - (1/2)\max\{d_{ik}, d_{jk}\} - (1/2)d_{ij}$$

	1,2	3	4	5	6
1,2	0.0	.7325	4.180	3.017	2.892
3		0.0	6.859	4.794	5.365
4			0.0	2.759	1.545
5				0.0	1.415
6					0.0

a) Distances after First Join

	1,2,3	4	5	6
1,2,3	0.0	2.474	1.762	1.289
4		0.0	2.759	1.545
5			0.0	1.415
6				0.0

b) Distances after Second Join

	1,2,3,6	4	5
1,2,3,6	0.0	.436	.597
4		0.0	2.759
5			0.0

c) Distances after Third Join

	1,2,3,6,4	5
1,2,3,6,4	0.0	-.702
5		0.0

d) Distance at Which Last Join Is Made

Figure 2. Example 1 Concerning the Consequences of the Parameter Quadruple (1/2, 1/2, -1/2, -1)

$$d_{(12)3} = .7325 ,$$
$$d_{(12)4} = 4.18 ,$$
$$d_{3,4} = 6.859 ,$$

but

$$d_{(12)3} + d_{(12)4} = 4.9125 \neq 6.859 = d_{3,4} .$$

The ultimate consequence of choosing the parameter quadruple (1/2, 1/2, -1/2, -1), however, is that the final joining distance (Figure 2d) is negative, which is a highly undesirable characteristic for a distance between two clusters to have.

A second example using the same generated data set and the same resultant set of Euclidean distances, which are given in Figure 1, as used for the first example will demonstrate some of the consequences which may occur when the sequence of joining distances is monotone increasing. Choosing the parameter quadruple (3/4, 3/4, -1/2, 1/2), then

$$\begin{aligned} d_{(ij)k} &= (3/4)d_{ik} + (3/4)d_{jk} - (1/2)d_{ij} + (1/2)|d_{ik} - d_{jk}| \\ &= (3/4)d_{ik} + (3/4)d_{jk} - (1/2)d_{ij} + (1/2)\max\{d_{ik}, d_{jk}\} \\ &\qquad - (1/2)\min\{d_{ik}, d_{jk}\} \\ &= (5/4)\max\{d_{ik}, d_{jk}\} + (1/4)\min\{d_{ik}, d_{jk}\} - (1/2)d_{ij} \end{aligned} \tag{3.3}$$

Derived from the recursive application of Equation (3.3) to the sets of distances, Figure 3 depicts the sequence of clusterings and their associated sets of distances, conveniently displayed in two-way tables. As in Figure 2, the joining distance for each successive clustering is circled in each set of distances. It should be noted that the sequence of joining distances is monotone increasing, which is a desirable

$$d_{(ij)k} = (5/4)\max\{d_{ik}, d_{jk}\} + (1/4)\min\{d_{ik}, d_{jk}\} - (1/2)d_{ij}$$

	1,2	3	4	5	6
1,2	0.0	2.895	7.852	5.496	5.768
3		0.0	6.859	4.794	5.365
4			0.0	2.759	1.545
5				0.0	(1.415)
6					0.0

a) Distances after First Join

	1,2	5,6	3	4
1,2	0.0	7.876	(2.895)	7.852
5,6		0.0	7.197	3.128
3			0.0	6.859
4				0.0

b) Distances after Second Join

	1,2,3	5,6	4
1,2,3	0.0	10.197	10.082
5,6		0.0	(3.128)
4			0.0

c) Distances after Third Join

	1,2,3	4,5,6
1,2,3	0.0	(13.704)
4,5,6		0.0

d) Distance at Which Last Join Is Made

Figure 3. Example 2 Concerning the Consequences of the Parameter Quadruple (3/4, 3/4, -1/2, 1/2)

characteristic for a sequence of joining distances to possess. However, as in Example 1, even for the set of distances obtained after the first join (in Figure 3a), the triangle inequality does not hold for every possible choice of three clusters. For example,

$$d_{(12)4} = 7.852 ,$$
$$d_{(12)6} = 5.768 ,$$
$$d_{4,6} = 1.545 ,$$

but

$$d_{(12)6} + d_{4,6} = 7.313 \not\geq 7.852 = d_{(12)4} .$$

The ultimate consequence of choosing the parameter quadruple $(3/4, 3/4, -1/2, 1/2)$, however, is that the final joining distance is approximately twice as large as the largest initial distance, which surely indicates that some type of distortion is being perpetrated on the initial set of distances by the application of this member of the family of agglomerative clustering algorithms to the sets of distances.

In Figure 2, the sequence of clusterings provides an example of complete chaining as each single-point cluster in turn joins $Y_{(12)}$. In Figure 3, however, the sequence of clusterings provides an example of the direct opposite to complete chaining, i.e., the case where at each join the tendency is to form equal-sized clusters. Thus, two quite different hierarchies are derived from the same set of data by specifying two different members of the family of agglomerative clustering algorithms. Lance and Williams (1966) made the following similar observation concerning the consequences of arbitrarily choosing parameter quadruples for Equation (3.1):

> The extent of clustering is thus not an inherent property of data; a given set of data may now, by varying the parameters, be made to appear as sharply clustered as a user may desire.

Therefore, it seems relevant to study the properties of the sequence of distances, $d_{(ij)k}$, as a means to exploring the amount of distortion which might result from the application of an agglomerative clustering method to a set of data.

A Two Parameter Sub-Family of Agglomerative Clustering Algorithms

A two parameter sub-family of agglomerative clustering algorithms may be derived from the four parameter family by placing a suitable set of constraints on the parameters given in Equation (3.1). If the constraints are given by

$$\alpha_i = \alpha_j = \alpha \ ,$$

$$\alpha_i + \alpha_j + \beta = 1 \ ,$$

then a member of the four parameter family of agglomerative clustering algorithms that has parameter values which satisfy the constraints can be represented by the ordered pair (β, γ).

Without loss of generality, it will be assumed that

$$d_{ij} < d_{ik} < d_{jk} \ .$$

Noting that the two constraints imply that

$$\alpha_i = \alpha_j = \frac{1 - \beta}{2} \ ,$$

then Equation (3.1) becomes

$$d_{(ij)k} = \frac{1 - \beta}{2} d_{ik} + \frac{1 - \beta}{2} d_{jk} + \beta d_{ij} + \gamma |d_{ik} - d_{jk}| \ .$$

Since

$$d_{ij} < d_{ik} < d_{jk} \quad ,$$

then

$$d_{(ij)k} = \frac{1 - \beta + 2\gamma}{2} d_{jk} + \frac{1 - \beta - 2\gamma}{2} d_{ik} + \beta d_{ij} \quad . \qquad (3.4)$$

Thus, Equation (3.4) characterizes a sub-family of agglomerative clustering algorithms which shall be referred to as the (β, γ) family, and each member of this sub-family shall be referred to as a (β, γ) algorithm. Consequently, it is possible to represent each member of the (β, γ) family of agglomerative clustering algorithms as a point in the (β, γ) Cartesian coordinate plane. It is also worth noting that single linkage, complete linkage, unweighted average linkage, and the flexible strategy given by Lance and Williams (1967) are members of the (β, γ) family of agglomerative clustering algorithms.

If

$$D^*_{(\beta, \gamma)} = \{d_{(ij)k} \text{ at } (\beta, \gamma) | d_{ij} < d_{ik} < d_{jk}\} \quad ,$$

then the essential properties to consider for (β, γ) algorithms are given by Definitions 8, 9, 10, and 11.

Definition 8. A (β, γ) algorithm is monotone increasing iff for each $d_{(ij)k} \in D^*_{(\beta, \gamma)}$, $d_{(ij)k} > d_{ij}$.

Definition 9. A (β, γ) algorithm is space-conserving iff for each $d_{(ij)k} \in D^*_{(\beta, \gamma)}$, $d_{ik} < d_{(ij)k} < d_{jk}$.

Definition 10. A (β, γ) algorithm is space-contracting iff

$$g.\ell.b.(D^*_{(\beta, \gamma)}) \leq d_{ik} \quad .$$

Definition 11. A (β, γ) algorithm is space-dilating iff

$$\ell.u.b.(D^*_{(\beta, \gamma)}) \geq d_{jk} \quad .$$

It is of interest to explore the properties of $D^{*}_{(\beta, \gamma)}$ over various regions of the (β, γ) plane, and this investigation will be presented in its entirety in the next section.

A Study of the Properties of the (β, γ) Family of Agglomerative Clustering Algorithms

The regions of the (β, γ) plane investigated in this study originate in a natural way as a part of the overall development. The three primary boundary lines result from considering the values of the parameters for which each coefficient in Equation (3.4) is equal to zero. Hence, the following points are relevant:

(i) $\frac{1 - \beta + 2\gamma}{2} = 0$, if $\gamma = \frac{\beta - 1}{2}$;

(ii) $\frac{1 - \beta - 2\gamma}{2} = 0$, if $\gamma = \frac{1 - \beta}{2}$;

(iii) $\beta = 0$ on the γ-axis .

The seven regions to be investigated in this study are shown in Figure 4.

Region I is defined by the intersection of the following inequalities:

(i) $0 < \beta < 1$,

(ii) $\frac{\beta - 1}{2} < \gamma < \frac{1 - \beta}{2}$.

The boundary lines for Region I shall be labeled as follows:

A. $\beta = 0$ & $(\beta - 1)/2 < \gamma < (1 - \beta)/2$;

B. $\gamma = (1 - \beta)/2$ & $0 < \beta < 1$;

C. $\gamma = (\beta - 1)/2$ & $0 < \beta < 1$.

The three vertices of the triangular Region I are worthy of separate

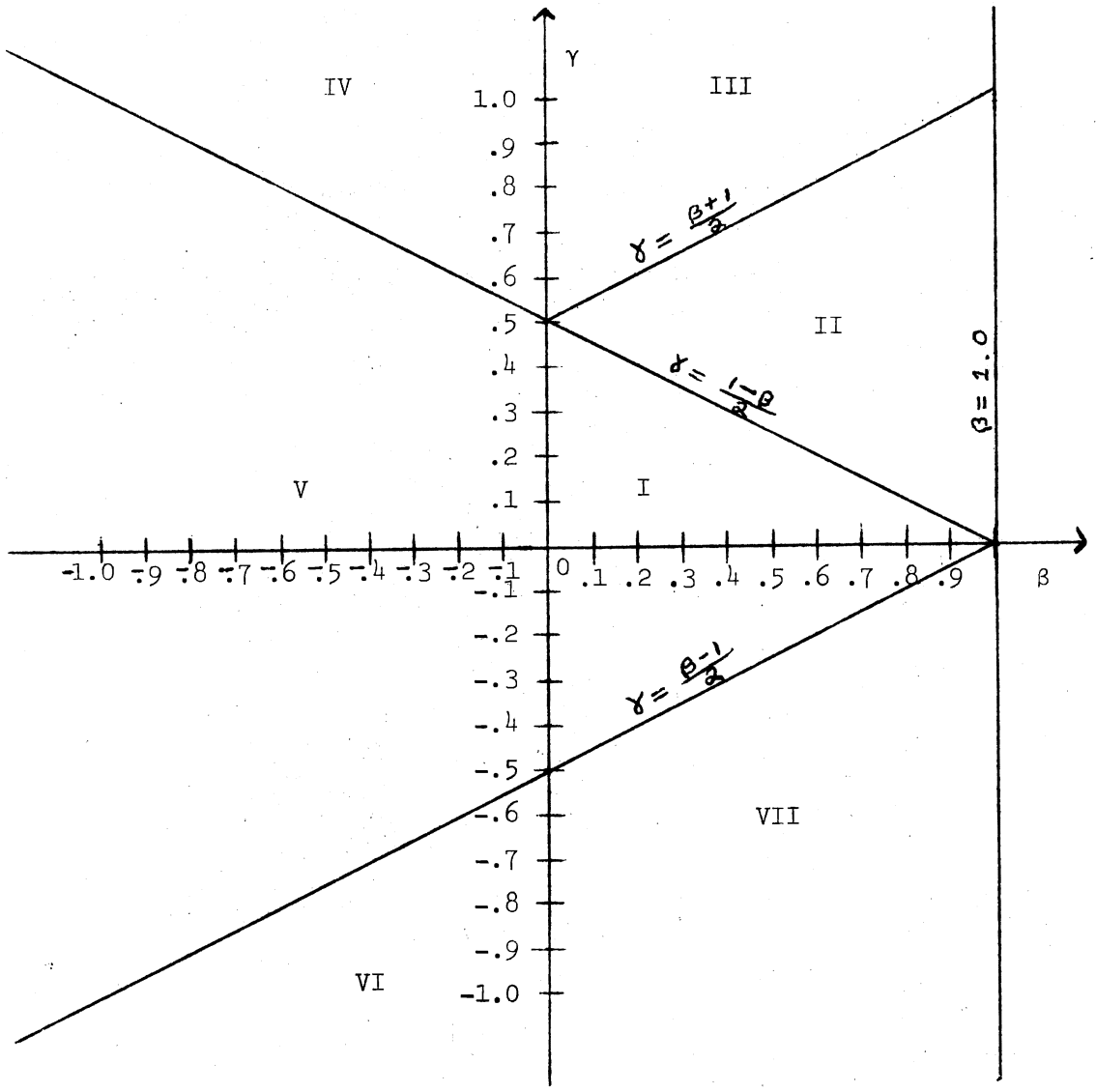


Figure 4. The Seven Regions of the (β, γ) Plane

consideration before exploring the properties of Region I and its boundary lines.

The point (0.0, .5) represents the complete linkage algorithm; and for this algorithm,

$$d_{(ij)k} = d_{jk} \quad .$$

The point (0.0, -.5) represents the single linkage algorithm; and for this algorithm,

$$d_{(ij)k} = d_{ik} \quad .$$

The point (1.0, 0.0) designates an algorithm for which

$$d_{(ij)k} = d_{ij} \quad .$$

The properties of the algorithms lying along the boundary lines for Region I will be considered before the properties of the algorithms lying inside Region I are considered. Since $\beta = 0$ along Boundary A,

$$d_{(ij)k} = \frac{1 + 2\gamma}{2} d_{jk} + \frac{1 - 2\gamma}{2} d_{ik} \quad . \tag{3.5}$$

An upper bound for $D^*_{(\beta, \gamma)}$ along Boundary A results from adding the positive number,

$$\frac{1 - 2\gamma}{2} (d_{jk} - d_{ik}) \quad ,$$

to the right side of Equation (3.5). Therefore,

$$d_{(ij)k} < \frac{1 + 2\gamma}{2} d_{jk} + \frac{1 - 2\gamma}{2} d_{jk} = d_{jk} \quad .$$

A lower bound for $D^*_{(\beta, \gamma)}$ along Boundary A results from adding the negative number,

$$\frac{1 + 2\gamma}{2} (d_{ik} - d_{jk}) \quad ,$$

to the right side of Equation (3.5). Therefore,

$$d_{(ij)k} > \frac{1 + 2\gamma}{2} d_{ik} + \frac{1 - 2\gamma}{2} d_{ik} = d_{ik} \quad .$$

Consequently, for each $d_{(ij)k} \in D^*_{(\beta, \gamma)}$, where $\beta = 0$ and $(\beta - 1)/2 < \gamma < (1 - \beta)/2$,

$$d_{ik} < d_{(ij)k} < d_{jk} \quad ,$$

and therefore the set of agglomerative clustering algorithms defined by Boundary A is a set of space-conserving algorithms.

Since $\gamma = (1 - \beta)/2$ along Boundary B,

$$d_{(ij)k} = (1 - \beta)d_{jk} + \beta d_{ij} \quad . \tag{3.6}$$

An upper bound for $D^*_{(\beta, \gamma)}$ along Boundary B may be derived by adding the positive number,

$$\beta(d_{jk} - d_{ij}) \quad ,$$

to the right side of Equation (3.6). Thus,

$$d_{(ij)k} < (1 - \beta)d_{jk} + \beta d_{jk} = d_{jk} \quad .$$

A lower bound for $D^*_{(\beta, \gamma)}$ along Boundary B results from adding the negative number,

$$(1 - \beta)(d_{ij} - d_{jk}) \quad ,$$

to the right side of Equation (3.6). Therefore,

$$d_{(ij)k} > (1 - \beta)d_{ij} + \beta d_{ij} = d_{ij} \quad .$$

Consequently, for each $d_{(ij)k} \in D^*_{(\beta, \gamma)}$, where $\gamma = (1 - \beta)/2$ and $0 < \beta < 1$,

$$d_{ij} < d_{(ij)k} < d_{jk} \quad .$$

Since the only other potentially interesting lower bound for $D^*_{(\beta,\ \gamma)}$ along Boundary B is given by

$$d_{(ij)k} > (1 - \beta)d_{jk}$$

and since for each $0 < \beta < 1$, there exists $d_{ik} < d_{jk}$ such that

$$(1 - \beta)d_{jk} < d_{ik} \quad ,$$

then along Boundary B,

$$g.\ell.b.(D^*_{(\beta,\ \gamma)}) \leq d_{ik} \quad .$$

Hence, the set of agglomerative clustering algorithms defined by Boundary B is a set of space-contracting algorithms.

Since $\gamma = (\beta - 1)/2$ along Boundary C,

$$d_{(ij)k} = (1 - \beta)d_{ik} + \beta d_{ij} \quad . \qquad (3.7)$$

An upper bound for $D^*_{(\beta,\ \gamma)}$ along Boundary C may be derived by adding the positive number,

$$\beta(d_{ik} - d_{ij}) \quad ,$$

to the right side of Equation (3.7). As a result,

$$d_{(ij)k} < (1 - \beta)d_{ik} + \beta d_{ik} = d_{ik} \quad .$$

A lower bound for $D^*_{(\beta,\ \gamma)}$ along Boundary C may be derived by adding the negative number,

$$(1 - \beta)(d_{ij} - d_{ik}) \quad ,$$

to the right side of Equation (3.7). Hence,

$$d_{(ij)k} > (1 - \beta)d_{ij} + \beta d_{ij} = d_{ij} \quad .$$

Therefore, for each $d_{(ij)k} \in D^*_{(\beta,\ \gamma)}$, where $\gamma = (\beta - 1)/2$ and $0 < \beta < 1$,

$$d_{ij} < d_{(ij)k} < d_{ik} \quad .$$

Since the only other potentially interesting lower bound for $D^*_{(\beta,\ \gamma)}$ along Boundary C is given by

$$d_{(ij)k} > (1 - \beta)d_{ik}$$

and since for each $0 < \beta < 1$, $(1 - \beta)d_{ik} < d_{ik}$, then along Boundary C,

$$g.\ell.b.(D^*_{(\beta,\ \gamma)}) \leq d_{ik} \quad .$$

Thus, the set of agglomerative clustering algorithms defined by Boundary C is a set of space-contracting algorithms.

To derive the properties for the algorithms lying inside Region I, Equation (3.4) must be considered. An upper bound for $D^*_{(\beta,\ \gamma)}$ inside Region I may be derived by adding the positive number,

$$\frac{1 - \beta - 2\gamma}{2}(d_{jk} - d_{ik}) + \beta(d_{jk} - d_{ij}) \quad ,$$

to the right side of Equation (3.4). Therefore,

$$d_{(ij)k} < \frac{1 - \beta + 2\gamma}{2} d_{jk} + \frac{1 - \beta - 2\gamma}{2} d_{jk} + \beta d_{jk} = d_{jk} \quad .$$

A lower bound for $D^*_{(\beta,\ \gamma)}$ inside Region I results from adding the negative number,

$$\frac{1 - \beta + 2\gamma}{2}(d_{ij} - d_{jk}) + \frac{1 - \beta - 2\gamma}{2}(d_{ij} - d_{ik}) \quad ,$$

to the right side of Equation (3.4). Hence,

$$d_{(ij)k} > \frac{1 - \beta + 2\gamma}{2} d_{ij} + \frac{1 - \beta - 2\gamma}{2} d_{ij} + \beta d_{ij} = d_{ij} \quad .$$

Consequently, for each $d_{(ij)k} \in D^*_{(\beta, \gamma)}$, where $0 < \beta < 1$ and $(\beta - 1)/2 < \gamma < (1 - \beta)/2$,

$$d_{ij} < d_{(ij)k} < d_{jk} \quad .$$

Although there are other possible lower bounds for $D^*_{(\beta, \gamma)}$ inside Region I, in a manner similar to that used for Boundary A, it can be shown that inside Region I

$$g.\ell.b.(D^*_{(\beta, \gamma)}) \leq d_{ik} \quad .$$

Consequently, the set of agglomerative clustering algorithms defined by Region I is a set of space-contracting algorithms.

Region II is defined by the intersection of the following inequalities:

(i) $0 < \beta < 1$,

(ii) $\frac{1 - \beta}{2} < \gamma < \frac{\beta + 1}{2}$.

The boundary lines for Region II shall be labeled as follows:

B. $\gamma = (1 - \beta)/2$ & $0 < \beta < 1$;

D. $\gamma = (\beta + 1)/2$ & $0 < \beta < 1$;

E. $\beta = 1$ & $(1 - \beta)/2 < \gamma \leq (\beta + 1)/2$.

The properties of the algorithms lying along the boundary lines for Region II will be considered before the properties of the algorithms lying inside Region II are considered; and since Boundary B was discussed in conjunction with Region I, the discussion of the properties of the algorithms lying along the boundary lines of Region II will begin with Boundary D.

Since $\gamma = (\beta + 1)/2$ along Boundary D,

$$d_{(ij)k} = d_{jk} - \beta d_{ik} + \beta d_{ij} \quad . \qquad (3.8)$$

An upper bound for $D^*_{(\beta,\ \gamma)}$ along Boundary D may be derived by adding the positive number,

$$\beta(d_{ik} - d_{ij}),$$

to the right side of Equation (3.8). Thus,

$$d_{(ij)k} < d_{jk} - \beta d_{ij} + \beta d_{ij} = d_{jk} \quad .$$

A lower bound for $D^*_{(\beta,\ \gamma)}$ along Boundary D may be derived by adding the negative number,

$$(d_{ij} - d_{jk}) + \beta(d_{ik} - d_{ij}) \quad ,$$

to the right side of Equation (3.8). Hence,

$$d_{(ij)k} > (1 - \beta)d_{ij} + \beta d_{ij} = d_{ij} \quad .$$

Consequently, for each $d_{(ij)k} \in D^*_{(\beta,\ \gamma)}$, where $\gamma = (\beta + 1)/2$ and $0 < \beta < 1$,

$$d_{ij} < d_{(ij)k} < d_{jk} \quad .$$

Since the only other potentially interesting lower bounds for $D^*_{(\beta,\ \gamma)}$ along Boundary D are given by

$$d_{(ij)k} > (1 - \beta)d_{jk} > (1 - \beta)d_{ik}$$

and since for each $0 < \beta < 1$, there exists $d_{ik} < d_{jk}$ such that

$$(1 - \beta)d_{ik} < (1 - \beta)d_{jk} < d_{ik} \quad ,$$

then along Boundary D,

$$g.\ell.b.(D^*_{(\beta,\ \gamma)}) \leq d_{ik} \quad .$$

Therefore, the set of agglomerative clustering algorithms defined by Boundary D is a set of space-contracting algorithms.

Since $\beta = 1$ along Boundary E,

$$d_{(ij)k} = \gamma d_{jk} - \gamma d_{ik} + d_{ij} \quad . \tag{3.9}$$

An upper bound for $D^*_{(\beta, \gamma)}$ along Boundary E may be derived by adding the positive number,

$$(d_{ik} - d_{ij}) + (1 - \gamma)(d_{jk} - d_{ik}) \quad ,$$

to the right side of Equation (3.9). As a consequence,

$$d_{(ij)k} < \gamma d_{jk} + (1 - \gamma) d_{jk} = d_{jk} \quad .$$

A lower bound for $D^*_{(\beta, \gamma)}$ along Boundary E may be derived by adding the negative number,

$$\gamma(d_{ik} - d_{jk}) \quad ,$$

to the right side of Equation (3.9). Thus,

$$d_{(ij)k} > \gamma d_{jk} - \gamma d_{jk} + d_{ij} = d_{ij} \quad .$$

Consequently, for each $d_{(ij)k} \in D^*_{(\beta, \gamma)}$, where $\beta = 1$ and $(1 - \beta)/2 < \gamma \leq (\beta + 1)/2$,

$$d_{ij} < d_{(ij)k} < d_{jk} \quad .$$

Since the only other possible lower bounds for $D^*_{(\beta, \gamma)}$ along Boundary E are smaller than d_{ij} and since $d_{ij} < d_{ik}$, then along Boundary E,

$$g.\ell.b.(D^*_{(\beta, \gamma)}) \leq d_{ik} \quad .$$

Therefore, the set of agglomerative clustering algorithms defined by Boundary E is a set of space-contracting algorithms.

To derive the properties for the algorithms lying inside Region II, Equation (3.4) must be considered. An upper bound for $D^*_{(\beta, \gamma)}$ inside Region II may be derived by adding the positive number,

$$\beta(d_{ik} - d_{ij}) + \frac{1 + \beta - 2\gamma}{2} (d_{jk} - d_{ik}) \quad ,$$

to the right side of Equation (3.4). Hence,

$$d_{(ij)k} < \frac{1 - \beta + 2\gamma}{2} d_{jk} + \frac{1 + \beta - 2\gamma}{2} d_{jk} = d_{jk} \quad .$$

A lower bound for $D^*_{(\beta, \gamma)}$ inside Region II may be derived by adding the negative number,

$$\frac{1 - \beta + 2\gamma}{2} (d_{ik} - d_{jk}) + (1 - \beta)(d_{ij} - d_{ik}) \quad ,$$

to the right side of Equation (3.4). Hence,

$$d_{(ij)k} > (1 - \beta)d_{ij} + \beta d_{ij} = d_{ij} \quad .$$

Consequently, for each $d_{(ij)k} \in D^*_{(\beta, \gamma)}$, where $0 < \beta < 1$ and $(1 - \beta)/2 < \gamma < (\beta + 1)/2$,

$$d_{ij} < d_{(ij)k} < d_{jk} \quad .$$

The other potentially interesting lower bounds for $D^*_{(\beta, \gamma)}$ inside Region II are the same as the ones given for $D^*_{(\beta, \gamma)}$ along Boundary D. Therefore, it can be shown that inside Region II

$$g.\ell.b.(D^*_{(\beta, \gamma)}) \leq d_{ik} \quad .$$

Consequently, the set of agglomerative clustering algorithms defined by Region II is a set of space-contracting algorithms.

Region III is defined by the intersection of the following inequalities:

(i) $0 < \beta < 1$,

(ii) $\gamma > \frac{\beta + 1}{2}$.

The boundary lines for Region III shall be labeled as follows:

D. $\gamma = (\beta + 1)/2$ & $0 < \beta < 1$;

F. $\beta = 0$ & $\gamma > (\beta + 1)/2$;

G. $\beta = 1$ & $\gamma > (\beta + 1)/2$.

The properties of the algorithms lying along the boundary lines for Region III will be considered before the properties of the algorithms lying inside Region III are considered; and since Boundary D was discussed in conjunction with Region II, the discussion of the properties of the algorithms lying along the boundary lines of Region III will begin with Boundary F.

Since $\beta = 0$ along Boundary F,

$$d_{(ij)k} = \frac{1 + 2\gamma}{2} d_{jk} + \frac{1 - 2\gamma}{2} d_{ik} \quad . \tag{3.10}$$

An upper bound for $D^*_{(\beta, \gamma)}$ along Boundary F may be derived by adding the positive number,

$$\frac{2\gamma - 1}{2} d_{ik} \quad ,$$

to the right side of Equation (3.10). Thus,

$$d_{(ij)k} < \frac{1 + 2\gamma}{2} d_{jk} \quad .$$

A lower bound for $D^*_{(\beta, \gamma)}$ along Boundary F may be derived by adding the negative number,

$$\frac{2\gamma - 1}{2} (d_{ik} - d_{jk}) \quad ,$$

to the right side of Equation (3.10). Hence,

$$d_{(ij)k} > \frac{1 + 2\gamma}{2} d_{jk} + \frac{1 - 2\gamma}{2} d_{jk} = d_{jk} \quad .$$

Consequently, for each $d_{(ij)k} \in D^*_{(\beta, \gamma)}$, where $\beta = 0$ and $\gamma > (\beta + 1)/2$,

$$d_{jk} < d_{(ij)k} < \frac{1 + 2\gamma}{2} d_{jk} \quad .$$

Since any other upper bounds for $D^*_{(\beta, \gamma)}$ along Boundary F are larger than $((1 + 2\gamma)/2)d_{jk}$ and since for each $\gamma > .5$,

$$\frac{1 + 2\gamma}{2} d_{jk} > d_{jk} \quad ,$$

then along Boundary F,

$$\ell.u.b.(D^*_{(\beta, \gamma)}) \geq d_{jk} \quad .$$

Therefore, the set of agglomerative clustering algorithms defined by Boundary F is a set of space-dilating algorithms.

Since $\beta = 1$ along Boundary G,

$$d_{(ij)k} = \gamma d_{jk} - \gamma d_{ik} + d_{ij} \quad . \tag{3.11}$$

An upper bound for $D^*_{(\beta, \gamma)}$ along Boundary G results from adding the positive number,

$$\gamma d_{ik} - d_{ij} \quad ,$$

to the right side of Equation (3.11). Thus,

$$d_{(ij)k} < \gamma d_{jk} \quad .$$

A lower bound for $D^*_{(\beta, \gamma)}$ along Boundary G results from adding the negative number,

$$\gamma(d_{ik} - d_{jk}) \quad ,$$

to the right side of Equation (3.11). Thus,

$$d_{(ij)k} > d_{ij} \quad .$$

Consequently, for each $d_{(ij)k} \in D^*_{(\beta, \gamma)}$, where $\beta = 1$ and $\gamma > 1.0$,

$$d_{ij} < d_{(ij)k} < \gamma d_{jk} \quad .$$

To derive the properties for the algorithms lying inside Region III, Equation (3.4) must be considered. An upper bound for $D^*_{(\beta, \gamma)}$ inside Region III may be derived by adding the positive number,

$$\frac{\beta + 2\gamma - 1}{2} d_{ik} - \beta d_{ij} \quad ,$$

to the right side of Equation (3.4). Hence,

$$d_{(ij)k} < \frac{1 - \beta + 2\gamma}{2} d_{jk} \quad .$$

A lower bound for $D^*_{(\beta, \gamma)}$ inside Region III may be derived by adding the negative number,

$$\frac{\beta - 2\gamma - 1}{2} d_{jk} + \frac{\beta + 2\gamma - 1}{2} d_{ik} + (1 - \beta) d_{ij} \quad ,$$

to the right side of Equation (3.4). Hence,

$$d_{(ij)k} > (1 - \beta) d_{ij} + \beta d_{ij} = d_{ij} \quad .$$

Consequently, for each $d_{(ij)k} \in D^*_{(\beta, \gamma)}$, where $0 < \beta < 1$ and $\gamma > \frac{\beta + 1}{2}$,

$$d_{ij} < d_{(ij)k} < \frac{1 - \beta + 2\gamma}{2} d_{jk} \quad .$$

It can be shown that bounds for $D^*_{(\beta, \gamma)}$ inside Region III are such that an agglomerative clustering algorithm represented by a point inside Region III might be space-conserving, space-contracting, or space-dilating depending upon the relative magnitudes of d_{ij}, d_{ik}, and d_{jk}.

Region IV is defined by the intersection of the following inequalities:

(i) $\beta < 0$,

(ii) $\gamma > \frac{1 - \beta}{2}$.

The boundary lines for Region IV shall be labeled as follows:

F. $\beta = 0$ & $\gamma > (\beta + 1)/2$;

G. $\gamma = (1 - \beta)/2$ & $\beta < 0$.

The properties of the algorithms lying along the boundary lines for Region IV will be considered before the properties of the algorithms lying inside Region IV are considered; and since Boundary F was discussed in conjunction with Region III, the discussion of the properties

of the algorithms lying along the boundary lines of Region IV will consist of Boundary H.

Since $\gamma = (1 - \beta)/2$ along Boundary H,

$$d_{(ij)k} = (1 - \beta)d_{jk} + \beta d_{ij} \quad . \tag{3.12}$$

An upper bound for $D^*_{(\beta, \gamma)}$ along Boundary H may be derived by adding the positive number,

$$-\beta d_{ij} \quad ,$$

to the right side of Equation (3.12). As a consequence,

$$d_{(ij)k} < (1 - \beta)d_{jk} \quad .$$

A lower bound for $D^*_{(\beta, \gamma)}$ along Boundary H may be derived by adding the negative number,

$$-\beta(d_{ij} - d_{jk}) \quad ,$$

to the right side of Equation (3.12). As a consequence,

$$d_{(ij)k} > (1 - \beta)d_{jk} + \beta d_{jk} = d_{jk} \quad .$$

Consequently, for each $d_{(ij)k} \in D^*_{(\beta, \gamma)}$, where $\gamma = (1 - \beta)/2$ and $\beta < 0$,

$$d_{jk} < d_{(ij)k} < (1 - \beta)d_{jk} \quad .$$

Since any other upper bounds for $D^*_{(\beta, \gamma)}$ along Boundary H are larger than $(1 - \beta)d_{jk}$ and since for each $\beta < 0$,

$$(1 - \beta)d_{jk} > d_{jk} \quad ,$$

then along Boundary H,

$$\ell.u.b.(D^*_{(\beta, \gamma)}) \geq d_{jk} \quad .$$

Therefore, the set of agglomerative clustering algorithms defined by Boundary H is a set of space-dilating algorithms.

To derive the properties for the algorithms lying inside Region IV, Equation (3.4) must be considered. An upper bound for $D^*_{(\beta,\ \gamma)}$ inside Region IV may be derived by adding the positive number,

$$\frac{2\gamma + \beta - 1}{2} d_{ik} - \beta d_{ij} \quad ,$$

to the right side of Equation (3.4). Thus,

$$d_{(ij)k} < \frac{1 - \beta + 2\gamma}{2} d_{jk} \quad .$$

A lower bound for $D^*_{(\beta,\ \gamma)}$ inside Region IV may be derived by adding the negative number,

$$\frac{2\gamma + \beta - 1}{2} (d_{ik} - d_{jk}) - \beta(d_{ij} - d_{jk}) \quad ,$$

to the right side of Equation (3.4). Thus,

$$d_{(ij)k} > \frac{1 - \beta + 2\gamma}{2} d_{jk} + \frac{1 - \beta - 2\gamma}{2} d_{jk} + \beta d_{jk} = d_{jk} \quad .$$

Consequently, for each $d_{(ij)k} \in D^*_{(\beta,\ \gamma)}$, where $\beta < 0$ and $\gamma > \frac{1 - \beta}{2}$,

$$d_{jk} < d_{(ij)k} < \frac{1 - \beta + 2\gamma}{2} d_{jk} \quad .$$

Since any other upper bounds for $D^*_{(\beta,\ \gamma)}$ inside Region IV are larger than $((1 - \beta + 2\gamma)/2)d_{jk}$ and since for each $\gamma > (1 - \beta)/2$,

$$\frac{1 - \beta + 2\gamma}{2} d_{jk} > d_{jk} \quad ,$$

then inside Region IV,

$$\ell.u.b.(D^*_{(\beta,\ \gamma)}) \geq d_{jk} \quad .$$

Therefore, the set of agglomerative clustering algorithms defined by Region IV is a set of space-dilating algorithms.

Region V is defined by the intersection of the following inequalities:

(i) $\beta < 0$,

(ii) $\frac{\beta - 1}{2} < \gamma < \frac{1 - \beta}{2}$.

The boundary lines for Region V shall be labeled as follows:

A. $\beta = 0$ & $(\beta - 1)/2 < \gamma < (1 - \beta)/2$;

H. $\gamma = (1 - \beta)/2$ & $\beta < 0$;

J. $\gamma = (\beta - 1)/2$ & $\beta < 0$.

The properties of the algorithms lying along the boundary lines for Region V will be considered before the properties of the algorithms lying inside Region V are considered; and since Boundary A was discussed in conjunction with Region I and Boundary H was discussed in conjunction with Region IV, the discussion of the properties of the algorithms lying along the boundary lines of Region V shall consist of a discussion of Boundary J.

Since $\gamma = (\beta - 1)/2$ along Boundary J,

$$d_{(ij)k} = (1 - \beta)d_{ik} + \beta d_{ij} \quad . \tag{3.13}$$

An upper bound for $D^*_{(\beta, \gamma)}$ along Boundary J may be derived by adding the positive number,

$$-\beta d_{ij} \quad ,$$

to the right side of Equation (3.13). Hence,

$$d_{(ij)k} < (1 - \beta)d_{ik} \quad .$$

A lower bound for $D^*_{(\beta, \gamma)}$ along Boundary J may be derived by adding the negative number,

$$-\beta(d_{ij} - d_{ik}) \quad ,$$

to the right side of Equation (3.13). Hence,

$$d_{(ij)k} > (1 - \beta)d_{ik} + \beta d_{ik} = d_{ik} \quad .$$

Consequently, for each $d_{(ij)k} \in D^*_{(\beta, \gamma)}$, where $\beta < 0$ and $\gamma = (\beta - 1)/2$,

$$d_{ik} < d_{(ij)k} < (1 - \beta)d_{ik} \quad .$$

Since any other upper bounds for $D^*_{(\beta, \gamma)}$ along Boundary J are larger than $(1 - \beta)d_{ik}$ and since for each $\beta < 0$, there exists $d_{ik} < d_{jk}$ such that

$$(1 - \beta)d_{ik} > d_{jk} \quad ,$$

then along Boundary J,

$$\ell.u.b.(D^*_{(\beta, \gamma)}) \geq d_{jk} \quad .$$

Therefore, the set of agglomerative clustering algorithms defined by Boundary J is a set of space-dilating algorithms.

To derive the properties of the algorithms lying inside Region V, Equation (3.4) must be considered. An upper bound for $D^*_{(\beta, \gamma)}$ inside Region V may be derived by adding the positive number,

$$\frac{1 - \beta - 2\gamma}{2} (d_{jk} - d_{ik}) - \beta d_{ij} \quad ,$$

to the right side of Equation (3.4). Thus,

$$d_{(ij)k} < \frac{1 - \beta + 2\gamma}{2} d_{jk} + \frac{1 - \beta - 2\gamma}{2} d_{jk} = (1 - \beta)d_{jk} \quad .$$

A lower bound for $D^*_{(\beta, \gamma)}$ inside Region V may be derived by adding the negative number,

$$\frac{1 - \beta + 2\gamma}{2} (d_{ik} - d_{jk}) - \beta(d_{ij} - d_{ik}) \quad ,$$

to the right side of Equation (3.4). Thus,

$$d_{(ij)k} > \frac{1 - \beta + 2\gamma}{2} d_{ik} + \frac{1 - \beta - 2\gamma}{2} d_{ik} + \beta d_{ik} = d_{ik} \quad .$$

Consequently, for each $d_{(ij)k} \in D^*_{(\beta, \gamma)}$, where $\beta < 0$ and $(\beta - 1)/2 < \gamma < (1 - \beta)/2$,

$$d_{ik} < d_{(ij)k} < (1 - \beta)d_{jk} \quad .$$

Since any other upper bounds for $D^*_{(\beta, \gamma)}$ inside Region V are larger than $(1 - \beta)d_{jk}$ and since for each $\beta < 0$,

$$(1 - \beta)d_{jk} > d_{jk} \quad ,$$

then inside Region V,

$$\ell.u.b.(D^*_{(\beta, \gamma)}) \geq d_{jk} \quad .$$

Therefore, the set of agglomerative clustering algorithms defined by Region V is a set of space-dilating algorithms.

Region VI is defined by the intersection of the following inequalities:

(i) $\beta < 0$,

(ii) $\gamma < \frac{\beta - 1}{2}$.

The boundary lines for Region VI shall be labeled as follows:

J. $\gamma = (\beta - 1)/2$ & $\beta < 0$;

K. $\beta = 0$ & $\gamma < (\beta - 1)/2$.

The properties of the algorithms lying along the boundary lines for Region VI will be considered before the properties of the algorithms lying inside Region VI are considered; and since Boundary J was discussed in conjunction with Region V, the discussion of the properties of the algorithms lying along the boundary lines of Region VI shall consist of a discussion of Boundary K.

Since $\beta = 0$ along Boundary K,

$$d_{(ij)k} = \frac{1 + 2\gamma}{2} d_{jk} + \frac{1 - 2\gamma}{2} d_{ik} \quad . \qquad (3.14)$$

An upper bound for $D^*_{(\beta, \gamma)}$ along Boundary K may be derived by adding the positive number,

$$\frac{-1 - 2\gamma}{2} (d_{jk} - d_{ik}) \quad ,$$

to the right side of Equation (3.14). Hence,

$$d_{(ij)k} < \frac{1 + 2\gamma}{2} d_{ik} + \frac{1 - 2\gamma}{2} d_{ik} = d_{ik} \quad .$$

A lower bound for $D^*_{(\beta, \gamma)}$ along Boundary K may be derived by adding the negative number,

$$\frac{2\gamma - 1}{2} d_{ik} \quad ,$$

to the right side of Equation (3.14). Hence,

$$d_{(ij)k} > \frac{1 + 2\gamma}{2} d_{jk} \quad .$$

Consequently, for each $d_{(ij)k} \varepsilon D^*_{(\beta, \gamma)}$, where $\beta = 0$ and $\gamma < (\beta - 1)/2$,

$$\frac{1 + 2\gamma}{2} d_{jk} < d_{(ij)k} < d_{ik} \quad .$$

Since any other lower bounds for $D^*_{(\beta, \gamma)}$ along Boundary K are smaller than $((1 + 2\gamma)/2)d_{jk}$ and since for each $\gamma < -.5$,

$$\frac{1 + 2\gamma}{2} < 0.0 \quad ,$$

then along Boundary K, $\exists\, d_{(ij)k} \varepsilon D^*_{(\beta, \gamma)} \ni$

$$d_{(ij)k} < d_{ij} \quad .$$

Therefore, the set of agglomerative clustering algorithms defined by Boundary K is a set of algorithms which are not monotone increasing. It should also be noted that $D^*_{(\beta, \gamma)}$ along Boundary K can contain negative distances.

To derive the properties of algorithms lying inside Region VI, Equation (3.4) must be considered. An upper bound for $D^*_{(\beta, \gamma)}$ inside Region VI may be derived by adding the positive number,

$$\frac{1 - \beta + 2\gamma}{2} (d_{ik} - d_{jk}) - \beta d_{ij} \quad ,$$

to the right side of Equation (3.4). As a consequence,

$$d_{(ij)k} < \frac{1 - \beta + 2\gamma}{2} d_{ik} + \frac{1 - \beta - 2\gamma}{2} d_{ik} = (1 - \beta) d_{ik} \quad .$$

A lower bound for $D^*_{(\beta, \gamma)}$ inside Region VI may be derived by adding the negative number,

$$\frac{\beta + 2\gamma - 1}{2} d_{ik} - \beta d_{ij} \quad ,$$

to the right side of Equation (3.4). Thus,

$$d_{(ij)k} > \frac{1 - \beta + 2\gamma}{2} d_{jk} \quad .$$

Consequently, for each $d_{(ij)k} \in D^*_{(\beta, \gamma)}$, where $\beta < 0$ and $\gamma < (\beta - 1)/2$,

$$\frac{1 - \beta + 2\gamma}{2} d_{jk} < d_{(ij)k} < (1 - \beta) d_{ik} \quad .$$

Since any other lower bounds for $D^*_{(\beta, \gamma)}$ inside Region VI are smaller than $((1 - \beta + 2\gamma)/2) d_{jk}$ and since for each $\gamma < (\beta - 1)/2$,

$$\frac{1 - \beta + 2\gamma}{2} < 0.0 \quad ,$$

then inside Region VI, $\exists\ d_{(ij)k} \in D^*_{(\beta, \gamma)}\ \exists$

$$d_{(ij)k} < d_{ij} \quad .$$

Therefore, the set of agglomerative clustering algorithms defined by Region VI is a set of algorithms which are not monotone increasing, and the application of anyone of these algorithms to a set of metric distances may result in negative joining distances for some of the joins in the formation of the hierarchy.

Region VII is defined by the intersection of the following inequalities:

(i) $0 < \beta < 1$,

(ii) $\gamma < \frac{\beta - 1}{2}$.

The boundary lines for Region VII shall be labeled as follows:

C. $\gamma = (\beta - 1)/2$ & $0 < \beta < 1$;

K. $\beta = 0$ & $\gamma < (\beta - 1)/2$;

L. $\beta = 1$ & $\gamma < (\beta - 1)/2$.

The properties of the algorithms lying along the boundary lines for Region VII will be considered before the properties of the algorithms lying inside Region VII are considered; and since Boundary C was discussed in conjunction with Region I and Boundary K was discussed in conjunction with Region VI, the discussion of the properties of the algorithms lying along the boundary lines of Region VII shall consist of a discussion of Boundary L.

Since $\beta = 1$ along Boundary L ,

$$d_{(ij)k} = \gamma d_{jk} - \gamma d_{ik} + d_{ij} \quad . \tag{3.15}$$

An upper bound for $D^*_{(\beta, \gamma)}$ along Boundary L may be derived by adding the positive number,

$$-\gamma(d_{jk} - d_{ik}) \quad ,$$

to the right side of Equation (3.15). Hence,

$$d_{(ij)k} < d_{ij} \quad .$$

A lower bound for $D^*_{(\beta, \gamma)}$ along Boundary L may be derived by adding the negative number,

$$\gamma d_{ik} - d_{ij} \quad ,$$

to the right side of Equation (3.15). Hence,

$$d_{(ij)k} > \gamma d_{jk} \quad .$$

Consequently, for each $d_{(ij)k} \in D^*_{(\beta,\ \gamma)}$, where $\beta = 1$ and $\gamma < (\beta - 1)/2$,

$$\gamma d_{jk} < d_{(ij)k} < d_{ij} \quad .$$

Since any other lower bounds for $D^*_{(\beta,\ \gamma)}$ along Boundary L are smaller than γd_{jk} and since $\gamma < 0$, then along Boundary L,

$$\exists\, d_{(ij)k} \in D^*_{(\beta,\ \gamma)} \ni$$

$$d_{(ij)k} < d_{ij} \quad .$$

Therefore, the set of agglomerative clustering algorithms defined by Boundary L is a set of algorithms which are not monotone increasing. It should also be noted that $D^*_{(\beta,\ \gamma)}$ along Boundary L can contain negative distances.

To derive the properties of the algorithms lying inside Region VII, Equation (3.4) must be considered. An upper bound for $D^*_{(\beta,\ \gamma)}$ inside Region VII may be derived by adding the positive number,

$$\frac{1 - \beta + 2\gamma}{2}(d_{ik} - d_{jk}) + \beta(d_{ik} - d_{ij}) \quad ,$$

to the right side of Equation (3.4). Thus,

$$d_{(ij)k} < \frac{1 - \beta + 2\gamma}{2} d_{ik} + \frac{1 - \beta - 2\gamma}{2} d_{ik} + \beta d_{ik} = d_{ik} \quad .$$

A lower bound for $D^*_{(\beta,\ \gamma)}$ inside Region VII may be derived by adding the negative number,

$$\frac{\beta + 2\gamma - 1}{2} d_{ik} - \beta d_{ij} \quad ,$$

to the right side of Equation (3.4). Thus,

$$d_{(ij)k} > \frac{1 - \beta + 2\gamma}{2} d_{jk} \quad .$$

Consequently, for each $d_{(ij)k} \in D^*_{(\beta,\ \gamma)}$, where $0 < \beta < 1$ and $\gamma < (\beta - 1)/2$,

$$\frac{1 - \beta + 2\gamma}{2} d_{jk} < d_{(ij)k} < d_{ik} \quad .$$

Since any other lower bounds for $D^*_{(\beta, \gamma)}$ inside Region VII are smaller than $((1 - \beta + 2\gamma)/2)d_{jk}$ and since for each $\gamma < (\beta - 1)/2$,

$$\frac{1 - \beta + 2\gamma}{2} < 0.0 \quad ,$$

then inside Region VII, $\exists\ d_{(ij)k} \in D^*_{(\beta, \gamma)}\ \ni$

$$d_{(ij)k} < d_{ij} \quad .$$

Therefore, the set of agglomerative clustering algorithms defined by Region VII is a set of algorithms which are not necessarily monotone increasing. It should also be noted that $D^*_{(\beta, \gamma)}$ inside Region VII can contain negative distances.

The properties of the (β, γ) family of agglomerative clustering algorithms are summarized in Figure 5 and Figure 6. In Figure 5, a range of values for $D^*_{(\beta, \gamma)}$ inside each of the seven regions and along their boundary lines is given. In Figure 6, each of the seven regions is labeled according to Definitions 8, 9, 10, and 11.

Choosing the Agglomerative Clustering Algorithms for the Comparative Study

Initially, the objective of investigating the properties of the (β, γ) family of agglomerative clustering algorithms was to choose a "good" set of agglomerative clustering algorithms for the comparative study which is presented in Chapter V. Since a (β, γ) algorithm which is not monotone increasing also results in a $D^*_{(\beta, \gamma)}$ which may contain negative distances, then the (β, γ) algorithms from Regions VI and VII and Boundaries K and L were immediately eliminated from consideration for the comparative study. It remained to be determined whether

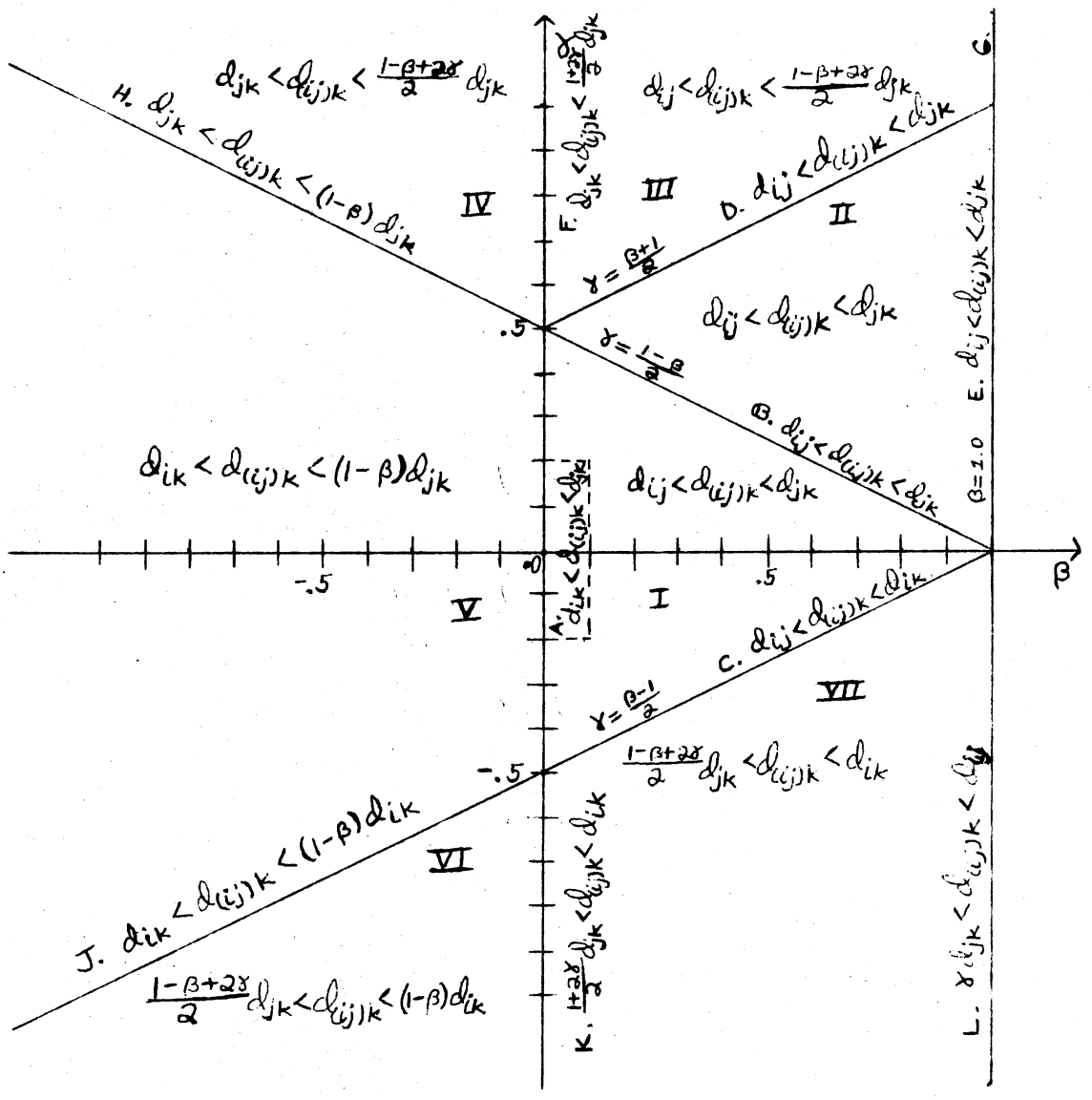


Figure 5. A Range of Values for $D^*_{(\beta,\ \gamma)}$ over Various Regions of the (β, γ) Plane

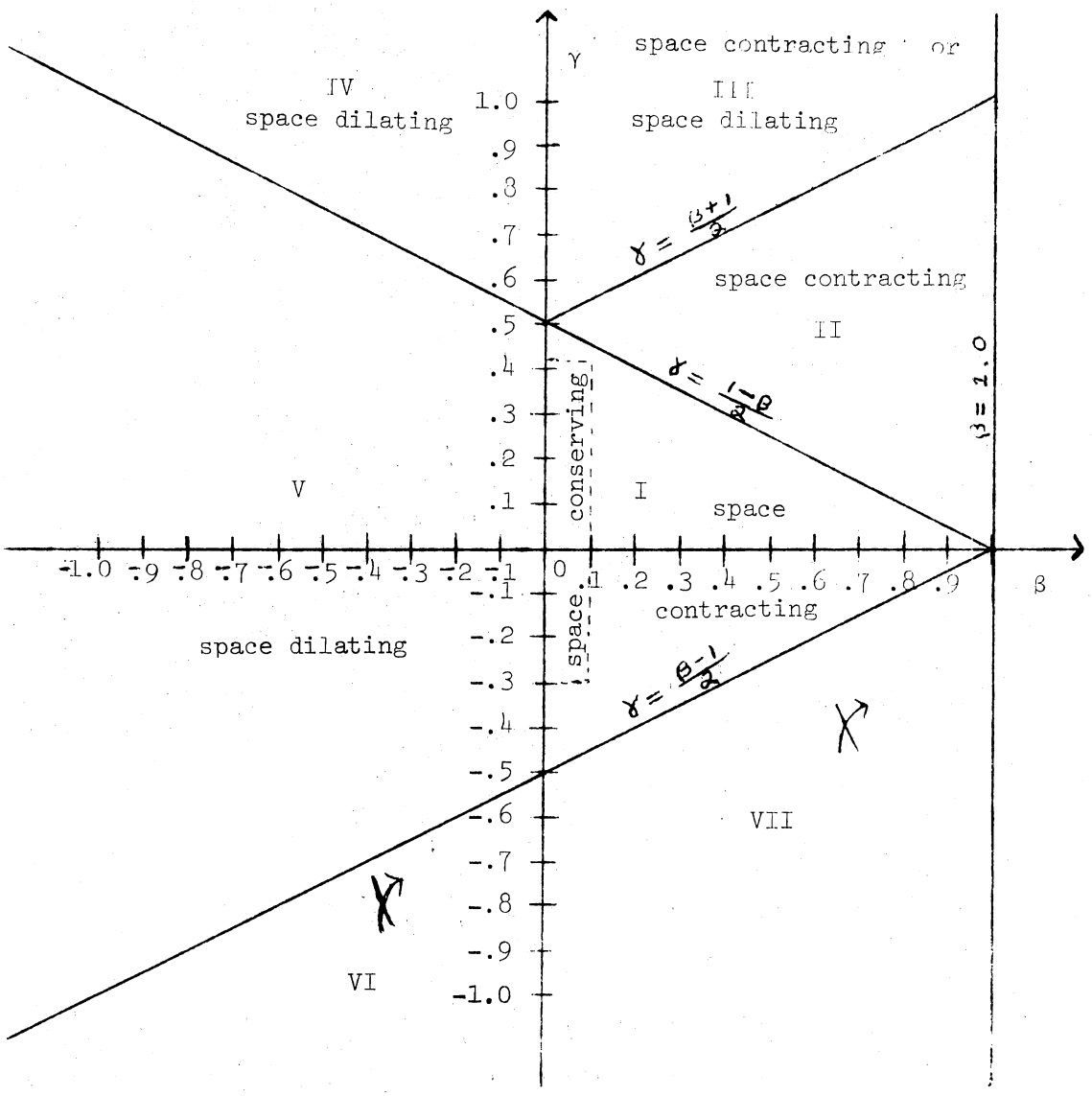


Figure 6. A Classification of the (β, γ) Family of Agglomerative Clustering Algorithms

space-conserving, space-contracting, or space-dilating algorithms produce "better" results when used in conjunction with the metric of Euclidean distance on multivariate normal data sets.

To further limit the set of agglomerative clustering algorithms being considered for the comparative study, a preliminary investigation was devised using multivariate normal data sets, Euclidean distance, and representative (β, γ) algorithms from each of the five remaining regions and from most of the remaining boundary lines. The following two important observations emanated from the preliminary investigation:

1. Algorithms which are close together in the (β, γ) plane produce very similar results when applied to the same set of distances;
2. Space-contracting algorithms produce relatively "poor" results with the metric of Euclidean distance on multivariate normal data sets.

Thus, the set of agglomerative clustering algorithms being considered for the comparative study was reduced to Regions IV and V and Boundaries A, F, H, and J by the preliminary investigation.

The final choice of the subset of the (β, γ) family of agglomerative clustering algorithms to be used in the comparative study was made by balancing the following objectives:

(i) Include all of the well-known algorithms from the (β, γ) family;

(ii) Include some space-conserving algorithms (Boundary A);

(iii) Include some space-dilating algorithms from both Region IV and Region V;

(iv) Include only (β, γ) algorithms which are relatively far apart in the (β, γ) plane;

(v) Include some (β, γ) algorithms from each side of the γ-axis.

Single linkage at the point (0.0, -.5) is a space-contracting algorithm, and complete linkage at the point (0.0, .5) is a space-dilating algorithm. It should be noted that the two points, (0.0, -.5) and (0.0, .5), are the endpoints of Boundary A which is the space-conserving region of the (β, γ) plane. Thus, single linkage and complete linkage are sometimes referred to as boundary algorithms, since the space-conserving algorithms lie between them along Boundary A. Average linkage at the point (0.0, 0.0) is a space-conserving algorithm. Hence, single linkage, complete linkage, and average linkage formed a basis for choosing six equally spaced algorithms along the γ-axis, which would satisfy all of the objectives except (iii). The six algorithms chosen are given in order from negative to positive along the γ-axis as follows:

(1.1) Single linkage at (0.0, -.5),

(1.2) (0.0, -.25),

(1.3) Average linkage at (0.0, 0.0),

(1.4) (0.0, .25),

(1.5) Complete linkage at (0.0, .5),

(1.6) (0.0, .75).

To determine a matching set of six algorithms in the space-dilating regions of the (β, γ) plane and thereby satisfying objective (iii) also, it was noted that the flexible strategy (Lance and Williams, 1967) is represented by the point (-.25, 0.0). Thus, it was decided to choose six equally spaced algorithms along the line $\beta = -.25$ such that this

second set of six points would be paired horizontally with the first set of six points. The six algorithms chosen are given in order from negative to positive along the line $\beta = -.25$ as follows:

(2.1) (-.25, -.5),

(2.2) (-.25, -.25),

(2.3) Flexible strategy at (-.25, 0.0),

(2.4) (-.25, .25),

(2.5) (-.25, .5),

(2.6) (-.25, .75).

To satisfy the five previously stated objectives, a set of twelve agglomerative clustering algorithms from the (β, γ) family was chosen for the comparative investigation which is presented in Chapter V. Before the comparative study is presented, however, a discussion of the comparative statistic to be employed in the comparative study will be presented in Chapter IV.

CHAPTER IV

A COMPARATIVE STATISTIC

Equivalent Forms of the Comparative Statistic

Since the primary objective of this thesis is to compare clustering methods, a comparative statistic is required to quantify each comparison of clustering methods. Rand's (1969, 1971) c statistic is a very general and versatile statistic which may be used to compare clustering methods based on how they partition the object space. Essentially, c measures the similarity between clusterings derived from any source. However, if two clusterings are produced by the application of two different clustering methods to the same object space, then c is a measure of the similarity between the two clustering methods through their resultant clusterings. As motivation for the comparisons presented in Chapter V, discussion of Rand's development of the c statistic is presented in this section.

Rand (1971, p. 847) makes the following three reasonable assumptions concerning the nature of a general clustering problem as a rationale for the development of the c statistic:

> First, clustering is discrete in the sense that every point is unequivocably assigned to a specific cluster. Second, clusters are defined just as much by those points which they do not contain as by those points which they do contain. Third, all points are of equal importance in the determination of clusterings.

Thus, Rand (1971) points out that a basic unit of comparison between two clusterings is how pairs of points are clustered.

To facilitate the definition of the c statistic, Definition 12 concerning the similar assignment of point-pairs is tendered.

Definition 12. Given an object space X consisting of N data points, $X_1, X_2, \ldots, X_N$, and two clusterings of X, $Y = \{Y_1, Y_2, \ldots, Y_{K_1}\}$ and $Y' = \{Y'_1, Y'_2, \ldots, Y'_{K_2}\}$, then a similar assignment in clusterings Y and Y' of a pair of data points, X_i and X_j, results if and only if either of the following two conditions holds:

(i) $\exists$ k and k' $\ni$ $X_i, X_j \in Y_k$ and $X_i, X_j \in Y'_{k'}$;

(ii) $\exists$ k and k' $\ni$ $X_i \in Y_k, Y'_{k'}$, and $X_j \notin Y_k, Y'_{k'}$.

Basically, if the elements of an individual point-pair are placed together in a cluster in each of two clusterings, or if they are assigned to different clusters in both clusterings, then a similar assignment of the point-pair has been made in the two clusterings. In essence, the c statistic gives a normalized count of the number of similar assignments of point-pairs between two clusterings as designated in Definition 13.

Definition 13. Given an object space X consisting of N data points, $X_1, X_2, \ldots, X_N$, and two clusterings of X, $Y = \{Y_1, Y_2, \ldots, Y_{K_1}\}$ and $Y' = \{Y'_1, Y'_2, \ldots, Y'_{K_2}\}$, then the c statistic between Y and Y' is defined as follows:

$$c(Y, Y') = \frac{\sum_{i<j} \eta_{ij}}{\binom{N}{2}} , \qquad (4.1)$$

where

$$\eta_{ij} = \begin{cases} 1, \text{ if there is a similar assignment of } X_i \text{ and } X_j \text{ in } Y \text{ and } Y', \\ 0, \text{ otherwise.} \end{cases}$$

Hence, c is a measure of similarity on $\mathcal{Y}$, the set of all possible clusterings of X.

Rand (1971) also gives a computational form for the c statistic, which is related to incidence matrix concepts. If the clusters within each clustering are arbitrarily numbered and if n_{ij} represents the number of data points which are simultaneously in the $i^{\underline{th}}$ cluster of Y and the $j^{\underline{th}}$ cluster of Y', then

$$c(Y, Y') = \frac{\binom{N}{2} - \frac{1}{2}\left[\sum_i \left(\sum_j n_{ij}\right)^2 + \sum_j \left(\sum_i n_{ij}\right)^2\right] + \sum_{i,j} n_{ij}^2}{\binom{N}{2}} . \tag{4.2}$$

Another formulation of Rand's c statistic is worth noting. According to Anderberg (1973), the c statistic is equivalent to the simple matching coefficient. The simple matching coefficient, which was originally introduced to numerical taxonomy by Sokal and Michener (1958), is a binary measure of association based on 2X2 contingency tables. To demonstrate the equivalence relationship between Rand's c statistic and the simple matching coefficient, a particular form of the simple matching coefficient will be developed.

The simple matching coefficient may be used to assess the amount of agreement between any two binary vectors of the same length, where a binary vector is defined in Definition 14.

Definition 14. A vector $V = (v_1, v_2, \ldots, v_n)$ is a binary vector if and only if for each $i = 1, 2, \ldots, n$, $v_i = 1$ or $v_i = 0$.

To compute the simple matching coefficient, it is necessary to define a match between two binary vectors as indicated in Definition 15.

Definition 15. A match between the corresponding components of two

binary vectors, $U = (u_1, u_2, \ldots, u_n)$ and $V = (v_1, v_2, \ldots, v_n)$, occurs if and only if either of the following two conditions hold:

(i) $u_i = 0$ & $v_i = 0$;

(ii) $u_i = 1$ & $v_i = 1$.

If the number of matches between two binary vectors of length n is denoted by m, then a definition for the simple matching coefficient is given as Definition 16.

Definition 16. The simple matching coefficient between two binary vectors, U and V, of length n is given by the following formula:

$$s(U, V) = \frac{m}{n} \quad , \tag{4.3}$$

where m is the number of matches.

Thus, the simple matching coefficient represents a normalized count of the number of matches between two binary vectors.

If a clustering can be represented as a binary vector, then a simple matching coefficient between clusterings can be computed. A binary representation of a clustering can be obtained by constructing a binary vector, U, consisting of $n = \binom{N}{2}$ components, where each component of U indicates whether a pair of data points are together or apart in the clustering. Letting X be an object space consisting of N data points, then a more precise formulization of a binary representation of a clustering is given in Definition 17.

Definition 17. The binary vector,

$$U = (u_{12}, u_{13}, \ldots, u_{1n}, u_{23}, \ldots, u_{2n}, \ldots, u_{n-1,n}) \quad ,$$

is a binary representation of clustering $Y = \{Y_1, Y_2, \ldots, Y_K\}$ if and only if for each $i < j$,

$$u_{ij} = \begin{cases} 1, & \text{if } \exists\, k \ni X_i, X_j \in Y_k \\ 0, & \text{otherwise} \end{cases} .$$

Therefore, if U is a binary representation of clustering Y and V is a binary representation of clustering Y', then

$$s(U, V) = \frac{m}{n} = \frac{m}{\binom{N}{2}} = \frac{\sum_{i<j} n_{ij}}{\binom{N}{2}} = c(Y, Y') .$$

Consequently, Rand's (1969, 1971) c statistic is equivalent to the simple matching coefficient.

The c statistic has the following three fundamental properties as noted by Rand (1969, 1971):

1. c is a measure of similarity with $0 \leq c \leq 1$;
2. $1 - c$ is a metric on the set of all possible clusterings of X ;
3. c is a random variable.

It should be noted that Rand (1969) provides a proof of the fact that $1 - c$ is a metric on $\mathcal{Y}$ in his thesis. Fundamental property 3 is the subject of the remainder of Chapter IV.

Since c is a random variable, under certain assumptions, c possesses a probability distribution. However, Rand (1969, p. 39) comments on the distribution of c as follows: "This is a complicated distribution and analytic expression of it is not attempted here." Logically, part of the complication with respect to the distribution of c concerns the choice of the space on which initial distributional assumptions should be placed. Conceptually, X is a subset of Euclidean p-space with cardinality N ; a clustering method maps X into $\mathcal{Y}$; and

$$c: \mathcal{Y} \times \mathcal{Y} \longrightarrow [0, 1] .$$

The present research effort includes some work on the distribution of the c statistic, and this effort is reported in the next section.

A Method for Deriving the Exact Distribution of the Comparative Statistic

Since both the c statistic and the number of matches, m, are discrete random variables on $\mathcal{Y} \times \mathcal{Y}$ and since m and c are proportional by a proportionality factor of $n = \binom{N}{2}$, then m and c have the same probability distribution under a fixed set of assumptions pertaining to $\mathcal{Y}$. Theoretically, given a fixed value of N, if a probability distribution for $\mathcal{Y}$ and a sampling scheme on $\mathcal{Y}$ are specified, then the probability distribution of c (or equivalently m) may be derived by a procedure which shall be referred to as the method of complete enumeration. In this section, under a reasonable set of assumptions which simulate the hypothetical phenomenon of obtaining clusterings from two random clustering methods, the method of complete enumeration is demonstrated for small values of N, and the exact probability distribution of the c statistic is given for $N = 3, 4$, and 5.

Letting L_N denote the cardinality of the set $\mathcal{Y}$ of all possible clusterings of object space X which consists of N data points, then the probability distributions of the c statistic are derived under the following two fundamental assumptions:

1. The clusterings $Y \in \mathcal{Y}$ have a discrete uniform distribution; i.e.,
$$\forall\, Y \in \mathcal{Y} \quad , \quad P(Y) = 1/L_N \quad .$$
2. The two clusterings, Y and Y', are drawn at random from $\mathcal{Y}$ with replacement.

Therefore, if the ordered pair (Y, Y') represents an element of $\mathcal{Y} \times \mathcal{Y}$, then
$$P[(Y, Y')] = (1/L_N)^2 \quad .$$

Case 1, N = 3

Figure 7 illustrates the method of complete enumeration for $N = 3$. Figure 7a presents the $L_3 = 5$ clusterings in $\mathcal{Y}$, which are arbitrarily labeled with a small letter to facilitate the derivation of the distribution of the number of matches. Figure 7b provides the binary representation of each clustering in $\mathcal{Y}$, where the vector length of each binary representation is $n = 3$. In Figure 7c, the distribution of the number of matches (conveniently displayed in a two-way table) for each pair of clusterings in $\mathcal{Y} \times \mathcal{Y}$ is given , where each clustering is identified by its arbitrary label.

K = 1	a.	$(X_1\ X_2\ X_3)$	a.	(1 1 1)
	b.	$(X_1\ X_2)\ (X_3)$	b.	(1 0 0)
K = 2	c.	$(X_1\ X_3)\ (X_2)$	c.	(0 1 0)
	d.	$(X_2\ X_3)\ (X_1)$	d.	(0 0 1)
K = 3	e.	$(X_1)\ (X_2)\ (X_3)$	e.	(0 0 0)

a) Clusterings b) Binary Representations

Y \ Y'	a	b	c	d	e
a	3	1	1	1	0
b	1	3	1	1	2
c	1	1	3	1	2
d	1	1	1	3	2
e	0	2	2	2	3

c) Number of Matches for Each Pair of Clusterings

Figure 7. For $N = 3$, the Set of All Possible Clusterings and the Distribution of the Number of Matches for Pairs of Clusterings from $\mathcal{Y}$

Recalling that

$$c(Y, Y') = \frac{m}{n} \quad , \tag{4.4}$$

then the distribution of the values of the c statistic for $N = 3$ is derived from the distribution of m by dividing each element in the two-way table given in Figure 7c by 3. Consequently, for $N = 3$, the probability mass function (p.m.f.) of the c statistic is given by the following expression:

$$f(c;\ N = 3,\ n = 3,\ L_3 = 5) = \begin{cases} 2/25 & , \text{ if } c = 0 \\ 12/25 & , \text{ if } c = 1/3 \\ 6/25 & , \text{ if } c = 2/3 \\ 5/25 & , \text{ if } c = 1 \\ 0 & , \text{ otherwise.} \end{cases}$$

Therefore, when $N = 3$,

$$E(c) = \frac{13}{25} = .52 \quad ;$$

$$VAR(c) = \frac{56}{625} = .0896 \quad .$$

Case 2, $N = 4$

Figures 8 and 9 Illustrate the method of complete enumeration for $N = 4$. Figure 8a presents the $L_4 = 15$ clusterings in $\mathcal{Y}$, which are arbitrarily labeled with a small letter to facilitate the derivation of the distribution of the number of matches. Figure 8b provides the binary representation of each clustering in $\mathcal{Y}$, where the vector length of each binary representation is $n = 6$. In Figure 9, the distribution of the number of matches (conveniently displayed in a two-way table) for each pair of clusterings in $\mathcal{Y} \times \mathcal{Y}$ is given, where each clustering is identified by its arbitrary label from Figure 8.

K = 1	a.	$(X_1$	X_2	X_3	$X_4)$	a.	1	1	1	1	1	1
K = 2	b.	$(X_1$	X_2	$X_3)$	(X_4)	b.	1	1	0	1	0	0
	c.	$(X_1$	X_2	$X_4)$	(X_3)	c.	1	0	1	0	1	0
	d.	$(X_1$	X_3	$X_4)$	(X_2)	d.	0	1	1	0	0	1
	e.	$(X_2$	X_3	$X_4)$	(X_1)	e.	0	0	0	1	1	1
K = 2	f.	$(X_1$	$X_2)$	$(X_3$	$X_4)$	f.	1	0	0	0	0	1
	g.	$(X_1$	$X_3)$	$(X_2$	$X_4)$	g.	0	1	0	0	1	0
	h.	$(X_1$	$X_4)$	$(X_2$	$X_3)$	h.	0	0	1	1	0	0
K = 3	i.	$(X_1$	$X_2)$	(X_3)	(X_4)	i.	1	0	0	0	0	0
	j.	$(X_1$	$X_3)$	(X_2)	(X_4)	j.	0	1	0	0	0	0
	k.	$(X_1$	$X_4)$	(X_2)	(X_3)	k.	0	0	1	0	0	0
	l.	$(X_2$	$X_3)$	(X_1)	(X_4)	l.	0	0	0	1	0	0
	m.	$(X_2$	$X_4)$	(X_1)	(X_3)	m.	0	0	0	0	1	0
	n.	$(X_3$	$X_4)$	(X_1)	(X_2)	n.	0	0	0	0	0	1
K = 4	o.	(X_1)	(X_2)	(X_3)	(X_4)	o.	0	0	0	0	0	0

a) Clusterings

b) Binary Representations

Figure 8. For $N = 4$, the Set of All Possible Clusterings of X

Thus, from Equation (4.4), it follows that the distribution of the values of the c statistic for $N = 4$ can be derived from the distribution of m by dividing each element in the two-way table given in Figure 9 by six. Consequently, for $N = 4$, the p.m.f. of the c statistic is given by the following expression:

Y \ Y'	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
a	6	3	3	3	3	2	2	2	1	1	1	1	1	1	0
b	3	6	2	2	2	3	3	3	4	4	2	4	2	2	3
c	3	2	6	2	2	3	3	3	4	2	4	2	4	2	3
d	3	2	2	6	2	3	3	3	2	4	4	2	2	4	3
e	3	2	2	2	6	3	3	3	2	2	2	4	4	4	3
f	2	3	3	3	3	6	2	2	5	3	3	3	3	5	4
g	2	3	3	3	3	2	6	2	3	5	3	3	5	3	4
h	2	3	3	3	3	2	2	6	3	3	5	5	3	3	4
i	1	4	4	2	2	5	3	3	6	4	4	4	4	4	5
j	1	4	2	4	2	3	5	3	4	6	4	4	4	4	5
k	1	2	4	4	2	3	3	5	4	4	6	4	4	4	5
l	1	4	2	2	4	3	3	5	4	4	4	6	4	4	5
m	1	2	4	2	4	3	5	3	4	4	4	4	6	4	5
n	1	2	2	4	4	5	3	3	4	4	4	4	4	6	5
o	0	3	3	3	3	4	4	4	5	5	5	5	5	5	6

Figure 9. For $N = 4$, the Distribution of the Number of Matches for Pairs of Clusterings from $\mathcal{Y}$

$$f(c;\ N = 4,\ n = 6,\ L_4 = 15) = \begin{cases} 2/225 & , \text{ if } c = 0 \\ 12/225 & , \text{ if } c = 1/6 \\ 48/225 & , \text{ if } c = 2/6 \\ 64/225 & , \text{ if } c = 3/6 \\ 60/225 & , \text{ if } c = 4/6 \\ 24/225 & , \text{ if } c = 5/6 \\ 15/225 & , \text{ if } c = 1 \\ 0 & , \text{ otherwise.} \end{cases}$$

Therefore, when $N = 4$,

$$E(c) = \frac{5}{9} = .5556 \quad ;$$

$$VAR(c) = \frac{19}{405} = .0469 \quad .$$

Case 3, N = 5

For $N = 5$, Figure 10 presents the binary representations for the $L_5 = 52$ clusterings in $\mathcal{Y}$, where the vector length of each binary representation is $n = 10$. From these binary representations, the distribution of m and thus, of c can be derived by applying the method of complete enumeration and by considering certain patterns and shortcuts learned from the previous cases.

Thus, for $N = 5$, the p.m.f. of the c statistic is given by the following expression:

$$f(c;\ N = 5,\ n = 10,\ L_5 = 52) = \begin{cases} 2/2704 & \text{if } c = 0 \\ 20/2704 & \text{if } c = 1/10 \\ 30/2704 & \text{if } c = 2/10 \\ 120/2704 & \text{if } c = 3/10 \\ 440/2704 & \text{if } c = 4/10 \\ 480/2704 & \text{if } c = 5/10 \\ 600/2704 & \text{if } c = 6/10 \\ 560/2704 & \text{if } c = 7/10 \\ 300/2704 & \text{if } c = 8/10 \\ 100/2704 & \text{if } c = 9/10 \\ 52/2704 & \text{if } c = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, when $N = 5$,

$$E(c) = \frac{1594}{2704} = .5895 \quad ;$$

$$VAR(c) = .02897 \quad .$$

The Cardinality of $\mathcal{Y}$

Theoretically, the method of complete enumeration could be applied for $N = 3, 4, 5, 6, 7, 8, \ldots$; and cumulative distribution function (C.D.F.) tables could be constructed. However, the cardinality of $\mathcal{Y}$

$$
\begin{pmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\
1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 \\
1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 1 \\
0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\
0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}
$$

Figure 10. For N = 5, the Binary Representations of y

increases rapidly thereby making the construction of C.D.F. tables for the c statistic costly with respect to computer time. As an indication of the counting problems related to deriving the probability distribution of the c statistic by the method of complete enumeration, a brief discussion of the cardinality of $\mathcal{Y}$ for specified values of N is relevant.

It should first be noted that any problem related to cluster analysis which requires the complete enumeration of all possible clusterings for a specified value of N as a part of its solution approaches practical impossibility in terms of numerical enormities for even relatively small values of N. In addition, for a specified value of N, the derivation of the probability distribution of the c statistic by the method of complete enumeration requires $(L_N)^2$ pairwise comparisons of the clusterings from $\mathcal{Y}$. However, the number of pairwise comparisons necessary to derive the probability distribution of the c statistic by the method of complete enumeration can be substantially reduced by noting that c is a symmetric function on $\mathcal{Y} \times \mathcal{Y}$, i.e.,

$$c(Y, Y') = c(Y', Y)$$

and that

$$c(Y, Y') = 1 \text{ if and only if } Y = Y' .$$

Therefore, only

$$\frac{L_N(L_N - 1)}{2}$$

pairwise comparisons of clusterings from $\mathcal{Y}$ are required to derive the probability distribution of the c statistic for a specified value of N.

From a practical point of view, it is the size of L_N which restricts the derivation of the probability distribution of the c statistic by the method of complete enumeration to "small" values of N

(e.g., $N = 3$, 4, and 5). Duran and Odell (1974) show that for each specification of N and K, the number of possible clusterings of size K, denoted by $S(N, K)$, is a Stirling number of the second kind. As a consequence,

$$L_N = \sum_{K=1}^{N} S(N, K) \quad .$$

Hence, the cardinality of $\mathcal{Y}$ for each specification of N is the sum of Stirling numbers of the second kind.

Computing Stirling numbers of the second kind is tedious. Duran and Odell (1974) prove that Stirling numbers of the second kind may be computed by the following formula:

$$S(N, K) = \frac{1}{K!} \sum_{j=0}^{K} \binom{K}{j} (-1)^j (K - j)^N \quad . \qquad (4.5)$$

By definition,

$$S(N, 0) = 0 \quad ;$$

$$S(N, N + i) = 0 \quad , \text{ if } \quad i > 0 \quad .$$

Obviously,

$$S(N, 1) = 1 \quad .$$

It can also be shown that

$$S(N, N) = 1 \quad .$$

In addition, Duran and Odell (1974) give the following recursive relationship between Stirling numbers of the second kind, which may be employed in deriving a table of Stirling numbers of the second kind:

$$S(N + 1, K) = K[S(N, K)] + S(N, K - 1) \quad . \qquad (4.6)$$

Using the above properties of Stirling numbers of the second kind, Duran and Odell (1974) derive a two-way table of Stirling numbers of the second kind from $S(1, 1)$ through $S(8, 8)$, which aptly depicts the immensity of the numerical problem of complete enumeration of $\mathcal{Y}$.

In this section, the exact probability distributions of the c statistic for $N = 3$, 4, and 5 were derived by the method of complete enumeration. For $N = 6$,

$$L_N = 203 \quad ,$$

which implies that

$$\frac{L_N(L_N - 1)}{2} = 20,503$$

pairwise comparisons of clusterings from $\mathcal{Y}$ are necessary to derive the probability distribution of the c statistic by the method of complete enumeration. Thus, for large values of N, an alternative procedure for deriving or approximating the probability distribution of the c statistic is necessitated.

The Relationship of the Distribution of the Simple Matching Coefficient to the Distribution of the Comparative Statistic

An alternative to the method of complete enumeration for deriving the probability distributions of the c statistic for specified values of N is to construct, under a set of "reasonable" assumptions, a population model for the c statistic, which yields general formulas for the p.m.f. and the moments of the distribution. The set of "reasonable" assumptions should adequately and correctly characterize the population of interest.

Goodall (1967) derives a theoretical distribution for the simple matching coefficient under a set of assumptions which may be delineated as follows:

1. Each binary vector, $U = (u_1, u_2, \ldots, u_n)$, is randomly selected from a population of binary vectors of length n, where the probabilities of the alternatives in the population for each component, u_j, $j = 1, 2, \ldots, n$, of U are given by the following formulation:

$$f_{1j} = P(u_j = 1) \quad ;$$

$$f_{0j} = P(u_j = 0) \quad ;$$

$$f_{0j} + f_{1j} = 1 \quad .$$

2. The components, u_j, $j = 1, 2, \ldots, n$, of each binary vector U are mutually independent.

From the above assumptions, it follows that the probability, p_j, that two randomly chosen binary vectors, U and V, of length n match on their $j^{\underline{th}}$ components is derived as follows:

$$\begin{aligned} p_j &= P(u_j = v_j) \\ &= P(u_j = 1)P(v_j = 1) + P(u_j = 0)P(v_j = 0) \\ &= f_{1j}^2 + f_{0j}^2 \quad . \end{aligned}$$

As a consequence, Goodall (1967) states that the probability distribution of the simple matching coefficient, s, is a special case of the Poisson binomial distribution. Therefore,

$$E(s) = \frac{1}{n} \sum_{j=1}^{n} p_j = \bar{p} \quad ;$$

$$VAR(s) = \frac{\bar{p}(1 - \bar{p})}{n} - \frac{VAR(p_j)}{n} \quad .$$

It is also noted by Goodall (1967) that if f_{1j} is constant for all u_j, $j = 1, 2, \ldots, n$, then the Poisson binomial distribution

degenerates to the binomial distribution. Thus, under certain restrictions, the simple matching coefficient has a binomial distribution.

It was previously shown that the c statistic and the simple matching coefficient are equivalent. Since each component of a binary representation of clustering Y indicates whether a particular pair of data points occur together or apart in clustering Y and since each pair of data points has the same likelihood of occurring together in a randomly chosen partition of X, then over the set of all possible clusterings of N data points, f_{1j} must be a constant for all components of the binary representation of clustering Y. Hence, according to Goodall's (1967) development of the distribution of the simple matching coefficient, the probability distribution of the c statistic should be binomial. The relationship of the binomial distribution to the previously derived exact probability distributions of the c statistic for $N = 3$, 4, and 5 requires further exploration.

Case 1, $N = 3$

When $N = 3$, $n = 3$; and from Figure 7b, it is obvious that for all $j = 1, 2, 3$,

$$f_{1j} = \frac{2}{5} \quad \text{and} \quad f_{0j} = \frac{3}{5} .$$

Therefore,

$$p_j = \left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2 = \frac{13}{25} , \quad j = 1, 2, 3 ;$$

$$\bar{p} = \frac{13}{25} .$$

Hence, if $N = 3$, then

$$E(c) = \bar{p} = \frac{13}{25} ,$$

and the population mean for the c statistic as obtained by the method of complete enumeration agrees with the mean of the binomial distribution.

If the variance of the c statistic for the binomial formulation is denoted by $VAR_b(c)$, then

$$VAR_b(c) = \frac{\left(\frac{13}{25}\right)\left(\frac{12}{25}\right)}{3} = \frac{52}{625} \neq \frac{56}{625} = VAR(c) \quad .$$

It should be noted that for $N = 3$, the variance of the binomial distribution underestimates the exact variance of the c statistic as derived by the method of complete enumeration. It is also easily observed that for $N = 3$, the probability distribution of the c statistic is not derivable from the binomial p.m.f.

Case 2, N = 4

When $N = 4$, $n = 6$; and from Figure 8b, it is obvious that for all $j = 1, 2, \ldots, 6$,

$$f_{1j} = \frac{1}{3} \quad \text{and} \quad f_{0j} = \frac{2}{3} \quad .$$

Therefore,

$$p_j = \left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2 = \frac{5}{9}, \quad j = 1, 2, \ldots, 6 \; ;$$

$$\bar{p} = \frac{5}{9} \quad .$$

Hence, if $N = 4$, then

$$E(c) = \bar{p} = \frac{5}{9} \quad ,$$

and the population mean for the c statistic as obtained by the method of complete enumeration agrees with the mean of the binomial distribution.

If the variance of the c statistic for the binomial formulation is denoted by $VAR_b(c)$, then

$$VAR_b(c) = \frac{\left(\frac{5}{9}\right)\left(\frac{4}{9}\right)}{6} = \frac{10}{243} \neq \frac{19}{405} = VAR(c) \quad .$$

It should be noted that for $N = 4$, the variance of the binomial distribution underestimates the exact variance of the c statistic as derived by the method of complete enumeration. It is also easily observed that for $N = 4$, the probability distribution of the c statistic is not derivable from the binomial p.m.f.

Case 3, $N = 5$

When $N = 5$, $n = 10$; and from Figure 10, it is obvious that for all $j = 1, 2, \ldots, 10$,

$$f_{1j} = \frac{15}{52} \quad \text{and} \quad f_{0j} = \frac{37}{52} \quad .$$

Therefore,

$$p_j = \left(\frac{15}{52}\right)^2 + \left(\frac{37}{52}\right)^2 = \frac{1594}{2704} \quad , \quad j = 1, 2, \ldots, 10 \; ;$$

$$\bar{p} = \frac{1594}{2704} = .5895 \quad .$$

Hence, if $N = 5$, then

$$E(c) = \bar{p} = .5895 \quad ,$$

and the population mean for the c statistic as obtained by the method of complete enumeration agrees with the mean of the binomial distribution.

If the variance of the c statistic for the binomial formulation is denoted by $VAR_b(c)$, then

$$VAR_b(c) = \frac{(.5895)(.4105)}{10} = .0242 \neq .02897 = VAR(c) \quad .$$

It should be noted that for $N = 5$, the variance of the binomial distribution underestimates the exact variance of the c statistic as derived by the method of complete enumeration. It is also easily observed that for $N = 5$, the probability distribution of the c statistic is not derivable from the binomial p.m.f.

Reconciling the Disparity

Although the c statistic and the simple matching coefficient are equivalent, the c statistic represents a restricted application of the simple matching coefficient. The assumption of mutual independence among the components, u_j, $j = 1, 2, \ldots, n$, of each binary vector U is fundamental to Goodall's (1967) derivation of the theoretical distribution of the simple matching coefficient. However, the components of a binary representation of clustering Y are necessarily dependent because the classification of a particular subset of object space X into clusters is sufficient to determine clustering Y. For example, in clustering Y, if data points X_1 and X_2 occur together in cluster Y_k and data points X_1 and X_3 occur together in cluster $Y_{k'}$, then $Y_k = Y_{k'}$ and data points X_2 and X_3 also occur together in cluster Y_k; this is a consequence of overlapping clusters being disallowed in a partition of the object space. Thus, Goodall's (1967) second fundamental assumption is invalid for the restricted application of the simple matching coefficient to the comparison of clusterings of the object space. Hence, the c statistic does not have a binomial probability distribution.

A partition of the object space represents a strong condition with respect to the composition of $\mathcal{Y}$. However, the condition of being a partition is difficult to quantify in general terms. If, for the purposes of this discussion, N and X are fixed and $n = \binom{N}{2}$, then the cardinality of the population of binary vectors of length n is 2^n, and the cardinality of the set of binary representations of $\mathcal{Y}$ is

$$L_N = \sum_{K=1}^{N} S(N, K) < 2^n \quad .$$

Consequently, the binary representations of $\mathcal{Y}$ are only a subset of the population of binary vectors of length n ; and on this larger population of binary vectors, the c statistic would have a binomial probability distribution. However,

$$2^n - L_N$$

of the members of the population of binary vectors are eliminated from the set of binary representations of $\mathcal{Y}$ by the condition that a clustering must be a partition of object space X . Therefore, the probability distribution of the c statistic on $\mathcal{Y}$ must be derived by a conditional probability argument, but so far this approach has proven to be intractable in general terms. The special cases where N = 3, 4, and 5 were given previously in this chapter.

For the purposes of the comparative study presented in the next chapter, three observations concerning the c statistic will suffice:

1. $.5 < E(c) < 1.0$;
2. The closer c is to 1.0, the more similar are the two clusterings;
3. If
$$c(Y, Y') > c(Y, Y'') \quad ,$$
then Y and Y' are more similar than Y and Y'' are.

CHAPTER V

A COMPARATIVE STUDY OF TWELVE AGGLOMERATIVE CLUSTERING METHODS

Rationale for the Comparative Study

A clustering results from the interaction of the lineaments of the data with a clustering method, but distinct clustering methods often produce different clusterings when applied to the same data. One explanation for this phenomenon is that different clustering methods are affected by different aspects of the structure (or the lack of it) within the data. Consequently, a comparative study of clustering methods should also provide for an investigation of the effect of controlled structural changes within the data on the resultant clusterings. Thus, a basis for comparing clustering methods is induced by giving operational interpretations to the fundamental concepts of "retrieval" and "noise".

The philosophical genesis of the concept of "retrieval" may be traced to the Aristotelian postulation of the existence of "natural" structure in the universe. A clustering method is purported to be a functional mechanism for finding or "retrieving" "natural" structure within data. Hence, the degree to which a clustering method "retrieves" known structure within generated data is an important characteristic of the clustering method. To quantify the "retrieval" ability of a clustering method, N data points are generated from K "well-separated"

populations, and the clustering of size K which groups together data points which are generated from the same population is denoted by Y. Letting Y' denote the clustering which results from applying a specific clustering method to the N data points, then the value of $c(Y, Y')$ is a measure of the "retrieval" ability of the clustering method (subject to random variation in the generated data).

In engineering terms, the concept of "noise" is used to describe detectable interferences in a signal. Thus, "noise" in terms of the performance of a clustering method might be viewed as any anomaly in the data which interferes with the ability of the clustering method to "retrieve" the "natural" structure present in the data. The simulation of various types of "noise" has been an important aspect of many recent, empirical comparative studies as indicated in Chapter II. Empirical, comparative studies concerning the perturbation of data points as described by Rand (1969, 1971) or the perturbation of initial ranks as described by Cunningham and Ogilvie (1972) or Baker (1974) represent attempts to investigate the effect of a particular type of "noise" on the performance of a clustering method. Rand (1971, p. 848) gives the following motivation for investigating the sensitivity of a clustering method to perturbation of the data:

> In many applications it is not known whether the data are good representations of their respective populations. The changes of clustering which result from slight movement of points are therefore of critical importance in both choice of methods and interpretations of results.

Hence, these perturbation studies might be viewed as investigations of the sensitivity to measurement errors or the sensitivity to resampling of a clustering method. Another form of "noise" is simulated by the addition of uninformative variables to the set of p informative

variables which locate the data points in p-space as described by Mrachek (1972).

The simulation of different levels of "noise" by means of changing the correlation between variables embodies the essence of the ideas presented in each of the previously mentioned "noise" studies. For simplicity, only bivariate data will be considered in this discussion; i.e., all data points will consist of two variables and only two variables. If ρ represents the population correlation between the two variables within a single population of data points, then the level of "noise" existent in this population to obscure the clustering of data points from this population into the same cluster is quantified by the specification of a value of ρ. Thus, a specification of $\rho \neq 0$ implies that each variable within the single population of data points is semi-informative rather than completely informative or completely uninformative. It should also be noted that increasing ρ, $\rho \geq 0$, for an otherwise fixed population of data points causes the data points within this population to be systematically shifted from an approximately circular configuration to a more elliptical configuration. Since it has been demonstrated that some clustering methods opt for circular clusters, a relevant, comparative characteristic of a clustering method is its robustness to increasing non-circularity in the population of data points. Hence, a study of the effect of increasing ρ, $\rho \geq 0$, on the "retrieval" ability of a clustering method provides a measure of the degree to which a clustering method imposes structure on the data rather than "retrieving" structure from the data, and it provides a measure of the effect of a particular type of "noise" on the resultant clusterings.

For convenience, the important considerations in any extensive, systematic comparison of clustering methods shall be termed structural parameters; a structural parameter is any variable which controls some aspect of the structure of the data. For the purposes of the comparative study presented in this chapter, the primary structural parameter of interest is ρ as discussed above. However, the set of structural parameters for a comparative study of clustering methods should consist of all variable features within the data which might affect the resultant clusterings. Some of the possible structural parameters which require controlled change to make a comparative study "dynamic" are delineated as follows:

1. N, the number of data points in X;
2. p, the number of variables defining each data point; i.e., the dimensionality of the Euclidean p-space in which X is embedded;
3. K, the number of populations from which the data points are generated;
4. The type of population or the probability distribution from which each of the K populations of data points is generated;
5. μ_k , k = 1, 2, ..., K, the mean vector for each population of data points;
6. Σ_k , k = 1, 2, ..., K, the variance-covariance structure for each population of data points;
7. δ_i , i = 1, 2, ..., $\binom{K}{2}$, the distance between each pair of population mean vectors;

8. The relative location of the population mean vectors or the spatial configuration of the population mean vectors;
9. The split or n_k , $k = 1, 2, \ldots, K$, the number of data points generated from each population of data points.

In any comparative study of clustering methods, some of the structural parameters in the set of possible structural parameters must remain fixed, and a few of the structural parameters of special interest may be extensively studied over a range of meaningful settings for a fixed set of clustering methods. The primary objective of the comparative study presented in the remainder of this chapter is to investigate the effect of increasing the correlation between variables within the populations of data points on the "retrieval" ability of twelve agglomerative clustering methods. However, a limited investigation of the effect of changes in the settings of two other structural parameters is also presented. In the next section, the particular structural parameters of interest for the comparative study of twelve agglomerative clustering methods are specified, and the fixed and variable settings for these structural parameters are given.

Design of the Comparative Study

In terms of the design of the comparative study, initially, it is necessary to specify the setting for each of the fixed structural parameters and the range of settings for each of the variable structural parameters. For the purposes of the comparative study, the probability distribution from which each of the K populations of data points was generated was fixed to be multivariate normal (MVN). A brief discussion of the basic generating procedure used should suffice. For the purpose

of efficient discussion, MVN populations with the same variance-covariance matrix will be termed "similar". MVN vectors may be generated from a population having a mean vector of zero and any specified positive definite, symmetric variance-covariance matrix by calling subroutine GGNRM from the IMSL catalogued programs. Generation from other similar MVN populations may be accomplished by adding a fixed constant vector to each vector generated from the GGNRM subroutine. This procedure simulates the generation of vectors from a MVN population with a mean vector equal to the fixed constant vector which was added to each of the generated vectors and the same variance-covariance matrix as was originally specified.

Because of the necessity to operate within certain cost constraints, the number of data points, the number of variables per data point, and the number of MVN populations of data points in X were fixed at the following values:

(i) $N = 21$;

(ii) $p = 2$;

(iii) $K = 3$.

The choice of $N = 21$ was arbitrary subject to its divisibility by three. However, since the primary purpose of the comparative study was to investigate the effect of increasing the correlation between variables on the "retrieval" ability of twelve agglomerative clustering methods, the choice of $p = 2$ was necessary to simplify the design of the comparative study and to enhance the interpretability of the results from the comparative study. One rationale for choosing $K = 3$ is that to maintain the information content of the variables within a population of data points throughout X , it is important to choose $K > p$. The

choice of $K = 3$ was also related to the choice of a potentially interesting spatial configuration for the population mean vectors.

To facilitate the controlled change of the structural parameters δ_i , $i = 1, 2, \ldots, \binom{K}{2}$, it was apropos to quantify the distance between population mean vectors by a single structural parameter, δ ; i.e.,

$$\forall \; i = 1, 2, \ldots, \binom{K}{2}, \quad \delta_i = \delta \quad .$$

Consequently, since K was fixed at three and since the representation of the distance between the population mean vectors by a single structural parameter implies that the population mean vectors are equally spaced in the plane, the spatial configuration for the population mean vectors was automatically fixed so that the three population mean vectors were always placed at the vertices of an equilateral triangle. It should be noted that the specification of a value for δ in conjunction with the equilateral triangle configuration for the population mean vectors is sufficient with respect to locating the population mean vectors in Euclidean two-space since the actual location of the equilateral triangle in the plane does not affect the performance of an agglomerative clustering method. Thus, N, p, K , the generating probability distribution, and the spatial configuration of the population mean vectors remained fixed at the previously mentioned settings throughout the comparative study of agglomerative clustering methods.

The three structural parameters subject to controlled variation in the comparative study were δ , split, and ρ. The settings for the structural parameter δ , the distance between each pair of population mean vectors, were $\delta = 4.0$ and $\delta = 5.0$; these two settings were deemed worthy of further consideration for the equilateral triangle

spatial configuration of population mean vectors after a preliminary investigation with respect to some agglomerative clustering methods and various settings for some of the other structural parameters. It has been aptly demonstrated by other investigators (e.g., Everitt, 1974) that some clustering methods opt for equal sized clusters. Thus, a limited investigation of the robustness of the twelve agglomerative clustering methods to unequal sized clusters was attempted by contrasting the equal sized clusters setting for split, 7-7-7, with an unequal sized clusters setting for split, 11-7-3.

The variance-covariance structure for the bivariate normal (BVN) populations of data points was of primary importance in the comparative study. Since the structural parameter of interest in the variance-covariance structure was ρ as indicated in the discussion given in the previous section, the data points forming the object space X were generated from three similar BVN populations with a specified value of ρ and unit variances; i.e.,

$$\forall\ k = 1, 2, 3, \quad \Sigma_k = \Sigma = \begin{bmatrix} 1.0 & \rho \\ \rho & 1.0 \end{bmatrix},$$

where $\rho = 0.0, .1, .2, \ldots, .9$.

Consequently, the effect of correlated variables ("noise") on the "retieval" ability of agglomerative clustering methods may be investigated by fixing all structural parameters in the framework which was developed in this section except ρ which is systematically varied across its range of settings.

In Figure 11, the actual population mean vectors used in the comparative study are portrayed for $\delta = 4.0$ and the equilateral triangle spatial configuration of population mean vectors. Letting Σ be the

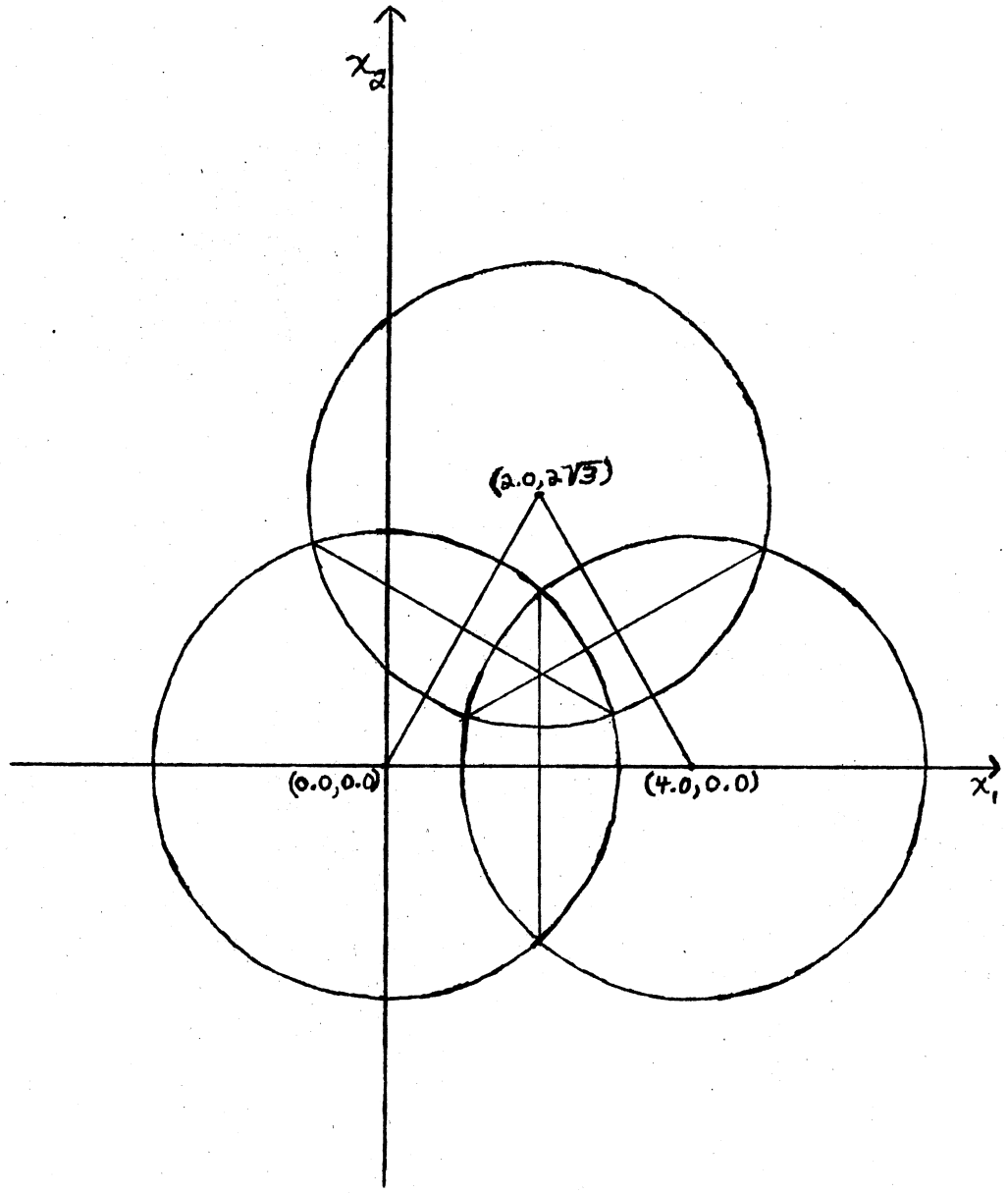


Figure 11. An Example from the Structural Framework Developed for the Comparative Study

identity matrix, then the three circles represent the 3σ contours for each of the BVN populations. Generated data points from this structural framework which, because of random variation, fall in the overlapping regions of the three circles are likely to be clustered with data points generated from a different BVN population than the one from which they were generated. This observation, of course, illustrates only one of the possible reasons that a clustering method fails to "retrieve" the exact structure as generated.

A brief summary of the data structures for the comparative study of agglomerative clustering methods may be outlined as follows:

$$X_i \sim BVN(\mu_k, \Sigma) \quad ,$$

where: $i = 1, 2, \ldots, 21$ with splits into the $K = 3$ populations of either 7-7-7 or 11-7-3 ;

: μ_k, $k = 1, 2, 3$, is constrained by an equilateral triangle spatial configuration and $\delta = 4.0, 5.0$;

: $\Sigma = \begin{bmatrix} 1.0 & \rho \\ \rho & 1.0 \end{bmatrix}$, $\rho = 0.0, .1, .2, \ldots, .9$.

To apply an agglomerative clustering method to a set of data points, it is necessary to specify both a measure of distance and an agglomerative clustering algorithm. For the purposes of the comparative study, the measure of distance was fixed to be Euclidean distance since a preliminary comparative investigation using some of the same agglomerative clustering algorithms later chosen for use in the comparative study in conjunction with Euclidean distance and three other measures of distance indicated that the measure of distance is not as important in determining

the resultant clusterings as the algorithm is. The agglomerative clustering algorithms chosen for the comparative study are discussed in Chapter III. To briefly reiterate the agglomerative clustering algorithms chosen for the comparative study, it should be noted that the twelve agglomerative clustering algorithms form natural groups of two or six algorithms. Thus, the (β, γ) values which define the twelve agglomerative clustering algorithms are conveniently delineated in two groups of six algorithms as follows:

(1) $\beta = 0.0$ with $\gamma = -.5, -.25, \ldots, .75$;

(2) $\beta = -.25$ with $\gamma = -.5, -.25, \ldots, .75$.

One of the basic considerations in designing the comparative study was the choice of a logical running sequence which would produce each of the sets of results necessary to compare the twelve agglomerative clustering methods with respect to their ability to "retrieve" the generated data structure. Each setting of the triple $(\rho, \delta, \text{split})$ of variable structural parameters characterizes a different replication (rep) of the comparative study of agglomerative clustering methods. For each setting of the triple $(\rho, \delta, \text{split})$, the following sequence of steps was utilized to generate twelve values of $c(Y, Y')$, where each value of $c(Y, Y')$ quantified the "retrieval" ability of one of the twelve agglomerative clustering methods:

1. An object space X of data points was generated for the complete set of structural parameters;
2. The Euclidean distance between each pair of data points in X was computed and stored in standard lower triangular matrix order by rows as the vector D ;

3. Each of the twelve agglomerative clustering algorithms was applied to D to produce a hierarchy, H_a , $a = 1, 2, \ldots, 12$;
4. For each of the twelve agglomerative clustering algorithms, the three cluster clustering, $(Y')_a$, was chosen as the representative clustering from H_a , where $a = 1, 2, \ldots, 12$;
5. Each of the representative clusterings, $(Y')_a$, $a = 1, 2, \ldots, 12$, was compared by means of the c statistic to clustering Y of size three, which clustered together all data points generated from the same population of data points.

Thus, by means of the above sequence of steps, a value of $c(Y, Y')$ was assigned to each of the twelve agglomerative clustering methods. For each setting of the triple $(\rho, \delta, \text{split})$, the above sequence of steps was replicated 100 times, and the following statistics were computed for each of the twelve agglomerative clustering methods:

1. $\bar{c}$, the sample mean of the c statistic for the sample of 100 reps;
2. s_c , the sample standard deviation for the 100 c values;
3. The % of the 100 clusterings which corresponded exactly with the generated data structure, i.e., the number of times that $c(Y, Y')$ was equal to one in the 100 reps.

Consequently, for each setting of the triple $(\rho, \delta, \text{split})$ of variable structural parameters and for each of the twelve agglomerative clustering methods, the triple $(\bar{c}, s_c, \%)$ results from 100 reps to quantify

the "retrieval" ability of each of the agglomerative clustering methods, and these triples also provide a means for comparing the performance of the twelve agglomerative clustering methods at the particular settings specified for the complete set of structural parameters. The results from the comparative study of agglomerative clustering methods are discussed in the next section.

A Discussion of the Results from the Comparative Study

Tables I-VIII in the Appendix give the results from the comparative study of agglomerative clustering methods. In these eight tables, the results are given in the form of a triple $(\bar{c}, s_c, \%)$ computed over 100 reps for each setting of the triple of variable structural parameters $(\rho, \delta, \text{split})$ and for each of the twelve agglomerative clustering methods. To simplify the discussion, since Euclidean distance was used in conjunction with each of the twelve agglomerative clustering algorithms, the differences and similarities observed among the agglomerative clustering methods will be discussed in terms of the different algorithms, but this convenience is not intended to imply that the results are independent of the measure of distance employed. An observed difference or similarity among the agglomerative clustering algorithms should be interpreted as a difference or similarity among the agglomerative clustering methods formed by combining the same algorithms with Euclidean distance. The results from the comparative study are also not independent of the fixed structural parameters which were specified in the previous section, but the results will be discussed in terms of the variable structural parameters. Thus, all results from the comparative

study will be discussed in terms of changes in the variable structural parameters $(\rho, \delta, \text{split})$ and changes in the ordered pair (β, γ) which defines the agglomerative clustering algorithm. To enhance the interpretation of the results from the comparative study, Figures 12-29 in the Appendix portray various comparative aspects of the performance of the twelve agglomerative clustering methods. The tables and figures given in the Appendix will be discussed in detail in this section.

Tables I and II display the results for the twelve algorithms in two groups of six and for $\rho = 0.0, .1, .2, \ldots, .9$ with $\delta = 4.0$ and a split of 7-7-7. Table I presents the results for the six algorithms which lie along $\beta = 0.0$, and these results are graphically portrayed in Figures 12-14. In Figure 12, $\bar{c}$ is graphed across the values of ρ for each of the six algorithms lying along $\beta = 0.0$. It should be noted that the single linkage algorithm produces a uniformly smaller $\bar{c}$ than the other algorithms. The highest $\bar{c}$ value occurs at $\rho = 0.0$ with the $(0.0, .25)$ algorithm. Except at a value of $\rho = .9$, either the $(0.0, .25)$ algorithm or the complete linkage algorithm has the highest $\bar{c}$ value. At $\rho = .9$, the average linkage algorithm produces the highest $\bar{c}$ value. Increasing ρ appears to have the greatest effect on the $\bar{c}$ value for the single linkage algorithm.

In Figure 13, s_c is graphed across the values of ρ for each of the six algorithms lying along $\beta = 0.0$. It should be noted that the single linkage algorithm produces a uniformly larger s_c than the other algorithms except at $\rho = .9$ where it has the smallest s_c value. The lowest s_c value occurs at $\rho = 0.0$ with the $(0.0, .75)$ algorithm. In general, the complete linkage and the $(0.0, .75)$ algorithms produce the smallest s_c values.

In Figure 14, the % of the 100 reps for which the agglomerative clustering method "retrieved" the generated data structure exactly is graphed across the values of ρ for each of the six algorithms lying along $\beta = 0.0$. It should be noted that the single linkage algorithm produces a uniformly smaller % than the other algorithms except at $\rho = .9$. The highest % occurs at $\rho = .7$ with the $(0.0, -.25)$ algorithm. The % appears to be less stable across ρ for these six algorithms than either $\bar{c}$ or s_c.

Table II presents the results for the six algorithms which lie along $\beta = -.25$ for $(\rho, \delta = 4.0, 7\text{-}7\text{-}7)$, and these results are graphically portrayed in Figures 15-17. In Figure 15, $\bar{c}$ is graphed across the values of ρ for each of the six algorithms lying along $\beta = -.25$. It should be noted that the $(-.25, -.5)$ algorithm produces a uniformly smaller $\bar{c}$ than the other algorithms. The highest $\bar{c}$ value occurs at $\rho = .7$ with the $(-.25, .25)$ algorithm. For $\rho \geq .3$, the $(-.25, .25)$ algorithm produces the highest values of $\bar{c}$, and for $\rho \leq .2$, the flexible strategy algorithm produces slightly higher values of $\bar{c}$ than the $(-.25, .25)$ algorithm. In general, increasing ρ appears to have only a slight effect on the $\bar{c}$ values produced by the six algorithms lying along $\beta = -.25$ when $\delta = 4.0$ with a 7-7-7 split.

In Figure 16, s_c is graphed across the values of ρ for each of the six algorithms lying along $\beta = -.25$. It should be noted that the $(-.25, -.5)$ algorithm produces a uniformly larger s_c than the other algorithms. The smallest s_c value occurs at $\rho = 0.0$ with the $(-.25, .5)$ algorithm. In general, increasing ρ appears to have only a slight effect on the s_c values produced by the six algorithms lying along $\beta = -.25$ when $\delta = 4.0$ with a 7-7-7 split.

In Figure 17, the % of the 100 reps for which the agglomerative clustering method "retrieved" the generated data structure exactly is graphed across the values of ρ for each of the six algorithms lying along $\beta = -.25$. It should be noted that the highest % obtained with these six algorithms was 22% which occurs at $\rho = .5$ with the $(-.25, .25)$ algorithm, at $\rho = .7$ with the $(-.25, .25)$ and the $(-.25, .5)$ algorithms, and at $\rho = .9$ with the $(-.25, -.25)$ algorithm. The % appears to be less stable across ρ for these six algorithms than either $\bar{c}$ or s_c .

Tables III and IV display the results for the twelve agglomerative clustering algorithms in two groups of six and for $\rho = 0.0, .1, \ldots, .9$ with $\delta = 4.0$ and an 11-7-3 split. Table III presents the results for the six algorithms which lie along $\beta = 0.0$, and these results are graphically portrayed in Figures 18-20. In Figure 18, $\bar{c}$ is graphed across the values of ρ for each of the six algorithms lying along $\beta = 0.0$. It should be noted that the single linkage algorithm produces a uniformly smaller $\bar{c}$ than the other algorithms except at $\rho = .9$ where it has the largest value of $\bar{c}$. The highest $\bar{c}$ value occurs at $\beta = .1$ with the complete linkage algorithm. Except at a value of $\rho = .9$, either the $(0.0, .25)$ algorithm or the complete linkage algorithm has the highest $\bar{c}$ value. Increasing ρ appears to have the greatest effect on the $\bar{c}$ value for the single linkage algorithm.

In Figure 19, s_c is graphed across the values of ρ for each of the six algorithms lying along $\beta = 0.0$. It should be noted that the single linkage algorithm produces a uniformly larger s_c than the other algorithms except at $\rho = .9$. The lowest s_c value occurs at $\beta = 0.0$ with the complete linkage algorithm. In general, the complete

linkage and the (0.0, .75) algorithms produce the smallest s_c values.

In Figure 20, the % of the 100 reps for which the agglomerative clustering method "retrieved" the generated data structure exactly is graphed across the values of ρ for each of the six algorithms lying along $\beta = 0.0$. It should be noted that the single linkage algorithm produces a uniformly smaller % than the other algorithms except at $\rho = .9$. The highest % occurs at $\rho = .9$ with the average linkage algorithm. The % appears to be less stable across ρ for these six algorithms than either $\bar{c}$ or s_c.

Table IV presents the results for the six algorithms which lie along $\beta = -.25$ for $(\rho, \delta = 4.0, 11\text{-}7\text{-}3)$, and these results are graphically portrayed in Figures 21-23. In Figure 21, $\bar{c}$ is graphed across the values of ρ for each of the six algorithms lying along $\beta = -.25$. It should be noted that the (-.25, -.5) algorithm produces a uniformly smaller $\bar{c}$ value than the other algorithms except at $\rho = .8, .9$. The highest $\bar{c}$ value occurs at $\rho = .9$ with the flexible strategy algorithm. Across ρ, the algorithms that produce the higher values of $\bar{c}$ are the flexible strategy, (-.25, .25), and (-.25, .5). In general, increasing ρ appears to have a relatively small effect on the $\bar{c}$ values produced by the six algorithms lying along $\beta = -.25$ when $\delta = 4.0$ with an 11-7-3 split.

In Figure 22, s_c is graphed across the values of ρ for each of the six algorithms lying along $\beta = -.25$. It should be noted that the (-.25, -.5) algorithm produces a uniformly larger s_c that the other algorithms. The smallest s_c value occurs at $\rho = .4$ with the (-.25, .25) algorithm. In general, increasing ρ appears to have only

a slight effect on the s_c values produced by the six algorithms lying along $\beta = -.25$ when $\delta = 4.0$ with an 11-7-3 split.

In Figure 23, the % of the 100 reps for which the agglomerative clustering method "retrieved" the generated data structure exactly is graphed across the values of ρ for each of the six algorithms lying along $\beta = -.25$. It should be noted that the highest % occurs at $\rho = .8$ with the $(-.25, -.25)$ algorithm. Across ρ, the flexible strategy algorithm usually produces the highest value of %. The % appears to be less stable across ρ for these six algorithms than either $\bar{c}$ or s_c.

To enhance the interpretation of the results presented in Tables I-IV, Figures 24-29 provide graphical portrayals across ρ of the performance of the twelve agglomerative clustering methods in six groups of two algorithms for the two different splits. In Figure 24, the % of the 100 reps for which the agglomerative clustering method "retrieved" the generated data structure exactly is graphed across the values of ρ for each of the two algorithms lying along $\gamma = -.5$ with each of the two splits. For either the 7-7-7 split or the 11-7-3 split, the $(-.25, -.5)$ algorithm produces a uniformly higher % across ρ than the single linkage algorithm. In general, the values of % are higher for both algorithms with the 11-7-3 split than with the 7-7-7 split. It is also interesting to note that for these two algorithms, increasing ρ affects the % more with the 11-7-3 split.

In Figure 25, the % of the 100 reps for which the agglomerative clustering method "retrieved" the generated data structure exactly is graphed across the values of ρ for each of the two algorithms lying along $\gamma = -.25$ with each of the two splits. For the 11-7-3 split

only, the $(-.25, -.25)$ algorithm produces a uniformly higher % across ρ than the $(0.0, -.25)$ algorithm. In general, the values of % are higher for both algorithms with the 11-7-3 split than with the 7-7-7 split.

In Figure 26, the % of the 100 reps for which the agglomerative clustering method "retrieved" the generated data structure exactly is graphed across the values of ρ for each of the two algorithms lying along $\gamma = 0.0$ with each of the two splits. Except at $\rho = 0.0$ with the 11-7-3 split, the flexible strategy algorithm produces a uniformly higher % across ρ than the average linkage algorithm produces for both the 7-7-7 split and the 11-7-3 split. Increasing ρ appears to have very little effect on the values of % produced by either the flexible strategy algorithm or the average linkage algorithm when the 7-7-7 split is used. In general, the values of % are higher for both algorithms with the 11-7-3 split as opposed to the 7-7-7 split.

In Figure 27, the % of the 100 reps for which the agglomerative clustering method "retrieved" the generated data structure exactly is graphed across the values of ρ for each of the two algorithms lying along $\gamma = .25$ with each of the two splits. For either the 7-7-7 split or the 11-7-3 split, the $(-.25, .25)$ algorithm produces a higher % if $\rho > .4$, when being compared to the $(0.0, .25)$ algorithm. In general, the values of % are higher for both algorithms with the 11-7-3 split as opposed to the 7-7-7 split. It is also interesting to note that for these two algorithms, increasing ρ affects the % more with the 11-7-3 split.

In Figure 28, the % of the 100 reps for which the agglomerative clustering method "retrieved" the generated data structure exactly is

graphed across the values of ρ for each of the two algorithms lying along $\gamma = .5$ with each of the two splits. For either the 7-7-7 split or the 11-7-3 split, the $(-.25, .5)$ algorithm usually produces a value of % at least as large as the value of % which the complete linkage algorithm produces. In general, the values of % are higher for both algorithms with the 11-7-3 split as opposed to the 7-7-7 split.

In Figure 29, the % of the 100 reps for which the agglomerative clustering method "retrieved" the generated data structure exactly is graphed across the values of ρ for each of the two algorithms lying along $\gamma = .75$ with each of the two splits. For the 7-7-7 split only, the $(-.25, .75)$ algorithm produces a higher % when $\rho \geq .2$ in comparison to the $(0.0, .75)$ algorithm. In general, the values of % are higher for both algorithms with the 11-7-3 split as opposed to the 7-7-7 split.

Tables V and VI display the results for the twelve algorithms in two groups of six and for $\rho = 0.0, .1, .2, \ldots, .9$ with $\delta = 5.0$ and a split of 7-7-7 . Table V presents the results for the six algorithms which lie along $\beta = 0.0$. The results presented in Table V are similar to the results presented in Table I. However, the $\delta = 5.0$ setting, in general terms, causes the values of $\bar{c}$ and % to be larger and the values of s_c to be smaller for all values of ρ in comparison with the values of $(\bar{c}, s_c, \%)$ which resulted for $\delta = 4.0$. It is also interesting to note that $\bar{c}$ and s_c are more stable across ρ when $\delta = 5.0$ than when $\delta = 4.0$. However, the % is much more variable across ρ when $\delta = 5.0$ than when $\delta = 4.0$ for all six of the algorithms. Table VI presents the results for the six algorithms which lie along $\beta = -.25$. The results presented in Table VI are similar to the

results presented in Table II, and all of the general comments made pertaining to differences between the results presented in Tables I and V also hold for differences between the results presented in Tables II and VI.

Tables VII and VIII display the results for the twelve agglomerative clustering algorithms in two groups of six and for $\rho = 0.0, .1, \ldots, .9$ with $\delta = 5.0$ and a split of 11-7-3 . Table VII presents the results for the six algorithms which lie along $\beta = 0.0$. The results presented in Table VII are similar to the results presented in Table III. Table VIII presents the results for the six algorithms which lie along $\beta = -.25$. The results presented in Table VIII are similar to the results presented in Table IV. Also for the 11-7-3 split, when $\delta = 5.0$, the values of $\bar{c}$ and % are larger and the values of s_c are smaller than when $\delta = 4.0$, and this appears to hold for all values of ρ and for all twelve algorithms. It is also interesting to note that $\bar{c}$ and s_c are more stable across ρ when $\delta = 5.0$ than when $\delta = 4.0$ for all twelve of the agglomerative clustering algorithms. However, the % is much more variable across ρ when $\delta = 5.0$ than when $\delta = 4.0$ for all twelve of the algorithms.

In the final chapter of this thesis, some general conclusions will be drawn from the comparative study of agglomerative clustering methods, and some possible directions for the extension of the comparative study will be indicated.

CHAPTER VI

GENERAL TRENDS AND POSSIBLE EXTENSIONS

The stated objective of the research presented in this thesis is: To compare agglomerative clustering methods. However, because of the number of structural parameters requiring controlled variation to make the comparative study "dynamic" and because of the infinite number of possible agglomerative clustering methods which might be chosen for inclusion in the comparative study, the realization of the above objective was necessarily limited in its scope. The comparative study of agglomerative clustering methods presented in this thesis, however, is at least a source for structuring future comparative studies of clustering methods.

Observations and conclusions from the comparative study of agglomerative clustering methods must be made with respect to (wrt) the settings (MVN, N = 21, p = 2, K = 3, equilateral triangle spatial configuration) used for the fixed structural parameters and also with respect to the fixed metric of Euclidean distance; generalizations beyond these settings are of a purely hypothetical nature. Some general trends observable in the results as specified by the triple $(\bar{c}, s_c, \%)$ will be indicated in terms of the triple $(\rho, \delta, \text{split})$ of variable structural parameters and in terms of the ordered pairs (β, γ) which define the agglomerative clustering algorithms. However, these trends were evidenced only for the setting (MVN, N = 21, p = 2, K = 3, equilateral triangle spatial

configuration) of the fixed structural parameters and for the fixed metric of Euclidean distance.

In the context of the triple $(\bar{c}, s_c, \%)$ of results from the comparative study, a "good" agglomerative clustering algorithm for a specified $(\rho, \delta, \text{split})$ might be designated as one that produces a high (close to 1.0) value of $\bar{c}$, a low (close to 0.0) value of s_c, and a high (close to 100) value of $\%$. To explicate "good" algorithms in comparative terms, some convenient notation and terminology is required. For a fixed setting of the triple $(\rho, \delta, \text{split})$ of variable structural parameters, $\bar{c}[A]$ shall denote a $\bar{c}$ value produced by algorithm A; $s_c[A]$ shall denote an s_c value produced by algorithm A; and $\%[A]$ shall denote a $\%$ value produced by algorithm A. Algorithm A will be termed "better" wrt $\bar{c}$ than algorithm B or algorithm B will be termed "worse" wrt $\bar{c}$ than algorithm A iff

$$\forall\, \rho, \quad \bar{c}[A] \geq \bar{c}[B] \quad \text{and} \quad \exists\, \rho \ni \bar{c}[A] > \bar{c}[B],$$

where $\rho = 0.0, .1, \ldots, .9$ and the pair (δ, split) is fixed.

Algorithm A will be termed "better" wrt s_c than algorithm B or algorithm B will be termed "worse" wrt s_c than algorithm A iff

$$\forall\, \rho, \quad s_c[A] \leq s_c[B] \quad \text{and} \quad \exists\, \rho \ni s_c[A] < s_c[B],$$

where $\rho = 0.0, .1, \ldots, .9$ and the pair (δ, split) is fixed.

Algorithm A will be termed "better" wrt $\%$ than algorithm B or algorithm B will be termed "worse" wrt $\%$ than algorithm A iff

$$\forall\, \rho\,, \quad \%[A] \geq \%[B] \quad \text{and} \quad \exists\, \rho \ni \%[A] > \%[B]\,,$$

where $\rho = 0.0, .1, \ldots, .9$ and the pair (δ, split) is fixed.

Thus, given the previously mentioned settings for the fixed structural parameters and a metric of Euclidean distance, some general observations with respect to the settings for the variable structural parameters and the agglomerative clustering algorithms included in the comparative study will be offered for the triple $(\bar{c}, s_c, \%)$ of measured statistics.

The single linkage algorithm, which is the only space-contracting algorithm included in the comparative study, was conspicuously different from all of the other algorithms wrt $(\bar{c}, s_c, \%)$ for all settings of the triple $(\rho, \delta, \text{split})$ used in the comparative study. The single linkage algorithm was in general (with a few exceptions when ρ was close to 1.0) the worst algorithm wrt $(\bar{c}, s_c, \%)$ for all settings of (δ, split). The single linkage algorithm was the only algorithm on which increasing ($\nearrow$) ρ had a marked effect with respect to its performance. The following general trends should be noted for the single linkage algorithm wrt ρ for all settings of the pair (δ, split) used in the comparative study:

(i) $\rho \nearrow \longrightarrow \bar{c} \nearrow$;

(ii) $\rho \nearrow \longrightarrow s_c \swarrow$;

(iii) $\rho \nearrow \longrightarrow \% \nearrow$.

Thus, the performance of the single linkage algorithm improves wrt $(\bar{c}, s_c, \%)$ as ρ increases for all settings of the pair (δ, split). The observations concerning the single linkage algorithm seem to imply that space-contracting algorithms are worse at "retrieving" the

generated structure than either space-conserving or space-dilating algorithms when MVN data and Euclidean distance are employed; this is not surprising considering the theoretical research on agglomerative clustering algorithms presented in Chapter III.

The three space-conserving algorithms -- $(0.0, -.25)$, average linkage, and $(0.0, .25)$ -- lie along the line $\beta = 0.0$. The boundary algorithm on the lower end of the space-conserving region is the single linkage algorithm which is a space-contracting algorithm. It has already been noted that the performance of the single linkage algorithm is better wrt $(\bar{c}, s_c, \%)$ when ρ is close to 1.0 than when ρ is close to 0.0 . The boundary algorithm on the upper end of the space-conserving region is the complete linkage algorithm which is a space-dilating algorithm. The other space-dilating algorithm along the line $\beta = 0.0$ is the $(0.0, .75)$ algorithm. It should be noted that the performance of the complete linkage and the $(0.0, .75)$ algorithms is worse wrt $(\bar{c}, s_c, \%)$ when ρ is close to 1.0 than when ρ is close to 0.0 for the settings of the pair (δ, split) used in the comparative study. In contrast, the space-conserving algorithms are relatively stable across ρ wrt $(\bar{c}, s_c, \%)$ for all settings of the pair (δ, split) used in the comparative study.

From the results of the comparative study, the best algorithms wrt $(\bar{c}, s_c, \%)$ appear to be those lying along the line $\beta = -.25$, and all six of these algorithms are space-dilating algorithms. One of the algorithms lying along the line $\beta = -.25$ is always the best wrt $\bar{c}$ and s_c for all settings of the pair (δ, split) used in the comparative study. However, the performance of all twelve agglomerative clustering algorithms wrt % is somewhat erratic. All six of the algorithms lying along

the line $\beta = -.25$ show relatively little change in their level of performance (i.e., they are relatively stable) across ρ wrt $\bar{c}$ and s_c for all settings of the pair (δ, split) used in the comparative study. For each pair of algorithms with the same γ value, the algorithm with $\beta = -.25$ is generally (a few exceptions exist wrt $\bar{c}$) better wrt $\bar{c}$ and s_c for all settings of the pair (δ, split) used in the comparative study. Consequently, in a future comparative study of agglomerative clustering algorithms in conjunction with Euclidean distance, it would be interesting to explore the performance with respect to their "retrieval" of MVN data structure of a set of six algorithms along the line $\beta = -.5$ with the same γ values as the sets of six algorithms along $\beta = 0.0$ or $\beta = -.25$ which were employed in the comparative study presented in this thesis.

A few general observations with respect to the settings of the pair (δ, split) used in the comparative study can also be made. Apparently, as δ increases, the performance of the algorithms becomes more stable across ρ wrt $\bar{c}$ and s_c for each setting of the structural parameter split; this observation is not surprising since the clusters become more distinct as the population means move further apart. It should be noted that the performance of the algorithms becomes more erratic across ρ wrt % for each setting of the structural parameter split when δ increases. Overall, increasing δ from 4.0 to 5.0 causes an increase in $\bar{c}$ and the % values and a decrease in the s_c values produced by each of the twelve algorithms for all settings of the pair (ρ, split). The two different splits have a greater effect on the performance of the algorithms wrt % than they do wrt $\bar{c}$ and s_c. As an overall conclusion, ρ does not greatly affect the performance of the agglomerative

clustering algorithms wrt $\bar{c}$ and s_c for the two different splits with the effect becoming less for increasing δ.

There are a myriad of possible extensions for the comparative study of agglomerative clustering methods presented in this thesis in terms of changing a setting for any of the specified structural parameters, including both the fixed structural parameters and the variable structural parameters. Obviously, in future comparative investigations of agglomerative clustering methods, a larger value of N should be chosen, and at least a limited comparative investigation of the effect of correlated variables on the "retrieval" ability of the agglomerative clustering methods should be attempted when $p = 3$. Of course, the populations of data points could be generated from probability distributions other than the MVN probability distribution, but the choice of a MVN data structure for each of the populations of data points seems reasonable. However, it would be enlightening to attempt a limited comparative investigation of agglomerative clustering methods when each MVN population of data points represented in X has a different variance-covariance matrix.

A great deal of flexibility in a limited extension of the comparative study of agglomerative clustering methods could be achieved by making the spatial configuration a variable structural parameter while keeping the settings for the other structural parameters (both fixed and variable) the same as specified in Chapter V. An effective method for obtaining a systematic variation of the spatial configuration would be to consider isosceles triangles with the two equal sides having length δ, and since the length of the third side of the isosceles triangle is a function of the measure of the included angle between the two equal sides of the isosceles triangle, the "new" variable structural parameter

could be designated as the measure of the included angle between the two equal sides of the isosceles triangle, which would then be allowed to vary between 0 and π radians. Some theoretical work with respect to the "size" of the overlapping regions for the equilateral triangle spatial configuration and for some of the possible isosceles triangle spatial configurations would represent a valuable contribution towards understanding the "retrieval" results provided by the agglomerative clustering methods, when MVN populations of data points are utilized. The consideration of non-triangular spatial configurations requires the specification of a larger value of K, which should be accompanied by an increase in the value of N to provide for potentially interesting settings of the structural parameter for split. It should also be noted that an increase in the value of p should be accompanied by an increase in the value of K to maintain the information content within the generated populations throughout object space X.

If the settings for the fixed and variable structural parameters other than δ and split remain the same as specified in Chapter V, then the range of potentially interesting settings for δ should be between 3.0 and 6.0; and the two different splits, 7-7-7 and 11-7-3, are probably sufficient to indicate any changes in the performance of the agglomerative clustering methods with respect to equal vs. unequal cluster sizes, considering the relatively small value of N. Since the values of ρ close to 1.0, in general, affected the performance of the agglomerative clustering methods the most, a larger number of values of ρ close to 1.0 (such as .85, .95, .96, .97, .98, .99) might be chosen for inclusion in an extension of the comparative study. It should also be noted that any extension of the comparative study should include a larger number of

replications at each setting of the variable structural parameters for each of the agglomerative clustering methods.

Two extensions of the theoretical work presented in this thesis are also worth noting. The classification of agglomerative clustering algorithms into the classes of space-contracting, space-conserving, and space-dilating algorithms could be repeated for a different set of constraints on the quadruple $(\alpha_i, \alpha_j, \beta, \gamma)$ of parameters which determine $d_{(ij)k}$ in Equation (3.1); i.e., in the general linear combinatorial strategy originated by Lance and Williams (1966). It was also noted in Chapter IV that C.D.F. tables could be constructed for Rand's (1969, 1971) c statistic. However, it is necessary to provide the probability distribution of the c statistic for each special application of the c statistic; e.g., the probability distribution of the c statistic is needed when $N = 21$, $K = 3$, and all clusterings are to be compared to one "correct" clustering. Another interesting paradox results when possible null hypotheses to be tested with respect to the c statistic are tendered. For example, if the pair of hypotheses,

$$H_0: \ c = 1.0 \quad ,$$

$$H_A: \ c < 1.0 \quad ,$$

were of interest in terms of "retrieval" of some generated data structure, it would be desirable to accept H_0.

In conclusion, two justifications for cluster analyzing a data set are offered. Dubes and Jain (1975, p. 20) make the following comment concerning the usefulness of cluster analysis:

> A user must remember that a clustering program is a tool for discovery, not an end in itself. A cluster analysis is really a preprocessing step that should generate ideas and help the user form hypotheses. A cluster analysis should be supplemented by other descriptive techniques... The utility of a cluster analysis is more in the questions raised than in the questions answered.

Finally, Kendall (1973, p. 183) provides a philosophical justification for the research presented in this thesis:

> Over the past fifty years mathematics has tended to discount subjective impressions gained from visual inspection, but the practising statistician cannot afford to neglect any method of feeling his way in p dimensions, however intuitive and however empirical.

A SELECTED BIBLIOGRAPHY

Anderberg, Michael R.

1973 Cluster Analysis for Applications. New York: Academic Press.

Baker, Frank B.

1974 "Stability of Two Hierarchical Grouping Techniques Case 1: Sensitivity to Data Errors." JASA, Vol. 69 (June), p. 440-446.

Ball, Geoffrey H.

1965 "Data Analysis in the Social Sciences: What About the Details?" Fall Joint Computer Conference (In AFIPS Conference Proceedings), Vol. 27, p. 533-559.

Beale, E. M. L.

1969 "Euclidean Cluster Analysis." Bulletin of the International Statistical Institute, Vol. 43, p. 92-94.

Bolshev, L. N.

1969 "Cluster Analysis." Bulletin of the International Statistical Institute, Vol. 43, p. 411-425.

Boyce, A. J.

1969 "Mapping Diversity: A Comparative Study of Some Numerical Methods." In Numerical Taxonomy. Editor A. J. Cole. New York: Academic Press, p. 1-32.

Chaddha, R. L. and L. F. Marcus.

1968 "An Empirical Comparison of Distance Statistics for Populations with Unequal Covariance Matrices." Biometrics, Vol. 24 (September), p. 683-694.

Cormack, R. M.

1971 "A Review of Classification." Journal of the Royal Statistical Society, Vol. 134, p. 321-367.

Cunningham, K. M. and J. C. Ogilvie.

1972 "Evaluation of Hierarchical Grouping Techniques: A Preliminary Study." Computer Journal, Vol. 15, p. 209-213.

Dubes, Richard and Anil K. Jain.

1975 "Clustering Techniques: The User's Dilemma." Technical Report [TR 75-01]. Place: Department of Computer Science, College of Engineering, Michigan State University, November.

Duran, Benjamin S. and Patrick L. Odell.

1974 Cluster Analysis; A Survey. In Econometrics (100). Managing Editors M. Beckman and H. P. Kunzi. Lecture Notes in Econ.

Edwards, A. W. F. and L. L. Cavalli-Sforza.

1965 "A Method for Cluster Analysis." Biometrics, Vol. 21 (June), p. 362-376.

Engelman, L. and J. A. Hartigan.

1969 "Percentage Points of a Test for Clusters." JASA, Vol. 64 (December), p. 1647-1649.

Everitt, Brian.

1974 Cluster Analysis. New York: Halsted Press, Division of John Wiley & Sons.

Farris, James S.

1969 "On the Cophenetic Correlation Coefficient." Systematic Zoology, Vol. 18, p. 279-285.

Fisher, Lloyd and John W. Van Ness.

1971 "Admissible Clustering Procedures." Biometrika, Vol. 58 (April), p. 91-104.

Fisher, Walter D.

1958 "On Grouping for Maximum Homogeneity." JASA, Vol. 53 (December), p. 789-798.

Fleiss, Joseph L. and Joseph Zubin.

1969 "On the Methods and Theory of Clustering." Multivariate Behavioral Research, Vol. 4 (April), p. 235-250.

Fortier, J. J. and H. Solomon.

1966 "Clustering Procedures." In Multivariate Analysis (Proceedings of an International Symposium Held in Dayton, Ohio, June 14-19, 1965). Editor Paruchuri R. Krishnaiah. New York: Academic Press, p. 493-507.

Friedman, H. P. and J. Rubin.

1967 "On Some Invariant Criteria for Grouping Data." Journal of the American Statistical Assoc., Vol. 62 (December), p. 1159-1179.

Goodall, D. W.

1967 "The Distribution of the Matching Coefficient." Biometrics, Vol. 23, p. 647-656.

Goodman, L. A. and W. H. Kruskal.

1954 "Measures of Association for Cross Classification." JASA, Vol. 49 (December), p. 732-764.

Gower, J. C.

1967 "A Comparison of Some Methods of Cluster Analysis." Biometrics, Vol. 23 (December), p. 623-637.

Gower, J. C. and G. J. S. Ross.

1969 "Minimum Spanning Trees and Single Linkage Cluster Analysis." Applied Statistics, Vol. 18, p. 54-64.

Hartigan, J. A.

1967 "Representation of Similarity Matrices by Trees." JASA, Vol. 62 (December), p. 1140-1158.

1970 "Clustering a Data Matrix." Paper Presented at the ASA Meetings, Denver.

Hubert, Lawrence.

1974 "Approximate Evaluation Techniques for the Single-Link and Complete-Link Hierarchical Clustering Procedures." JASA, Vol. 69 (September), p. 698-704.

Jardine, N. and R. Sibson.

1968 "The Construction of Hierarchic and Non-Hierarchic Classifications." The Computer Journal, Vol. 11 (August), p. 177-184.

Johnson, Stephen C.

1967 "Hierarchical Clustering Schemes." Psychometrika, Vol. 32 (September), p. 241-255.

Kendall, M. G.

1938 "A New Measure of Rank Correlation." Biometrika, Vol. 30, p. 81-93.

1948 Rank Correlation Methods. London: Charles Griffin and Company, Limited.

1973 "The Basic Problems of Cluster Analysis." In Discriminant Analysis and Applications. Editor T. Cacoullos. New York: Academic Press, p. 179-191.

King, Benjamin.

1967 "Step-Wise Clustering Procedures." JASA, Vol. 62 (March), p. 86-102.

Kuiper, F. Kent and Lloyd Fisher.

1975 "A Monte Carlo Comparison of Six Clustering Procedures." Biometrics, Vol. 31 (September), p. 777-783.

Lance, G. N. and W. T. Williams.

1966 "A Generalized Sorting Strategy for Computer Classifications." Nature, Vol. 212, p. 218.

1967 "A General Theory of Classificatory Sorting Strategies. 1. Hierarchical Systems." The Computer Journal, Vol. 9 (February), p. 373-380.

Ling, R. F.

1973 "A Probability Theory of Cluster Analysis." JASA, Vol. 68 (March), p. 159-164.

Maronna, Ricardo and Pablo M. Jacovkis.

1974 "Multivariate Clustering Procedures with Variable Metrics." Biometrics, Vol. 27, (September), p. 499-507.

Marriott, F. H. C.

1971 "Practical Problems in a Method of Cluster Analysis." Biometrics, Vol. 27 (September), p. 501-515.

Mayer, Lawrence S.

1971 "A Method of Cluster Analysis When There Exist Multiple Indicators of a Theoretic Concept." Biometrics, Vol. 27 (March), p. 143-155.

Mrachek, Roger J.

1972 "Some Statistical Aspects of Clustering Procedures." (Unpub. M. S. thesis, Iowa State University.)

Norton, James Michael.

1975 "Some Statistical Procedures to Aid in the Evaluation of a Cluster Analysis." (Unpub. Ph.D. thesis, Oklahoma State University.)

Rand, William Medden.

1969 "The Development of Objective Criteria for Evaluating Clustering Methods." (Unpub. Ph.D. thesis, University of California at Los Angeles.)

1971 "Objective Criteria for the Evaluation of Clustering Methods." JASA, Vol. 66 (December), p. 846-850.

Rubin, Jerrold.

1967 "Optimal Classification into Groups: An Approach for Solving the Taxonomy Problem." Journal of Theoretical Biology, Vol. 15, p. 103-144.

Scott, A. J. and M. Knott.

1974 "A Cluster Analysis Method for Grouping Means in the Analysis of Variance." Biometrics, Vol. 30 (September), p. 507-512.

Scott, A. J. and Michael J. Symons.

1971 "Clustering Methods Based on Likelihood Ratio Criteria." Biometrics, Vol. 27 (June), p. 387-397.

Sneath, P. H. A.

1969 "Evaluation of Clustering Methods." In Numerical Taxonomy. Editor A. J. Cole. New York: Academic Press, p. 257-271.

Sneath, Peter H. A. and Robert R. Sokal.

1973 Numerical Taxonomy. San Francisco: W. H. Freeman & Co.

Sokal, Robert R.

1974 "Classification: Purposes, Principles, Progress, Prospects." Science, Vol. 185 (September), p. 1115-1123.

Sokal, R. R. and C. D. Michener.

1958 "A Statistical Method for Evaluating Systematic Relationships." Univ. Kansas Sci. Bull., Vol. 38, p. 1409-1438.

Sokal, Robert R. and F. James Rohlf.

1962 "The Comparison of Dendrograms by Objective Methods." Taxon, Vol. XI, #2 (February), p. 33-40.

Sokal, Robert R. and Peter H. A. Sneath.

1963 Principles of Numerical Taxonomy. San Francisco: W. H. Freeman and Co.

Van Ness, John W.

1973 "Admissible Clustering Procedures." Biometrika, Vol. 60 (August), p. 422-424.

Ward, Joe H., Jr.

1963 "Hierarchical Grouping to Optimize an Objective Function." JASA, Vol. 58 (March), p. 236-244.

Warde, William D.

1975 Personal Communication. Oklahoma State University.

Williams, W. T., J. M. Lambert, and G. N. Lance.

1965 "Multivariate Methods in Plant Ecology." "V. Similarity Analyses and Information - Analysis." Journal of Ecology, Vol. 54, p. 427-445.

Wishart, David.

1969a "An Algorithm for Hierarchical Classifications." Biometrics, Vol. 25 (March), p. 165-170.

1969b "Mode Analysis: A Generalization of Nearest Neighbour which Reduces Chaining Effects." In Numerical Taxonomy. Editor A. J. Cole. New York: Academic Press, p. 282-308.

APPENDIX

RESULTS FROM THE COMPARATIVE STUDY OF TWELVE AGGLOMERATIVE CLUSTERING METHODS

TABLE I

A COMPARISON ACROSS ρ OF SIX ALGORITHMS
ALONG $\beta = 0.0$ WHERE $\delta = 4.0$
WITH A 7-7-7 SPLIT

ρ		Single	(0, -.25)	Average	(0, .25)	Complete	(0, .75)
0	$\bar{c}$	.66829	.81243	.86395	.88314	.87586	.87848
	s_c	.17675	.15176	.10594	.09396	.10382	.08941
	%	6	12	19	20	20	18
.1	$\bar{c}$	.67929	.82648	.85957	.87924	.88181	.87362
	s_c	.18289	.15090	.12373	.09757	.09233	.09657
	%	7	15	18	18	18	17
.2	$\bar{c}$	.70614	.83857	.86243	.87581	.88148	.87810
	s_c	.18323	.14086	.12166	.11083	.10101	.09539
	%	8	18	16	16	16	16
.3	$\bar{c}$	.70795	.83852	.86767	.88286	.88152	.87581
	s_c	.18708	.13631	.11744	.09232	.09287	.09507
	%	9	19	16	17	14	14
.4	$\bar{c}$	.72029	.82471	.86838	.86524	.87805	.86190
	s_c	.17541	.13515	.10526	.11458	.10085	.10566
	%	9	14	16	16	17	14
.5	$\bar{c}$	.71919	.81929	.85357	.86452	.86790	.86438
	s_c	.17753	.14881	.13125	.10797	.10471	.10461
	%	10	16	16	14	15	14
.6	$\bar{c}$	.73057	.83981	.85524	.86600	.86438	.86010
	s_c	.17688	.12886	.12126	.11198	.10959	.10953
	%	10	18	17	14	12	14
.7	$\bar{c}$	.74986	.84105	.86857	.86686	.85257	.85767
	s_c	.18247	.14573	.11406	.11665	.11916	.11819
	%	13	21	17	17	15	15
.8	$\bar{c}$	.77338	.83810	.85590	.85433	.85552	.83924
	s_c	.16066	.13937	.12471	.12769	.11947	.12798
	%	13	18	17	16	18	14
.9	$\bar{c}$	.80505	.84348	.85795	.84667	.82767	.80886
	s_c	.11648	.13362	.12240	.13521	.13097	.13673
	%	13	19	18	16	14	12

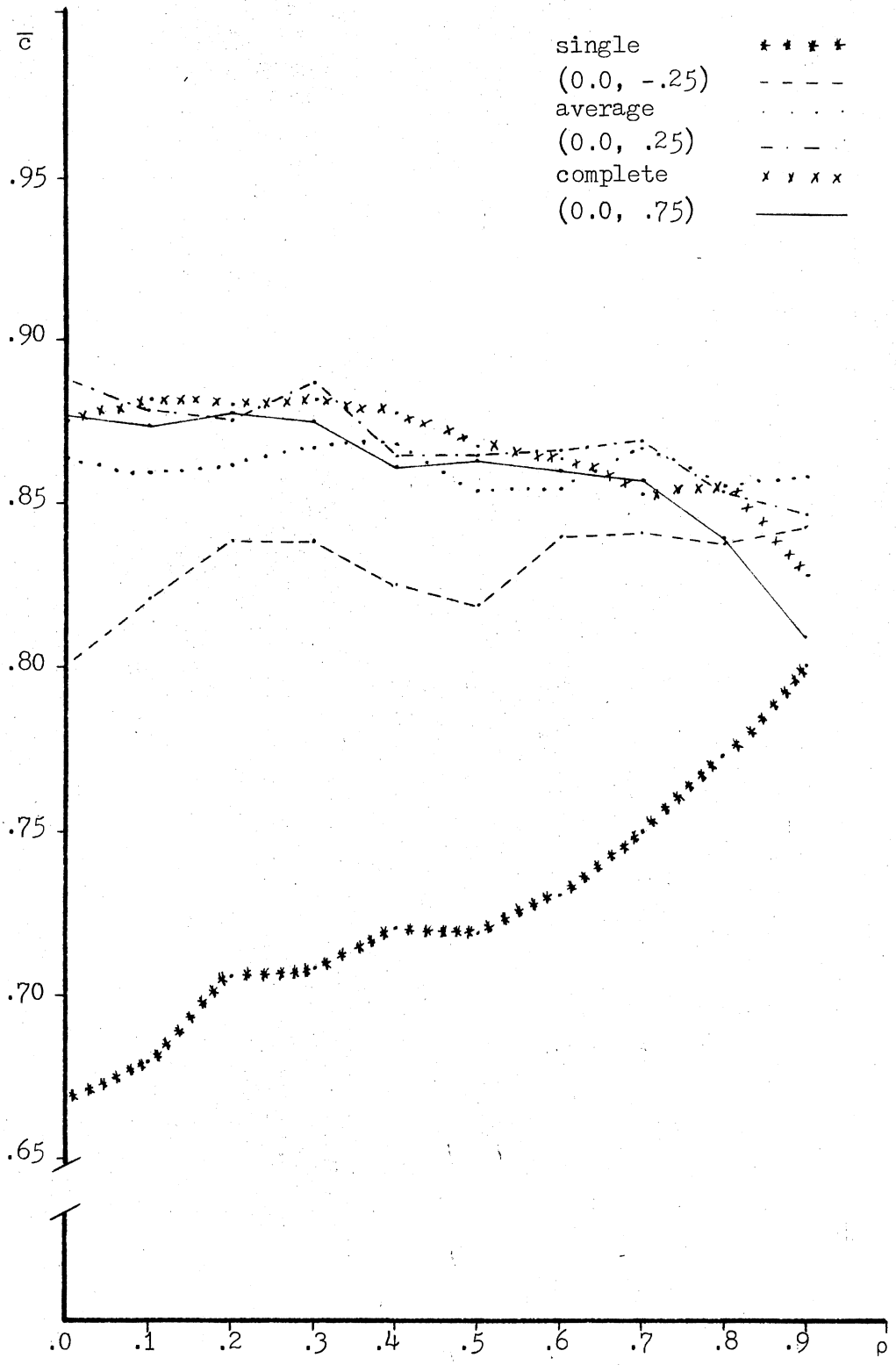


Figure 12. Using $\bar{c}$, a Graphical Comparison across ρ of Six Algorithms along $\beta = 0.0$ where $\delta = 4.0$ with a 7-7-7 Split

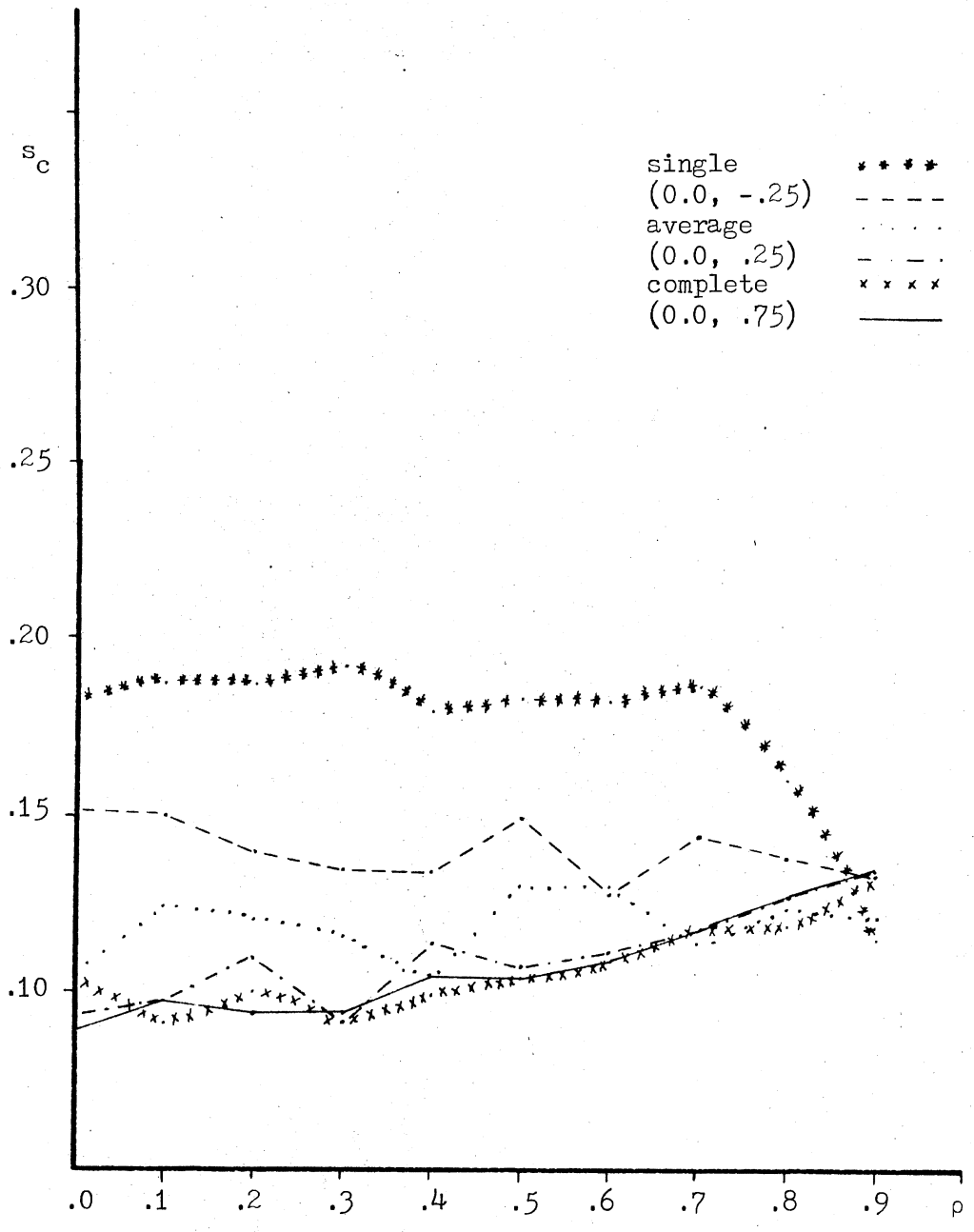


Figure 13. Using s_c, a Graphical Comparison across ρ of Six Algorithms along $\beta = 0.0$ where $\delta = 4.0$ with a 7-7-7 Split

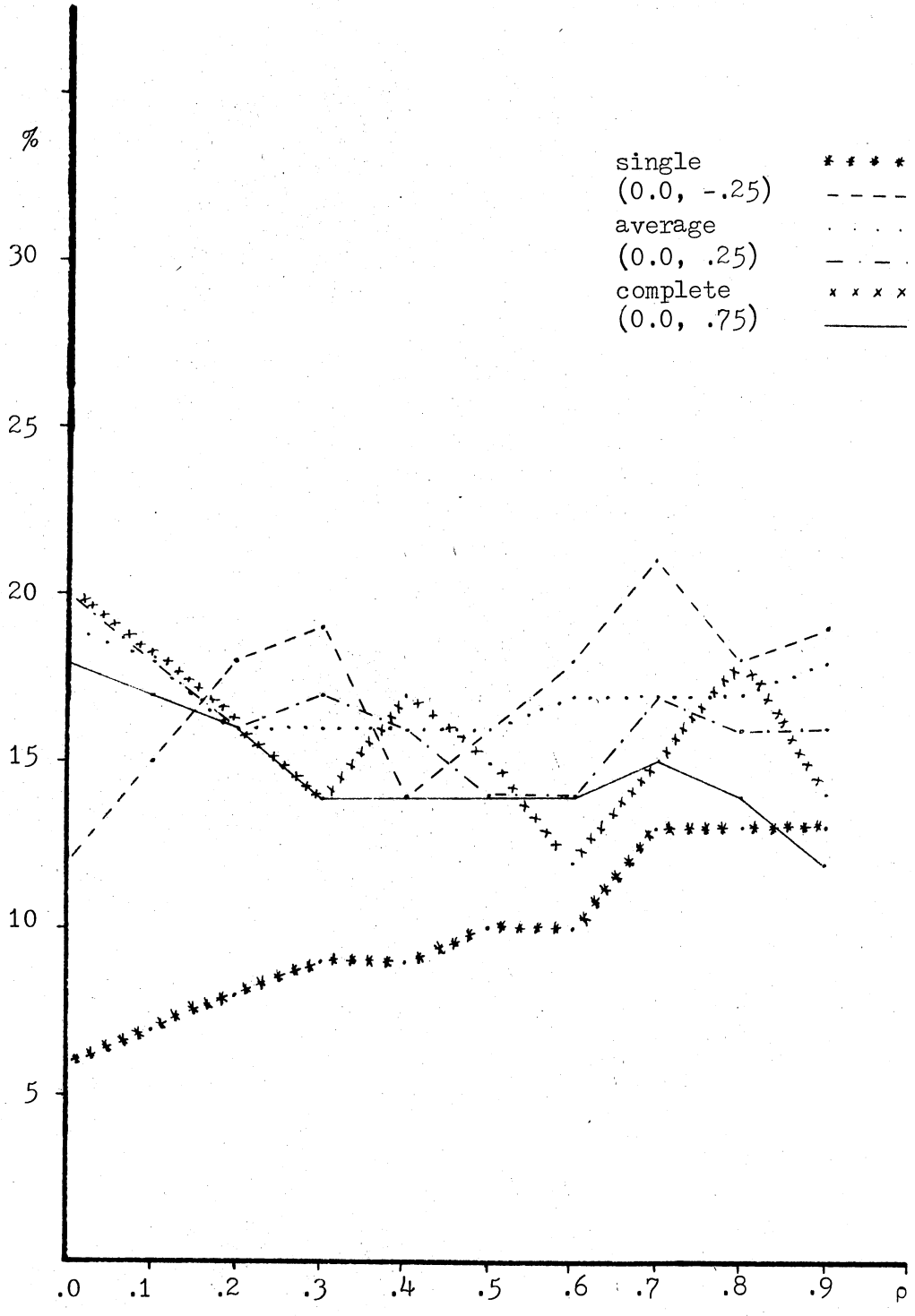


Figure 14. Using % Correctly Classified, a Graphical Comparison across ρ of Six Algorithms along $\beta = 0.0$ where $\delta = 4.0$ with a 7-7-7 Split

TABLE II

A COMPARISON ACROSS ρ OF SIX ALGORITHMS
ALONG $\beta = -.25$ Where $\delta = 4.0$
WITH A 7-7-7 SPLIT

ρ		(-.25,-.5)	(-.25,-.25)	Flexible	(-.25,.25)	(-.25,.5)	(-.25,.75)
0	$\bar{c}$	.82781	.88010	.89810	.89490	.89776	.88448
	s_c	.12637	.09604	.07909	.08001	.06989	.07708
	%	13	19	20	18	17	15
.1	$\bar{c}$	.83495	.88195	.89843	.89581	.89281	.87857
	s_c	.11385	.08855	.08062	.07858	.07125	.08352
	%	12	16	21	18	15	13
.2	$\bar{c}$	.83700	.88552	.89676	.89557	.89200	.88924
	s_c	.12231	.09292	.08317	.07969	.07534	.08557
	%	12	17	19	18	17	18
.3	$\bar{c}$	.84148	.88510	.89305	.89867	.89652	.89362
	s_c	.12222	.09690	.09184	.08387	.07998	.08076
	%	16	17	19	20	19	18
.4	$\bar{c}$	.83605	.88819	.89271	.90110	.89595	.89086
	s_c	.12897	.09595	.08871	.08285	.08543	.08784
	%	15	20	21	21	21	20
.5	$\bar{c}$	.84057	.87910	.89290	.90005	.89471	.89210
	s_c	.12682	.10915	.08836	.08787	.08476	.08656
	%	16	19	19	22	20	19
.6	$\bar{c}$	.84671	.87610	.89776	.90467	.89848	.88562
	s_c	.12647	.10749	.08448	.07729	.08386	.09058
	%	18	16	19	19	20	17
.7	$\bar{c}$	.84657	.88995	.89867	.90614	.90381	.89867
	s_c	.13640	.10026	.08130	.07944	.07663	.08196
	%	18	19	19	22	22	21
.8	$\bar{c}$	.85871	.88962	.89624	.90571	.89510	.89819
	s_c	.12323	.08948	.08204	.07021	.07580	.07719
	%	21	19	19	19	16	17
.9	$\bar{c}$	.86481	.89295	.89800	.89957	.89748	.88724
	s_c	.11027	.09593	.07921	.07492	.07597	.08713
	%	19	22	21	20	19	20

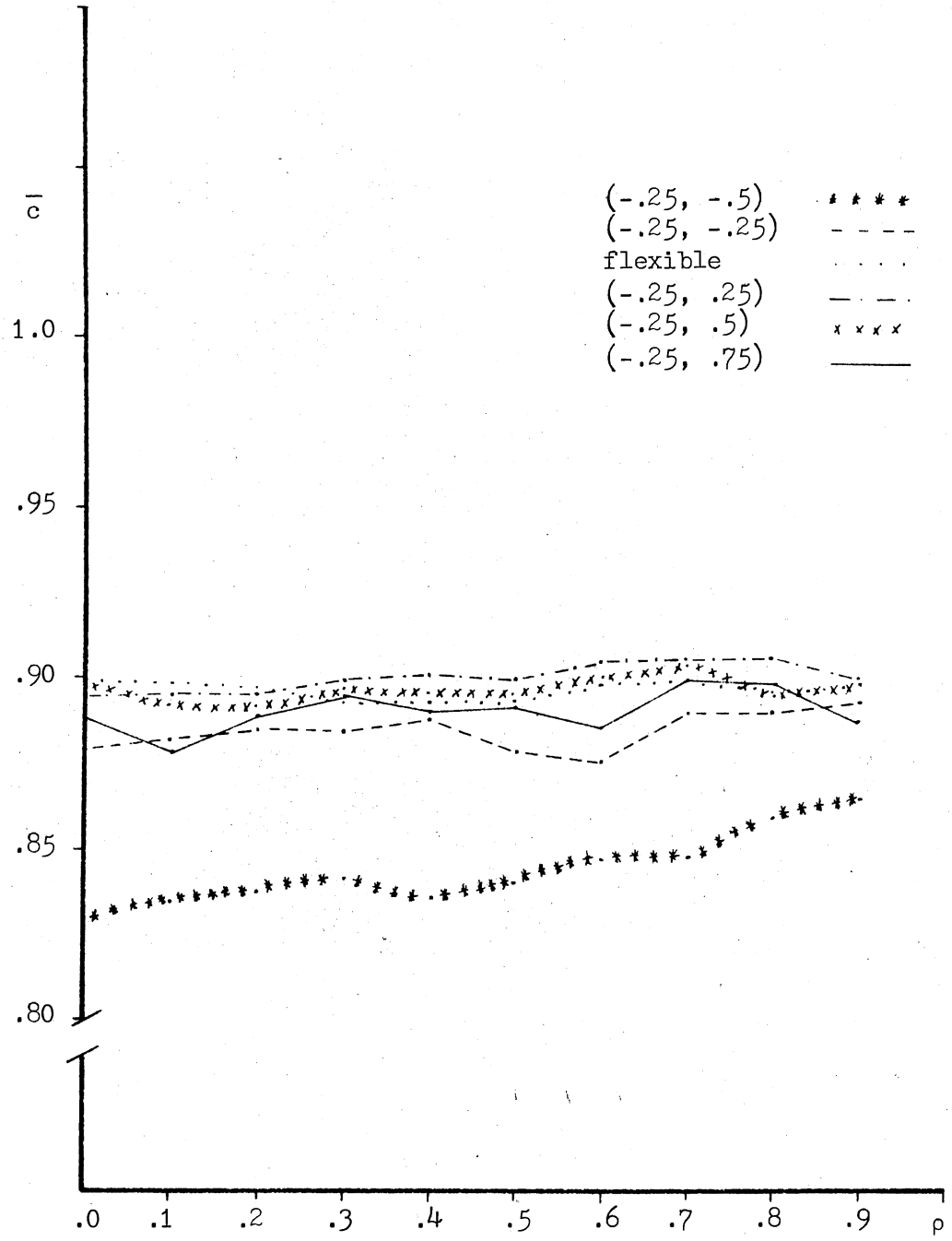


Figure 15. Using $\bar{c}$, a Graphical Comparison across ρ of Six Algorithms along $\beta = -.25$ where $\delta = 4.0$ with a 7-7-7 Split

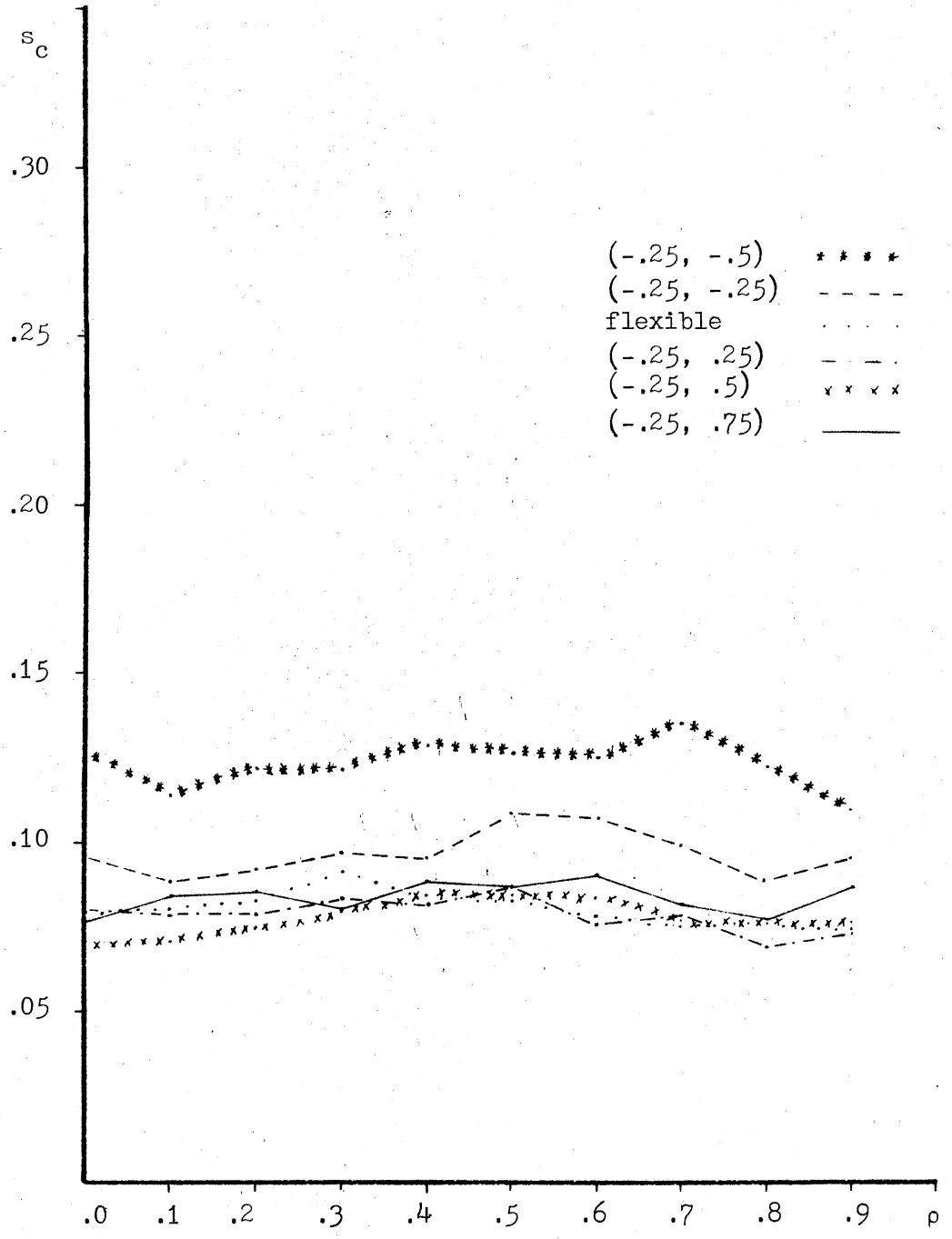


Figure 16. Using s_c, a Graphical Comparison across ρ of Six Algorithms along $\beta = -.25$ where $\delta = 4.0$ with a 7-7-7 Split

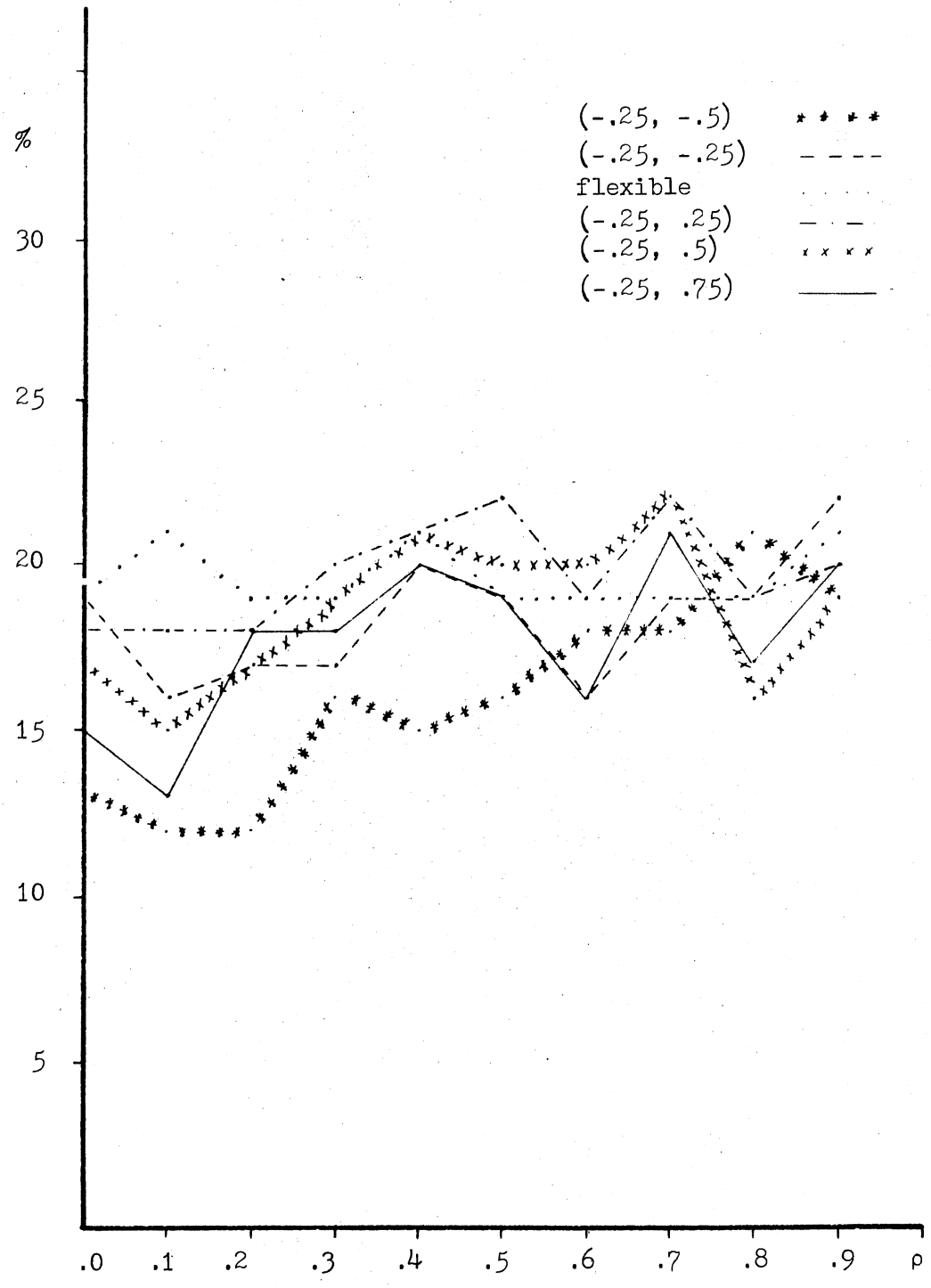


Figure 17. Using % Correctly Classified, a Graphical Comparison across ρ of Six Algorithms along $\beta = -.25$ where $\delta = 4.0$ with a 7-7-7 Split

TABLE III

A COMPARISON ACROSS ρ OF SIX ALGORITHMS
ALONG $\beta = 0.0$ WHERE $\delta = 4.0$
WITH AN 11-7-3 SPLIT

ρ		Single	(0, -.25)	Average	(0, .25)	Complete	(0, .75)
0	$\bar{c}$	.73690	.84286	.88281	.87762	.89029	.88100
	s_c	.17293	.13059	.10336	.10947	.08978	.09757
	%	6	14	21	21	21	18
.1	$\bar{c}$	.73262	.85729	.86405	.87986	.89357	.87876
	s_c	.16995	.12401	.11236	.10782	.09674	.09793
	%	7	15	18	23	25	21
.2	$\bar{c}$	.72110	.84195	.86400	.88029	.88438	.87124
	s_c	.17196	.13244	.11760	.10299	.09652	.10533
	%	7	15	20	23	23	20
.3	$\bar{c}$	.73095	.84843	.86552	.88271	.87495	.87514
	s_c	.17258	.13080	.11502	.10155	.09843	.10310
	%	11	17	19	21	19	21
.4	$\bar{c}$	.74524	.84238	.85714	.87629	.87867	.87067
	s_c	.16973	.13408	.12262	.10773	.10304	.11281
	%	11	17	19	21	21	22
.5	$\bar{c}$	.74405	.85262	.88186	.87748	.88443	.87576
	s_c	.17147	.12990	.10624	.10165	.09875	.10954
	%	11	19	22	20	23	24
.6	$\bar{c}$	.74929	.83976	.86414	.88324	.87943	.88276
	s_c	.16572	.13096	.11609	.10914	.10610	.10174
	%	11	17	21	24	24	23
.7	$\bar{c}$	.78281	.84157	.84390	.87952	.88543	.86476
	s_c	.15686	.14743	.12904	.11422	.10445	.12061
	%	13	23	19	26	27	23
.8	$\bar{c}$	.83529	.86248	.85048	.86257	.87290	.86100
	s_c	.14194	.13307	.13348	.13085	.11145	.11017
	%	20	26	23	25	25	21
.9	$\bar{c}$	.87200	.86633	.86900	.85443	.85210	.83200
	s_c	.12752	.13556	.12885	.12608	.11884	.13003
	%	27	24	29	25	22	19

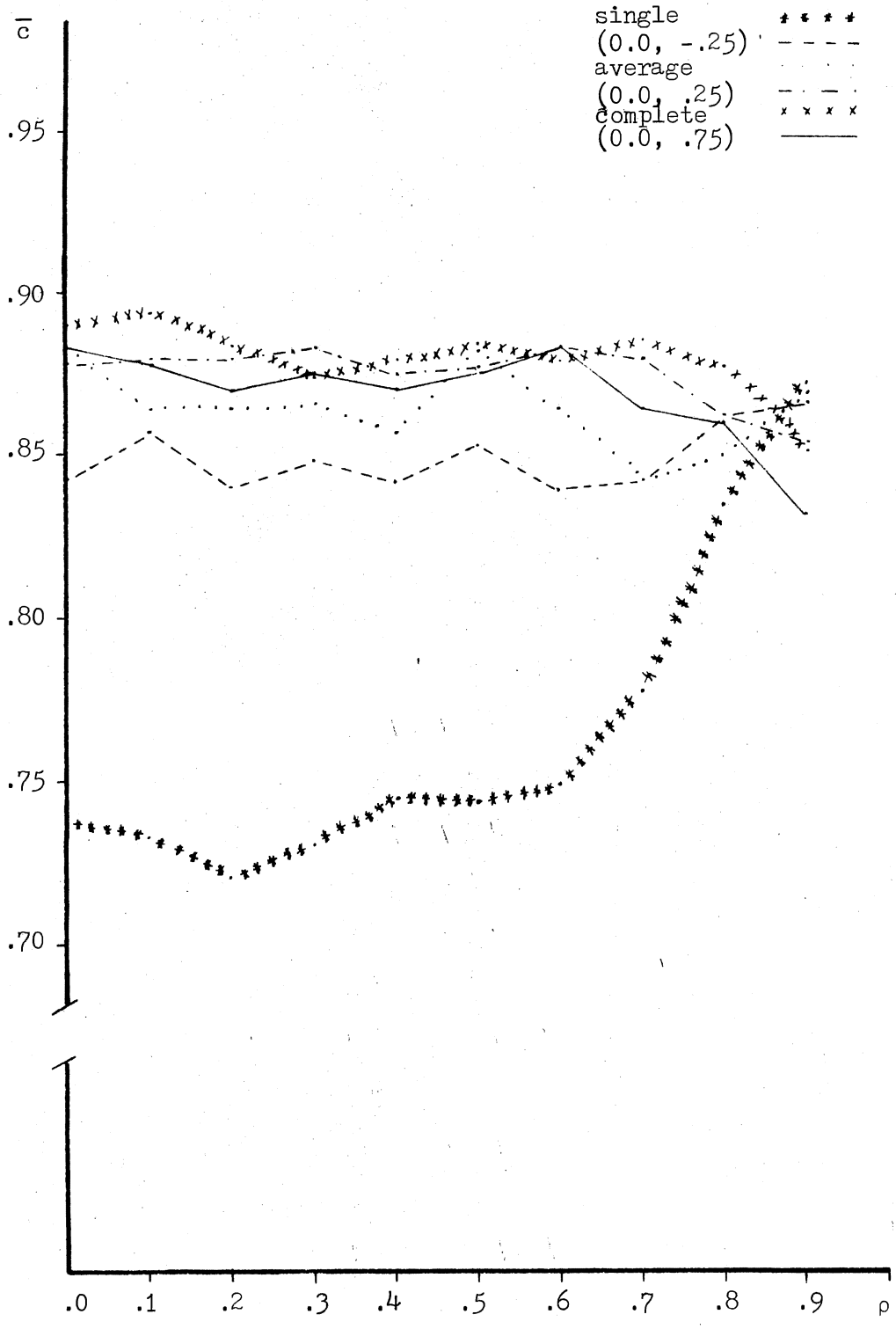


Figure 18. Using $\bar{c}$, a Graphical Comparison across ρ of Six Algorithms along $\beta = 0.0$ where $\delta = 4.0$ with an 11-7-3 Split

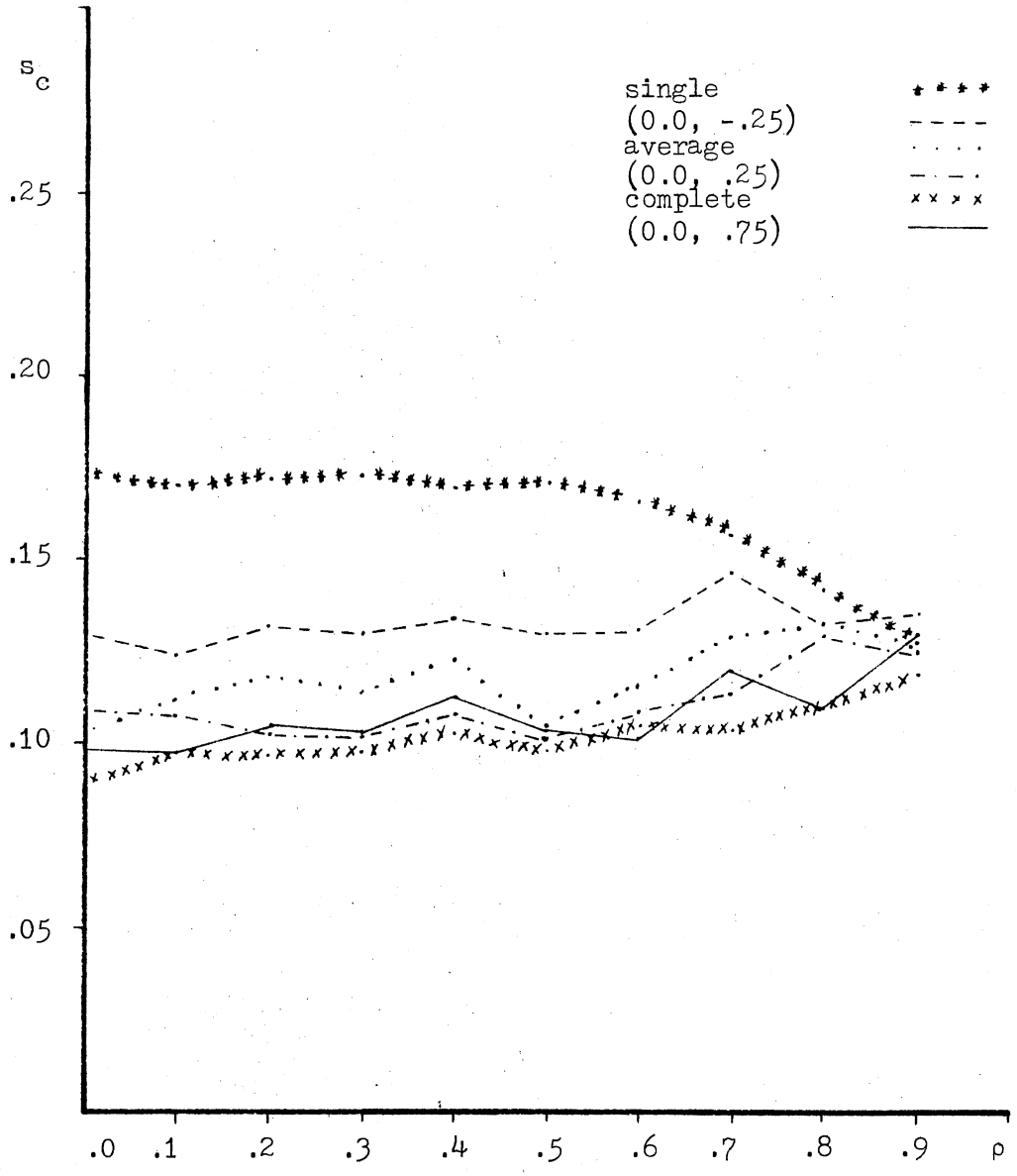


Figure 19. Using s_c, a Graphical Comparison across ρ of Six Algorithms along $\beta = 0.0$ where $\delta = 4.0$ with an 11-7-3 Split

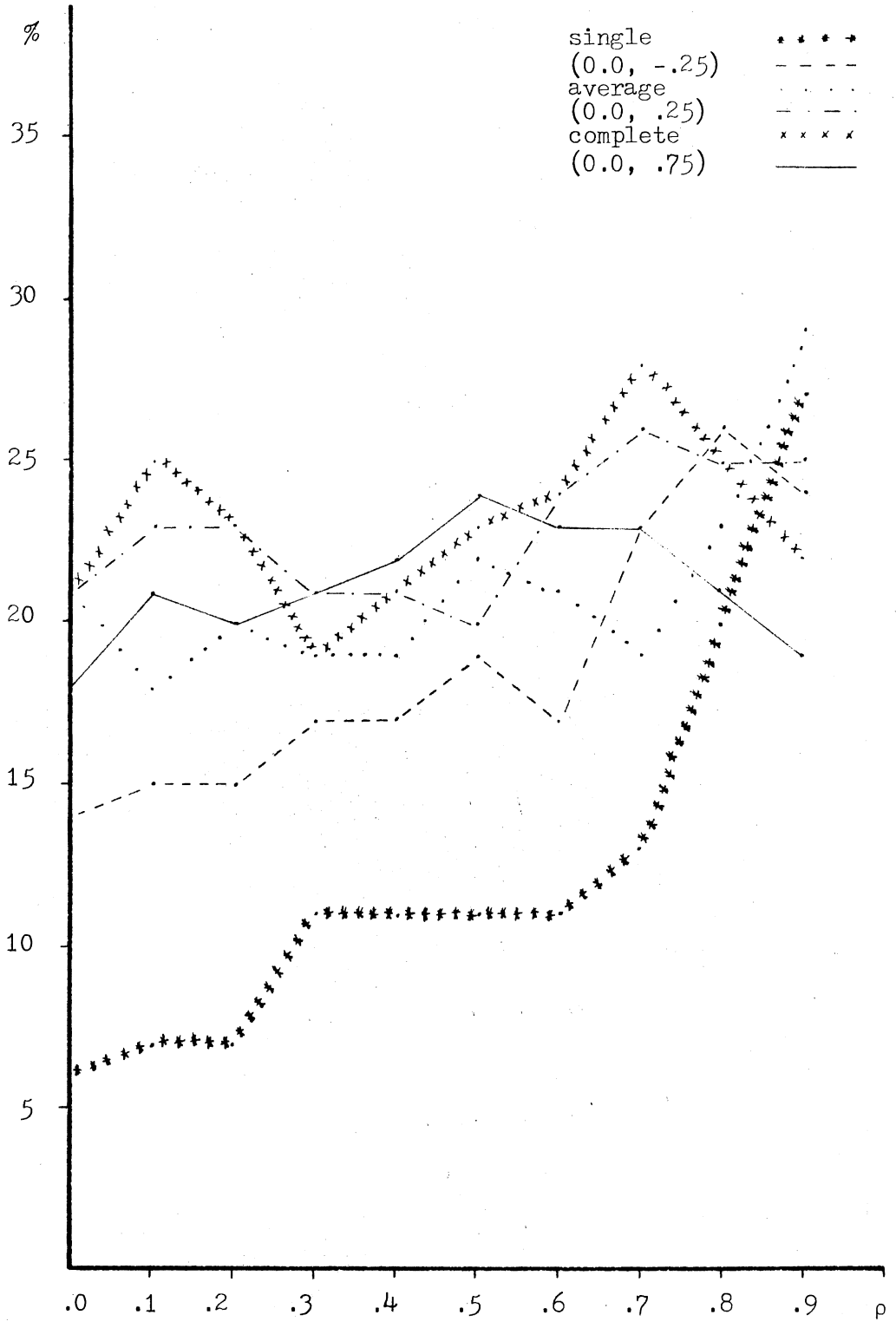


Figure 20. Using % Correctly Classified, a Graphical Comparison across ρ of Six Algorithms along $\beta = 0.0$ where $\delta = 4.0$ with an 11-7-3 Split

TABLE IV

A COMPARISON ACROSS ρ OF SIX ALGORITHMS
ALONG $\beta = -.25$ WHERE $\delta = 4.0$
WITH AN 11-7-3 SPLIT

ρ		(-.25,-.5)	(-.25,-.25)	Flexible	(-.25,.25)	(-.25,.5)	(-.25,.75)
0	$\bar{c}$	.83552	.88552	.89800	.89595	.89071	.87229
	s_c	.13198	.09389	.08262	.08388	.08845	.09822
	%	15	17	20	18	19	17
.1	$\bar{c}$	.84790	.88852	.90071	.89810	.89133	.88338
	s_c	.12291	.09478	.08675	.08026	.08712	.08962
	%	16	21	25	22	21	19
.2	$\bar{c}$	.84214	.87071	.88952	.89457	.89748	.88214
	s_c	.13501	.11219	.09458	.08516	.07950	.09379
	%	18	22	26	25	23	20
.3	$\bar{c}$	.84929	.87824	.88738	.88967	.89343	.88481
	s_c	.12425	.10136	.08492	.08146	.08677	.08972
	%	21	21	21	22	23	18
.4	$\bar{c}$	.85981	.88490	.89248	.89000	.88862	.88352
	s_c	.11194	.09562	.08493	.07691	.08670	.09158
	%	21	23	23	20	21	19
.5	$\bar{c}$	.85224	.89114	.90252	.89490	.89300	.89076
	s_c	.12235	.09682	.08494	.08126	.09421	.09031
	%	22	24	28	25	25	24
.6	$\bar{c}$	.84952	.88205	.90071	.90314	.89371	.87648
	s_c	.12927	.10486	.08894	.08126	.09421	.09680
	%	22	25	29	27	27	24
.7	$\bar{c}$	.86824	.88190	.90114	.90462	.88795	.87519
	s_c	.11063	.10596	.08835	.08561	.09593	.10285
	%	23	27	30	29	25	22
.8	$\bar{c}$	.87190	.89938	.90443	.89952	.89362	.87133
	s_c	.11083	.09689	.09092	.08929	.09450	.09899
	%	25	33	32	30	30	23
.9	$\bar{c}$	.88719	.89724	.90752	.89814	.88933	.87429
	s_c	.10214	.09238	.08170	.09080	.09413	.09923
	%	29	31	32	30	28	25

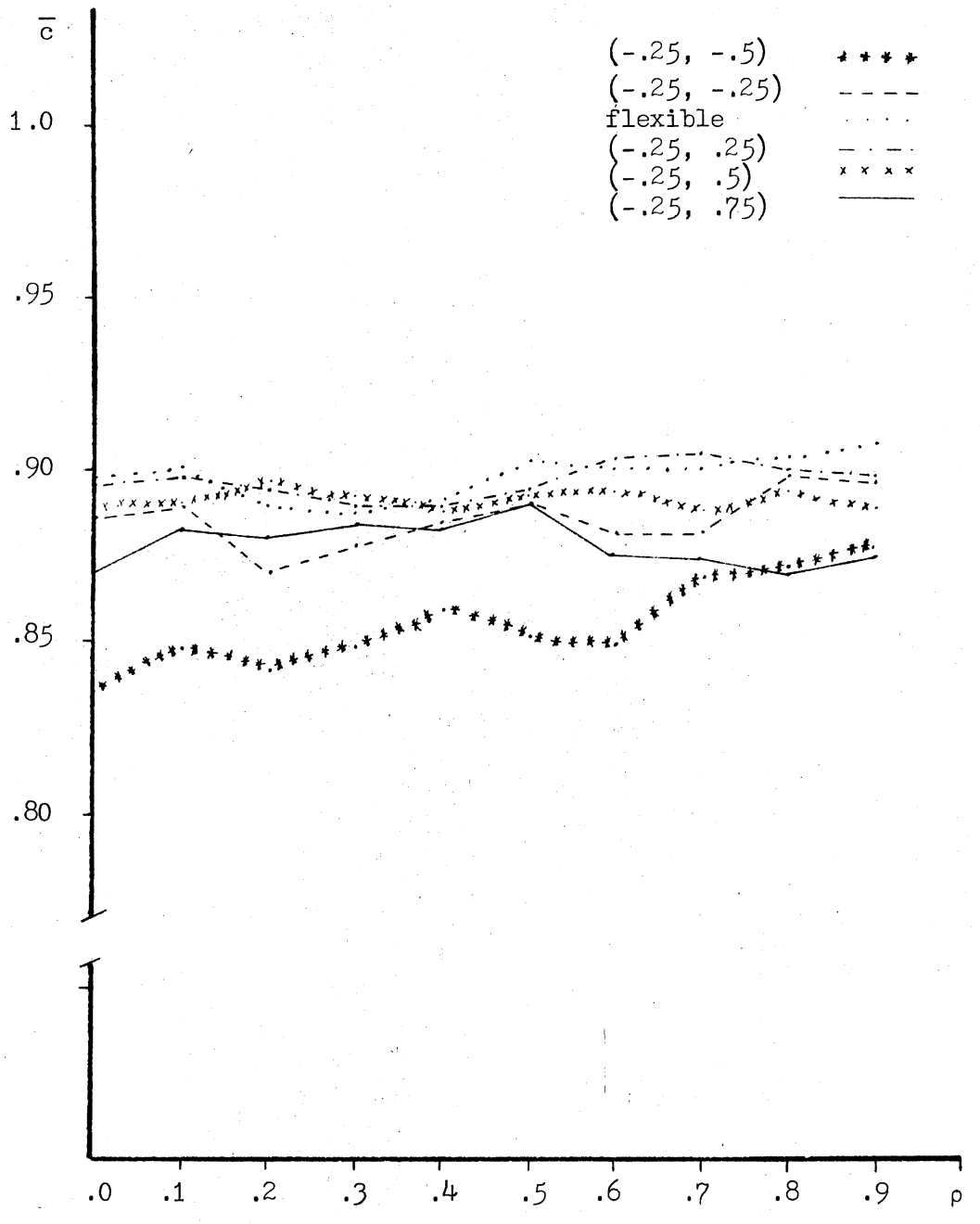


Figure 21. Using $\bar{c}$, a Graphical Comparison across ρ of Six Algorithms along $\beta = -.25$ where $\delta = 4.0$ with an 11-7-3 Split

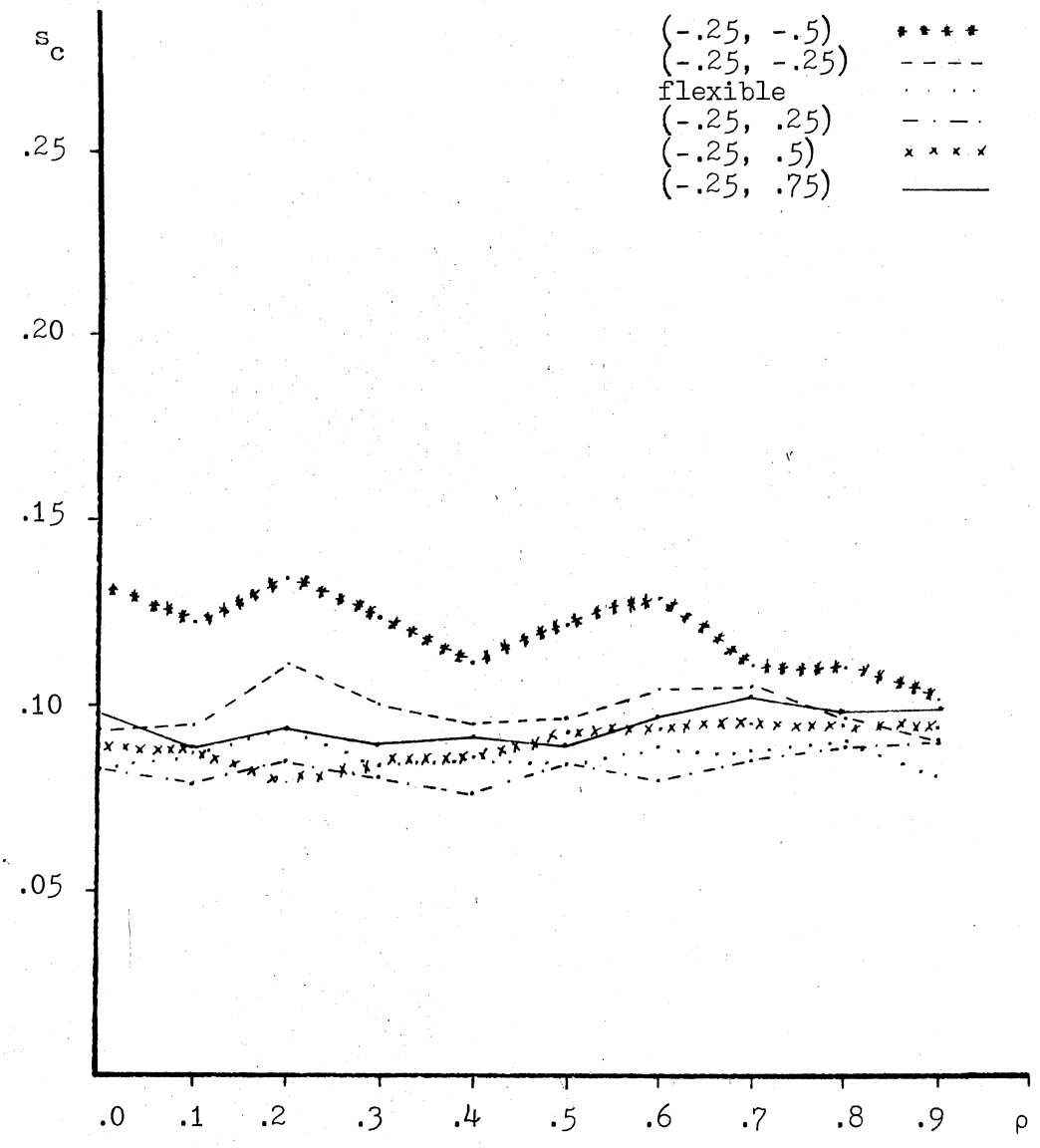


Figure 22. Using s_c, a Graphical Comparison across ρ of Six Algorithms along $\beta = -.25$ where $\delta = 4.0$ with an 11-7-3 Split

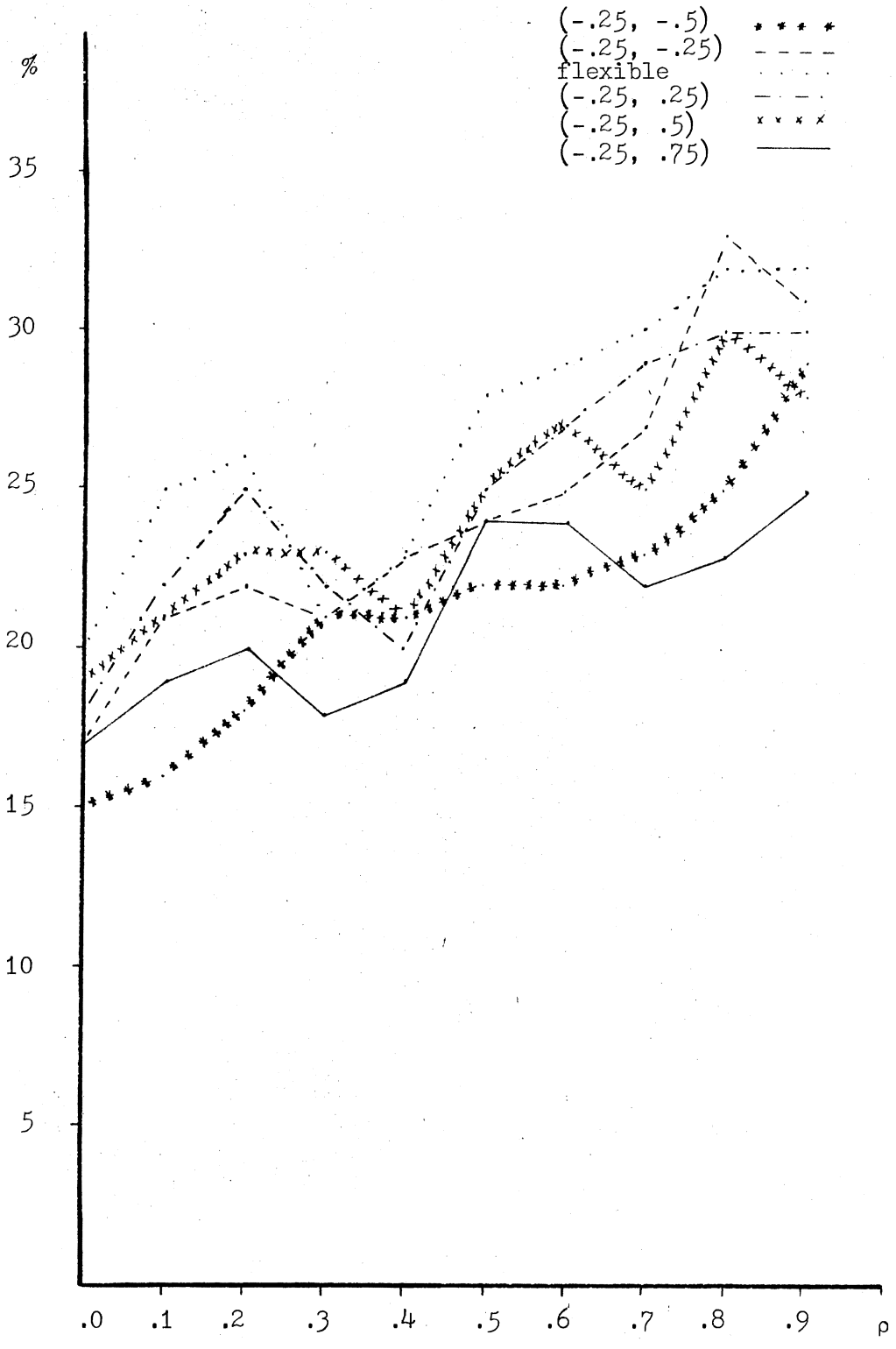
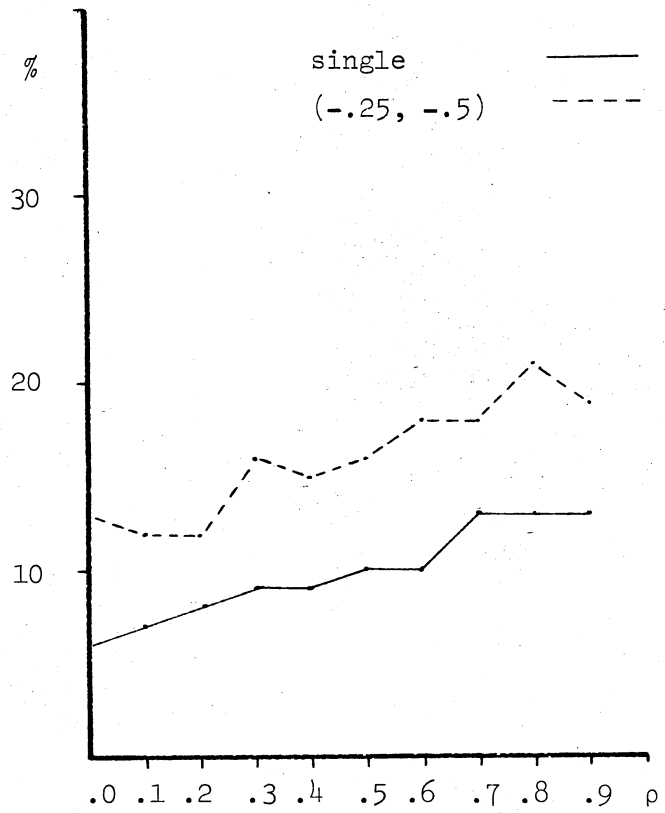
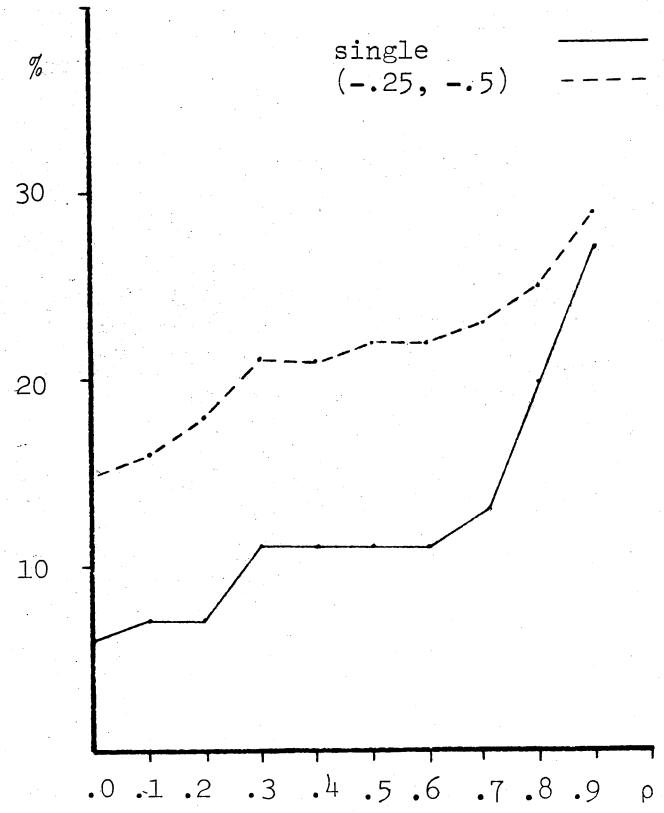


Figure 23. Using % Correctly Classified, a Graphical Comparison across ρ of Six Algorithms along $\beta = -.25$ where $\delta = 4.0$ with an 11-7-3 Split

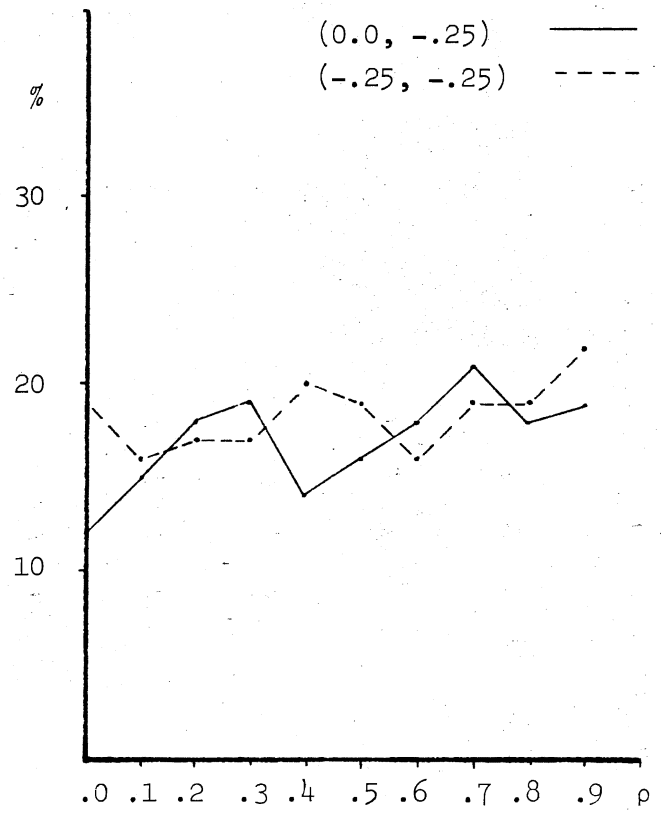


a) 7-7-7 Split

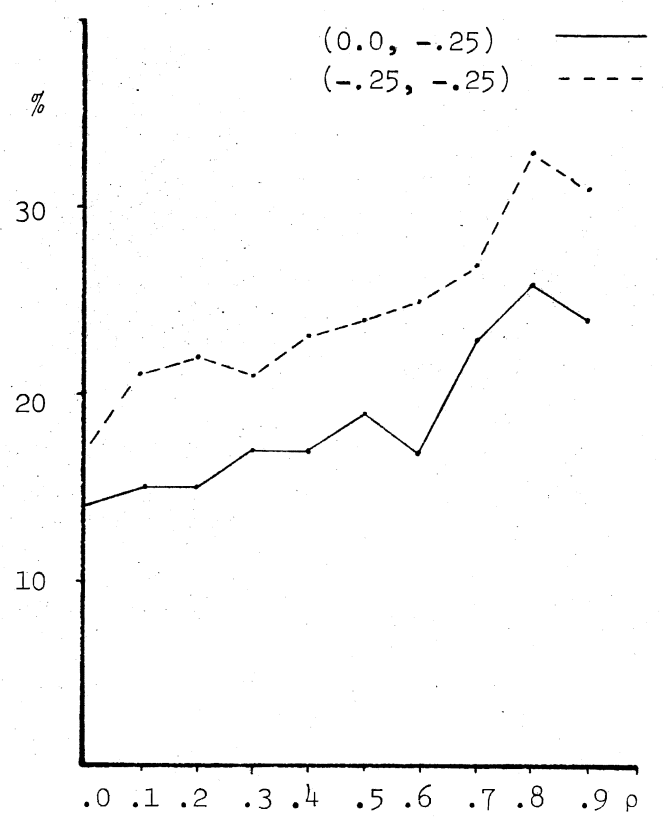


b) 11-7-3 Split

Figure 24. Using % Correctly Classified, a Graphical Comparison across ρ of Two Algorithms along $\gamma = -.5$ where $\delta = 4.0$ with Two Different Splits



a) 7-7-7 Split



b) 11-7-3 Split

Figure 25. Using % Correctly Classified, a Graphical Comparison across ρ of Two Algorithms along $\gamma = -.25$ where $\delta = 4.0$ with Two Different Splits

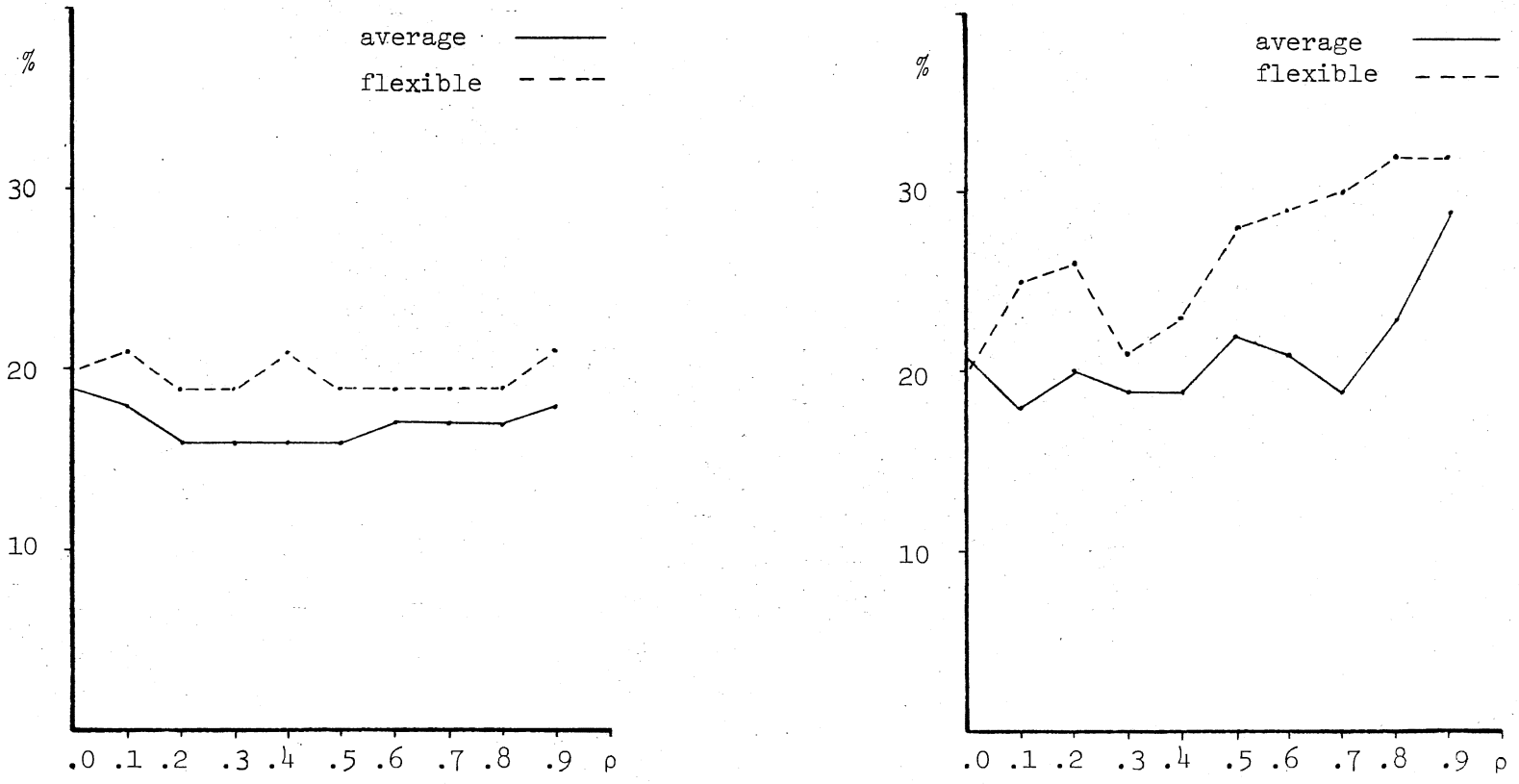


Figure 26. Using % Correctly Classified, a Graphical Comparison across ρ of Two Algorithms along $\gamma = 0.0$ where $\delta = 4.0$ with Two Different Splits

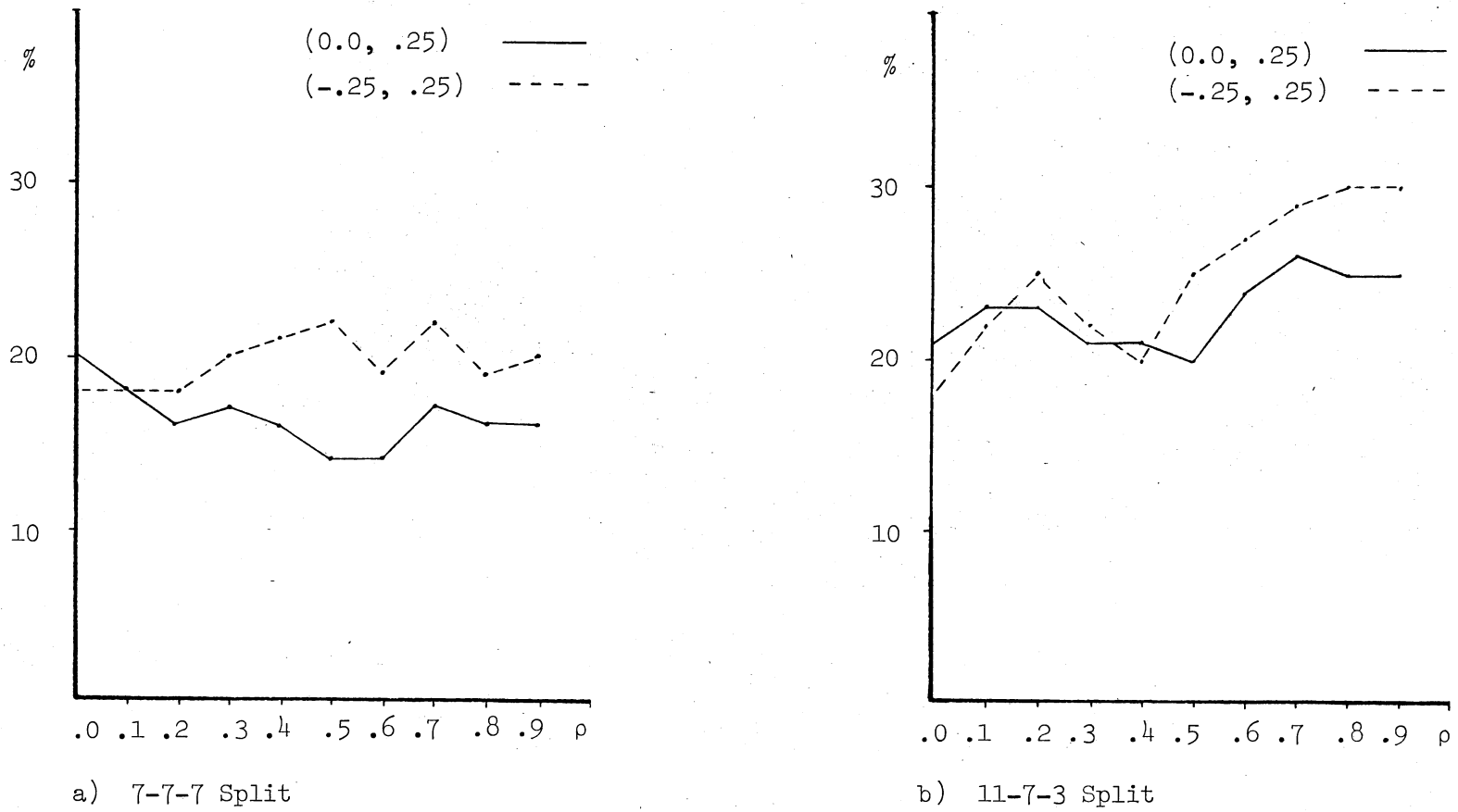
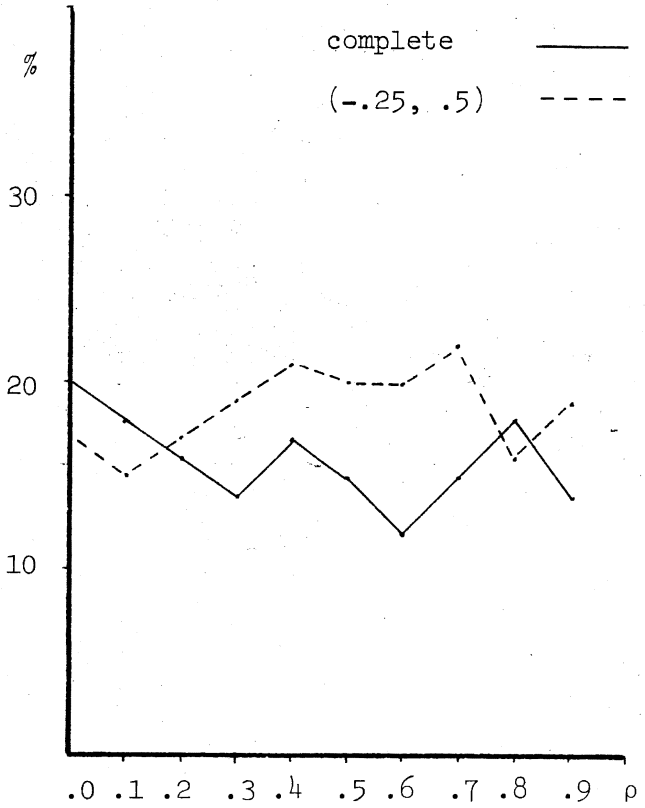
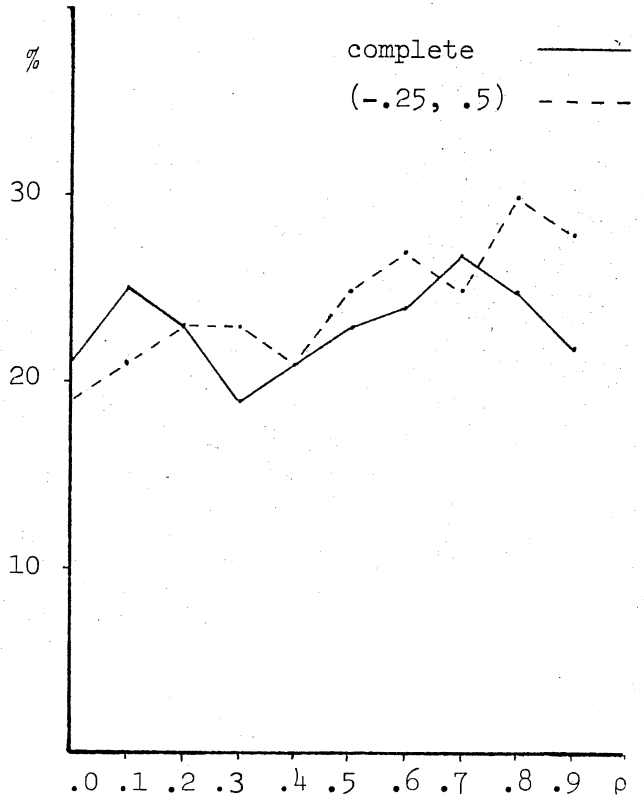


Figure 27. Using % Correctly Classified, a Graphical Comparison across ρ of Two Algorithms along $\gamma = .25$ where $\delta = 4.0$ with Two Different Splits



a) 7-7-7 Split



b) 11-7-3 Split

Figure 28. Using % Correctly Classified, a Graphical Comparison across ρ of Two Algorithms along $\gamma = .5$ where $\delta = 4.0$ with Two Different Splits

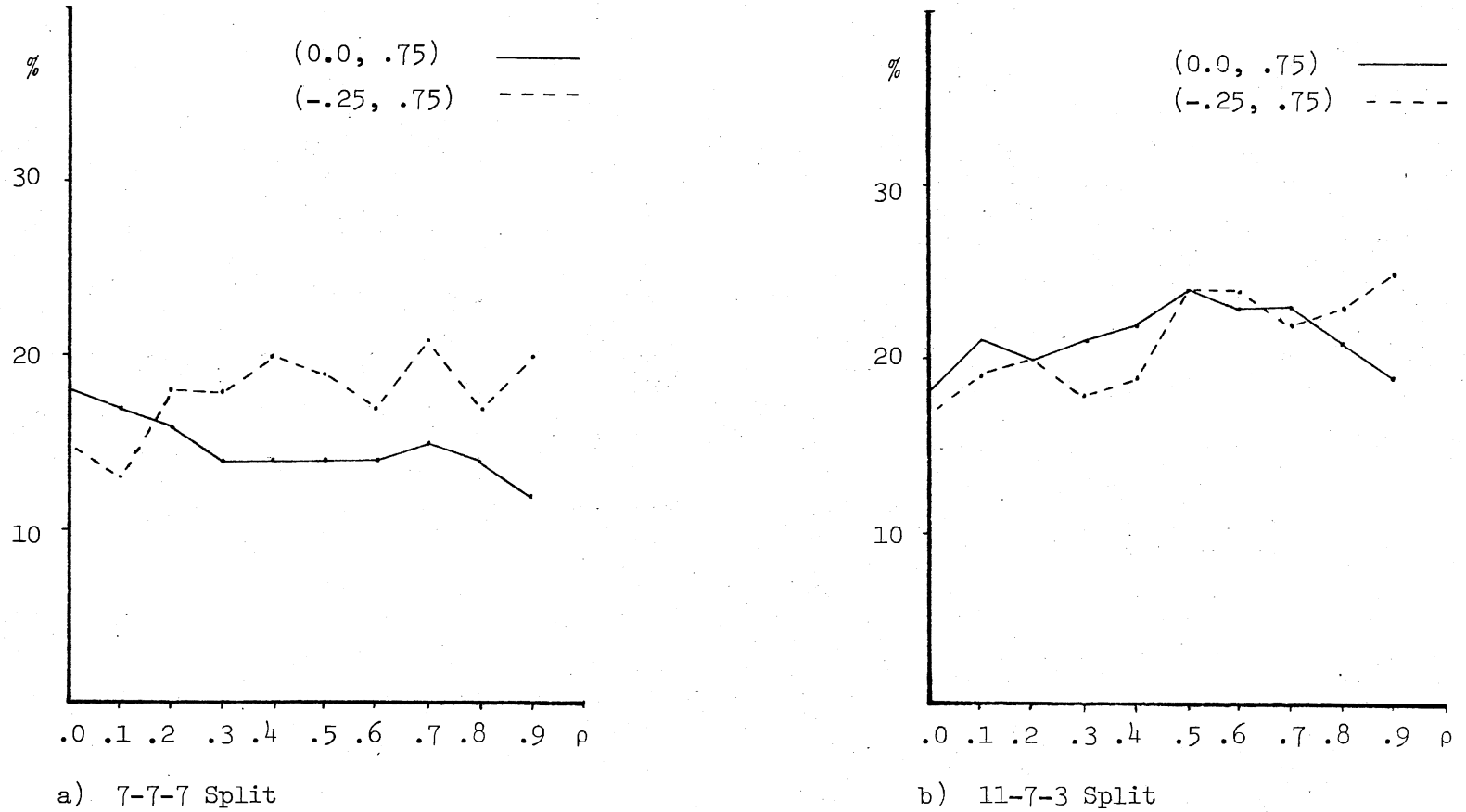


Figure 29. Using % Correctly Classified, a Graphical Comparison across ρ of Two Algorithms along $\gamma = .75$ where $\delta = 4.0$ with Two Different Splits

TABLE V

A COMPARISON ACROSS ρ OF SIX ALGORITHMS
ALONG $\beta = 0.0$ WHERE $\delta = 5.0$
WITH A 7-7-7 SPLIT

ρ		Single	(0, -.25)	Average	(0, .25)	Complete	(0, .75)
0	$\bar{c}$	.87171	.94667	.96290	.96252	.95614	.94705
	s_c	.13392	.08379	.06205	.06132	.06006	.07082
	%	42	62	66	66	57	51
.1	$\bar{c}$	.86271	.94190	.95795	.96357	.95695	.95681
	s_c	.13347	.09258	.06772	.06050	.06785	.06320
	%	39	63	65	66	60	58
.2	$\bar{c}$	.85962	.93690	.95467	.96052	.95690	.96171
	s_c	.13815	.09992	.07146	.05933	.07098	.05827
	%	38	62	62	61	61	60
.3	$\bar{c}$	.87548	.94552	.95700	.96329	.96029	.95571
	s_c	.12904	.09369	.06591	.05593	.06575	.06233
	%	40	64	61	61	60	56
.4	$\bar{c}$	.88438	.93790	.95467	.96314	.96743	.95776
	s_c	.13350	.11070	.06841	.05671	.04814	.06192
	%	43	63	59	60	62	58
.5	$\bar{c}$	.88681	.94352	.95648	.95600	.95743	.95457
	s_c	.14621	.09828	.06861	.06680	.06740	.06387
	%	49	61	61	59	59	57
.6	$\bar{c}$	.88876	.94300	.94552	.96243	.95529	.94357
	s_c	.14520	.09929	.09054	.06091	.06885	.08795
	%	49	60	57	62	57	55
.7	$\bar{c}$	.89052	.92814	.93881	.95405	.94457	.94990
	s_c	.14643	.11581	.10415	.06853	.07957	.06969
	%	51	55	53	57	54	54
.8	$\bar{c}$	.89648	.91824	.93614	.94738	.92248	.92448
	s_c	.12383	.11955	.10803	.07986	.09936	.09660
	%	48	51	54	57	43	45
.9	$\bar{c}$	.89929	.92871	.93800	.93243	.93405	.92829
	s_c	.11412	.10384	.09300	.08470	.07949	.08617
	%	48	54	52	49	45	43

TABLE VI

A COMPARISON ACROSS ρ OF SIX ALGORITHMS
ALONG $\beta = -.25$ WHERE $\delta = 5.0$
WITH A 7-7-7 SPLIT

ρ		(-.25,-.5)	(-.25,-.25)	Flexible	(-.25,.25)	(-.25,.5)	(-.25,.75)
0	$\bar{c}$	.94195	.96581	.96648	.96833	.96019	.95338
	s_c	.08604	.05629	.04841	.04657	.05231	.05631
	%	59	65	61	62	55	50
.1	$\bar{c}$	.95324	.96362	.97019	.96771	.96286	.95714
	s_c	.07265	.06129	.04640	.04927	.04858	.05278
	%	62	66	66	63	57	52
.2	$\bar{c}$	.94752	.96800	.96752	.96700	.96319	.95605
	s_c	.08401	.05144	.05365	.04441	.04696	.05264
	%	62	66	63	59	56	51
.3	$\bar{c}$	.94171	.97090	.96800	.96500	.96395	.96400
	s_c	.08818	.04985	.05184	.04858	.04971	.04884
	%	57	68	63	59	59	58
.4	$\bar{c}$	.94267	.96467	.96743	.96748	.96319	.96276
	s_c	.09771	.05466	.04490	.04650	.05009	.04970
	%	59	63	59	61	58	57
.5	$\bar{c}$	.94657	.96319	.97129	.96976	.96981	.96467
	s_c	.08193	.05580	.04487	.05165	.04842	.04591
	%	58	62	64	65	64	57
.6	$\bar{c}$	.94105	.96581	.97248	.96881	.96867	.96543
	s_c	.08731	.05572	.04424	.04692	.04825	.04897
	%	57	65	66	62	62	60
.7	$\bar{c}$	.94005	.96352	.96519	.96538	.96295	.96100
	s_c	.08707	.06039	.05171	.05148	.05214	.05245
	%	56	62	61	60	58	57
.8	$\bar{c}$	.93433	.96171	.96419	.95962	.95857	.96038
	s_c	.08639	.06093	.05068	.05427	.05650	.05409
	%	52	62	60	56	56	57
.9	$\bar{c}$	.93448	.95424	.96157	.96119	.95857	.95357
	s_c	.08883	.06481	.05463	.04886	.05251	.05712
	%	55	56	59	56	55	52

TABLE VII

A COMPARISON ACROSS ρ OF SIX ALGORITHMS
ALONG $\beta = 0.0$ WHERE $\delta = 5.0$
WITH AN 11-7-3 SPLIT

ρ		Single	(0, -.25)	Average	(0, .25)	Complete	(0, .75)
0	$\bar{c}$	.87486	.95157	.94243	.95181	.94976	.94743
	s_c	.15043	.07208	.07570	.06617	.07148	.08022
	%	37	55	49	52	52	54
.1	$\bar{c}$	.87552	.94724	.94424	.95824	.95595	.94805
	s_c	.14722	.07961	.07742	.05738	.06466	.07155
	%	36	53	49	54	53	51
.2	$\bar{c}$	.86819	.93619	.95400	.95519	.94743	.95043
	s_c	.15052	.09763	.07589	.06076	.06533	.06329
	%	35	52	60	56	50	52
.3	$\bar{c}$	.87500	.94248	.94110	.94833	.94986	.94519
	s_c	.14639	.09228	.09010	.07245	.06531	.07225
	%	37	56	54	54	53	50
.4	$\bar{c}$	.87419	.94348	.94443	.94119	.95090	.95124
	s_c	.15423	.09129	.09151	.09685	.07312	.06500
	%	39	58	57	55	55	55
.5	$\bar{c}$	.88448	.95176	.95810	.95590	.95843	.94457
	s_c	.15004	.08072	.06115	.08109	.07877	.08266
	%	43	59	55	61	65	54
.6	$\bar{c}$	.91062	.95290	.95043	.95610	.95890	.94695
	s_c	.12887	.06645	.08622	.08537	.08038	.08290
	%	46	57	58	61	65	59
.7	$\bar{c}$	.91776	.95776	.94671	.94629	.95419	.94843
	s_c	.11683	.05958	.08745	.09814	.08210	.07543
	%	48	57	55	58	60	55
.8	$\bar{c}$	.91738	.95290	.93714	.94257	.94367	.93438
	s_c	.09531	.06533	.09608	.09512	.08903	.09616
	%	41	53	51	55	54	52
.9	$\bar{c}$	.92833	.94500	.94514	.93871	.93014	.92457
	s_c	.07115	.06700	.09176	.10219	.10415	.09939
	%	39	48	53	54	53	47

TABLE VIII

A COMPARISON ACROSS ρ OF SIX ALGORITHMS
ALONG $\beta = -.25$ WHERE $\delta = 5.0$
WITH AN 11-7-3 SPLIT

ρ		(-.25,-.5)	(-.25,-.25)	Flexible	(-.25,.25)	(-.25,.5)	(-.25,.75)
0	$\bar{c}$	.94486	.95010	.95324	.95781	.96133	.95086
	s_c	.07604	.06937	.06276	.05372	.04670	.06602
	%	52	52	52	51	53	50
.1	$\bar{c}$	.94762	.95276	.96024	.96152	.95776	.94771
	s_c	.07525	.06684	.05923	.05341	.05296	.07142
	%	54	53	58	56	53	50
.2	$\bar{c}$	.95057	.95214	.95929	.95500	.95571	.94838
	s_c	.07398	.06778	.05769	.06174	.05759	.06808
	%	54	54	57	54	55	49
.3	$\bar{c}$	.95105	.95457	.95610	.95757	.95224	.94781
	s_c	.07425	.06058	.05702	.05892	.06074	.06484
	%	56	53	56	57	52	49
.4	$\bar{c}$	.94619	.95552	.95905	.95890	.95290	.94676
	s_c	.08210	.05995	.05478	.05803	.06238	.06916
	%	55	56	58	59	55	50
.5	$\bar{c}$	.94838	.95614	.96181	.96110	.96133	.95381
	s_c	.07736	.06100	.05659	.05853	.06057	.06261
	%	57	58	62	62	62	54
.6	$\bar{c}$	.95210	.96295	.96876	.96019	.96067	.95467
	s_c	.06818	.05552	.05231	.05979	.05844	.06466
	%	58	61	67	62	60	56
.7	$\bar{c}$	.95548	.96090	.96110	.95933	.95800	.94519
	s_c	.06690	.06083	.06079	.06312	.06161	.07307
	%	59	61	62	61	59	52
.8	$\bar{c}$	.95876	.95733	.95338	.95524	.95524	.93743
	s_c	.05935	.05960	.06549	.06476	.06417	.07902
	%	58	56	55	57	57	50
.9	$\bar{c}$	.95024	.95010	.94995	.95381	.94395	.93657
	s_c	.06541	.07012	.06868	.06529	.07275	.07800
		52	55	55	57	53	50

VITA

Janice Lynn DuBien

Candidate for the Degree of

Doctor of Philosophy

Thesis: COMPARATIVE TECHNIQUES FOR THE EVALUATION OF CLUSTERING METHODS

Major Field: Statistics

Biographical:

Personal Data: Born in Joliet, Illinois, June 5, 1947, the daughter of Mr. and Mrs. Harold L. DuBien.

Education: Graduated from Lincoln-Way Community High School, New Lenox, Illinois, in June, 1965; received the Bachelor of Science degree with a major in mathematics from Illinois State University, Normal, Illinois, in June, 1969; received the Master of Science degree with a major in statistics from Oklahoma State University, Stillwater, Oklahoma, in May, 1973; completed requirements for the Doctor of Philosophy degree at Oklahoma State University in July, 1976.

Professional Experience: Teaching experience in the Department of Mathematics at Bloom Township High School, Chicago Heights, Illinois, from September, 1969 to June, 1971; graduate teaching assistant, Oklahoma State University, from September, 1971 to July, 1976.