N O T I C E

THIS DOCUMENT HAS BEEN REPRODUCED FROM MICROFICHE. ALTHOUGH IT IS RECOGNIZED THAT CERTAIN PORTIONS ARE ILLEGIBLE, IT IS BEING RELEASED IN THE INTEREST OF MAKING AVAILABLE AS MUCH INFORMATION AS POSSIBLE

NASA

Technical Memorandum 80721

A HILL-SLIDING STRATEGY FOR INITIALIZATION OF GAUSSIAN CLUSTERS IN THE MULTIDIMENSIONAL SPACE

John K. Park, Yung H. Chen,
Daryl B. Simons and Lee D. Miller

(NASA-TM-80721) A HILL-SLIDING STRATEGY FOR INITIALIZATION OF GAUSSIAN CLUSTERS IN THE MULTIDIMENSIONAL SPACE (NASA) 29 p HC A03/MF A01 CSCL 09B N80-31072
Unclas
G3/61 31318

July 1980

National Aeronautics and
Space Administration

Goddard Space Flight Center
Greenbelt, Maryland 20771

SEP 1980
RECEIVED
NASA STI FACILITY
ACCESS DEPT.

A HILL-SLIDING STRATEGY FOR INITIALIZATION OF GAUSSIAN CLUSTERS IN THE MULTIDIMENSIONAL SPACE[1]

John K. Park[2], Yung H. Chen[3], Daryl B. Simons[3], and Lee D. Miller[4]

Introduction

Many diverse techniques have been devised to discover structure within complex bodies of data by unsupervised fashion, i.e., cluster analysis (Ball, 1965; Cormack, 1971; Anderberg, 1973; Duran et al., 1974; Everitt, 1974). The techniques attempt to group data points, usually in a multidimensional space, into cluster such that all points within a cluster possess intrinsic similarity relatively distinct from the others.

Hence, application of the techniques to the data often reveals unexpected characteristics inhibited in the data structure. But it has often suffered from lack of adequate mathematical description, and either too many suboptimal solutions, or requirements of astronomical enumerations, in the course of searching for the optimal solution.

The objective of this paper was to present an approach for extraction of Gaussian clusters using discrete probability density estimates as the first step for an iterative unsupervised classification. The approach is based on the presumpti that there are parts of the feature space in which data populations are y dense, separated by parts of low density. An attempt was made to devis. a method suitable for processing a moderate volume of multivariate measurements, such as satellite multispectral scanner (MSS) data.

Parameterization for Clustering Function

Clustering is often the first step in analyzing a set of data whose characteristics have not yet been revealed. It is common to begin with the assumption of normal distribution if no knowledge about the data structure is available. In multivariate mixture distribution, the normality means multimodal Gaussian distribution in multidimensional space. Data surrounding each mode can be interpreted as a cluster. A group of data representing a real class may consist of two or more unimodal clusters and have multimodal distribution. Such data are divided into two or more subgroups so that the unimodal distribution can be applied to each subgroup.

---

1. This paper was presented at the joint Soil and Machine Processing of Remotely Sensed Data Symposia, Purdue University, West Lafayette, Indiana, June 3-6, 1980.
2. Earth Survey Applications Division, National Aeronautics and Space Administration/Goddard Space Flight Center (NASA/GSFC) Greenbelt, Maryland.
3. Colorado State University, Fort Collins, Colorado.
4. Texas A&M University, College Station, Texas.

Under the assumption of unimodal normality in a cluster, the probability density function is given by (Duda and Hart, 1973)

$$p_i(\underline{x}) = \frac{P_i}{(2\pi)^{d/2}|C_i|^{1/2}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu}_i)^T C_i^{-1}(\underline{x}-\underline{\mu}_i)} \tag{1}$$

where

P_i	=	a priori probability of cluster i (w_i)
C_i	=	d-by-d covariance matrix of cluster i
C_i^{-1}	=	inverse of the covariance matrix C_i
$\underline{x}$	=	pattern (d-component column) vector
$\underline{\mu}_i$	=	mean vector of patterns of cluster i
$(\ )^T$	=	tranpose of a matrix
d	=	dimension of feature space (integer)
e	=	base of natural logarithm.

This is the multivariate normal distribution function for the cluster called "w_i". The clustering process is carried out by finding all sets of cluster parameters: mean vector $\underline{\mu}_i$, covariance matrix C_i and a priori probability P_i for all the clusters. For a given set of N measurements, $\chi = \{\underline{x}_n\}_{n=1}^{N}$, the multivariate mixture probability $p(\underline{x})$ may be estimated and then postulated as the sum of all the cluster probabilities $p_i(\underline{x})$:

$$p(\underline{x}) = \sum_{\text{all } i} p_i(\underline{x}), \tag{2}$$

It may be computed from discrete measurement data by

$$\hat{p}(\underline{x}) \cong \frac{\text{sum of population in a volume element } \Delta V \text{ (or cell)}}{\text{total population (N)}} \tag{3}$$

This is the probability density of the mixture of all probable clusters. Decomposition of the mixture probability into a set of subgroups having unimodal distribution is the task of the clustering process. It is intuitive to divide the region into two parts by the boundary where

$$p_i(\underline{x}) = p_j(\underline{x}) \ , \ i \neq j \ . \tag{4}$$

Furthermore, it is likely to declare that $\underline{x}$ belongs to the cluster i (i.e., $\underline{x} \in w_i$) if

$$p_i(\underline{x}) > p_j(\underline{x}) \quad \text{for all } j \neq i \tag{5}$$

and, otherwise, $\underline{x}$ does not belong to the cluster i (i.e., $\underline{x} \notin w_i$). This is the maximum likelihood classification or decision rule (Duda et al., 1975). The region R_i is defined by the subspace where the inequality (Eq. 5) is satisfied. This intuition will be exploited to extract a cluster from the data set by a clustering function proposed as

$$G_i(\underline{x}) = \ln \frac{p(\underline{x})}{p_i(\underline{x})} \tag{6}$$

where $\ln$ denotes natural logarithm.

The usefulness of the clustering function for the maximum likelihood decision rule may be seen in the following properties:

1. $G_i(\underline{x}) = 0$ for the distribution of a single (unmixed) class, (7)

2. $G_i(\underline{x}) \geq 0$ for any mixture distribution of two or more different clusters, (8)

3. For a two-cluster mixture distribution,

   a. $G_1(\underline{x}) = G_2(\underline{x}) = \ln 2$ on the boundary between the clusters, (9)

   b. $\left.\begin{array}{l} G_i(\underline{x} \in R_i) < \ln 2 \\ G_i(\underline{x} \notin R_i) > \ln 2 \end{array}\right\} i=1,2.$ (10)

PROOF:

1. It is evident by the definition since $p(x)=p_i(\underline{x})$ for the single class data,

2. $p(\underline{x}) = \sum_{\text{all } j} p_j(\underline{x}) \geq p_i(\underline{x})$.

   then, (q.e.d.)

   $$G_i(\underline{x}) = \ln \frac{p(\underline{x})}{p_i(\underline{x})} \geq 0.$$

3a. On the boundary between the two clusters,

$$p_1(\underline{x}) = p_2(\underline{x}) = p_i(\underline{x}).$$

$$G_i(\underline{x}) = \ln \frac{p_1(\underline{x}) + p_2(\underline{x})}{p_i(\underline{x})} = \ln \frac{2p_i(\underline{x})}{p_i(\underline{x})} = \ln 2 \text{ for } i=1,2.$$

That is,

$$G_1(\underline{x}) = G_2(\underline{x}) = \ell n2. \qquad \text{(q.e.d.)}$$

3b. Let $i \neq j$. And the maximum likelihood decision rule shows

$$p_i(\underline{x}\epsilon R_i) > p_j(\underline{x}\epsilon R_i).$$

Therefore,

$$G_i(\underline{x}\epsilon R_i) = \ell n \frac{p_i(\underline{x}\epsilon R_i) + p_j(\underline{x}\epsilon R_i)}{p_i(\underline{x}\epsilon R_i)}$$

$$= \ell n \left[1 + \frac{p_j(\underline{x}\epsilon R_i)}{p_i(\underline{x}\epsilon R_i)} \right] \qquad (11)$$

$$< \ell n2.$$

Similarly, for the two-cluster mixture

$$p_j(\underline{x}\notin R_i) = p_i(\underline{x}\epsilon R_j) < p_j(\underline{x}\epsilon R_j) = p_j(\underline{x}\notin R_i).$$

$$G_i(\underline{x}\notin R_i) = \ell n \left[1 + \frac{p_j(\underline{x}\notin R_i)}{p_i(\underline{x}\notin R_i)} \right]$$

$$> \ell n2. \qquad \text{(q.e.d.)}$$

The last property of the clustering function suggests that a feature space R can be divided into two regions: 1) a region belonging to cluster i and 2) another out of the region, in the case of two-cluster mixture, as such $\underline{x} \in w_i$ (cluster i) if

$$G_i(\underline{x}) < \ell n2 \qquad (12)$$

and otherwise, $x \notin w_i$. This criterion will be utilized in this study for the extraction of one-cluster data from the whole set of data. It requires only knowledge of a set of the parameters for a single cluster each time. Elements of a prospective cluster can be extracted from the set of data without knowing characteristics of the other clusters based on this criterion. This fact is the beauty of the clustering function $G_i(\underline{x})$.

Hill-Sliding Strategy

The clustering function, however, cannot be evaluated unless the set of cluster characteristic parameters are estimated. The first problem in clustering is to find good initial estimates of the parameters employed in most cases. It is intuitively viewed that a cluster has a mode, which has the highest probability

density in the cluster. The location of a cluster mode depends on the characteristics of distribution type or governing law of the distribution, but it is usually observed near the gravitational center (centroid) of the cluster (Fig. 1).

Suppose this presumption is acceptable in a set of data to be analyzed. Then at least one candidate mode which has the highest probability density can be picked up. Such a mode initiates the first clustering by fusing all probability cell points that may be categorized into one cluster.

A major problem faced in clustering is that the types of data distribution are generally not known in advance; thus each cluster may have a different characteristic shape in its distribution. Due to the absence of prior knowledge on characteristics of expected clusters, each cluster was initiated with the assumption of isotropic normal distribution, at least in the immediate neighborhood of a mode candidate. The group of data initially coalesced into a cluster reflects the distribution characteristics of the forming cluster in some degree since the theoretical shape of its probability contour surface maintains near the mode as well as throughout the region of a cluster. Distortion of its shape may be observed usually in the regions of its tails or valleys where distributions are affected by neighbor clusters. A simple Euclidean distance measure between measurement points can be used in clustering data near an apparent centroid without introducing large trial errors. The question is where to terminate the fusing process to avoid picking up data points probably originated from different clusters. The values of parameters for termination of initial fusion process are threshold values of clustering.

One of the threshold parameters is derived in d-variate space. A differential volume $\Delta V(r)$ at radius r from a cluster center is defined by

$$\Delta V(r) \propto r^{d-1}\,\Delta r \tag{13}$$

where Δr is a small segment of the radius r and symbol $\propto$ denotes the proportionality. This differential volume can be viewed as a hyper-shell (called simply shell hereafter) enclosed by two concentric hyperspherical surfaces (Fig. 2). A series of concentric shells around a mode are drawn with increasing r by Δr. The number of cluster elements, $\Delta N(r^2)$, in a shell is proportional to the volume of the shell multiplied by average population density in the shell:

$$\Delta N(r^2) \propto \exp\left(-\frac{r^2}{2\sigma^2}\right)\Delta V(r) \tag{14}$$

The exponential term in Eq. 14 is that of an isotropic normal distribution with standard deviation σ. This relation can be rearranged by employing squared radius r^2 as

$$\frac{\Delta N}{r^{d-2}\,\Delta r^2} \propto \exp\left(-\frac{r^2}{2\sigma^2}\right) \tag{15}$$

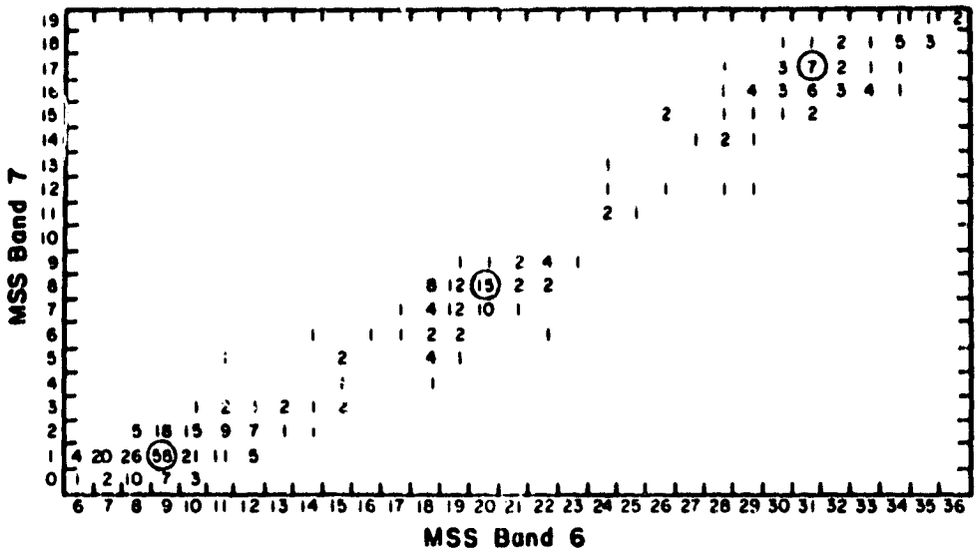
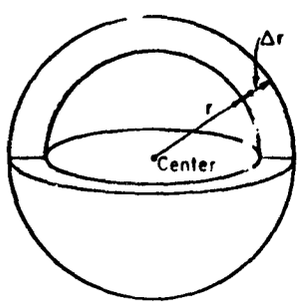


Fig. 1. Bivariate Population Distribution of 400 Landsat Data and the First Three Most Probable Candidates for Modes of Clusters (in circles). These typical example data were taken from the Landsat data over a portion of Korean west coast (Park and Miller, 1978). The numbers are occurrences of bivariate data in each block formed by both discrete MSS bands 6 and 7 data. Visually, three clusters and their probable mode positions were distinguished without difficulty.



Example Shell Volumes:

$\Delta V_{4\text{-}d} \propto r^3 \Delta r \propto r^2 \Delta r^2$

$\Delta V_{3\text{-}d} \propto r^2 \Delta r \propto r \Delta r^2$

$\Delta V_{2\text{-}d} \propto r \Delta r \propto \Delta r^2$

$\Delta V_{1\text{-}d} \propto \Delta r \propto \Delta r^2 / r$

Fig. 2. A Differential Volume (Hyperspherical Shell) in Three-Dimensional Space. Example formulas are given for one- through four-dimensional shells. Subscript i-d denotes i-dimension.

where the term in the left-hand side is a generalized mean population density in a shell at distance r. The parameter σ^2 is the variance in the population distribution. The right-hand side of this relationship is a monotonically decreasing function with increasing r^2 (Fig. 3).

Plots of $\ln\left(\frac{\Delta N}{r^{d-2}\Delta r^2}\right)$ vs. r^2 may reveal a family of straight lines having the slope of $-\frac{1}{2\sigma^2}$ (Fig. 3b). A group of data can be considered as originating from the same class if the estimates of shell population densities fall near a straight line. The slope of shell population data will remain fairly constant near the center of a cluster, but may change significantly when populations of other clusters enter into the shell. The squared radius at which the first significant change of the slope is detected is the threshold value (r_t^2) for initiation of clustering. Such a change occurs when the sequential searching point attempts to cross a valley and then to climb a hill consisting of other cluster data (Fig. 4).

There are several possibilities which may introduce slope changes in the case of anistropic distribution of the data. A plot of two-dimensional population distribution will be utilized to visualize some of these causes (Fig. 4). One of them is the case where a cluster is well separated from the others even though the isotropic assumption is employed and the threshold value covers nearly the whole region where most of cluster data are located (for example, cluster A in Fig. 4). Another case is when the group of one-cluster data are closely neighbored with the others (for example, cluster B in Fig. 4). Considerable overlaps between clusters may exist in this case.

The threshold value r_t^2 was given by the largest value of r^2 satisfying that

$$\theta(r^2) < \mathrm{Min}\ (\theta_c,\ 0) \tag{16}$$

where

$$\theta(r^2) \equiv -\frac{1}{2\sigma^2}$$

$$= \frac{1}{r^2}\ \ln\left[\frac{\Delta N}{r^{d-2}\ \Delta r^2}\right] \tag{17}$$

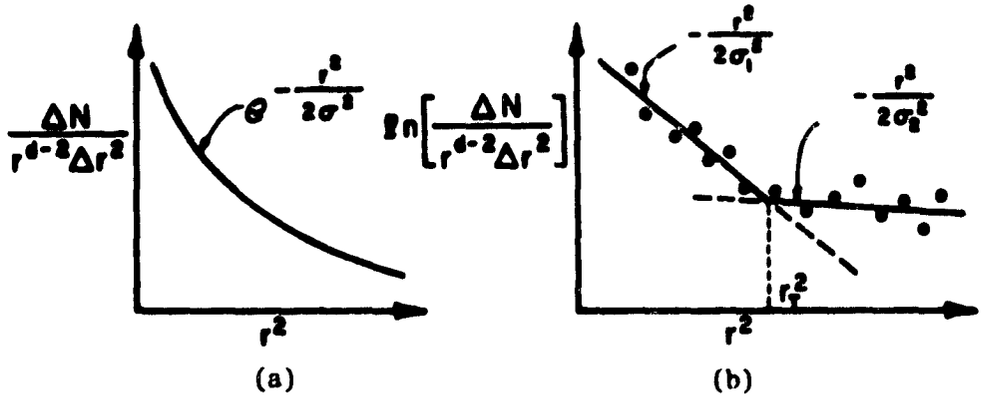


Fig. 3. A Curve of Eq. 15 and Data of a Hypothetical Two-Cluster Mixture. Data of a unimodal isotropic distribution yields an exponential curve shown in (a). Discrete data of a two-cluster mixture may produce a plot shown in (b), where two straight lines are approximate moving averages of two parts divided by r_t^2. The first straight line represents the population distribution of the first cluster with the parameter σ_1^2. r_t^2 will be used as a threshold value for the cluster.

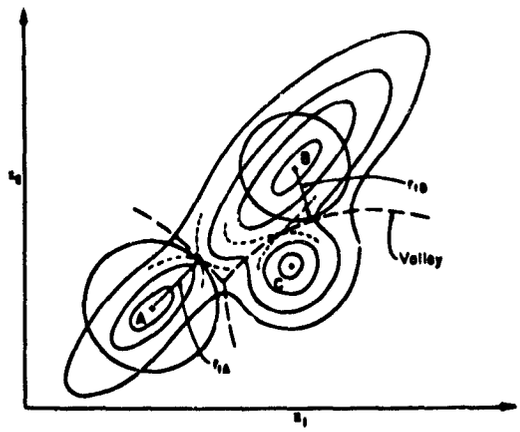


Fig. 4. Contours of a Mixture Probability Density Function (or Population Distribution) in a Two-Dimensional Feature Space. The mixture p.d.f. consists of three unimodal p.d.f.'s. A, B and C points are the modes of the clusters. r_{tA} or r_{tB} may be one of the threshold values estimated by the method shown in Fig. 3 (b). They are interpreted as the shortest Euclidean distance to a valley, which is the natural boundary between the cluster and its neighbor one.

$$\theta_c = \bar{\theta} + f_\theta S_\theta$$

and

θ_c = updated critical slope up to the previous estimate

$\bar{\theta}$ = updated average slope of all previous estimates

f_θ = positive empirical constant (about 2)

S_θ = updated standard deviation of θ.

The first criterion ($\theta<\theta_c$) given in Eq. 16 prevents other cluster cells from merging into the cluster under formation. The second criterion ($\theta<0$) distinguishes the cluster cells from the others which may cause violation of the normal distribution laws when they merge into the clusters. Values of the slope parameter must be less than zero for normally distributed data. Once the threshold value is found, a new initial cluster is formed by fusing cells closer than the distance corresponding to the threshold value. This initial cluster leads to computation of a set of parameters which will characterize the early stage of the cluster.

Updating the Initial Cluster

A group of data immediately surrounding (i.e., within the threshold value r_t^2 from) a mode candidate formed an initial cluster under the assumption of the isotropic normal distribution. The parameters estimated for this initial cluster reflects to a certain degree the distribution characteristics even at its early stage, since the shape of its probability contour is retained throughout all the region of a cluster. Hence, the isotropic assumption is no further needed when the clustering function given by Eq. 6 is evaluated.

To compute the clustering function, the a priori probability of the cluster should be known as well as the other parameters. It is one of the uncertain parameters especially at this very initial step. The a priori probability of a cluster in the mixture distribution is computed by

$$P_i = \frac{\text{population in cluster } i}{\text{total population}} \tag{19}$$

This value changes whenever any data are merged into or deleted from a cluster. Other changing parameters are the position vector of the cluster centroid (mean) and covariance matrix. All of these parameters as well as random components of the data, contribute to fluctuation of G_i estimates.

It was shown that the expected value of the clustering function would be smaller than $\ln 2$ for any cell data within a well-defined cluster region. Values of $G_i(\underline{x})$ computed at this stage, however, may not be close to those expected at the final stage. They may range from negative to large positive values mainly because initial estimates of cluster characteristic parameters deviate from reasonable values and/or because the data contain random or noisy components.

To allow for a certain level of fluctuations in G_i estimates, especially at a formative stage, a flexible criterion value rather than $\ln 2$ as in Eq. 9 is employed as

$$G_c = \bar{G}_i + f_G * S_G \tag{20}$$

where

G_c = critical value of $G_i(\underline{x})$

$\bar{G}_i$ = average value of $G_i(x)$

f_G = empirical constant (about 2.)

S_G = standard deviation of $G_i(\underline{x})$

A cell is tested for the membership of the cluster under formation by the criterion:

$$G_i(\underline{x}) \leq \mathrm{Max}(G_c, \ln 2) \tag{21}$$

It will be rejected if this inequality is not satisfied. The criterion value is continuously updated as a new member is merged into the cluster. The empirical constant f_G as well as f_θ in Eq. 18 is not a sensitive parameter, but selection of its value depends upon the detail required in cluster divisions in the end result.

The criterion test is always applied first to the cell having the highest probability density among the remaining cells and then the next test is performed. In this way the present estimate of $G_i(\underline{x})$ would be very close to those of the sample points that just joined the group in the previous steps, if the point is a strong candidate of the cluster. Otherwise, it would be of far greater value than those in the region, especially at the earlier stage of cluster formation. In this case it is thought that the point is picked up from a hill side of another group (Fig. 1). The way to jump up and down from a hypothetical hill to the others creates distinct distance (G_i value in a precise term) gaps between points within the cluster being formed and that of other cluster candidates. These distance gaps allow gradual updating of cluster characteristic parameter values by first merging cells only closer to the centroid. Gradual updating is important in this approach since estimated parameters at the earlier stage have larger uncertainty factors than those at later or final steps. Testing membership candidacy for each point, merging or rejecting, and updating of the parameters continue until the last point is checked. After all the above-mentioned steps are processed, searching for the next cluster is repeated. The test stops if no single cell element is left over (or if the maximum number of clusters set up in the program is reached). This procedure is similar in manner to "hill-sliding." One who is sliding down from the highest point on a hill will eventually

arrive at the bottom. Geometrical interpretation of the algorithm developed here in multidimensional space is not directly comparable to the pathway of hill-sliding by an object. But the general procedure may be considered as a "hill-sliding" aspect. Actual paths from the present position to the next lower density point will be zigzag motion due to randomness of the estimated probability density function in the discrete space. The pathway is always descending or leveling, ending at the bottom of a valley.

Cluster Compactness

Most of the measures to examine goodness of clustered results give relative comparisons on the basis of original data structure or among clusters themselves. The sum of squared errors within clusters and divergence between clusters are typical examples of such measures. The former evaluates the deviation of cluster samples from each centroid, while the latter measures separability between two clusters. In either example, the quantities of the measures increase with increasing dimensions of the feature space (Tou et al., 1974). Thus, difficulties are encountered in standardizing criteria of these measures. It is desirable to formulate a cluster measure independent of the number of variables and the number of sample data employed for clustering.

A measure is proposed to evaluate the goodness of an individual cluster as

$$L_i = \left[\frac{|C_i|}{N_i - d}\right]^{1/d} \Bigg/ \left[\frac{|T|}{N-d}\right]^{1/d} \tag{22}$$

where

L_i = compactness of cluster i

N_i = population of cluster i

d = dimension of feature space

T = total scatter matrix of the data

and the determinants ($|C_i|$ and $|T|$) of both matrices, C_i and T, exist.

The cluster compactness parameter L_i is a dimensionless quantity. The denominator is constant for a given set of data and has the dimension of length-square. It can be considered a characteristic value of the data (say C_L). The determinant of a scatter matrix is proportional to the product of the variances in the direction of the principal axes, which are defined by the canonical transform of the scatter matrix. It is the volume of a hyper-ellipsoid defined by the unit Mahalanobis distance (i.e., $(\underline{x}-\underline{\mu}_i)\ T_{C_i}{-1}\ (\underline{x}-\underline{\mu}_i)=1$) from the cluster centroid. The volume measures the average scatterness (or squared Euclidean distance) of the pattern vectors within the cluster around their mean pattern vector. The length between the centroid and a point on the hyper-ellipsoid

may be interpreted as the mean squared-error in the direction of the feature space. For this reason, the hyper-ellipsoidal volume defined by the determinant of a cluster covariance matrix will be called simply the "scatterness volume" of the cluster. The value in the bracket of Eq. 22 is approximately proportional to the average volume per cluster element if the number of elements defining the covariance matrix is sufficiently larger than that of dimensions. Subtraction by d from N or N_i in the denominator of each bracket is devised for the unbiased estimation of the parameter. Note that pattern vectors less than or equal to the number of dimensions cannot form any hypervolume and the covariance matrix of those pattern vectors is always singular (Duda et al., 1973). However, the value of d may be any nonnegative value, if desired for the purpose of defining the parameter only.

Analysis of the Landsat data in this study indicates that clusters having a compactness parameter less than 0.4 are distinctly separable from others and that those with a parameter larger than 1 are scattered around in a region rather than distributed normally.

Separability between Clusters

Distinctness of a cluster against the rest of the data has been evaluated in terms of various measures, such as Mahalanobis distance and divergence (Duran and Odell, 1974). Mahalanobis distance was introduced for a measure of metric distance between two population centroids (Atchley et al., 1975). Its original definition is different from the concept employed here, which is a distance measure between a pattern vector and a cluster centroid. The original formula uses a pooled covariance matrix of two distributions. Application of this formula to all possible pairs of classes requires considerable computational time if the number of classes is large.

Divergence is another commonly used measure of dissimilarity between two distributions (Tou et al., 1974; Swain, 1972). It is defined by the sum of expectations of log-likelihood ratios in favor of one class against the other:

$$D_{ij} = \int_{\underline{x}} [p_i(\underline{x}) - p_j(\underline{x})] \ln \frac{p_i(\underline{x})}{p_j(\underline{x})} d\underline{x} \qquad (23)$$

The divergence is inferred as the total average information for discrimination between two classes. Higher values of the divergence estimates indicate better separability between the pair. It also possesses other interesting properties (Tou et al., 1974; Swain, 1972).

The divergence is used in this paper to analyze the clustering performance. The major reason for employing the parameter is that it can be computed by a simpler formula under Gaussian assumption. For two Gaussian classes with unequal a priori probabilities, Eq. 23 is reduced to

$$D_{ij} = \tfrac{1}{2}\,\mathrm{tr}\,[(P_iC_i - P_jC_j)\,(C_j^{-1} - C_i^{-1})]$$
$$+ \tfrac{1}{2}\,\mathrm{tr}\,[(P_iC_i^{-1} + P_jC_j^{-1})\,(\underline{\mu}_i - \underline{\mu}_j)\,(\underline{\mu}_i - \underline{\mu}_j)^T]$$
$$+ (P_i - P_j)\,\ln \frac{P_i|C_j|^{\frac{1}{2}}}{P_j|C_i|^{\frac{1}{2}}}, \qquad (24)$$

where tr denotes the trace of the matrix in the bracket. This is an extended formula of the relationship usually seen in the literature (Tou et al., 1974) in the case of two distributions with different mixing proportions. It is noteworthy that the Mahalanobis generalized distance is the divergence between two Gaussian populations with unequal mean vectors but equal a priori probabilities and covariance matrices (Tou et al., 1974). The divergence is normalized by the sum of two class a priori probabilities as $2D_{ij}/(P_i + P_j)$. The additive property of divergence for independent variables indicates that no universal value of a divergence criterion is acceptable for any combinations of multivariate measurements. It is desirable to reduce the effects of dimensionality as well as the sample size in cluster analysis. For this reason, the estimates of divergence divided by the entropy of the data is used in this study, whenever any comparison is made regarding divergence. The entropy $E(\chi)$ is a statistical measure of uncertainty defined by (Young et al., 1974)

$$E(\chi) = \int_{\underline{x}} p(\underline{x})\,\ln\,[1/p(\underline{x})]\,d\underline{x}. \qquad (25)$$

It is interpreted as the expected value of an information unit, $\ln[1/p(\underline{x})]$, that is, the average uncertainty of the information source. As indicated by its functional form similar to that of divergence, Eq. 23, the entropy possesses properties similar to those for divergence. The normalized divergence to the entropy given by

$$G_{ij} = \frac{2D_{ij}}{(P_i+P_j)\,E(\chi)} \qquad (26)$$

is comparable in any combination of variables. This value can determine relative separability of one cluster against the other regardless of the number of variables employed. Higher values indicate distinctive separability between the pair of clusters while smaller ones mean high resemblance of the pairs in their data characteristics.

On the Overall Objective of Partitioning

Remote sensing data of natural scenes may contain countless subcategorical information on natural land-cover/land-use classes. One of the best partitioning in an established mathematical frame may not satisfy a user (or analyst)

who desires the class categorical information at a certain level. Tuning of the mathematical goal at a user's desired level is not easily achievable by a numerical scale. Existence of various levels for classification schemes (Anderson et al., 1976) inevitably introduces heuristic parameters to obtain the desired level of the resultant classification or clustering.

To evaluate the performance of clustering, the following overall objective function was set forth:

$$F = \sum_{i=1}^{I_c} (N_i - d) L_i^d \tag{27}$$

subject to

$$G_i(\underline{x}_n \in w_i) \le G_j(\underline{x}_n \in w_i) \text{ for all } i, j \text{ and } n, \tag{28}$$

$$1 \le I_c, \tag{29}$$

$$0 < L_i \le L_c \text{ if } \operatorname{Min} G_{ij} < D_s \text{ for all } i \text{ and } j, \tag{30}$$

$$M_c < M_i \le N \text{ for all } i, \tag{31}$$

where I_c, D_s, L_c and M_c are the number of resultant clusters, the minimum acceptable value of the normalized divergence, the maximum acceptable value of the compactness parameter, and the minimum number of probability cells in a cluster, respectively. M_i is the number of cells in cluster i. The minimum number M_c of cells in a cluster should be larger than the number of variates (dimensions) d, so that a covariance matrix might not be singular. This is a better statement than that $N_c < N_i \le N$ where N_c is the minimum number of identities required in a cluster. The reason is that $M_i < N_i$ and hence it gives better assurance of a covariance being nonsingular. Note that a covariance matrix is always singular if $N_i \le d$ or $M_i \le d$ (Duda et al., 1973). Parts of the constraints: 1) $I_c \ge 1$, 2) $L_i > 0$, and 3) $M_i \le N$ are self-evident and there is no requirement for specification of these criteria in the algorithm. However, an investigator may input any other desired values which do not exceed the limits as parameters. It is also worthwhile to note that cluster or class identities less than ten times the dimensionality d will usually lead to an increase in probability of error if predictions are made based on their covariance matrices (Ball, 1965). The constraint Eq. 28 is equivalent to the decision rule of the maximum likelihood classification (Eq. 11), since the only other variable in clustering function $G_i(\underline{x}_n)$ defined by Eq. 6 is $p(\underline{x}_n)$, which is common in both sides. Hence, the inequality, Eq. 28, can be called the "maximum likelihood constraint" for each data point.

The objective function F can be expressed in terms of covariance matrices:

$$F = \frac{N-d}{|T|} \sum_{i=1}^{I_c} |C_i| \tag{27a}$$

Minimizing F is equivalent to minimizing the sum of the determinants of individual cluster covariance matrices. Therefore, the objective of clustering is to obtain a partitioning of the data which minimizes the sum of cluster scatterness volumes under the imposed constraints. The objective function is generally nonlinear and its usual multidimensional form cannot be described in easily manageable terms.

There are substantial differences between the present formulation and those which use frequently-cited clustering criterion function $|W|$, where

$$W = \sum_{i=1}^{I_c} C_i \ . \tag{32}$$

The simple algebraic sum of all the cluster covariance matrices, W, is commonly referred to the total intragroups (or pooled-within clusters) scatter matrix (Friedman et al., 1967; Fukunaga et al., 1970; Duda et al., 1973). It has been shown that the determinant of the matrix is invariant to nonsingular linear transformations of the data and is able to produce well-definable natural cluster boundaries when it is used as a clustering criterion (Fukunaga et al., 1970). The determinant of the scatter matrix alone is of no use as a clustering criterion function if the number of clusters is not known in advance, since more subdivisions of the data space tend to reduce the value of the determinant. An essential difference between the present objective function F and the determinant of the total intragroups scatter matrix, $|W|$, as a clustering criterion comes from the fact:

$$|W| = \left| \sum_{i=1}^{I_c} C_i \right| \neq \sum_{i=1}^{I_c} |C_i| \ .$$

The determinant $|W|$ has been used as a measure of compactness of the clusters (Duda et al., 1973), but this interpretation is somewhat misleading. The two simple examples (Fig. 5) illustrate the inappropriateness of using $|W|$ as a clustering objective function or an overall cluster compactness measure.

$$C_1 = C'_1 = C_2 = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}, \qquad C'_2 = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix},$$

$$W = C_1 + C_2 = \begin{pmatrix} 2 & 0 \\ 0 & 8 \end{pmatrix}; \qquad |W| = 16,$$

$$W' = C'_1 + C'_2 = \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}; \qquad |W'| = 25,$$

$$|C_1| + |C_2| = |C'_1| + |C'_2| = 4 + 4 = 8.$$

The first case is that two-dimensional covariance matrices of two clusters are identical except for their locations, and the second that the two have the same scatterness volumes (determinants) but different orientations and locations.

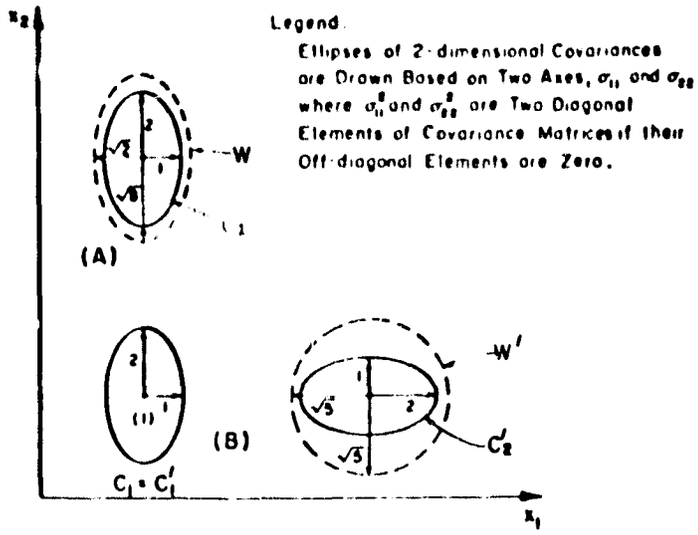


Fig. 5. Illustration of the Total Intragroups Scatter Matrices for Two Separable Cluster in Two-Dimensional Space. Each cluster has the same scatterness volume but different mean (centroid) from the other in either case. The pooled covariance matrices $W(=C_1 + C_2)$ and $W'(=C_1' + C_2')$ are different from each other since cluster orientations are not the same in both cases, even though $|C_1| = |C_2| = |C_1'| = |C_2'|$.

Under the assumption that two clusters are completely separated in both cases, their total intragroups scatter matrices have different shapes and determinant values. The first case yields smaller determinant values of the resultant scatter matrix than the latter does. This indicates that minimizing the determinant of the total intragroup scatter matrix W forces all the clusters to be partitioned in the shapes and orientations as similarly as possible. Any set of separable clusters would not make much difference whatever their natural shapes or orientations. This is another drawback in using $|W|$ criterion for clustering. The present clustering formulation has been devised to circumvent these difficulties by employing the sum of the determinants of individual cluster covariance matrices as the objective function. The set of constraints has provided some guidelines to overcome various undesirables aspects commonly encountered in clustering the heterogeneous natural scene data in this formulation.

The global solution to this optimization problem may be found by a systematic but exhaustive enumeration of all partitioning alternatives. Search of the solution by such an enumeration is often not permitted due to requirement of

excessive computation and memory storage for a large volume of data. It is noted that the objective function is proportional to (N-d). Hence, a data partition index to evaluate the clustered results is proposed as a sample-size-independent indicator:

$$PI = [F/(N-d)]^{1/d}$$

$$= \left(\frac{1}{|T|}\sum_{i} |C_i|\right)^{1/d} \tag{33}$$

This index is an overall measure of the goodness for the data partitioning. The smaller the index value is, the better optimal partitioning is achieved.

Results and Discussion

The proposed cluster initialization method was programmed and tested using the Landsat data of May 11, 1976, over Chippewa River Basin, Wisconsin. Sample data of the Multi-Spectral Scanner (MSS) bands 4 and 5 were extracted from the satellite's computer-compatible tape by the Landsat Mapping System of Colorado State University, Fort Collins, Colorado (Park et al., 1979). The data covered two different locations of the study area. Of the total 200 pixels (picture elements), 111 population cells were identified (Fig. 6). The test was made with the input parameters of $f_\theta = 2.7$ and $f_G = 1$ for the criterion functions, Eqs. 18 and 20, respectively. Fifteen initial clusters were formed (Fig. 7), but one of them which did not meet the criterion, Eq. 31, was excluded from various cluster evaluations (Table 1).

The results suggested that there would be

a compact cluster if $L_i < 0.2$,

a scattered cluster if $L_i > 0.4$, and

a well-separated cluster from the others if $G_{ij} > 50$ for all $j \neq i$.

Several test runs with varying input values of both parameters yielded similar partitioned patterns of the data and most of significant clusters (like clusters 1 through 5) were identified.

Summary

A method was developed to extract significant Gaussian clusters on the basis of discrete multivariate probability density estimates. The key idea was that a proposed clustering function could distinguish elements of a cluster under formation from the rest without knowledge of the other clusters. The algorithm described here showed effectiveness in extracting Gaussian-type clusters from Landsat MSS data. The partitioning obtained by this technique was not optimal, but it could be used as reasonable input data to other iterative clustering programs for further improvement.

A dimensionless cluster compactness parameter was set forth as a universal measure of cluster goodness and used favorably in test runs. Separability between clusters was examined by employing a normalized divergence which was defined by the divergence divided by the entropy of the entire sample data. The overall clustering objective function and the partition index were formulated in terms of cluster covariance matrices. They were good indicators of the overall achievement for evaluation of the partitioned results obtained under various conditions or at different stages.

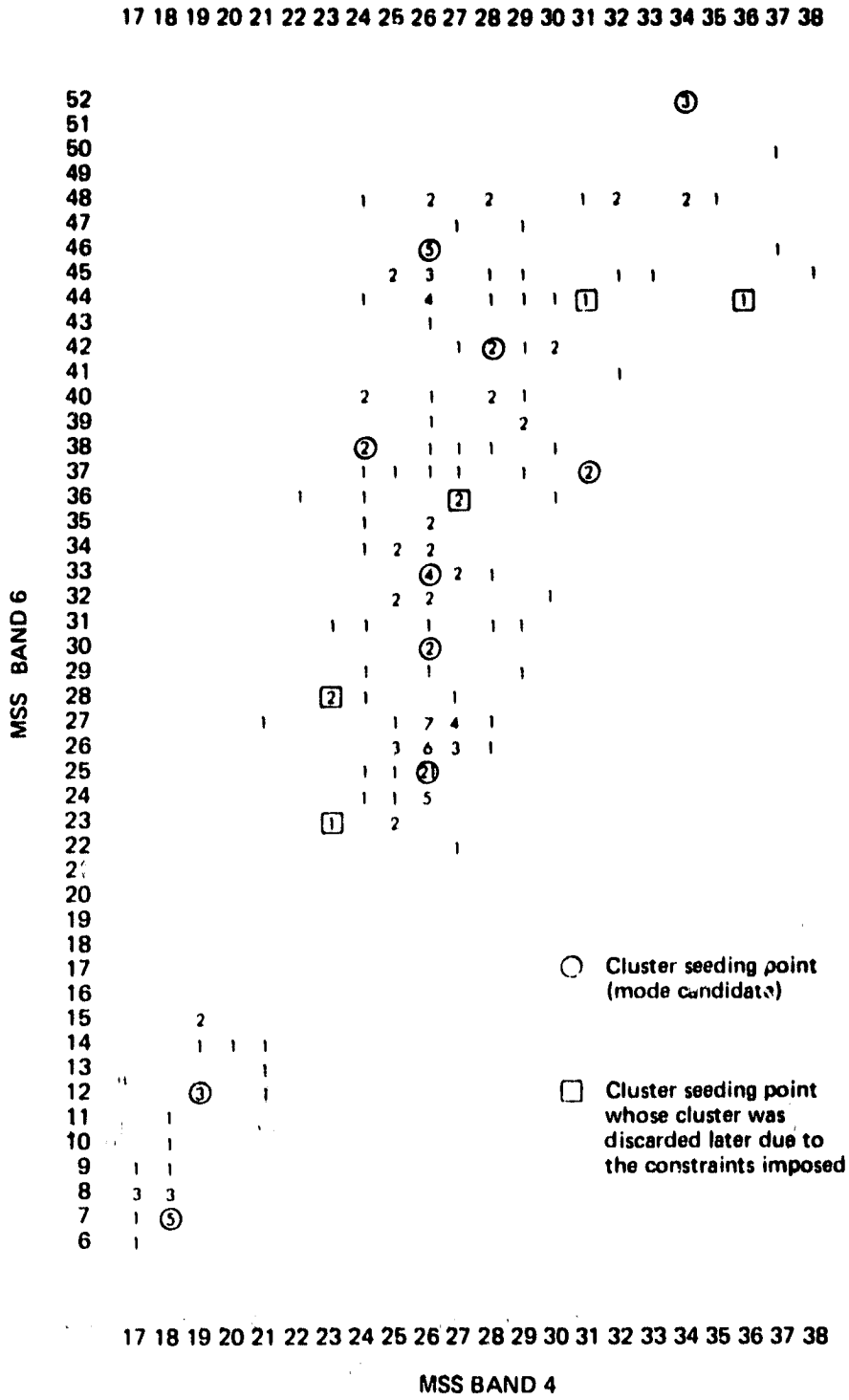


Fig. 6. Bivariate Population Distribution of Landsat Sample Data. Cluster seeding points were identified by the hill-sliding algorithm.

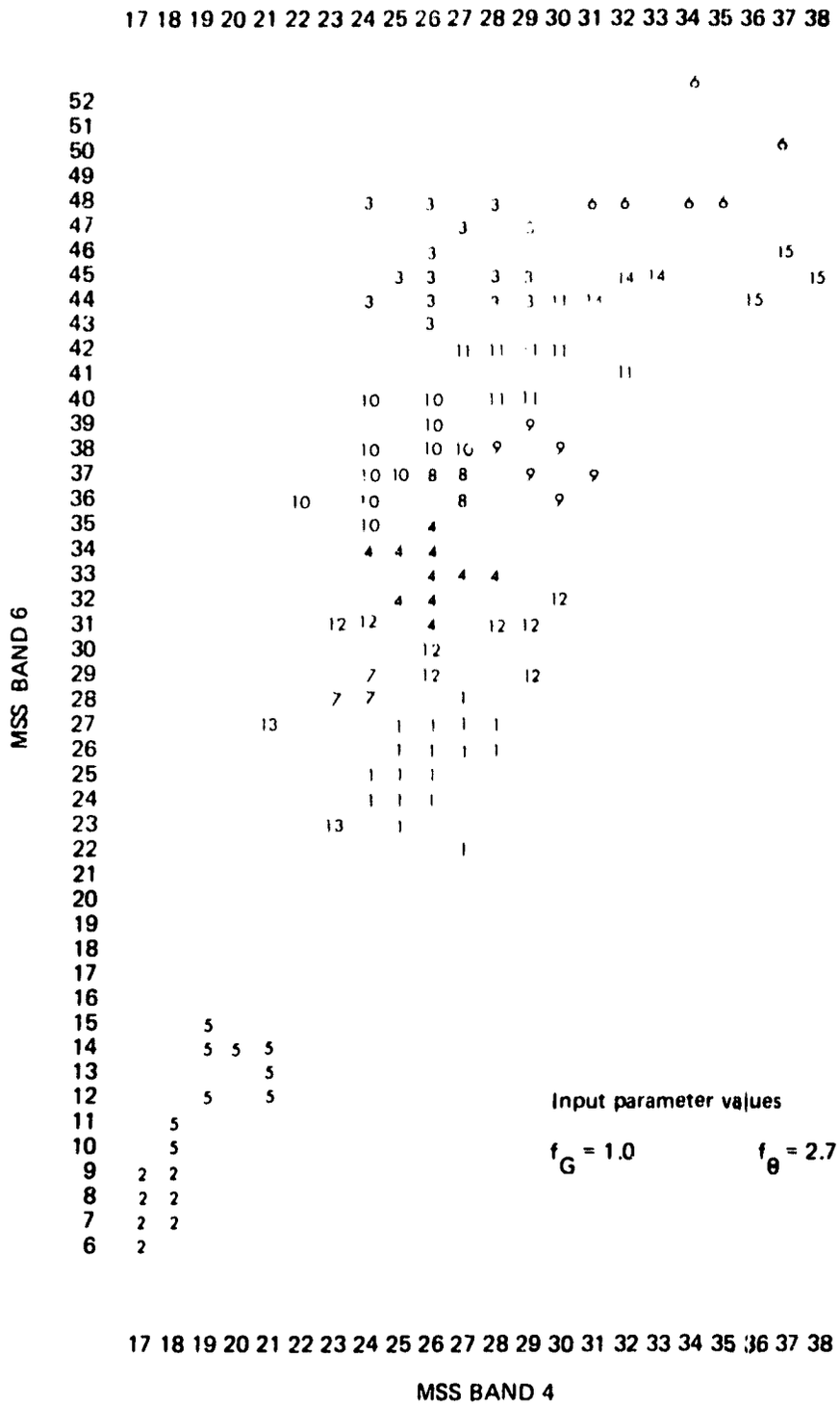


Fig. 7. Initial Clusters of the Test Data. The numbers are cluster labels.

TABLE 1. COMPARISON OF INITIAL CLUSTERS

Cluster No. (i)	N_i	M_i	r_t^2	L_i	G_{ij} three smallest values (j cluster)			Map symbol in Fig. 8	Identified Class
1	60	17	11.5	.058	20(12)	33 (7)	47 (4)	*	crop 1
2	15	7	7.0	.064	35 (5)	340 (1)	850(12)	W	water
3	27	15	15.0	.20	13(11)	25(10)	29 (6)	X	crop 2
4	19	10	6.5	.11	9.1(12)	18(10)	24 (8)	-	mixture
5	12	9	183.0 (18.0) ①	.25	35 (2)	130 (1)	210(12)	#	shallow water
6	10	6	22.5 (21.6) ①	.52	28(15)	29 (3)	34(11)	1,Z ②	healthy veg.
7	4	3	3.5	.098	29(12)	33 (1)	44 (4)	(+) ③	
8	4	3	3.0	.092	23(10)	24 (4)	33 (9)	(.,·) ①	
9	8	6	11.5	.19	14(11)	24(10)	33 (8)	-	mixture
10	13	11	14.5	.29	14(11)	18 (4)	23 (8)	.	mixture
11	11	8	33.0 (23.6)	.28	13 (3)	14(10)	14 (9)	/	mixture
12	9	8	139.5 (26.3)	.44	9.1(4)	20 (1)	29 (7)	+	mixture
13	2	2 ④	262.5 (29.1)	-	-	-	-	(*,+) ③	
14	3	3	15.0	.14	32(11)	60(15)	63 (3)	(Z,/) ③	
15	3	3	-	.42	28 (6)	43(11)	60(14)	(1,Z) ③	
		F ⑤ = 4.17 PF ⑤ = 0.146							

Note.

① Adjusted value in the bracket was used.

② Cluster was split into two parts when Figure 8 was produced because of high L_i.

③ Elements of the cluster were merged into other clusters in Figure 8.

④ Cluster was discarded and merged into other clusters due to the singularity of its covariance matrix.

⑤ Cluster no. 13 had no contribution to this value.

```
COLUMN          0000000000
AND LINE        3333333334
NUMBERS         1234567890
IN THE AREA
MAPS
     41         XX/=******
     42         XX/+******
     43         XX/+******
     44         XXX=******
     45         XXX=******
     46         XXX=******
     47         XXX.******
     48         XXX+******
     49         XX/+******
     50         X//++*****+
```

(a)

CROP FIELDS

```
                0000000000
                0001111111
                7890123456
    151         ../...=#WW
    152         --..++WW#=
    153         -=*#+#W#+-
    154         =*#WWW+/-/
    155         *#WWW#..=/
    156         #WWW#+.==.
    157         W##+===X1/
    158         -/=.==.X11
    159         1%==.-/%11
    160         11==-%%%%1
```

(b)

RIVER AND SURROUNDINGS

Fig. 8. Cluster Maps of the Study Area with Final 11 Clusters at a Scale 1: 24,000.

REFERENCES

Anderberg, M. R., 1973, Cluster Analysis for Applications, Academic Press, New York, 359 p.

Anderson, J. R., E. E. Hardy, J. T. Roach, and R. E. Witmer, 1976, A Land Use and Land Cover Classification System for Use with Remote Sensor Data, Geological Survey Professional Paper 964, U.S. Gov't Printing Office, Washington, D.C., 28 p.

Atchley, W. R., and E. H. Bryant (ed.), 1975, Multivariate Statistical Methods, vol. I, Among-Groups Covariation, Editor's Comments on Paper 8 through 16, Dowden, Hutchinson & Ross, Inc., Stroudsburg, Pennsylvania.

Ball, G. H., and D. J. Hall, 1965, ISODATA, a Novel Method of Data Analysis and Pattern Classification, A.D. 699616, Stanford Research Inst., Menlo Park, California.

Cormack, R. M., 1971, A Review of Classification, J. of the Royal Statistical Society, Series A (general), vol. 134, part 3, pp. 321-367.

Duda, R. O., and P. E. Hart, 1973, Pattern Classification and Scene Analysis, John Wiley & Sons, New York, 482 p.

Duran, B. S., and P. L. Odell, 1974, Cluster Analysis, A Survey, Lecture Notes in Economics and Mathematical Systems, vol. 100, Springer-Verlag, New York, 137 p.

Everitt, B., 1974, Cluster Analysis, John Wiley & Sons, New York, 122 p.

Friedman, H. P., and J. Rubin, 1967, On Some Invariant Criteria for Grouping Data, American Statistical Assoc. J., vol. 62, pp. 1159-1178.

Fukunaga, K., and W. L. G. Koontz, 1970, A Criterion and an Algorithm for Grouping Data, IEEE Transactions on Computers, vol. C-19, no. 10, pp. 917-923.

Park, (John) K. Y., and L. D. Miller, 1978, Korean Coastal Water Depth/Sediment and Land Cover Mapping (1:25,000) by Computer Analysis of LANDSAT Imagery. NASA Technical Memorandum 79546, NASA/Goddard Space Flight Center, Greenbelt, Maryland, 21 p.

Park, J. K., Y. H. Chen, and D. B. Simons, 1979, Cluster Analysis Based on Density Estimates and its Application to Landsat Imagery, Hydrology Papers No. 98, Colorado State University, Fort Collins, Colorado, 39 p.

Swain, P. H., 1972, Pattern Recognition: A Basis for Remote Sensing Data Analysis, LARS Information Note 111572, the Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana, 41 p.

Tou, J. T., and R. C. Gonzalez, 1974, Pattern Recognition Principles, Addison-Wesley Publishing Company, Inc., Reading, Massachusetts, 377 p.

Young, T. Y., and T. W. Calvert, 1974, Classification, Estimation and Pattern Recognition, American Elsevier, New York, 366 p.