NOTICE

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

NASA

Technical Memorandum 80732

ICAP: An Interactive Cluster Analysis Procedure for Analyzing Remotely Sensed Data

Stephen W. Wharton

(NASA-TM-80732) ICAP: AN INTERACTIVE CLUSTER ANALYSIS PROCEDURE FOR ANALYZING REMOTELY SENSED DATA (NASA) 27 p
HC A03/MF A01 CSCL 05B
N81-20494
Unclas
G3/43 31321

JULY 1980

National Aeronautics and
Space Administration

Goddard Space Flight Center
Greenbelt, Maryland 20771

SEP 1980
RECEIVED
NASA STI FACILITY
ACCESS DEPT.

TM 80732

ICAP: AN INTERACTIVE CLUSTER ANALYSIS PROCEDURE FOR ANALYZING REMOTELY SENSED DATA*

Stephen W. Wharton
Earth Resources Branch

July 1980

*This is a preprint of a manuscript submitted to Remote Sensing of Environment.

Goddard Space Flight Center, Greenbelt, Maryland 20771

ICAP: AN INTERACTIVE CLUSTER ANALYSIS PROCEDURES FOR ANALYZING REMOTELY SENSED DATA

I. INTRODUCTION

The LANDSAT Multispectral Scanner measures the intensity of radiation reflected by the earth's surface in four spectral bands at a ground resolution of approximately 80 m. Ground objects reflect radiation in a characteristic pattern of intensities, according to the object's physical properties. This pattern may be defined in terms of radiance means and a covariance matrix (i.e. training statistics) for a particular cover type. These statistics may then be used to train a classifier which recognizes patterns in a new environment by classifying the radiance data for each resolution element (pixel) into one of the pattern classes (cover types) under consideration. A thematic map can be produced to show the spatial distribution of the categories identified. Such maps can provide valuable information for use in mapping and monitoring natural resources.

Training statistics describing various land cover types can be developed using a supervised or unsupervised approach. Supervised methods involve the derivation of signature statistics from the analysis of picture elements within areas of spectral uniformity. These "training" areas must be located for each land cover category of interest. It may sometimes be difficult or impossible to specify a full list of the categories to be identified or to define training areas for all of the important features in a scene, especially for small, irregular or sparsely distributed features. Unsupervised methods such as cluster analysis can be used to estimate training statistics without the use of training areas and to map features in a scene without predetermining their identity.

The purpose of cluster analysis is to group data with a minimum of a priori knowledge. Since it is probable that universal objective clustering criterion exists (Fukunaga and Koontz, 1970), many different clustering approaches have been defined. Anderberg (1973) gives comprehensive coverage of the theoretical background and methodologies of cluster analysis. Hartigan (1975) presents program listings and describes various clustering and related algorithms. Dubes and Jain

(1976) tested and compared eight representative clustering programs and listed guidelines for program selection by potential users.

II. CLUSTERING METHODS USED IN REMOTE SENSING

Procedures used to cluster remotely sensed data can be divided into two groups based upon the methods used to control the clustering process. Those used by Turner (1972), Su and Cummings (1972), Kan et al. (1973), and the ISODATA algorithm as used by Zobrist (1976) require that the user manually specify various parameters to control the clustering process. These parameters are varied and the programs run in an iterative fashion until the output set of clusters meets the analyst's criteria.

Other procedures given by LeBoucher and Lowitz (1976), Borriello and Capozza (1974), Eigen et al. (1974), Fromm and Northouse (1976), and Goldberg and Shlien (1978) require a minimum of user input or determine the control parameters automatically from the data itself. This automatic group of procedures are most effective in producing an initial scene classification since the analyst is presumed to be unfamiliar with the scene and cannot intelligently select control parameters.

Most cluster analysis procedures used to process remotely sensed data invoke an iterative two step process. The first step deals with centroid location and cluster formation or growth. The information relevant to this initial step is quantitative since all of the entities to be manipulated are expressed numerically. A set of numerical rules are defined to regulate the formation of new centroids and to determine those data points which will be assigned to a given centroid. For example, the creation of new centroids can be controlled by defining a threshold distance from all existing centroids that a candidate point must exceed before becoming a new centroid. The minimum euclidian distance criterion can be used to determine the point membership of each centroid. A data point is assigned to the cluster whose centroid is nearest to that point in P space, where P is the dimensionality of the data set.

The second logical step within an iteration is the evaluation of the clusters produced by the first step. Once formed, clusters must be evaluated to determine if the present configuration is optimal or whether modifications are necessary. Most procedures define a fixed set of criteria by which clusters are evaluated and subsequently modified. For example, the ISODATA algorithm (Ball and Hall, 1965) is designed to split any cluster whose standard deviation exceeds a split threshold, delete any cluster with less than a specified number of members, and lump together cluster pairs whose centroids are less than a specified distance apart. The various thresholds are determined by the analyst.

A disadvantage of these indirect evaluation methods (indirect in the sense that the analyst manipulates parameters rather than the clusters) is that no one set of rules can be defined to cover all of the possible analytical objectives of the data analysis. In addition, the analyst cannot effectively extrapolate prior information about the category structure into the selection of control parameters. Consider a situation in which the objective is to map different types of forested areas, such as hardwood or conifers, within a scene. Ideally, the analyst could encourage the development of forest signatures by focusing attention on clusters whose centroids resemble typical forest responses and suppress clusters which appear to belong to irrelevant categories. Such a selective clustering process cannot be performed by existing procedures since the clusters are collectively evaluated according to fixed criteria.

III. The ICAP Algorithm

An Interactive Cluster Analysis Procedure (ICAP) was developed to avoid the inflexibility imposed by fixed cluster evaluation criteria, via a direct evaluation process in which each cluster is appraised and modified independently of the other clusters. ICAP combines the rapid numerical processing capacity of the computer with the human ability to integrate qualitative information to form a supervised clustering procedure. Control of the clustering process alternates between ICAP which examines data, locates new centroids and forms clusters; and the analyst who can request a cluster summary table and determine and execute the modifications, if any, to be made to the cluster configuration.

This shared control approach has two major advantages: ICAP does not have to optimize the cluster configuration, thus simplifying the program and reducing its execution time; effective use is made of subjective judgement since the analyst's judgement becomes an integral part of the clustering process, thus qualitative information can be used as a natural part of the analysis.

The methodology used in ICAP combines the concept of a cluster acceptance region (Muccairdi and Gose, 1972) with cluster manipulation techniques adopted from the ISODATA algorithm (Ball and Hall, 1965), and incorporates them into an interactive scheme. ICAP can be logically divided into three stages:

1. Data Preprocessing The data are examined and the overall distance threshold (ODT) is computed. The ODT is used to control the resolution (number and relative size) of the clusters to be produced in Supervised Clustering (SCLUS). If initial centroids are not specified, the mean of the scanned data is used as a starting centroid.

2. Supervised Clustering (SCLUS) Control of the clustering process alternates between ICAP, which scans the data, locates new centroids and forms clusters, and the analyst, who can evaluate and elect to modify the cluster structure. Thus, the analyst interacts with ICAP and controls the frequency of this interaction by specifying the maximum number of data points to be processed at once. The capability of modifying the cluster structure after processing arbitrarily sized segments of the data enables the analyst to closely supervise the clustering process. Clusters can be deleted, lumped together pairwise, or new centroids can be added. A summary of the cluster statistics can be requested to facilitate cluster manipulation.

3. Data Classification (DCLASS) The data are classified using centroids which remain fixed for a complete pass through the data. After each pass, new centroids are computed

to be the mean of their respective clusters. In addition to the modifications listed in SCLUS, the analyst can elect to split clusters.

A data set need only be preprocessed once. Stages 2 and 3 can be used to iteratively perform a global-local analysis similar to the approach proposed by Northouse et al. (1973). The methods of approach used in the three stages are described below.

Data Preprocessing

This stage locates the initial data centroid(s) and computes an overall distance threshold (ODT). The data are scanned and the sample mean, standard deviation, and maximum and minimum responses are computed for each of the P dimensions of the data. Upper and lower bounds are located on each dimension of the data to include the main concentration of data and to exclude outliers. These bounds are given by the dimension mean plus or minus 2.5 standard deviations. This interval should include approximately 99 percent of the data assuming they are normally distributed data. If either computed bound exceeds the actual range of the data, the appropriate bound is reset to be the actual maximum or minimum response. The volume (V) of the data is found by taking the product of the dimensional ranges.

ODT is a function of V, the approximate volume of the data space excluding outliers, and R, the user defined resolution or desired number of clusters to be examined in SCLUS (equation 1).

$$\mathrm{ODT} = \left(\frac{V}{R}\right)^{1/P} \qquad (1)$$

where P is the dimensionality of the data. Conceptually, ODT is the side length of a hypercubical cell selected such that V can be partitioned into R such cells. ODT is also equal to the minimum distance between the centers of neighboring hypersheres inscribed within the hypercubes. It is used in SCLUS to define the radius of a hyperspherical acceptance region which is centered about each centroid. All data points within an acceptance region are joined to the appropriate cluster. Data points outside all acceptance regions become the initial centroids for new clusters.

For uniformly distributed data, this scheme should allow approximately R clusters to be generated in SCLUS. It can be expected in practice, that more than R clusters will be produced, since outlier points would form additional clusters, and because the ODT is individually weighted for each cluster.

This procedure does not attempt to optimize the computation of the ODT beyond identifying reasonable ranges in each dimension, nor does it attempt to detect clusters which violate the assumptions made about the cell structure. The initial centroid(s) can be supplied by the analyst or the mean of the scanned points may be used. Figure 1 illustrates the above computations in a simple two dimensional case.

Supervised Classification (SCLUS)

SCLUS requires an overall distance threshold (ODT) and at least one initial centroid. These parameters can be supplied by the analyst if known a priori or can be determined by preprocessing the data. Hyperspherical acceptance regions are centered about the cluster centroids with radii equal to ODT times the local cluster density (described below) for each cluster. Each data point within a segment is examined in turn. If the point falls within the acceptance region of a centroid, it is grouped with that centroid. Otherwise, the point becomes a new centroid and immediately begins to accumulate its own points. This method of centroid determination tends to promote a fairly uniform distribution of centroids over the data space.

Cluster proliferation is encouraged in areas of relative low cluster density and inhibited in areas of high cluster density by weighting the ODT by the local cluster density. This selectively changes the acceptance region size. The local cluster density for the ith cluster is equal to the average distance between the ith centroid and all other centroids, divided by the average distance between all centroid pairs. This radio is greater than unity for regions with high cluster density and less than unity for low density regions.

After each data segment is processed, a listing can be requested to summarize the current cluster configuration. Statistics (see Table II) including the centroid locations, number of member points, index of the nearest and farthest centroid, distance to the nearest centroid, and the average distance to other centroids are given to help the analyst determine which modifications if any, are necessary. Based upon this evaluation, the analyst can elect to lump clusters together by pairs, delete clusters, add new centroids, or leave the configuration as is. Any modification of the cluster structure within an iteration makes it impossible to compute the cluster standard deviation. Since the standard deviation is used as a criterion for cluster splitting the option to split clusters is deferred to the DCLASS stage. The analyst may perform any combination of the above modifications as long as sufficient clusters remain to be manipulated. Additional summaries can be requested to aid this process. Upon completion of the modifications, control is returned to ICAP which then continues to process additional segments and alternate control with the analyst until all of the scene has been examined.

Data Classification (DCLASS)

DCLASS requires an input set of centroids and does not allow any change in the number of position of the centroids during one complete pass through the data. Cluster memberships are determined by the minimum euclidian distance rule, subject to the constraint that a point must be no further than DNC from its nearest centroid to be joined to that centroid's cluster. DNC is the distance from the centroid under consideration to its nearest neighboring centroid. This constraint prevents outlier data from being joined to inappropriate clusters. After each pass new centroids are computed to be the mean of their respective clusters. DCLASS can be run in an iterative fashion until the process converges; that is until there is no significant point reallocation among clusters between subsequent passes.

The standard deviation, ADG, and ADL are computed for each dimension of all clusters. ADG is the distance from the centroid to the mean of all points in the cluster greater than the centroid. ADL is the corresponding distance from all points less than the centroid. A cluster summary identical to that described in SCLUS and the cluster standard deviations are listed. The analyst can direct that certain clusters be split, based on the information provided. ICAP splits a cluster by first defining two new centroids which are identical to the original except in the dimension to be split. The values for this dimension are determined by adding the ADG and subtracting the ADL from the original centroid value. In addition to cluster splitting, the modifications detailed for SCLUS can also be performed.

Selection of R and SCLUS Segment Sizes

A goal of the analysis is the recognition and location of natural groups within the data. Depending upon the resolution factor R used in ICAP, a given natural group may be represented by several clusters, by one cluster, or it may share a cluster with other natural groups. In the second case, no corrective action is necessary. The error in the first case can be corrected by lumping clusters together, and the error in the third case can be corrected by splitting clusters.

A logical method of lumping clusters would be to join the pair with nearest centroids as determined from examination of the pairwise distances between all centroids. The number of computations required for this correction is a function of the number of clusters. Candidates for splits can be identified by reviewing the standard deviation for each dimension of all clusters. The number of computations is a function of the number of data points. Since the number of clusters is usually much less than the number of data points, the splitting operation uses more computer resources than the lumping operation. The need for splitting clusters can be largely eliminated in SCLAS by slecting R to be somewhat larger than the expected number of clusters. An R of 1.5 – 2.0 times the desired number of clusters was used in the ICAP tests reported in this paper.

The analyst controls the frequency of interaction within SCLUS by specifying that the image be processed by segments. The capability of examining and modifying the cluster structure at varying intervals within one pass of the data allows the analyst to moniter the formation of new centroids and subsequent cluster growth. The principal advantage of this approach is that unwanted clusters can be promptly eliminated. This improves the efficiency of the clustering process since the number of centroids to be examined is reduced.

The maximum rate of centroid proliferation can be expected during the initial stages of data processing. This rate should diminish as the number of existing centroids increases. To prevent the formation of two many centroids at once, the initial segments should be relatively small compared to the size of the data set (ie. the smaller of 500 points, or 5 percent of the data set size). The segment size should then be gradually increased during the latter stages of processing. Although the segment size selection is an arbitrary process, a rule of thumb can be given. Experience from testing ICAP has shown that 3 – 10 new centroids is a "comfortable" number to consider after segment processing. Let LSEG be the number of points processed in the last segment, and NCEN be the number of new centroids created. If NCEN is less than 3, the next segment size should be twice LSEG. If NCEN is greater than 10, the next segment size should be half LSEG.

IV. IMPLEMENTATION AND TESTING OF ICAP

The ICAP algorithm is designed to function in an interactive mode in which the analyst directly interacts with the computer, supplying input at the request of the program and receiving output as it is computed. The procedure is coded in APL (A Programming Language), which supports this interaction. APL, originally developed by Iverson (1962), is a concise and powerful language in which operations on single items (scalars) extend naturally to matrices of any size and shape. A large number of operators enable single APL instructions to perform operations requiring many statements in other languages. Single instructions can be combined into expressions that can be grouped into APL programs. This, lengthy procedures in other languages can often be succinctly expressed in APL with much fewer lines of code. The use of APL is described by Gilman and Rose (1976). ICAP was implemented on an IBM 370/3033 computer at the Pennsylvania State University, University Park, Pa. Various programs from a software system developed by the Office for

for the Remote Sensing of Earth Resources (ORSER) at the Pennsylvania State University (Turner et al, 1978) were used to evaluate ICAP's performance.

Two different Landsat scenes were used to test ICAP's clustering abilities. The first, in which the analyst was assumed to have no prior knowledge of the data, required an initial categorization type of analysis in which the clusters were formed more or less automatically with a minimum of user input. The second, in which the analyst was assumed to have partial knowledge of the important groups in the data employed a selective clustering type of analysis. Using this approach, the analyst focused attention and enhanced the development of clusters of interest and inhibited the development of clusters of little interest. The testing of the selective clustering approach is described in detail since it better illustrates the interactive use of ICAP.

A. Selective Clustering

The data used in this test are from an unpublished study by Turner (1978) which described the mapping of gypsy moth forest defoliation damage in central Pennsylvania using two merged scenes of Landsat imagery. The July 19, 1976 Landsat scene (data dimensions 5 to 8) had no defoliation. The June 19, 1977 scene (data dimensions 1 to 4) showed defoliation. The two scenes were geometrically corrected and registered to one another using the VICAR image processing program package at the NASA Goddard Space Flight Center, Greenbelt, Md. The test site included a mountain covered by hardwood forest, surrounded by agricultural lands. Since the goal of this analysis was to map canopy defoliation, the non-forest areas were not considered when developing training statistics or assessing classification accuracy It was known beforehand that hardwood forest vegetation at the test site had typical response of about 16, 14, 52, and 35 in Landsat bands, 4, 5, 6 and 7 respectively, on both dates.

The reference signatures for the accuracy comparison were developed using a supervised analysis. Training statistics were derived from training areas covering healthy, moderately and severely defoliated forest. These training areas were located through the use of the ORSER Uniformity Mapping Program UMAP, (Turner, et al. 1978) in conjunction with U-2 color aerial photography. Although no quantitative accuracy assessment was performed, the thematic map produced by classifying the scene with the reference signatures using the ORSER minimum euclidian distance classifier CLASS, (Turner, et al. 1978) appeared to correspond to the U-2 photography. A description of the analysis performed with the ICAP and CLUS programs is give: below.

ICAP Analysis

The data were first preprocessed to determine the overall distance threshold and to locate an initial centroid (Table I). It was believed that 4 to 6 categories were sufficient to map subclasses within the forest canopy category. A larger resolution factor of 10 was selected to reduce the potential for cluster splitting.

The SCLUS stage was used to locate an initial data partition. A cluster summary was requested after each segment was processed to determine what modifications might be necessary. Eight centroids were grown during the processing of the first segment which contained 500 points. The cluster summary is listed in Table II.

The forest clusters, recognized on the basis of a priori information, were always left unchanged. At this point, the major task of the analyst was to limit the number of non-forest clusters. This was done by lumping together similiar non-forest cluster pairs. For example, clusters 6-9 in Table II seemed to be forest clusters and were not altered. This similar non-forest clusters, pairs (1, 5) and (2, 3) were lumped together. Nine clusters remained after the last segment was processed. Seven of these belonged to the forest category. The other two clusters appeared to typify the non-forest categories response (believed to be agricultural lands) and were retained in the analysis. This was done to limit the proliferation of spurious non-forest clusters since non-forest responses would more likely be grouped with either or these two categories rather than cause new centroids to be created.

An additional pass through the data was made using DCLASS to refine the centroids produced in SCLUS (Table III). Clusters 6-9 appeared to be non-forest and the pairs (6, 8) and (7, 9) were lumped together. The three forest clusters, 1, 2, and 4, with the highest standard deviation were split in dimensions 7, 3, and 7, respectively, to form additional forest clusters. Another pass using DCLASS was made to refine the new centroids. The change in point allocation among clusters was judged to be minor and the ICAP clustering was terminated. The ICAP analysis took about 40 minutes of user time to complete and used 103 seconds of CPU time.

CLUS Analysis

The scene was also clustered with the ORSER CLUS program, using the default parameters described in the program documentation (Turner et al. 1978). It was necessary to run the program three times, adjusting the control parameters according to suggested guidelines in the documentation, until a satisfactory classification map was obtained. The CLUS analysis took about 10 minutes of user time to complete and used 66 seconds of CPU time.

Comparison of Results

The ORSER program CLASS was used to produce character classification maps for the reference, ICAP, and CLUS signatures. The performance of ICAP and CLUS was assessed by noting the number of pixels classified as being in agreement with the reference map. The ORSER program MAPCOMP (Turner, et al. 1978) was used to automate this comparison. The MAPCOMP program compares two character maps element by element and produces a comparison map and accompanying summary tables. Any differences in the number of categories between the test and reference maps were resolved by adjusting the symbols used to indicate a particular category. The severe and moderate defoliation categories were assigned unique mapping symbols. Other areas were ignored and mapped as blanks.

The test results (Tables IV and V) indicated that ICAP more accurately duplicated the reference map in locating the defoliation categories (70.7 versus 57.2 percent agreement for CLUS). Visual comparison of the test maps revealed that both ICAP and CLUS had difficulty in resolving the boundary between the severely and moderately defoliated categories.

B. Initial Categorization

A test procedure similar to the one described above was used to analyse data from part of a study by Merembeck (1978). He mapped forest cover and small openings in northwestern Pennsyivania using four channel Landsat data. The reference signatures for the larger homogeneous cover types were derived from training areas. Signatures for the smaller sparesely distributed cover types had been derived from the application of the ORSER CLUS program to the portions of the scene left unclassified by the supervised analysis. Merembeck devised a set of 34 signatures which he grouped into 13 categories. No accuracy assessment was performed. The goal of the test was to map as many of these categories as possible with ICAP and CLUS, and derive the best initial classification of the scene. The results of the unsupervised classification using ICAP and CLUS were compared to Merembeck's results.

It was known from visual examination of the Landsat imagery that portions of the scene were under considerable cloud cover. These areas were identified by their higher responses, typically above 45, 45, 45, and 30 in Landsat bands 4, 5, 6, and 7 respectively. These responses were considered to be noise and were ignored in the analysis. The test was made under the assumption that nothing was known about the cover type categories, other than a general familiarity with cover types in similiar regions of Pennsylvania.

It was believed that as many as 10 to 15 categories might be represented in the scene and a resolution (R) of 20 was selected. Since no specific a priori knowledge was assumed, the modifications performed in SCLUS were limited in scope to the reduction of noise (cloud) clusters. After an additional pass of the data was made with DCLASS, the ICAP clustering was terminated. The ICAP analysis took about 30 minutes of user time to complete, using 237 seconds of CPU time, and produced 7 spectral classes.

The scene was also clustered using the ORSER CLUS program, using the default parameters. An examination of the classification map revealed the the five clusters appeared to categorize the data into meaningful patterns and no further processing was done. The CLUS analysis took about 10 minutes of user time to complete and used 28 seconds of CPU time.

Comparison of Results

The ORSER program CLASS was again used to generate three classification maps for each set of signatures. The reference map was altered for comparison purposes by mapping similiar categories with the same mapping symbol. The ICAP and CLUS programs were compared (using MAPCOMP) with versions of the reference map altered to a resolution of seven and five categories, respectively.

The test results (Tables VI and VII) indicated that ICAP produced a higher resolution (seven versus five categories) and matched the reference map more accurately than CLUS (81.9 versus 70.7 percent agreement). Visual examination of the test comparison maps revealed that the major difference was that ICAP more accurately located the category boundaries, particularly in the Northwest Aspect Forest and Small Stream categories.

V. CONCLUSIONS

The general methodology used in cluster analysis and several of the techniques used in remote sensing applications have been reviewed. The existing algorithms for clustering remotely sensed data were considered to have limited flexibility, and cannot perform selective clustering since the clusters are evaluated collectively, thus preventing the analyst from effectively utilizing a priori knowledge about the data. A new procedure called ICAP was developed which allows the user to form clusters automatically or to interactively control the clustering process. Unlike existing procedures, this control is implemented by direct manipulations of the clusters themselves. No processing parameters are necessary. The flexibility of ICAP was evaluated using data from different Landsat scenes that represent two situations: one in which the user has limited prior knowledge about the category structure and wishes to have the clusters formed more or less automatically, and the other in which the user has a fairly complete knowledge about the existing categories in the data and wishes to use that information to closely supervise the clustering process.

For comparison, an existing clustering method CLUS by Turner (1972) was also applied to the same data sets. ICAP performed appreciably better than the CLUS program in matching the reference classification maps for the two test areas. For these scenes at least, the results indicate that ICAP is at least as good or better than the CLUS procedure in terms of accuracy. The results support the conclusion that the flexibility of ICAP can be effectively utilized to perform cluster analysis, regardless of the amount of a priori knowledge available.

The ICAP program used more CPU and analyst time than did the CLUS program in processing the test areas. It is difficult and perhaps unwise to draw general conclusions about the analyst time and CPU time required for the ICAP and CLUS analyses. The amount of CPU time used is dependent upon either the number of CLUS runs or the number of passes made through the data in ICAP. Both of these may vary widely for any given data set since the determination of a satisfactory result is largely subjective. However, it would appear that ICAP offers a more productive use of time since the user is always in direct contact with the clustering process. This supports a continuous learning process, unlike other procedures which function in a batch mode, in which the user must select control parameters and wait for results.

REFERENCES

Anderberg, M. R. 1973. Cluster Analysis for Applications. Academic Press, New York, N. Y.

Ball, G. H. and D. J. Hall. 1965. ISODATA, a novel method of data analysis and pattern classification. AD 699616. Stanford Res. Inst., Menlo Park, Cal. Cited in M. R. Anderberg. 1973. Cluster Analysis for Applications. Academic Press, New York, N. Y.

Borriello, L. and F. Capozza. 1974. A clustering algorithm for unsupervised crop identification. Proc. 9th Intl. Symp. on Remote Sensing of Environment, Vol. I. Env. Res. Inst. of Mich., Ann Arbor, Mich. pp. 181-188.

Dubes, R. and A. K. Jain. 1976. Clustering techniques: the user's dilemma. Pattern Recognition 8(4): 247-260.

Eigen, D. J., F. R. Fromm, and R. A. Northouse. 1974. Cluster analysis based on dimensional information with applications to feature selection and classification. IEEE Trans. Syst., Man, and Cybern. SMC-4(3): 284-294.

Fromm, R. F. and R. A. Northouse. 1976. CLASS: a nonparametric clustering algorithm. Pattern Recognition 8(3): 107-114.

Fukunaga, K. and W. L. G. Koontz. 1970. A criterion and an algorithm for grouping data. IEEE Trans. Computers C-19(7): 917-923.

Gilman, L. and A. J. Rose. 1976. APL, An Interactive Approach. 2nd Ed. John Wiley and Sons, Inc., New York, N. Y.

Goldberg, M. and S. Shlien. 1978. A clustering scheme for multispectral images. IEEE Trans. Syst., Man, and Cybern. SMC-8(2): 86-92.

Hartigan, J. A. 1975. Clustering Algorithms. John Wiley and Sons, Inc., New York, N. Y.

Iverson, K. E. 1962. A Programming Language. John Wiley and Sons, Inc., New York, N. Y.

Kan, E. P., W. A. Holley, and H. D. Parker. 1973. The JSC clustering program ISOCLS and its applications. 1973 IEEE Conf. on Machine Processing of Remotely Sensed Data, The Laboratory for Applications of Remote Sensing, Purdue Univ., West Lafayette, Ind. pp. 4B-36 to 4B-45.

Leboucher, G. and G. E. Lowitz. 1976. What can a histogram really tell the classifier? IEEE Intl. Joint Conf. on Pattern Recognition, Coronado, Cal. pp. 689-695.

Merembeck, B. F. 1978. Small-Area Mapping and Spectral Signature Extension in Pennsylvania Hardwood Forests Using Landsat-1 Multispectral Scanner Data. M.S. thesis, the Pennsylvania State Univ., University Park, Pa.

Muccairdi, A. N. and E. E. Gose. 1972. An automatic clustering algorithm and its properties in high dimensional spaces. IEEE Trans. Syst., Man, and Cybern. SMC-2(2): 247-254.

Su, M. Y. and R. E. Cummings. 1972. An unsupervised classification technique for multispectral remote sensing data. Proc. 8th Intl. Symp. on Remote Sensing of Environment, Vol. II. Env. Res. Inst. of Mich., Ann Arbor, Mich. pp. 861-879.

Turner, B. J. 1972. Cluster analysis of multispectral scanner remote sensor data. Pages 538-549 in F. Shahrokhi, Ed., Remote Sensing of Earth Resources, Vol. I. Space Inst., Univ. Tennessee, Tullahoma, Tenn.

______. 1978. Personal communication. Assoc. Prof. For. Mgmt., School of Forest Resources, The Pennsylvania State Univ., University Park, Pa.

______, D. N. Applegate, and B. F. Merembeck. 1978. Satellite and Aircraft Multispectral Scanner Digital Data User Manual. ORSER Tech. Rept. 9-78, The Pennsylvania State Univ., University Park, Pa.

Zobrist, G. W. 1976. On clustering of ERTS data to develop a character map. IEEE Milwaukee Symp. on Automatic Computation and Control, Milwaukee, Wis. pp. 205-210.

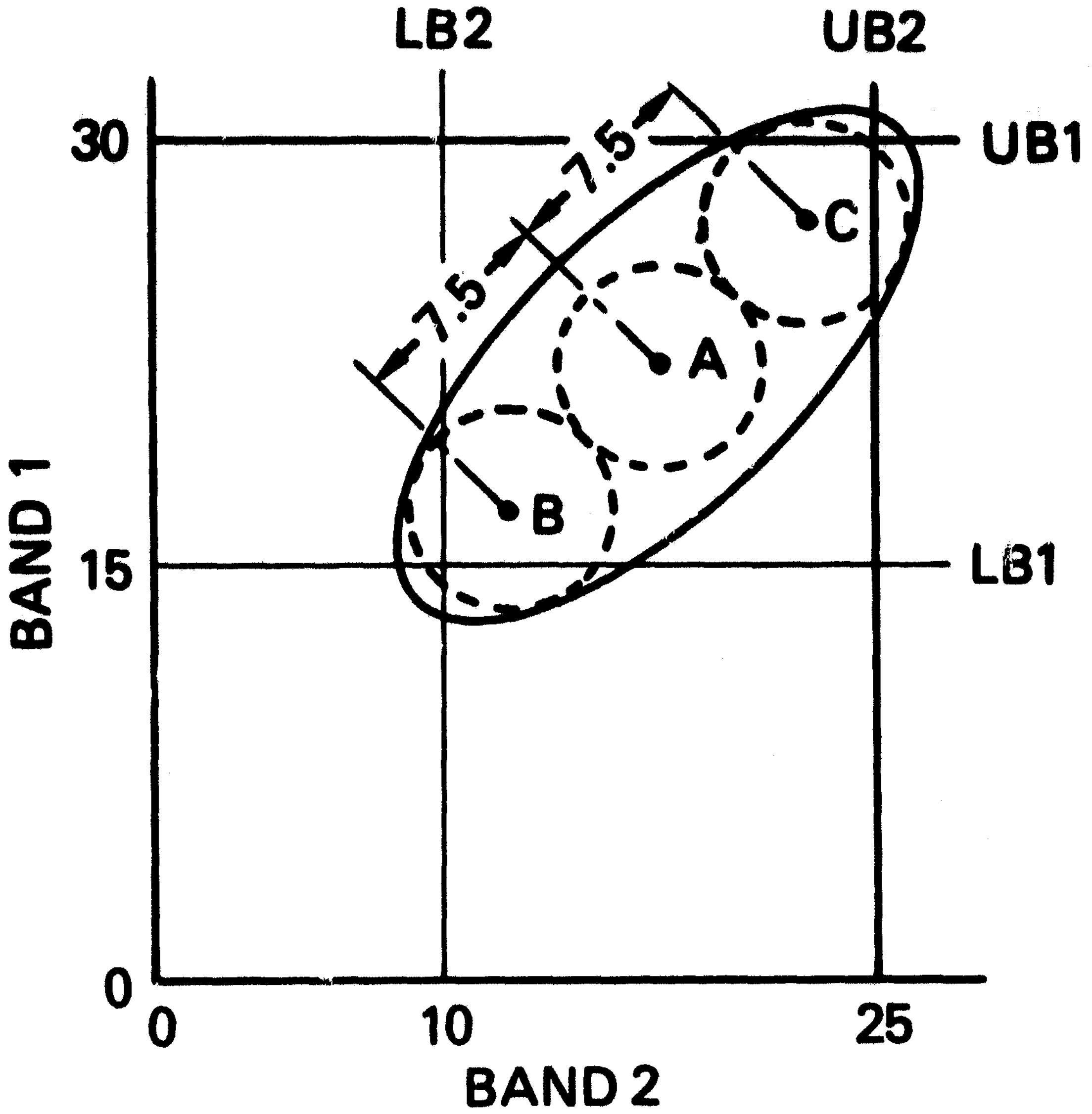

Figure I. Calculation of ODT in a two dimensional case.

$$
\begin{aligned}
V &= (UB2 - LB2) \cdot (UB1 - LB1) \\
&= (25 - 10) \cdot (30 - 15) \\
&= 225
\end{aligned}
\qquad\qquad
\begin{aligned}
ODT &= \left(\frac{V}{R}\right)^{1/P} \\
&= \\
&= \left(\frac{225}{4}\right)^{1/2} \\
&= 7.5
\end{aligned}
$$

With the elliptical data distribution shown with mean at A, new centroids would be grown at B and C.

Table I. Statistics from preprocessing the data.

	Dimensions							
	1	2	3	4	5	6	7	8
Mean	18.1	17.2	53.3	28.0	17.8	15.2	58.9	32.9
Standard deviation	2.2	4.0	7.5	5.2	2.8	4.1	4.1	2.7
Minimum	14.0	12.0	35.0	16.0	15.0	11.0	35.0	14.0
Maximum	31.0	36.0	73.0	42.0	34.0	39.0	78.0	42.0

Table II. Cluster summary after processing the first segment.[a]

							Dimensions							
#	CCNT	CW	DNC	ADOC	NC	FC	1	2	3	4	5	6	7	8
1	201	202	8.2	12.0	5	9	17.6	15.9	54.6	29.0	18.1	15.7	58.5	32.2
2	66	67	8.1	16.5	3	9	17.3	15.0	62.1	34.1	17.7	15.1	63.4	35.6
3	97	98	8.1	15.4	2	9	19.0	17.6	61.9	32.4	21.7	20.5	62.9	32.7
4	29	30	10.0	15.0	3	9	20.6	21.3	55.9	27.8	22.8	23.3	59.9	29.9
5	14	15	8.2	14.7	1	9	17.1	13.9	58.4	31.5	19.2	17.7	53.8	28.2
6	3	3	11.9	17.1	7	2	16.0	13.7	45.0	23.3	21.0	22.3	56.3	28.3
7	20	21	7.6	16.4	9	3	17.3	16.3	47.3	24.1	16.6	13.0	52.0	29.2
8	11	12	11.1	16.9	2	9	17.2	14.5	54.8	29.8	19.8	17.9	69.7	36.9
9	52	53	7.6	19.1	7	2	18.2	19.0	42.4	20.8	15.8	13.2	55.2	31.3

[a] CCNT = cluster count; CW = cluster weight; DNC = distance to nearest cluster; ADOC = average distance to other clusters; NC = nearest cluster; FC = farthest cluster.

Table III. Cluster summary after the cluster statistics pass.[a]

							Dimensions							
#	CCNT	CW	DNG	ADOC	NC	FC	1	2	3	4	5	6	7	8
1	786	786	8.4	14.7	5	6	17.3	15.8	49.6	25.8	16.3	13.1	56.2	32.1
2	801	801	10.9	20.5	1	6	18.6	20.8	41.6	20.4	16.3	13.0	56.7	32.4
3	833	833	7.4	14.4	7	6	16.9	14.3	56.6	31.1	16.7	13.5	58.0	33.4
4	642	642	9.3	19.1	3	6	16.9	14.1	62.7	35.1	16.9	13.7	62.9	36.4
5	51	51	8.3	16.9	1	6	17.1	14.3	50.7	26.6	17.6	15.1	50.3	27.0
6	43	43	10.4	22.4	8	2	23.1	23.9	57.2	27.4	28.4	32.0	60.8	28.1
7	142	142	5.6	13.0	9	2	18.5	17.2	57.0	30.2	19.9	18.0	61.6	33.1
8	128	128	9.0	16.0	9	4	21.8	23.2	52.7	25.4	23.6	24.7	58.8	29.0
9	153	153	5.6	13.8	7	2	20.5	20.4	57.5	29.0	22.1	21.0	62.0	32.1

[a] CCNT = cluster count; CW = cluster weight; DNC = distance to nearest cluster; ADOC = average distance to other clusters; NC = nearest cluster; FC = farthest cluster.

Table IV. ICAP confusion table indicating percentage agreement and disagreement between categories identifed by ICAP and similar categories using the reference map signatures.

Reference Categories	ICAP Categories			
	Moderate	Severe	Other	Total
Moderate	30.0	4.0	16.5	50.5
Severe	0.5	40.7	8.4	49.6
Total	30.5	44.7	24.9	

Total percentage agreement = 70.7

Table V. CLUS confusion table indicating percentage agreement and disagreement between identified by CLUS and similar categories using the reference map signatures.

Reference Categories	CLUS Categories			
	Moderate	Severe	Other	Total
Moderate	30.5	0.0	20.0	50.5
Severe	7.6	26.7	15.2	49.5
Total	38.1	26.7	35.2	

Total percentage agreement = 57.2

Table VI. ICAP confusion table indicating percentage agreement and disagreement between categories identified by ICAP and similar categories using the reference map signatures.

Reference Categories	ICAP Categories								
	NW	SE	Open	Water	Creek	HSE2	Edge	Other	Total
NW	45.6	0.3	0.0	0.0	0.0	0.0	0.0	0.0	45.9
SE	0.0	9.8	0.0	0.0	0.0	0.0	0.0	0.6	10.4
Open	1.4	0.6	0.3	0.0	0.6	0.0	0.8	2.2	5.9
Water	0.0	0.0	0.0	3.5	0.0	0.0	0.0	0.0	3.5
Creek	6.7	0.0	0.0	0.0	13.7	0.0	0.0	0.0	20.4
HSE2	0.0	0.0	0.0	0.0	0.0	0.3	0.0	1.2	1.5
Edge	0.0	0.2	0.0	0.0	0.0	0.0	0.6	0.5	1.3
Other	0.5	0.1	0.3	0.2	1.3	0.5	0.2	8.1	11.2
Total	54.2	11.0	0.6	3.7	15.6	0.8	1.6	12.6	

Total percentage agreement = 81.9

Table VII. CLUS confusion table indicating percentage agreement and disagreement between categories identified by CLUS and similar categories using the reference map signatures.

	CLUS Categories						
Reference Categories	NW	SE	Water	Open	Creek	Other	Total
NW	31.1	12.0	0.0	0.0	2.8	0.0	45.9
SE	0.0	8.8	0.0	0.0	0.0	1.7	10.5
Water	0.0	0.0	3.5	0.0	0.0	0.0	3.5
Open	0.6	1.1	0.0	0.6	0.7	4.0	7.0
Creek	0.1	0.0	0.0	3.3	17.0	0.0	20.4
Other	0.2	0.4	0.1	2.1	0.0	9.9	12.7
Total	32.0	22.3	3.6	6.0	20.5	15.6	

Total percentage agreement = 70.7