Reprinted from

Seventh International Symposium

Machine Processing of

Remotely Sensed Data

with special emphasis on

Range, Forest and Wetlands Assessment

June 23 - 26, 1981

Proceedings

Purdue University
The Laboratory for Applications of Remote Sensing
West Lafayette, Indiana 47907 USA

AN UNSUPERVISED CLASSIFICATION APPROACH FOR ANALYSIS OF LANDSAT DATA TO MONITOR LAND RECLAMATION IN BELMONT COUNTY, OHIO

JAMES O. BRUMFIELD

Marshall University
Huntington, West Virginia

HUBERTUS H.L. BLOEMER

Ohio University
Athens, Ohio

WILLIAM J. CAMPBELL

NASA/Goddard Space Flight Center
Greenbelt, Maryland

I. ABSTRACT

Surface mining for coal is a major economic activity in east central Ohio. Ohio has strict reclamation laws which require that the mining companies return the mined land to environmentally acceptable conditions. During the decade of the seventies, a particularly conscious effort has been made in Ohio to enforce the reclamation laws. Monitoring the reclamation efforts and progress via traditional means is time consuming, expensive, and often subjective. LANDSAT multispectral data provides a means to eliminate some of the negative aspects of the above.

A nontraditional unsupervised classification procedure has been devised using a clustering algorithm with a NASA modification of the canonical analysis algorithm as implemented on the Pennsylvania State University ORSER system. The algorithms are implemented on the ERRSAC IDIMS/HP 3000 at NASA/Goddard Space Flight Center in Greenbelt, Md. for use in the unsupervised classification approaches. A standard unsupervised clustering/maximum likelihood algorithm sequence is compared to a nontraditional unsupervised clustering/canonical transformation/clustering algorithm sequence in delineation of land cover categories in surface mining areas. This nontraditional unsupervised classification approach demonstrates appreciable improvement in spectral category groupings when compared to the traditional unsupervised classification approach and land cover information.

II. INTRODUCTION

A methodology was developed to demonstrate and compare a traditional unsupervised clustering/maximum likelihood algorithm sequence and a non-traditional unsupervised clustering/canonical transformation/clustering algorithm sequence applied to monitor land reclamation. These algorithm sequences are applied to LANDSAT data of a surface mining area in east central Ohio. East central Ohio, particularly Belmont county (figures 1 & 2) is nearly synonymous with surface mining. Belmont County has the highest percentage of land churned over in Ohio since the introduction of surface mining in the county in 1918.

There are several seams of coal throughout Belmont County. Although some seams are extracted through deep mining most of the coal is surface mined. Most of that is done via the contour method (figure 3) as opposed to area surface mining.

In the contour method, a bench is cut into the coal seam that crops out along a valley. Earth removal continues into and around the hill until the overburden removal costs make the amount of coal gained no longer economically feasible. After the coal mining procedures are terminated, a surface mined area is traditionally characterized by a steep uncut face, the highwalls, a relatively flat bench that follows the contour of the hill, and an adjacent spoil pile consisting of previously removed overburden. Lakes or ponds form frequently on the bench between the spoil and the highwall. Traditionally, the monitoring has been accomplished by aerial and ground survey at considerable cost, time, and historically sparse infrequent coverage. LANDSAT data can be used for monitoring and techniques have been established by many researchers. (Rogers et al., 1974; Anderson et al., 1977; Russell, 1977; Spisz and Dooley, 1980; Irons et al., 1980; Middleton and Bly, 1981). A synoptic overview of the literature can be found in Bloemer et al., 1981.

With the assistance of ERRSAC at NASA/Goddard Space Flight Center and their computerized Interactive Digital Image Manipulation System (IDIMS)/HP3000, the authors utilized LANDSAT data to monitor the reclamation of such areas. However, the authors' approach used feature extraction techniques for classification procedures. These feature extraction techniques apparently increase the level of accuracy in classification of land cover categories as determined by comparison of data collected from the corresponding locations on the ground.

III. DISCUSSION OF METHOD

Two unsupervised approaches of classification are compared for agreement with data collected from the corresponding locations on the ground. These two proce-

dures are 1) the traditional clustering/ maximum likelihood algorithm sequence, which assumes spectral groupings in the LANDSAT data in n dimensional spectral space and 2) a nontraditional approach which also looks at the spectral groupings in the data swarm in n dimensional space. The latter method includes an additional feature extraction technique involving canonical analysis which appears to provide an apparent advantage in information extraction not available in the traditional approach. The canonical transformation translates, rotates, and rescales the data based on the within cluster and among clusters group variability. The among group variability hierarchically diminishes along each additional transformed axis, which is orthogonal in the previously developed axis. This results in maximizing among clusters separability while reducing the dimensionality of the data for the classification procedure. Commonly, 90% to 98% plus of the variability in the data can be accounted for the first two transformed axes (Merembeck, 1977; Lachowski and Borden, 1973).

There appears to be an advantage of the canonical transformation over the Karhunen-Loeve transformation (Merembeck, 1977). The advantage derives from the rescaling of the data along each of the orthogonal axes to minimize the within cluster variance to unity. This rescaling factor in the transformation is, of course, applied to the entire data swarm in n dimension spectral space. Therefore, it seems reasonable to assume that while the rescaling may increase the variance of some clusters in the data swarm along some axes, the nature of the transformation would be to decrease the within cluster variability of most of the clusters along the transformed axes. This should particularly be the case in which the within cluster variability is maximum along those axes of greatest among clusters variability. In the application of the cluster algorithm to the original data, the sparsity of the data in those localized regions of the n dimensional space may have resulted in fewer clusters being placed in those regions than the number of categories represented there. The only solution to the dilemma is to force the clustering algorithm to delineate a large number of very subtle clusters. This would have apparently resulted in very subtle clusters in the data with larger within cluster variability not having been expressed in the application of the original cluster algorithm. The effect of the transformation rescaling is to increase the density within the localized regions of interest by developing the axes of maximum among clus-

ter variability. At the same time each axis is rescaled so that the within cluster variability is minimized. The net effect is to make possible the identification of new clusterings of the data with apparently more direct correspondence to ground cover of categories and lesser within cluster variability formed in the data swarm. The effect of this change is that the second application of the clustering algorithm (ISOCLS, Idims Ref. Man., 1978) now places more clusters in these newly formed dense regions of the spectral space. The identified clusters were then evaluated for informational value.

IV. DISCUSSION OF PROCEDURE

An area, near Piedmont Lake in Belmont County, Ohio was chosen for the study. An August 1976 LANDSAT computer compatible tape was purchased from the Eros Data Center from which the study area subset was extracted. The August date was chosen because 1:125,000 NASA/U2 CIR photography for the same month was available for ground truth comparison. This summer date provided maximum information regarding vegetated and barren areas which is of particular interest in monitoring of land reclamation progress. To facilitate evaluation for areal extent of categories to be compared with the classification procedures, one of the NASA/U2 CIR photographs was optically enlarged and registered to two 7½' USGS topographic sheets of the study sites by General Electric Laboratory in Greenbelt, Md. This provided the registered data base for which mylar overlays could be used to outline category areas and measured with a Keuffel and Esser Polar Compensating Planimeter for the category acreage estimates. Each of the category areas was carefully identified, outlined, and then planimetered three times for an averaged value of acreage values for each of the categories.

The algorithm sequence in figure 4 illustrates the algorithms and sequence of applications for both the supervised classification approaches, as well as the development of graphic representation and classification evaluations of the LANDSAT data.

An iterative clustering algorithm (ISOCLS) was applied to the subset area data for the four spectral bands, with tightly applied parameters on the clustering (i.e. 2 standard deviations about the mean, 8 iterations through the data, and 30 clusters) (IDIMS Reference Manual, 1978). The statistics file generated from this clustering containing the means and covariance matrices, was subjected to the two

classification procedures illustrated in the flow chart in figure 4. This statistics file (JBLCLS76.STATS) was input to a maximum likelihood classifier (CLASFY ref. IDIMS Man. Ref., 1978) for further refinement of this classification. The same statistics file was also input to the canonical analysis program. This program is a modification of the ORSER CANAL program (Turner et al., 1978) and is implemented on the NASA/GSFC/-ERRSAC HP 3000 computer as a utility program. The modified canonical program develops a transformation matrix, transformed means, and transformed covariance matrices. The transformed means, and transformed covariance matrix are retained in a statistics file. In the IDIMS program, Kltrans, is an option which allows that program to be used as a matrix multiplier upon inputting the transformation matrix. Thus the original data set is canonically transformed. This canonically transformed data set was again subjected to ISOCLS with the parameters set as above with only 20 clusters (the researchers determined 20 clusters in the transformed data provided the categories of interest, thereby not necessitating the extra CPU time and the regrouping of clusters into information categories involved with 30 clusters).

Finally, each of the statistics files was subjected to a utility program (Compareg: ESL, 1978) to plot the means and one standard deviation about the means for any two axes or bands. Further, from each classified image a 10% sample including the informational categories of interest was subset corresponding to the mylar delineated areas in the enlarged and registered NASA/U2 CIR photography. These sampled areas included the same area and informational categories which could be planimetered with accuracy for comparative evaluation. This subset study site was then subjected to a pixel count and the clusters grouped according to categories for comparison.

V. DISCUSSION OF RESULTS AND CONCLUSIONS

A. GRAPHIC REPRESENTATION OF RESULTS

The Compareg utility program, discussed in the previous section on procedure, plots the mean of each cluster and the distribution of data values within one standard deviation about the mean using the statistics files for each classification. The mean is illustrated as the cluster number of the centroid and the one standard deviation (one sigma) by the ellipse about the centroid for any two spectral bands or axes.

In figure 5A (JBLCL76.STATS) a plot of the means and the one sigma ellipse about the mean is illustrated for the spectral bands 2 vs. 4 of the original clustered data. Of the 6 band plot combination for this clustering classification, 2 vs. 4 shows the maximum observed cluster separability in 2 dimensional spectral space for these data. The relative isolation of cluster number 16 (water category) and 17,1 (banding/unclassified) from the other cluster groupings in the data is apparent in figure 5A(JBLCL76 2vs4). There is also confusion of clusters within one sigma for the mean for categories involving forest/agriculture and reclaimed vegetation (see figures 5A & 6 2vs4 and cluster comparison. However, some of the clusters for forest, agriculture and unreclaimed barren mining separate in band 2 vs 4 plot figure 5A 2vs4.

In figure 5B (JBCNL76.STATS), a plot of the means and the one sigma ellipse about the mean is illustrated for the canonically transformed axes 1 and 2. In as much as these two transformed axes count for over 90% of the variability in the data, cluster graphing illustrates maximum separability of the clusters. The cluster number of the transformed data (figure 5A JBLCL76) are the same as the original clustered data (figure 6 cluster comparison). In reference to figures 5A & 5B JBLCL76 and JBCNL76 the relative positions and conditions of clusters 16, 17, and 2 are such that 1) the axes of the transformed data has been rotated and translated, 2) the rescaling of the transformed data results in more circular clusters about the centroid, 3) the reduction of confusion of clusters in the region near cluster number 2 maximizing the separability among clusters is apparent in the transformed data. For further interpretability, the region in figure 5A JBLCL76, about cluster number 16 represents relatively low reflectance in 2 dimensional spectral space; the region about cluster number 2 represents relatively low reflectance in bands 1 and 2 and relatively high reflectance in bands 3 and 4, and the region about cluster 9 represents relatively high reflectance in all four bands.

As illustrated in figure 5B JBCNL76, a few clusters were still inseparable after the transformation in the region near cluster 2. Therefore, rescaling in the transformation resulted in other clusters in the data not previously identified that are of informational value. The reclustering of the transformed data set is illustrated in figure 5C JBCNIS76.STATS. The region about cluster number 16 (figure 5C JBCNIS76) was then represented by four clusters, numbers 17, 13, 22, 2 in the reclustered data fig-

ure JBCNIS76 rather than only one cluster. These four categories related to sedimentation patterns, shallow water, and non-transformed reclustered classification. In figure 5B JBCNL76 the cluster in the region about cluster 9 was better defined in spectral space with greater separability of clusters and the inseparability in the original clustered region figure 5A JBLCL76 near cluster 2 is simplified and resolved in the reclustered transformed data (figure 5C JBCNIS76). An increase in the number of clusters to 30 for the transformed reclustering showed good separation of clusters and the additional clusters not only include the water categories but include additional clustering in the region about cluster 21 (see figure 5D JBLTNIS76.STATS). Since the objective of this study was to select informational categories applied to surface mining and land reclamation, and to demonstrate the two unsupervised classification approaches in this aspect, no attempt was made to further deliniate forest, agriculture and other land cover categories. Research is currently in progress with emphasis on stages of revegetation in reclamation, agriculture, forestation and analyze other classification confusion problems (Irons et al., 1980; Middleton and Bly, 1981) as additional ground truth information becomes available for evaluation.

B. CLASSIFICATION AGREEMENT

Prior to classification, the LANDSAT scene was geometrically and radiometrically corrected. Therefore, the classified products are based on a 1.1 acre pixel size. This allows for planimetric acreage comparisons. Figure 7 illustrates the various informational categories evaluated and also summarizes the results for each of the classification techniques as a percentage of level of agreement. As is illustrated in the figure, the clustering (ISOCLS) and maximum likelihood (CLASFY) algorithms are within a few percent of agreement for the classification categories. The clustering/transformation/clustering algorithm sequence demonstrates significant increases in the percentage agreement for each of the categories except water. The water category in the nontraditional unsupervised classification approach, however, identifies water categories not evident in the traditional unsupervised classification approach. The traditional classification approach did not include these additional water categories and the NASA/U2 CIR photography was not exactly the same date as the LANDSAT data. The dynamic condition of the sediment patterns in the water could vary considerably due to any new precipitation and therefore

comparative evaluation could not be reasonably assumed.

Perhaps the most significant results in terms of figure 7 (LANDSAT vs ground truth) is the 20% plus improvement in the classification by the nontraditional procedure for the categories being considered. Research is ongoing in identification and illustration of additional categories as more detailed ground truth information becomes available. Further research includes plans to evaluate this nontraditional unsupervised classification procedure for Thematic Mapper Simulator data in data reduction and classification.

VI.   SUMMARY OF CONCLUSIONS

1) The rescaling factor in the canonical transformation rearranges the density patterns of spectral data in regions of n dimensional spectral space.

2) These rearrangements of density patterns may be identified with a sensitive iterative cluster algorithm such as ISOCLS.

3) Cluster separability increases as a result of reclustering the canonical transformed data set.

4) These newly identified clustered density patterns apparently have higher correlation to informational categories than some of the original clusters.

5) There is an apparent significant increase in the agreement with ground truth information by the nontraditional unsupervised reclustered transformed data technique.

REFERENCES

Bloemer, H.H.L., J.O. Brumfield, W.J. Campbell, R.G. Witt, and B. Bly, 1981. "Application of LANDSAT Data to Monitor Land Reclamation in Belmont County, Ohio." Second Eastern Regional Remote Sensing Application Conference, Danvers, Mass.

ESL Technical Manual, Interactive Digital Image Manipulation System Reference Manual. Sunnyvale, California. 1978.

Irons, J.R., H. Lachowski, and C. Peterson. 1980. "Remote Sensing of Surface Mines: A Comparative Study of Sensor Systems." The 14th International Symposium on Remote Sensing of Environment. San Jose, Costa Rica. May 1980.

Lachowski, H.M. and F.Y. Borden, 1973.
"Classification of ERTS-1 MSS Data
by Canonical Analysis," NASA SP-327-1,
AO. NASA/GSFC, Greenbelt, Maryland.
pp. 1243-1251.

Merembeck, F., F.Y. Borden, M.H. Podwysoc-
ki, and D.N. Applegate. 1977. "Ap-
plication of Canonical Anaysis to
Multispectral Scanner Data." 14th
Annual Symposium on the Application
of Computers in Mine and Industry,
University Park.

Middleton, E.M. and B.G. Bly. 1981. "Vir-
ginia Strip Mine Project: A LANDSAT
survey of Wise County." ERRSAC Pro-
ject Report Number 81-X. Goddard
Space Flight Center, Greenbelt, Mary-
land.

Rogers, R.H., W.A. Pettyjohn, and L.E.
Reed. 1974. "Automatic Mapping of
Strip Mine Operations from Space/Craft
Data." NASA Technical Memorandum
79268. 19 p.

Russell, O.R. 1977. "Application of
LANDSAT 2 Data to the Implementation
and Enforcement of the Pennsylvania
Surface Mining Conservation and Re-
clamation Act." Final Report Pre-
pared for National Aeronautic and
Space Administration. Goddard Space
Flight Center, Greenbelt, Maryland.

Spisz, E.W. and J.T. Dooley. 1980. "As-
sessment of Satellite and Aircraft
Multispectral Scanner Data for Strip
Mine Monitoring." NASA Technical
Memorandum 79268. 19 p.

Tanner, C.E. 1979. "Computer Processing
of Multispectral Scanner Data Over
Coal Strip Mines." EPA-600/7-79-080,
Energy/Environment R & D Program Re-
port. Las Vegas, Nevada.

Turner, B. et al. 1978. Satellite and Air-
craft Multispectral Scanner Digital
Data User Manual. Office for Remote
Sensing of Earth Resources. The Penn-
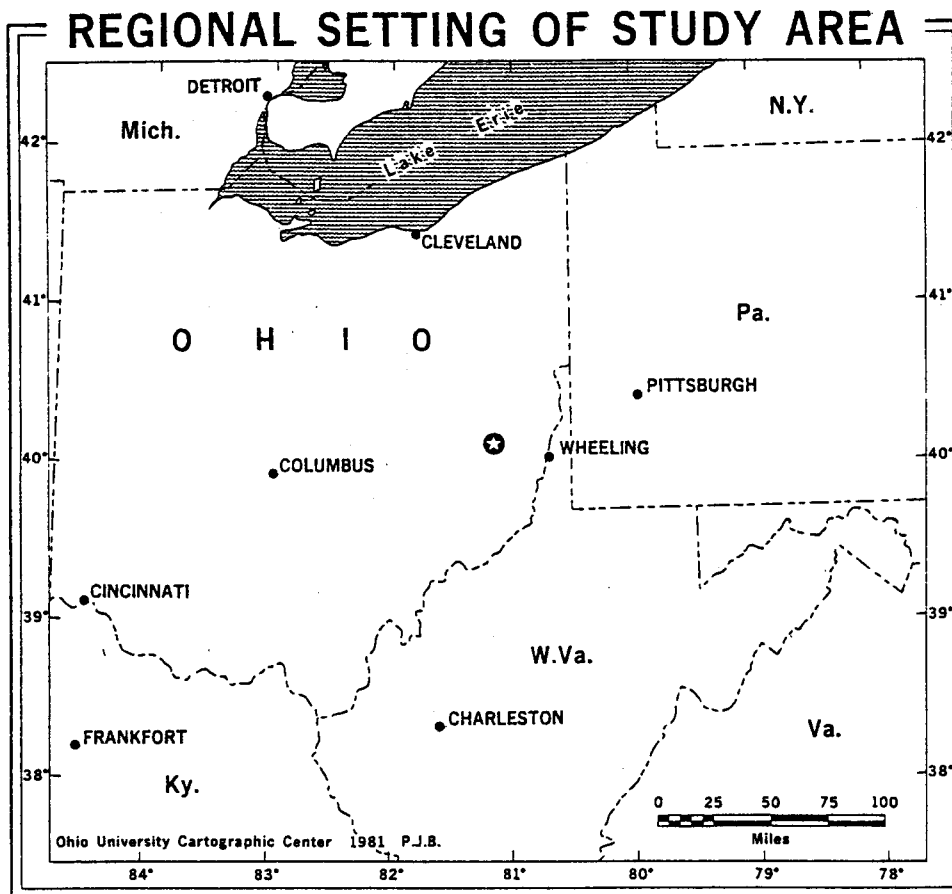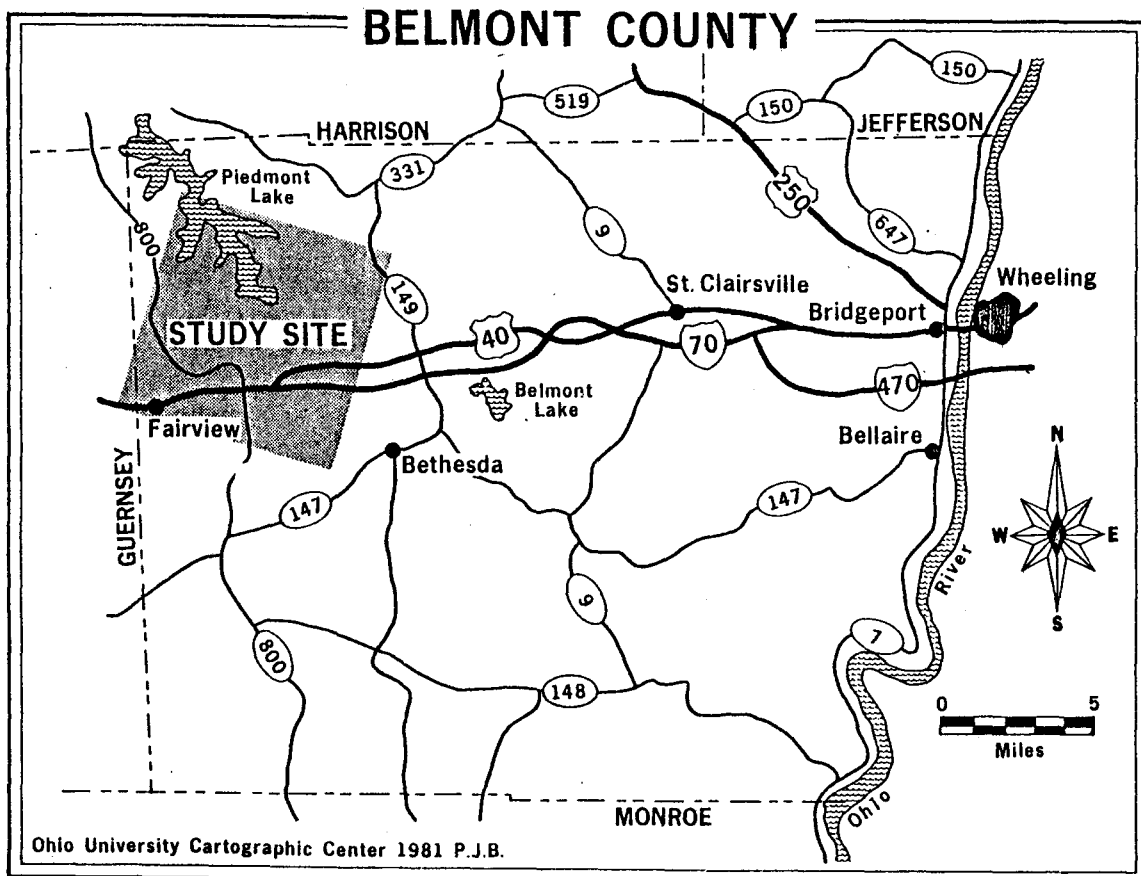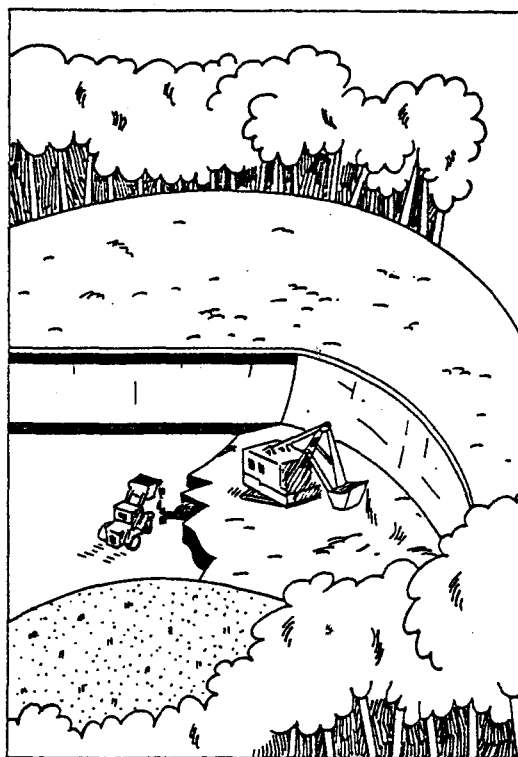sylvania State University, University
Park, Pa.

REGIONAL SETTING OF STUDY AREA

FIGURE 1

FIGURE 2

FIGURE 3

1981 Machine Processing of Remotely Sensed Data Symposium

ALGORITHM FLOW CHART
FOR
IDIMS — HP 3000
NASA / G.S.F.C. / ERRSAC — Greenbelt, Md.

FIGURE 4

Figure 5A
JBLCL76. STATS
(Original Data Clusters)

Figure 5B
JBCNL76. STATS
(Canonically Transformed Clusters)

Figure 5C
JBCNIS76. STATS
(Reclustered Canonically Transformed Data/20 Clusters)

Figure 5D
JBLTNIS 76. STATS
(Reclustered Canonically Transformed Data/30 Clusters)
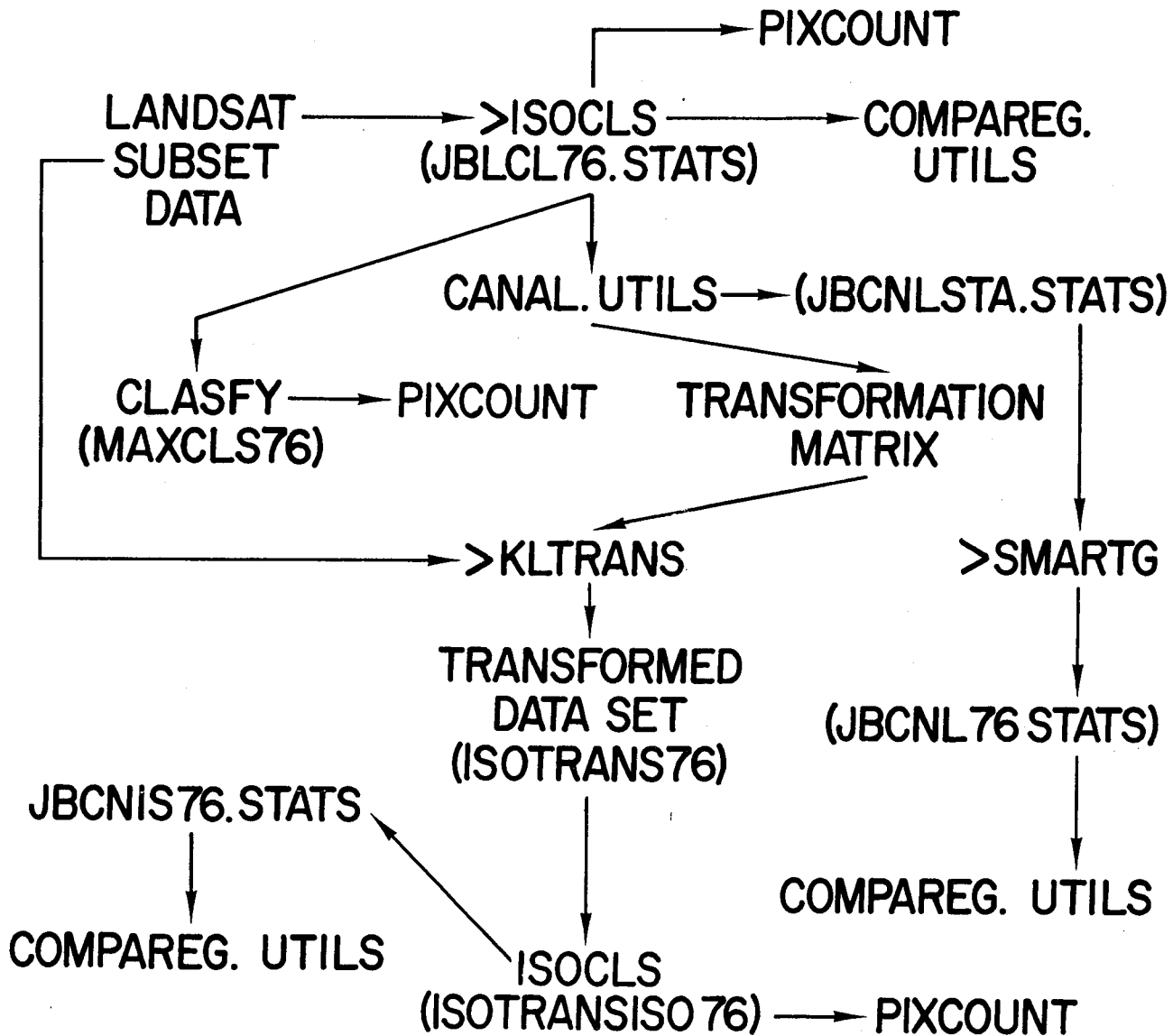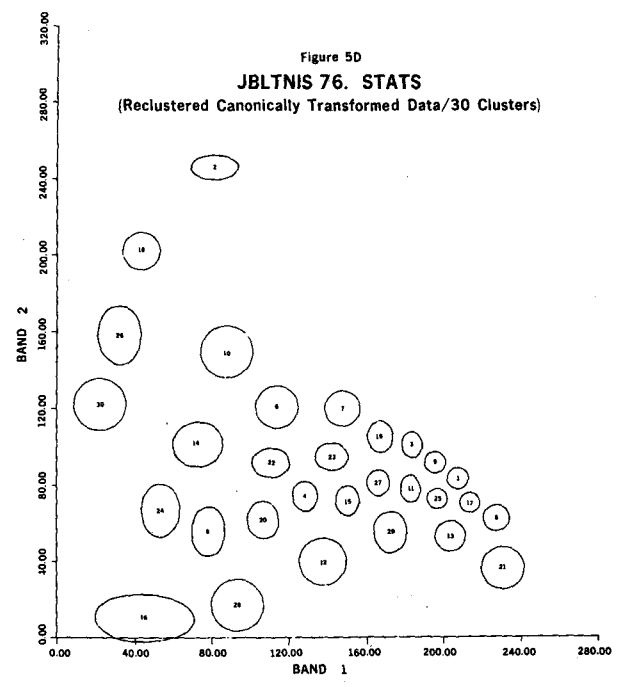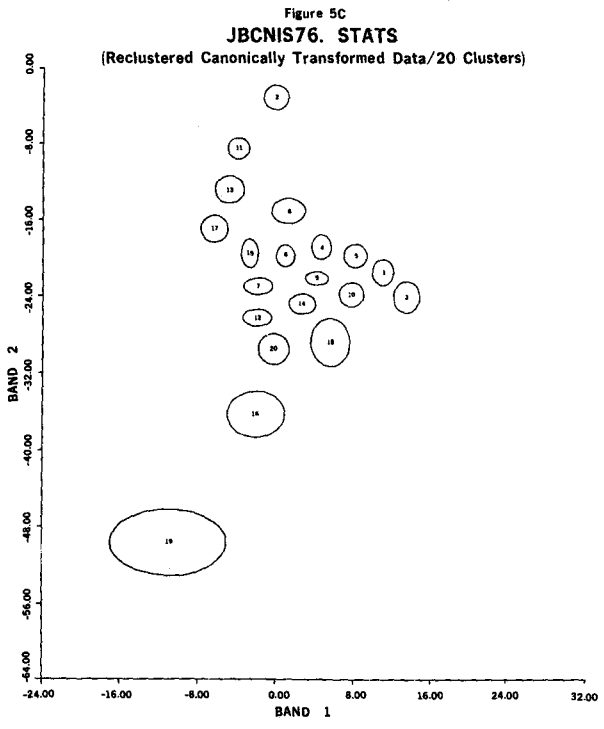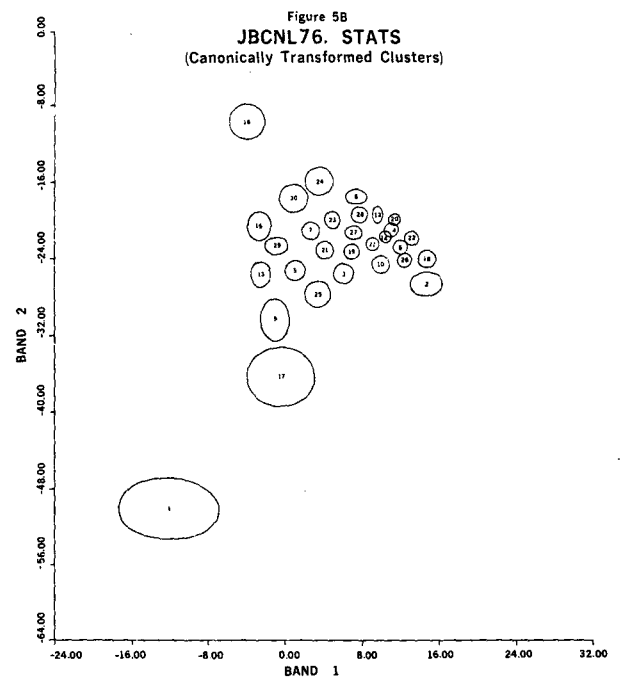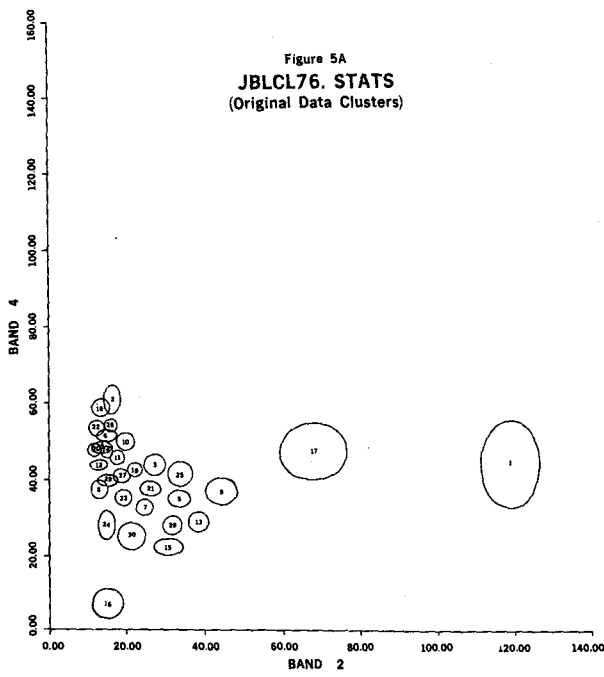
CLUSTER COMPARISON FOR CATEGORIES

ISOCLS 76 (JBLCL 76. STATS)		ISOTRANSISO 76 (JBCNIS 76. STATS)
NONE PRESENT	HIGH SEDIMENT SHALLOW H_2O	17
NONE PRESENT	MEDIUM SEDIMENT H_2O	13
16	H_2O	11, 2
13, 9, 25, 15	UNRECLAIMED BARREN	12, 7, 15, 20, 16
3, 11, 10, 5, 21, 19, 23, 27, 7 29, 30	RECLAIMED VEGETATED	10, 9, 14, 6, 4, 18
2, 4, 22, 20, 12, 18, 6, 14, 26	FOREST/AGRICULTURE	1, 5, 3
28, 8, 24, 17, 1	UNCLASSIFIED / BANDING	8, 19

FIGURE 6

LANDSAT DATA VS. GROUND TRUTH

ISOCLS		G.T.	LEVEL OF AGREEMENT
H_2O	290 PIXELS 319 ACRES	$22.8cm^2$ 324 ACRES	98.3%
RECLAIMED AREAS	1,684 PIXELS 1,852.4 ACRES	$108.6cm^2$ 1,544.53 ACRES	83%
UNRECLAIMED (BARREN)	276 PIXELS 303.6 ACRES	$30.5cm^2$ 433.78 ACRES	70%

MAXIMUM LIKELIHOOD CLASSIFICATION (MAXCLS)

H_2O	292 PIXELS 321 ACRES	99%
RECLAIMED AREAS	1,657 PIXELS 1,822.7 ACRES	72%
UNRECLAIMED (BARREN)	283 PIXELS 311.3 ACRES	84.7%

ISOCLS/CANONICAL TRANSFORM/ISOCLS

H_2O CLEAR	105 PIXELS 115.5 ACRES	
H_2O MED. SED.	13 PIXELS 14.3 ACRES	
H_2O HI. SED.	170 PIXELS 187 ACRES	
TOTAL 316.8 ACRES		97.7%
RECLAIMED AREAS	1,462 PIXELS 1,608.2 ACRES	96%
UNRECLAIMED (BARREN)	400 PIXELS 440 ACRES	98.58%

FIGURE 7