CLUSTER ANALYSIS BASED ON DENSITY ESTIMATES AND ITS APPLICATION TO LANDSAT IMAGERY

by
John Kyoungyoon Park,
Yung Hai Chen, and Daryl Baldwin Simons

September 1979

HYDROLOGY PAPERS
COLORADO STATE UNIVERSITY
Fort Collins, Colorado

98

Several departments at Colorado State University have substantial research and graduate programs oriented to hydrology. These Hydrology Papers, therefore, are intended to communicate quickly the current results of this research to specialists interested in these activities. The papers supply most of the background, research data, and results of this work. Shorter versions are usually published in appropriate scientific and professional journals, or presented at national and international scientific and professional meetings and published in the proceedings of these meetings.

This study was funded partly by the Federation of Rocky Mountain States, the U.S. Army Corps of Engineers, St. Paul District, Contract No. DAC 37-77-C-0133 and the Colorado State University Experiment Station 107.

Subscriptions and correspondence to these papers should be addressed to: Secretary of Hydrology Papers, Colorado State University, Fort Collins, Colorado 80523.

CLUSTER ANALYSIS BASED ON DENSITY ESTIMATES AND ITS APPLICATION TO LANDSAT IMAGERY

by
John Kyoungyoon Park
Yung Hai Chen
Daryl B. Simons

HYDROLOGY PAPERS
COLORADO STATE UNIVERSITY
FORT COLLINS, COLORADO 80523

September 1979

No. 98

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF PLATES

LIST OF SYMBOLS

Symbol	Definition
B	intragroup (or between groups) scatter matrix
C_i	d-by-d covariance matrix of cluster i
C_i^{-1}	inverse of the covariance matrix C_i
C_L	characteristic length of the entire (sample) data
D_c	criterion value of the normalized divergence for refining clusters
$D_i(\underline{x})$	(squared) Mahalanobis distance from a cluster centroid
D_{ij}	divergence between two clusters i and j
D_s	criterion value of the normalized divergence for discarding clusters
d	dimension of the feature space (number of variates)
$E(\chi)$	entropy of the entire sample data χ
e	base of natural logarithm
F	objective function for clustering, weighted sum of cluster compactness values for all the clusters
f_G	empirical constant regarding G_c
f_θ	empirical constant regarding θ_c
G_c	criterion value of clustering function $G_i(\underline{x})$
$G_i(\underline{x})$	clustering function for cluster i
G_{ij}	normalized divergence based on a priori probabilities of clusters i and j
I	identity matrix
I_c	number of clusters
I_M	maximum number of clusters which can be processed by the program
k	number of the clusters involved in a boundary region
L_c	criterion value of cluster compactness for discarding clusters
L_i	cluster compactness of cluster i
L_s	criterion value of cluster compactness for splitting clusters
ℓn	natural logarithm
M_c	minimum number of probability cells in a cluster
M_i	number of probability cells in cluster i
N	number of sample data (total population)
N_c	minimum number of data elements in a cluster
N_i	number of data elements in cluster i (population of cluster i)
P_i	a priori probability (or mixing proportion) of cluster i
$p(\underline{x})$	mixture probability density (esimtate or function)at the point of pattern vector $\underline{x}$
$p_i(\underline{x})$	probability density function of cluster i at the point of pattern vector $\underline{x}$
R	feature space (general)
R^d	d-dimensional feature space
R_i	subspace (region) of R belonging to cluster i
r	radius from a cluster centroid (Euclidean distance)
r_i^2	sequentially arranged values of all the estimates of r^2 between sample data points and the cluster centroid (under formation) in increasing order
r_m^2	mean of the two r_i^2 and r_{i+1}^2
r_t^2	threshold value of r^2 for clustering the data
S_G	standard deviation of clustering function $G_i(\underline{x})$
S_θ	standard deviation of θ
T	total scatter matrix of the entire sample data
V	volume in a multi-dimensional space
W	total intragroup (or pooled-within groups) scatter matrix

Symbol	Definition
w_i	cluster i (status of nature)
$\underline{x}$	pattern (d-component column) vector
x_n	n-th component of pattern vector $\underline{x}$
$\underline{x}_n$	n-th pattern vector

Greek Letter

ε	$(A \varepsilon B)$ = A is a subspace (or member) of B
$\notin$	$(A \notin B)$ = A is not a subspace (or member) of B
θ	coefficient of the exponent in an isotropic normal density function; $-1/(2\sigma^2)$
θ_c	criterion value of θ
θ_{new}	newly estimated value of θ
$\underline{\mu}_i$	mean vectors of patterns belonging to cluster i; i.e., centroid of cluster i
ρ^d	volume of d-dimensional hypercube, i.e., volume of a probability cell
σ^2	variance of the sample population distribution in an isotropic normal distribution
χ	set of pattern vectors, i.e., $\underline{x}_n \; {}^{n=N}_{n=1}$

(Others)

$\Delta(\quad)$	small increment of the variable in the bracket
$\overline{(\quad)}$	average of the variable in the bracket
$(\hat{\ })$	estimate of the variable in the bracket
$(\quad)^{-1}$	inverse of the matrix in the bracket
$(\quad)^{T}$	transpose of the matrix in the bracket
$tr(\quad)$	trace of the matrix in the bracket
$\lvert(\quad)\rvert$	determinant of the matrix in the bracket
$\propto$	$(A \propto B)$: A is proportional to B

(Abbreviation)

CCT	Computer Compatible Tape
LMS	LANDSAT Mapping System
LMSE	Least-mean-squared error
MSS	Multispectral Scanner
p.d.f.	probability density function

ACKNOWLEDGEMENT

The paper is based on the Ph.D. thesis of John Kyoungyoon Park. Dr. Yung Hai Chen, Assistant Professor of Civil Engineering and Dr. Lee D. Miller, Research Scientist and Professor, Texas A&M University, were his advisors. The other committee members were Dr. Daryl B. Simons, Associate Dean for Research and Professor of Civil Engineering, Dr. Eugene L. Maxwell, Associate Professor of Earth Resources, and Dr. James A. Smith, Associate Professor of Forest and Wood Sciences. The acknowledgment goes also to Dr. Craig Tom, Senior Environmental Analyst, HRB Singer, Inc., Dr. Nguyen Duong, Research Associate of Civil Engineering, Dr. Kaew Nualchawee and Mr. Tienpo Chang, for their help in completing the research.

This study was funded partly by the Federation of Rocky Mountain States, the U.S. Army Corps of Engineers, St. Paul District, Contract No. DAC 37-77-C-0133 and the Colorado State University Experiment Station 107.

ABSTRACT

Most of the available clustering (unsupervised classification) techniques suffer either from a lack of adequate mathematical descriptions or from inefficiency in handling a large volume of multivariate data. The primary objective of this study was to develop a clustering algorithm with practical applicability to remotely sensed natural scene data.

The population density distribution of data often leads to an intuitive notion of the cluster, such as hill-like groups of data usually seen in one- or two-dimensional histograms. This concept of the cluster is consistent with the maximum likelihood decision rule in the statistical decision theory. A hill-sliding technique was devised to extract such natural clusters from the sample data based on this well-known notion.

Difficulties commonly encountered in computing and storing discrete multivariate probability densities were circumvented by utilizing the idea of lexicographic probability cells. Reduction of the computer memory storage requirement by this technique was significant in processing population distributions of LANDSAT multispectral scanner data.

The underlying assumption throughout the clustering process was that each cluster possess a unimodal normal distribution. A dimensionless cluster compactness parameter was proposed as a universal measure of cluster goodness and was used satisfactorily in test runs. A normalized divergence, defined by the cluster divergence divided by the entropy of the entire sample data, was utilized as a general separability measure between clusters.

The overall clustering objective function was set forth in terms of cluster covariance matrices, from which the cluster compactness measure could be deduced. Status of improvement in data partitioning could be evaluated by this objective function. The objective function was optimized with improvement of cluster compactness followed by repetitive operations of splitting or abandoning the clusters which do not meet user-supplied constraints. A desired level of end products could be reflected through these constraints.

The secondary objective of the research was to demonstrate the applicability of the clustering program to LANDSAT imagery. Cluster analysis of prototype-class imagery data of the Denver metropolitan area showed promising results. Subcategorical information on known land use/land cover classes was drawn from this analysis. The program was successfully applied to Chippewa River Basin areas in estimating the aerial extent of various land use/land cover classes with the aid of the LANDSAT Mapping System of Colorado State University. Performance of the hill-sliding clustering program was compared with that of the ISOCLAS, a version of the ISODATA family program. The hill-sliding program employed less heuristic input parameters and yielded more reasonable partitioning of the sample data of the Chippewa River Basin than did the ISOCLAS.

The hill-sliding clustering technique developed herein has applicability for use in decomposing any multivariate normal mixture distribution into a number of unimodal distributions, i.e., those of natural clusters. A subcategorical data structure can be inferred from these natural groups of data.

Chapter I
INTRODUCTION

1.1 General

There are many causes and variations to nature's countless phenomena. It is a purist's dream to isolate a single phenomenon out of all other complexities so that unique inference of the subject may be drawn. Statisticians try to pool such disarrays and to abstract common properties with the hope of describing a portion of nature. The entangled surroundings, however, always hinder meaningful abstraction and its justification. Mathematicians attempt to extend known eloquent principles which are seldom applicable to real nature. Such attempts often end up just as daydreams. Nevertheless, a better grasp of these natural phenomena has always been the goal of scientists.

Ironically, nature provides the most abundant source of information on the earth's resources in the form of electromagnetic energy (radiation) emanating from the scenes of its complex constituents. Thus, the information is embodied in a great degree of spectral, spatial and temporal variations of the radiance. This perplexing information has drawn great interest from many scientific disciplines over the last few decades, owing to recent development of remote sensing technology as well as a growing demand for ever-diminishing resources (Langrebe, 1976). The question exists as to whether the technology produces data and information to satisfy the needs of the user community. The effectiveness and the cost-savings associated with a particular data source, its timeliness, and compatibility with other information, are key factors in determining its usefulness.

Much work has been carried out to explore the potential uses of remotely sensed data since airborne sensing devices were first introduced. Pattern recognition plays a central role in extracting (or processing) useful information from this inexhaustible source. Mathematically, pattern recognition is a classification problem. Through classification procedures, each entity of the data can be identified by its intrinsic property based on class categories of the data structure, which are gained from either prototype classes or self-organized groups of data (clusters). It has been the experience that costly, time-consuming elaboration is required to obtain sufficient class-categorical information on prototype classes through selected training field data unless class patterns are very simple. Clustering of natural scene patterns has offered a promising alternative approach. 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.

1.2 Study Objective

The probability density distribution of the data often leads to an intuitive notion of the cluster as usually seen in one- or two-dimensional histograms. No clustering technique presently applicable to remote sensing data has been adequately described in terms of this well-known notion as far as the investigator knows. The objective of this study was to develop a new practical technique for unsupervised classification (clustering) of remote sensing data on the basis of probability density estimation. The technique was pursued as a suitable method to a moderate volume of multivariate measurements, such as satellite multispectral scanner (MSS) data.

1.3 Approach

The major scheme of the present approach is that after the most prospective mode is found in a mixture probability distribution, this mode enables separation of elements (or entities) of the cluster containing it from the whole set of data. Decomposition of a mixed-class distribution was sought by this scheme.

To cope with multidimensional problems, a chain of sequentially-arranged nonzero density estimates was constructed. Each multivariate probability density estimate in such a chain was equivalent to that obtained by the Parzen probability density function (p.d.f.) estimator (Fehlauer and Einstein, 1978). It was assumed in this research that these p.d.f. estimates constitute a multimodal multivariate normal distribution which is considered a mixture of two or more unimodal multivariate normal distributions. Separation of elements of each cluster from the others was attempted under the hypothesis that a unimodel distribution represents a cluster bounded by probabilistic valleys consisting of local minima of the mixture p.d.f.

A strategy for clustering entities of the density estimates was devised in the manner of "hill sliding." The term "hill" was considered the portion of a hypersurface consisting of the higher-valued p.d.f. than those of its surroundings. Clustering in this strategy was implemented by closely incorporating both the discrete density estimate and the location (measurement coordinates) of each entity. Through the implementation procedure, minor hills were disregarded after initial partitioning of all the data on the basis of cluster compactness and separability (or distinctness) from the others.

The hill-sliding technique in this paper was primarily not iterative. However, better partitioning of identities was sought repetitively after dissociation of parts or entire bodies of ill-defined clusters and merging action of those decomposed into other clusters.

1.4 Paper Organization

The salient characters of available clustering techniques were reviewed briefly in Chapter II. A clustering function in the parametric form was introduced in Chapter III, and set foth the strategy to implement clustering based on multivariate probability density estimates. Chapter IV described application of the algorithm to LANDSAT multispectral scanner data for acquiring timely land-cover/land-use information, analysis of the results obtained from Denver metropolitan data, and comparative study with results by a version of ISODATA (Ball and Hall, 1965) program. The hill-sliding algorithm developed herein for clustering, and analysis of the results obtained by the algorithm were summarized in Chapter V. The appendices included a brief overview of the LANDSAT Mapping System (LMS), which was used throughout the study for preprocessing satellite data and end products of classification. Key features of the ISODATA algorithm were also described in the Appendices. Glossary of terms was provided in the last Appendix.

Chapter II
BACKGROUND

2.1 On Satellite Remote Sensing of Hydrologic Variables

Utilization of remote sensing technology for hydrologic problems relies largely on the capability to provide updated information for calibration of the parameters employed in hydrologic models. Many parameters in physical hydrologic models, such as evapotranspiration, infiltration, and overland flow resistance, are related to time-varying phenomena of land cover on the ground surface (see Simons et al., 1975; Li, 1974). It was shown by Ragan and Jackson (1976) that LANDSAT imagery could provide better land cover information essential to evaluation of the runoff curve number in the Soil Conservation Service (SCS) model. The imagery was used for land use classification in the watershed by photointerpreters. The results based on the satellite imagery compared well with those obtained in published example problems using the conventional categories.

Better values of the parameter are the key of the model performance. Many investigators (Blanchard, 1975; Khorram, 1976) attempted to use LANDSAT multispectral scanner (MSS) data in improving simulation results of hydrologic models. Blanchard identified that linear combination of MSS band data were related to a parameter in the SCS storm runoff model. LANDSAT color composite imagery was used in estimating snow water content and evapotranspiration water loss over the watershed with the aid of low altitude aerial photography and topographic data by Khorram (1976). He introduced a concept of multi-stage sampling to utilize remote sensing information. He pointed out that real-time information could be generated for the entire watershed.

Many approaches reported in this line are categorically similar to those mentioned above. Land-use/land-cover classes are mapped by a specific classification algorithm(s) using remotely sensed data (Miller et al., 1977; Park et al., 1978). Then better values of parameters employed in hydrologic models are estimated by correlating mapped information with the parameters. A critical point in this procedure is how reasonably remote sensing-derived information reflects the real world at the time for the intended purposes. Subsequent questions are also raised concerning the best approach to retrieve such information.

Taking into account time-varying phenomena and spatial variations of remotely sensed data were other difficult tasks involved in this line of research. Phenology of vegetation is the most important factor in time-varying phenomena of vegetated lands. Mixture and ecological tones of vegetation as well as slope and aspect of terrains, make it complicated to recognize patterns of naturally vegetated lands in mountainous regions (Maxwell et al., 1977). Oversimplication of phenological factors or ignorance of slope and aspect influence were commonly introduced in most of the investigations reported. Hence, many conclusions based on such assumptions were often far removed from general applicability to real situations.

Use of digital remote sensing data may be divided into two categories: classification and regression analysis. Regression analysis is carried out by comparing known variables with particular signature values (original measurements, or values transformed into special coordinates from the measured). Information quantities of the variable can be deduced from the data using regression relationships within a certain confidence limit. For instance, surface water turbidity or biomass of particular vegetation types may be evaluated by this method. The classification approach is widely used in mapping land-use/land-cover, and in estimating their acreages. The basic idea of classification schemes in remote sensing is to divide the feature space into non-overlapping regions, each of which is to be designated for one of the class categories. Different classification schemes differ mainly in the criteria employed to establish these subspaces.

There are two basically different approaches in classification: supervised and unsupervised. Supervised classification is based on known information about prototype classes. A classifier as a set of discriminant functions is devised to recognize a predetermined pattern by means of various adaptive schemes through available training samples (ground truth data). Literature surveys and historical remarks on this approach can be found in many recent publications (Cormack, 1971; Duda et al., 1973; Das Gupta, 1973; and Tou et al., 1974). In the past, the main approaches for classification of remote sensing data were based on a variety of supervised techniques. Selection of a classification algorithm relies on the characteristics of the data structure and the quality to be processed.

It has often been experienced in the space remote sensing that the supervised approach requires the analyst to select training samples to represent all possible variations in spectral response for each prototype class (Fleming et al., 1975). Proper selection of such a training data set proved very difficult in many cases due to complex vegetation types and rugged terrain over the target areas (Maxwell et al., 1977; Fleming et al., 1975). Considerable human judgment and intervention with time-consuming iterations are unavoidable until satisfactory results are produced.

Many unsupervised classification algorithms employ self-clustering techniques to group the multispectral (generally, multivariate) data into a number of classes. A subdivision of the feature space is achieved by identifying natural groupings (or clusters) of the data. The nature of the classes thus found is determined afterwards on the basis of known information. In this respect, the procedures of the unsupervised classification (clustering) are in the reverse order of the supervised. A definite advantage of unsupervised approaches is their ability to alleviate the problem of categories with multimodal distribution (Nagy, 1972). A large amount of the literature has been accumulated in this area over the past decade (Duran et al., 1974; Duda et al., 1972; Anderberg, 1973; Everitt, 1974). In remote sensing, unsupervised techniques are usually applied to those areas where ground truth data are not readily available or where the information provided is not sufficient for supervised classification. This approach often helps a user learn existence of unexpected group(s) of classes or variations of known group properties such as temporal changes of earth resources.

Most of the algorithms employed in both supervised and unsupervised approaches categorically fall in the statistical analysis. Many investigators

are in favor of statistical treatment of data in pursuing inherent characteristics of classes or clusters from noisy original data. Minor spatial and temporal variations of signatures may well be processed statistically. Apparent noise contained in the data had been filtered in preprocessing stages (Maxwell et al., 1977). However, correction of path radiance distorted through the medium must be made on the basis of classical radiative transfer theory, which is a deterministic approach (see Rogers et al., 1973). Both statistical and deterministic approaches may have to be employed for analyzing a set of remotely sensed data. This is another difficult aspect to cope with in the space platform data.

2.2 Clustering Techniques

Many diverse techniques have been devised to discover structure within complex bodies of data in 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. In cluster analysis, all that is available is a collection of data whose category memberships are unknown. The operational objective is to discover category structure which fits the data. A general strategy for this objective is implemented by defining a clustering criterion and constructing an algorithm which consists of a set of operations. Such operations can be consistently applied to the clustering problem. An efficient algorithm may assemble data into clusters which prior misconception or ignorance would otherwise preclude. Hence, application of the techniques to the data often reveals unexpected characteristics inhibited in the data structure.

It has been known, however, that clustering techniques are tools for discovery rather than ends in themselves (Anderberg, 1973; Dubes et al., 1976). No universal clustering criterion has been found. Different clustering techniques produce different results. Slightly varying tactics even under the same criterion are often found in great variations of the results. Difficulties in attempting to fit the intuitive nature of clustering techniques into any meaningful mathematical framework have been described by many investigators (Anderberg, 1973; Cormack, 1971; and Dubes et al., 1976).

Most of the early works in cluster analysis were in the fields of biology and zoology, where numerical taxonomy is a frequent substitute. Initially taxonomy was an art rather than a scientific method. Later numerical techniques have gradually become widespread when digital computers have served as common tools. A variety of techniques were developed and applied in many fields. For the last three decades cluster analysis has been a multidisciplinary technique of data analysis (Anderberg, 1973). A comprehensive overview was given by Ball (1965). Other recent reviews by Cormack (1971), Duran et al. (1974), and Everitt (1974) reported brief descriptions of methodologies and extensive references. Clustering techniques have been broken down into various categories; but they themselves may be classified into two approaches: (1) hierarchical and (2) nonhierarchical.

2.2.1 Hierarchical clustering techniques

In the hierarchical technique, the similarity measures are often used to construct a similarity matrix representing all pairwise associations among the entities (samples). The techniques operate on this matrix to construct a tree characterizing relationships among the entities. One starts with N entities and groups the two most similar (nearest) ones into a cluster, thus reducing the number of clusters to N-1. By repeating this procedure, all the entities form one cluster in the final step. For every hierarchical clustering there is a corresponding tree, called a dendogram (or tree diagram). It shows the diagram of grouping the entities throughout all the steps. The method has had the widest use in ecological studies (Everitt, 1974). A classic example is the grouping of biological samples into species, species into genera, genera into families, and so on. An excellent source of reference in this area is the book by Sokal and Sneath (1963).

A major drawback of hierarchical clustering is the massive storage requirements for the similarity matrix, which consists of N(N-1)/2 elements, if the number of samples "N" is large. In most remote sensing applications, the amount of data is quite large. For example, it requires data of about 30,000 resolution elements to cover the area of one USGS 7 1/2 minute quadrangle topographic map, in cases of processing LANDSAT digital data. Even though one-tenth random samples of the total data are analyzed to produce statistics of clusters, it is necessary to compute 4,500,000 elements of the similarity matrix for the hierarchical clustering. Application of these techniques to remote sensing is nearly prohibitive due to the large volumes of data. Extensive discussions on other disadvantages as well as general aspects of this approach can be found in the literature (Cormack, 1971; Everitt, 1974; Duran et al., 1974).

2.2.2 Nonhierarchical clustering techniques

Nonhierarchical clustering methods may include all of the clustering techniques not necessary for the calculation and storage of the similarity matrix. Hence, the methods are generally suitable to much larger problems than the hierarchical methods. A variety of techniques have been reported in this category. A common scheme in most of nonhierarchical clustering techniques is to assign some initial partition of the data units and then, if necessary, improve cluster memberships under given instructions. The various algorithms which have been proposed differ as to criterion for defining the best partition or the way for achieving a better partition. Based on key algorithms employed in nonhierarchical clustering, the following subclasses may be categorized:

1) Optimum partitioning techniques, in which the entities are grouped into mutually exclusive clusters which optimize a clustering criterion.

2) Density-search techniques, in which the entities are grouped into several subgroups by searching for regions having a relatively high probability density.

3) Clumping techniques, in which the classes (clumps) can overlap, and a class and its complement are treated as a different type of class.

4) Other techniques which do not fall clearly into one of any previous type, or which may be a mixed type of two or more techniques.

2.2.2.1 Optimum partitioning techniques. Clustering by partitioning techniques is often carried

out by four distinct procedures as follows: (a) initiation of clusters; (b) allocation of entities to initial clusters; (c) evaluation of objective function (or criterion); and (d) reallocation of entities to other clusters to achieve the optimal value of the objective function. In most of these approaches, the last two steps are repeated until the results are satisfactory. Various methods employ different objective functions or different strategies in each step.

Most of the objective functions for optimum partitioning are derived from the well-known matrix identity:

$$T = W + B \tag{2.1}$$

where T is the total scatter matrix of the samples, W is the total intragroup (or pooled-within groups) scatter matrix, and B is the intergroup (or between groups) scatter matrix (Friedman et al., 1967). The intragroup scatter is a measure of dispersion of the members in the group. For any given data set the matrix T is constant, and so the function of either B or W is sought as clustering criterion. It can be illustrated in the simple case of one variable (dimension) that Eq. 2.1 is a scalar equation and a good criterion is to minimize the total intragroup scatter quantity W. This is equivalent to maximizing the intergroup scatter B. For more than one variable, the matrix equation should be transformed into scalar relationships, from which a clustering criterion can be deduced. The following are commonly cited relationships:

$$tr(T) = tr(W) + tr(B) \tag{2.2}$$

and

$$\frac{|T|}{|W|} = |I + W^{-1} B| \tag{2.3}$$

where

$tr(\ )$ = trace of the matrix in the bracket

$|\ |$ = determinant of the matrix

$(\ )^{-1}$ = inverse of the matrix in the bracket

I = identity matrix.

Based on the above two equations, the following criteria are derived:

1) minimization of $tr(W)$ or maximization of $tr(B)$

2) minimization of $|W|$ or maximization of $|T|/|W|$

3) maximization of $tr(W^{-1} B)$

The first criterion is interpreted as the least mean-squared-error (LMSE), sum-of-squared-error, or often within-groups-sum-of-squares (WGSS) criterion, which has been exploited by many investigators (Duda et al., 1973). The value of $tr(W)$ is invariant under an orthogonal transform of the feature space, and the algorithms based on this are found most appropriate for fairly concentrated clusters (Nagy, 1968). However, the LMSE partition might change if the variables are scaled, since it is not invariant under such a transform. It is important to note that this criterion function does not take into account effects of correlations between variables, which are commonly observed, for example, in multispectral scanner (MSS) data. Other drawbacks experienced are that the optimal value of the criterion function depends on the number of clusters and that the solutions are frequently suboptimal (Duda et al., 1973; Everitt, 1974).

The second criterion is similar to the first one that minimizes $tr(W)$ but the two need not be the same. The determinant of a scatter matrix measures the square of the scattering volume, since it is proportional to the product of the variances in the direction of the principal axes. Hence, this criterion has a good physical interpretation of cluster compactness. It has been shown by Koontz and Fukunaga (1972) that this as well as the third criterion has better performance requirements but has computational complexity. This criterion cannot be employed if the expected number of clusters is not larger than the dimensionality since the determinant of the matrix W will be singular. The third criterion shows similar performance as the first one (Duda et al., 1973).

Another key feature in optimum-partitioning techniques is the intial partitioning of the data units into groups (i.e., initiating clusters). It is important since the solutions to the clustering problems depend upon initial configurations of seeding clusters in many cases (Everitt, 1974). A majority of techniques begin with some mutually exclusive points as cluster centers. Such an arbitrary set of initial cluster centers often affects on convergence of the iterative solutions as well as computational time. Some other techniques attempt to search reasonable configuration of initial clusters in the beginning.

After a set of initial clusters is found, search for the optimal solution continues by reallocating entities. Many different algorithms have been devised to implement this step. Major differences between algorithms stem from differences in criterion functions or instructions for achieving improvements. Instructions depend on criterion functions. Some algorithms employ nearest centroid sorting techniques and others employ mathematical programming techniques. Typical examples in the former case are K-means algorithm by MacQueen (1967) and ISODATA (Iterative Self-Organizing Data Analysis Technique) by Ball and Hall (1965). These two algorithms are iterative procedures. K-means algorithm starts with the first arbitrary chosen K samples as initial K cluster centroids and then revises cluster centroids as receiving new data, such that the sum of the squared distances from all points to the new cluster centers is minimized. Several extended versions of the ISODATA program exist at present (Anderberg, 1973). They are all similar in principle to the K-means procedure, but work in cooperation with numbers of empirical parameters such as lumping, splitting and chaining parameters. It has been understood that the application of ISODATA algorithms to a set of moderately complex data often requires extensive experimentation before meaningful conclusions are drawn. Although considerable insight into the structure of the data can be gained through the information obtained in each iteration, no convergence to the optimal solution is guaranteed in these types of approaches (Anderberg, 1973; Tou et al. 1974).

Mathematical programming for systematic search of improved partitioning to the optimal solution has been exploited by Jensen, Vinod and other investigators (Duran et al., 1974). The idea of this method is to evaluate the objective function for each choice of clustering alternatives and then to choose the partition yielding an optimal solution. In this line, some applications of dynamic programming and integer programming to cluster analysis were reviewed by Duran,

et al. A shortcoming in mathematical programming techniques is that tedious computation and excessive storage of all possible optimal solutions at numerous transitional stages are unavoidable when the number of data and clusters are large. It seems inadequate to use these techniques for a large volume of data.

2.2.2.2 Density search techniques. The clustering techniques based on probability density estimation may lead to a well-defined notion of cluster. There is a natural tendency that, when the data are distributed in a feature space, there should be parts of the space in which data populations are very dense, separated by parts of low density. This concept could form the basis of the definition of a natural cluster. Many probabilistic cluster-seeking techniques search regions of high density or mode based on this presumption. Modes are local maxima of the probability density function. Major efforts in this technique are made to search for a local optimum of the criterion, by rearranging existing partitions, keeping the new arrangement only if it gives an improved criterion value. Procedures devised to implement these techniques are often called mode-seeking algorithms. The mode analysis by Wishart (1969) and "hill-climbing" technique by Bryan (1971) are typical examples of this approach. In mode-seeking algorithms, the number of resultant clusters depends on given parameters. Sometimes only one cluster may be formed but usually the analysis reaches a point at which a maximum number of clusters are isolated. These techniques are often considered significant for a moderate volume of data, even though they suffer from the problem of containing various empirical input parameters (Everitt, 1974). In addition, there appear to be several solutions for the small set of data (say, less than 100) which form a multivariate mixture distribution in which clusters are not widely separated. In multivariate problems the storage requirement becomes serious, even for a medium size of data. It also requires a considerable amount of computation to update all of the multivariate density parameters unless assumptions on the underlying distribution are simplified (Ball, 1965).

Another approach in the probabilistic cluster-seeking techniques is decomposition of mixture distribution. The basic idea behind decomposition techniques is that separable clusters have distinctively different distribution characteristics. These techniques attempt to find the estimates of the parameters of the density function for each separable cluster as well as its mixing proportion (a priori probability). A classical work on this subject is that by Stanat (1968), in which decomposition of multivariate normal and multivariate Bernoulli mixtures was investigated. It was shown that the parameters of multivariate density function for any distinct class can be estimated in principle by means of characteristic function (Fourier transform) of the mixture distribution. Computational burden, however, is quite severe, especially for estimation of characteristic function of multivariate normal distribution and its storage. Hence, application of this approach to remote sensing seems almost impractical if variables (dimensions) are more than two. Use of maximum likelihood and Bayesian approaches for decomposition problems had been investigated by Day (1969), Hasselblad (1966), Wolfe (1965-1970), and many others. In many cases, solutions to these problems are obtained by iterative approximation with slow convergence. Hence, the techniques would eventually be time-consuming unless the number of parameters to be computed were small. Methods also suffer from the problem of suboptimal solutions, since there may be more than one solution to the maximum likelihood equations (Everitt, 1974).

2.2.2.3 Clumping techniques. Most clustering techniques yield disjoint groups. In clumping techniques, regions of clusters must have overlaps if they are to be of any value. As seen in language, most words have more than one meaning; and when complemented with other words, they take on another meaning. Clumping partitioning is carried out on the basis of the similarity matrix or correlation coefficient matrix (Ball, 1965). The techniques require a large amount of computation for estimation of all pairwise distances and storage for them. Thus, the applicability is limited with a small sample size.

2.2.2.4 Miscellaneous techniques. The approaches discussed so far constitute perhaps the major framework of cluster analysis. There remain, however, numerous other clustering techniques which do not fall clearly into any of the previous categories. Strategies in some algorithms involve using several clustering methods together (or iteratively and sequentially) in order to gain a more extensive appreciation of the structure in the data set. Others might be devised for particular problems with particular tactics which appear somewhat different from those categorized. Some of these techniques which have been applied to remote sensing data will be described.

Spatial clustering techniques generate clusters based on the distributions of image data in both spatial domain and feature space. Efforts in this area have been made by Nagy et al. (1972), Haralick and his colleague (1969, 1975), Kettig (1975). The central idea is to analyze gradient images or data distribution of neighborhood blocks and then assign spatially homogeneous patterns to clusters. Boundary detection is often incorporated with spatial clustering procedures. These algorithms aim mainly at remote sensing application and show promising results. However, their present versions tend to yield excessive numbers of clusters when applied to the relatively heterogeneous areas, such as naturally vegetated lands. Their applicability appears very limited.

Another interesting technique, which has some resemblance to the approach of this study, is a method of mode separation proposed by Kittler (1976). A chain of hypercubic (or hyperspheric) cells is constructed by sequentially arranging neighbor cells (points having non-zero probability density estimates) according to their density values. In this way, the multivariate density cells are aligned in the sequential (i.e., one dimension) probability densities. Modes are separated by the local minima of probability densities in the chain since the majority of cells from each mode tend to be successive elements of the sequence. However, finding neighbor cells requires computation of all the pairwise distances among the cell points, which is not desirable for a large amount of data. Hence, the chaining operation seems to require laborious computation in cases of remote sensing application. Considering that distinct clusters are separable through a set of valleys (not a point) consisting of local density minima in the multivariate mixture distribution, the concept of separation of clusters is not clear in the approach. The use of a discriminant classification technique is actually suggested to find a separating surface between the two modes. This inconsistency is another drawback of Kettler's method. Thus, this method seems

to have its significance only in the capability to initiate clusters regardless of distribution types.

Cluster analysis (or pattern recognition) of multispectral images has frequently been faced with the curse of dimensionality. Traditionally, investigators have used either principal components analysis or factor analysis to alleviate the difficulty of too many dimensions. Many clustering algorithms, mentioned earlier, have worked with the aid of such a dimensional reduction procedure. Recently, new interest has centered on the possibility of incorporating analyses of both entities and variables simultaneously for better understanding of the underlying structure. Procedures for simultaneous clustering of entities and variables were described by Good (1965) and Hartigan (1972). However, principal components and multidimensional scaling do not often give adequate representations of some sets in lower dimensions, and so may lead to graphical misinterpretation (Duda et al., 1973; Everitt, 1974). Multidimensional scaling may yield unwanted suboptimal solutions to clustering problems.

2.3 Summary

The space platform imagery and associated interpretations at various levels of sophistication have been widely accepted as a means of generating the spatial input data for hydrologic models (Miller et al., 1977). Remote sensing products provide updated information and the data bases for planning large- and small-scale hydrologic developments. A critical question commonly raised in the remote sensing community, however, is how consistently the remote-sensing-derived information reflects the real world. The difficult task is to take into account time-varying phenomena and heterogeneous spatial distribution of target objects. Phenology, ecological tones and mixtures of vegetation as well as slope and aspect of the terrain make recognition of patterns of land covers complicated. Statistically meaningful abstraction of such patterns requires a collection of sufficient amount of sample data and consequently, laborious computation.

A traditional approach for recognition of land cover/land-use is that of classification: supervised and unsupervised. Supervised classification is based on information about prototype classes. Proper selection of training data proves very difficult when the target areas consist of complex land-cover types and rugged terrain. Unsupervised classification (clustering) is suitable to classification of most of the remotely sensed data, at least for the first examination of the data structure. An efficient clustering algorithm may assemble data into clusters which prior misconception or ignorance would otherwise preclude. Unsupervised classification has been divided into two broad categories: hierarchical and nonhierarchical. Hierarchical clustering techniques work on the similarity matrix which represents all pairwise associations among the sample data requiring excessive computer storage if the number of data is large. These techniques generally are not adequate for the classification of a large volume of remote sensing data.

Among various techniques for nonhierarchical clustering, optimum partitioning techniques are the most frequently cited in the literature. The algorithms along these lines have been described in better mathematical terms (e.g., objective function) than others have. However, the solutions have frequently been found to be suboptimal and dependent upon the number of clusters specified in advance. No convergence to the optimal solution is guaranteed in most of these types of approaches. Furthermore, the number of iterative computations needed to find the solutions at a satisfactory level (or rate of convergence) depends heavily upon configuration of initial clusters. The initial configuration is often critical in leading the solution to the global one. It is also pointed out that mathematical programming techniques require excessive enumeration to check all possible sets of optimal solutions at every transitional step unless the size of data is small.

The probabilistic cluster-seeking techniques are based on the presumption that there should be parts of the feature space in which data populations are very dense, separated by parts of low density. Mode-seeking algorithms and decomposition of mixture distribution are typical examples along this line. The mode-seeking techniques are often considered significant for a moderate volume of data, in spite of the problem of containing numerous empirical input parameters and several optimal solutions. Decomposition of multivariate normal mixtures in Fourier domain has been theorized, but its computational complexity seems to remain unresolved.

Another promising area for remote sensing application is the spatial clustering approach. It operates by incorporating the distributions of image data in both spatial domain and feature space. However, the present versions of this approach tend to yield an excessive number of clusters when applied to the nonhomogeneously-vegetated lands. Their applicability is very limited.

This brief review on the various available approaches to the utilization of remote sensing technology for hydrologic problems reveals the wide disparities in both the basic approaches and their useful results. In light of the significant discrepancies among the numerous attempts, extensive efforts to extract consistent and useful information from remotely sensed data should be made. The growing applications of clustering methodology to remote sensing studies warrant further investigation to devise a practical clustering algorithm. The usefulness of such an algorithm must be subjected to adequate justification in the real field data.

Chapter III
CLUSTERING OF MULTISPECTRAIL SCANNER DATA

3.1 General Description

Multispectral scanners (MSS) loaded in satellite or aircraft platforms view the earth's surface and record levels of radiance emanating from a resolution element of the target area. Heterogeneous characteristics of natural scenes over the area often result in large variations in the observed data. Such variations are due mainly to mixtures of countless scene components which fall within a resolution element and which have different spectral responses. Hence, MSS data obtained from a satellite or aircraft platform can reasonably be assumed to have multimodal normal distribution. Each mode in the mixture distribution is considered as that of a cluster. Modes in such a distribution may reflect status of several class mixtures or phenological variations of vegetation covered over the area. Thorough analyses of phenology related with remote sensing data reveal that no standardized framework can be established in the measured or extracted signature space and no universal characteristics can be described by any supervised approach. It is often necessary to learn general characteristics of the data structure by naturally grouping (or clustering) before application to supervised classification for analysis.

The clustering techniques based on probability density estimation lead to a well-defined notion of cluster. It is often intuitively observed that, when the data are distributed in a feature space (which consists of the measured and/or transformed data coordinates), there are parts of the space where data populations are very dense, separated by parts of low density. This concept forms the common definition of a natural cluster. In this chapter a new clustering algorithm is formulated on the basis of the probability density distribution. To start with, a clustering function is proposed in the parametric form; a discussion of its characteristics near boundaries between clusters in the mixture distribution is included. The hill-sliding strategy is devised to implement clustering. An iterative procedure to improve the sets of clusters initially obtained is described.

3.2 Parameterization for Clustering Function

Clustering is often the first step in analyzing a set of data whose characteristics have not yet been examined. 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 one or more unimodal clusters and have multimodel distribution. Such data are divided into two or more subgroups so that the unimodal distribution can be applied to each subgroup.

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

$$p_i(\underline{x}) = \frac{P_i}{(2\pi)^{d/2} |C_i|^{1/2}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu}_i)^T C_i^{-1}(\underline{x}-\underline{\mu}_i)}, \quad (3.1)$$

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

$( )^T$ = transpose of a matrix

d = dimension of feature space (integer)

e = base of natural logarithm.

This is a multivariate normal distribution function. The density function is expressed in a compact form:

$$p_i(\underline{x}) = p_i(\underline{\mu})\, e^{-\frac{D_i(\underline{x})}{2}} \quad (3.1')$$

In this expression,

$$p_i(\underline{\mu}) \equiv \frac{P_i}{(2\pi)^{d/2} |C_i|^{1/2}} \quad (3.2)$$

is the proability density at the centroid of cluster i and

$$D_i(\underline{x}) \equiv (\underline{x} - \underline{\mu}_i)^T C_i^{-1} (\underline{x} - \underline{\mu}_i) \quad (3.3)$$

is often called Mahalanobis distance (or squared Mahalanobis distance).

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 $\chi = \{\underline{x}_n\}_{n=1}^{n=N}$, the multivariate mixture probability $p(\underline{x})$ may be estimated and then postulated as the sum of all the cluster probabilities $p_i(\underline{x})$:

$$p(\underline{x}) = \sum_{\text{all } i} p_i(\underline{x}) \quad (3.4)$$

It may be computed by

$$\hat{p}(\underline{x}) \cong \frac{\text{sum of population in a volume element } \Delta V}{\text{total population } (N)} \quad (3.5)$$

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. The present study concentrated on decomposition of normal (Gaussian) mixture density. For this purpose, a clustering function is proposed as

$$G_i(\underline{x}) = \ell n \frac{p(\underline{x})}{p_i(\underline{x})} \quad (3.6)$$

where ℓn denotes natural logarithms.

The clustering function is always theoretically non-negative:

$$G_i(\underline{x}) \geq 0 \tag{3.7}$$

since

$$p(\underline{x}) = \sum_{\text{all } j} p_j(\underline{x}) \geq p_i(\underline{x}) \tag{3.8}$$

It is also evident that

$$G_i(\underline{x}) = 0 \tag{3.9}$$

for the data of a single unmixed cluster. Probability densities of discrete measurements, however, do not often exist in some regions where $G_i(\underline{x})$ yields negative infinite value. The regions of nonexisting probability density are out of consideration in this study.

An unfortunate point on this formulation is that the clustering function $G_i(\underline{x})$ cannot be evaluated until the a posteriori probability density function $p_i(\underline{x})$ is estimated. This difficulty may be overcome by initializing a reasonable cluster center and by grouping data near the center. The clustering function shows a useful characteristic for discriminant analysis when the parametric representation of a cluster is found. Suppose that estimates of a set of Gaussian cluster parameters, p_i, $\underline{\mu}_i$, C_i are made and then $G_i(\underline{x})$ is evaluated. It is possible to extract data belonging to the cluster from the data pool according to the following fact. The clustering function $G_i(\underline{x})$ is rewritten as

$$G_i(\underline{x}) = \ell n \frac{\sum_{\text{all } j} p_j(\underline{x})}{p_i(\underline{x})} = \ell n[1 + \sum_{\text{all } j \neq i} \frac{p_j(\underline{x})}{p_i(\underline{x})}] \tag{3.10}$$

Near a cluster center in the mixture distribution, Eq. 3.10 yields an immediate approximation given by Eq. 3.9 since

$$p_i(\underline{x}) >> \sum_{\text{all } j \neq i} p_j(\underline{x}) \tag{3.11}$$

Consider a problem of decomposing the data having bimodal distribution into two clusters. It is intuitive to separate the region into the two by the boundary where

$$p_1(\underline{x}) = p_2(\underline{x}) \tag{3.12}$$

In such a boundary, Eq. 3.10 leads to

$$G_1(\underline{x}) = G_2(\underline{x}) = \ell n\ 2 \tag{3.13}$$

And in each region, say R_i $(i=1,2)$

$$G_i(\underline{x} \in R_i) = \ell n[1 + \frac{p_j(\underline{x} \in R_i)}{p_i(\underline{x} \in R_i)}] < \ell n\ 2 \tag{3.14}$$

since

$$p_i(\underline{x} \in R_i) > p_j(\underline{x} \in R_i) \tag{3.15}$$

where $(\underline{x} \in R_i)$ denotes that $\underline{x}$ is a vector point in a subspace R_i. This suggests that a feature space R can be divided into two regions: a region belonging to cluster i and the other out of the region, in the case of two cluster mixtures as such $\underline{x} \in w_i$ (cluster i) if

$$G_i(\underline{x}) < \ell n\ 2 \tag{3.16}$$

and otherwise, $x \notin w_i$. Here w_i denotes the state of nature (or class) and $\underline{x} \in w_i$ means that $\underline{x}$ belongs to class w_i while $\notin$ is the opposite of symbol "$\in$". This criterion was utilized in this study for the extraction of one-cluster data from the whole set of data. The criterion 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 $G_i(\underline{x})$.

Clustering in multicluster mixture regions follows a similar procedure. An extended clustering rule in a known k-cluster mixture region would be that $\underline{x} \in w_i$ if

$$G_i(\underline{x}) < \ell n\ k \tag{3.17}$$

and otherwise, $\underline{x}$ be rejected from being merged into the group (i.e., $\underline{x} \notin w_i$). It is a direct generalization of the criterion given in Eq. 3.14 considering the boundary surface where

$$\left.\begin{aligned} &p_i(\underline{x}) = p_{i+1}(\underline{x}) = \ldots = p_{i+k-1}(\underline{x}) \\ &p_j(\underline{x}) = 0 \quad \text{for all the others.} \end{aligned}\right\} \tag{3.18}$$

Such a boundary surface (point in the univariate space and line in the bivariate space) is not observed in a univariate feature space. The boundary surface of k (more than 2) clusters may exist only in two or higher dimensional space. Multicluster boundary surfaces (or lines or points) are hardly detected even in computerized clustering procedures since discretization of the space with a finite interval results in shifting of its probable exact location. Hence, application of this generalization to recognition of multicluster boundary is questionable. A practical rule for clustering in such a mixture region may be set up as

$$\ell n(k-1) < G_i(\underline{x}) < \ell n\ k \tag{3.19}$$

in which k is a positive integer larger than one. The region unclassified by the criterion inequality Eq. 3.16 may be merged in the cluster region R_i when the inequality Eq. 3.19 is satisfactory.

3.3 Hill-Sliding Strategy

The clustering function proposed in the previous section 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, but it is usually observed near the gravitational center (centroid) of the cluster. Its actual location may deviate from its theoretical position due to the randomness of measured data.

This study extracts cluster data mainly from the LANDSAT Computer Compatible Tapes (CCT) containing multispectral scanner (MSS) data. Many investigators have shown that the LANDSAT MSS data can be reasonably treated under the assumption of normality for each cluster/class (Maxwell, 1976; Maxwell et al., 1977). Most of the centroids within probable clusters are located at the approximate center of each group (Fig. 3.1). Suppose this presumption is acceptable in a set of data to be analyzed. Then at least one probable mode which has the highest probability density can be picked up in any distribution space. Such a mode initiates the first clustering by fusing all probability cell points that may be categorized into one cluster while a scanning pointer moves down from the present highest density position to the next highest. 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. A perspective view of a bivariate normal probability density function (p.d.f.) is shown in Fig. 3.2. The estimates of p.d.f. deduced from discrete measurements do not form smooth topographic surfaces as shown in this figure. A schematic (ideal) progress of the hill-sliding path in the algorithm is visualized in the figure.

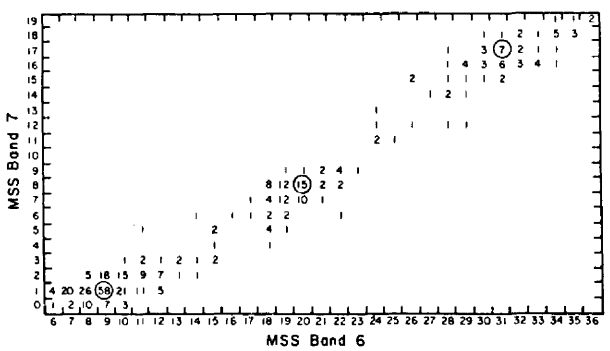


Fig. 3.1 Bivariate population distribution of 400 LANDSAT data and the first three most probable candidates for modes of clusters (in circles). These typical example data were taken from the LANDSAT data over a portion of Korean west coast (Park and Miller, 1978). The numbers are occurrences of bivariate 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.

Actual paths from the present position to the next lower density point will be zigzag motion due to randomness of the estimated p.d.f. in the discrete space. The pathway is always descending or leveling, ending at the bottom of a valley. Such an end point seems obvious from the schematic diagram even though it depends on the topographic shape. The topograph in the present clustering approach is that of multivariate probability density. Its shape can be characterized by the probabilistic laws governing the data at hand.

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

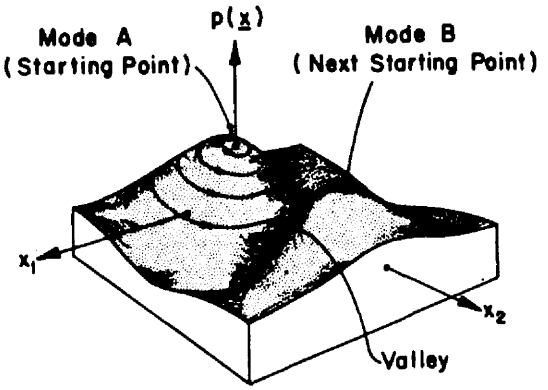


Fig. 3.2. Topographic surface of a bimodal bivariate normal distribution function and schematic presentation of "hill-sliding" pathway. The hill-sliding algorithm starts with the highest probability density point (MODE) by fusing each highest density point until a valley is reached.

characteristic shape to its distribution. Assumption of normal (Gaussian) distribution was employed in this study.

The LANDSAT data reveal high correlations between neighbor channels, with characteristics of anisotropic distribution (Fig. 3.1). This study attempted to group the anisotrophic normally distributed data in unsupervised manner. Due to the absence of prior knowledge on characteristics of expected clusters, each cluster was initiated with the assumption of isotropic normal distribution until its fusing process stops at a certain level. 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 core center without introducing much of 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. A set of threshold parameters will be employed in this approach.

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

$$\Delta V(r) \propto r^{d-1} \Delta r \qquad (3.20)$$

where Δr is a small segment of the radius r and symbol $\propto$ denotes the proportionality. This differential volume can be viewed as a hypershell (called simply shell hereafter) enclosed by two concentric hyperspherical surfaces (Fig. 3.3). A series of concentric shells around a mode are drawn with increasing r^2 by Δr^2. 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:

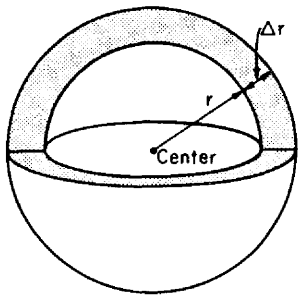
$$\Delta N(r^2) \propto \overline{p\ (|\underline{x}-\underline{\mu}|)}\ \Delta V(r)$$

$$\propto \exp\left(-\frac{r^2}{2\sigma^2}\right)\ r^{d-1}\ \Delta r \qquad (3.21)$$

The exponential term in Eq. 3.21 is that of an isotropic normal distribution with standard deviation σ. This relation can be rearranged by employing squared radius r^2 as

$$\frac{\Delta N}{r^{d-2}\ \Delta r^2} \propto \exp\left(-\frac{r^2}{2\sigma^2}\right) \qquad (3.22)$$

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. 3.4). Plots of $\ell n(\Delta N/r^{d-2}\Delta r^2)$ vs. r^2 may reveal a family of straight lines having the slope of $-1/2\sigma^2$ (Fig. 3.4b). 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.



Example Shell Volumes:

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

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

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

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

Fig. 3.3. A differential volume (hypersperhical shell) in three-dimensional space. Example formulas are given for one- through four-dimensional shells. Subscript i-d denotes i-dimension.

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. 3.5). One of them is the case where a cluster is well separated from the others even though the isotropic assumption is employed and that the threshold value covers nearly the whole region where most of the cluster data are located (for example, cluster A in Fig. 3.5). Another case is when the group of one cluster data are closely neighbored with the others (for example, cluster B in Fig. 3.5). Considerable overlaps between clusters may exist in this case.

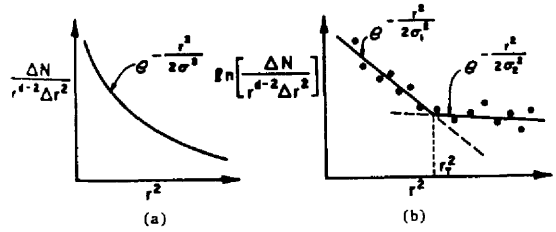


Fig. 3.4. A curve of Eq. 3.22 and data of a hypothetical two-cluster mixture. Data of a unimodal isotropic distribution yields an exponential curve shown in (a). Discrete data of a two-cluster mixture may produce a plot shown in (b), where two straight lines are approximate moving averages of two parts divided by r_t^2. The first straight line represents the population distribution of the first cluster with the parameter $\sigma_1^2\ r_t^2$ will be used as a threshold value for the cluster.

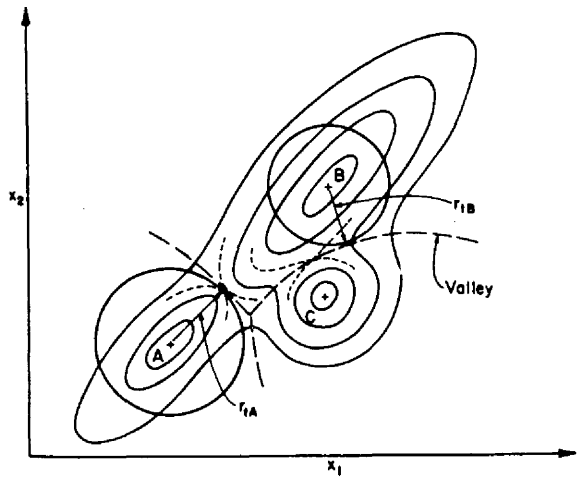


Fig. 3.5. Contours of a mixture probability density function (or population distribution) in a two-dimensional feature space. The mixture p.d.f. consists of three unimodal p.d.f.'s. A, B and C points are the modes of the clusters. r_{tA} or r_{tB} may be one of the threshold values estimated by the method shown in Fig. 3.4(b). They are interpreted as the shortest Euclidean distance to a valley, which is the natural boundary between the cluster and its neighbor one.

Extensive discussion of such possibilities is beyond the scope of this study. The purpose of estimating this threshold value is to initialize formation of a cluster by fusing data points within the neighborhood closer than this value. Euclidean distance is the most favorable distance measure from a probable mode position to each measurement point since any knowledge about a prospective cluster is not acquired at the moment.

3.4 Implementation of Hill-Sliding Strategy

3.4.1 Transform of LANDSAT MSS data into a probability space

Estimation of probability densities is required for implementation of hill-sliding strategy. Many interesting characteristics of the data structure can be directly observed when a set of data in the feature space is transformed into the probability space. Modes of clusters may be found simply by examining a probability histogram or searching local maxima of the density estimates. It is, however, not easy to compute probability densities from multivariate discrete measurements for the whole ranges of data in many cases. One of the most common difficulties is the requirement of excessive memory storage in a computer if the number of discretized levels is large. Another difficult fact is that most computer systems handle no more than three dimensional subscripts. Such a computer may not be efficiently manageable for the estimation and storage of multivariate probability densities having more than three variates.

The LANDSAT multispectral scanner (MSS) data are typical examples that often encounter the above difficulties. The first three channels of the LANDSAT data have a 7 bit range (0 through 127) while channel 4 has a 6 bit range (0 through 63). Hence, required memory storage to handle the whole range of probability densities often exceeds the capacity of commonly available computers if more than two channels of data are processed. These difficulties may be overcome for the LANDSAT data by utilizing their high correlations between neighboring spectral variables (channels).

The inherent nature of the spectral bands generally results in high correlations among bands (Fig. 3.1). Variables consisting of highly correlated channel data leave large portions of the multivariate space unused. Elimination of such unused spaces can reduce the required size of the central memory storage necessary to store discrete probability density estimates. A way to do this is to put an identification number on each point having non-zero probability density and to store only non-zero probability data. The sequential identification number for each non-zero probability datum refers to the corresponding compartment (cell) in the multivariate space. Multidimensional difficulty in multivariate probability estimates can also be overcome in this approach, since identity numbering applies to any place of the space wherever non-zero discrete density estimates exist.

A procedure to fulfill this idea is to divide the whole feature space where data range into several compartments (subcells) at first and compute data populations within each compartment. The next step is to eliminate any compartment which is empty (i.e., having zero population) and divide nonempty compartments into several smaller subcompartments. Repeat this procedure until the desired size (d-dimensional hypercube with volume ρ^d) in individual compartments is reached. The compartment obtained at the final stage is called a "cell" in this study. Each cell is labeled by a unique identification number and can be traced back to its original position in the feature space by the number. The population within a cell is converted to probability density in the way of the Parzen p.d.f. estimator (Fehlauer and Einstein, 1978).

Kittler (1976) investigated an approach similar to the hill-sliding algorithm. He constructed a chain of hypercubic (or hyperspheric) cells by sequentially arranging neighbor cells with lower or higher density values in accordance with hill-descending or hill-ascending stage. In this way, the multivariate density cells were aligned in the sequence of probability densities, something like one-dimensional distribution. Non-zero density cell points only appeared in the chain. This chain is similar to the results obtained by manipulation of the subcell algorithm in this study. However, finding neighbor cells requires computation of all the pairwise distances between the cell points, which is not desirable for large amounts of data. Hence, the operation of chaining all the cells in a desired sequence seems to involve laborious computation and excessive computer storage for most remote sensing applications.

3.4.2 Formation of initial clusters

It was discussed in the previous section that the location of the cell having the highest probability density would be the most probable position of the mode of a prospective cluster for normally distributed data. Once a probable mode is found, a series of shells around the mode are defined by r^2 and $r^2 + \Delta r^2$ in the multidmensional space where r is the Euclidean distance from the mode to the inner surface of the shell and Δ denotes the differential increment. The total population ΔN within a shell is computed by summing up the populations of all the cells contained in the shell. Generalized mean population densities defined by the left hand side of Eq. 3.22 are computed. Computation of the mean densities can be carried out by averaging the populations at two sequential observation points. Say, $r_i^2 < r_{i+1}^2$, where r_i^2 is the sequentially arranged value in increasing order. Shell populations corresponding to r_i^2 and r_{i+1}^2 are $\Delta N(r_i^2)$ and $\Delta N(r_{i+1}^2)$ respectively. The the generalized population density can be approximated

$$\frac{\Delta N(r_m^2)}{r_m^{d-2}\,\Delta r_m^2} \simeq \frac{\Delta N(r_i^2) + \Delta N(r_{i+1}^2)}{2} \Big/ \left[(r_m^2)^{d/2-1}\,\Delta r_m^2\right] \tag{3.23}$$

where

$$r_m^2 = \frac{r_i^2 + r_{i+1}^2}{2}, \tag{3.24}$$

$$\Delta r_m^2 = r_{i+1}^2 - r_i^2, \tag{3.25}$$

Estimates of the generalized mean population densities at two or more different r_m^2 values lead to computation of σ^2, coefficient of the exponent in Eq. 3.22, by

$$\sigma_{mij}^2 = \frac{1}{2}\,(r_{mj}^2 - r_{mi}^2)/\ell n\left[\frac{\Delta N(r_{mi}^2)}{\Delta N(r_{mj}^2)}\left(\frac{r_{mj}}{r_{mi}}\right)^{d-2}\frac{\Delta r_{mj}^2}{\Delta r_{mi}^2}\right] \tag{3.26}$$

The coefficient value obtained by this equation may be interpreted as that valid only in the segment between r_{mi} and r_{mj}. Different segments may yield different values of this coefficient depending on data distribution. It may not be judged simply that different coefficient estimates for two sets of data mean differences of class-origins. Such a decision would be based on statistical inference of the data structure due to random components contained within it. One of

the commonly applicable methods is that of the least sum of squared errors in evaluating σ^2 of Eq. 3.22. The statistically averaged value of the coefficient can be obtained by use of the method based on a set of data points rather than only two. General trends (i.e., slope of the line in Fig. 3.4b) of the coefficient changes are evaluated if proper sets of data are chosen for consecutive intervals (Fig. 3.6). The significant turning points indicate that shapes of the population distribution changes at the indicated distance levels and that other groups of data begin to influence the computed distribution at the levels. The distance at the first significant turning point will be used as the threshold parameter for a prospecting cluster.

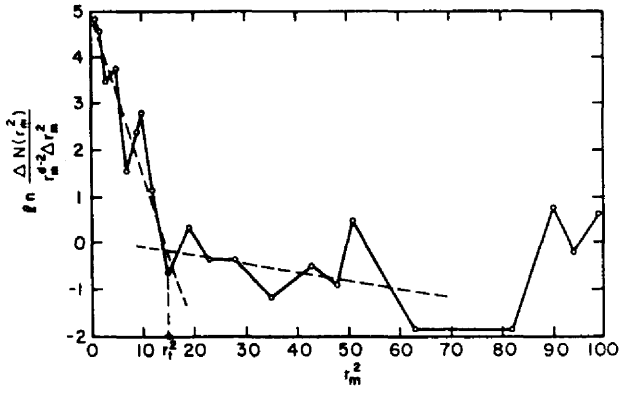


Fig. 3.6. The generalized density (Eq. 3.23) against the squared Euclidean distance from a centroid (Eq. 3.24). Each pair of consecutive two points in the figure defines the slope given by Eq. 3.27, which is the coefficient of the exponent in Eq. 3.22. The first straight line as an approximate moving average of the curve up to r_t^2 indicates that the data within this range have a unimodeal normal distribution forming a cluster. This curve was computed from 400 LANDSAT data over a Korean west coast (Park and Miller, 1978).

Search of the threshold value is carried out employing a transformed parameter, slope of the data distribution instead of the coefficient σ^2. The slope is defined by

$$\theta = -\frac{1}{2\sigma^2}$$

$$= \frac{1}{r_{mj}^2 - r_{mi}^2} \ln \left[\frac{\Delta N(r_{mi}^2)}{\Delta N(r_{mj}^2)} \left(\frac{r_{mj}}{r_{mi}}\right)^{d-2} \frac{\Delta r_{mj}^2}{\Delta r_{mi}^2}\right] \quad (3.27)$$

This is the parameter rearranged from Eq. 3.26. Average values of the parameter θ are expected as negative in a cluster (Fig. 3.6). The criterion function to find the threshold value of r^2 in clustering is formulated as

$$\theta_c = \bar{\theta} + f_\theta S_\theta \quad (3.28)$$

where

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 slope parameter θ of a cluster distribution is a function of Euclidean distance r. It is estimated as a moving average of the slopes for four consecutive increment steps of r_m^2 (i.e., for $\{r_{mi}^2\}_{i=\ell}^{i=\ell+3}$). Its next estimate is made for the next four r_{mi}^2 data after two increment steps forward (i.e., for $\{r_{mi}^2\}_{i=\ell+2}^{i=\ell+5}$) in this study. The threshold value r_t^2 is set by $r_{m_{i+1}}$ if one of the following criteria is met:

$$\theta_{new} > \theta_c \quad (3.29)$$

or

$$\theta_{new} \geq 0 \quad (3.30)$$

The first criterion given in Eq. (3.29) prevents other cluster cells from merging into the cluster under formation. The second criterion 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 in 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.

3.4.3 Refinement of initial clusters

The first application of hill-sliding algorithm to the data in the form of probability densities yields an initial cluster in each step discussed in the previous section. Initial estimates of cluster characteristic parameters can be deduced from this early stage of the cluster. Those values enable the computation of the clustering function G_i given by Eq. 3.6 for each probability cell. The function can be expressed in terms of normal density parameters by

$$G_i(\underline{x}) = \frac{D_i(\underline{x})}{2} - \ln p_i(\underline{\mu}_i) + \ln p(\underline{x})$$

$$= \frac{D_i(\underline{x})}{2} - \ln P_i + \frac{1}{2} \ln |C_i| + \frac{d}{2} \ln (2\pi) + \ln p(\underline{x}) \quad (3.31)$$

It was shown in the earlier sections that the expected value of the clustering function would be smaller than $\ln 2$ for any cell data within well-defined (separable) clusters. Computed values of $G_i(\underline{x})$ however, may not be close to the theoretically expected value at the final stage if Eq. 3.31 is applied to cell data at the forming stage of a cluster. 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.

One of the parameters uncertain in the initial steps of the hill-sliding approach is the a priori probability of the cluster. The a priori probability

of a cluster in the mixture distribution is computed by

$$P_i = \frac{\text{population in cluster } i}{\text{total population}} \qquad (3.32)$$

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

To allow for a certain level of fluctuations in G_i estimates, especially at a forming stage, a flexible criterion value rather than $\ell n2$ as in Eq. 3.14 is employed as

$$G_c = \bar{G}_i + f_G * S_G \qquad (3.33)$$

in which

G = 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 $G_i(\underline{x})$

The functional form is the same as in Eq. 3.28. A cell is tested as being a member of the cluster under formation by the criterion:

$$G_i(\underline{x}) \leq G_c \qquad (3.34)$$

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 in the hill-sliding algorithm. The empriccal constant f_G as well as f_θ in Eq. 3.28 is not a sensitive parameter, but selection of its value depends upon the detail required in cluster divisions in the end result.

Hill-sliding strategy in this research employed the cell having the highest probability density at hand as the most probable "mode" cell for a prospective cluster. It utilizes the natural tendency of the same cluster cells being more frequently located near the centroid. The next highest probability density point (cell position) being examined would be very close to the sample point just joining the group in the previous steps if the new point is a strong candidate for the group. Otherwise, it would be far from the cluster region since it was picked up from a hill side of another group (Fig. 3.7). The way to jump up and down from a hill to other hypothetical hills creates distinct distance gaps between the cluster being formed and a cell point originating from the other cluster candidates. These distance gaps allow gradual updating of cluster characteristic parameter values by first merging cells closer to the centroid. Gradual updating is important in this approach since estimated parameters at the earlier cluster-forming 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 in the manner of hill-sliding. Hill-sliding strategy stops if no single cell element is left over

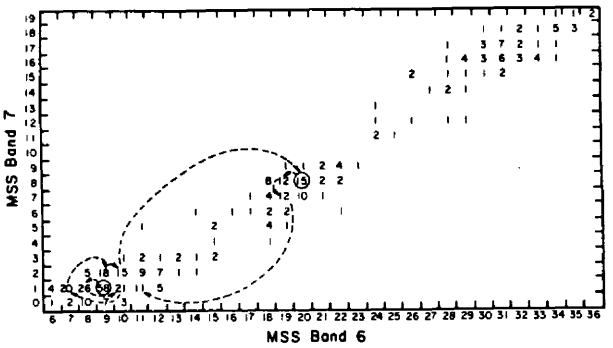


Fig. 3.7. Bivariate population distribution and sequence of testing membership candidacy for each population data point in hill-sliding algorithm. The highest population point (probable mode of the cluster) serves as a seeding point of the forming cluster. The sequence of testing membership candidacy follows the descending order of population data. The first several sequences are shown by marking arrows and two big distance jumps located between two possible cluster candidates are observed. These jumps reduce chance that elements to be of the clusters are mixed up with others while the characteristic parameters of the forming cluster are updated. The data are the same as in Fig. 3.1.

or if the maximum number of clusters set up in the program is reached.

3.5 Improvement of Overall Clusters

3.5.1 Evaluation of cluster compactness

Most of the measures on goodness of clustered results give relative comparison 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 meaures 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 in this thesis to evaluate the goodness of an individual cluster as

$$L_i = \left[\frac{|C_i|}{N_i - d}\right]^{1/d} \Big/ \left[\frac{|T|}{N - d}\right]^{1/d} \qquad (3.35)$$

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

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., $D_i(\underline{x}) = 1$ in Eq. 3.3) 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. 3.35 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 hyper-volume and the covariance matrix of those pattern vectors is always singular (Duda et al., 1973). However, the value of d may be any non-negative 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. This study employs a critical value of this parameter as one of input data. Initially formed clusters will be eliminated in the final consideration if the cluster compactness values exceed the critical value read in. The elements in the eliminated clusters are reevaluated in later steps. Another consideration made regarding cluster compactness is that, if the estimates of the compactness exceeds a certain value, say L_s, less than the critical value L_c, the cluster is refined by examining individual estimates of clustering functions for each constituent element. That is: a cluster will be 1) discarded or refined if

$$L_i < L_c \quad \text{(Criterion A)} \tag{3.36}$$

or 2) refined if

$$L_s < L_i < L_c \quad \text{(Criterion B)} \tag{3.37}$$

These criteria are coupled with additional conditions discussed in the next section to save excessive computation. Refinement will be described in a later section.

3.5.2 Evaluation of divergence 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 (see Eq. 3.3). 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 et al., 1972). It is defined by the sum of expectations of log-likelihood ratios in favor of a class against the other:

$$D_{ij} = \int_{\underline{x}} p_i(\underline{x}) \ell n \frac{p_i(\underline{x})}{p_j(\underline{x})} d\underline{x} + \int_{\underline{x}} p_j(\underline{x}) \ell n \frac{p_j(\underline{x})}{p_i(\underline{x})} dx$$

$$= \int_{\underline{x}} [p_i(\underline{x}) - p_j(\underline{x})] \ell n \frac{p_i(\underline{x})}{p_j(\underline{x})} d\underline{x} \tag{3.38}$$

The divergence is inferred as the total average information for discrimination between two classes. It possesses the following properties:

1. $D_{ij} = 0$ for $i = j$ (identical distribution)
2. $D_{ij} = D_{ji}$
3. $D_{ij} > 0$ for non-identical distributions
4. D_{ij} is additive for independent variates;

   $$D_{ij}(x_1, x_2, \ldots, x_d) = \sum_{k=1}^{d} D_{ij}(x_k)$$

5. Adding new variates never decreases the divergence

   $$D_{ij}(\underline{x} \varepsilon R^d) \leq D_{ij}(\underline{x} \varepsilon R^{d+1})$$

   where R^d denotes d-dimensional feature space.

It is noted in theory that the divergence is positive infinity ($D_{ij} = \infty$) when the two classes are perfectly separable. Higher values of the divergence estimates indicate better separability between the pair.

The divergence is used in the hill-sliding algorithm to analyze the clustering performance and to improve computational efficiency by cutting down unnecessary computations. The major reason for employing the parameter in this study is that it can be computed by a simpler formula under Gaussian assumption. For two Gaussian classes with unequal a priori probabilities, Eq. 3.35 is reduced to

$$D_{ij} = \frac{1}{2} \mathrm{tr}[(P_i C_i - P_j C_j)(C_j^{-1} - C_i^{-1})]$$

$$+ \frac{1}{2} \mathrm{tr}[(P_i C_i^{-1} + P_j C_j^{-1})(\underline{\mu}_i - \underline{\mu}_j)(\underline{\mu}_i - \underline{\mu}_j)^T]$$

$$+ (P_i - P_j) \ell n \frac{P_i |C_j|^{1/2}}{P_j |C_i|^{1/2}} \tag{3.39}$$

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). A working equation is formulated by normalizing with the sum of two class a priori probabilities as

$$G_{ij} = 2D_{ij}/(P_i + P_j) \tag{3.40}$$

where G_{ij} is called a normalized divergence. Refinement of initial clusters is carried out based on this information.

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).

$$E(\chi) = \int_{\underline{x}} p(\underline{x}) \, \ell n \, [1/p(\underline{x})] d\underline{x} \tag{3.41}$$

where $\chi = \{\underline{x}_n\}_{n=1}^{n=N}$ represents a set of pattern vectors (data). It is interpreted as the expected value of an information unit, $\ell n[1/p(\underline{x})]$, that is, the average uncertainty of the information source. As indicated by its functional form similar to that of divergence, Eq. 3.38, the entropy possesses properties similar to those for divergence. It always yields nonnegative values for information with discrete probability and has the maximum value for uniformaly distributed outcomes. The additive property for independent variables is also valid in this relationship (Maxwell, 1975). The ratio of the divergence to the entropy, $G_{ij}/E(x)$ or $D_{ij}/E(x)$, is comparable in any combination of variables. This value can determine relative separability of a 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. The objective of the clustering algorithm is to produce optimum partitioning of the data so that all the clusters lead to the separability values which are as high as possible.

Refinement of existing clusters for better partitioning of their elements entails expensive computing costs. Even a systematic search of optimum partitioning from every possible combination, like dynamic programming techniques, requires excessive computation unless numbers of both data and clusters are small (Duran and Odell, 1974). It is desirable to avoid irrelevant computation in improvement of partitioning. The following criteria are utilized for this purpose: a cluster will

(1) be saved if

$$\text{Min } G_{ij}/E(\chi) > D_s \quad \text{(Criterion C)} \tag{3.42}$$

even if Criterion A is satisfied, and

(2) not be reevaluated for refinement if

$$\text{Min } G_{ij}/E(\chi) > D_c \quad \text{(Criterion D)} \tag{3.43}$$

even if Criterion B is satisfied. Here, D_c and D_s are empirical values which are $D_c > D_s$ in general.

3.5.3 Valley refinement

Valley refining (or valley seeking) routine is devised to reevaluate probability cells which are generally located near boundary regions of clusters as well as those left over from discarded previously ill-defined clusters. These are simply classification procedures based on statistics of existing clusters. A cell having pattern vector $\underline{x}$ will be classified as $\underline{x} \in w_i$, in which the clustering function $G_i(\underline{x})$ is minimized among all i clusters. The previous membership of a cell (or data identities) may be changed through this reevaluation procedure. As such a membership change occurs, characteristic parameters of clusters are revised by deleting changed cells from the old cluster and merging them into the new cluster. The step of revising cluster memberships and statistics is carried out repetitively until no single element is moved or the maximum iteration is completed. Convergence has been achieved within 5 or less iterations in most cases. Slow convergence indicates that some of the clusters are not separable from the others or they are not defined well. In such a case, cluster statistics are revised by splitting cluster cells or deleting some of them in outer regions. Then valley refining is attempted interatively.

The present valley refining procedure should not be confused by the name with the valley-seeking technique described by Koontz and Fukunaga (1972). The latter is based on nonparametric density estimates in the neighborhood of each data point. A valley is sought as a clustering procedure and interpreted as the boundary between clusters lying in regions of low density. Classification of each point follows the rule of minimizing the fixed neighborhood penalty measured by the Euclidean distance.

3.5.4 Improvement in terms of the overall objective

Contrary to most clustering techniques which attempt to find the solution of the optimum partition directly through inexhaustible enumeration, the algorithm developed here does not aim at the global optimum solution. 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.

The objective of the present algorithm in this paper is to find a solution which minimizes

$$F = \sum_{i=1}^{I_c} (N_i - d) L_i^d \tag{3.44}$$

subject to

$$G_i(x_n \in w_i) \leq G_j(x_n \in w_i) \quad \text{for all } i, j \text{ and } n \tag{3.45}$$

$$1 \leq I_c \leq I_M \tag{3.46}$$

$$0 < L_i \leq L_c \quad \text{if } \min G_{ij}/E(\chi) < D_s \quad \text{for all } i \text{ and } j \tag{3.47}$$

$$M_c < M_i \leq N \quad \text{for all } i \tag{3.48}$$

where I_c, I_m and M_c are the number of resultant clusters, the maximum number of clusters, 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 $N_c < N_i \leq N$ 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 $N_i \leq d$ or $M_i \leq d$ (Duda et al., 1973). Parts of the constraints: 1) $I_c \geq 1$, 2) $L_i > 0$, and 3) $M_i \leq N$ are self-evident and there is no requirement for specification of these criteria in the algorithm. However, a user (or analyst) 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. 3.45 is equivalent to the decision rule of the maximum likelihood classification:

$$p_i(x_n \in w_i) \geq p_j(x_n \in w_i) \tag{3.49}$$

for all i, j and n, since the only other variable in clustering function $G_i(x_n)$ defined by Eq. 3.6 is $p(x_n)$, which is common in both sides. Hence, the inequality, Eq. 3.45, can be called the "maximum likelihood constraint" for each data point.

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

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

$$= \sum_{i=1}^{I_c} |C_i| / C_L^D$$

$$= \frac{N-d}{T} \sum_{i=1}^{I_c} |C_i| \tag{3.50}$$

Here the data characteristic length C_L is constant as well as the total number of data N and the total scatter matrix T. Hence, 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

$$W = \sum_{i=1}^{I_c} C_i \tag{3.51}$$

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). No efficient clustering algorithm has been proposed on the basis of this matrix, however. One of the most discouraging facts in using $|W|$ criterion is that the matrix will be singular unless either the number of clusters is greater than the dimensionality or the total number of the data is greater than the sum of the dimensionality and the number of clusters (Duda et al., 1973). 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 scatter matrix $|W|$ has been used as a measure of compactness of the clusters (Duda et al., 1973), but this interpretation is somewhat misleading. Following two simple cases illustrate inappropriateness of using $|W|$ as a clustering objective function or an overall cluster compactness measure (Fig. 3.8):

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

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

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

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

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

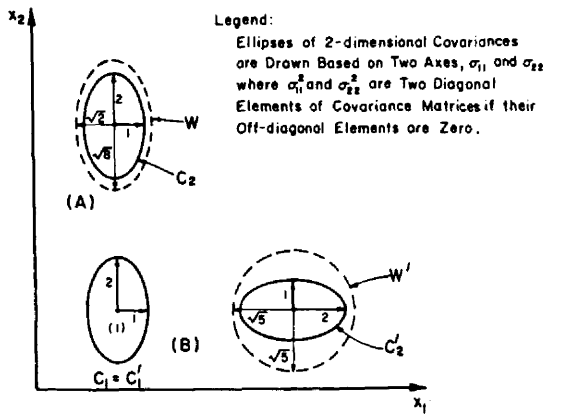


Fig. 3.8. Illustration of the total intragroups scatter matrices for two separable clusters in two-dimensional space. Each cluster has the same scatterness volume but different mean (centroid) from the other in either case. The pooled covariance matrices $W(= C_1 + C_2)$ and $W'(= C_1' + C_2')$ are different from each other since cluster orientations are not the same in both cases, even though $|C_1| = |C_2| = |C_1'| = |C_2'|$.

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 similar as possible. Any set of separable clusters would not make much difference whatever their natural shapes or orientations. This is another drawback to 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 matrice as the objective function. The set of constraints has provided some guidelines to overcome various undesirable 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. The solution sought by the present algorithm is an optimal (usually suboptimal) solution of the objective function, which satisfies the imposed conditions given by Eq. 3.45 through 3.48. The condition of Eq. 3.46 is mandatory unless the program is modified. The minimum number M_c of probability cells in a cluster should be larger than the number of variables (dimensions) d, so that a covariance matrix is nonsingular with better assurance. Hence, it is required that $M_c > d$. The limitation of cluster compactness value L_c has the greatest flexibility among the constraints. Its upper limit is totally up to an analyst's choice. However, a higher value may allow accepting clusters having elements scattered around other clusters. Too small value of L_c may not be achievable due to other coupled constraints. In other words, a set of unreasonable constraints may lead to no optimum solution of the problem.

This study does not intend to try all the possible enumerations to search for an improvement in partitioning, particularly because of probable excessive computation in handling multivariate probability density function. Instead, most computational procedures concentrate on achieving the results which satisfy the constraints while the objective function is minimized. Any local optimum solution detected will terminate the searching process.

The major procedure for searching the solution in the present algorithm is to redefine clusters by discarding and splitting clusters which do not satisfy the given constraints or which are considered as ill-defined clusters. Cluster goodness is based on satisfaction of criteria A through D described in Section 3.5.1 and 3.5.2. Splitting and redefining unsatisfactory clusters are carried out by separating cells in outer regions from those near their centroids. An intermediate cluster is formed by the cells left over by the separation process; it is then tested for cluster goodness. The separation process is performed iteratively until a satisfactory cluster is formed or no cells are left over to form a cluster. Cells separated from those near the centroid attempt to form another good cluster; if the new cluster(s) fail to satisfy given criteria, they will be merged with other existing clusters by the classification and valley refining procedure. Consequently, optimality of the new clusters as a solution to the problem is evaluated. Redefining clusters and optimality tests are repeated until an optimum solution is found. It should be understood that the solution obtained by this procedure may not necessarily be globally optimal. Ultimate satisfaction of the clusters as a solution is up to the analyst.

3.6 Outline of the Clustering Program

The basic step in detecting a distinguishable cluster from a set of data is to find any cohesiveness or discontinuity of the patterns distributed in a spatial or feature domain. This idea is exploited on the basis of the multivariate probability density estimation in the feature space. A major assumption in this formulation is that the mixture of various natural scene data constitutes a multivariate multimodal normal distribution. A group of the data surrouding a mode forms a cluster. A mode is a local maximum of a density function. Hence, the cell (or compartment in the feature space) having the highest density estimate is the seeding element for a cluster to be formed. Individual clusters are extracted one by one from the data set by the hill-sliding algorithm outlined in earlier sections until every cell is merged into one of the clusters.

The random nature of measurement data introduces a certain degree of uncertainty in computed results, causing misplacements of some cluster elements. The clusters obtained by the hill-sliding strategy are refined by reclassification of each cell which is weakly associated with its group or left over from discarded clusters which have loose compactness and low separability values. Reclassification of each cell is based on the maximum likelihood decision rule which uses a priori probabilities of individual clusters. The refining process is terminated when an optimality of the overall clustering objective function is found or the maximum iteration (of an input value) for refining is completed. The program is also devised to reuse the previously obtained results for further refining the clustered partitions, if desired (Fig. 3.9).

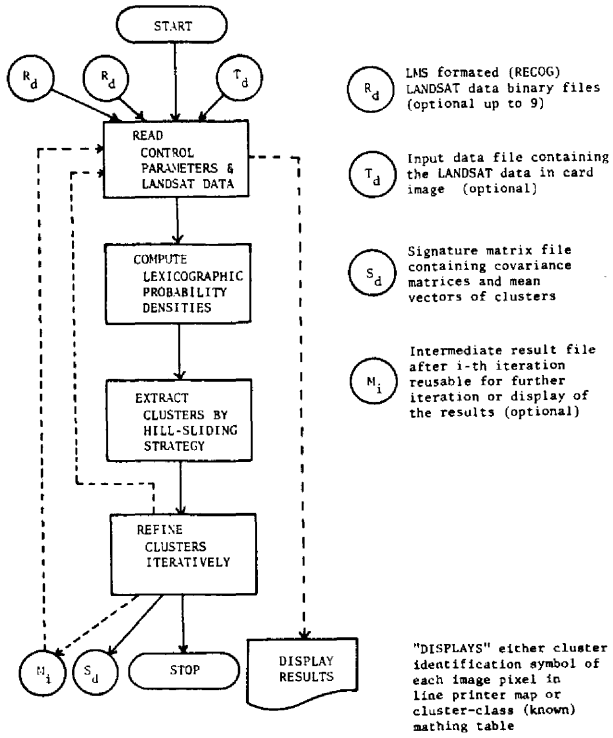


Fig. 3.9. A schematic view of the general flow structure in the hill-sliding clustering program. Solid lines indicate major flow directions of the clustering algorithm while dash lines show optional flow paths which can be repeatedly applicable.

Performance of a clustering program often relies heavily on the size of the sample data being processed. Each program has its lower and upper size limits which are internally set depending upon the algorithms employed and the available computing facility. Clustering algorithms based on density estimation are expected to produce better results if estimated densities reflect better characteristics of the assumed distribution. It is commonly accepted that more sample data leads to better characterization of their distribution and classification categories. Especially, the reasonable cluster sample size would be more than ten times dimensionality if predictions are made on the basis of covariance matrices as in the maximum likelihood method under normality assumption (Ball, 1965). Therefore, the minimum sample size for reasonable performance of the program developed here may be estimated by $10 \times d \times I_c$, where d and I_c are numbers of dimensions and expected clusters, respectively. This requirement may appear as a limiting factor in the use of a clustering program in multi-dimensional (multivariate) cases.

The present clustering program can read up to 1011 samples of maximum four variates (Table 3.1). It has been experienced that this upper limit of the sample size is a severe restriction in cases of four or more variate data with more than twenty expected clusters for analysis of the LANDSAT imagery. The LANDSAT computer compatable tape data over heterogeneous land cover areas vary in wide ranges. Probability density estimates within unit hypercubic cells (i.e., without increasing descretization interval) are widely scattered over the feature domain. Individual

Table 3.1. Limits of various input data parameters in the hill-sliding clustering program.

PARAMETERS	LIMIT (Max. in numbers)
INPUT RELATED PARAMETERS	
Samples	1,011
Variates (Channels or Dimensions)	4
Known classes: (if used for cluster-class matching table)	30
Disk data files (presently in LMS data format)	9
OUTPUT RELATED OR INTERNAL PARAMETERS	
Clusters	59
Nonzero probability cells (nonzero population cells in counts)	899
OTHER LIMITING SPECIFICATION	
Type of sample data	INTEGER

cells seldom have more than two pixels (picture elements) in population counts (alternative way to show probability density estimates). It is recommended in such a case that the probability densities be estimated within two or larger unit hypercubic cells.

Nature has provided many perplexing constituents of class/cluster information. Pooling a huge amount of natural scene data may produce only one or a few recognizable clouds if observations are made in limited scales. A huge cloud may not be deducible to what a user wants in detail. Selection of a reasonable sample size is pretty much heuristic. It should be based on statistical and logical grounds within permissible ranges of the program.

A clustering strategy may employ numerous techniques to implement its goal. The results obtained under the same strategy may differ from an algorithm to others depending upon the techniques employed. The present version of the hill-sliding algorithm was written for research and development. Hence, various alternative options were provided for analyses of the results by different options (Table 3.2). An investigator can trace the ways in which individual data points or probability cells move one cluster to another. A two-dimensional display of the intermediate results can also be obtained. Too many options, however, may cause the program to become too large, using more computer memory space and time.

Two major sets of data can be supplied for clustering analysis: one in the form of card images with labels of known class types; the other in the RECOG format, which is the standard input data format in the LANDSAT Mapping System (LMS) of Colorado State University (Appendix 1). The former produces a cluster-class matching table; while the latter displays the clustered results in the usual map-like format by the computer line printer. Statistics of resultant clusters are also obtained for examination and further use in classifying each identity of the sample data taken from various areas (Fig. 3.9). The usefulness of the results must be justified through real field data and user satisfaction.

Table 3.2 Key optional features in addition to the basic approach. The basic approach (in circle) was implemented in this research. However, the optional approach was provided to further analyze the particular performance of the program under such a condition or combination of such conditions.

Refining clusters based on

1 Maximum likelihood in terms of Gaussian probability density

2. Mahalanobis distance

Objective function:

1 Sum of cluster scatterness volumes

2. Sum of square roots of cluster scatterness volumes

Estimation of merging criterion values:

1 Updating

2. No updating

Splitting at each time of iteration among clusters which do not meet given compactness criterion

1 All clusters

2. The worst cluster only

3. The one having the largest scatterness volume

Time saving by

1 Limited search for values of threshold distance square

2. Limited checking possible movement of cells to other clusters

3. Do every case

Chapter IV
APPLICATION TO LANDSAT IMAGERY DATA

4.1 Processing Modules with the LANDSAT Mapping System

The applicability of the hill-sliding algorithm to an unsupervised classification of the LANDSAT imagery data will be evaluated in this chapter. The LANDSAT multispectral scanner (MSS) data over portions of the Chippewa River Basin, Wisconsin, and the Denver Metropolitan area were analyzed using the algorithm. A picture element (pixel) recorded on the LANDSAT computer compatible tape (CCT) represents a ground area of about 79 meter (E-W) by 79 meter (N-S) parallelogram inclined about 12 degrees east of north. Each pixel in the LANDSAT CCT contains the four band digital MSS data (Table 4.1). The values of MSS data are mainly the recorded levels of instantaneously reflected solar radiation from the scene within a resolution element, and they range 0 to 127 (7 bits) in bands, 4, 5, 6 and 0 to 63 (6 bits) in band 7.

Table 4.1 Spectral ranges of LANDSAT multispectral scanner bands.

Band	Wavelength (micrometer)	Color Range
4	.5 - .6	Green
5	.6 - .7	Red
6	.7 - .8	Near Infrared
7	.8 - 1.1	Infrared

The data array in the LANDSAT CCT is not arranged in the north-south geometric orientation as desired for spatially registered overlays of readily available map information, due mainly to orbital inclination of the satellites. The test data used for the study are rectified by the geometric rectification module in the preprocessing program of the LANDSAT Mapping System (LMS) package developed at Colorado State University (Appendix I). This LMS module employs the nearest-neighbor resampling technique with a uniform, completely filled output grid at a desired scale. The data elements resulting from this process can correctly represent a ground area of known geographic position. However, a significant mismatch may occur if the desired new map scale for either six or eight vertical lines per inch displays results in extensive oversampling or undersampling (Fig. 4.1). The optimal rectification can be achieved at the 8 x 10 line printer display of near 1:24,000 map scale in this module. Test data for the Chippewa River basin were rectified at a map scale of 1:24,000 for 8 by 10 line printer display; those of the Denver metro area were for 1 by 1 square block (microfilm) display. Additional preprocessing with the LMS used in this study includes reformation of the data file structure to efficiently store and manage the desired portion of the whole LANDSAT MSS data.

The clustering program developed here can produce a cluster map in the same format as the LMS classification map. It can also provide statistics of resultant clusters in the format of input data to the LMS supervised classification program. The LMS software

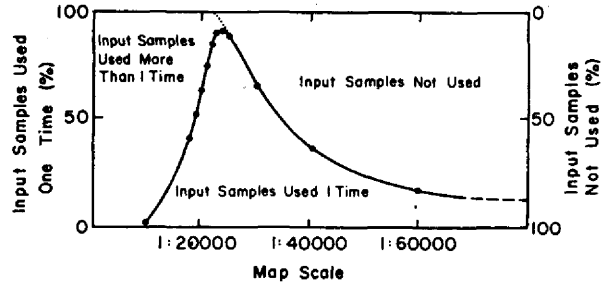


Fig. 4.1. Resampling efficiencies of the geometric rectification. The vertical axes show frequencies of usage or disusage of a LANDSAT data element when the nearest-neighbor approach is applied to resampling at various map scales. The curve applies to maps resampled in the ratio 8 N-S to 10 E-W for display at the scales shown on the 8 line/inch printer (Sung and Miller, 1977).

embraces several major image-processing phases including the preprocessing phase described above (see more in Appendix I). Its first phase, preprocessing, and last phase, classification were used in this study. The LMS classification module is a supervised classification algorithm based on the Gaussian likelihood method. An underlying assumption of the algorithm is that all the input classes have the same a priori probabilities (in other words, equally-divided mixing proportions in a mixture distribution). The best performance of this classifier can be expected when all the classes have the distinctive unimodal normal distributions with equally-divided mixing proportions in the available multidimensional feature space (Park and Miller, 1978; Maxwell et al., 1977). It is, however, seldom probable to have the same mixing proportions for all the natural clusters or all known classes in the land-use/land cover classification. The hill-sliding algorithm computes the mixing proportions by naturally grouping the data. There would be substantial differences between both results obtained by the hill-sliding algorithm and by the LMS classification module even if the same cluster statistics were used (Fig. 4.2). This argument will be tested and analyzed in the following sections.

The program developed herein can be used for analyzing the sample data whose class categories are known. Such data are to be read as punched card images in the same format as in output data of the LMS "PRINT/PUNCH" program (Fig. 4.3). The results are given in a table of matching cluster-classes.

4.2 Cluster Analysis of Denver Metropolitan Area

A cluster is defined as a group of data bounded by a chain of probabilistic valley in the multidimensional feature space. The cluster may be a natural land cover/land-use class or a class subgroup. It may represent a hardly separable mixture of two or more natural classes. The task is to separate each distinguishable cluster, rather than to decompose those unseparable mixed classes. An identifiable cluster is defined here as a group of data within the region bounded by the probabilistic valleys, which constitutes a series of local minima in probability density

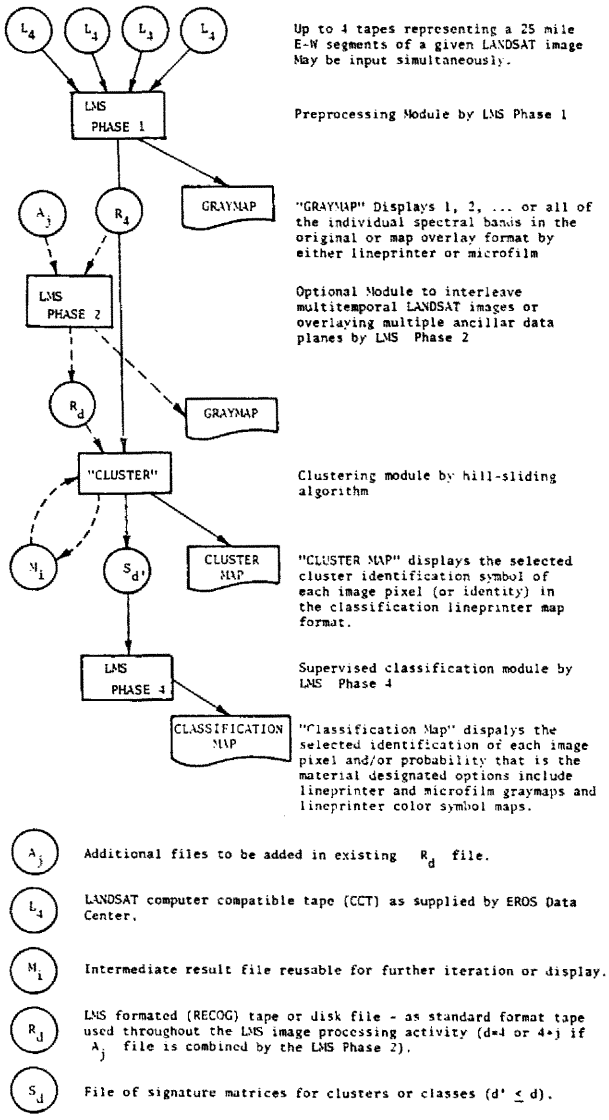


Fig. 4.2. Option-1 flow chart of input/output files and major computer processing modules to produce cluster/classification maps.

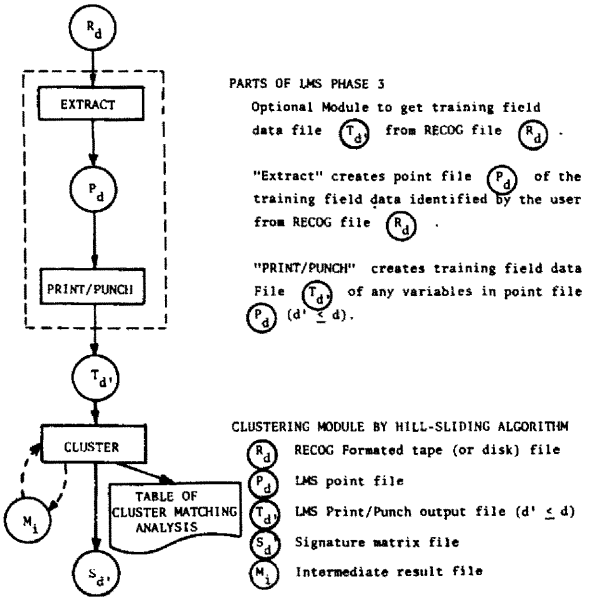


Fig. 4.3. Option-2 flow chart of input/output files and major computer processing modules to produce cluster-class matching tables. The input data file contains prototype class information for each identity (so called "training field data") and is read as simple card images which are in the format of the output from PRINT/PUNCH Program of the LMS.

estimates. A cluster may cause some conflict when it is compared with the commonly-adopted USGS land-use/land cover classification scheme (Table 4.2). Clusters of measurement data are formed on the basis of the cohesiveness or structural similarity in their measured quantities, while land-use/land cover classes are mainly for user purpose. Therefore, the latter may have many structural subcategories, some of which may be quite similar to those of other classes. The degree of confusion strongly depends upon the resolution the data possess. For instance, some parts of urban residential areas would reveal the data spectrum similar to those of grasslands. There is little difference between lakes and reservoirs in the values of the LANDSAT MSS data. One of the beneficial points in clustering data is that the result may reveal previously unknown subclasses which may be meaningful in interpreting various aspects of the data structure.

The data used for testing the clustering program were selected from 4,100 geometrically rectified ground truth samples of LANDSAT imagery over Denver Metropolitan area on August 15, 1973. These sample data were compiled and used as a part of the data base for a series of extensive landscape/land-use inventory modeling efforts (Miller et al., 1977; Tom et al., 1977). The base data were resampled from the LANDSAT computer compatible tape data to yield 1.11 acres per square cell and overlay 1:24,000-scale topographic maps by the LMS preprocessing routine. They covered a square of 24 by 24 miles in cellurized 576 rows (or lines) by 576 columns. The ground truth samples were extracted from the base data by a self-verifying, uniform-grid-sampling procedure. The set of training data represented a one-ninth by one-ninth (1/81) sampled image of known land use by reference to the 1972/1973 USGS land use. This systematic point sampling process yielded the ground truth data of land-use types proportional to frequencies of their actual occurrences in the field (Table 4.2).

One of the major drawbacks in uniform-grid sampling is that numbers of some class samples approximately proportional to their population in the area might turn out too few to produce statistically sound characterization. When prediction is made based on covariance matrice, samples of at least ten times dimensionality are required to reduce chances of error (Ball, 1965). Examination of the Denver data (Table 4.2) immediately reveals that some of the prototype classes do not satisfy this requirement especially for two or higher multivariate analysis. Those samples may fail to form clusters even if they have very distinguishable characteristics from others.

Another difficulty in analyzing this set of data is due to limitation imposed in the clustering program. Suppose that all four LANDSAT MSS data are analyzed and more than twenty clusters are expected

Table 4.2 Hierarchical land-use/land cover classification scheme and number of samples selected from the LANDSAT imagery of Denver Metropolitan area. The various levels of USGS Circular 671 System (Anderson et al., 1972) with minor changes as Professional Paper 964 (Anderson et al., 1976) are shown. The numbers of sample picture elements (pixels) were extracted proportionally to frequencies of land-use types acquired by manual air photo interpretation and automated LANDSAT image analysis (Tom et al., 1977).

Digital Codes	FIRST-ORDER LAND-USE/LAND-COVER Second-Order Land-Use/Land-Cover Third-Order Land-Use/Land-Cover	Sample (pixel)
1	URBAN AND BUILT-UP LAND	
11	Residential	1245
12	Commercial and Services	142
121	Recreational	160
13	Industrial	151
14	Extractive	57
15	Transportation, Communications, and Utilities	76
151	Utilities	12
16	Institutional	344
17	Strip and Clustered Development	
18	Mixed Urban	
19	Open and Other Urban	284
191	Solid-Waste Dump	4
192	Cemetery	25
2	AGRICULTURAL LAND	
21	Cropland and Pasture	
211	Nonirrigated Cropland	589
212	Irrigated Cropland	7
213	Pasture	430
22	Orchards, Groves, and other Horticultural Areas	
23	Feeding Operations	
24	Other Agricultural Land	
3	RANGELAND	
31	Grass	305
32*	Savannas	
33	Chapparal (taken as brushland)	39
34*	Desert Shrub	
4	FOREST LAND	
41	Deciduous	
411	Deciduous/Intermittent Crown	11
42	Evergreen (Coniferous and Other	
421	Coniferous/Solid Crown	50
422	Coniferous/Intermittent Crown	2
43*	Mixed Forest Land	
5	WATER	
51	Streams and Waterways	4
52	Lakes	60
53	Reservoirs	18
54*	Bays and Estuaries	
55	Other Water	
6	NONFORESTED WETLAND	
61	Vegetated	
	Bare	
7	BARREN LAND	
71*	Salt Flats	
72*	Beaches	
73	Sand Other Than Beaches	24
74	Bare Exposed Rock	
741	Hillslopes	61
75	Other Barren Land	
8*	TUNDRA	
81*	Tundra	
9*	PERMANENT SNOW AND ICEFIELDS	
91*	Permanent Snow and Icefields	
	TOTAL	4100

* Land-use/land-cover type not found in the Denver Metropolitan Area

in the results. Then at least 800 (10x4x20) samples are necessary for reliable results. However, some cluster sizes are significantly larger than others. Thus, much more than such a minimum number of required samples are needed for better performance of the clustering process. The present clustering program developed for this paper can read up to 1011 samples of maximum four variates (Table 3.1). These limits are due mainly to the computer central memory capacity (250K) available at Colorado State University. Exclusion of any external memory devices, as well as inclusion of many optional result-checking routines, made this restriction more severe. The present version has been written purely for development of the algorithm and needs fast turnaround for testing intermediate results in various steps. Input limitation can be improved by employing external memory devices and deleting optional routines in a future version.

These upper limits of input data parameters force one to use only a portion of the collected ground truth data (Table 4.2). The following are criteria for selection of samples from the data pool to evaluate the effectiveness of the clustering program:

1) use all the samples in classes which have less samples than 10 x d.

2) select samples roughly proportional to the class sample populations if they are more than 10 x d.

Performance of the program was tested for chosen data in terms of

1) ability to decompose two apparently separable class mixture distribution, and

2) ability to decompose all-class mixture distribution.

A test of the first case was attempted using all the samples of rangeland-grass (305 pixels) and forest-evergreen with solid crown (55 pixels). Two populations were well separated in MSS bands 5 and 7 (Fig. 4.4). Only four of 183 cells of the originally discretized unit have elements originated from both classes. Most of the evergreen class data were distributed in the region having lower values of both bands while those of the grass were in that of higher values, particularly in band 5.

Clustering these two class samples was carried out following the first two sets of input parameters:

Clustering Parameter	Run 1	Run 2	Run 3
Number of samples	355 (2 classes)	355 (2 classes)	975 (24 classes)
Used bands	5, 7	5, 7	5, 7
Probability cell size in each band	1	2	1
f_θ (Eq. 3.28)	2.7	2.7	2.7
f_G (Eq. 3.33)	2.0	2.0	2.0
L_c (Eq. 3.36)	0.99	1.6	1.6
L_s (Eq. 3.37, fixed in the program)	$\frac{1}{4} L_c$	$\frac{1}{4} L_c$	$\frac{1}{4} L_c$
D_s (Eq. 3.42)	3.0	3.0	3.0
D_c (Eq. 3.43, fixed in the program	$10 D_s$	$10 D_s$	$10 D_s$
M_c/No. of bands (Eq. 3.48)	2.5	2.5	2.5
Condition to stop refining			
F_{old}/F_{new} -1 (F in Eq. 3.44)	≤ 0.001	≤ 0.001	≤ 0.001
Number of changes in cluster element	$< 0.5\%$	$< 0.5\%$	$< 0.5\%$
Maximum iteration	4	4	7

In the result of Run 1, eleven clusters were formed (Fig. 4.5, Table 4.3). Three of them had pixels assigned in both classes. Total commission error, which was computed by the sum of erroneously assigned pixels divided by the total number of pixels, was 2.5 percent. This error rate was strikingly small. Hence, it indicated that the two classes were well separated. The characteristic length of the whole data was 1.6 while most individual cluster compactness values were less than one except for the ninth cluster (Table 4.4). The sum of determinants of cluster covariance matrices were far smaller than the determinant of the overall data covariance (or total scatter) matrix. This meant that most of the unoccupied spaces went out of consideration as the clusters were formed.

The second RUN differed in cell size from the first. The cell size was two by two discretized units of two band data (Fig. 4.6). Six clusters were formed from 100 probability cells (Fig. 4.7, Table 4.5). The error rates and characteristic lengths were nearly the same in both cases. The determinant of the total

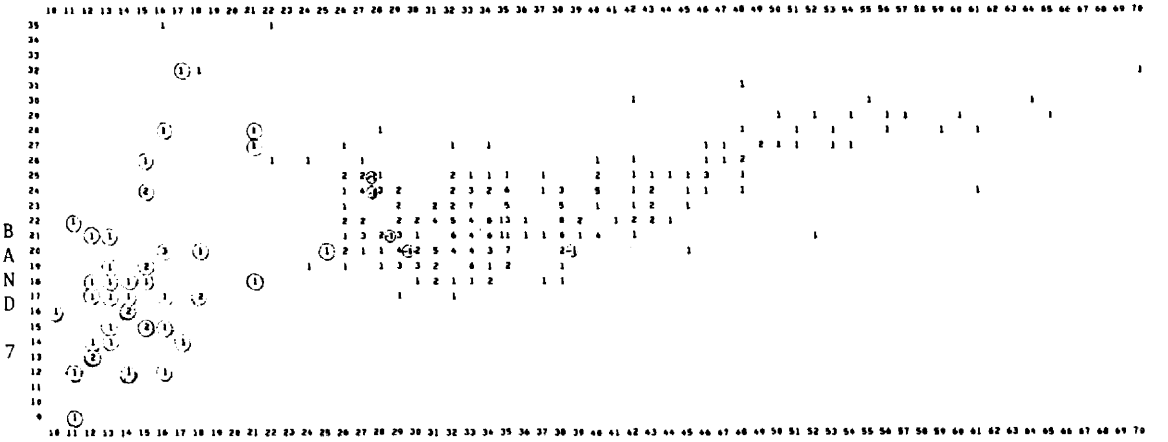


Fig. 4.4. Mixture population distribution of rangeland-grass and forest-evergreen class samples in the LANDSAT MSS bands 5 and 7 for Run 1. The numbers are the sum of populations of two classes in the cell which has corresponding values in bands 5 and 7. The numbers in circles are populations from the forest-evergreen class only and "-" sign indicates that the portions are from the forest-evergreen class among the sum in the cell.

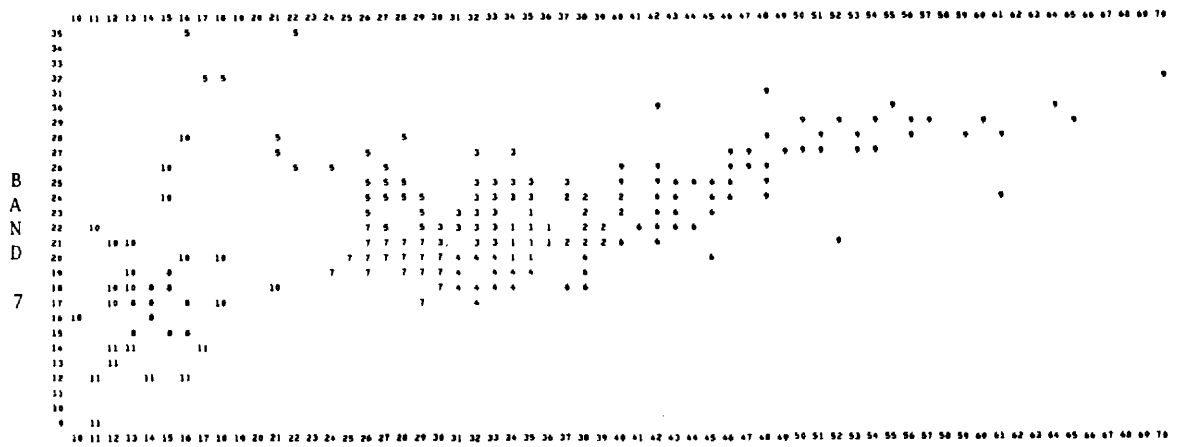


Fig. 4.5. Display of the resultant clusters in Run 1. The numbers are cluster labels. The numbering was made sequentially when it was formed.

Table 4.3 (a) Class-cluster matching matrix in Run 1. The numbers are those of picture elements in each cluster. The clusters are listed in ascending order of the mean values of MSS band 4 data (1). Commission errors (2) were computed by (confused numbers/sum) * 100.

Land Use/Land Cover Class	Cluster Number (1)											Sum
	11	8	4	10	7	1	6	2	3	5	9	
Rangeland-grass	0	0	31	0	30	55	31	33	58	28	39	305
Forest-evergreen	9	13	0	19	3	0	1	0	0	5	0	50
Sum	9	13	31	19	33	55	32	33	58	33	39	355
Commission Error (2) (percent)	0	0		0	9.1	0	3.1	0	0	15.2	0	2.5

Table 4.3 (b) Summary table in Run 1.

Land Use/Land Cover Class	Clusters (1)		Sum
	11,8,10	4,7,1,6 2,3,5,9	
Rangeland-grass	1	305	305
Forest-evergreen	41	9	50
Commission Error (2) (percent)	0	2.9	2.5

Table 4.4. Some characteristic values pertinent to clusters in Run 1. The determinant of cluster covariance matrix is the scatterness volume of the cluster. Smaller values of cluster compactness indicate more compact clusters.

i	Cluster Number ① 11	8	4	10	7	1	6	2	3	5	9	Sum	Overall Data
Determinant of Covariance, $\|C_i\|$	10	2.2	1.2	12	3.1	.30	10	.94	3.6	45	183	271	900
Cluster Compactness L_i	76	.28	.12	.53	.20	.047	.36	.11	.16	.75	1.4	-	1.6 ③

① Refer to Table 4.3

③ Characteristic Length

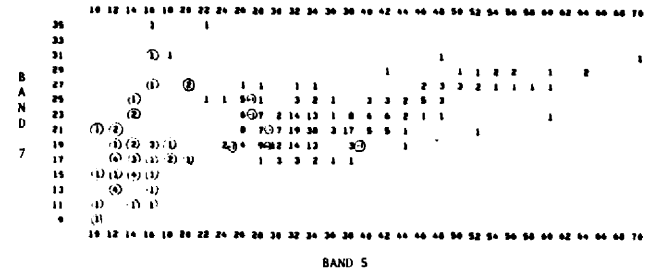


Fig. 4.6. Mixture population distribution of rangeland-grass and forest-evergreen class samples in the LANDSAT MSS band 5 and 7 with discretization interval 2. The numbers are the sum of populations of two classes in the cell which has corresponding values in bands 5 and 7. The numbers in circles are populations from the forest-evergreen class only and "-" sign indicates that the portions are from the forest-evergreen class among the sum in the cell.

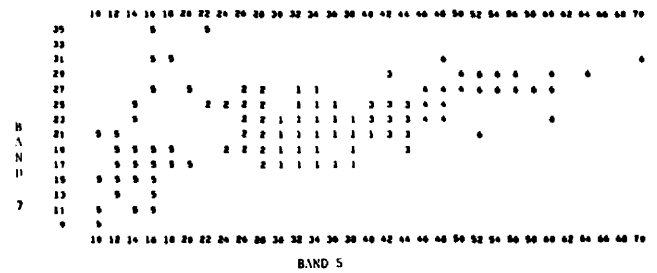


Fig. 4.7. Display of the resultant clusters in Run 2. The numbers are cluster levels. The numbering was made sequentially when it was formed.

Table 4.5 (a) Class-cluster matching table in Run 2. The numbers are those of picture elements in each cluster. Refer to Table 4.4 for the notes of ① and ② .

Land Use/Land Cover Class	Cluster Number ① 5	1	2	3	4	6	Sum
Rangeland-grass	3	186	49	30	18	19	305
Forest-evergreen	44	1	5	0	0	0	50
Sum	47	187	54	30	18	19	355
Commission Error ② (percent)	6.4	.5	9.3	0	0	0	2.5

Table 4.5 (b) Summary table in Run 2.

Land Use/Land Cover Class	Clusters 1,2,3 4,6	5	Sum
Rangeland-grass	302	3	305
Forest-evergreen	6	44	50
Sum	308	47	355
Commission Error ② (percent)	1.9	6.4	2.5

scatter matrix was about 920, which was two percent larter than it was in RUN 1 (Table 4.6). The results of both runs compared fairly well except for computer central processing times, in which RUN 2 used only one-third of the time of RUN 1 (Table 4.7). These two simple cases indicated that the size of discretization interval has a significant effect on computation time when discrete probability density estimates are used. It is unnecessary to use too small discretization intervals unless the mixture distribution characteristics are significantly changed by increasing the

Table 4.6 Some characteristic values pertinent to clusters in Run 2. Refer to Table 4.4. for further explanation.

i	Cluster Number 5	1	2	3	4	6	Sum	Overall Data
Determinant of Covariance, $\|C_i\|$	220	24	11	7.5	3.8	13	276	920
Charactoristic Length, L_i	1.4	.22	.28	.32	.30	.54	-	1.6

Table 4.7. Summary of performance of the clustering program in three runs. Run 1 and Run 2 were for two classes (rangeland-grass and forest-evergreen), 355 samples and different discretization intervals. Both results were fairly comparable except for computer central processing time. Run 3 was for 975 samples of 24 classes. The estimate of commission error in Run 3 was not quite clear since class-cluster matchings were not clearly determinable.

Description	RUN 1	RUN 2	RUN 3
Probability cell	183	100	472
Determinant of total scatter matrix (scatterness volume)	900	920	5100
Characteristic length	1.6	1.6	2.3
Cluster	11	6	36
Sum of determinants of cluster covariances	272	276	1438
Objective function	14.8	11.9	24.0
Commission error	2.5	2.5	73.2
Approximate computer CP time (second) for 1st 4 iteration	38	13	151

interval. The results in these two test runs can be summarized as: 1) the clustering program decomposed two class samples satisfactorily; 2) numbers of subgroups in each class depended upon the discretization interval of data for probability density estimation; and 3) computer processing time varied with the discretization level.

Another test (Run 3) was performed with 975 samples chosen from all the classes of Denver Metropolitan area for two MSS bands 5 and 7 (Table 4.8). Populations of the sample data were widely scattered in the region where the band 5 data ranged from 8 to 75 and the band 7's were from 0 to 48 (Fig. 4.8).

The clustering program yielded 36 clusters based on 472 discrete probability density estimates (Table 4.8, Fig. 4.9). Not many clusters could be claimed as distinctively (say, having more than 50 percent commission accuracy) representing a class or a subgroup of a class in the result of this test run. Elements of some clusters came from various classes rather than any single (representative) class. This fact indicated that most of the established classes were not well separated from each other, at least in these two MSS bands. Samples of urban, agricultural and rangeland classes were spread in many clusters. Forest and water classes are well distinguished from the other. No cluster was found that distinguishably represented classes of utilities, solid-waste dump, irrigated cropland, grass, deciduous forestland, evergreen/intermittent crown, stream or rock. It was interesting that the number of grass samples was not small compared with others. Most of this class pixels, however, were confused with other classes, especially the pasture class. This suggested that any attempt to classify other pixels based on information derived from these training samples might lead to unpredictable results. Major causes of such a conflict seemed to come from samples of complex urban classes. It may be summarized through this result that many urban-type classes were not adequately definable in these spectral band signatures, specifically for the resolution level (about one acre) of the LANDSAT. This summary could be convincing by the low commission accuracies in the result (Table 4.8). The overall commission accuracy for all the classes was 26.8 percent. The second- and first-order class-cluster matching matrix showed 29 and 49.4 percent commission accuracies, respectively.

Performance of a clustering program might be evaluated by visual examination of displayed clusters in a one- or two-dimensional feature space (Figs. 4.5, 4.7, 4.9). The hill-sliding program demonstrated the ability to decompose the relatively complex multivariate normal mixture distribution in these test runs.

4.3 Mapping Land Cover/Land-Use of a Chippewa River Basin Area

Predicting land cover/land-use activities is a major effort in the utilization of LANDSAT imagery since the first LANDSAT (formerly Earth Resources Technology Satellite) series was launched into a near polar orbit on July 23, 1972. Technology to further predict agricultural crop harvests or to inventory natural resources has been greatly advanced in recent years. However, a basic question remains unresolved; how much of man's intervention is required for implementing such a task. Cluster analysis is a way to lessen man's burden in this task. An experiment was carried out for this study using LANDSAT imagery data over the Chippewa River Basin (Simons and Chen, 1978).

The present cluster program can generate statistics of clusters, especially for the maximum likelihood classification. Unsupervised classification maps can be produced for any size area based on these statistics, for example, using the LANDSAT Mapping System (LMS) of Colorado State University (Fig. 4.2). The LANDSAT I imagery from May 11, 1976, was analyzed to estimate areal extent of land cover/land-use classes over the lower Chippewa River Basin area. The LANDSAT computer compatible tape data of eight rectangular regions along the river from Lake Pepin to Eau Claire, Wisconsin, were preprocessed by the LMS with geometric rectification at the scale of 1:24,000 for computer line printer displays. Clustering was performed by the hill-sliding algorithm developed herein using 902 uniform grid samples among the total 178,519 pixels (picture elements). All four MSS band data were used for this cluster analysis. Statistics of 25 clusters obtained by this analysis were applied to the LANDSAT data of the whole area to produce cluster maps by the LMS classification module.

A portion of the cluster maps produced by both the clustering program and the LMS classification module compared to evaluate differences of basic underlying assumptions on the a priori probabilities of clusters proportional to individual cluster populations, while the results by the maximum likelihood classifier of the LMS were based on equal a priori probabilities for all classes. Substantial differences are expected due to those of mixing proportions. The results showed a difference of five points among 130 points in total (3.8 percent) over the area (Fig. 4.10, Table 4.9). This amount of difference is strikingly small when it is compared with many other uncertainty factors in measurement and distribution estimation. The use of the maximum likelihood method with equal mixing proportions in all the cluster/classes might be justified by this finding.

Table 4.8. Class-cluster matching matrix for 975 samples chosen from all the classes. The numbers are those of picture elements (pixels) in each class-cluster matching box. The class(es) having the largest number of pixels in each cluster (in circle) may be claimed as representative class(es).

First Order Land Use/ Land Cover	Digital Code	Brief Description	CLUSTER NUMBERS 25 8 10 20 25	21 31 30 15 36	14 33 22	17 35	19 16	7 12 2	29	9 6 5 29	11 26	27 24 34	13 32	18	3 1	SUM
Urban and Built-Up Land	11	Residential	⑩ ⑭ ⑦ ⑦ ⑤	1	2 1	4 1	3 1	1 2 1	1	1 4 7	1	1			1 5	83
	12	Commercial	1 1 1	⑦ 2	4 1	2	2 2	1 5 2	1	1 1 2 1	2				3 1	43
	121	Recreational	3 1 1 2	⑨ ⑥ 3 2			1	2 3 3	1	3 2 2	2				3 1	54
	13	Industrial	1 2	3	⓪ 3 2	8 4	1 1	1 1	1	1 2 4 1	1		2 2		2 1	51
	14	Extractive	1 2	2 1	4 2	⑪ ⑤	2 2	1 3 1		1 1 3	4 1				3 2	57
	15	Transportation	1 1	1 2	2 1	3 3	④ 3	4 2	1	1 3 2	1	1			2	38
	151	Utilities		1 2	1	1		2	1	1 1	1					12
	16	Institutional	2 2	2	1	2	1	⑩ ⑦ ⑥	5	1 2 6	7				5 5	69
	19	Open, Urban	4 1 2	2	1 1	3 4	3	2 2 2	5	4 1 3 1	4				2 7	57
	191	Dumping			1	1		1	1	1						4
	192	Cemetery	2 1 2					1 2	1	1 2 2	6				2	25
Agricultural Land	211	Crop, nonirr.	3 2 2 1 1	1 2 2	4	8 3	2 2	4 4	3	4 ⑩ 4 1	3				3 2	73
	212	Crop, irrig.	1 1	2 1				1	1	1	1					7
	213	Pasture	4 2 1 1	1 3	1 1	1	1	7 4 5	4	3 5 ⑩ ⑤	3	1	1		2 5	72
Rangeland	31	Grass	2 4	2 2 1	1	1 2	1 1	6 5 2	1	2 5 7 1					6 11	61
	33	Brushland	1 3 1					3 1 1	1	4 1	⑭ ⑤	1 1			1 5	39
Forest Land	41	Deciduous F.	1 2					2			4					11
	421	Evergreen, S.	1		1			2 1		2	8 2	⑫ ⑧ ⑤				50
	422	Evergreen, I.								1	1					2
Water	51	Streams			1		1		1	1						4
	52	Lakes	1 1		3			6 1		2 2	3 2		⑤ ⑥	4	2	60
	53	Reservoirs			1				1	1			3 5	⑦	1 1	18
Barren Land	74	Rock	3					2 1		4 4	5	1			2	24
	741	Hillslopes					3	4 3 5		8 7	4 1	2 1	1		⑥ ⑨	61
SUM			40 29 28 14 7	14 10 9 19 7	31 13 9	46 23	18 19	55 52 32	29	22 58 70 10	73 11	18 17 7	29 15	11	40 62	975
Commission Accuracy (percent)		third order	40 48 25 50 43	50 70 67 16 29	35 23 22	24 22	22 16	20 13 19	19	18 17 14 30	19 27	83 76 71	79 47	64	15 25	26.8
		second order	39.2	42.4	30.2	23.2	18.9	17.3	21	24.4	20.2	78.6	68.2		23.5	29.0
		first order	63.5										89.1			49.4

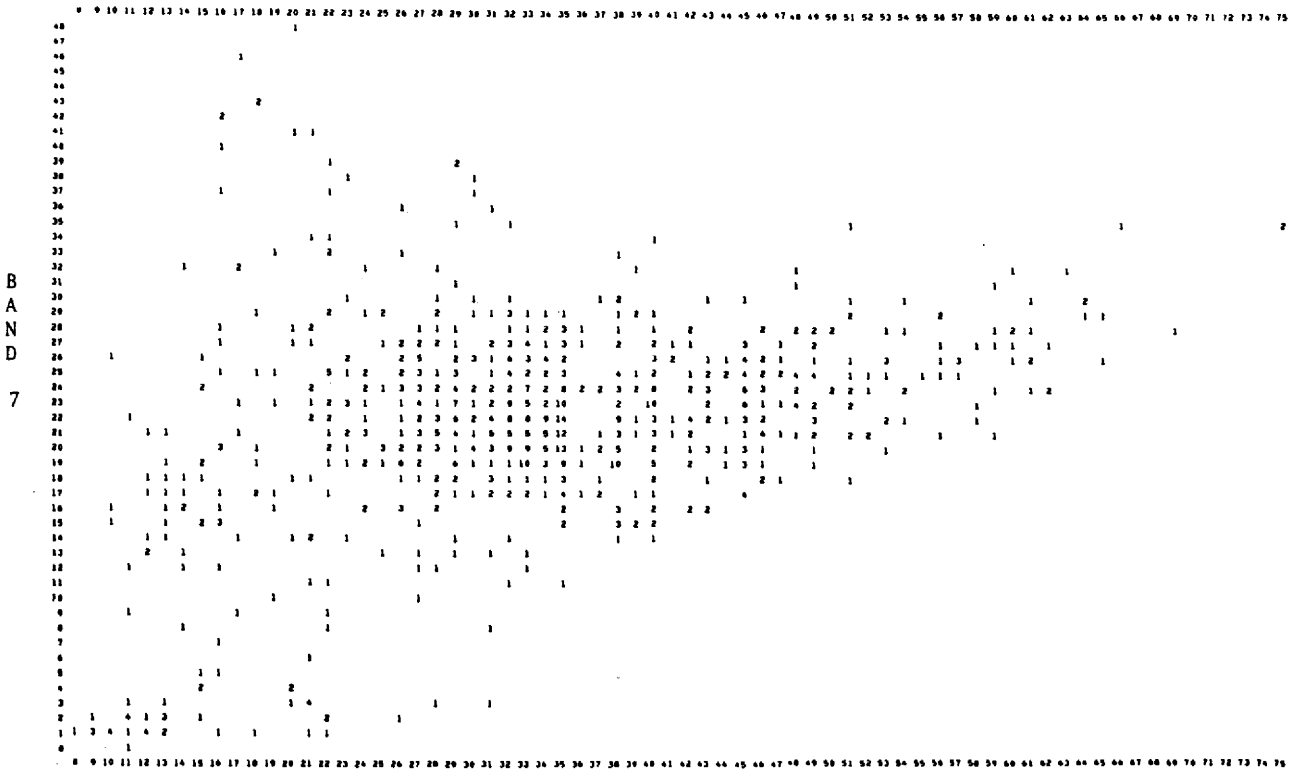


Fig. 4.8. Mixture population distribution of sample data for Run 3. 24 classes are present in the display of the LANDSAT MSS bands 5 and 7 data. The numbers are frequencies of occurrence in each two-dimensional cell. 472 discrete population points (cells) are shown here.

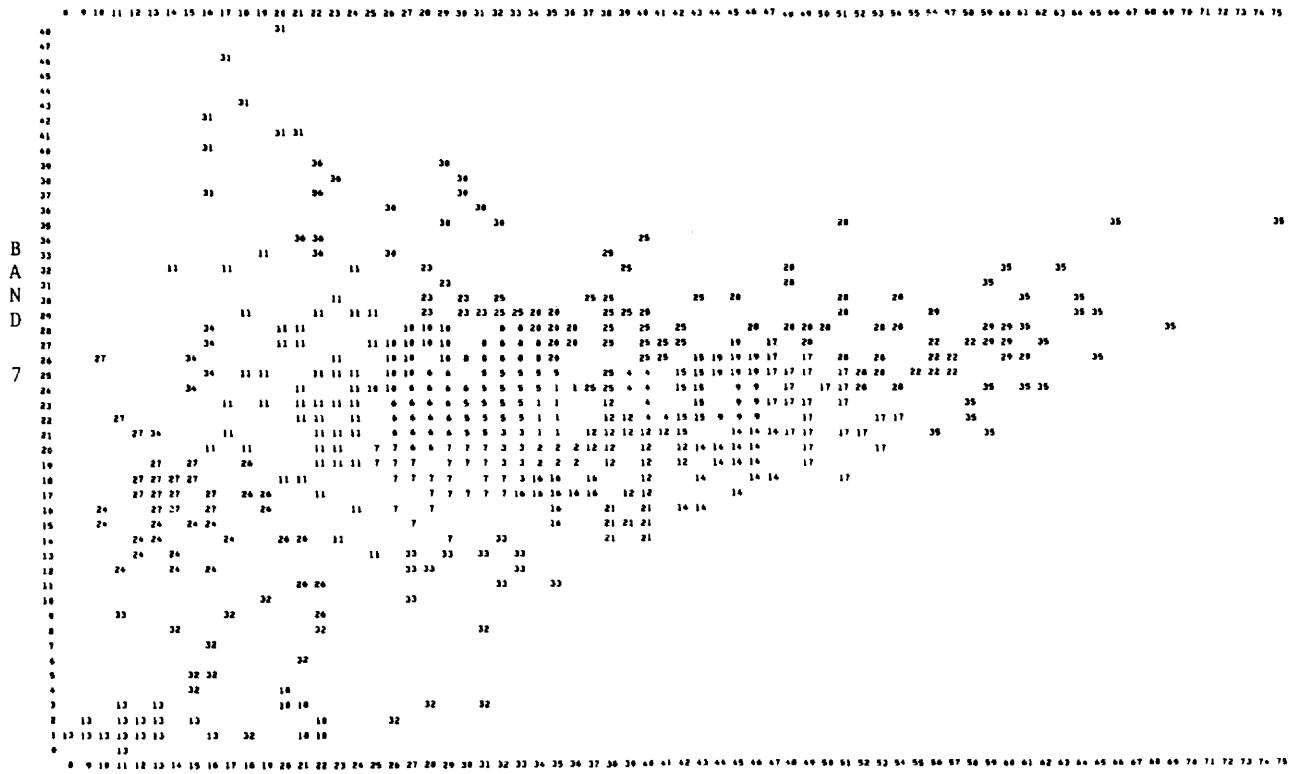
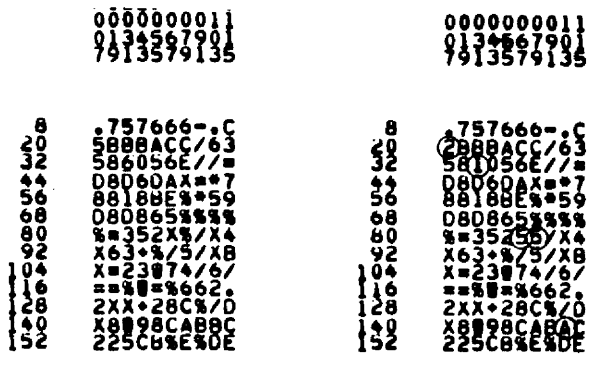


Fig. 4.9. Display of the resultant clusters in Run 3. The numbers are cluster labels. The numbering was made sequentially when it was formed.



(a) (b)

Fig. 4.10. Comparison of both cluster maps produced by (a) the hill-sliding algorithm and (b) the LMS classification, respectively. Five pixels (in circles) among 130 pixels in total are different between both cluster maps. The difference is due to that of assumptions regarding a priori probability (mixing proportion) in the mixed class/cluster distribution. Note that the cluster map by the LMS classification module was based on the cluster statistics obtained by the hill-sliding algorithm. The map data points are the one-twelfth by one-twelfth (1/144) sampled image pixels of Durand Quadrangle, Wisconsin. Interpretation of symbols are given in Table 4.10.

Table 4.9. Contingency table of cluster-classification displayed in Fig. 4.10. Classification was performed by the LMS maximum likelihood classifier using the cluster statistics obtained by the hill-sliding clustering program. Existence of off-diagonal numbers indicates that both methods employ different assumptions. The difference is due to that of assumptions regarding a priori probability (mixing proportion) in the mixture distribution of class/cluster. The numbers are those of picture elements in that category.

SYMBOL / CLASSIFIED AS	SUM
0 5	5
1 0	0
2 2	2
3 2	2
4 7	7
5 1	1
C - 6 3	5
L / 7 9	9
U X 8 1	9
S \ 9 12 1	15
T 1 A 1	1
E 2 B 7	7
R 3 C 4	4
E 4 D 2	2
D 5 E 1 8	9
6 1 11	12
7 4	4
8 10	10
9 2	2
A 0 1	1
S A 3	3
B 1 7	8
C 7	7
D 7	7
E 4	4
SUM	130

A line printer cluster map at a scale of 1:24,000 was reproduced for a small portion (4.18 square kilometers, 1.6 square miles) of the river basin (Figs. 4.11 and 4.12). The site, located north of Duran City, Wisconsin, includes a part of Chippewa River (northwest corner), north side of the city residential area, agricultural lands and naturally vegetated flood plain areas. Limited ground information was available for this study. One source was the Maps of Vegetation, Land, and Water Surface Conditions on the Upper Navigable Portion of the Mississippi River, 1973, prepared by the IAFHE Remote Sensing Laboratory University of Minnesota (Plate 1). All the land cover/land-use classes shown in the maps of the area were forbs, buttonbush, river (water), pond (water), sand, mud, agricultural land and developed area (urban). The other available information was black and white aerial photographs from May 4, 1978, provided by the Corps of Engineers at a scale of 1:24,000 (Plate 1). A broad category of land-use/land cover patterns was made based on these photos.

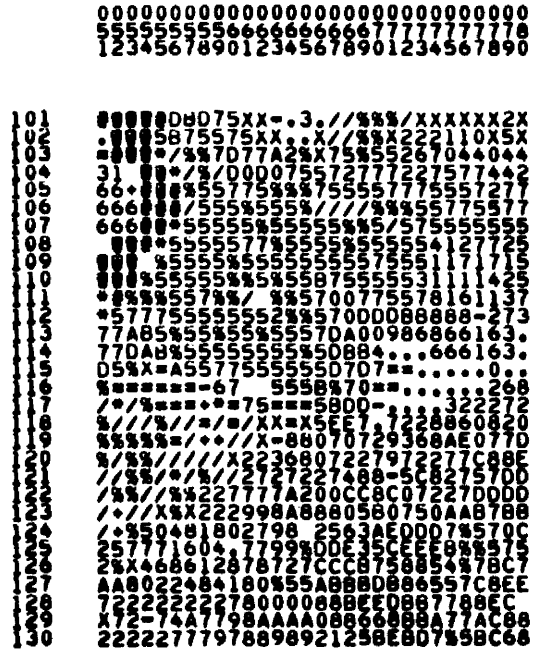


Fig. 4.11. Unsupervised classification map of land cover types near north of Durand, Wisconsin. The line printer map at the scale of 1:24,000 was produced by the LMS classification module based on the cluster statistics obtained by the hill-sliding clustering program using four MSS band data of LANDSAT imagery on May 11, 1976. Each symbol represents a cluster/class. Blanks were left over as unclassified if the pixels have lower probabilities to be associated with any cluster than a threshold value (5 percent in this case).

The cluster map shown here reveals several subgroups in land cover/land-use classes when compared with existing ground truth information. Each agricultural field (farm) might have different practice shown as spectrally different clusters. The water of the Chippewa River consists of two subgroups: deep water and shallow or near-bank water. Spectral characteristics of shallow or near-bank water depend

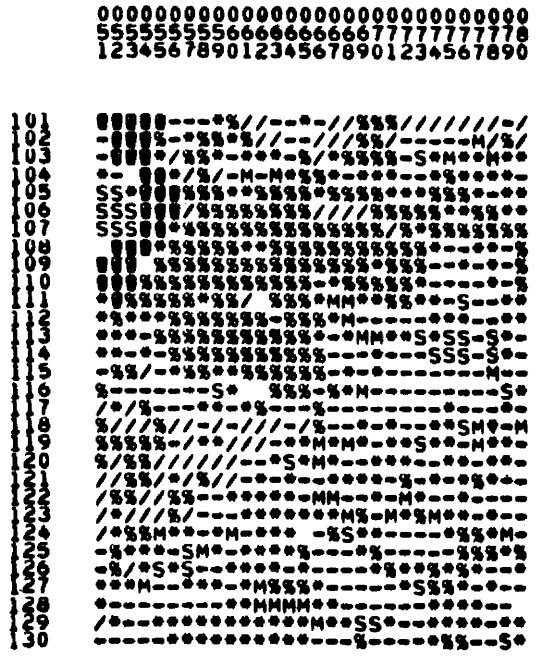


Fig. 4.12. Unsupervised classification map of land cover types having a unified symbol for all the subgroups of each class based on the result shown in Figure 4.11. Descriptions of symbols were given in Table 4.10.

upon the proportion of contribution by the river bottom or included lands in a picture element (pixel). Such a mixed picture element may have weak association with its parent classes/clusters and have lower probability of being classified as any existing clusters. The classification may display such a pixel as blank (Fig. 4.11).

A user may not need detailed subgroup maps in final products. Unified symbols for all the probable subgroups of individual land cover/land-use classes help in clearly visualizing spatial distributions of broad categorical classes (Fig. 4.12). This experimental map was produced by incorporating existing ground information. Identifying origin of subgroups is another difficult job in practice. Such a job is almost impossible without ground truth data. An association (or similarity) measure between clusters as used in hierarchical clustering techniques may indicate their mutual relationships but may not lead to user-oriented parent classes from lower level classes such as the second or third order land-use/land cover classification scheme (Table 4.2).

Another notable fact in this cluster map is that urban lands picked up almost every cluster in the area. A picture element of the LANDSAT has about one acre resolution and hence can pick up almost any combination of their complex constituents in residential or industrial areas. Such heterogeneous areas are often categorized as disturbed lands. It seems inadequate to distinguish individual pixels in such a disturbed land based on the resolution and precision of the LANDSAT multispectral scanner data.

Table 4.10. Aerial extents of land-cover type clusters displayed in Figs. 4.11 and 4.12. The interpretation was made based on black and white aerial photographs and some existing land-use/land cover classification maps for a portion of the study site.

Class	Class Symbol ①	Cluster Symbol ②	Assigned Number of Points ③	Percent
Water	W	W	19	2.11
Water	W	W	12	1.33
Disturbed Lands or Mixed Class	*	*	9	1.00
Disturbed Lands or Mixed Class	*	+	6	.67
Agricultural	-	=	23	2.56
Agricultural	-	-	7	.78
Agricultural	-	.	29	3.22
Buttonbush	/	/	46	5.11
Buttonbush	/	X	26	2.89
Forb	%	%	75	8.33
Agricultural	-	1	22	2.44
Agricultural	-	2	68	7.56
Disturbed Lands or Mixed Class	*	3	12	1.33
Distrubed Lands or Mixed Class	*	4	17	1.89
Forb	%	5	162	18.00
Sand	S	6	31	3.44
Disturbed Lands or Mixed Class	*	7	117	13.00
Disturbed Lands or Mixed Class	*	8	52	5.78
Disturbed Lands or Mixed Class	*	9	12	1.33
Mud	M	0	35	3.89
Disturbed Lands or Mixed Class	*	A	21	2.33
Agricultural	-	B	30	3.33
Agricultural	-	C	15	1.67
Agricultural	-	D	31	3.44
Agricultural	-	E	15	1.67
TOTAL POINTS			900	

① Symbols shown Fig. 4.5
② Symbols shown Fig. 4.4
③ 1 point (pixel or picture element) = 1.148 acres (4,646 m^2)

Plate 1. Black and white aerial photograph of the test site at a scale of 1:24,000 with overlay showing land-cover/land-use classes.

4.4 Comparison with the Results by an ISODATA Family Program

The ISODATA (Iterative Self-Organizing Data Analysis Technique) method was developed at Stanford Research Institute over a period of several years (Anderberg, 1973). It has been widely used for unsupervised classification of remote sensing data and ramified into various versions since it was first introduced by Ball and Hall (1965). The method searches iterative improvement of data partitioning following instructions given in terms of a set of heuristic parameters. Splitting and lumping parameters play major roles in the iterative process. Clusters having the maximum standard deviation greater than a threshold value are forced to split into two groups. Clusters will be combined if Euclidean distances between two cluster centroids are closer than the given lumping parameter. Each identity will be merged into the nearest cluster seeding point or existing cluster centroid in each iteration process. Other important parameters in the ISODATA family are the maximum number of clusters and the minimum number of data elements in a cluster. Most ISODATA family programs use Euclidean or city block distance which requires less computational steps. A version, called ISOCLAS (Senkus, 1976), has been installed as a computer library routine at Colorado State University. This version uses the city block distance measure, which is the sum of absolute differences between each band data, to assign each data point to a cluster. The use of this distance measure in the ISOCLAS is the major difference from that of the maximum likelihood decision rule in the hill-sliding algorithm.

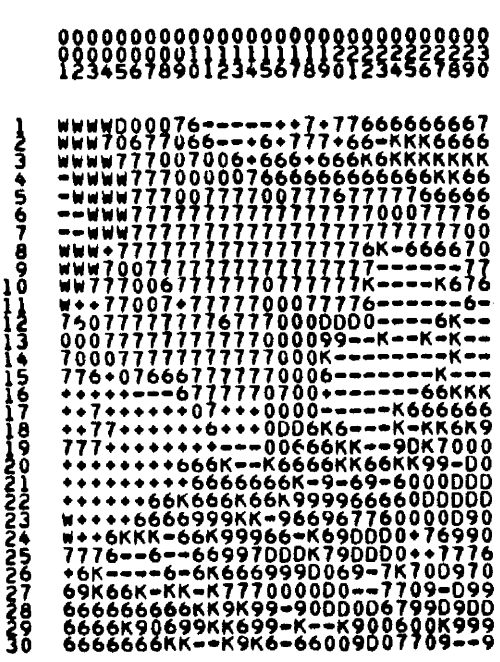
Performance of the hill-sliding program was compared with that of the ISOCLAS using the LANDSAT MSS band 5 and 7 data of the test area described in the previous section. The hill-sliding program produced 26 clusters after 3 initial and 3 additional iterations for refining clusters while the ISOCLAS seemed to repeat splitting and combining operations after 18th iteration with 10 clusters (Table 4.11). The results by the ISOCLAS failed to separate the sand and buttombush classes from some of the agricultural lands contrary to the hill-sliding program (Fig. 4.13). More additional mandatory parameter input may be required to extract these rare classes in the area. Further manipulation with varying parameter values was not attempted in this study because of the basic difference between two algorithms. The comparison of the results were provided to demonstrate the performances of both methods (Fig. 4.13). Spatial patterns of the dominant and distinct classes as water and forbs were fairly comparable in both maps.

Table 4.11. Comparison of performances of the hill-sliding and ISOCLAS programs. 900 LANDSAT pixels over a Chippewa River Basin area were used for clustering MSS bands 5 and 7 data.

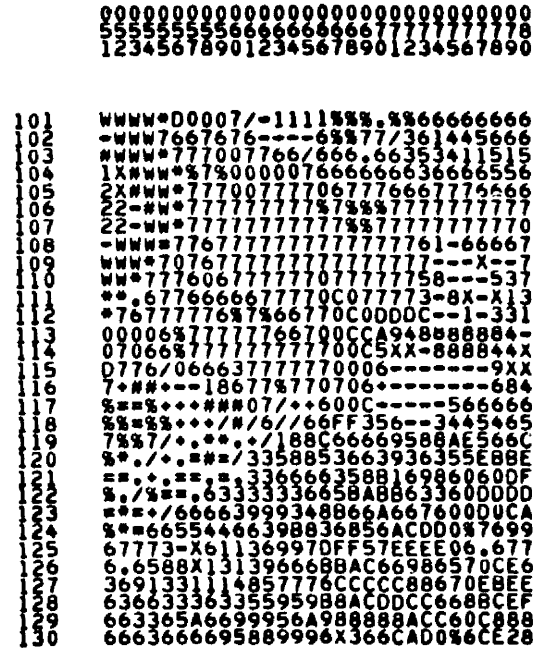
	Hill-sliding	ISOCLAS
Iteration	3 + 3 ①	20 ②
Number of clusters	26	10
CP time (second)	60 + 10	18

Note: ① Two consecutive runs with 6 iterations of refining clusters in total. ② After 18-th iteration splitting and combining seemed to repeat without improvement.

Definite judgment on both methods was not made due to insufficient ground truth data.



(a) ISOCLAS CLUSTER MAP



(b) HILL-SLIDING CLUSTER MAP

Fig. 4.13. Comparison of the cluster maps by both (a) ISOCLAS and (b) hill-sliding programs using MSS bands 5 and 7. The symbols in the ISOCLAS cluster map were matched with those in the hill-sliding program. The sand ("2" in (b)) and buttonbush ("=" in (b)) classes were not separated from agricultural classes ("-" and "+") in the map (a).

Chapter V
SUMMARY AND CONCLUSIONS

5.1 Summary

The primary objective of this study was to develop a practical technique for unsupervised classification of remote sensing data based on probability density estimates. The hill-sliding algorithm was developed for clustering normally distributed data incorporating with the maximum likelihood decision rule. The hill-sliding program has three major parts in implementing the clustering objective: 1) computation of lexicographic population distributions to effectively deal with discrete multivariate probability density estimates for wide ranges of data values; 2) extraction of initial clusters by the hill-sliding strategy; and 3) refinement of clusters by improving cluster compactness as well as optimizing the overall clustering objective function.

A "subcell model" program has been devised to alleviate multidimensional (multivariate or multi-indexing) problems in computation of the multivariate population densities. The storage requirement for density estimates is minimized by eliminating unused parts of the multidimensional feature space.

Initial clusters are extracted one by one according to the hill-sliding tactics. Separation of initial cluster elements from a set of data is the major framework of the clustering program. Further separation of cluster elements and their minor (say, cluster tail) refining process are carried out based on the newly proposed clustering function, which can be deduced from the maximum likelihood decision rule. Goodness of a cluster is measured by a dimensionless cluster compactness parameter.

The overall clustering objective function proposed is optimized with improvement of cluster compactness followed by repeated operations of splitting or abandoning the clusters which do not meet given constraints. A search for the globally optimal value of the objective function had not been devised in the present study. The result obtained by the clustering program can be reused for further improvement of the objective function criterion by applying a new set of constraints to the existing partition. The ultimate satisfaction of the results is up to the user or analyst.

Some other features of the hill-sliding algorithm are:

1) The expected number of clusters does not need to be specified.

2) It is suitable for an intermediate size of samples (say, about ten times the number of variates times number of expected clusters).

3) Mahalanobis distance classification option is provided, especially for a smaller set of data.

4) Input parameters are statistically rationalized values, which are not sensitive to the data structure.

Analysis of prototype class data over the Denver Metropolitan area showed promising results. Subcategorical information on known classes could be drawn through the analysis. The analysis also revealed that many urban disturbed class data of the LANDSAT multispectral scanners might not be suitable for classification of individual picture elements (pixels) by the hill-sliding algorithm or this type of cluster analysis.

The algorithm has been successfully applied to the Chippewa River Basin area to estimate aerial extents of land cover/land-use classes, i.e., by the unsupervised classification. Spatial distributions of vegetated and exposed land-type classes were confirmed satisfactorily based on very limited available ground information. However, pixel-by-pixel confirmation of urban land-use classes seemed not to have much meaning. Nevertheless, locating disturbed urban lands could be made by identifying the highly heterogeneous spatial distributions of various classes/clusters in confined areas.

Performance of the clustering program was compared with that of the ISOCLAS, a version of the ISODATA family program, for unsupervised classification of the LANDSAT data. The hill-sliding program yielded more detailed subcategorical clusters than those obtained by ISOCLAS in the simple runs, i.e., without much elaboration for adjusting heuristic input parameters. The dominant or well-separable classes were consistent with similar spatial distribution patterns in cluster maps produced by both programs. Relatively less populated classes appeared distinctively in the results by the hill-sliding program while those were mixed up with other classes in the ISOCLAS results.

The hill-sliding program used more computer central processing (CP) time than the ISOCLAS did. A main reason for this difference was that the former generated natural clusters more than twice those by the ISOCLAS. Each cluster is formed by itself based on its unimodal distribution characteristics without much interference by input parameters in the hill-sliding program. On the contrary, the ISOCLAS generates a cluster based on a set of mandatory given criterion parameters of distance measures, which force natural clusters either to be split or to be combined. Because of this difference, precise figures of CP times necessary to generate comparable results by both programs were not evaluated for this comparison.

5.2 Conclusions

The hill-sliding program developed herein has proved to effectively decompose multivariate mixture distributions of remote sensing data into a number of unimodal distributions, i.e., those of natural clusters. Inference of subcategorical structure on land-use cover classes can be drawn based on these natural groups of data.

Difficulties commonly encountered in computing and storing discrete multivariate probability densities were circumvented by utilizing the idea of lexicographic probability cells. Reduction of the computer memory storage requirement by this technique was significant in processing population distributions of LANDSAT multispectral scanner data.

The proposed dimensionless cluster compactness parameter has shown its universality as a measure of cluster goodness in various test runs. A merit of

this parameter is that it is less dependent on a varying number of dimensions or on wide ranges of data spread. Another advantage is the direct linkage to the overall clustering objective function.

A rationalized divergence measure between a pair of clusters was utilized successfully in the clustering program. This new measure is defined by the divergence (or general divergence which accounts for clusters mixing proportions) divided by the entropy of the entire sample data. The test runs demonstrated it has great promise as a general separability measure among clusters.

A new clustering function has been set forth in terms of individual cluster covariance matrices, from which a measure of cluster compactness can be deduced. Status of improvement in data partitioning can be evaluated solely by this function. An attempt to achieve the optimum partitioning was devised incorporating a set of user supplied constraints, which reflect the desired level of end products.

One of the drawbacks in this clustering program is that initial larger clusters (i.e., having larger determinant of cluster covariance matrix and/or larger a priori probability estimate) tend to grow faster than smaller ones. This is due to the maximum likelihood constraints, in which larger clusters yield larger values in density estimation of individual data point. It is often found that a smaller or compact cluster is located in the middle of a larger or loose cluster cloud. This situation may be natural and also acceptable. But a user may not find such a large cluster desirable. Further refining operations may break it down into two or more smaller clusters.

Another disadvantage of the present hill-sliding algorithm is that so-called "flat distributions" lead to production of an initial set of too many unnecessarily small clusters and consequently result in longer computational processing until the optimum partitioning is reached. The "flat distribution" is the term that describes where no hill-like population density estimate appears in the feature space. This occurs when too few data are used in comparison with dimensionality and discretized level.

5.3 Suggestions for Future Study

While the hill-sliding clustering program was developed, several areas of promise for further investigation emerged as follows.

- Use of the optimum discretization interval can avoid having flat distributions in significant parts of the data distribution space so that better performance of the hill-sliding program can be achieved. It is highly desirable to devise a way to estimate such an interval.

- Higher order statistics such as skewness (third central moment) and kurtosis (fourth central moment), should be utilized for rationale of splitting a cluster into two or more.

- Optimum critical values of cluster characteristic length and divergence should be found on rationale basis for their use as criterion parameters for splitting or discarding clusters.

- Devising systematic search for the global optimum solution to a clustering problem is highly desirable for use of the proposed objective function.

- Use of predetermined initial cluster centroids (or seeding points) might be worthwhile, especially when the number of samples is not large enough for a confident result. Inclusion of such an option is desired in the present program.

- Further scrutiny for efficient use of computer central memory and processing is recommended. Computer processing cost can be reduced by cutting down comparison or computation of insigificant parameters.

- Option for flexible sampling strategy would be helpful to investigators.

- Construction of better tabular form to display the results would make it easier for an analyst to evaluate cluster-class matching matrix.

There exist tradeoffs, however, between benefits and risks as a consequence of adding any new feature in the existing program. Such consequences should be taken into account.

REFERENCES

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

Anderson, J. R., E. E. Hardy, and J. T. Roach, 1972. A Land-Use Classification System for Use with Remote Sensor Data, Geological Survey Circular 671, U.S. Gov't Printing Office, Washington, D.C., 16 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.

Andrews, H. C., 1972. Introduction to Mathematical Techniques in Pattern Recognition, John Wiley & Sons, New York, 242 p.

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

Ball, G. H., 1965. Data Analysis in the Social Sciences: What about the Details? in Proceedings of the Fall Joint Computer Conferences, Stanford, Macmillan, New York, pp. 533-559.

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.

Blanchard, B. J., 1975. Remote Sensing Techniques for Prediction of Watershed Runoff. Proceedings of the NASA Earth Resources Survey Symposium, vol. I-D, Houston, Texas, pp. 2379-2406.

Bryan, J. K., 1971. Classification and Clustering Using Density Estimation, Ph.D. Dissertation, Univ. of Missouri, Columbia, Missouri.

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

Das Gupta, S., 1973. Theories and Methods in Classification: A Review, in Discriminant Analysis and Applications, T. Cacoullos, ed., Academic Press, New York, pp. 77-137.

Day, N. E., 1969. Estimating the Components of a Mixture of Normal Distributions, Biometrika, vol. 56, no. 3, pp. 463-474.

Dubes, R., and A. K. Jain, 1976. Clustering Techniques: the User's Dilemma, Pattern Recognition, vol. 8, pp. 247-260.

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.

Fehlauer, J., and B. A. Einstein, 1978. Structural Editing by a Point Density Function, IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-8, no. 5, pp. 262-270.

Fleming, M. D., J. S. Berkebile, and R. M. Hoffer, 1975. Computer-Aided Analysis of LANDSAT-1, MSS Data: A Comparison of Three Approaches, Including a "Modified Clustering" Approach, LARS Information Note 072475, Purdue Univ., West Lafayette, Indiana, 9 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.

Good, I. J., 1965. Categorization of Classification, in Mathematics and Computer Science in Biology and Medicine, H.M.S.O.

Haralick, R. M., and G. L. Kelly, 1969. Pattern Recognition with Measurement Space and Spatial Space for Multiple Images, Proceedings of the IEEE, vol. 57, no. 4, pp. 654-665.

Haralick, R. M., and J. Dinstein, 1975. A Spatial Clustering Procedure for Multi-image Data, IEEE Transactions on Circuits and Systems, vol. CAS-22, no. 5, pp. 440-450.

Hartigan, J. A., 1972. Direct Clustering of a Data Matrix, J. of the American Statistical Assoc., vol. 67, no. 337, pp. 123-129.

Hasselblad, V., 1966. Estimation of Parameters for a Mixture of Normal Distributions, Technometrics, vol. 8, pp. 431-444.

Kettig, R., 1975. Computer Classification of Remotely Sensed Multispectral Image Data by Extraction and Classification of Homogeneous Objects, LARS Information Note 050975, Purdue Univ., West Lafayette, Indiana, 184 p.

Khorram, S., 1976. Remote Sensing-Aided Information System Design for Water Resources Management, presented at the First Annual Earth Science Symposium, California Region, Forest Service, U.S.D.A. Region Five, 14 p.

Kittler, J., 1976. A Locally Sensitive Method for Cluster Analysis, Pattern Recognition, vol. 8, pp. 23-33.

Koontz, W. L. G., and K. Fukunaga, 1972. A Nonparametric Valley-Seeking Technique for Cluster Analysis, IEEE Transactions on Computers, vol. C-21, no. 2, pp. 171-178.

Landgrebe, D., 1976. Computer-Based Remote Sensing Technology - A Look to the Future, Remote Sensing of Environment, vol. 5, pp. 229-246.

Li, R. M., 1974. Mathematical Modeling of Response from Small Watershed, Ph.D. Dissertation, Civil Engineering Dept., Colorado State Univ., Fort Collins, Colorado, 212 p.

MacQueen, J., 1967. Some Methods for Classification and Analysis of Multivariate Data, Proceedings, Fifth Berkeley Symposium on Probability and Statistics, Univ. of California Press, Berkely, AD 669871, pp. 218-297.

Maxwell, E. L., 1975. Information Theory Applied to Remote Sensing, presented at the Fourth Annual Remote Sensing of Earth Resources Conference, Univ. of Tennessee, Tullahoma, Tennessee, March 24-26.

Maxwell, E. L., 1976. A Remote Rangeland Analysis System, J. of Range Management, vol. 29, no. 1, pp. 66-73.

Maxwell, E. L., T. C. Hart, R. L. Riggs, and L. D. Miller, 1977. Land Use Classification for Six Rocky Mountain States - Using LANDSAT Multispectral-Multitemporal Data, Final Report, prepared under Contract to the Federation of Rocky Mountain States under NASA Contract NAS-5-22338, Depts. of Earth Resources and Civil Engineering, Colorado State Univ., Fort Collins, Colorado.

Miller, L. D., K. Nualchawee, and C. Tom, 1978. Analysis of the Dynamics of Shifting Cultivation in the Tropical Forests of Northern Thailand Using Landscape Modeling and Classification of LANDSAT Imagery, NASA Technical Memorandum 79545, NASA/Goddard Space Flight Center, Greenbelt, Maryland, 19 p.

Miller, L. D., C. Tom, and K. Nualchawee, 1977. Remote Sensing Inputs to Landscape Models Which Predict Future Spatial Land Use Patterns for Hydrologic Models, NASA preprint X-023-77-115, Goddard Space Flight Center, Greenbelt, Maryland, 41 p.

Nagy, G., 1968. State of the Art in Pattern Recognition, Proceedings of the IEEE, vol. 56, no. 5, pp. 836-862.

Nagy, G., and J. Tolaba, 1972. Nonsupervised Crop Classification through Airborne Multispectral Observations, IBM J. Res. Develop., vol. 16, no. 2, pp. 138-153.

Nagy, G., 1972. Digital Image-Processing Activities in Remote Sensing for Earth Resources, Proceedings of the IEEE, vol. 60, no. 10, pp. 1177-1200.

Park, (John) K. Y., and L. D. Miller, 1978. Korean Coastal Water Depth/Sediment and Land Cover Mapping (1:24,000) by Computer Analysis of LANDSAT Imagery. NASA Technical Memorandum 79546, NASA/Goddard Space Flight Center, Greenbelt, Maryland, 21 p.

Regan, R. M., and T. J. Jackson, 1976. Hydrograph Synthesis Using LANDSAT Remote Sensing and the SCS Models, NASA-TM-X-71175: X-913-76-161 Preprint, NASA/Goddard Space Flight Center, Greenbelt, Maryland, 57 p.

Rogers, R. H., K. Peacock, and N. J. Sah, 1973. A Technique for Correcting ERTS Data for Solar and Atmospheric Effects, paper 1-7 presented at Third ERTS Symposium, NASA/GSFC, Washington, D.C., December 10-14, 18 p.

Senkus, W. M., 1976. ISOCLAS-User's Guide, version 1.1, Remote Sensing Research Program, Univ. of California, Berkeley, California, 30 p.

Simons, D. B., R. M. Li, and M. A. Stevens, 1975. Development of Models for Predicting Water and Sediment Routing and Yield from Storms on Small Watersheds, prepared for USDA Forest Service, Rocky Mountain Forest and Range Experiment Station, Civil Engineering Dept., Colorado State Univ., Fort Collins, Colorado, 130 p.

Simons, D. B., and Y. H. Chen, 1978. Investigation of the Effects of Chippewa River Erosion and Silt Reduction Measures, Phase IIA, prepared for U.S. Army Engineers District, St. Paul, Minnesota, Civil Engineering Dept. Colorado State Univ., Fort Collins, Colorado.

Sokal, R. R., and P. H. A. Sneath, 1963. Princples of Numerical Taxonomy, Freeman, San Francisco, California.

Stanat, D. F., 1968. Unsupervised Learning of Mixtures of Probability Functions, in Pattern Recognition, L. V. Kanal, ed., Thompson Book Company, Washington, D.C., pp. 357-389.

Sung, Q., and L. D. Miller, 1977. Land Use/Land Cover Mapping (1:25,000) of Taiwan, Republic of China by Augomated Multispectral Interpretations of LANDSAT Imagery, NASA preprint X-923-77-210, Goddard Space Flight Center, Greenbelt, Maryland, 168 p.

Tom, C., L. D. Miller, and J. W. Christenson, 1978. Spatial Land-Use Inventory Modeling and Projection/Denver Metropolitan Area, with Input from Existing Maps, Air photos and Landsat Imagery, NASA/Goddard Space Flight Center, Technical Paper, Greenbelt, Maryland, 212 p.

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

Wishart, D., 1969. Numeral Classification Method for Deriving Natural Classes, Nature (London), vol. 221, pp. 97-98.

Wolfe, J. H., 1965. A Computer Program for the Maximum Likelihood Analysis of Types, Technical Bulletin, 65-15, U.S. Naval Personnel Research Activity, San Diego.

Wolfe, J. H., 1967. NORMIX: Computation Methods for Estimating the Parameters of Multivariate Normal Mixtures of Distributions, Research Memorandum, SRM 69-17, Naval Personnel Research Activity, San Diego.

Wolfe, J. H., 1970. Pattern Clustering by Multivariate Mixture Analysis, Multivariate Behav. Res., vol. 5, pp. 329-350.

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

APPENDIX I

LANDSAT MAPPING SYSTEM (LMS)

The LANDSAT Mapping System (LMS) package has been developed at Colorado State University for specific use with both LANDSAT imagery inputs and composite mapping (Reference 1). The package is a total rewriting of the REGOCnition Mapping System or RECOG, which was designated principally for instructional purposes (References 2 through 4). This new software is compatible with the REGOC. Advantages cited for the LMS include flexibility in operation, exportability to other computers and high volume production.

The LMS software consists of four major image-processing phases. The first is the preprocessing phase, which prepares the standard data file RECOG to be used throughout all the phases, by inputting LANDSAT computer-compatible tapes (Fig. A1). The preprocessing phase contains modules of 1) conversion of the LANDSAT data into those in the RECOG format, 2) geometric rectification of the same data in a given scale for the line printer map, and 3) spatial filtering.

The second phase interleaves images from various dates and/or adds ancillary data to form a multivariate file of a specific map area (Fig. A2). The third phase is the computation of optimized statistical signatures of the materials to be mapped by classification (Fig. A3). The program in this module performs feature extraction, optimization of signature definition of prototype classes, and computation of statistical signatures. The fourth and final phase classifies each identity (or picture element) in desired mapping areas based on given class statistics by the maximum likelihood decision rule and then displays the classification maps via some visual media, such as microfilm or a line printer symbol map (Fig. A4).

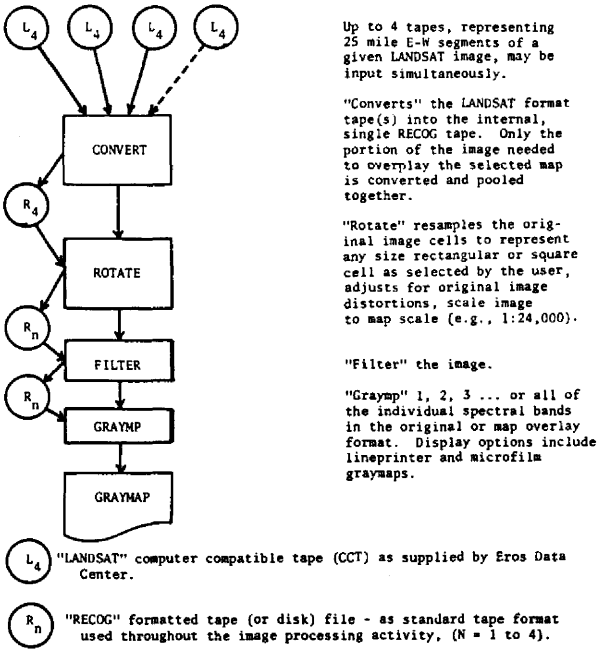


Figure A1. Phase 1 preprocesses the LANDSAT imagery.

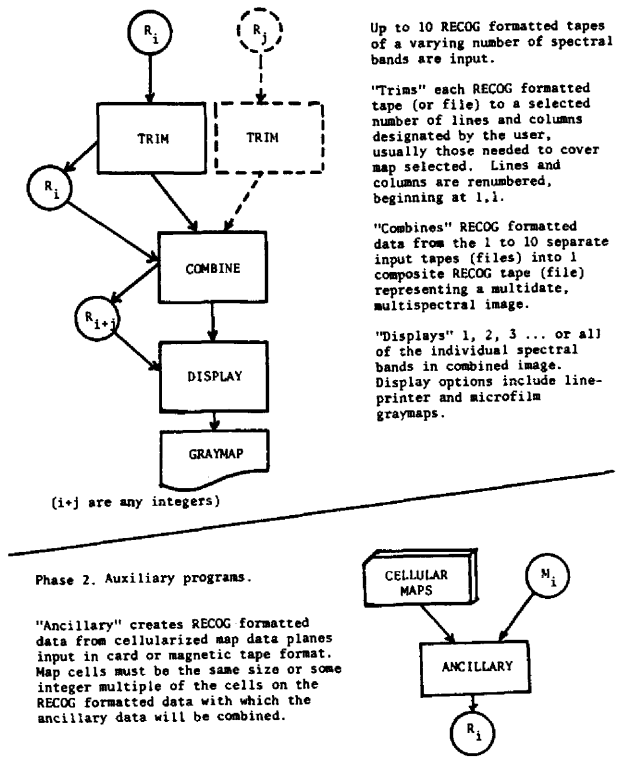


Figure A2. Phase 2 interleaves images from various data files.

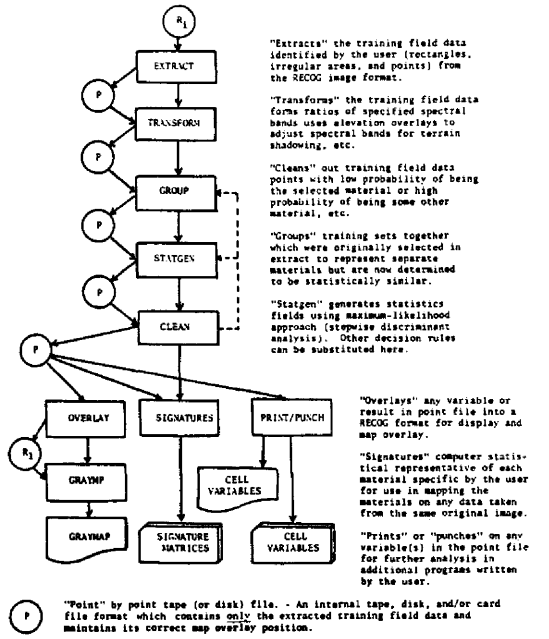


Figure A3. Phase 3 computes statistical signatures of materials to be mapped.

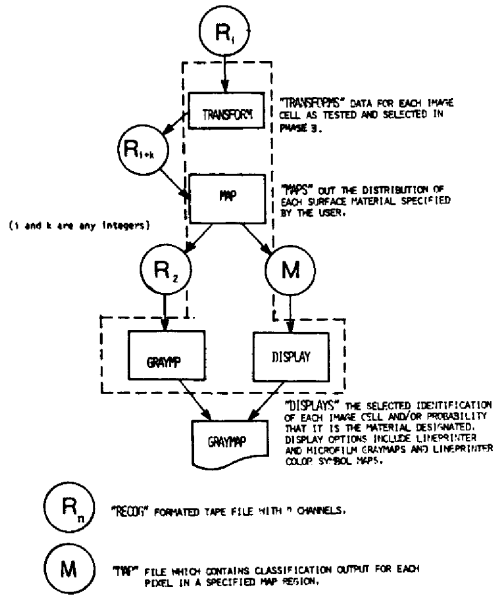


Figure A4. Phase 4 maps distribution of each material.

REFERENCES

1. Maxwell, E. L., T. C. Hart, R. L. Riggs, and L. D. Miller, 1977. Land Use Classification for Six Rocky Mountain States - Using LANDSAT Multispectral-Multitemporal Data, Final Report, prepared under Contract to the Federation of Rocky Mountain States under NASA Contract NAS-5-22338, Depts. of Earth Resources and Civil Engineering, Colorado State Univ., Fort Collins, Colorado.

2. Ells, T., L. D. Miller, and J. A. Smith, 1972a. User's Manual for RECOG (Pattern RECOGnition Programs). Science Series 3B, Dept. of Watershed Sciences, Colorado State Univ., Fort Collins, Colorado, 216 p.

3. Ells, T., L. D. Miller, and J. A. Smith, 1972b. Programmer's Manual for RECOG (Pattern RECOGnition Programs). Science Series 3C, Dept. of Watershed Sciences, Colorado State Univ., Fort Collins, Colorado, 216 p.

4. Smith, J. A., L. D. Miller, and T. Ells, 1972. Pattern Recognition Routines for Graduate Training in the Automatic Analysis of Remote Sensing Imagery-RECOG, Science Series 3A, Dept. of Watershed Sciences, Colorado State Univer., Fort Collins, Colorado, 80 p.

APPENDIX II

ISOCLAS

The ISOCLAS is a modified version of the ISODATA (Iterative Self-Organizing Data Analysis Technique) which was originally developed by Ball and Hall at Stanford Research Institute (Reference 1). Most later modifications in the ISOCLAS were made by Kan and his colleague (References 2 through 4) for use by NASA/ Manned Spacecraft Center, Houston, Texas. The new version is called ISOCLS (Iterative Self-Organizing Clustering Program) at Manned Space Center, while it is called ISOCLAS at the Remote Sensing Research Program, University of California, Berkeley, California (Reference 1).

The ISODATA family programs can be categorized as the nearest centroid sorting method. As the distance measured for sorting, the ISOCLAS employs the so-called city block distance:

$$\ell_i = \sum_{j=1}^{d} |x_j - \mu_{ij}|$$

where μ_{ij} is the j-th component of the estimated centroid of cluster i and other notations follow those of the main text. The method consists of the following steps (Fig. A5);

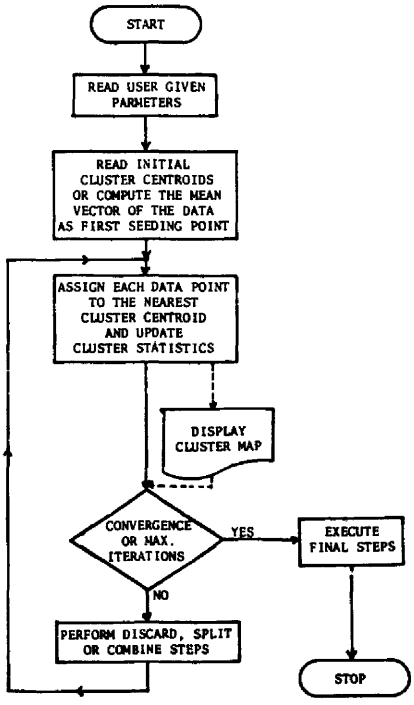


Figure A5. A general flow chart of the ISODATA family program.

1. Choose for key control parameters, such as
   - FEATURE(S): Coordinate(s) in a data vector to be analyzed.
   - ISTOP: Maximum number of iterations (default = 10).
   - LNCAT: Number of initial cluster (default = 1; initial cluster centroid must be supplied if LNCAT > 1).
   - STDMAX: Maximum standard deviation in a coordinate allowed before splitting a cluster (default = 3.0).
   - SEP: Distance to separate clusters upon splitting (default = maximum of coordinate standard deviations in cluster).
   - DLMIN: Minimum distance between clusters before combining (default = 3.2).
   - NMIN: Minimum number of members allowed in any cluster (default = 30).
   - MAXCLS: Maximum number of clusters (default = 50; MAXCLS $\leq$ 50).
2. Assign each data unit to the cluster whose centroid is the nearest to the data point and update cluster statistics.
3. Discard any cluster whose members are less than NMIM.
4. Perform either a splitting of a combining iteration according to the rules:
   a. Split any cluster i if $[\sigma_{jj}]_i$ > STDMAX for j=1,..., d. and I_c < MAXCLS where $[\sigma_{jj}]_i$ is the standard deviation of j-th component data of cluster i and I_c is the number of clusters. Then update cluster statistics.
   b. Combine any two clusters i and i' if
   $$\left[\sum_{j=1}^{d} \frac{1}{[\sigma_{jj}]_i [\sigma_{jj}]_{i'}} (\mu_{ij} - \mu_{i'j})^2\right]^{1/2} < \text{DLMIN}$$
   and then update cluster statistics.
5. Repeat steps 2, 3, and 4 until either the process converges or iterations reach ISTOP, or go the next step.
6. Execute final steps, such as chaining option and generation of punch file(s).

REFERENCES

1. Senkus, W. M., 1976. ISOCLAS - User's Guide, version 1.1, Remote Sensing Research Program, Univ. of California, Berkeley, California, 30 p.
2. Kan, E. P. F., and W. A. Holley, 1972a. More on Clustering Techniques with Final Recommendation on ISODATA, Lockhead Electronics Co., Inc., HASD, Houston, Texas, Tech. Rep. 640-TR-112 (May).
3. Kan, E. P. F., and W. A. Holley, 1972b. ISOCLS (ISODATA) Clustering: A Well Defined Problem, Lockhead Electronics Co., Inc., HASD, Houston, Texas, LEC/HASD No. 640-TR-152 (December).
4. Kan, E. P. F., 1973. The JSC Clustering Program ISOCLS and Its Applications, Lockhead Electronics Co., Inc., HASD, Houston, Texas, LEC-0483 (July).

APPENDIX III

GLOSSARY OF TERMS

a posteriori probability: Probability determined from measurements of the corresponding relative frequencies.

a priori probability: Probability of being a status (or cluster/class) given in advance. Term "mixing proportion" is often used as substitute in the mixture distribution.

algorithm: A statement of the steps to be followed in the solution of a problem.

band: A selection of wavelengths.

cell: An array of digitized elements in the feature space, hypercubic in shape.

centroid: The point whose coordinates are the mean values of the coordinates of the points in the set.

cluster compactness: A measure indicating that elements in a cluster are located closely together around the centroid (mean vector).

CCT (Computer Compatible Tape): A magnetic tape, containing data representing the image observed by the satellite. The data are arranged in a format which is directly readable by a computer.

commission accuracy: The proportion of the elements which originally came from the same class/cluster.

commission error: The proportion of the elements which originally came from other classes/clusters.

convergence: The act or condition of tending to one point or focus.

display: The graphic presentation of the output data of a device or system.

divergence: A measure of distance (or separability) between two class (or cluster) populations, defined by the sum of expectations of log-likelihood-ratios based on each distribution.

feature space: The space spanned by axes representing the measurement and/or transformed data.

Gaussian: A statistical term that refers to a normal distribution of values.

ground truth data: Term coined for data/information obtained on surface/subsurface features to aid in interpretation of remotely sensed data.

hypercube: A generalization of the concept of cube in three-dimensional Euclidean space to cube in d-dimensional space.

hypershell: A generalization of the concept of concentric shell in three-dimensional Euclidean space to shell in d-dimensional space.

hypervolume: A generalization of the concept of volume in three-dimensional Euclidean space to volume in d-dimensional Euclidean space.

imagery: The products of image-forming instruments (analogous to photography).

isotropic: Pertaining to a state in which a quantity or spatial derivatives thereof are independent of direction.

LANDSAT: Satellite(s) designed to make repetitive multispectral images of the earth's surface and relay data from remote automatic sensor stations at fixed locations on the ground, formerly known as the Earth Resources Technology Satellite (ERTS).

lexicographic probability cells: A chain of sequentially-arranged (or labeled) hypercubic compartment units with nonzero discrete multivariate probability density estimates.

mode: The most frequent value of a set of numbers or local maximum of the probability (or population) distribution.

module: A one-package assembly of functionally associated parts, usually a plug-in unit, so arranged as to function as a system or subsystem, or black box.

multispectral scanner: A remote sensing device which operates on the same principle as the infrared scanner except that is capable of recording data in the ultraviolet and visible portions of the spectrum as well as the infrared.

orientation: Direction or arrangement with respect to other detail.

parameter: A constant or variable in a mathematical expression, which distinguishes various specific cases and which may be assigned more or less arbitrary values for purposes of the problem at hand.

pattern: (1) In a photo image, the regularity and characteristic placement of tones or textures. (2) The relations between any parameters of a response.

pattern vector: Multidimensional quantity of measurements on various characteristics of a pattern. In the text, d-component column vector.

pixel: Discrete picture element.

rectification: The process of projecting a tilted or oblique photograph onto a horizontal reference plane.

remote sensing: The measurement or acquisition of information of some property of an object or phenomenon, by a recording device that is not in physical or intimate contact with the object or phenomenon under study.

resolution: The ability of an entire remote sensor system, including lens, antenna, display, exposure, processing, and other factors, to render a sharply defined image.

scatter matrix: Covariance matrix of the data.

scatterness volume: Term used in the text as a substitute of the determinant of a cluster/class covariance matrix.

signature: Any characteristic or series of characteristics by which a material may be recognized.

spectral band: An interval in the electromagnetic spectrum defined by two wavelengths, frequencies, or wave numbers.

topographic surface: The configuration of a surface including its relief and the position of its natural and man-made features.

Key Words: Cluster Analysis, Computers, Land Use, LANDSAT Imagery, Mapping, Natural Resources, Planning, Probability Density Estimates, Remote Sensing, Statistical Analysis

Abstract: A clustering algorithm with practical applicability to remotely sensed natural scene data was developed. A hill-sliding technique was devised to extract such natural cluster from the sample data. The computer memory storage required in processing population distributions of LANDSAT multispectral scanner data was significantly reduced by using this technique. The hill-sliding clustering program developed herein was applied to the Denver metropolitan area and the Chippewa River Basin areas in estimating the aerial extent of various land use/land cover classes. Performance of the hill-sliding clustering

Key Words: Cluster Analysis, Computers, Land Use, LANDSAT Imagery, Mapping, Natural Resources, Planning, Probability Density Estimates, Remote Sensing, Statistical Analysis

Abstract: A clustering algorithm with practical applicability to remotely sensed natural scene data was developed. A hill-sliding technique was devised to extract such natural cluster from the sample data. The computer memory storage required in processing population distributions of LANDSAT multispectral scanner data was significantly reduced by using this technique. The hill-sliding clustering program developed herein was applied to the Denver metropolitan area and the Chippewa River Basin areas in estimating the aerial extent of various land use/land cover classes. Performance of the hill-sliding clustering

Key Words: Cluster Analysis, Computers, Land Use, LANDSAT Imagery, Mapping, Natural Resources, Planning, Probability Density Estimates, Remote Sensing, Statistical Analysis

Abstract: A clustering algorithm with practical applicability to remotely sensed natural scene data was developed. A hill-sliding technique was devised to extract such natural cluster from the sample data. The computer memory storage required in processing population distributions of LANDSAT multispectral scanner data was significantly reduced by using this technique. The hill-sliding clustering program developed herein was applied to the Denver metropolitan area and the Chippewa River Basin areas in estimating the aerial extent of various land use/land cover classes. Performance of the hill-sliding clustering

Key Words: Cluster Analysis, Computers, Land Use, LANDSAT Imagery, Mapping, Natural Resources, Planning, Probability Density Estimates, Remote Sensing, Statistical Analysis

Abstract: A clustering algorithm with practical applicability to remotely sensed natural scene data was developed. A hill-sliding technique was devised to extract such natural cluster from the sample data. The computer memory storage required in processing population distributions of LANDSAT multispectral scanner data was significantly reduced by using this technique. The hill-sliding clustering program developed herein was applied to the Denver metropolitan area and the Chippewa River Basin areas in estimating the aerial extent of various land use/land cover classes. Performance of the hill-sliding clustering

program was compared with that of the ISOCLAS, a version of the ISODATA family program. The hill-sliding program employed less heuristic input parameters and yielded more reasonable partitioning of the sample data of the Chippewa River Basin than did the ISOCLAS.

Reference: Park, John Kyoungyoon; Chen, Yung Hai; and Simons, Daryl Baldwin; Colorado State University; Hydrology Paper No. 98 (September 1979); Cluster Analysis Based on Density Estimates and Its Application to LANDSAT Imagery.

program was compared with that of the ISOCLAS, a version of the ISODATA family program. The hill-sliding program employed less heuristic input parameters and yielded more reasonable partitioning of the sample data of the Chippewa River Basin than did the ISOCLAS.

Reference: Park, John Kyoungyoon; Chen, Yung Hai; and Simons, Daryl Baldwin; Colorado State University; Hydrology Paper No. 98 (September 1979); Cluster Analysis Based on Density Estimates and Its Application to LANDSAT Imagery.

program was compared with that of the ISOCLAS, a version of the ISODATA family program. The hill-sliding program employed less heuristic input parameters and yielded more reasonable partitioning of the sample data of the Chippewa River Basin than did the ISOCLAS.

Reference: Park, John Kyoungyoon; Chen, Yung Hai; and Simons, Daryl Baldwin; Colorado State University; Hydrology Paper No. 98 (September 1979); Cluster Analysis Based on Density Estimates and Its Application to LANDSAT Imagery.

program was compared with that of the ISOCLAS, a version of the ISODATA family program. The hill-sliding program employed less heuristic input parameters and yielded more reasonable partitioning of the sample data of the Chippewa River Basin than did the ISOCLAS.

Reference: Park, John Kyoungyoon; Chen, Yung Hai; and Simons, Daryl Baldwin; Colorado State University; Hydrology Paper No. 98 (September 1979); Cluster Analysis Based on Density Estimates and Its Application to LANDSAT Imagery.