Paper L 1

COMPUTER-IMPLEMENTED LAND USE CLASSIFICATION WITH PATTERN RECOGNITION SOFTWARE AND ERTS DIGITAL DATA

Armond T. Joyce, *NASA Earth Resources Laboratory, Mississippi Test Facility, Bay St. Louis, MS 39520* and Thomas W. Pendleton, *Ibid*

ABSTRACT

Significant progress has been made in the classification of surface conditions (land uses) with computer-implemented techniques based on the use of ERTS digital data and pattern recognition software. The supervised technique presently used at the NASA Earth Resources Laboratory is based on maximum likelihood ratioing with a digital table look-up approach to classification. After classification, colors are assigned to the various surface conditions (land uses) classified, and the color-coded classification is film recorded on either positive or negative 9 1/2" film at the scale desired. Prints of the film strips are then mosaicked and photographed to produce a land use map in the format desired. Computer extraction of statistical information is performed to show the extent of each surface condition (land use) within any given land unit (e.g. township, county, drainage, etc.) that can be identified in the image. Evaluations of the product indicate that classification accuracy is well within the limits for use by land resource managers and administrators. Classifications performed with digital data acquired during different seasons indicate that the combination of two or more classifications offer even better accuracy. Future emphasis will include adaptation of software to general purpose computers, development of low-cost hardware for image display, and establishment of ground truth logistics for widespread implementation over large areas, e.g. statewide.

N74 30729

INTRODUCTION

ERTS-1 data offers the land use analyst several new dimensions. A single ERTS pass results in the collection of data over a swath approximately 100 nautical miles wide, whereas imagery acquired with mapping cameras flown in aircraft commonly cover swaths from two to fifteen nautical miles. ERTS repetitive coverage on an eighteen day cycle provides possibilities for a rapid detection of cultural changes on the earth's surface as well as seasonal differences in vegetation and land use practices. In addition, the digital form of the data is conducive to automated data processing based on computerized systems.

The objective of the study reported in this paper was to perform computer-implemented land use classifications utilizing ERTS digital data and pattern recognition software for two sets of data, each pertaining to a different season of the year, and to compare the two classifications as to their portrayal of seasonal differences in vegetation and agricultural practices. ERTS digital data acquired over the Mississippi coastal plains on August 5, 1972, and again on January 16, 1973 were selected for the study.

*Nasa Earth Resources Laboratory, Mississippi Test Facility, Bay St. Louis, MS 39520

**Ibid.

DATA PROCESSING

Land use classification at the Earth Resources Laboratory is performed using a Data Analysis Station (DAS) and UNIVAC 1108 software. The DAS consists of a Varian 620f computer with 16,000 16 bit words, two nine track digital tape decks, a color television display device (CRT) with light pen capability, a Singer color film recorder, a card reader and a line printer. The UNIVAC 1108 software consists of several modules which constitute a supervised maximum likelihood classification scheme based on Gaussian statistics. The modules are a statistical module, a training sample separation module, and a classification module.

The initial stage of data processing consists of reformatting on the DAS the nine track ERTS computer compatible bulk data tapes received from the Goddard Space Flight Center. The reformatting operation produces a data tape in a format suitable for the 1108 software and a display tape which can be used to drive the DAS CRT or the DAS film recorder. Using the display tapes and the light pen capability of the DAS CRT, the coordinates of surface areas with known land use, called training samples, are determined. These scan line and scan line element coordinates allow the training sample areas to be located in the data in a supervised classification system. Using the training sample coordinates and the reformatted bulk data tapes, the training sample data is extracted and stored on a training sample edit tape which can be used with the UNIVAC 1108 software.

The statistical module on the UNIVAC 1108 is used to compute means and covariance matrices and to plot histograms for each training sample. The information output by the statistical module is used to edit the training sample data and is used for input to the separation module. The separation module computes a measure, "divergence," of the similarity of pairs of training samples. The measure, while quantitative, is difficult to relate to physical processes. However, it is known that the larger the measure the greater the difference between the training samples. ERL uses the measure to determine which training samples can be grouped to form a training class and which training samples cannot be grouped but must be treated as subclasses. In particular, for the classification of the two subject ERTS frames, the divergences between all training samples which potentially belonged to a single class were computed. Those training samples which had a divergence of less than approximately 15 were grouped into a single subclass.

As an example of training sample grouping we could consider the class "forest" from the 7 August 1972 data set as shown in Figure 1. The training information for the "forest" class consisted of fifteen training samples identified as pine and twelve samples identified as hardwood. The divergence criteria grouped these training samples into three subclasses of pine and four subclasses of hardwood. Hence, the forest classification was derived from seven forest subclasses. In general, the six class classification was derived from twenty-three subclasses which were three soybean subclasses, one corn subclass, two exposed soil subclasses, two grass subclasses, one pasture subclass, three marsh subclasses, three water subclasses, one urban industrial subclass and the previously mentioned seven forest subclasses.

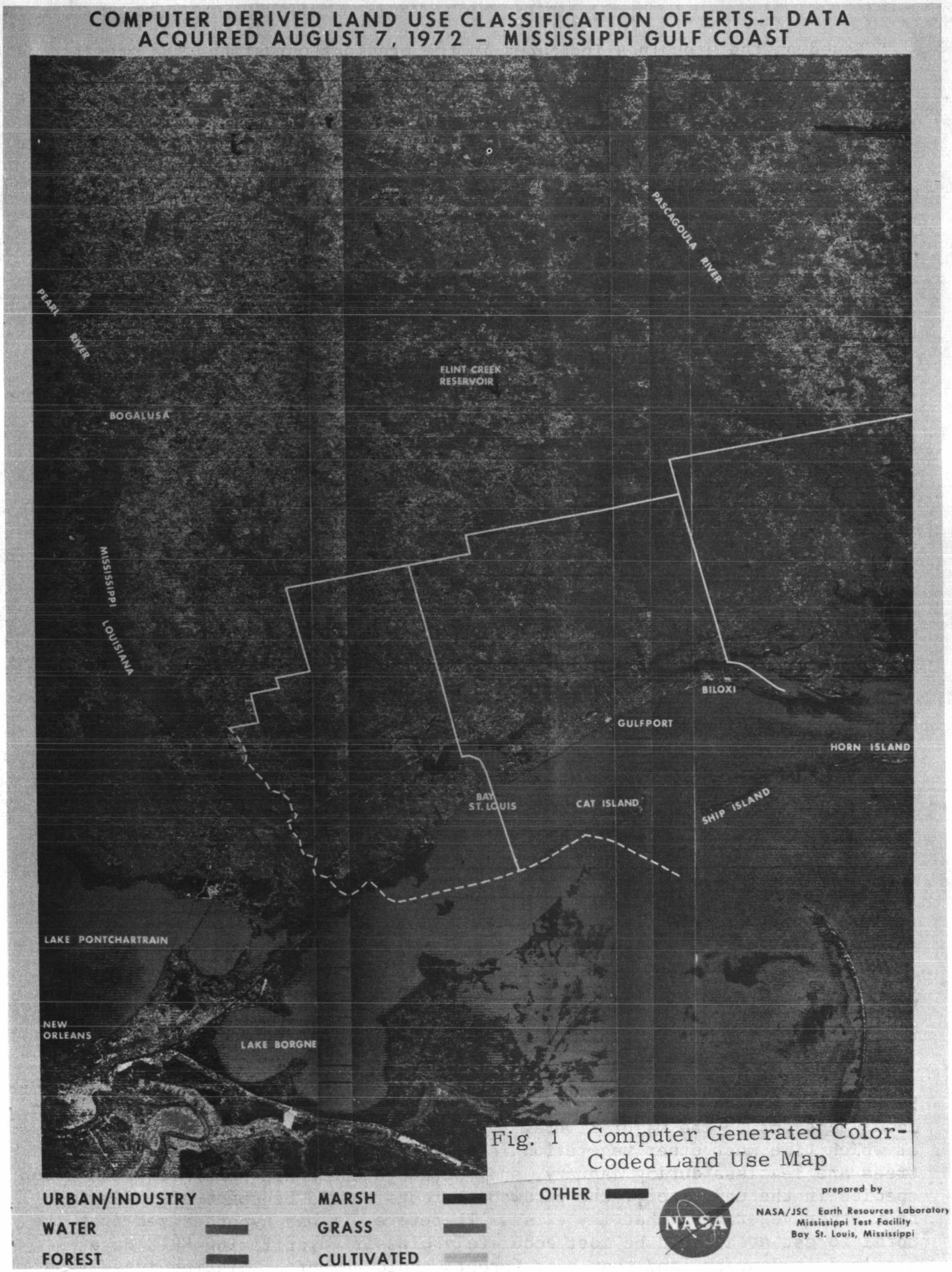

Fig. 1 Computer Generated Color-Coded Land Use Map

Based on the groupings indicated in the previous paragraph, the statistical module was used to generate information used by the classification module to classify the reformatted ERTS bulk data tapes. The classification algorithm is based on pre-storing in the computer a representation of each data element and the class to which it is to be assigned. This technique eliminates the need to compute for re-occurring data elements the probability that the data element belongs to each subclass, and the comparison of all such probabilities. The classification algorithm can process one ERTS computer compatible tape in eight minutes. However, the algorithm is limited to twelve classes. Since we used twenty-three classes, two passes were required per data tape. Therefore, four tapes or one ERTS frame requires about one hour to process. The resulting classification is stored on tape as a color-coded classification symbol for each data element.

The classification tape is displayed in false color on the DAS CRT and is displayed on the DAS film recorder. When the classification is displayed on the film recorder, rectification allows overlaying the classification data with a map of desired scale. The rectification technique considers scan angle, scan rate, sample rate, V/H ratio of the platform, rotation rate of the earth, and the characteristics of the film recorder. A quantitative evaluation of the rectification has not yet been made, but rectified data has been overlayed with a 1:250,000 scale map on a Traverse Mercator Projection. The match between the rectified data and the map appears to be very good in a region 25 x 100 nautical miles which corresponds to 1/4 of an ERTS frame and it is expected that the entire ERTS frame will match equally as well.

DATA ANALYSIS

The classifications for both sets of data were performed using the same geographic locations for training sample areas. There were eighty-two training sample areas which together encompassed 4376 resolution cells. The results of the classifications within training sample areas are shown in Table 1.

Overall, the classification within training sample areas indicates that the August 1972 data yielded a more accurate classification than the January 1973 data (94.8% versus 89.0%), but both classifications are well within the limits for use by land resource managers, administrators, or planners. The ground evaluation is still in progress, but preliminary findings show that classification accuracy within the entire test area (six counties) is not substantially different from the results of the classification within training sample areas. However, in viewing the statistics in Table 1, it is evident that certain surface conditions were classified more accurately with one set of data than with the other.

The forested areas of the Mississippi coastal plains are mainly pine forests, but there are also large areas covered by swamp hardwood forest in the bottomlands adjacent to the major rivers. Pine tree foliage is green during January at which time most other vegetation is either dead or leafless. Most hardwood trees are leafless during January, and, although there are some evergreen brush species in the understory, the hardwood forests should be spectrally distinct from other vegetation that may exist. Therefore, it was hypothesized that the total forest area could be most accurately classified utilizing ERTS data acquired in January, and that pine forests could be better distinguished from

Table 1. Classifications within training sample areas expressed as percentage of total cells representing a given surface condition classified as pertaining to that surface condition.

Surface Condition	Aug. 1972	Jan. 1973
(Land Use)	data	data
Total forested area	92.4	97.8
Pine forest	81.4	91.5
Hardwood forest	72.7	88.5
Total cropland area	84.6	84.2
Soybeans	80.0	-
Corn	96.0	-
Exposed soil	92.0	-
Winter ryegrass	-	89.1
Stubble	-	69.8
Grass (improved and unimproved pasture)	89.0	80.4
Marsh (non-forested wetlands)	94.9	67.4
Water	97.6	98.9
Inert materials (asphalt, concrete, metal, etc.)	94.9	65.9
Overall	94.8	89.0

hardwood forests with the same data. The validity of this hypothesis is indicated by the classification results shown in Table 1 which show that the classification within forest training sample areas was 92.4% for August data and 97.8% for January data. Furthermore, the scorecard classification within pine forest training samples was 81.4% for August data versus 91.5% for January data.

During August 1972, on the Mississippi coastal plains, the main agricultural crop was soybeans, although there was some corn and some exposed soil in cultivated areas. Soybeans and corn are spectrally similar, and both crops are spectrally similar to grass vegetation. However, exposed soil is spectrally distinct from all green vegetation during August. During January, some of the cultivated area contains winter ryegrass in a green growing condition; and the remainder of the cultivated area contains stubble (dead soybean or corn stalks) or dead vegetation that is spectrally distinct from ryegrass but spectrally similar to dead grass and dead marsh vegetation. Therefore, it was hypothesized that there was not likely to be a significant difference in land use classification accuracy between January data and August data for the total cultivated area, but that the ryegrass classification attained with the January data was likely to be more accurate than the soybean classification attained with the August data. The validity of this hypothesis is also indicated by the scorecards. The classification within all training samples for cultivated areas was 84.6% for August data and 84.2% for January data, but the soybean classification with August data was 80.0% and the ryegrass classification with January data was 89.1%.

Marsh vegetation is in a vigorously growing state during August; whereas, except for a few evergreen brush species in some areas, marsh vegetation is dead during January, and should be spectrally similar to dead grass, stubble and leafless hardwood. Therefore it was hypothesized that the most accurate classification of marsh vegetation could be attained with August data. The scorecard results show 94.4% for August data and 67.4% for January data. The low accuracy for the marsh classification with January data is attributed to a similarity between a marsh spectral signature and a spectral signature for a flooded condition under leafless hardwood swamp forest. In the January classification, 21.7% of "marsh" training sample cells were classified as "hardwood."

A computer-derived classification, as used in this study, is based on separating surface conditions that have different spectral characteristics caused by differences in reflected energy as measured from above. Urban, commercial, industrial, and residential land uses can be separated from the other land uses only in-as-much as their surface conditions with inert materials (asphalt, concrete, metal, wood, etc.) are spectrally different from vegetation or other material (sand, water, etc.). Residential or urban areas that have foliated trees overtopping the buildings as well as lawns and shrubs occupyng surface areas as seen from above are likely to be classified as vegetation. However, the color assigned to inert materials (asphalt, concrete, etc.) and the colors assigned to other surface conditions (grass, trees, etc.) that may be in the urban environment will form color patterns on a color presentation that can be interpreted so as to enable delineations of urban areas, especially to separate urban commercial and industrial centers with large concentrations of inert surface materials, residential areas with associated vegetation, and other land uses. In this context, the classification results within training sample areas is meaningless to the accuracy of the classification of the total urban commercial, industrial, or residential area. However, in-as-much-as the hardwood trees that overtop one-story or two-story buildings are leafless and lawn grasses are dead during January, a larger portion of the urban areas outside of training sample areas was classified as having inert material (asphalt, concrete, etc.) on the surface when utilizing ERTS data acquired in January than when utilizing data acquired in August when all vegetation is green.

CONCLUSIONS

In summary, it was shown that August data yielded the best classification to determine the extent of the area occupied by marsh vegetation, soybeans, and exposed soil; whereas, the January data yielded the best classification of the forested areas (as well as separating pine forest from hardwood forest), winter ryegrass and urban areas. It was also shown that there were no significant differences.

between the August and January classifications of the total agricultural area, or water bodies. These findings suggest that even better classifications may be attained for certain surface conditions utilizing data acquired during other seasons (e.g. for the total cultivated area during May when most fields are in some phase of soil preparation and have no significant amounts of green or dead vegetation). If this hypothesis is valid, it may be possible to attain the most accurate classification of surface conditions by integrating land use classifications made with ERTS data acquired during each season of the year.

Work is proceeding to produce a land use classification with ERTS data acquired during the spring when nearly all cropland is in some state of soil preparation, and during the fall when deciduous species have changed leaf color but have not shed their leaves. A comparison of four classifications of data acquired during the four seasons will determine the extent to which land use classification can be improved by combining two or more classifications for different seasons.

Future work will also include adaptation of software to general purpose computers, and establishment of ground truth logistics for the implementation of computer implemented land use classifications over areas larger than the 100 nautical mile by 100 nautical mile area encompassed by this study.