PROJECT INITIATION

Date: 9/6/72

Project Title: Analytical Chemistry

Project No.: 7-157

Project Director: Dr. J. H. Spann

Sponsor: Georgia Department of Natural Resources

Effective: 6/6/72, Estimated to run until 9/30/72

Type of Agreement: Standard Industrial dated 6/25/72

Amount: $2,860.00

REPORT REQUIRED: Final Report

SPONSOR CONTACT PERSON: Dr. Richard (Dick) Bailey
Georgia Department of Natural Resources
Atlanta, Georgia 30311

Assigned to: Special Techniques Division

COPIES TO:

- Project Director
- Director
- Assistant Director
- GTRI
- Division Chief
- Service Groups
- Patent Coordinator
- Photographic Laboratory
- Security, Property, Reports Coordinator
- EES Accounting
- EES Supply Services
- Library
- Rich Electronic Computer Center
- Project File
PROJECT TITLE: "Computer Analysis of ERTS Imagery"

PROJECT NO.: A-1557

PROJECT DIRECTOR: Mr. G. W. Spann

SPONSOR: Georgia Department of Natural Resources

TERMINATION EFFECTIVE: 9-30-73 (Contract Expiration)*

Charges should clear accounting by: All funds expended

Final report submitted 4-8-74.

CONTRACT CLOSEOUT ITEMS REMAINING: None

SPECIAL TECHNIQUES DIVISION

COPIES TO:
Project Director
Director
Associate Director
Assistant Director
Division Chief
Branch Head
Accounting
Engineering Design Services

General Office Services
Photographic Laboratory
Purchasing
Report Section
Library
Security
Rich Electronic Computer Center
FINAL TECHNICAL REPORT
PROJECT A-1557

COMPUTER LAND USE MAPPING FROM ERTS

By
G. William Spann and N. L. Faust

Prepared for
The Department of Natural Resources
Office of Planning and Research
Capitol Square
Atlanta, Georgia

25 March 1974

ENGINEERING EXPERIMENT STATION
Georgia Institute of Technology
Atlanta, Georgia 30332
COMPUTER LAND USE MAPPING FROM ERTS

FINAL TECHNICAL REPORT

PROJECT A-1557

G. William Spann and N. L. Faust

March 25, 1974

Prepared for

The Department of Natural Resources
Office of Planning and Research
Capitol Square
Atlanta, Georgia

By

Engineering Experiment Station
Georgia Institute of Technology
Atlanta, Georgia  30332
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>II. AREAS PROCESSED</td>
<td>3</td>
</tr>
<tr>
<td>III. PROCESSING TECHNIQUES</td>
<td>5</td>
</tr>
<tr>
<td>IV. IDENTIFICATION OF LAND USES</td>
<td>7</td>
</tr>
<tr>
<td>V. COST ANALYSIS OF THE AUTOMATIC PROCESSING OF ERTS DATA</td>
<td>13</td>
</tr>
<tr>
<td>VI. ADVANTAGES OF AUTOMATIC PROCESSING</td>
<td>17</td>
</tr>
<tr>
<td>VII. SUMMARY AND CONCLUSIONS</td>
<td>19</td>
</tr>
</tbody>
</table>
Figure 1. A Portion of the Electromagnetic Spectrum.
II. AREAS PROCESSED

The initial area of processing was chosen in close coordination with the Georgia Department of Natural Resources (DNR) and included an area centered on downtown Atlanta and encompassing all major highway systems of Atlanta including the perimeter (I-285). A large area of rural land to the west of Atlanta was specifically included to test the complexity of the clustering of a rural versus an urban scene. The study was intentionally conducted in an interactive atmosphere so that any changes in the areas covered or the types of processing desired by DNR could be accomplished with a minimum of confusion.

Initially several printouts were shown to DNR to ascertain that EES was producing results in a manner acceptable to them. After discussion with DNR, it was decided that a grey level printout of the results of the clustering programs was the preferred method of display. (Hand coloring of the printouts proved too time consuming and too costly.) A grey level presentation of the data was then developed. Next, a short study was made to determine which channel of data contained information related to highway systems and urban areas. As discussed later, channel 4 (.5 - .6 \mu m) from the ERTS tapes readily allowed discrimination of major cultural features of Atlanta including the highway systems.

Channel 4 data was used subsequently in one channel clustering over the whole test area. These results were then forwarded to DNR for study. Some time was spent in conjunction with DNR in identifying the clusters using aerial photos as ground truth data. Then, at the request of DNR, the area of study was restricted to the district west of Atlanta which includes Interstate Highway 285, the Fulton Industrial District, and much rural land. An intensive study was conducted in this area using one channel and multichannel clustering to observe the effects of multichannel data on the discrimination of classes which could be directly related to the ground truth data. Four channel clustered data was found to be superior to any one channel of data or other combinations of two or three channels.
Other areas of intensive study and multichannel clustering were the downtown Atlanta-Peachtree Road area, and the Atlanta Airport area.
III. PROCESSING TECHNIQUES

The capability was developed to automatically process remote sensing data via supervised classification or unsupervised clustering. The basic unsupervised technique employed in this study is a clustering embodied in the computer program ASTEP (Algorithm Simulation Test and Evaluation Program). One data base used was ERTS (Earth Resources Technology Satellite) digital data. Since ERTS data is composed of four observations in different spectral bands of a particular resolution element on the ground (57 meters by 79 meters), this information may be represented as a four element vector associated with each resolution element.

This four element vector may be compared in four dimensional (spectral) space to the vectors associated with all resolution elements in a particular area. The distance in spectral space is computed to each other vector and is used as a basis for comparison of the data from different resolution elements. If a vector is within a specified four dimensional distance of another vector, the values for the two vectors are combined and a mean vector is computed. This mean value denotes the center point of the cluster containing the two vectors. Each data vector is considered in turn and tested against the already established clusters. If the distance to any of the cluster means is beyond a preset tolerance, a new cluster containing only that vector is generated. This proceeds until all the data are processed or until the specified maximum number of clusters is exceeded. In the latter case, a statistical method is employed to decide which cluster should be eliminated so that the procedure may continue.

One can see that since the data is processed sequentially, this technique may impart a bias to the data depending on the point at which the data reduction is initiated. This drawback is eliminated by processing all the data a second time and using the resultant mean vectors from the first pass through the data as weighted initial mean vectors for the second pass.

Once the data are clustered, the major problem of how to display the data in an efficient manner remains. During the course of this study, several methods were attempted. Initially the ASTEP program printed letters
that represented the cluster into which a particular patch of ground was classified. This was found unsuitable because it was difficult to distinguish one cluster from another using only letters of the alphabet, A, B, C, D, etc., on the computer printout. A normal printout contained 500 lines of letters with 120 letters per line. Thus, a new method for visual discrimination of clusters was employed. This involved color coding each cluster letter so that it would stand out as being different from other clusters. This was satisfactory from a cluster detection point of view but was entirely unsuitable from a manpower standpoint since much painstaking effort was needed to produce a single computer printout. Therefore, a new automatic method for cluster display was sought. The method finally decided upon was a grey scale differentiation of the different clusters by assignment of various letters or groups of letters to each cluster. This involved programming the line printer to over print some characters with other letters so that the grey scale effect would be achieved. This addition involved increasing the processing time, but it did produce grey scales which were easily differentiable by the eye.

The final problem encountered was that the computer printouts represented different scales in the horizontal and vertical directions, and, therefore, could not be used to overlay maps of the area. Several attempts were made to find a way to adapt the line printer so that at least the scales in the horizontal and vertical directions would be the same. However, mechanical problems were involved, and no funds were available for purchase and modification of a computer printer.

This problem was next attacked by attempting to use the Cal-Comp plotter system at Georgia Tech to reproduce the clustering printouts at a scale suitable for mapping. This process is an order of magnitude more expensive than line printer output, but early results have shown that it produces good results at map scales.

Recently, EES has tried a new approach to the scaling problem. A method of internal scaling on the original ERTS data base has been tested and proven to be an economical way to produce a map scale output. The user may input the exact map scale at which he wishes to have the output produced. The output medium is the computer line printer. The additional computer costs associated with this scaling are nominal.
IV. IDENTIFICATION OF LAND USES

One of the purposes of this pilot study was to determine the effectiveness of computer processed ERTS data in the discrimination of land use patterns in Georgia. This study was entered into by the parties involved with a knowledge of the limitations on the resolution of ERTS data, i.e., at maximum resolution one element represents an area on the ground 57 meters by 79 meters. Thus, many of the classes that are detected by the clustering technique may be mixtures of two or more things—one cluster may represent a mixture of trees, driveways, and houses. The problem was approached with these factors in mind.

During the initial phases of this study, it was discovered that the ERTS MSS channel #4 (.5 - .6 \( \mu \)m wavelength) could be used effectively for the recognition of manmade objects such as highways, groups of buildings, office parks, and large parking lots. In the first processing of the Atlanta area, all major roads were located as well as major shopping centers, industrial areas, and airports. The clustering technique using only channel #4 data resulted in six different classes which were interpreted using 1973 aerial photos as ground truth to be:

1. cleared land (low vegetation),
2. green vegetation/trees,
3. highways, light industrial/commercial/multi-family, residential,
4. central business district/heavy industrial,
5. bare ground, and
6. concrete.

Channel #7 (.8 - 1.1 \( \mu \)m) is the channel that is farthest into the near infrared portion of the spectrum. Channel #6 is also in the near infrared (.7 - .8 \( \mu \)m). These channels show water very well and may be used to indicate some water courses and major lake boundaries.

Each channel of ERTS data contains some information that is not present in other channels. It is not surprising, then, that the clustering of all four channels of ERTS MSS data yields more clusters and significantly more
information than any single channel of data alone. The clustering of 2 and 3 channel data sets was attempted but was found to be less desirable than using all four channels. The maximum number of classes that were allowed in this study was 20—each ideally representing something significantly different (in a spectral sense) than any other cluster. A general breakdown of the clusters using all four channels is given below.

(1) Bare Ground - (7) Parking Lots
(2) Grass Areas & Farmland (8) Airport Runways (concrete)
(3) Trees (Conifers) (9) Mixed Forest
(4) Trees (Hardwoods) (10) Roof Tops, Large Buildings
(5) Water (11) Vegetated Residential (Single Family)
(6) Highways (12) Unvegetated Residential (Multi Family)

The land use classes enumerated above are the ones that represent identifiable features. It is probable that there are other unique classifications existing in the data. However, it has not yet been possible to make a complete evaluation of all land use categories due to the lack of ground truth taken simultaneously with the ERTS imagery. Nevertheless, EES has been able to differentiate and identify enough land uses to conclude that the techniques do provide separable land use categories.

In the supervised classification, the same land use categories were identified as in the unsupervised clustering. However, the incidence of misclassification was slightly lower with the supervised approach. Exact accuracies for the two types of classification cannot be calculated because the ground truth data were obtained several months after the ERTS data. Accuracies on the order of 80% for the unsupervised classification and 90% for the supervised classification are thought to be correct; this could vary depending on the scene analyzed. Accuracies were estimated using a sampling of points from the area around Cumberland Mall/I-285/Chattahoochee River.

Figures 2a, 2b, and 2c compare the land use discrimination available in ERTS photographs, ERTS computer processed data, and low altitude aerial photographs. A supervised classification was performed on the area around the I-75 - I-285 interchange and the Chattahoochee river.
Figure 2a. ERTS Photographs (scale approximately 1:500,000) of Atlanta and Gainesville/Lake Lanier. The Study Area is in the Lower Left Portion of the Picture.
Figure 2b. Computer Print Out (scale 1:50,000) of Study Area Which Includes the Chattahoochee River, I75/I285 Interchange and Cumberland Mall. The Computer Symbols Represent the Following Land Use Categories.

- WATER
- PINE FOREST
- COMMERCIAL
- HARDWOOD FOREST
- MULTI FAMILY RESIDENTIAL
- GRASSY AREAS
- SINGLE FAMILY HOUSING
- TRANSPORTATION
- BARE GROUND
- CONCRETE
Figure 2c. Aerial Photomosaic (scale 1:50,000) of Study Area Which Includes the Chattahoochee River, I75/I285 Interchange and Cumberland Mall. Photographs Furnished by Georgia Department of Transportation.
The computer program used to process the ERTS multispectral data was ASTEP (Algorithm Simulation Test and Evaluation Program). This program and several others for processing remotely sensed data are operational on the Univac 1108 time sharing system at Georgia Tech. It has the capability to perform clustering (unsupervised classification) as well as supervised classification. Since it was designed as a testing program for various remote sensing processing techniques, it is built in modules and automatically records the amount of Central Processing Unit (CPU) time spent in each module of the program. Thus, processing times for input of data, clustering, and output may be considered separately. The data are detailed below.

(1) Input
(500 lines @ 120 char/line) (a) 1 channel ~ 9.8 secs
(b) 2 channels ~ 15.6 secs
(c) 3 channels ~ 20.6 secs
(d) 4 channels ~ 28.3 secs.

(2) Clustering

The processing time under this heading depends on many factors. Some of these are number of classes, number of channels of data, and type of initialization for the clustering. The ASTEP program asks the user to specify the maximum number of classes that will be allowed. Unless the maximum is reached at some time during the processing, the program is free to select the number of classes warranted by the data. If the user wishes he may input vectors to act as the initial mean vectors for the clusters. Otherwise the initial means will be taken as zero. This input step is necessary when correlation between separate clustering runs is needed, i.e., if a class that has been identified on one run is to have the same symbol representing it on a subsequent run in a different area. The
The table below gives the CPU time as a function of the number of channels and number of clusters.

**TABLE 1**

<table>
<thead>
<tr>
<th>Number of Clusters</th>
<th>Number of Channels of Data</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1.0 x 10^{-3}</td>
<td>1.4 x 10^{-3}</td>
<td>1.9 x 10^{-3}</td>
<td>2.4 x 10^{-3}</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1.6 x 10^{-3}</td>
<td>2.4 x 10^{-3}</td>
<td>3.1 x 10^{-3}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>1.8 x 10^{-3}</td>
<td>2.7 x 10^{-3}</td>
<td>3.7 x 10^{-3}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>2.0 x 10^{-3}</td>
<td>3.2 x 10^{-3}</td>
<td>4.6 x 10^{-3}</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

By plotting the data from Table 1, it can be seen that there is an almost linear relation between number of clusters and computation time in the ASTEP program. The slope of this linear relation, however, changes with the number of channels of data being used. For example, the slope for two channels of data is approximately 40 microseconds per cluster per acre. For three channels of data, the slope is approximately 75 microseconds per cluster per acre, and for four channels of data, the slope was found to be approximately 150 microseconds per cluster per acre.

Because of the unavailability of current ground truth information, a supervised classification of the test area could not be accomplished; however, from previous experience with supervised classification, the classification time per acre was estimated to be .01 sec/acre or greater with 10 classes and four channels of data. Thus, for 10 classes, the unsupervised approach is approximately a factor of three faster than the supervised techniques, although there is expected to be some improvement in accuracy by using the supervised scheme. The time estimate for supervised classification does not
include the time required to train the classifier by computing the statistics of many training fields of data.

(3) Output

The normal output mode for the ASTEP program at the start of this study was a line printer printout. The average time for a 500 line output was 13 seconds of CPU time. Later, after it was decided to go to a grey level print instead of hand coloring the printouts, the program was rewritten to overprint each line of the printout to allow greater grey level discrimination. This increased output time for 500 lines to 30.0 seconds for three overprints and 38.2 seconds for four overprints as requested in the state symbol set requested by DNR.

As stated above, a major objection to this type of output was that the scale in the horizontal direction was not the same as the scale in the vertical direction. Therefore, other methods of output were investigated. The Cal-Comp plotter on the Georgia Tech Univac 1108 system allows one to specify scales in two directions, and this method has been incorporated into the ASTEP output package. The principal disadvantages to the Cal-Comp plots are the long plot times required and the slow turn-around of computer runs. By making the Cal-Comp plotter use symbols developed at EES instead of its own internal set of symbols, the plot time was reduced from 3 minutes per 120 character line to approximately 0.6 minutes per 120 character line. Even with this reduction, Cal-Comp plots are at least a factor of 10 more expensive than line printer outputs. However, plots from the ASTEP program currently may be overlaid on 1:24,000 scale maps, and changes to the scale factors in ASTEP are easily accomplished using the Cal-Comp plotter.

The internal scaling method which was discussed in the processing techniques section has proved to be the most economical way to attack the scaling problem. The total computer time is not increased significantly by this method since the line printer can be used.
VI. ADVANTAGES OF AUTOMATIC PROCESSING

There are many books, papers, and articles that deal with the advantages of automatic processing techniques. No attempt will be made to reiterate what is contained in these sources. However, a brief discussion of the land-use analysis problem in Georgia is presented.

There are approximately 50,000 square miles in Georgia—the largest state east of the Mississippi River. This is 32 million acres. When considering a state-wide or regional land-use analysis, an acre is generally a reasonable size unit of area to work with. To produce a state-wide land-use analysis at this scale would then require 32 million decisions as to the primary land-use at each point in Georgia. Even if a sufficient number of trained photo-interpreters were available, the state could hardly afford the time and money required to produce such a study manually.

However, a computer based land-use analysis of Georgia using four-band ERTS data and unsupervised classification techniques would take approximately 50 hours of U1108 time. At the current rate of $400 per hour, this is $20,000 in computer time. For any region of the state containing a reasonable area (at least 100 square miles), the same computer cost ratio would hold, that is, the computer processing costs would be approximately $ .40 per square mile or about $ .15 per square kilometer.

The amount of man-time needed for the analysis of the computer printouts is directly dependent on the complexity of the problem. For example, a user interested only in the regional characteristics of the data would need less interpretation time than would a city planner who would desire detailed urban data. Using automatic techniques and extrapolation from the pilot study, it is estimated that the data processing effort associated with producing a state-wide land use map would consume one or more man-years using ERTS data. The effort above one man-year would depend on the amount of analysis of the data that was involved.

There are also advantages to the use of digital tapes from ERTS as opposed to ERTS photographs. From the tapes the minimum area that can be analyzed is restricted only by the sensors on the ERTS. From the ERTS
photographs, however, for some categories the minimum area that can be analyzed is three to five times as large. In addition, there are a maximum of 16 grey-levels per band in an ERTS photograph, whereas there are 256 grey-levels per band in the digital data. Thus, not only the geometric fidelity but also the spectral fidelity is greater in the digital data.

As an example of the greater utility of the digital data, consider the scene which includes the Fulton Industrial District on the previously furnished printouts (see Figure 2). Vastly more detail is visible in the computer printouts than is available in the ERTS photography.

Additional information concerning the use of ERTS data for all disciplines is given in the "Proceedings of the Symposium on Significant Results Obtained from Earth Resources Technology Satellite - 1," Volumes I-IV, March 5-9, 1973, published by NASA, Goddard Space Flight Center.
VII. SUMMARY AND CONCLUSIONS

The Engineering Experiment Station has demonstrated its capability to process and display in a convenient form remote sensing data from the ERTS. The programs available at the Engineering Experiment Station can also routinely process aircraft remote sensing data and, with minor changes, Skylab data.

The results of this study indicate that the ERTS data and automatic processing of that data may be useful in some disciplines in the State of Georgia. Six land-use categories were identified using only channel four of the multispectral scanner data from ERTS; however, much greater discrimination (twelve or more categories) was realized using all four channels of ERTS data.

It should be noted that a limitation on the application of ERTS data to state problems is its maximum ground resolution of a little over an acre. One purpose of this pilot study was to demonstrate the capability at EES to process remote sensing data. Since the only multispectral data that was available was ERTS data, this data was processed as an example. When additional data such as high altitude MSS aerial data is obtained, it is expected that the computer processing of that data with its excellent resolution will allow a much greater level of discrimination than the aerial photos from that altitude. Thus, the machine processing of high flight MSS data may be applicable to a multitude of problems in Georgia that ERTS data cannot solve because of its resolution.

The cost per unit area of unsupervised classification is shown to increase linearly as a function of the number of clusters (land-use categories) for any subset of data channels. A comparison of this cost to previous supervised classification costs indicates at least a factor of three cost savings by using the unsupervised approach with the same number of classes. The accuracy is high for both of the methods. However, it appears that the accuracy of the supervised classification is slightly higher than the accuracy of the unsupervised classification.
Even though there is a cost advantage to using fewer than four channels of data, the Engineering Experiment Station recommends that in future studies all four ERTS channels be used because of the superior discrimination afforded by using four channels of data.