LARS Contract Report 060780

Quarterly Progress Report

Evaluation of SLAR and Thematic Mapper MSS Data for Forest Cover Mapping Using Computer-Aided Analysis Techniques

Contract No. NAS 9-15889

Reporting Period: March 1, 1980 - May 31, 1980

Submitted to: Exploratory Investigations Branch
NASA Lyndon B. Johnson Space Center

Prepared by: Laboratory for the Applications of Remote Sensing
Purdue University
West Lafayette, Indiana 47906

Technical Monitor: Dr. D. L. Amsbury
NASA Mail Code SF5
Exploratory Investigations Branch
Houston, Texas 77058

Principal Investigator: Dr. Roger M. Hoffer
Ecosystems Program Leader
LARS/Purdue University
West Lafayette, Indiana 47906

Table of Contents

List of Tables

I. Activities of the Past Quarter

A. Definition of Sampling Scheme for Test Data Set

A considerable effort was spent on defining an appropriate sampling scheme for developing the test data set. A key element in this process was to define the data set in which the same test areas could be used for each of the four different spatial resolutions being evaluated (i.e., 15 x 15 meter, 30 x 30 meter, 45 x 45 meter, and 60 x 75 meter). Several possibilities were considered. One that appeared to be among the most logical did not prove feasible. This approach would have allowed the basic sampling strategy to be defined according to the Landsat spatial resolution, and then all pixels at the higher spatial resolutions, within the Landsat pixel, would have been included as test data. However, this system could not be used for two primary reasons. First, the cover type identification for a particular pixel will change as a function of spatial resolution. For example, at the Landsat resolution (0.46 hectares) one might identify a particular pixel as being "cut-over forest land," but at a 15 meter spatial resolution, one would define some pixels within that same 0.46 hectare area as being cut-over forest land and other pixels as being brush windrow. Even in undisturbed forest land areas, at Landsat resolution one would define a pixel as being forest land whereas at 15 meter resolution one might define some pixels as being tree crowns and other pixels as being shadow areas in between the tree crowns. Such shadow areas would be tabulated as errors in the classification results tables if they were included in the "forest cover" category. It is apparent that the identification of cover types will vary as a function of spatial resolution, and there is a much greater level of detail in the definition of cover types as the spatial resolution becomes smaller.

The second reason for rejecting the approach of using all smaller spatial resolution pixels within a Landsat resolution element for the test data set was that such a large number of pixels would be generated for the 15 meter resolution that the LARSYS program would have to be modified.

The ideal allocation of sampling effort would result in sample sizes assigned to each cover type for each resolution such that the variances on the estimated error rate would be: 1) small, and 2) equal. This would provide accurate and comparable (between resolutions) error rate estimates. However, since the error estimates are based on normalized frequencies of the occurrence of an error their distribution is hypergeometric and, with sufficiently large number of sample points, can be assumed binomial. Therefore the optimal allocation of sampling effort would result in equal error estimate variances where

$$\hat{\sigma} = \frac{pq}{n} \; * \; .$$

This would direct a larger number of samples for those cover types having a large or very small frequency of error. Since no error frequency estimate is available (if we assume that, in the absence of information to the contrary, they will be approximately equal for the various cover types and resolutions) then the suitable course would be one such that approximately equal numbers of samples are taken for each cover type and each resolution.

---

* where p = probability of error, q = 1-p, and n = number of samples.

The approach that was adopted is basically a systematic sampling technique specifically designed to provide the ground cover identity for pixels of all resolutions being studied for each observation. Since all data sets of resolution coarser than that of the original aircraft data are simply unweighted arithmetic means of some "cell" of aircraft pixels, the line-column coordinates of any given point can be defined for each data set by the aircraft pixel count by line and by column for the cell. This results in certain restrictions when the precise location of an area or boundary is desired.

The sampling technique employs a grid of cells corresponding to the number of aircraft pixels in the 60 x 75 meter resolution data set (i.e., 4 x 5 pixels). The spacing between "candidate" test cells is determined by the mathematic relationships between the coordinate locations of a common area.

The actual location in the data of each "candidate" test cell will then be provided by the grid and its location relative to a COMTAL Vision One/20 image. The identity of the ground cover will then be determined through the use of the COMTAL image and the CIR aerial photographs. Each "candidate" cell component satisfying the selection criteria (i.e., some threshold of homogeneity with respect to cover type) will then be identified and recorded. The work on test data compilation is progressing satisfactorily.

The following software procedures were therefore developed during the past quarter:

1. Appropriate blocks of MSS data were reformatted in order to be displayed on the COMTAL display unit. Tables 1 and 2 show the current run number, tape number and file designation for all data sets involved in the current investigation.

2. A program was written to generate a grid for defining the sampling interval compatible with the display capabilities of the COMTAL. As indicated, the grid to be generated involved every third column and every sixth line of Landsat data spatial resolution. The same grid would define every fourth column and every tenth line in the 45 x 45 meter resolution data set, every sixth column and fifteenth line in the 30 x 30 meter data set, and every twelfth column and thirtieth line in the 15 x 15 meter data set.

3. A program was written to transfer the defined grid into a graphics plane in the COMTAL.

4. A program had to be written to translate the COMTAL coordinates into MIST format coordinates in order to implement the information obtained on the COMTAL in the MIST format compatible LARSYS processors.

5. A program was written to translate the coordinates among the various resolution data sets. In this case one could define a coordinate for the 30 x 30 meter spatial resolution and be able to locate the same place in the data on any of the other spatial resolution data sets and vice versa.

Table 1. Locational Information for Preprocessed NS 001 MSS Data Obtained on May 2, 1979

Flight Line	Original Data As It Arrived From NASA/JSC August 2, 1979	Lines and Columns Reversed (Reformatting Group)	Columns Reversed and Geometrically Adjusted	Response Level Adjusted by Column
1N	R(78006401)T(2616)F(2)[1/]	R(78006404)T(2663)F(1)	R(78006404)T(5234)F(1)	R(78006404)T(5235)F(2)
1S	R(78006400)T(2617)F(2)	R(78006403)T(2662)F(1)	R(78006403)T(5235)F(1)	R(78006403)T(5233)F(2)
2N	R(78006501)T(2651)F(1)	R(78006504)T(2673)F(1)	R(78006504)T(5236)F(1)	R(78006504)T(5234)F(2)
2S	R(78006500)T(2650)F(1)	R(78006503)T(2666)F(1)		
3N	R(78006402)T(2616)F(3)	R(78006405)T(2665)F(1)		
3S	R(78006502)T(3460)F(2)	R(78006505)T(2668)F(1)		

[1/] R=Run No.; T=Tape No.; F=File No.

Table 2. Locational Information for Final Data Sets (4 Spatial Resolutions) of NS 001 MSS Data Obtained on May 2, 1979 in South Carolina.

Flight Line	Spatial Resolution			
	15.3 x 15.3m	30.6 x 30.6m	45.9 x 45.9m	61.2 x 76.5m
1N	R(78006404)T(5235)F(2)[1/]	R(78006404)T(5238)F(1)	R(78006404)T(5238)F(4)	R(78006404)T(5238)F(3)
1S	R(78006403)T(5233)F(2)	R(78006403)T(5239)F(1)	R(78006403)T(5239)F(4)	R(78006403)T(5239)F(3)
2N	R(78006504)T(5234)F(2)	R(78006504)T(5240)F(1)	R(78006504)T(5240)F(4)	R(78006504)T(5240)F(3)

[1/] R=Run No.; T=Tape No.; F=File No.

All of these programs have been generated and the compilation of test data sets is progressing satisfactorily. The use of the COMTAL display system has been found to be very effective and very helpful for this phase of the investigation.

B. Training Statistics Development - Feature Evaluation

Initially the training statistics were to be developed using a Multi-Cluster Blocks (MCB) approach, which had been found to be very effective in previous investigations involving Landsat data. However, in the current study, because of the previously mentioned difficulties with identification of a particular pixel at the higher spatial resolutions, and because of the very large number of different spectral classes which were being defined in the 15 and 30 meter spatial resolution data sets, we found that the MCB approach had some distinct limitations in terms of the purpose of this particular investigation.

On behalf of the feature evaluation portion of the study, it appeared incongruous to conduct an evaluation of subsets of channels employing a "distance" measure between classes which were defined by the LARSYS CLUSTER (which iteratively minimizes the distance between "cluster center" and all cluster member points) using data from those same channels. The objective of the feature evaluation portion of this study is to determine how well a set of cover types can be differentiated on the basis of their spectral response patterns in a given set of wavelength bands. It was considered more appropriate to define the cover types with their corresponding spectral characteristics on a basis other than spectral characteristics alone.

Areas in the data of twelve different broad cover types were located. The ground cover condition and the line-column coordinates were determined for 247 such areas. These were initially sorted into 120 different ground cover condition classes (e.g., bare soil, saturated bare soil, emergent crops, ...) on the basis of ground cover condition information alone. These were pooled into 83 ground cover condition classes and the means and covariance matrices for all 83 classes were computed and punched to disk by LARSYS *STATISTICS.

Examining the histograms and the variances revealed that the pixels of many of the "fields," while being all of one information or ground cover class, were not of one spectral class. That is, many "fields" had a large amount of spectral variability within the "field" itself. Added variability often resulted from the combining of "fields" of equal or similar ground cover condition on the basis of that information alone.

At this point, it appeared necessary to somehow partition the pixels of the same ground cover into different groups to provide spectral classes covering the spectral variability within each cover class yet reducing the variability within each spectral class. Providing the feature evaluation analysis with spectral classes which do not accomodate the spectral variability naturally occurring within the various ground cover classes would not provide proper assessment of the separability between cover classes based on the channel subsets employed. Low transformed divergence values in

this case may well indicate the similarity among broad spectral classes more than the capability of discriminating among a set of ground cover classes based on spectral response in the various channel subsets. This problem could be particularly acute where a given cover type has several spectral classes in one channel due to the different amounts of radiation reflected from the surface in that waveband where the ground cover is in different states or conditions (e.g., wet vs dry, flowering vs vegetative), yet in other wavebands there are no such corresponding differences. If the multi-class waveband provides discrimination between that cover type and others of interest when the number of spectral classes is sufficient to accomodate the variable reflectance, but results in low separability measures when the variability is represented in a much reduced number of spectral classes, then that and similar bands will be under-evaluated.

To avoid or reduce this problem it was decided that some method of partitioning the spectral vectors associated with each cover type into "optimal" groups or spectral classes was desired in order to properly evaluate waveband combinations while minimizing the effect of spectral variability associated with the various states or conditions of the ground cover. The LARSYS *CLUSTER was then used to basically sort the vectors into spectral classes to reduce this within spectral class variability. The clustering algorithm employed has been shown to be sensitive to scaling or unequal linear transformations among channels [Anuta, 1979, Bartolucci, 1979]. It is currently unknown how the inequality of within class variance among the various channels affects the cluster class composition. The studies by Anuta have shown that those channels with the greater variance will tend to determine which cluster class any particular vector will be assigned. This seems consistent with the mathematical nature of the Euclidean distance measure employed by CLUSTER. Much work is needed to direct more appropriate unsupervised spectral class compilation techniques. At this point in time all of the spectral classes have been defined for the computation of the transformed divergence values used to evaluate the waveband combinations. Transformed divergence was selected as the measure of separability based on the work by Swain, Robertson, and Wacker (1971) which displayed correlation between probability of correct classification (PCC) and transformed divergence.

Divergence being,

$$D_{(ij)} = \tfrac{1}{2}\,\mathrm{tr}\,[(\Sigma_i^{-1} - \Sigma_j^{-1})(\Sigma_i - \Sigma_j)] + \tfrac{1}{2}\,\mathrm{tr}\,[(\Sigma_i^{-1} + \Sigma_j^{-1})(U_i - U_j)(U_i - U_j)T]$$

and the transformed divergence is

$$TD_{ij} = 2000\,[1 - \exp(-D_{ij}/8)]$$

where:

Σ = the covariance matrix of the respective class i and j

U = the mean vector of the respective class i and j

tr = the trace of a matrix (i.e., the sum of the diagonal components)

It should be apparent that transformed divergence is a measure of the statistical difference between two classes based on both the difference between their covariance matrices and their mean vectors in relation to the summed variance in each vector component. This measure is then "saturated" through use of an exponential transformation such that classes which differ beyond a certain level (which corresponds with a leveling of the probability of misclassification (PMC)) do not result in increasing transformed divergence values. Work on the waveband combination is progressing satisfactorily.

C. Data Quality Evaluation

After looking at the spectral variability within some of the data sets involved in the analysis, some questions were generated concerning the basic signal-to-noise characteristics within the data set being used. For this reason, a data quality evaluation test was run in which series of 4 x 5 blocks of data were defined over the Wateree Reservoir. The variance within each block would give an indication of the noise characteristics in the data assuming that the water was fairly calm (as was indicated by the photos) so that such variance would not be due to differences in reflection from waves. Because Wateree Reservoir extends in a somewhat northwest-southeast direction, we were also able to examine the variability of response level consequent to the radiometric adjustment procedure by carrying out an ANOVA on the response level in each channel, by line and by column. These results show that the variability by column was much reduced over that present prior to the radiometric adjustment processing. The current level is considered sufficiently small to render variations due to ground cover differences detectable. Variability by line was difficult to assess as silt content is highly correlated with proximity to the dam which is correlated with line. However, the level of variability in calibration values (corresponding to a non-reflective surface and "calibrated" lamp) was very low and indicated an acceptable level of quality for the purpose of this investigation. It should be pointed out that this portion of the study received proportionately little attention as the objective was to merely ascertain whether any noticable anomalies existed which would distract from the inferences made from the results.

D. Processing of the Landsat Data

Preliminary analysis of the Landsat MSS data revealed that the data set itself did not conform to the quality indications obtained from the Sioux Falls Data Center, primarily because a large portion of the test area contained a layer of high cirrus clouds which essentially render this particular set of Landsat data unusable. We are therefore in the process of evaluating other data sets of potential use in the analysis. It appears that we will have to revert to an anniversary data set because of the lack of reasonable quality data at a time frame close to that in which the aircraft data was obtained.

E. Spring 1980 Aircraft Mission

Frequent communication with the Aircraft Mission Group at NASA/JSC has been maintained during the last several weeks. A potential mission in mid-April had to be canceled because of a relatively late spring in the test site area and a concern that because many of the deciduous species were not fully leafed out, the data set would be very atypical for many of the cover types involved. The alternate dates of April 28-May 15 were then defined for obtaining the aircraft data. However, due to many instrumentation problems (particularly with the radar system), weather problems, and other flight committments, no data was obtained over the test site during this time period. After May 15, the aircraft were down for maintenance, and both aircraft will not be available again until the period starting June 16. It is hoped that a satisfactory data set with both the NC-130 and RB-57 can be obtained during the last two weeks of June.

II. Problems Encountered

The major problems of significance encountered during the past quarter have involved the poor quality of the Landsat MSS data and the difficulties experienced in collecting aircraft MSS and radar data during the period between April 28 and May 15. It is hoped that both a new Landsat CCT data set and the aircraft data will be obtained during the next quarter.

III. Personnel Status

The following personnel committed the respective percentages of time to the project during the past quarter.

Name	Position	Average Monthly Effort (%)
Anuta, Paul	Reformatting/preprocessing	3
Bartolucci, Luis	Research Physicist	7
Goodrick, F. E.	Professional Assistant	11
Hoffer, Roger	Principal Investigator	40
Latty, Rick	Research Associate	75
Peterson, Carol	Research Statistician	17
Peterson, John	Associate Director	5
Prather, Brenda	Secretary	50

IV. Anticipated Accomplishments

The following are the anticipated accomplishments of the forthcoming quarter (June 1, 1980 - August 31, 1980):

1) Definition of the Training Statistics for each of the different spatial resolution data sets.

2) Generation of the final set of test data for each of the different spatial resolution data sets.

3) Wavelength band selection analysis using the feature selection processor within LARSYS.

4) Continuation of the analysis of the four different spatial resolution data sets.

5) Conducting the field work in conjunction with the first 1980 aircraft data mission.

No major problems are anticipated during the forthcoming quarter.

References

Anuta, Paul E.,and Nim-Yau Chu, 1979. "Multidimensional Scaling for Clustering of Dissimilar Data Types," LARS Information Note 050279; NSF Grant # ENG-7614400. 27 pp.

Bartolucci, Luis A. and Ramon Bermudez de Castro, 1979. "Clustering of Landsat MSS Data - Certain Limitations," LARS Technical Report 060679. 17 pp.

Cochran, William G., 1963. Sampling Techniques. Second Edition, John Wiley and Sons, New York. 413 pp.

Swain, P.H., T.V. Robertson, and A.G. Wacker, 1971. "Comparison of the Divergence and B-distance in Feature Selection," LARS Information Note 020871. 12 pp.

Swain, P.H., and R.C. King, 1973. "Two Effective Feature Selection Criteria for Multispectral Remote Sensing," LARS Information Note 042663. 5 pp.