1. INTRODUCTION AND SUMMARY

In this report, we describe some techniques that can be utilized for verification of clusters in cluster analysis and also for detection of mixtures of probability density functions. Some analyses of simulated and real data sets are presented, including a preliminary analysis of the classical iris data of R.A. Fisher.

Let \( F_X(x) \) be a cumulative distribution function on \( E_k \), \( k \)-dimensional Euclidean space, \( k \geq 1 \). We assume that \( F_X(x) \) is absolutely continuous with respect to \( k \)-dimensional Lebesgue measure and denote the corresponding probability density function by \( f_X(x) \). Assume that a random sample of size \( n \) has been obtained from \( F_X(x) \) and denote the realizations by \( x_1, x_2, \ldots, x_n \).

In cluster analysis, similar objects are to be placed in the same cluster. We will interpret similarity as being close with respect to some distance on \( E_k \). The relationship between graph theory and cluster analysis has been described in the books by Bock (1974) and Godehardt (1990). Mathematical results related to those used here are given in Eberl and Hafner (1971), Hafner (1972), Godehardt and Harris (1995), Godehardt and Harris (1998) and Maehara (1990).

In order to proceed, we need to introduce some notions from graph theory.

2. GRAPH THEORETIC CONCEPTS
A graph $G_n = (V,E)$ is defined as follows. $V$ is a set with $|V| = n$ and $E$ is a set of (unordered) pairs of elements of $V$. The elements of $V$ are called the vertices of the graph $G$ and the pairs in $E$ are referred to as the edges of the graph $G$. With no loss of generality, we can assume $V = \{1,2,\ldots,n\}$. For the purposes at hand, we choose a distance $\rho$ on $E_k$ and a threshold $d > 0$. Then for $i \neq j$, place $(i,j) \in E$ if $\rho(x_i, x_j) < d$. Since $x_1, x_2, \ldots, x_n$ are realizations of random variables, the set $E$ is a random set and the graph $G$ is a random graph. In particular, these graphs are generalizations of interval graphs. Specifically, if $I_1, I_2, \ldots, I_n$ are intervals on the real line, then the interval graph $G(I_n)$ is defined by $V = \{1,2,\ldots,n\}$ and $(i,j) \in E$ if $I_i \cap I_j \neq \emptyset, 1 \leq i < j \leq n$. Thus, for the model under consideration, if $k = 1$, the intervals $I_i, i = 1, 2, \ldots, n$, are the intervals $[x_i - d/2, x_i + d/2]$. Let $V_m \subset V$ with $|V_m| = m < n$. $K_{md}$ is a complete subgraph of order $m$ if all $\binom{m}{2}$ pairs of elements of $V_m$ are in $E$. If $m = 1$, then $K_{1,d}$ is a vertex, if $m = 2$, then $K_{2,d}$ is an edge and if $m = 3$, then $K_{3,d}$ is called a triangle. A vertex has degree $\nu, \nu = 0, 1, 2, \ldots, n - 1$, if there are exactly $\nu$ edges incident with that vertex. If $\nu = 0$, then that vertex is said to be an isolated vertex.