EPA/620/R-94/020
March 1994

# EMAP

# Chesapeake Bay Watershed

# Pilot Project

Project Manager - L. Dorsey Worthy
Advanced Monitoring Systems Division
EMSL Las Vegas, NV

EMAP Research and Assessment Center
Environmental Monitoring and Assessment Program
Office of Research and Development
U.S. Environmental Protection Agency
Research Triangle Park, NC 27711

# NOTICE

The information in this document has been funded wholly or in part by the U.S. Environmental Protection Agency. It has been subjected to the Agency's peer and administrative review process. It has been approved for publication as an EPA document. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

and industrial activities, all potentially affect the condition of the watershed and the Bay. The Bay extends 333 km (200 miles) from its northern extent at Susquehanna Flats, Maryland, south to its southern extent between Cape Henry and Cape Charles, Virginia. The Bay is fed by 48 major rivers and more than 100 smaller tributaries, draining 165,800 km$^2$ (64,000 mi$^2$). Watershed tributaries reach north to Cooperstown, New York, south to Norfolk, Virginia, west to the Appalachian and Allegheny Mountains, and east into the State of Delaware. Many Bay monitoring activities, including non-point source water-quality assessments and process modelling, require recent land cover/use inventory data to adequately assess the status and trends of this dynamic watershed and Bay.

## OBJECTIVES

The major objective of the EMAP Chesapeake Bay Watershed Pilot Project was the development and testing of methods for producing detailed digital land cover and land use data over large geographic areas using commercially available satellite imagery. The land cover/use map generated by this project is intended to be used in the CBPO non-point source water quality model and will replace the currently used and outdated map. This project was also intended to compliment other similar remote sensing data products being generated for the Chesapeake Bay area. These include the National Oceanic and Atmospheric Administration (NOAA) Coastal Change Analysis Program (C-CAP) Chesapeake Bay Project, and efforts by both the states of Maryland and Virginia to map Bay area resources

A secondary goal of this project was the evaluation of the applicability of the EMAP hexagon sampling frame. This frame consists of a systematic, uniformly distributed grid of continuous 640 km$^2$ hexagonal units covering the United States. Smaller 40 km$^2$ Stage Two hexagons are centered within each of these larger hexagons. These smaller hexagons have been recommended as standard sampling units for EMAP resource groups.

Results of this and similar projects have led to the formation of the Multi-Resolution Land Characteristics (MRLC) consortium, an interagency cooperative effort established to pool expertise and defray costs of production of a satellite based digital land cover/use database for the conterminous United States. The MRLC consists of representatives from USGS EROS Data Center (EDC), USGS National Water Quality Assessment Program (NAWQA), NOAA Coastal Change Assessment Program (C-CAP), EPA North American Landscape Characterization Project (NALC), and EMAP-LC. This project will, therefore, serve as a model for large-scale projects using remote sensing for environmental monitoring.

## DATA SOURCES

Landsat Thematic Mapper (TM) digital multispectral imagery was selected as the source image data due to its relatively high spatial and spectral resolution. Spatial resolution refers to the level of spatial detail inherent in an image. Spectral resolution relates to the widths of wavelength bands and positions of bands in the electromagnetic spectrum measured by the sensor. Sixteen TM scenes from the Landsat 5 satellite were used in the project. Figure 1-2 illustrates the positions of the scenes across the watershed.

The TM sensor system has the capability to differentiate reflected and emitted electromagnetic radiation (EMR) in seven discrete wavelength bands or channels. These capabilities include spectral sensitivity in the visible (three bands), near and mid-infrared (three bands), and thermal portions of the EMR spectrum. These band combinations provide quantitative spectral values which can be used to discriminate land cover/use types.

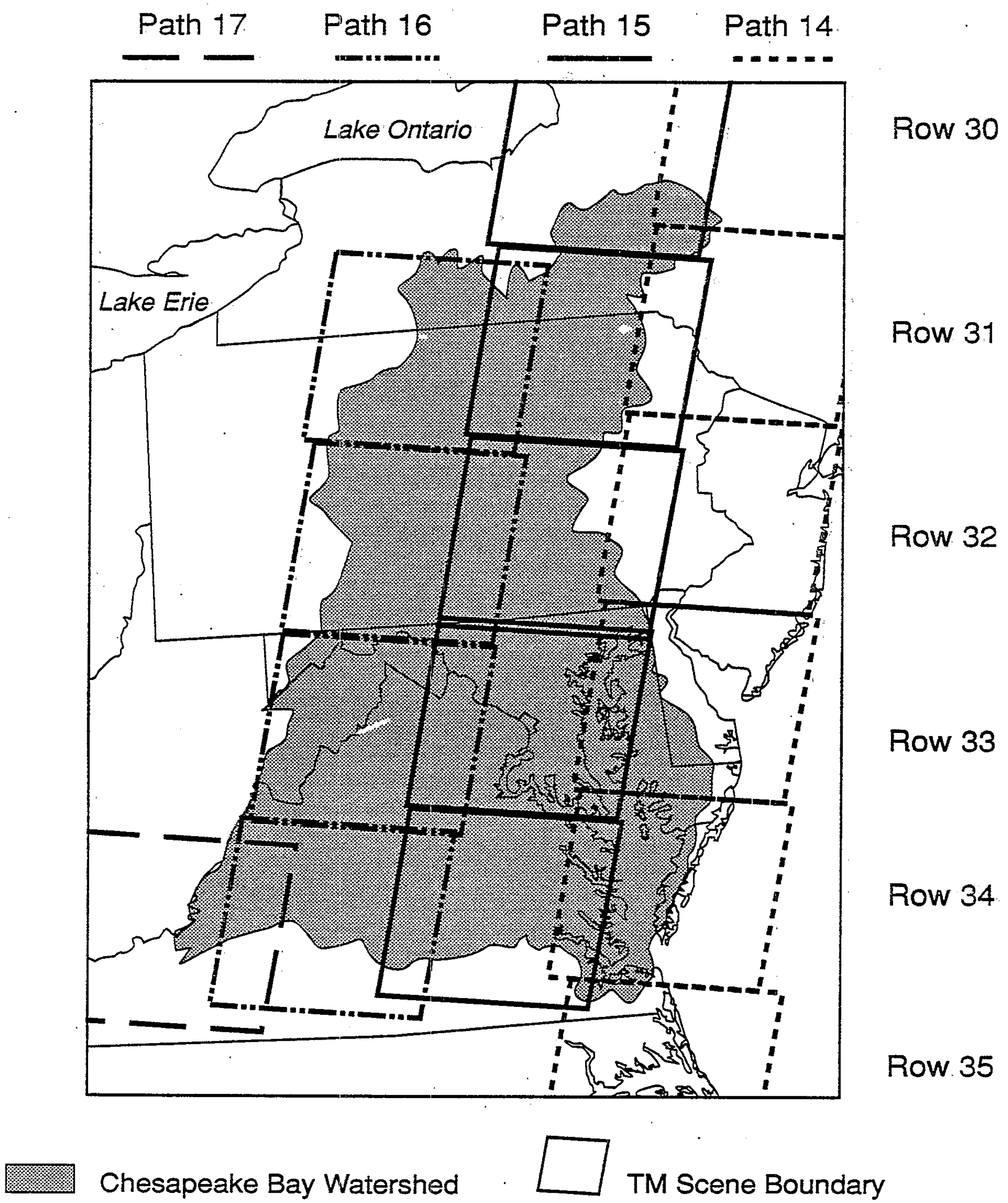


Figure 1-2. Chesapeake Bay Watershed Study
(Location of Landsat Thematic Mapper Scenes)

# CHAPTER 2

# CLASSIFICATION SYSTEM

The EPA EMAP-LC participated in the development of an interagency ecological land cover and land use classification system. Collaborators included EPA EMAP, US Geological Survey (USGS), US Fish and Wildlife Service - National Wetlands Inventory (USFWS-NWI), National Oceanic and Atmospheric Administration (NOAA) - National Marine Fisheries Service (NMFS), University of Delaware, Oak Ridge National Laboratory, Salisbury State University, and Florida Department of Natural Resources. The system was developed to be hierarchial, with broad categories at the basic level, and increasing detail at subsequent levels.

## CLASSIFICATION CRITERIA

Portions of Anderson (1971), Anderson *et al.* (1976) and Cowardin *et al.*, (1979) were incorporated into the effort. These systems were used to take advantage of their strengths, to avoid duplication of effort, to provide commonality between the current system and generally accepted systems, and to facilitate understanding of the final classification product by a variety of users. The following criteria, modified from Anderson (1971) and Anderson *et al.* (1976), were included in the development of this system.

1. The minimum level of interpretation accuracy in the identification of land use and land cover categories should be 85% correct overall.
2. The accuracy of interpretation for the several categories should be essentially equal.
3. Repeatable or repetitive results should be obtained from one interpreter to another, and from one time of sensing to another.
4. The classification system should be applicable over extensive areas.
5. The classification system should permit vegetation and other types of land cover categories to be used as surrogates for land use activity.
6. The classification system should be suitable for use with remote sensor data obtained at different times of the year.
7. Subcategories (finer detail) obtained from ground surveys or from the use of larger scale or enhanced remote sensor data can be used effectively.
8. Categories can be and should be aggregated when appropriate.
9. Current and future land use data can be compared.
10. Multiple uses of land can be recognized.

Initially, 19 categories were selected, including three forest categories, four wetlands categories, and two agricultural categories. Inconsistencies and low accuracies resulted in the eventual aggregation or elimination of many of these original categories, resulting in the final six category data set. Table 2-1 shows the original categories and the final database categories to which they were merged.

**TABLE 2-1. Original and Final Land Cover/Land Use Categories**

| Level 0 | Original Level 1 | Original Level 2 | Final Database Category |
|---|---|---|---|
| UPLAND | 1 DEVELOPED | 11 HIGH INTENSITY | 11 DVL. HIGH INTENSITY |
| | | 12 LOW INTENSITY | 12 DVL. LOW INTENSITY |
| | 2 CULTIVATED | 21 WOODY | 30 WOODY |
| | | 22 HERBACEOUS | 20 HERBACEOUS |
| | 3 GRASSLAND | 31 HERBACEOUS | 20 HERBACEOUS |
| | 4 WOODY | 41 DECIDUOUS | 30 WOODY |
| | | 42 MIXED | " |
| | | 43 EVERGREEN | " |
| | 5 EXPOSED | 51 SOIL | 40 EXPOSED |
| | | 52 SAND | " |
| | | 53 ROCK | " |
| | | 54 EVAPORITE DEPOSITS | " |
| | 6 SNOW & ICE | 61 SNOW & ICE | - NONE - |
| WETLAND | 7 WOODY | 71 DECIDUOUS | 30 WOODY |
| | | 72 MIXED | " |
| | | 73 EVERGREEN | " |
| | 8 HERBACEOUS | 81 HERBACEOUS | 20 HERBACEOUS |
| | 9 NONVEGETATED | 91 NONVEGETATED | 40 EXPOSED |
| WATER AND SUBMERGED LAND | 10 WATER | 100 WATER | 60 WATER |

The original intent was to use the USFWS National Wetlands Inventory (NWI) digital coverages to define wetlands. However, the NWI maps for the Chesapeake Bay watershed were incomplete. Therefore, an attempt was made to identify wetlands using the clustered spectral data. Early verification efforts showed these classifications to be inaccurate, and they were merged with identifiable component categories. In addition, pasture, grassland, and cultivated herbaceous categories were found to be indistinguishable and were merged. The mixed woody category accuracy was also unacceptable, and required the merging of all woody categories, since it could not be merged to evergreen or deciduous categories. The final six categories represented the greatest thematic detail meeting EMAP-LC Data Quality Objectives (DQO).

## DEFINITIONS

The remainder of this section provides descriptions for each of the final category or class in the classification system for the Chesapeake Bay Pilot study. The definitions are a combination of the project comments and text from Anderson (1971) and Anderson *et al.* (1976): *A Land Use and Land Cover Classification System for Use with Remote Sensor Data.*

Land use is defined here as spatial divisions based on anthropomorphic activity or utilization. Land use may or may not be identifiable from above the earths surface, and generally requires *a priori* knowledge of the feature. Examples include agricultural orchards and row crops, urban and residential areas, commercial zoning, and transportation networks.

Land cover is defined here as the substance existing on and visible and recognizable from above the earth's surface. This generally requires limited or no *a priori* knowledge of the feature for recognition or differentiation. Examples include vegetation, exposed or barren land, water, and snow and ice.

### UPLAND DIVISION

The Upland Division includes all categories other than open water. It is divided into four Level 1 categories: Developed, Herbaceous, Woody, and Exposed Land.

#### 10 DEVELOPED

This category is composed of areas of anthropogenic use, with much of the land covered by structures and other artificial and impervious surfaces. This category does not include cultivated or other agricultural land.

*Included in this category are cities; towns; villages; strip developments along highways; transportation, power, and communications facilities; and areas such as those occupied by mills, shopping centers, industrial and commercial complexes,* *and institutions that may, in some instances, be isolated from urban areas.* (Anderson *et al.*, 1976).

Developed Lands are divided into two Level 2 groups: 1.1 - High Intensity (Solid Cover) and 1.2 - Low Intensity (Mixed Cover).

### 11 HIGH INTENSITY DEVELOPED

This class refers to built-up urban areas, and it contains areas primarily composed of a solid cover of human-made materials. They contain few mixed (human-made materials and vegetation) areas. They may have a variety of land uses. A significant portion of the surface is covered by concrete, asphalt, and other artificial materials, and contain little vegetation. Examples are apartments, large buildings, shopping centers, factories, and industrial areas. This class often occurs in city centers. Some major highway systems are also included in this category.

### 12 LOW INTENSITY DEVELOPED

The Low Intensity class refers to areas that contain a mixture of human-made materials and other land cover/use resources. They are typically single family housing areas, and are often called suburban or residential. The category also contains roadways where pixels consist of a mix of highway materials and other resources. As the intensity of the human-made materials decreases, this category grades into 2.0-Herbaceous, 3.0-Woody, and other appropriate categories.

## 20 HERBACEOUS

The Herbaceous category includes lands covered by either natural or managed herbaceous cover, including agricultural row crops and pasture. The Herbaceous class is defined as land where the potential natural vegetation is predominantly grasses, grass-like plants, and forbs. Also included in this class are lawns and other landscaped grassy areas such as parks, cemeteries, golf courses, and road and highway right-of-ways.

## 30 WOODY

The class Woody refers to land covered by shrubs or trees. This includes any species that has an aerial stem which persists for more than one season, and in most cases a cambium layer for periodic growth in diameter (Harlow and Harrar, 1969). The Woody category includes deciduous, evergreen, and mixed trees, and shrub-scrub vegetation.

### 40 EXPOSED

Exposed land includes naturally occurring areas that have limited ability to support plant life, or have been burned, cleared, or disturbed. In general, these areas are covered with soil, sand, or rocks. Vegetation, if present, is widely spaced. Naturally barren areas contain less than one-third vegetation or other cover. The exposed areas may be transitional or developed. These areas include lands cleared for a variety of purposes, e.g., construction of new buildings, quarries, landfills, gravel pits, strip mines, etc.

## WATER AND SUBMERGED LAND DIVISION

The Water and Submerged Land Division consists of areas of open water and land areas covered by water. It has one Level 1 category: Water.

### 60 WATER

Water is defined as areas that contain standing shallow and deep water, either natural or human-made. Water habitats include environments where surface water is permanent and often deep, so that water, rather than air, is the principal medium within which the dominant organisms live. Human-made areas of water would include reservoirs, impoundments, dikes, ponds, and canals.

# CHAPTER 1

# INTRODUCTION

The Chesapeake Bay Watershed Pilot Project was initiated to develop and test methods for generating digital land cover and land use products from satellite based remotely sensed imagery. These methods and the resulting data products were intended to fulfill specific program requirements of the Landscape Characterization component of the Environmental Monitoring and Assessment Program (EMAP-LC) and the Chesapeake Bay Program Office (CBPO) of the U.S. Environmental Protection Agency (EPA). This report presents a standardized methodology for digitally classifying remote sensing data over relatively large and diverse geographic areas.

This report describes the methodology used to produce the land cover/use map of the Chesapeake Bay watershed. Chapter 2 discusses the development of the classification scheme and gives a general description of the classes. A technical description of all processing and quality assurance/quality control methods, including spatial and thematic accuracy assessments, is presented in Chapter 3. Chapter 4 presents results and summary statistics, and discusses difficulties, accomplishments, and recommendations for future research. Conclusions are presented in Chapter 5. The Chapters are followed by References, Acknowledgements, and Appendices.

## BACKGROUND

EMAP is an innovative research, monitoring, and assessment effort designed to report on the condition of the nation's ecological resources, including surface waters, agroecosystems, arid ecosystems, forests, and estuaries. This information, when combined with data from other monitoring programs, will provide a comprehensive view of the effectiveness of national environmental policies.

EMAP-LC is responsible for providing the geographic element of "condition," using consistent national methods now under development and testing. These methods are intended to provide a comprehensive, consistent, and statistically valid nationwide land cover/use product to assist in the overall EMAP effort. The use of satellite imagery to derive this coverage is viewed as the most cost effective and achievable approach.

The Chesapeake Bay Program Office has environmental stewardship over the nation's largest estuarine complex, a more than 165,800 km$^2$ (64,000 mi$^2$) watershed (Figure 1-1). A burgeoning population, combined with extensive forests, agricultural,


Figure 1-1 Location map showing the Chesapeake Bay Watershed.

CONTENTS

## LIST OF FIGURES



## LIST OF TABLES

# CHAPTER 3

# METHODOLOGY

The development of an overall methodology for executing a large area TM land cover/use classification was a primary goal of this project. A major effort went into the development and documentation of the techniques, taking into consideration the large number of TM scenes to be used and requirements of the various users within EPA and EPA-EMAP. The procedures and documentation were intended to facilitate efforts in similar large area or small-scale projects.

This section describes analytical procedures used in processing the Landsat TM data to produce land cover/use digital maps of the Chesapeake Bay watershed. This section is divided into five major topics: Quality Assurance and Quality Control; TM Band Selection; Classification Techniques; Accuracy Assessment; and Final Land Cover/Use Generation. Each section describes issues which relate to the major topic, followed by a discussion of the decision processes used to select the final methodology. Alternative approaches are discussed as they relate to methodology decisions and do not represent an exhaustive discussion of all available techniques.

The technical background of the methodology, including quality assurance/quality control procedures, is presented here. Detailed steps in data processing, including all tracking forms, are contained in Appendix A. A diagram of the image classification methodology is shown in Figure 3-1 and is discussed in this section of the report. A step-by-step instruction guide for data handling and processing is also found in Appendix A.

## QUALITY ASSURANCE AND QUALITY CONTROL PROCEDURES

A standard definition of the QA/QC terminology was adopted to avoid confusion. Taylor (1987) defines QA and QC as follows:

> *Quality Assurance: A system of activities whose purpose is to provide the producer or user of a product or a service the assurance that it meets defined standards of quality with stated level of confidence.*

> *Quality Control: The overall system of activities whose purpose is to control the quality of a product or service so that it meets the needs of users. The aim is to provide quality that is satisfactory, dependable, and economical.*

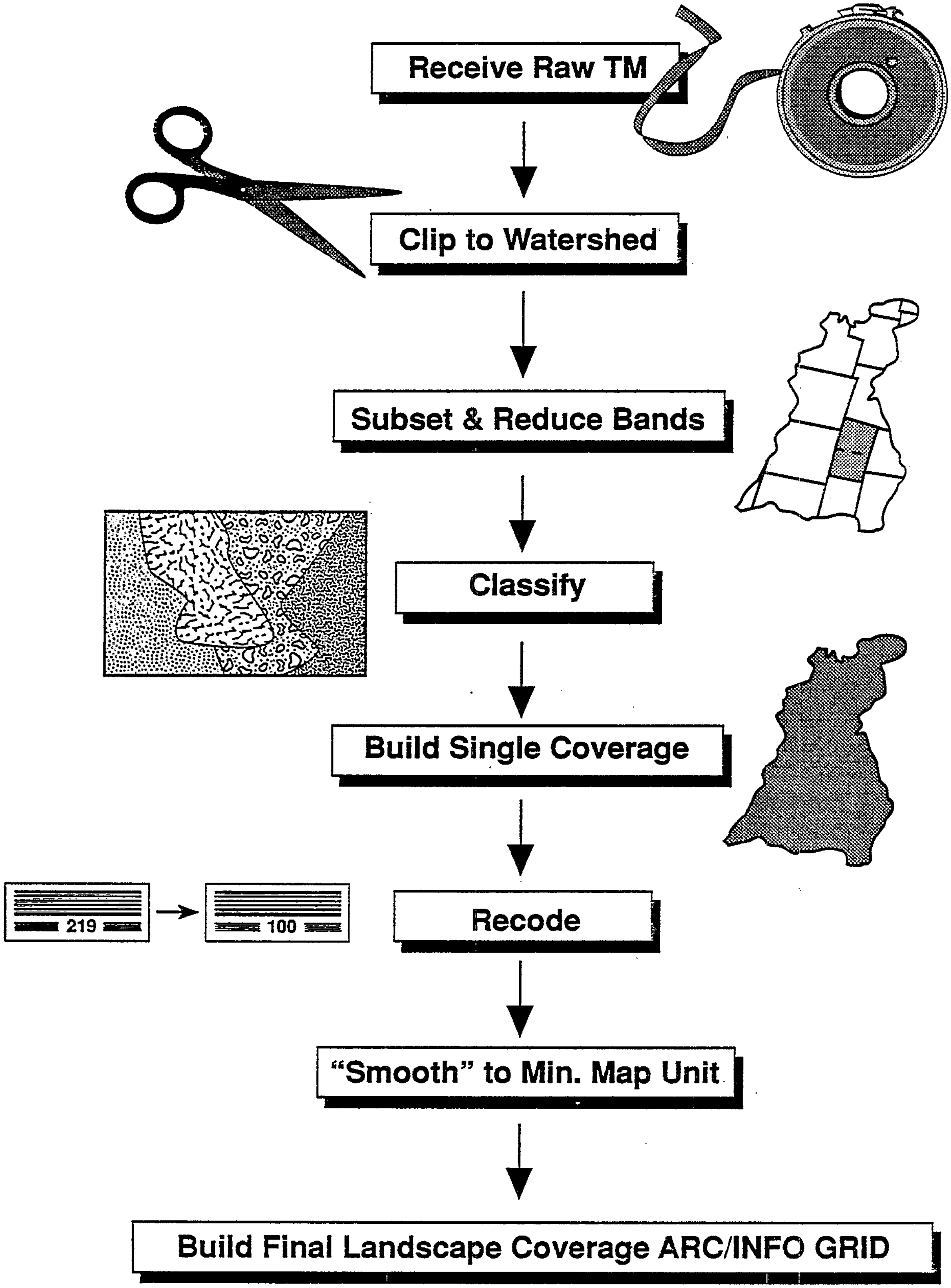


Figure 3-1 Flow chart illustrating the major phases in the Chesapeake Bay Watershed Pilot Project.

Production of all elements of the project, from receiving data, through evaluation of the final data classification, followed standardized and documented QA/QC procedures. The strategies and procedures encompassing management, personnel, problem areas, corrective actions, and products are discussed. The six major topics covered are: User Instructions and Training; Tracking Forms; Receiving TM Data; Spatial Accuracy Tests; and Combining and Edge Matching Classified TM Data.

**User Instruction and Training**

To complete a project of this magnitude, a trained staff and consistent procedures and analysis techniques were required. Standard Operating Procedures (SOP's) ensured that all TM scenes within the watershed were processed alike. Examples of these SOP's, along with the tracking forms, are found in Appendix A.

SOP's contained detailed, step-by-step procedures to help the analyst classify TM data from the project's beginning to end. The instructions included computer commands, data handling operations, file-naming conventions, file storage and retrieval, input parameters for statistical and analytical programs, and tracking forms. The four instruction guides in Appendix A are: Subset TM Scenes Instructions; Classification Instructions; Subset Editing Instructions; and Combining Subsets into the Master-Raster. These instruction guides ensure that all stages of data processing were carried out consistently for all data sets and by all analysts.

Project analysts were knowledgeable in remote sensing techniques. They also had a working knowledge of the hardware and software used in the project, including SUN and Silicon Graphics workstations, and UNIX, ERDAS, and Arc/Info software. Each analyst was familiar with the SOP's specified for the project.

**Tracking Forms**

Among the most important QA/QC elements developed for this project were the data tracking forms. Examples are included in Appendix A. The primary function of the tracking forms were to allow all steps to be reproduced, errors to be traced, and the results to be compared between different data subsets. Tracking forms were used for each of the following steps in processing: receiving TM data; visual inspection of TM data; spatial accuracy assessment; clipping TM scenes to watershed; subsetting TM scenes; classification of subsets; cluster labeling; combining files into the master-raster; editing subsets; and combining subsets into the entire file.

The tracking forms provided a mechanism so that errors could be traced to specific files, parameters, or procedures. The source of error could then be identified and corrected by repeating the processing steps and modifying the analysis. It was also instructive for the analysts to be able to compare results on tracking forms from different data sets. This comparison led to a better understanding of the procedures and how the results compared across different ecological regions.

## Receiving TM Data

A QA/QC check was performed to evaluate the spatial, radiometric, and general image quality of each Landsat TM scene as it was received. TM scenes were accompanied by a description sheet and header file associated with the digital TM data from EOSAT Corporation, the commercial landsat vendor. This header information contained unique parameters associated with each TM scene. Parameters were recorded for the purposes of documentation and later use in the project.

After the initial information from a TM scene was recorded, the quality of the product was checked visually. The scenes were viewed to note any clouds and cloud shadows, data dropouts, striping, or other irregularities in the image. This information was also recorded on the tracking form (Appendix A). If the percentage of cloud cover was greater than 10%, the scene was rejected and returned to EOSAT for a replacement.

## Spatial Accuracy Tests

An accuracy test was performed to check the spatial fidelity of the georeferenced TM data. The TM sensor collects data at a nominal spatial resolution of 28.8 m by 28.8 m. Data for this project were resampled by EOSAT to 25 m. As part of the reporting procedure EOSAT provided documentation with each TM scene which listed the Root Mean Square (RMS) error in point and linear coordinates, along with all points used or rejected in the resampling process.

A documentation check and a data validation process were used to confirm that the spatial accuracy of the Landsat TM data met the error specifications of ±15m RMS. The documentation check began with a review of the QA/QC procedures and the RMS error reported by EOSAT for the geocoded scenes. The project staff checked the reported RMS and the spatial accuracy of the TM scenes by using standard georeferencing methods. Points on the TM image were compared with the identical points on USGS topographic maps. Four 7.5-minute USGS quadrangle maps for each TM scene were randomly selected for validation; eight points that were identifiable in the image and on each map were digitized, and variation between image and map coordinates were determined.

The selection process for the 7.5-minute quadrangles was based on a random selection from a grid of 64 maps covering each TM scene. Each of these quadrangle areas was identified by an alphanumeric code related to the latitude/longitude of the southeast corner of the grid of 64 maps. Factors considered before the random selection included the date of each 7.5-minute quadrangle map and availability of identifiable points in the image and the map. There were a few 7.5-minute quadrangles that had not been updated in the 1980's, and the dates on these maps ranged from the 1950's to 1970's. In some of the rural and mountainous areas the availability of identifiable points posed a problem.

If the randomly selected 7.5-minute quadrangle did not contain a sufficient number of identifiable ground control points, then an adjacent 7.5-minute quadrangle was selected. The new quadrangle was chosen horizontal to the original quadrangle. If there were insufficient identifiable points on an adjacent horizontal quadrangle, then an adjacent vertical quadrangle was selected. If this failed to produce a sufficient number of identifiable points a diagonal quadrangle was chosen. This process was expanded and repeated until an adequate combination of map and image points could be identified. TM image coordinates were identified for each identifiable map point, and the information recorded on the Spatial Accuracy Assessment Tracking Form.

The map was then positioned on a digitizing table and test points selected to determine the accuracy of the map setup. Again this information was recorded on the tracking form. Map setup errors were generally found to be less than 7 m. Those areas with higher errors usually corresponded to older maps.

A difference (±) between the image and map coordinates was determined for each point identified in a TM scene. Also for each scene a mean difference was determined in the x and y directions. A graph was plotted for each TM scene indicating the difference between the image and map coordinates. This information was recorded in the project tracking book. The Spatial Accuracy Test indicated that EOSAT had met the required contract specifications. Overall, the spatial accuracy of the TM scenes was within ± 15 m.

## TM BAND SELECTION

Multidimensional data sets, such as TM data, contain redundant spectral information. This redundancy results from the fact that spectral reflective properties of land features in different parts of the electromagnetic spectrum are similar. Employing more than four TM bands for spectral feature extraction does not necessarily increase the clustering capability of computer-based classifications (Latty and Hoffer, 1981; Stenback and Congalton, 1990). Spectral clustering for TM land cover/use mapping typically utilizes band combinations which include at least one band from the visible (0.4 - 0.7 $\mu$m), near infrared (0.7 - 1.3 $\mu$m) and middle infrared (1.3 - 3.0 $\mu$m) spectral regions (Nelson *et al.*, 1984). The thermal band (TM band 6), due to it's lower spatial resolution and feature calibration requirements, is generally not used for spectral cluster development.

In order to determine the best band combination for use in classification algorithms, three techniques were investigated: Principal Components Analysis (PCA); Analysis of Correlation; and Optimum Index Factors (OIF). The PCA method was investigated, but was not used. Each scene would have had a different PCA transformation, making interpretation of composite images difficult, since similar resources would look different from scene to scene. In addition, rare land resources may not have been distinguishable until the higher components were analyzed.

Similarities between bands were measured by analyzing covariance and correlation matrices. Bands having high correlation values are considered to contain redundant information. The visible bands (1, 2, and 3) of the Chesapeake TM scenes were shown to be highly correlated (most were over 90%). This is typical for TM data because most surface features have similar reflectance properties in these wavelengths. Bands 5 and 7 were also highly correlated (most over 85%). Band 4 showed little correlation with the other bands.

Within-band variance provides an indication of the amount of information a band contains. A high variance results from different land-cover types reflecting various amounts of energy within that wavelength. It was desirable to use bands with high variance because they are more likely to provide high contrast between land features, making it easier to distinguish and segregate land cover/use types.

Chavez *et al.* (1982, 1984) described a technique for calculating Optimum Index Factor values (OIF) to determine best band combinations. The OIF combines the variance and covariance so that higher values result when within-band variance is high, and between-band covariance is low. Combinations of bands with higher OIF values are desirable because the amount of information (variance) is high, whereas the amount of redundancy (covariance) is low. This was the technique selected.

Two four-band combinations produced the highest OIF values for the majority of the selected TM scenes: bands 1, 4, 5, and 7; and bands 3, 4, 5, and 7. Band 1 was not included in the final band combination for the analysis because it measures a wavelength (0.45 $\mu$m - 0.52 $\mu$m) that is most affected by atmospheric scattering. Therefore, bands 3, 4, 5, and 7 were used to derive the spectral clusters.

## CLASSIFICATION TECHNIQUES

Automated spectral classification of remotely sensed digital data is performed using statistical analysis. There are two primary methods of automated statistical image classification: supervised and unsupervised. The two methods differ in the way that groupings of digital values are identified in spectral space. The TM bands describe a multi-dimensional data set where the digital values for a pixel describe its location in spectral space. Within spectral space, pixels with the same reflective properties will group together. The first step in the classification process is to identify statistics that differentiate these groupings.

In a supervised classification, the identity and location of representative land cover/use types are known *a priori*, through a combination of field work and/or the analysis of aerial photography and maps. These areas of known land cover/use, called training sites, are identified on the image. Multivariate statistical parameters (means, standard deviations, covariance matrices, correlation matrices, etc.) are calculated for each training site and are used to describe the land-cover categories in spectral space (Jensen, 1986). Enough training sites must be established so that land cover/use classes are identified for all spectral reflective conditions present in the scene. This requires identification of training sites for all existing land cover/use features under all topological, hydrological, and physiological conditions which might alter the spectral reflectance of the feature.

In an unsupervised classification, the identities of land-cover types to be specified as classes within a scene are not generally known *a priori*, either because ground truth is not available or surface features within the scene are not well defined (Jensen, 1986). An unsupervised clustering algorithm examines the spectral values of the imagery and statistically groups similar values into spectral clusters. A second algorithm then determines which cluster best represents the combination of spectral values for each individual pixel and assigns that cluster value to the pixel. It is then the responsibility of the analyst to assign a land cover/use label to each cluster using available reference materials.

An alternative to pure supervised or unsupervised classification is to use a combination of the two. A test was performed to determine whether a first-pass supervised classification could eliminate large, easily identified land features prior to an unsupervised classification. It was found that available reference material only allowed for identification of easiy recognized, very distinct features. These features also were easily differentiated with an unsupervised clustering algorithm. It was also found that a prohibitive number of training sites were required to adequately cover the widely varying spectral signatures of the classes.

The selection of training sites had to be made using the analysts' skills in manual interpretation of the TM data, since available reference material were collected at different times of the year. Analysts not familiar with the study area had a difficult time choosing representative training sites. The additional time spent choosing training sets was not worth the minimal gain.

Several alternative unsupervised clustering routines were considered and tested. The four criteria used to evaluate the options were: (1) validity of theory, (2) coverage of spectral space, (3) separability of resources, and (4) ease of use. A full discussion of all alternative unsupervised computer programs is beyond the scope of this report. After evaluation of available software, none of the routines was considered to be entirely adequate. Additional research efforts were made to refine an existing clustering routine. Therefore, a two-step clustering technique was developed for use in this project.

The two-step process clustering is illustrated in Figure 3-2. Cluster statistics were gathered within the scene subset. Each pixel was then assigned to a cluster using the allocation program described below. The resulting image was examined by the analyst and the spectral clusters were assigned to a land cover/use category. There were, inevitably, some clusters that represented more than one land cover/use type or remained unclassified. These "confusion" clusters were statistically analyzed again to break each cluster into several better defined spectral clusters. Pixels belonging to the "confusion" clusters were then assigned to one of these new clusters. These refined clusters were examined by the analyst and assigned to a land-cover category.

### Subset TM Scenes

A total of 16 Landsat TM scenes were required to cover the Chesapeake Bay watershed. Each TM scene included areas of overlap with its neighboring scenes and some included areas outside the watershed. Boundary scenes were clipped to eliminate those portions lying outside the watershed using a Geographic Information System (GIS) file provided by the Chesapeake Bay Liaison Office. Visual interpretation of the quality of the imagery was used to select which TM scene would be used in areas of overlap.

Computer processing limitations required that TM scenes be subset to practical image file sizes. A single TM scene contains approximately 38 million pixels. However, computer hardware available at EMSL-LV was limited to a maximum instantaneous display of 1024 X 1024 pixels. This restriction could be overcome by displaying an image at a reduced scale. However, the resulting image would no longer contain the full resolution of the original data, and would limit the ability to identify and label the clustered imagery. Limitations associated with the file server and backup facilities further restricted the ability to move, copy, analyze, backup, and otherwise maintain larger data files.

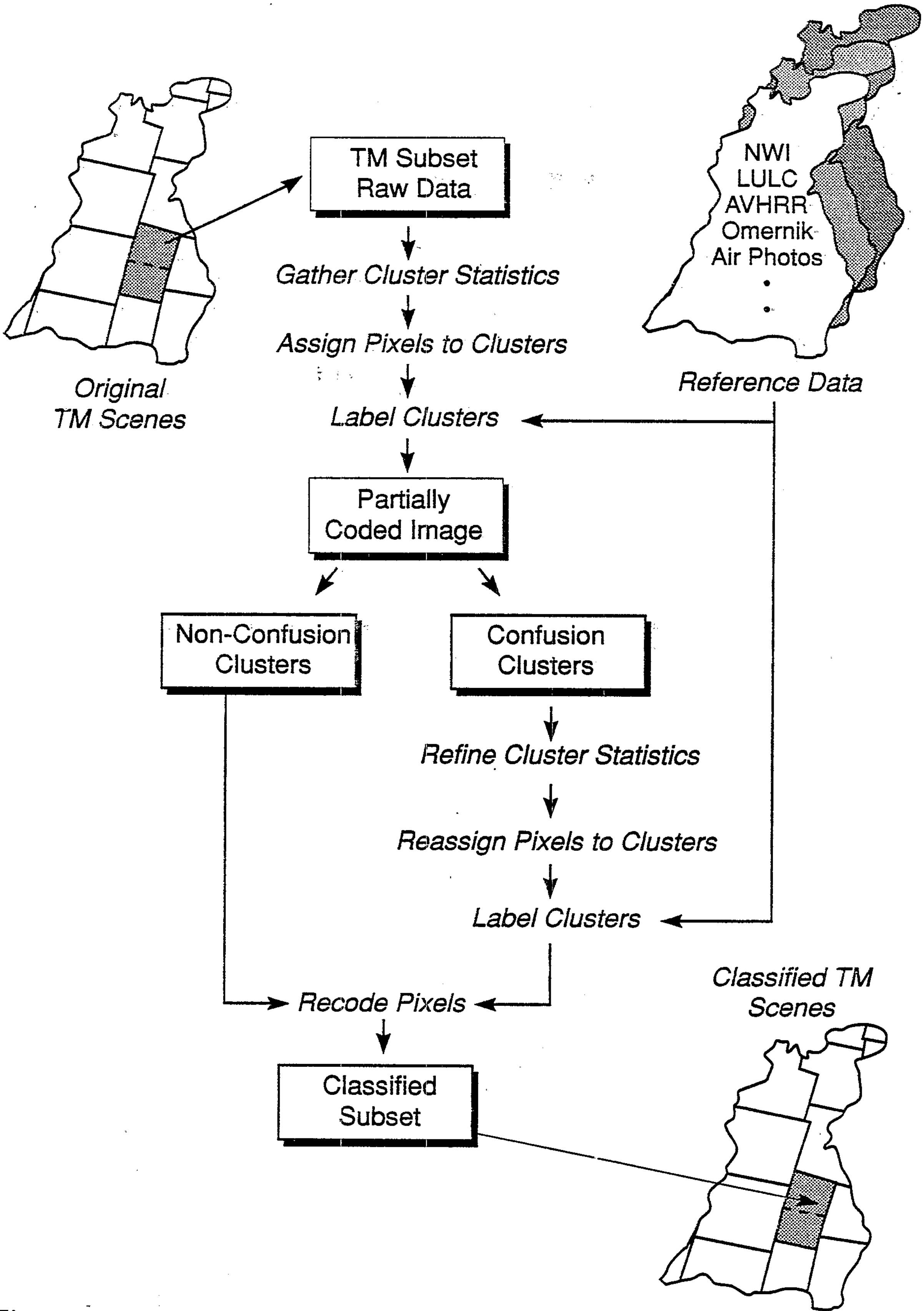


Figure 3-2. Flow chart outlining the steps in the image classification analysis.

In addition to the above, single land cover/use categories may have several spectral signatures depending on the characteristics of their physical location, satellite viewing angle, soil conditions, slope, aspect, atmospheric effects, etc. Larger data sets would generally contain a greater number of land cover/use types and consequently more spectral signatures. Also, large clusters tend to obscure spectral variability which may have otherwise generated seeds for new clusters. Therefore, an image subset of a quarter TM scene was chosen as a maximum file size for analysis.

**Defining Spectral Clusters**

Cluster statistics are the foundation of any unsupervised spectral classification and must accurately describe spectral information found within the image under analysis. The clustering programs used for this project were developed by programmers at EMSL-LV (Weerackoon and Mace, 1990). The algorithm used a non-overlapping 3x3 moving window as a basis for defining cluster statistics. It was desirable to collect the statistics of windows containing one cover type and reject windows containing more then one cover type. This was done by analyzing the variance between pixels within a window. Windows containing more than one land cover/use type should have a higher variance than those representing a single land cover/use type. Therefore, windows with lower variance were used to develop cluster statistics.

A separate routine was used to define maximum variance threshold levels for acceptable windows. This generated a report which showed the discrete cumulative distribution functions of the ceiling integer values of the variances for each band. Analysts identified variance threshold values for each band corresponding to the 50% cumulative level. The variance thresholds were then used as input parameters to the clustering program. If the variance of a window in any channel was greater than the user-defined threshold for that channel, the window was rejected. Lower threshold values resulted in fewer acceptable windows; higher thresholds resulted in more acceptable windows.

The clustering algorithm used the accepted windows to build spectral clusters. The first acceptable window became the first spectral cluster. From then on, each acceptable window was checked against the previously defined clusters. The closest cluster measured in Euclidean distance (in spectral space) to the new window was identified. If the mean of the new window was further than two standard deviations from the closest cluster, the window became a new cluster. If the mean of the new window was between one and two standard deviations from the closest cluster, it was rejected. If the mean of the window was within one standard deviation of the closest cluster, the statistics of the window were merged with that of the cluster. By analyzing the entire scene and checking all acceptable 3X3 windows, a set of initial spectral clusters was identified.

Because of the complexity of the TM scenes there were more initial clusters than were practical for an analyst to evaluate. Further statistical analysis was performed to reduce the number of spectral clusters. The user was able to specify the number of desired final clusters before running the program.

The variance threshold values, used to evaluate windows above, were used again in this final cluster merging step. An iterative process beginning with 0.1 times the standard deviation (the square root of the threshold variance) for each band was used. All distances between clusters (in spectral space) were checked. If any cluster distance was found to be within 0.1 times the standard deviation, the two clusters were merged.

The next iteration compared 0.2 times the standard deviation between the cluster distances, and the next iteration used 0.3 times the standard deviation. The iterative process of merging clusters stopped at 1.5 times the standard deviation. If, at any time during this iterative process, the number of clusters was less than or equal to the desired number of clusters specified by the user, the program stopped. The remaining clusters became the final set output by the program. If the number of clusters after the iteration using 1.5 times the standard deviation was still larger than specified, the program only output clusters built from the greatest number of pixels and deleted the remaining clusters..

## Image Classification - Pixel Allocation

The final step in the image classification, after generation of the spectral clusters, was allocation of individual pixels to the cluster class. The classification or allocation step assigned each pixel within the subscene to the cluster to which it had the highest likelihood of being a member. The most common and generally preferred method of pixel allocation is the maximum likelihood classifier (Richards, 1986). Other methodologies, such as nearest neighbor and parallelepiped, can be employed where computer time is at a premium. This project used an optimized maximum likelihood routine (Weerackoon and Mace, 1990) and relatively fast computers, minimizing processing time to the point that this was not a limiting factor.

## Cluster Labeling

During the labeling process, an analyst assigned a land cover/use class name to each of the spectral clusters. The use of reference data was necessary to aid the analyst in labeling the unsupervised clusters. Table 3-1 lists all reference data used by the analysts to label the clusters. The reference data sets fell into two categories,

**Table 3-1. Reference Material Used to Aid Classification Process.**

| Data Set | Format | Scale or pixel size | Notes |
|---|---|---|---|
| USGS Topographic maps | Hardcopy maps | 1:250,000<br>1:100,000<br>1:24,000 | General positional locations<br>Identification of urban areas, water, some wetlands |
| Aerial Photography<br>NAPP<br>NHAP | Hardcopy photos | <br>1:40,000<br>1:58,000 | Primary reference data set<br>Used to identify all land-cover types<br>Acquired NAPP where available |
| Landsat Thematic Mapper image maps | Hardcopy images | 1:100,000 | General positional locations<br>Visual interpretation |
| Agricultural Statistics Bulletins by state | Books | N/A | Indicates potential agriculture types<br>Growing seasons |
| Soil Survey Bulletins | Books and maps | 1:20,000<br>1:15,840 | Indicates soil regime<br>Potential cover types |
| County boundary coverage | Digital | 1 km pixels | Used with summary bulletins listed above |
| USGS 1972<br>Land Use Land Cover | Digital | 100 m pixels | General land cover/use types |
| AVHRR Land Cover Classification | Digital | 1 km pixels | General land cover type based on 1990 data |
| Omernik's Ecoregions | Digital | 1 km pixels | Suggests potential vegetation cover type |

those used to aid in geographic location and those used to identify land cover/use types. The USGS topographic maps and the TM image maps were used mainly to help analysts locate areas of interest. Most other data sets were used for labeling the unsupervised clusters.

The most important reference data set utilized for labeling clusters was aerial photography. A search was made to find a reference data set that was inexpensive, of high quality, and would give a consistent coverage across the study area. Aerial photographs satisfied these requirements. The TM images were acquired during leaf-on conditions of years 1988, 1989, and 1991. Color infrared 1:40,000 scale USGS National Aerial Photography Program (NAPP) photographs, acquired between 1987 to 1990, were available for most of the watershed. In areas of the watershed which lacked NAPP coverage, USGS National High Altitude Photography (NHAP) was substituted. The NHAP photographs were acquired from 1982 to 1985. The NAPP photography was preferred because it had higher resolution and more recent dates of acquisition.

The following types of areas were recognized as requiring aerial photographs for complete classification and labeling: 1) areas with clouds, 2) areas with cloud shadows, 3) areas of physiographic change (i.e.: gypsy moth defoliation), and 4) areas of special interest. After inspecting five scenes (16/34, 16/33, 15/33, 14/34, and 14/33), 130 to 200 areas per scene were identified which required additional information for categorization. However, due to cost constraints, photos were ordered for only 20 to 50 of the most problematic areas for each scene. In addition, stereo pairs were ordered for the center portion of each of the 291 EMAP hexagons located within the watershed. Additional photos, evenly distributed across the 1:100,000 map indexes, were ordered to cover general land cover/use types. This resulted in a photo coverage of roughly 35% of the watershed.

### Cluster Refinement

While labeling clusters, analysts sometimes found confused land surface features within a single spectral cluster. Clustering routines were developed to analyze the raw data of those pixels found in confusion clusters to form new spectral clusters. These programs are analogous to the clustering and maximum likelihood routines explained above. They used the same statistical principles but only work with those pixels identified as belonging to a confusion cluster. This refinement of the confusion clusters usually improved the separability of land cover/use classes.

### Post Classification Editing

After all of the spectral clusters were labelled for all subsets, the data were examined for classification errors. This procedure required analysts to examine every portion of the final classification at full resolution and manually update any areas that were unclassified, cloud covered, or incorrectly classified. The reference materials described in Table 3-1 were used by the analysts to make image corrections. Edits were necessary in areas where cloud cover or data anomalies obscured ground features. Common misclassifications included confusion between gypsy moth damaged forest and herbaceous clusters. Edits were made by manually drawing polygons around areas on the image screen and changing cluster values. In addition, major highways (limited access) were checked and vectors were drawn along the highways where they were not defined by at least a single row of pixels.

**Combining and Edge Matching Classified TM Data**

After all subsets in each of the individual TM scenes were labeled, they were mosaicked together to form a final coverage file. The original TM scenes were georeferenced and most were projected into UTM zone 18 coordinates. Four scenes (16-32, 16-33, 16-34, and 17-34) were referenced to UTM zone 17. All TM subsets were reprojected into the same Albers Equal Area Conic projection coordinates before building the final coverage.

Each subset of classified data was individually mosaicked into the final coverage. As each data subset was added it was checked for edge matching. The edge of each subset was viewed at full resolution to check that all land cover/land use classes were labelled consistently and matched across boundaries of subsets within scenes and between scenes. This assured that all the data along the edges was consistently labeled and that there were no gaps. Occasionally it was apparent that one of the land cover/use types was slightly over or under represented due to a mislabelling of one of the clusters. This usually happened when the cluster was a mixture of two surface types. In these cases the original cluster label was changed.

When visual inspections were made, most features lined up well between subsets. Geographic features crossing scene and subset boundaries matched very well. However, some clusters were found to have been labelled differently in adjacent subsets, even though they contained the same resource. This occurred primarily where clusters contained mixed cover types or there was confusion in interpreting between similar cover features. Whenever inconsistencies were discovered, clusters were reexamined and appropriate labels assigned.

After each TM subset was determined to be acceptable, the cluster values were recoded to the final range of class values. This was a two step process. Following classification and editing, each TM subset contained cluster values ranging from 1 to 150, depending upon the image subset. Each cluster of a common surface type was assigned to a common pixel value. After all TM subsets were combined into a single coverage the entire file was recoded to the class numbers as designated in the classification system in Chapter 2.

### Smoothing

A 1 hectare (ha = 10,000 $m^2$) Minimum Mapping Unit (MMU) was selected for this project. The MMU is the smallest contiguous area incorporated in the final digital product. The TM sensor collects data at a nominal spatial resolution of 28.8 m by 28.8 m. Each TM pixel area is approximately 830 $m^2$. This represents the closest approximation of the 10-to-1 ratio recommended by Congalton *et al.* (1992) and others to minimize propagation of spatial error, and to supply a mapping product which presents uniform and consistent delineation of land cover/use types over large areas. The TM data was smoothed to eliminate small groups of pixels of less than the MMU. Smoothing was executed on the data set after all subscenes were mosaicked into one final file.

A two step process was used to smooth the data: locating areas smaller than the MMU; and reassigning those pixels to other adjacent land cover/use values. The first step identified areas smaller than 1 hectare by identifying groups of adjoining pixels which had the same value and by counting the number of pixels in each group. A Minimum Mapping Unit was specified by choosing the minimum number of pixels a group should have. Groups of adjoining pixels smaller then the Minimum Mapping Unit were eliminated and areas larger than the Minimum Mapping Unit were retained. Adjacent pixels included horizontal, vertical, and diagonal strings of individual pixels, which helped to preserve narrow linear features.

The second step replaced old pixel values in areas smaller than the minimum map unit with values from adjoining pixels. For this study, a "majority rule" was used, where the new pixel value was changed to the value most frequently found in a 3x3 window surrounding the pixel.

## THEMATIC ACCURACY ASSESSMENT

The Chesapeake Bay watershed thematic accuracy assessment involved developing a methodology which was statistically valid and which adequately and efficiently represented each of the classes in the categorized data set. This included determining a representative sample number and deriving a method for comparing verification data to the categorized data set.

A preliminary accuracy assessment of a portion of the Chesapeake Bay watershed categorized imagery was undertaken at Towson State University, Towson, Maryland. This consisted of a limited ground verification of the categorized data for

Baltimore County. This initial assessment indicated certain ambiguities with a number of original class divisions, and resulted in revisions to the original classification scheme. This study underscored the need for a comprehensive accuracy assessment, which was subsequently accomplished.

## Accuracy Assessment Method Design

Accuracy assessment is a comparison of classified and labeled data to some true or known data. Designing the methods for assessing the thematic accuracy of a large categorized data set included identification and acquisition of verification data, design of a sampling scheme, and determining a means for comparing the categorized and verification data sets so that accuracies could be determined.

A common way to express the accuracy of such image or map data is by a statement of the percentage of the map area that has been correctly classified when compared with reference data or ground-truth (Story and Congalton, 1986). In accuracy assessments, the most common way to represent the classification accuracy of remotely sensed data is in the form of an error or confusion matrix as shown in Table 3-2 (Congalton, 1991).

**Table 3-2. An Example of an Error or Confusion Matrix.**

| CATEGORIES | Reference A | Reference B | Reference C | User's Accuracy |
|---|---|---|---|---|
| Image A | 65 | 5 | 10 | 65/80 = 0.81 |
| Image B | 25 | 85 | 15 | 85/125 = 0.68 |
| Image C | 10 | 10 | 75 | 75/95 = 0.79 |
| Producer's Accuracy | 65/100 = 0.65 | 85/100 = 0.85 | 75/100 = 0.75 | |
| Sum of major diagonal = 225 Overall accuracy = 225/300 = 75% | | | | |

An error matrix is a square array which expresses the number of sample units (i.e., pixels, clusters of pixels, or polygons) assigned to a particular category relative to the actual category as verified on the ground (Congalton, 1991). The columns in an error matrix represent reference or ground-truth data, while the rows represent the labeled data (Story and Congalton, 1986). The diagonal elements of this matrix represent agreement or correct classifications. The number of correct observations divided by the total number of points observed (times 100%) gives the overall percent of correct classifications, or the overall map accuracy. The error matrix is an effective way of visualizing errors of omission (exclusion) and commission (inclusion).

The accuracy for each class can be given in two ways: from the perspective of the user of the map, *i.e.*, the percentage of times that a class on the map correctly identified the class actually on the ground; or from the perspective of the producer, *i.e.*, the percentage of times that a class on the ground was correctly identified on the map. These two approaches can give very different results as is illustrated in Table 3-2, where the user's accuracy for category A is 81%, while the producer's accuracy for category A is only 65%.

A Kappa coefficient (Congalton 1991) is derived from the error matrix calculations and is used to measure the relationship of non-random categorization agreement versus expected disagreement. The calculation of Kappa assumes a multinomial sampling model and independence (Bishop *et al.*, 1988). It is used to monitor trends in reliability from one categorization to another. The Kappa coefficient equals zero when the agreement between the categorized data and ground truth equals chance or random agreement. Kappa increases to one as chance agreement decreases. Kappa equal to one occurs only when there is perfect agreement.

**Verification Data Set**

The verification data set consisted of medium to small-scale aerial photographs acquired by the US Department of Interior under the NAPP and NHAP programs. The photographic coverage included color infrared (CIR) aerial photographs over the entire United States at nominal scales of 1:40,000 and 1:58,000 respectively. Early in the categorization process a large quantity of NAPP photographs were acquired, both to assist in the labeling process, and to support the accuracy assessment effort. The accuracy assessment was performed using these previously acquired photographs. In order to avoid bias, photographs used in the labeling process were excluded from the data set used in the verification process. Limited reserve field data from earlier verification studies were used to validate the aerial photointerpretation of land cover/use.

Dates of the NAPP photographs ranged from September, 1987 to April, 1990. The dates of the NHAP photography ranged from April, 1981 to April, 1982. The temporal variation between the NAPP prints and the TM imagery did not appear significant. In the case of the NHAP photographs, which varied up to nine years in acquisition timing from the dates of the TM imagery, changes in land cover/use were sometimes noted. In these cases, the photos were used in conjunction with the original TM imagery to evaluate the land cover/use classifications.

**Sampling Scheme**

Adequately representing the categorized land cover/use data set was critical to a valid accuracy assessment. The sampling criteria employed in this study were based on the multinomial equation developed by Tortora (1978). Using this formula, and assuming a worst-case scenario (the majority land cover/use class representing 50% of the coverage), 71 samples were required for the majority land cover/use class to attain an 85% confidence interval. This was independent of the actual population size and was applied initially to three TM Scene areas and then to the watershed as a whole. The sample sizes for the remaining categories were evaluated by determining the proportional representation of each class within the scenes or watershed relative to the majority class. The proportional value was chosen since it was simpler to calculate and tended to produce a larger sample.

These calculations indicated that small sample sizes were appropriate for classes with low numbers of samples. In order to adequately assess these categories, an attempt was made to obtain a minimum of 5 samples per class. In some cases, however, repeating the sample selection process numerous times failed to produce five samples. As a result, some categories were not evaluated in certain data sets. This procedure was repeated for three representative TM scenes and for the watershed as a whole. Table 3-3 shows the combined total number of samples per category evaluated during the assessment (note that the three level 2 forest categories were aggregated into a single woody category in the final image data).

**TABLE 3-3. Total Sample Points Evaluated by Land Cover/Use Category**

| | Category | | Total Samples | |
|---|---|---|---|---|
| 11 | High-Density Developed | | 27 | |
| 12 | Low-Density Developed | | 61 | |
| 21 | Deciduous Woody | \ | 282 | \ |
| 22 | Evergreen Woody | (20 Woody) | 43 | (330) |
| 23 | Mixed Woody | / | 5 | / |
| 30 | Herbaceous | | 224 | |
| 40 | Exposed Land | | 20 | |
| 60 | Water | | 42 | |
| | | Total | 704 | |

## EMAP Hexagon Sampling Scheme

A review of the preliminary three scene accuracy assessment effort showed that screen digitizing the reference photographs occupied a significant amount of time. In the case of the preliminary scenes, as many as 96 reference photos were digitized. Building a corresponding photo data set for the watershed would have involved digitizing several hundred photos. One goal of this project was to test the effectiveness of the EMAP hexagons for characterizing land cover/use data. For this reason, a method of constraining the data set to existing EMAP Stage Two hexagons was devised.

The EMAP sampling grid consists of a series of continuous 640 $km^2$ hexagons covering the country. The EMAP Stage Two hexagons are an evenly spaced grid of 40 $km^2$ hexagons that fall at the center of every sampling grid cell. The Chesapeake Bay watershed contains 291 hexagons, uniformly distributed throughout the watershed. Figure 3-3 shows the location of these hexagons. In addition, NAPP and NHAP aerial photos were available for virtually all of these hexagons.

Before proceeding with the hexagon-constrained sampling routine, a test was performed to determine the adequacy of the EMAP hexagon grid for representing the watershed at large. The boundaries of the hexagons were superimposed over the categorized map image and summary statistics were derived for the hexagon areas, the photo coverage within the hexagons, and the watershed as a whole. The results of the classification for the whole watershed were compared to the results extracted from within the hexagons. Table 3-4 contains the summary statistics from the final classification, including the data within the 291 EMAP Stage 2 hexagons and the photo coverage used for the accuracy point sampling. For each of the 6 categories found in the watershed, areal extent is given as percentage, acres, and hectares.

The Woody category comprised half (54.7%) of the land cover/use within the watershed. The classes making up the remaining half of the watershed are: Herbaceous (32.9%), Water (7.5%), Low-Intensity Developed (4.0%), High-Intensity Developed (0.6%), and Exposed Land (0.4%). The results confirmed that the hexagon grid was an appropriate approach for a representative subsample of the Watershed land cover/use data.

## Assessment Procedure

Once the sampling scheme had been determined, the photo verification data set was assembled. This was accomplished by screen digitizing the four corner points of the selected photos onto the displayed TM imagery. The result was a coordinate file representing the selected photos and their coverages. Those areas of the classified data set which were within the photo coverages were then extracted

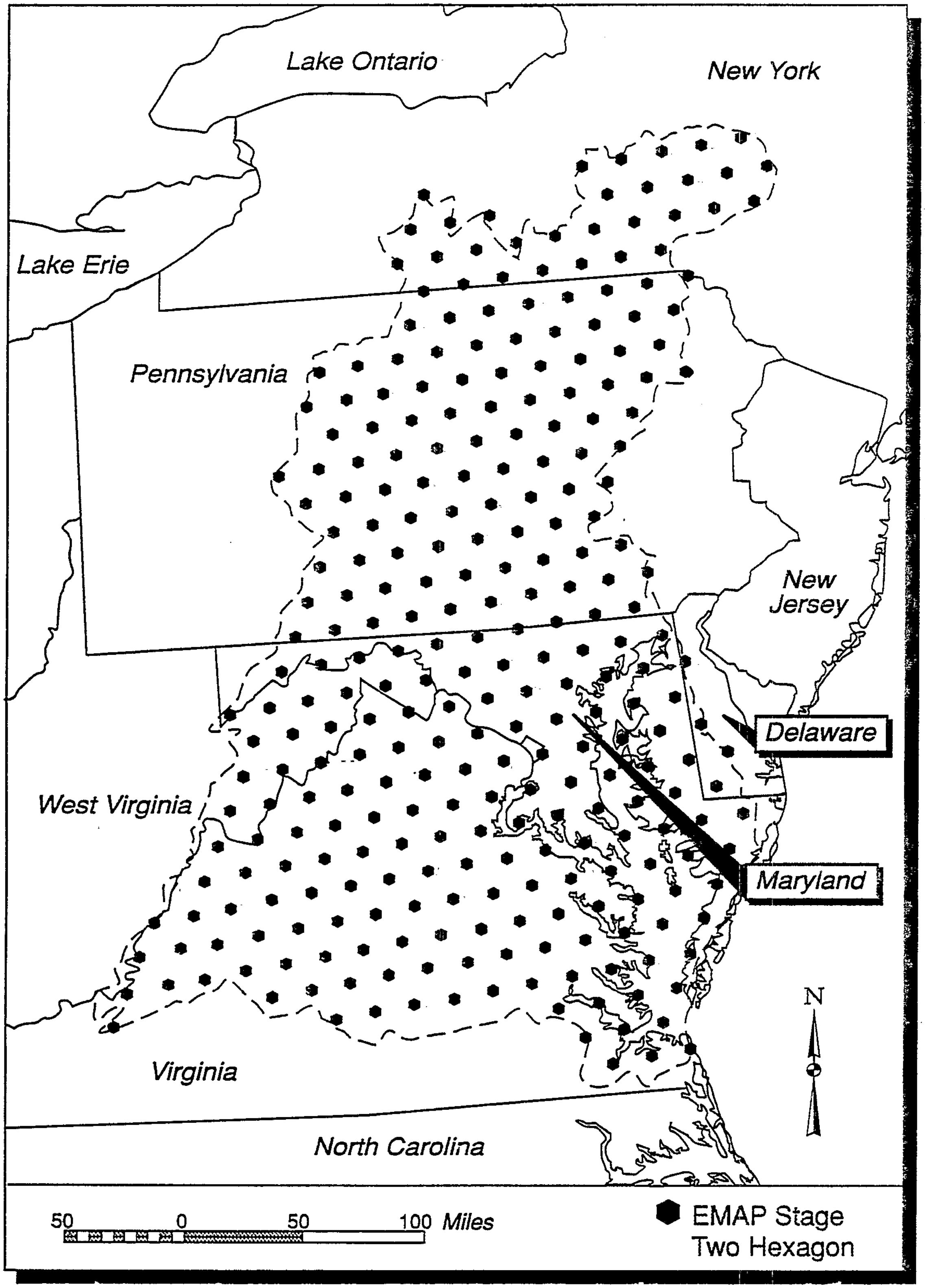


**Figure 3-3. Distribution of the 291 EMAP 40 km² Hexagons within the Chesapeake Bay Watershed.**

**Table 3-4. Land Cover/Use Statistics and Correlation Coefficients for Overall Chesapeake Bay Watershed, Sample Point Photo Coverage, and EMAP Stage 2 Hexagon Coverage.**

**Hexagon and Photo Coverage vs Overall Watershed**

**Overall Watershed**

| Category | Hectares | Acres | Sq.Miles | Percent |
|---|---|---|---|---|
| 11 | 101,302 | 250,323 | 391.13 | 0.57% |
| 12 | 714,992 | 1,766,783 | 2,760.60 | 4.03% |
| 20 | 8,158,817 | 20,160,877 | 31,501.37 | 54.74% |
| 30 | 5,838,413 | 14,427,034 | 22,542.24 | 32.88% |
| 40 | 49,495 | 122,306 | 191.10 | 0.28% |
| 60 | 1,333,399 | 3,294,901 | 5,148.28 | 7.51% |
| Total | 16,196,419 | 40,022,223 | 62,534.72 | |

**Hexagon Coverage**

| Category | Hectares | Acres | Sq.Miles | Percent |
|---|---|---|---|---|
| 11 | 7,089 | 17,517 | 27.37 | 0.63% |
| 12 | 46,171 | 114,090 | 178.27 | 4.13% |
| 20 | 624,417 | 1,542,969 | 2,410.89 | 55.87% |
| 30 | 358,233 | 885,212 | 1,383.14 | 32.05% |
| 40 | 3,260 | 8,055 | 12.59 | 0.29% |
| 60 | 78,611 | 194,251 | 303.52 | 7.03% |
| Total | 1,117,780 | 2,762,095 | 4,315.77 | |

**Coefficient of Correlation = 0.99190**

**Photo Coverage**

| Category | Hectares | Acres | Sq.Miles | Percent |
|---|---|---|---|---|
| 11 | 15,871 | 39,217 | 61.28 | 0.52% |
| 12 | 118,451 | 292,698 | 457.34 | 3.90% |
| 20 | 1,719,006 | 4,247,757 | 6,637.12 | 56.58% |
| 30 | 1,035,438 | 2,558,623 | 3,997.85 | 34.08% |
| 40 | 8,277 | 20,454 | 31.96 | 0.27% |
| 60 | 140,938 | 348,266 | 544.17 | 4.64% |
| Total | 3,037,981 | 7,507,015 | 11,729.71 | |

**Coefficient of Correlation = 0.99338**

Category 11 = High Intensity Developed
Category 12 = Low Intensity Developed
Category 20 = Woody
Category 30 = Herbaceous
Category 40 = Exposed Land
Category 60 = Water

from the overall watershed data. The resultant file, constrained to the photographic coverage, was the data set from which the random sample points were drawn.

Final sample point selection was performed by individuals not participating in the accuracy assessment process. This was done to avoid introducing bias into the photointerpretation process. In addition, the database land cover/use types were not revealed to the analyst until the assessment was completed and the results compiled.

Following the random point selection process, the area surrounding each point was observed and an interpretation made as to it's identity. The categorical value assigned to each 3x3 sample site was the majority value of the 3x3 pixel window centered at the selected point location. If no majority existed within the window, then the value assigned was that of the central pixel. The 3x3 pixel sample window was selected as it would provide a central pixel for point location. It also provided a sample size of 0.56 ha, or approximately half the size of the MMU, which facilitated an evaluation of small or narrow and irregularly shaped thematic units, as well as mixed or transition areas.

The final stage in the process involved displaying the outlines of the photo coverage as it overlaid on the original TM imagery. The 3x3 pixel sample site was displayed on the screen, located on the corresponding photograph, plotted on an acetate overlay attached to the photo, and an interpretation made as to the correct identification of the land cover/use. This interpreted land cover/use value was then recorded in a Classification Accuracy Table (CAT), which contained the sample number, its coordinates, the categorized value, and a data record field into which the verification (photo-interpreted) values were entered.

A Photographic Accuracy Assessment Form (Appendix A) was used to record the visually interpreted land cover/use category. This form allowed the analysts to make comments about the nature of features within the sample sites, the quality of the photographs, and to assess and assign a secondary category value if the sample was of mixed land cover/use types. A 3x3 pixel grid was employed to graphically represent the position of significant features within and surrounding mixed sample sites.

Results of the accuracy assessment of the watershed are presented in Table 3-5. Categorical detail was reduced from an original 17 classes to 6 classes to obtain a final overall accuracy of 80% with an 85% confidence level. Initial accuracies for many of the original classes were less than 50%. The Exposed Land category had the lowest accuracy, with 60% User's Accuracy. Further aggregation of Developed and Exposed classes does not improve the final individual class thematic accuracies. Note that the overall accuracy of the non-transition sample sites (those more than one pixel from a category boundary) was 90%.

**Table 3-5. Final Thematic Accuracies for the Chesapeake Bay Watershed Categorized Data Set**

| Image Classes | Photointerpretation Classes 11 | 12 | 20 | 31 | 40 | 61 | Total |
|---|---|---|---|---|---|---|---|
| 11 | 19 | 2 | 1 | 0 | 2 | 0 | 24 |
| 12 | 6 | 42 | 6 | 6 | 0 | 0 | 60 |
| 20 | 1 | 12 | 288 | 29 | 0 | 2 | 332 |
| 30 | 8 | 12 | 32 | 164 | 6 | 3 | 225 |
| 40 | 0 | 0 | 4 | 4 | 12 | 0 | 20 |
| 0 | 0 | 0 | 0 | 1 | 4 | 38 | 43 |
| Total | 34 | 68 | 331 | 204 | 24 | 43 | |

Combined Overall Watershed Accuracies

| Category | Producer's Accuracy | User's Accuracy |
|---|---|---|
| 11 = High-Intensity Developed | 19/34 = 55.88% | 19/24 = 79.16% |
| 12 = Low-Intensity Developed | 42/68 = 61.76% | 42/60 = 70.00% |
| 20 = Woody | 288/331= 87.00% | 288/332= 86.74% |
| 30 = Herbaceous | 164/204= 80.39% | 164/225= 72.88% |
| 40 = Exposed Land | 12/24 = 50.00% | 12/20 = 60.00% |
| 60 = Water | 38/43 = 88.37% | 38/43 = 88.37% |

Final Chesapeake Bay Watershed Categorized Data Set Results:

Sum of major diagonal = 563
Overall accuracy 563/704 = 79.97%
Kappa Coefficient (Khat) = 0.70155
Variance of Kappa = 0.00048

Watershed Non-Transition Site Results:

Sum of major diagonal = 79
Overall accuracy 79/88 = 89.77%
Kappa Coefficient (Khat) = 0.86093
Variance of Kappa = 0.00192

(Non-transition sites consisted of samples at least one pixel removed from any category boundary.)

## FINAL LAND COVER/LAND USE GENERATION

The final accuracy assessment results were used to refine the land cover/use classifications in the Chesapeake Bay watershed data set. Individual land cover/use accuracies that fell significantly below 60% were merged or aggregated into similar or more general classification categories. The original 17 land cover/land use types were first reduced to 12 classes. This action followed the preliminary accuracy assessments by personnel of Towson State University, MD. Subsequently, the classes were merged to eight classes prior to the start of the final accuracy assessment. Results of the final accuracy assessment necessitated a further reduction to six classes. The watershed data set was recoded to these final six classes. These actions produced a land cover/use data set of known and acceptable thematic accuracy.

The final data product was converted to Arc/Info GRID coverage, and will be archived and distributed on 8mm digital tapes. Arc/Info GRID is a raster data format that is integrated into the Arc/Info vector GIS database environment. The use of Arc/Info GRID will facilitate modeling efforts where synthesis of vector GIS data with raster image data is necessary or desirable. A reduced scale image of the final classified Chesapeake Bay watershed land cover/use map is shown in Figure 3-4. The final Arc/Info GRID coverage is approximately 80 Mbytes.

## SUMMARY

All methodologies described were implemented in the creation of the Chesapeake Bay watershed land cover/use data set. These methodologies cover the selection of TM Bands to use in analyses, quality assurance and quality control procedures, the subsetting of TM scene for analysis, classification techniques, accuracy assessment methodologies, and the generation of final products. The following paragraphs summarize the major approaches and results:

1 - TM bands 3, 4, 5 and 7 were selected for analysis because this band combination had a consistently high Optimum Index Factor (OIF), a measure of the information content of individual bands and band combinations;

2 - TM scenes were subset into quarter scenes due to system limitations and because coarser resolution data sets were found to be inadequate for reducing statistical confusion or separating general land cover/use features;

3 - A two-step, unsupervised approach was used, utilizing a custom clustering algorithm and an optimized maximum likelihood classifier to spectrally classify TM data;

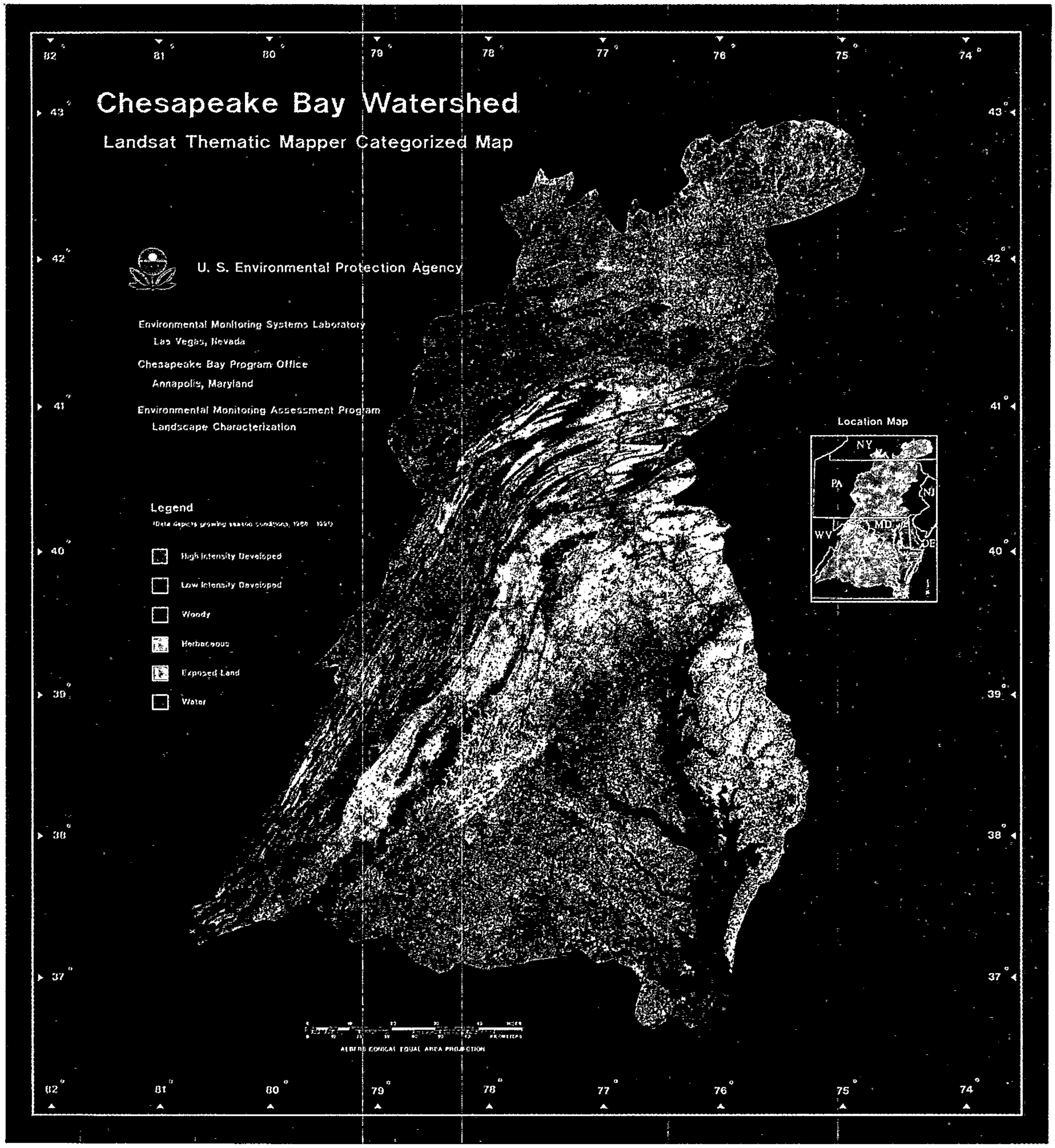


**Figure 3-4. Final classification of the Chesapeake Bay Watershed.**

4 - Spectral clusters were identified and assigned land cover/use labels primarily through reference to USGS NAPP and NHAP color infrared photographs. Aerial photographs were selected because they provided a reference data set that was relatively inexpensive, of high quality, and which provided consistent coverage across the study area;

5 - Following the first labeling step, a second unsupervised clustering and labeling routine was accomplished for all confusion clusters and unclassified pixels. This refinement was added because it improved the separation of land cover/use classes within the confusion clusters and reduced the number of unclassified pixels;

6 - All subsets of classified TM imagery were individually merged into a larger coverage, edge matched, and recoded to the final values of the classification system;

7 - All contiguous groups of pixels having the same land cover/use value were examined to eliminate areas smaller than the Minimum Map Unit of 1 ha. This was accomplished through use of a two step smoothing algorithm;

8 - An assessment of the thematic accuracy of the land cover/use was accomplished through photointerpretation of class-stratified random points. Land cover/use classes which did not meet accuracy DQO's were combined or aggregated into similar or more general classes and recoded;

9 - The final land cover/use classes were then converted to Arc/Info GRID format for archive and distribution.

# CHAPTER 4

# RESULTS AND DISCUSSION

The primary accomplishments of the Chesapeake Bay Watershed Pilot Project included a classification methodology, a digital land cover/use map, and a test of the EMAP Hexagon sampling scheme. This Chapter discusses the utility and limitations of the classification scheme and methodologies, and modifications are suggested. The final digital classification results and the test of the EMAP Hexagon sampling are also discussed.

## CLASSIFICATION SYSTEM

The classification system used for this project was a prototype of an interagency classification developed, as described in Chapter 2. However, not all features of the system were suited to the classification work performed in this project. The primary concern was the disparity between the categorization of land *cover* versus land *use*. The proposed system attempted to incorporate both.

Categories such as Woody, Herbaceous, and Water describe land cover. They identify basic features on the earth's surface. Categories such as Developed and Cultivated are land use descriptions. They identify an associated human use or interpretation of surface features. Wetlands represent conceptual environmental variables defined by soil types, topography, and species assemblages. These are often indistinguishable in spaceborne or aerial imagery. Satellite sensors can only measure reflected or emitted radiation from the earth's surface. It is the differentiation and identification of the resultant reflected or emitted spectra which drives traditional computer assisted digital-image classification. *A priori* knowledge is generally required to differentiate most land use classes. Detailed photointerpretation and field investigations are generally required to accurately delineate wetlands.

Results of the thematic accuracy assessment confirmed the difficulty of differentiating land cover and land use classes. Several of the original classes exhibited accuracies of less than 50%. This necessitated the aggregation of related land use and land cover classes into more general categories. For example, all herbaceous cover classes, including cultivated, were combined into one category. This process was repeated until the resulting overall accuracy met EMAP-LC Data Quality Objectives (DQO). While the remaining land cover/land use classes may not serve all intended purposes, they represent an accurate segmentation of surface features, within which more detailed differentiation and study can be accomplished.

There are a variety of potential methods which could produce greater thematic detail. One possibility would be to use the original raw spectral data to develop new spectral clusters within the final differentiated categories. This would eliminate some of the ambiguity incurred in developing signatures from an entirely undifferentiated TM scene or subscene, possibly improving the chances for separating desired features. Another possibility would be to use ancillary data such as municipal boundaries, transportation networks, population maps, tax maps, etc., in conjunction with the existing land cover classification to differentiate certain land use categories. The National Wetlands Inventory data from the United States Fish and Wildlife Service should serve to define wetland categories. There are ever increasing numbers and types of spatial data available which could be used to refine and improve the final categorization.

## CLASSIFICATION METHODOLOGY

The TM bands used in the unsupervised classification were selected based on an Optimum Index Factor (OIF), as described in Chapter 3. Based on the results of the OIF analysis, bands 3, 4, 5, and 7 were selected to maximize spectral information, while reducing processing complexity and time. This combination avoided problems of atmospheric scattering associated with the shorter wavelengths measured by band 1.

After completing the classification analysis, it was suspected that some of the confusion encountered between cover types during cluster labelling could have been reduced if TM bands 1 and 2 had been included. For example, a confusion between murky water and high intensity urban cover types might have been avoided. Given the widely varying types of surface features of interest in this study, it is suspected that using all reflective spectral information would yield better classification results. In addition, newer computer hardware has made the issue of processing complexity and speed less of a constraint.

### Subset TM scenes

The creation of Landsat TM scene subsets was the best available method for reducing the data volume at the time of the project. The relatively small and simple rectangular image blocks contained sufficient spectral variability (spectral signatures) for the development of spectral clusters.

The spectral statistics may have been improved had the scenes been subset by natural land divisions, such as ecological regions. The use of such boundaries to subset data introduces potential sources of problems and errors, since ecologic boundaries are likely to cross TM scene boundaries. Because of the spectral and temporal differences between TM scenes each should be classified independently. Problems of edge matching also occur. This process requires detailed human interaction. With small or irregular scene divisions the edge matching of the final data into one seamless coverage would be considerably more complex.

## Classification Techniques

The use of unsupervised clustering and maximum likelihood classification combined with photo interpretation and field work proved to be an effective and affordable method of covering the approximately 165,800 km² watershed. Processing and labelling instructions were as objective as possible,and were standardized in the Methods Guides (Appendix A). However, each data set had its own peculiarities and analysts occasionally had to modify procedures to maintain the quality of the clustering results.

The labelling of spectral clusters was the most subjective and time intensive stage in the classification process. The identification of a cover type can not be readily automated. An analyst must visually identify the cover type represented by each spectral cluster. This process depends on the experience of the analyst and the quality of the reference information (field notes, air photos, maps, etc.).

The consistency of the labelling process was apparent as the data subsets were pieced together to form the complete coverage. The occasional mismatched clusters were checked and relabeled if necessary before recoding to the final class values.

The cluster refinement process was effective in improving the discrimination between surface types. The ability to go back and redefine clusters that contained more than one surface type allowed the analyst to specify fewer clusters at the beginning. Since labelling clusters is so time consuming it was less tedious for the analyst to produce a smaller number of initial spectral clusters and then redefine cluster statistics for those clusters which contained mixed cover types. These clusters were indicated on the tracking forms, and editing suggestions were noted.

Post classification editing provided corrections for the final classification but little overall change to the classification of the subsets. These edits improved local areas cosmetically but changed only a small percent of the total pixels. Changes ranged from less than 1, to 3 percent of each image subset. An exception was in the region of forest gypsy moth infestation (Figure 4-1).

Gypsy moth defoliation of forest canopies resulted in an unobstructed satellite view of the shrub and herbaceous cover of the forest floor in these areas. This problem occurred primarily on the ridge tops in the Appalachians of the southwestern

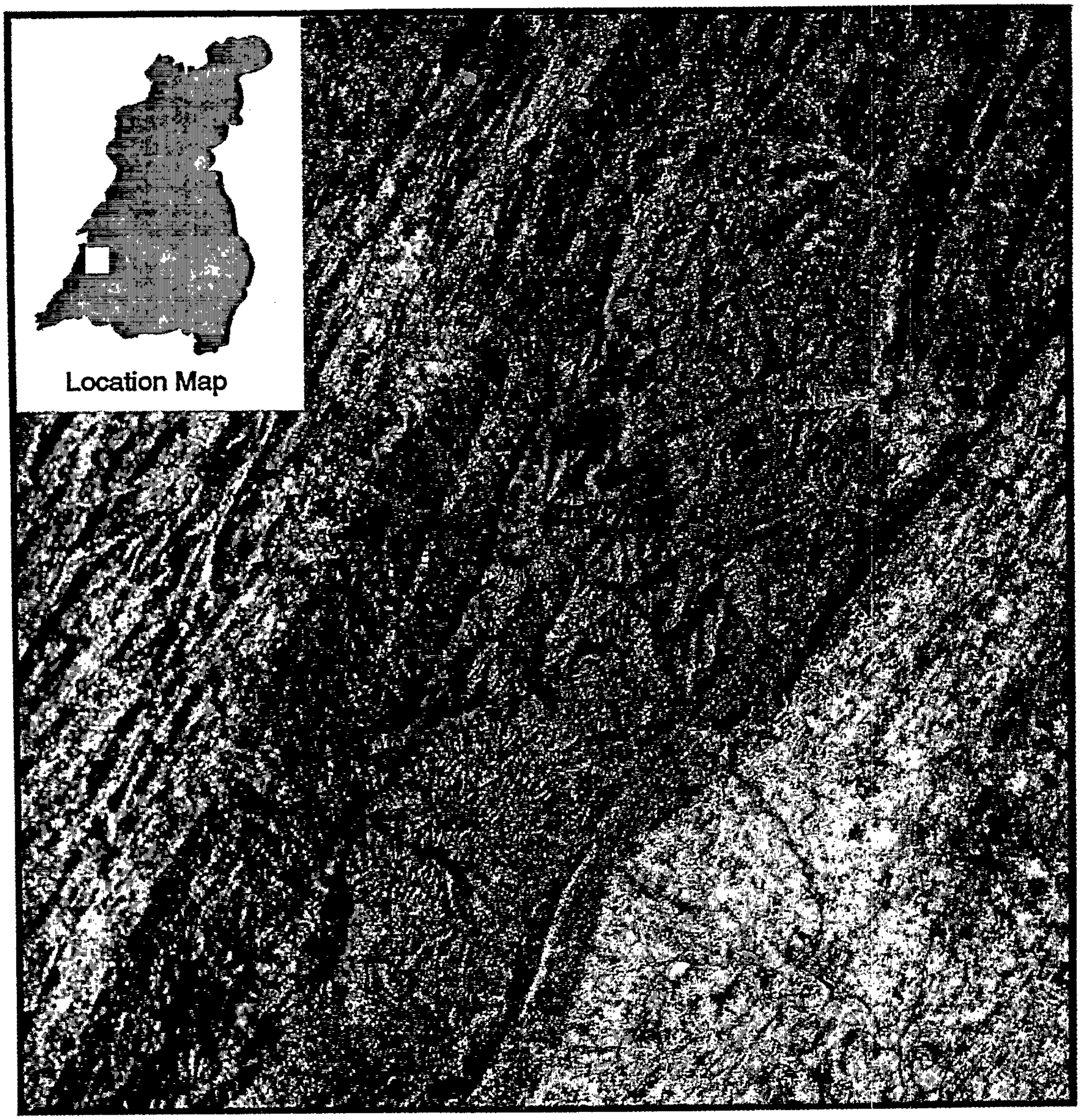


**Figure 4-1 Landsat Thematic Mapper color composite image of the Shenandoah Mountains, VA. Thematic Mapper bands 4, 5, and 3 are assigned to red, green, and blue, respectively. The blue areas along the ridges in the central portion of the image indicate areas of gypsy moth defoliation.**

and central western portions of the watershed. The resulting spectral signatures were indistinguishable from other areas of herbaceous cover. The striking difference between healthy forest and damaged forest can be seen in Figure 4-1. The false color composite image is composed of TM bands 5, 4, and 3 in red, green and blue, respectively. The Shenandoah Mountains run north-northeast across the image. The areas damaged by gypsy moths appear as blue tones along the ridges. There is a strong contrast between these areas and the reds and oranges of the healthy forest. The blue and blue-green areas in the image in the low lands either side of the ridge are primarily areas of herbaceous cover. The dark blue in the southwest corner of the image is the town of Harrisonburg, VA.

### Final Land Cover/Use Classification Product Generation

As each image subset was pieced into the master image file, mislabelled clusters (as discussed above) were relabeled. Each boundary between subsets was checked for classification consistency. Edges of the image subsets matched well, including subsets from within the same TM scene and from adjacent scenes.

The smoothing algorithm, used to eliminate pixel groups of less than the one hectare minimum mapping unit, effectively corrected edge-effect problems. Isolated groups of pixels, miss-classified because they lie on the edge of two resources and have a mixed signature, were essentially eliminated. The smoothing method as described in Chapter 3 finds and eliminates all features smaller than a combined pixel area of 1 ha. Unlike traditional smoothing filters which use a specific kernel size, linear features as narrow as one pixel wide were maintained as long as their combined area was equal to or greater than the minimum map unit. This allowed features such as roads and streams to remain.

### EMAP Hexagon Sampling Scheme

The EMAP Hexagon Sampling Scheme was developed as a spatial sampling system for a wide variety of point and spatial data. One goal of this sampling design was to test its utility in sampling land cover/use spatial data. The comparison of land cover/use statistics for the overall watershed and the hexagons (Table 4-2) indicated that the hexagons provided an adequate representation. The category percent coverages differed by less than 1% for all but the Woody class, which differed by 1.13%. The hexagon sampling scheme appears adequate, even for less common surface types. In summary, the EMAP Hexagon Sampling Scheme provided a useful estimate of the percent coverage of classes for the Chesapeake Bay watershed.

# CHAPTER 5

## CONCLUSIONS

The Chesapeake Bay Watershed Pilot Project was initiated to meet the needs of the Environmental Protection Agency (EPA) Environmental Monitoring and Assessment Program - Landscape Characterization (EMAP-LC) and the EPA Chesapeake Bay Program Office (CBPO). This Chapter summarizes the main points and recommendations from the report.

The classification system used for the project was developed by an interagency group for wide ranging purposes. As a land cover and land use system it was not ideally suited for use with a TM data classification. The classification of TM spectral signatures lacked the human interpretations necessary to categorize land use classes, as was explained in the Discussion section. The political and social importance of the land use categories such as croplands and urban areas are well understood. However, future work should carefully consider the spectral separability of these categories using TM data. It may be necessary to use additional data resources and/or a methodology other than traditional classification techniques to produce greater land use detail.

The methodology presented here effectively classified a number of land cover/use categories. However, accuracy DQOs were met only after simplifying and aggregating the original classes. Existing raster or vector coverages of specialized land use could be combined with the resultant classification to improve detail.

A significant accomplishment of this project was the development and implementation of tracking forms and instruction guides developed as part of the QA/QC procedures. From the beginning of this project tracking forms traced all procedures performed on the data. The tracking forms allowed for easy handling of the large number of TM scenes. The forms facilitated monitoring the completion of each step in the process for each data subset, comparison of results between data subsets, backup and retrieval of data, and tracing of errors. The instruction guides helped the analysts perform consistent analyses on all of the image subsets by providing step-by-step operating instructions.

## REFERENCES


Anderson, J. R. 1971. Land Use Classification Schemes Used in Selected Recent Geographic Applications of Remote Sensing. Photogrammetric Engineering 37(4):379-387.

Anderson, J. R., E. E. Hardy, J. T. Roach, and R. E. Witman. 1976. A Land Use and Land Cover Classification System for Use with Remote Sensor Data. U.S. Geological Survey Professional Paper 964, Washington, DC. 28 pp.

Bishop, Y.M.M., S.E. Fienberg, and P.W. Holland, 1988, *Discrete Multivariate Analysis, Theory and Practice*, The MIT Press, Cambridge, MA, 557 p.

Chavez, P. S., G. L. Berlin, and L. B. Sowers. 1982. Statistical Method for Selecting Landsat MSS Ratios. Journal of Applied Photographic Engineering 8:23-30.

Chavez, P. S., C. Guptill, and J. A. Bowell. 1984. Image Processing Techniques for Thematic Mapper Data. Technical Papers, 50th Annual Meeting of the American Society of Photogrammetry 2:728-742.

Congalton, R. G., 1991. A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data. Remote Sensing of the Environment 37(1):35-46.

Cowardin, L. M., V. Carter, F. C. Golet, and E. T. LaRoe. 1979. A Classification of Wetlands and Deepwater Habitats of the United States, Office of Biological Services, Fish and Wildlife Service, U.S. Department of the Interior, Washington, DC. 103 pp.

Harlow, W.M., and E.S. Harrar. 1969. *Textbook of Dendrology: Covering the Important Forest Trees of the United States and Canada*. McGraw-Hill Book Company, New York, NY. 511 pp.

Jensen, J. R., 1986. *Introductory Digital Image Processing*. Prentice-Hall, Englewood Cliffs, NJ. 379 pp.

Latty, R. S., and R. M. Hoffer. 1981. Waveband Evaluation of Proposed Thematic Mapper in Forest Cover Classification. Proceedings of American Society of Photogrammetry Fall Technical meeting, Niagara Falls, NY, pp. RS2-D:1-12.

Nelson, R. F., R. S. Latty, and G. Mott. 1984. Classifying Northern Forests Using Thematic Mapper Simulator Data. Photogrammetric Engineering and Remote Sensing 50(5):607-617.

Omernik, J. M., 1987. Ecoregions of the Conterminous United States. Annals of the Association of American Geographers 77(1):118-125.

Richards, J. A. 1986. *Remote Sensing Digital Image Analysis*. Springer-Verlag, New York, NY. 292 pp.

Sabins, F. F., 1987. *Remote Sensing Principles and Interpretation*. W.H. Freeman and Company, New York, NY. 449 pp.

Scheaffer, R. L., W. Mendenhall, and L. Ott. 1986. *Elementary Survey Sampling, 3rd edition*. P.W.S - Kent Publishing Company, Boston, MA. 324 pp.

Stenback, J. M. and R. G. Congalton, 1990. Using Thematic Mapper Imagery to Examine Forest Understory. Photogrammetric Engineering & Remote Sensing 56(9):1285-1290.

Story, M., and R. G. Congalton. 1986. Remote Sensing Brief - Accuracy Assessment: A User's Perspective. Photogrammetric Engineering & Remote Sensing 52(3):397-399.

Taylor, J. K. 1987. *Quality Assurance of Chemical Measurements*. Lewis Publishers, Inc., Chelsea, MI.

Weerackoon, R. D., and T. H. Mace. 1990. A Method of Optimizing the Maximum Likelihood Classifier Using Mathematical Transformations. Proceedings of the 1990 ACSM/ASPRS Annual Convention, Denver, CO. Vol. 4, pp. 464-474.

## ACKNOWLEDGMENTS

Many individuals have contributed to the success of this project. First we would like to acknowledge the many individuals who have contributed directly to the completion of this classification: Denice Shaw, EMAP-LC Technical Coordinator, Ross Lunetta, originating Project Officer and Douglas J. Norton, EMAP-LC Technical Director at the initiation of the project, all of the US EPA; Mark Finkbeiner and Steven R. Hoffer, Lockheed Environmental Systems and Technologies Company (LESAT); Janice L. Thompson,The Wilderness Society; Scott Thomasma, formerly of LESAT. Others who contributed significantly to the project include: John Lyon, Ohio State University; Russ Congalton, University of New Hampshire; Jay Morgan, Towson State University; William Aymard, PCI, Inc.; Mary E. Balogh, US Bureau of Reclamation; Edward Bright, Oak Ridge National Laboratory; Michael Cambers, US Geological Survey (USGS); James J. Chung, LESAT; Jerome E. Dobson, Oak Ridge National Laboratory; Lynn K. Fenstermaker, Desert Research Institute; Randolph L. Ferguson, NOAA/National Marine Fisheries Service (NMFS); Frank Golet, University of Rhode Island; Kenneth D. Haddad, Florida Department of Natural Resources; Jimmy Johnson, US Fish and Wildlife Service; Donley Kisner, the Bionetics Corporation; Richard Kleckner, USGS; Victor V. Klemas, University of Delaware; K. Peter Lade, Salisbury State University; Karen H. Lee, LESAT; Kathy Lins, USGS; James P. Thomas, NOAA/NMFS; and Bill O. Wilen, US Fish and Wildlife Service. Also contributing were: Triana N. Burchianti, LESAT; Dominic A. Fuccillo, LESAT; Lynda Liptrap, Computer Sciences Corporation; James Love, EOSAT; James R. Lucas, LESAT; Tom Mace, US EPA; John Nietling, LESAT; Lynn Schuler, US EPA - Chesapeake Bay Program Office; Kris Stout, the Bionetics Corporation; Ron Risty, USGS EROS Data Center; and Ridgeway D. Weerackoon, Desert Research Institute.

## CLIP TM SCENES TO WATERSHED

Path:______ Row:_____

BSTATS output (file_____)

File size of clipped file:
Number of rows:___________ columns:____________

PRINCO output (file_____)

OIF output (file_____)

Backup Tape:
Tape number:_______________________________________________

File name:__________________________________________________

Tape listing (file_____)

## LIST OF MAP NAMES

1:250,000:______________________ 1:250,000:______________________

1:100,000:___________________ 1:100,000:___________________

1:100,000:___________________ 1:100,000:___________________

1:100,000:___________________ 1:100,000:___________________

1:100,000:___________________ 1:100,000:___________________

1:250,000:______________________ 1:250,000:______________________

1:100,000:___________________ 1:100,000:___________________

1:100,000:___________________ 1:100,000:___________________

1:100,000:___________________ 1:100,000:___________________

1:100,000:___________________ 1:100,000:___________________

| | |
|---|---|
| 1:250,000:________________ | 1:250,000:________________ |
| 1:100,000:________________ | 1:100,000:________________ |
| 1:100,000:________________ | 1:100,000:________________ |
| 1:100,000:________________ | 1:100,000:________________ |
| 1:100,000:________________ | 1:100,000:________________ |

## APPENDIX A

## INSTRUCTIONS AND FORMS

Copies of instruction guides and tracking forms used for the image classification work are shown on the following pages. They appear in the order in which they are used in the analysis.

RECEIVING TM DATA

Path:_________ Row:_________ Scene ID:_________________________

***** From Billing Statement (file______)*************************************

Billing order number:________________________

Sequence number:_______________

Shipping date:___________________

***** From STX Header Information Sheet (file______)**************************

Acquisition date:_____________

UTM zone:______

Pixels per line:___________

Lines per image:___________

| | Latitude | Longitude | UTM-X | UTM-Y |
|---|---|---|---|---|
| UL: | _____°_____'________"W | _____°_____'________"N | __________ | __________ |
| UR: | _____°_____'________"W | _____°_____'________"N | __________ | __________ |
| LR: | _____°_____'________"W | _____°_____'________"N | __________ | __________ |
| LL: | _____°_____'________"W | _____°_____'________"N | __________ | __________ |
| | | Scene Center: | __________ | __________ |

Blocking factor:__________

Record length:__________

***** From STX Rectification Information (file______)*************************

Number of points in consensus set:_____

RMS X:________

RMS Y:________

RMS D:________

***** EMSL*LV Tape Library Information
****************************************

| | EMSL-LV 9-track tape number | TM bands | notes |
|---|---|---|---|
| 1 | _______________ | ________ | _______________________________________ |
| 2 | _______________ | ________ | _______________________________________ |
| 3 | _______________ | ________ | _______________________________________ |
| 4 | _______________ | ________ | _______________________________________ |

EMSL-LV 8mm number:________________ (file MTCOUNT output_____)

## SUBSET TM SCENES

1) Get disk assignment from supervisor. You will be working probably on drs4. Log in as "rsches" and move to that directory.
   sunss03% cd /drs4/rsches

2) Create a directory on that disk using the conventions outlined in the following example:
   sunss03% mkdir tm1533

3) Create a link at the home directory (/drs1/rsches) to the new directory. This link will help others find your files. Use the following commands:
   sunss03% cd
   sunss03% ln -s /drs4/rsches/tm1533 tm1533

4) Retrieve tape from library (tape numbers are listed in the tracking book) and insert in SUN390's 8mm tape drive. Shell to the server and cd to the new directory.
   sunss03% rsh sun390
   sun390% cd
   sun390% cd tm1533

5) Retrieve the file from tape. This step will take some time as the files are large. The "tar" command will read the entire tape, even though you will be requesting the first file on tape. The file size can be found on the tape listing in the file for this scene. You may monitor the status of the "tar" command by using another window to check the file size. After the entire file is read you may Ctrl-C to stop tape processing. (It will continue to scan the whole tape even if your file is the first one on the tape.) Use the following command to retrieve the file:
   sun390% tar -xvf /dev/rst1 cltm1533.lan

6) Run BSTATS, get header listing only, send output to printer and file results. Use the following example to respond to prompts:
   ERD> bstats
   Is this an Image or a GIS file? i
   Enter Image filename: cltm1533.lan
   Make listing go to Printer, Terminal, or Both? p
   Make a listing of the Statistics? y
   Make a listing of the Histogram? n
   Use the whole image? y
   Enter X skip factor: 1
   Enter Y skip factor: 1
   Count the zeros? n

7) Decide where to subset image. You will want a file roughly 4000x3000. Enter subset coordinate information on the tracking form for each subset.

8) Create subdirectories for each subset. Use the conventions outlined in the following examples:

sunss04% mkdir subset1
sunss04% mkdir subset2

9) Move to the appropriate subdirectory and use SUBSET to create subset image files. Remember to select only TM bands 3, 4, 5, and 7. Refer to the following example for prompt responses:

ERD> cd subset1
ERD> subset
Image or GIS file? i
Enter Input Image filename: ../cltm1533.lan
Use the whole image: n
Enter coordinates (X, Y) for upper left corner? 1829,712
Enter coordinates (X,Y) for lower right corner? 2829,1712
Enter Output Image filename: cltm1533sub1.lan
How many columns are to be in the output file? <default>
How may rows are to be in the output file? <default>
Enter coordinates of absolute upper-left corner of output file? <default>
Enter output file coordinates at which to place upper left corner of input subset? <default>
How may bands are to be in output file? 4
Copy all bands in order? n
For input band 1, enter output band? -1
For input band 2, enter output band? -1
For input band 3, enter output band? 1
For input band 4, enter output band? 2
For input band 5, enter output band? 3
For input band 6, enter output band? 4

10) Run BSTATS on subset image, get a statistics listing only, send output to printer and file results.

ERD> bstats
Is this an Image or a GIS file? i
Enter Image filename: tm1533sub1.lan
Overwrite the file? y
Make a statistics listing? y
Make a histogram listing? n
Listing to go to Printer, Terminal, or Both? p
Use the whole image? y
Enter X skip factor? 1
Enter Y skip factor? 1
Count zeroes in the statistics computation? n

11) Repeat steps 9 and 10 for each subset.

SUBSET TM SCENES

Path:______ Row:_____

Directory (full path):_____________________________________________

Tape number:_______________________

File name:_________________________

LISTIT (file_____)

Number of subsets:_____

******************** Subset 1 ********************

Directory (full path):________________________________________

Upper left file coordinate:____________,______________

Lower right file coordinate:__________,__________

Output file name:___________________________________

BSTATS (file_____)

******************** Subset 2 ********************

Directory (full path):________________________________________

Upper left file coordinate:____________,______________

Lower right file coordinate:__________,__________

Output file name:___________________________________

BSTATS (file_____)

******************** Subset 3 ********************

Directory (full path):________________________________________

Upper left file coordinate:____________,______________

Lower right file coordinate:__________,__________

Output file name:___________________________________

BSTATS (file_____)

Path:______ Row:_____

******************** Subset 4 ********************

Directory (full path):________________________________________

Upper left file coordinate:____________,______________

Lower right file coordinate:__________,__________

Output file name:___________________________________

BSTATS (file_____)

# CLASSIFICATION INSTRUCTIONS

The following instructions outline step by step how Landsat Thematic Mapper (TM) data are classified in the Chesapeake Bay Watershed Pilot Project. This document is designed to be used by an analyst during the data classification process. It does not explain the techniques or reasoning behind the different steps in the analysis (see the section on Methodology). Before proceeding with these steps, TM data must be subset to the proper area and reduced to four bands (bands 3, 4, 5, and 7). After completing the steps, the user will have created a classified image ready to be combined with other images and edited.

Computer programs written at the Environmental Systems Laboratory-Las Vegas (EMSL-LV) and Erdas software are required to complete the classification steps. Familiarity with the Unix operating system and Erdas software is assumed. The names of all programs are written in boldface.

<u>EMSL-LV programs</u>

| | | |
|---|---|---|
| **gisrain** | - | generates a rainbow file that imitates a three-band composite |
| **kluster** | - | generates statistical clusters from .lan file and input parameters |
| **maxopt** | - | assigns pixels to clusters generated in **kluster** |
| **printstf** | - | formats output from **kluster** to be printed. output file name, "printstf.log" |
| **unitvarne0** | - | calculates band variance thresholds from the .lan file, excludes zero values, output file, VFILE.DAT |
| **wckluster** | - | regenerates statistical clusters from mixed clusters from **maxopt** |
| **wcmaxopt** | - | assigns pixels to clusters generated in **wckluster** |

<u>Erdas programs</u>

| | | |
|---|---|---|
| **bstats** | - | generates image statistics |
| **colormod** | - | highlights clusters on the screen |
| **display** | - | displays .gis image |
| **electromap** | - | plots .lan and .gis image files |
| **gisedit** | - | edits screen values of .gis file and updates file values |
| **read** | - | displays .lan file |
| **recode** | - | recodes .gis file |
| **stitch** | - | attaches two geographically adjacent images into a single image |

The naming conventions for data files used in these instructions should be followed so that work may be traced easily. All examples in this document use "sub1" in the file names to indicate that subset 1 of a scene is being processed. The numbers 2, 3, or 4 should be substituted for the 1 in "sub1" to indicate the appropriate subset.

1) To begin classifying a new subset, fill out the top of a new classification tracking form and start a new file folder to hold the printouts. Remember to label all printouts with the scene and subset numbers and keep them in the folder (i.e. 1533 subset1). The tracking forms and the folder of data printouts must be kept organized for QA/QC checks.

2) Run **unitvarne0**

ERD> **unitvarne0** sub1.lan

3) Print VFILE.DAT, this listing should be filed in the scene folder.

ERD> lpr VFILE.DAT

4) Find the cumulative window count of 50% for each band on VFILE.DAT and mark on the printout. For example, if the percentage closest to 50% falls at the range of 11-12, then select the number 12 as your variance threshold.

5) Create a parameter file (.pfile) using textedit. Use the naming convention outlined in example below.

ERD> textedit sub1a.pfile &

6) The .pfile contains all information required by the **kluster** and **maxopt** programs. The file format must be exactly correct. Make sure you use a comma to separate items on a record line. The list below describes the items on each line.

Record 1: Image file name
Record 2, item 1: option of **kluster**
1= euclidian distance option, 2= quadratic threshold method
Record 2, item 2: # of windows to skip along columns
Record 2, item 3: # of windows to skip along rows
Record 3: output statistics file name (.stf)
Record 4, item 1: Unitvar variance threshold for channel 1
Record 4, item 2: Unitvar variance threshold for channel 2
Record 4, item 3: Unitvar variance threshold for channel 3
Record 4, item 4: Unitvar variance threshold for channel 4
Record 5, item 1: desired number of clusters
Record 5, item 2: segment of .stf file (use segment 1)
Record 6: output classified image file name
Record 7, item 1: standard deviation for **maxopt**, suggested value 2.1

The .pfile for subset 1 of a scene should look similar to the example below:

```
sub1.lan
1,0,0
sub1.stf
11.0,25.0,54.0,17.0
80,1
sub1a.gis
2.1
```

7) Run **kluster**.

SS03% kluster sub1a.pfile

8) Run **printstf** to create a listing of kluster results. The command line must contain the .stf file name and the number of channels. File the output in the scene folder.

ERD> printstf sub1.stf 4 (4 indicates the number of bands)
ERD> lpr printstf.log

Write the total number of clusters before the final merge (during **kluster**) at the bottom of the printstf.log. This number can be found at the bottom of the sub1.stf file.

9) The desired number of clusters is very scene-dependent and you will have to decide how many clusters you will need. A minimum of 80 clusters and as many as 95 may be useful. The output of the **printstf** will tell you, among other things, how many final clusters were found. To generate more clusters, edit record 5 in the .pfile and choose threshold values corresponding to the cumulative window count of 40% instead of 50%. To generate fewer clusters, use threshold values corresponding to the cumulative window count of 60% instead of 50%. In one example, only 52 clusters were originally found; consequently, the 50% values were replaced with 40% values in the pfile. Document this change on the printout of VFILE.DAT

10) Repeat steps 7 through 9 if required.

11) Put final printout of **printstf** in scene folder.

12) Run **maxopt**.

SS03% **maxopt** sub1a.pfile

13) Run **gisedit** to remove bad pixel values, if any, generated in the first row of the .gis file. Set the pixel values to zero.

14) Run **bstats** on the output .gis file. This listing must be kept in the scene folder.

ERD> **bstats**
Is this an Image or a GIS file: g
Enter GIS filename: sub1a.gis
Overwrite the file? y
Make a header information listing? y
Make a histogram listing? y
Make a listing of the color scheme? n
Listing to go to Printer, Terminal, or Both? b

15) Print a listing of the .pfile. File the printout in the scene folder.

ERD> lpr sub1a.pfile

16) Collect the appropriate hardcopy reference material for naming clusters.

U.S. Geological Survey Topographic maps
NAPP and NHAP aerial photography
U.S. Geological Survey Land Use/Land Cover Maps
U.S. Fish and Wildlife National Wetland Inventory maps
U.S. Department of Agriculture Agricultural Statistics Bulletins, by state
U.S. Soil Conservation Service Soil Survey Bulletins, by county
Landsat Thematic Mapper image maps

17) Label the clusters of the new .gis file created in **maxopt** (sub1.gis). First, use **gisrain** to generate a rainbow file containing color schemes that imitate three-band composite images. See **gisrain** help screen for use of this program. Run **display** to view the .gis file. Using **colormod,** update the trailer with the color scheme of your choice. Also in **colormod,** highlight each cluster one at a time to identify the level 2 class it belongs to. Use the reference materials obtained step 16. Record the class number for each cluster on the "Cluster Labels" tracking form. At this point do not edit the .gis file. If classes require editing, **note them on the tracking form**. Mark them on the image plotted in step 36. Clusters which include more than one cover type should be flagged for "re-cluster" on the tracking form. **Note any areas which should be field checked**. The following is a simplified list of the names and numbers of the classification system:

| Level 0 | Level 1 | Level 2 |
| --- | --- | --- |
| Upland | 1 Developed | 11 High Intensity<br>12 Low Intensity |
| | 2 Cultivated Land | 21 Woody<br>22 Herbaceous |
| | 3 Grassland | 31 Herbaceous |
| | 4 Woody | 41 Deciduous<br>42 Mixed<br>43 Evergreen |
| | 5 Exposed Land | 51 Soil<br>52 Sand<br>53 Rock<br>54 Evaporite Deposits |
| | 6 Snow & Ice | 61 Snow & Ice |
| Wetland | 7 Woody Wetland | 71 Deciduous<br>72 Mixed<br>73 Evergreen |
| | 8 Herbaceous Wetland | 81 Herbaceous |
| | 9 Nonvegetated Wetland | 91 Nonvegetated |
| Water and Submerged land | 10 Water and submerged land | 100 Water |

The classification system listed above was used for the Chesapeake Bay Watershed Pilot project. However, it was established and classification begun before the final EMAP classification system was determined. Future projects may wish to follow the final EMAP classification system, which was modified from the above in some categories.

18) Identify areas on the image for which air photos or other reference material is available. Record the file coordinates of several 1024x1024 windows and the photo numbers covered by this window on the "Cluster Labels" tracking form. This step may be done while **kluster** and **maxopt** are running.

19) Copy the old .gis to a new file name.

sunss04% cp sub1a.gis sub1b.gis

20) Copy the old .pfile to a new file name.

sunss04% cp sub1a.pfile sub1b.pfile

21) Edit the new sub1b.pfile. Make the following changes:

Record 5, item 3 - 13: add class number for those clusters needing re-clustering. Cluster 255 (unclassified pixels) may be included. Make sure the numbers are separated by commas. A maximum of 10 clusters may be listed

Record 6: specify the new output .gis file created in step 19

The file should look similar to the example below:

```
sub1.lan
1,0,0
sub1.stf
11.0,25.0,54.0,17.0
80,1,1,2,7,21,255
sub1b.gis
2.1
```

22) Run **wckluster**

SS03% **wckluster** sub1b.pfile

23) Use **printstf** to get a listing of **wckluster** results. The command line must contain the .stf file name and the number of channels.

ERD> **printstf** wckluster.stf 4 (4 indicates the number of bands)

ERD> lpr printstf.log

24) Run **wcmaxopt**

SS03% **wcmaxopt** sub1b.pfile

25) Update the trailer. Copy a trailer file containing names and numbers for the clusters into the working directory. This directory is useful when for printing the recode.aud file later on.

sunss04% cp /drs1/rsches/defaults/wckluster.trl sub1b.trl

26) Run **bstats** on the output .gis file. File the output in the scene folder.

ERD > **bstats**
Is this an Image or a GIS file: g
Enter GIS filename: sub1b.gis
Overwrite the file? y
Make a header information listing? y
Make a histogram listing? y
Make a color table listing? n
Listing to go to Printer, Terminal, or Both? b
Printer unit? 0

27) Print a hard copy listing of wcmaxopt.log. File it in the scene folder.

ERD > lpr wcmaxopt.log

28) Run **gisrain** again to create a new rainbow file for your new .gis image. Name the new clusters as in step 17. Write the class numbers on the "Cluster Labels" tracking form.

29) Run **recode** to create the final .gis file for this subset. Create an "audit" file of this step to document how the clusters are recoded to class values.

ERD > prep
Enter audit file name: recode.aud
ERD > **recode**

Enter output file name: sub1r.gis

ERD > noprep

30) Print a hardcopy of the audit file. Check the recode values on the hard copy and make any necessary corrections to the audit file with textedit. Make a new hard copy if necessary and put it in the scene folder. Run the audit file with a batch command.

ERD > lpr recode.aud
ERD > batch recode.aud

31) Copy the final trailer file. A standard trailer file that contains the color scheme and class names for the final .gis files was created at the start of the project. This file may be copied into your directory.

sunss04% cp /drs1/rsches/defaults/final.trl sub1r.trl

32) Run **bstats** on the output .gis file.

ERD> **bstats**
Is this an Image or a GIS file: g
Enter GIS filename: sub1a.gis
Overwrite the file? y
Make a header information listing? y
Make a histogram listing? y
Make a listing of the color scheme? n
Listing to go to Printer, Terminal, or Both? b

33) Use **display** to view the .gis image with the new trailer. Look at the image reduced to fit on the display screen and look for any problems.

34) Repeat steps 1 - 30 for each subset in the scene. If the other subsets are completed, compare them to your newly completed subset. Either display them side by side or temporarily **stitch** them together. Rerun any recodes that will improve the match of the subsets.

35) Plot a hard copy of the .lan file for your subset. Run **electromap** to plot a linear stretch of bands 2, 3, 1 (TM bands 3, 5, 4) as R, G, B. Use the default stretch options. Plot at a scale of 1:250,000. Your plot will probably use more than one strip of paper. Tape the strips together. Mark on the plot any unusual features and areas that will need to be edited in the classified .gis file and write an explanation in the margins. Store the plot in the map drawer marked RSCHES.

36) Generate a plot of the recoded .gis image to mark any major edits. First display your recoded .gis file. Run **colormod** and call up a rainbow file containing the class colors for the printer. The path name is /drs1/rsches/defaults/legend.rnb. Retrieve the look up table for the printer and update the trailer. Then run **electromap** and print the .gis file scaled to fit the page. Mark any areas that need editing. Circle areas with a heavy pen and mark in the margin the class numbers that are to be changed (for example 51 -> 22). Put this image in the tracking book following the "Cluster Labelling" tracking form.

37) When all of the subsets for a scene are completed, backup all files for the scene on 8mm tape and print a copy of the tape log. Use the following tar commands:

SUN390% tar -cv tm1533
SUN390% tar -tv > 1533back.up
SUN390% lpr 1533back.up

Record the tape number and the date on the tracking form and the printout of the backup log. File the backup log in the scene folder.

## CLUSTER LABELS

TM path:________ row:________ subset #________

file coordinates

| | X | Y | air-photo numbers |
|---|---|---|---|
| A: | | | |
| B: | | | |
| C: | | | |
| D: | | | |
| E: | | | |
| F: | | | |
| G: | | | |
| H: | | | |

| cluster | scan | wc | A | B | C | D | E | F | G | H | final | notes |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | | | |
| 2 | | | | | | | | | | | | |
| 3 | | | | | | | | | | | | |
| 4 | | | | | | | | | | | | |
| 5 | | | | | | | | | | | | |
| 6 | | | | | | | | | | | | |
| 7 | | | | | | | | | | | | |
| 8 | | | | | | | | | | | | |
| 9 | | | | | | | | | | | | |
| 10 | | | | | | | | | | | | |
| 11 | | | | | | | | | | | | |
| 12 | | | | | | | | | | | | |
| 13 | | | | | | | | | | | | |
| 14 | | | | | | | | | | | | |
| 15 | | | | | | | | | | | | |
| 16 | | | | | | | | | | | | |
| 17 | | | | | | | | | | | | |
| 18 | | | | | | | | | | | | |
| 19 | | | | | | | | | | | | |
| 20 | | | | | | | | | | | | |
| 21 | | | | | | | | | | | | |
| 22 | | | | | | | | | | | | |
| 23 | | | | | | | | | | | | |
| 24 | | | | | | | | | | | | |
| 25 | | | | | | | | | | | | |

| cluster | scan | wc | A | B | C | D | E | F | G | H | final | notes |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 26 | | | | | | | | | | | | |
| 27 | | | | | | | | | | | | |
| 28 | | | | | | | | | | | | |
| 29 | | | | | | | | | | | | |
| 30 | | | | | | | | | | | | |
| 31 | | | | | | | | | | | | |
| 32 | | | | | | | | | | | | |
| 33 | | | | | | | | | | | | |
| 34 | | | | | | | | | | | | |
| 35 | | | | | | | | | | | | |
| 36 | | | | | | | | | | | | |
| 37 | | | | | | | | | | | | |
| 38 | | | | | | | | | | | | |
| 39 | | | | | | | | | | | | |
| 40 | | | | | | | | | | | | |
| 41 | | | | | | | | | | | | |
| 42 | | | | | | | | | | | | |
| 43 | | | | | | | | | | | | |
| 44 | | | | | | | | | | | | |
| 45 | | | | | | | | | | | | |
| 46 | | | | | | | | | | | | |
| 47 | | | | | | | | | | | | |
| 48 | | | | | | | | | | | | |
| 49 | | | | | | | | | | | | |
| 50 | | | | | | | | | | | | |
| 51 | | | | | | | | | | | | |
| 52 | | | | | | | | | | | | |
| 53 | | | | | | | | | | | | |
| 54 | | | | | | | | | | | | |
| 55 | | | | | | | | | | | | |
| 56 | | | | | | | | | | | | |
| 57 | | | | | | | | | | | | |
| 58 | | | | | | | | | | | | |
| 59 | | | | | | | | | | | | |
| 60 | | | | | | | | | | | | |
| 61 | | | | | | | | | | | | |
| 62 | | | | | | | | | | | | |
| 63 | | | | | | | | | | | | |
| 64 | | | | | | | | | | | | |
| 65 | | | | | | | | | | | | |
| 66 | | | | | | | | | | | | |
| 67 | | | | | | | | | | | | |
| 68 | | | | | | | | | | | | |
| 69 | | | | | | | | | | | | |
| 70 | | | | | | | | | | | | |

## CLASSIFICATION OF SUBSETS

TM path:_____ row:_____ subset #_____

Directory:________________________________________________

File name:________________________________________________

UNITVARNE0

VFILE.DAT (file_____)

KLUSTER

final pfile name:_____________________________________

output stats file name:_______________________________

output of printstf (file _____)

MAXOPT

output GIS file name:_________________________________

BSTATS (file_____)

pfile after maxopt - with class names (file_____)

WCKLUSTER

new GIS file name:____________________________________

new pfile name:_______________________________________

output of printstf (file_____)

WCMAXOPT

output GIS file name:_________________________________

BSTATS (file_____)

WCMAXOPT.LOG - with class names (file_____)

RECODE

audit file name:______________________________(file_____)

output GIS file name:__________________________________

BSTATS (file_____)

FILE CHECK LIST

Image files

___ sub1.lan
___ sub1.sta
___ sub1a.gis
___ sub1a.trl
___ sub1b.gis
___ sub1b.trl
___ sub1r.gis
___ sub1r.trl

Miscellaneous files

___ VFILE.DAT
___ sub1a.pfile
___ sub1b.pfile
___ sub1.stf
___ printstf.log
___ wckluster.stf
___ wcmaxopt.log
___ recode.aud

(Your file names may vary from sub1, sub2, sub3, etc.)

PRINT OUT CHECK LIST

___ VFILE.DAT
___ printstf.log - after first clustering
___ bstats - for gis image after first clustering
___ .pfile
___ printstf.log - after "within class" clustering
___ bstats - for gis image after "within class" clustering
___ wcmaxopt.log
___ recode.aud
___ bstats - for gis image after recoding to the classification scheme

BACKUP

tape number: ______________________

date: ______________________

tar listing (file_____)

| cluster | scan | wc | A | B | C | D | E | F | G | H | final | notes |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 71 | | | | | | | | | | | | |
| 72 | | | | | | | | | | | | |
| 73 | | | | | | | | | | | | |
| 74 | | | | | | | | | | | | |
| 75 | | | | | | | | | | | | |
| 76 | | | | | | | | | | | | |
| 77 | | | | | | | | | | | | |
| 78 | | | | | | | | | | | | |
| 79 | | | | | | | | | | | | |
| 80 | | | | | | | | | | | | |
| 81 | | | | | | | | | | | | |
| 82 | | | | | | | | | | | | |
| 83 | | | | | | | | | | | | |
| 84 | | | | | | | | | | | | |
| 85 | | | | | | | | | | | | |
| 86 | | | | | | | | | | | | |
| 87 | | | | | | | | | | | | |
| 88 | | | | | | | | | | | | |
| 89 | | | | | | | | | | | | |
| 90 | | | | | | | | | | | | |
| 91 | | | | | | | | | | | | |
| 92 | | | | | | | | | | | | |
| 93 | | | | | | | | | | | | |
| 94 | | | | | | | | | | | | |
| 95 | | | | | | | | | | | | |
| 96 | | | | | | | | | | | | |
| 97 | | | | | | | | | | | | |
| 98 | | | | | | | | | | | | |
| 99 | | | | | | | | | | | | |
| 100 | | | | | | | | | | | | |
| 101 | | | | | | | | | | | | |
| 102 | | | | | | | | | | | | |
| 103 | | | | | | | | | | | | |
| 104 | | | | | | | | | | | | |
| 105 | | | | | | | | | | | | |
| 106 | | | | | | | | | | | | |
| 107 | | | | | | | | | | | | |
| 108 | | | | | | | | | | | | |
| 109 | | | | | | | | | | | | |
| 110 | | | | | | | | | | | | |
| 111 | | | | | | | | | | | | |
| 112 | | | | | | | | | | | | |
| 113 | | | | | | | | | | | | |
| 114 | | | | | | | | | | | | |
| 115 | | | | | | | | | | | | |

| cluster | scan | wc | A | B | C | D | E | F | G | H | final | notes |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 116 | | | | | | | | | | | | |
| 117 | | | | | | | | | | | | |
| 118 | | | | | | | | | | | | |
| 119 | | | | | | | | | | | | |
| 120 | | | | | | | | | | | | |
| 121 | | | | | | | | | | | | |
| 122 | | | | | | | | | | | | |
| 123 | | | | | | | | | | | | |
| 124 | | | | | | | | | | | | |
| 125 | | | | | | | | | | | | |
| 126 | | | | | | | | | | | | |
| 127 | | | | | | | | | | | | |
| 128 | | | | | | | | | | | | |
| 129 | | | | | | | | | | | | |
| 130 | | | | | | | | | | | | |
| 131 | | | | | | | | | | | | |
| 132 | | | | | | | | | | | | |
| 133 | | | | | | | | | | | | |
| 134 | | | | | | | | | | | | |
| 135 | | | | | | | | | | | | |
| 136 | | | | | | | | | | | | |
| 137 | | | | | | | | | | | | |
| 138 | | | | | | | | | | | | |
| 139 | | | | | | | | | | | | |
| 140 | | | | | | | | | | | | |
| 141 | | | | | | | | | | | | |
| 142 | | | | | | | | | | | | |
| 143 | | | | | | | | | | | | |
| 144 | | | | | | | | | | | | |
| 145 | | | | | | | | | | | | |
| 146 | | | | | | | | | | | | |
| 147 | | | | | | | | | | | | |
| 148 | | | | | | | | | | | | |
| 149 | | | | | | | | | | | | |
| 150 | | | | | | | | | | | | |

SUBSET EDITING INSTRUCTIONS

The following instructions outline step by step how classified Landsat Thematic Mapper (TM) data are edited in the Chesapeake Bay Watershed Pilot Project. This document is designed to be used by an analyst during the process of editing the imagery. It does not explain the methodology or results of processing the Chesapeake Bay imagery (see the section on Methodology). Before proceeding with these steps the TM data must be classified, projected into the final projection, and assigned a recode parameter file and rainbow file associated with it. After completing the steps the user will have an edited, classified image ready to be combined with other subsets for the final coverage.

Computer programs written at the Environmental Systems Laboratory-Las Vegas (EMSL-LV) and Erdas software are required to complete the classification steps. Familiarity with the Unix operating system and Erdas software is assumed. The names of all programs are written in boldface.

EMSL-LV programs
**classord** - generates a new parameter file for recode2 with clusters grouped by similar surface types and a new rainbow file
**recode2** - recodes cluster values in a .gis file

Erdas programs
**listit** - list file header information
**display** - displays a .gis image
**colormod** - load and modify rainbow files (color lookup tables)
**gisedit** - edit values in a .gis image

The naming conventions for data files should be followed so that work may be traced easily. All examples of image files use the scene path and row numbers followed by an "s" and a single digit for the subset number (for example, 1533s2: path 15, row 33, subset 2).

1) Three files should be present in a new working directory: the classified TM subset file (tm1633s3.gis); the recode parameter file used when balancing and recoding the subset to its surrounding subsets in the larger coverage (recode.pfile); and the rainbow file containing color schemes for the "old" classification, the "new" balanced classification, and the imitation of the color composite (tm1633s3.rnb). Retrieve these files from the tape backup of the master raster if necessary.

```
sunss03% rsh sun390
sunss03% cd tm1633
sun390% tar -xvf /dev/rst1 tm1633s3.gis recode.pfile tm1633s3.rnb
```

2) Copy the .gis file to a new file name for further processing:

ERD> cp tm1633s3.gis 1633s3ue.gis ("ue" for unedited)

3) Run **listit** to obtain the number of rows and columns in the .gis file. Send the output to the terminal only.

ERD> **listit**

4) Edit the existing recode parameter file for use in **classord**. The format of the file is as follows:

line 1: the new .gis file name (1633s3ue.gis)
line 2: -1 (no change)
line 3: 1, 1, number of rows, number of columns
line 4 to end: cluster number, recode number (no change)

The only changes to be made are the file name on line 1 (the file name) and on line 3 (numbers, rows, and columns).

sunss03% textedit recode.pfile

5) Run the program **classord**. This program creates a new recode parameter file that will be used to recode the .gis image so that clusters of the same surface type will be grouped together. It also creates a rainbow file adjusted to be used for the new image. **Classord** will run as follows:

sunss03% **classord**
enter input parameter file name: recode.pfile
enter output parameter file name: neworder.pfile
enter input rainbow file name: tm1633s3.rnb
enter output rainbow file name: neworder.rnb

6) Print a copy of the log file output from running **classord**.

sunss03% lpr classord.log

Copy the cluster range for each class in the space provided on the "Editing Subsets" tracking form. This will be useful during editing. Note that the first range of clusters (new value 200) includes clusters that were unclassified, reclustered in **wckluster**, and clusters missed in the labelling process.

7) Run the program **recode2**. This program uses the recode parameter file created in classord to recode the .gis image file. The program overwrites the original file. The resulting image will look identical; only the order of the clusters will be different. The program runs as follows:

sgws02% **recode2** neworder.pfile

8) Use Erdas **display** to display the .gis image after recoding.

9) Use Erdas **colormod** to access the rainbow file neworder.rnb created by **classord** in step 5. Update the trailer of the .gis image to display the "new" classified color scheme created in the last major step of balancing the subsets. The image should look just as it did before these editing steps were begun. To insure that all clusters have been assigned the correct colors, reassign colors to all clusters. Reassignment can be done very quickly in **colormod** by using the ranges of cluster values from the tracking form. Note the colors of the unclassified pixels (old cluster value 255). These may need to be changed to white to improve their color contrast on the image. Save the new lookup table in the "neworder.rnb" rainbow file as "new2." A color palette with the appropriate colors and code numbers can be found at the following path name:

/drs1/rsches/defaults/codes.dat

10) Retain the file with the "ue" unchanged in your directory. Copy the file and its trailer to new names and perform the edits on the new file.

sunss03% cp 1633s3ue.gis 1633s3e.gis
sunss03% cp 1633s3ue.trl 1633s3e.trl

11) If the image is in a region containing gypsy moth damage, record all cluster numbers that contain moth damage. Do this by using **colormod** to first display the look up table that imitates the color composite and then "flash" each cluster with the color palette. Record cluster numbers on the "Editing Subsets" tracking form.

12) Gather all reference material available to aid in image editing. This material may include air photos, 1:100,000 scale topographic maps, 1:24,000 topographic maps, soils maps, etc.

13) Set the Erdas image-display drivers and display portions of imagery to begin editing the data. The imagery will be edited at full resolution; therefore, it must be displayed in smaller overlapping sections. First be sure two image drivers are open; one should be 1024x1024, the other can be any size. Use **display** to display the upper left corner of the classified image in the 1024x1024 driver at a magnification of 1 (not a reduction of 1). Resize the second driver down to so that only the button panel on the right of the window is showing. Use **colormod** to load the "unclassified" rainbow file in the second driver. Note that the color scheme in the first driver changes to the lookup table used in the second driver because the display screen can only use one lookup table at a time. However, the RGB button on each driver window will return the lookup table of the window. All work will be done in the first driver, but note that by "clicking" on the RGB buttons of the two windows you can alternately view the classified and the color composite lookup tables.

14) Perform the image editing with **gisedit**. All edits will assign pixel values to the 201 to 219 range of pixel values to prevent overlapping with the pixel values of the clusters. See the list of class names and corresponding pixel values on the "Editing Subsets" tracking form. For large areas that can be outlined exactly, all pixels in a polygon may changed to a single new value. However, for the majority of the edits, scattered pixels of a particular class (color) will need to be changed to a different class. The range of the clusters for each class copied in step 6 will make these edits easier. For example, if scattered pixels labelled (colored) as "soil" in a field need to be changed to the label (color) of "cultivated," just circle the whole field and change the range of cluster values listed for "soil" to 204, the value for "cultivated."

    As sections of each subset are edited be sure to record the center coordinates of each section on the "Editing Subsets" tracking form. Space is also provided on the tracking form to note any unusual features or problems in each section. In addition to editing cluster values, vectors should be drawn down major roads (202), power lines (205), and airport runways (201 or 202) where features would be lost to smoothing. Do not bother drawing vectors on anything but major highways. Where a road with a vector crosses out of a subset, mark an arrow on the classified printout in the tracking book so that the adjacent subset will continue the vector.

15) Run **display** to view the entire subset after editing to look for mistakes and consistency in editing across the subset. Correct any errors that are found.

16) Run **bstats** on the final image. Save the hard copy output in the scene folder.

EDITING SUBSETS

TM Path: ______ Row: ______ Subset #: ______

old gis file name: ____________________

old recode parameter file name: ____________________

old rainbow file name: ____________________

new gis file name (before editing): ____________________

gis file - number of rows _____ number of columns _____

new recode parameter file name: ____________________

new rainbow file name: ____________________

print out classord.log (_____)

gis file name after editing ____________________

bstats from edited gis file (_____)

The range of cluster values for each class:

| Level 1 | Level 2 | class values old | new | cluster range |
|---|---|---|---|---|
| Unclassified | | 255 | 200 | |
| Developed . . . . . . . . . | high density | 11 | 201 | |
| | low density | 12 | 202 | |
| Cultivated land . . . . . . | woody | 21 | 203 | |
| | herbaceous | 22 | 204 | |
| Grassland . . . . . . . . . . . | herbaceous | 31 | 205 | |
| Woody . . . . . . . . . . . | .deciduous | 41 | 206 | |
| | mixed | 42 | 207 | |
| | evergreen | 43 | 208 | |
| Exposed Land . . . . . . | soil | 51 | 209 | |
| | sand | 52 | 210 | |
| | rock | 53 | 211 | |
| | evaporites | 54 | 212 | |
| Snow and Ice . . . . . . . | snow and ice | 61 | 213 | |
| Woody Wetland . . . . . | deciduous | 71 | 214 | |
| | mixed | 72 | 215 | |
| | evergreen | 73 | 216 | |
| Herbaceous Wetland . . . | herbaceous | 81 | 217 | |
| Nonvegetated Wetland . . | nonvegetated | 91 | 218 | |
| Water/Submerged Land . | water | 100 | 219 | |

Numbers of clusters with moth damage: ______________________________

____________________________________________________________

1024X1024 Edit Areas

| Area | Center Coordinate X | Center Coordinate Y | Comments |
| --- | --- | --- | --- |
| 1 | | | |
| 2 | | | |
| 3 | | | |
| 4 | | | |
| 5 | | | |
| 6 | | | |
| 7 | | | |
| 8 | | | |
| 9 | | | |
| 10 | | | |
| 11 | | | |
| 12 | | | |
| 13 | | | |
| 14 | | | |
| 15 | | | |
| 16 | | | |
| 17 | | | |
| 18 | | | |
| 19 | | | |
| 20 | | | |
| 21 | | | |
| 22 | | | |

| | | | |
|---|---|---|---|
| 23 | | | |
| 24 | | | |
| 25 | | | |

Check Lists

For each image subset, make sure the following printouts are in the scene folder, and the following files are left in the subset directory:

**Print Outs**

____ classord.log
____ bstats.out

**Files**

____ 1633s3ue.gis
____ 1633s3ue.trl
____ 1633s3e.gis
____ 1633s3e.trl
____ clasord.log
____ neworder.pfile
____ neworder.rnb
____ recode.pfile

## COMBINING SUBSETS INTO THE MASTER-RASTER

The following process steps are used to combine the subsets into a single raster file. This single file is called the "master-raster" throughout this document. The files that were output from the **wcmaxopt** program will be added to the master-raster. The labels previously assigned to the clusters will be compared to the subsets previously added to the master-raster. Modification may have to be made to the labels to match, as best as possible, the other subsets. Once final labels have been identified, the subset will be recoded to a classification numbering scheme that is unique to the master-raster. This numbering scheme uses values between 200 and 219, thus making it possible to manipulate the colors of cluster values (ranging from 1 to 199) without effecting the color surrounding completed data.

Computer programs written at the Environmental System Laboratory -Las Vegas (EMSL-LV) and ERDAS version 7.5 software are required to complete these steps. Familiarity with the UNIX operating system and ERDAS software is assumed. The names of all programs are written in boldface.

<u>EMSL-LV programs</u>
**dighead** - generates a ".dig" file from the header of a raster file
**listhead** - lists contents of the header of a raster file
**digcorners** - lists the extreme X and Y coordinates in a ".dig" file
**mapcon** - generates GCP points used to transform a raster file into a new projection
**georef** - uses the output of **mapcon** to calculate coefficients for projecting a raster file
**geomap** - uses the output of **georef** to create a raster file in a new projection
**recode2** - changes the pixel values in a ".gis" file

<u>ERDAS programs</u>
**ccvrt** - converts coordinates in a ".dig" file to a new projection
**subset** - overwrites a raster file with data from another raster file
**display** - displays a ".gis" file on the computer screen
**colormod** - allows interactive changes to colors associated with pixel values
**fixhed** - allows manipulation of the contents of the header in a raster file
**curses** - displays the pixel values currently displayed on the computer screen

Documentation of the ERDAS programs may be found in the ERDAS 7.5 software manuals. All EMSL-LV software is considered public domain and was developed for EPA under contract number 68-C0-0050 to Lockheed Engineering & Sciences Company. Source code is available upon request. Documentation of the EMSL-LV software is included in this final report of the Chesapeake Bay Pilot Project.

The naming conventions for data files used in these instructions should be followed to facilitate QC checks and process tracking. Most of the files have standard names. A few file names contain the TM path and row and subset number. Several of the examples in this document use "tm1533s1" within the file names to indicate that subset 1 of path 15, row 33 is being processed. The appropriate path, row and subset numbers should be substituted were applicable.

Retrieve File, Project to Albers, Subset, Recode, and Archive, a total of five major operations, comprised of 41 steps are described below.

<u>Retrieve File</u>

Retrieve the output of the classification from the tape archive to begin the procedure. You will be using the output from **wcmaxopt**. The pixels in this file contain the cluster numbers.

1) Record the TM path, row number, and the subset number on the top of the "Combine Files into Master Raster" tracking form.

2) Get 8-mm tape number from "Classification of Subsets" tracking form. Record it on the "Combine Files into Master Raster" tracking form.

3) The entire file name, including the path, must be specified to retrieve the file. This name can be found on the listing of the 8-mm tape. If a listing is not available, one can be made using a **tar** command. Insert the proper 8-mm tape into the tape drive, log onto the SUN390, and use the following command:

    sun390% **tar** -tvf /dev/rst1 > tape.log

    This command will create a file called "tape.log," which can be printed, containing a listing of the entire tape, specifying complete file names.

4) Record the ".gis" file name, including the path, on the "Combine Files into Master Raster" tracking form. If there is a rainbow (".rnb") file name instead of or in addition to the ".gis" file name, record its name on the tracking form. If no rainbow file exists, record the trailer (".trl") file name.

5) Insert the proper 8-mm into the tape drive, if you have not already done so. The **tar** command used to retrieve the files must include the complete file names with the path. Log onto the SUN390 and retrieve the files with a command line similar to the following (this is all one long command line):

    sun390% **tar** -xvf /dev/rst1 tm1533/subset1/sub1b.gis
    /dev/rst1/tm1533/subset2/sub1bgis.rnb

Your command line will vary, depending on the path and file name. If you are retrieving the trailer file rather than the rainbow file, the command will diff accordingly.

6) Record the output file name on the "Combine Files into Master Raster" tracking form.

Project to Albers

Project the subset from Universal Trans Mercator (UTM) to Albers by following these steps.

7) Record UTM zone on "Combine Files Into Master Raster" tracking form. The UTM zone may be found on the "Receiving TM Data" tracking form.

8) Run **dighead** program to create a ".dig" file of scene boundaries. The output file name should be specified as "utm1.dig." The command line should be similar to this:
    sunss03% **dighead** sub1b.gis utm1.dig

9) Record the upper left and lower right UTM coordinates, displayed by the **dighead** program, on the "Combine Files Into Master Raster" tracking form.

10) Run **listhead** to display the contents of the ".gis" file header. The corner UTM coordinates are listed so you may verify the coordinates recorded in step 9. Record the number of rows and number of columns in the file. These will be listed as number of columns and number of rows. They are recorded on the tracking form in reverse order (rows, columns) to simplify later processing steps.
    sunss03% listhead sub1b.gis

11) The program **mapcon** will be run to create a set of control points to be used to calculate the projection parameters. Create a parameter file called "geo1.pfile" to be used as input to the **mapcon** program. This file has the following parameters (correct entries on right):
    Line 1: Name of control point file . . . . . . . . . . . . . . . . . . . . geo.cfile
    Line 2: C, L, or W, where C requests a cubic fit, L a linear fit, and W a weighted linear fit . . . . . . . . . . . . . . . . . . . . . . . . . . . . . C
    Line 3: The following elements must be separated by commas:
      element 1: scan tolerance value measured in pixels . . . . . . . . . . . . 1.0
      element 2: element tolerance value measured in pixels . . . . . . . . . . 1.0
      element 3: UTM zone of input file . . . . . . . . . . . . . . . . . . see step 7
      element 4: latitude of the origin of output projection . . . . . . . . . . 0.0
      element 5: longitude of the origin of output projection . . -77.83333333333
      element 6: output projection type: 1 = Albers, 2 = Lambert . . . . . . . 1
      element 7: first standard parallel (Albers only) . . . . . . . . . . . . . . 38.0
      element 8: second standard parallel (Albers only) . . . . . . . . . . . . . 42.0
    Line 4: The word YES or the word NO, for listing of proceedings. If 'YES',

**georef** will generate a file named GEOPNTS.DAT and output relevant information . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . Yes

Line 5: The following elements must be separated by commas:
element 1: Name of input file . . . . . . . . . . . . . . . . . . . . . see step 6
element 2: Type of input file - ELAS or ERDAS . . . . . . . . . . . ERDAS
element 3: BL for a bilinear interpolation or NN for a nearest neighbor fitNN

Line 6: Name of output file . . . . . . . . . . . . . . . . . . . . . . tm*PPRR*s*#*.gis
where *PP* is the TM path, *RR* is the TM row, and # is the subset number.

Line 7: The following elements must be separated by commas:
element 1: Pixel width in meters . . . . . . . . . . . . . . . . . . . . . . . . 25.0
element 2: Pixel width in meters . . . . . . . . . . . . . . . . . . . . . . . . 25.0

Line 8: The following elements must be separated by commas:
element 1: UTM X coordinate of the upper left . . . . . . . . . . see step 9
element 2: UTM Y coordinate of the upper left . . . . . . . . . . see step 9
element 3: UTM X coordinate of the lower right . . . . . . . . . see step 9
element 4: UTM Y coordinate of the lower right . . . . . . . . . see step 9
element 5: line to start processing . . . . . . . . . . . . . . . . . . . . . . . . 1
element 6: number of lines to process . . . . . . . . . . . . . . . . . see step 10
element 7: element to begin processing . . . . . . . . . . . . . . . . . . . . . 1
element 8: number of elements to process . . . . . . . . . . . . . see step 10

All parameters for line 8 can be taken off the "Combine Files Into Master Raster" tracking form. The file should be called "geo1.pfile" and be similar to the following:

```
----------------------------------------
geo.cfile
C
1.0,1.0,18,0.0,-77.83333333333,1,38.0,42.0
YES
sub1b.gis, ERDAS, NN
tm1533s1.gis
25.0,25.0
265450.0,4397225.0,397675.0,4219050.0,1,7128,1,5290
----------------------------------------
```

Lines 1, 2, 4, and 7 will be exactly the same. Line 3 may differ depending on the input UTM zone only (element 3), all other parameters must be the same. Line 5 and 6 will differ only in the input and output file name, respectively. Line 8 will differ depending on the UTM coordinates and file size of the input file.

12) Run the program **mapcon** to create a set of control points to be used to calculate the projection parameters. Use the above parameter file (step 11).
sunss03% mapcon geo1.pfile

13) Make a hard-copy listing of both the parameter file and the log file. Mark the TM path and row and the subset number on these printouts and file them in the folder for this scene.

sunss03% lpr geo1.pfile
sunss03% lpr mapcon.log

14) Use the **ccvrt** program to convert the UTM coordinates in the "utm1.dig" file (see step 8) into Albers projection. The following example answers should help respond to the program prompts. All answers below will be the same for all subsets except the UTM zone.

ERD> **ccvrt**
Options: (D,T,A) [DIG file] : DIG file
Enter INPUT filename : utm1.dig
Enter UTM zone number ? [] : 18
North or South of the equator? (N,S) [North] : North
Enter spheroid number ? [] : 1
Enter OUTPUT filename : albers1
What is the OUTPUT Coordinate Type? 3 (Albers Conical Equal Area)
Enter LATITUDE of FIRST STANDARD PARALLEL? [] : 38
Enter LATITUDE of SECOND STANDARD PARALLEL? [] : 42
Enter LONGITUDE of CENTRAL MERIDIAN? [] : -77.83333333333 (use 10 3's)
Enter LATITUDE of ORIGIN of PROJECTION? [] : 0
Enter FALSE EASTING at CENTRAL MERIDIAN? [] : 0
Enter FALSE NORTHING at ORIGIN? [] : 0

15) Run **digcorners** program to obtain the new raster file coordinates. The command line should be exactly like this:

sunss03% digcorners albers1 albers2

16) Record the upper left and lower right Albers coordinates displayed by the **digcorners** program on the "Combine Files into Master Raster" tracking form. Verify these coordinates by comparing them to the "mapcon.log" file. Due to differences in orientation, the X and Y will not necessarily be listed in the same corner, and will probably be slightly different. The **mapcon** program simply projects the old UTM file corners. The output of **digcorners** is more exact and lists the <u>extreme</u> minimum and maximum X and Y coordinates. If you cannot find any coordinates in the "mapcon.log" file that are similar (within 10 m) to the coordinates output from **digcorners**, then check the contents of "geo1.pfile" (see step 11) and repeat steps 12 to 15. It is more likely that **mapcon** was run incorrectly than **dighead**, **ccvrt**, and **digcorners**. If you are sure **mapcon** is correct, rerun **dighead**, **ccvrt**, and **digcorners** (steps 8, 14, and 15, respectively). This is a critical step; do not proceed until you are sure everything was done correctly to this point.

17) Round off corners to multiples of 50 so they match the master-raster. Use the following conventions and be careful with negatives:

Upper left X: round DOWN to the nearest multiple of 50.
Upper left Y: round UP to the nearest multiple of 50.
Lower right X: round UP to the nearest multiple of 50.
Lower right Y: round DOWN to the nearest multiple of 50.

Remember that rounding a negative up results in a smaller negative number. Record these coordinates on the "Combine Files into Master Raster" tracking form.

18) Copy the above parameter file (see step 11) into a new file for editing. The new file should be called "geo2.pfile."

```
sunss03% cp geo1.pfile geo2.pfile
```

19) Edit line 8 of the parameter file and replace the UTM coordinates in line 8 with the new Albers coordinates from step 17. The new parameter file should be similar to the following:

```
----------------------------------------
geo.cfile
C
1.0,1.0,18,0.0,-77.83333333333,1,38.0,42.0
YES
sub1b.gis, ERDAS, NN
tm1533s1.gis
25,25
8350,4146250,146100,3963900,1,7128,1,5290
----------------------------------------
```

20) Make a hard-copy printout of the new parameter folder and save it in the file for this scene.

```
sunss03% lpr geo2.pfile
```

21) Before running the **georef** program (the next step), the file coordinates of the upper left corner must be 1,1. This position may be verified using the **listhead** program (see step 10). If the upper left file coordinates are not 1,1 use the ERDAS program **fixhed** to change it. Do not change any other header information.

22) The **georef** program uses the output from **mapcon** to calculate the necessary transformation coefficients to project the file from UTM into Albers projection. Run the **georef** program, using the above parameter file (step 19) as input.

```
sunss03% georef geo2.pfile
```

23) The **geomap** program uses the output from **georef** and actually creates the new raster file. Run **geomap** using the above parameter file (step 19) as input.

```
sunss03% geomap geo2.pfile
```

24) Record the output ".gis" file name (line 6, step 11) on the "Combine Files into Master Raster" tracking form.

25) Check the output using the ERDAS **display** program. If you were able to retrieve a rainbow file from tape archive (see steps 4 and 5) you may use **colormod** to retrieve a color scheme. Otherwise, the trailer will have to be copied to the new file name using the following command line as an example:

sunss03% cp sub1b.trl tm1533s1.trl

Create a new rainbow file using the ERDAS **colormod** program option "r - RAINBOW file I/O." Use a file name similar to that recorded in step 24 (for the above example the rainbow file name would be: "tm1533s1.rnb"). Save the color scheme that resembles the raw data and call it "unclassed."

When viewing the file, make sure none of the corners was inadvertently lost in the projection process. If they were lost, the rounded corner coordinates (step 17) were probably not computed correctly, or were entered into "geo.pfile" incorrectly (step 19). Repeat steps 17 to 24 until the entire file is transformed correctly.

Subset

Add the subset to the master-raster by the following steps.

26) Figure out the master-raster file coordinates for upper left and lower right corners of the subset. Use the rounded coordinates from above (step 9) and the following formulas:

Master X file coordinate = (257525 + Albers X) / 25
Master Y file coordinate = (4507925 - Albers Y) / 25

Record these coordinates on the "Combine Files into Master Raster" tracking form. Do these calculations carefully and double-check the results. A miscalculation can result in loss of data in the master-raster.

27) Use the ERDAS **subset** program to add the subset to the master-raster. Make sure that you use the proper file coordinates from step # above as the place for the upper left corner (line 5 below). Also specify that you want the output file overwritten (line 4 below), but that zero values should NOT overwrite existing data (line 6 below). Use care when answering the prompts because the Master-Raster will be overwritten and it will be difficult to correct mistakes. The following responses illustrate important correct answers to prompts:

Image or GIS file? GIS
Enter Input GIS filename : tm1533s1
Enter Output GIS filename : /drs4/rsches/master-raster/master
Overwrite the file? Yes
Enter output file coordinates at which to place
upper left corner of input subset? 10636 14468
Should input zero values overwrite data in the output file? No

28) Use the ERDAS **display** program to view the master-raster and assure that the subset fits into its proper place. Stay in the current directory and specify the full path name, "/drs4/rsches/master-raster/master.gis."

29) Use the ERDAS program **colormod**, option "r," to retrieve the color scheme created in step 25. This program will shade the new subset correctly, but black out the other values of the master-raster. Use **colormod** option "c - color palette entry" to reset the master-raster colors correctly. Use the "o - open new palette file" option and use the file: "/drs1/rsches/defaults/codes.dat." Modify the colors according to the following table:

| GIS value | Color name | Red | Green | Blue |
|---|---|---|---|---|
| 200 | sun-tan | 174 | 171 | 128 |
| 201 | 11 | 175 | 0 | 0 |
| 202 | 12 | 255 | 0 | 0 |
| 203 | 21 | 200 | 125 | 50 |
| 204 | 22 | 240 | 185 | 130 |
| 205 | 31 | 255 | 255 | 0 |
| 206 | 41 | 0 | 255 | 0 |
| 207 | 42 | 0 | 200 | 0 |
| 208 | 43 | 0 | 150 | 0 |
| 209 | 51 | 200 | 200 | 200 |
| 210 | 52 | 150 | 150 | 150 |
| 211 | 53 | 100 | 100 | 100 |
| 212 | 54 | 50 | 50 | 50 |
| 213 | 61 | 255 | 255 | 255 |
| 214 | 71 | 0 | 255 | 255 |
| 215 | 72 | 0 | 201 | 200 |
| 216 | 73 | 0 | 150 | 150 |
| 217 | 81 | 255 | 0 | 255 |
| 218 | 91 | 175 | 0 | 175 |
| 219 | 100 | 0 | 0 | 200 |

When you are done, the new subset should have colors resembling the original data, and the data previously entered into the master-raster should have the standard classification colors. Using the **colormod** program option, "r - RAINBOW file I/O," retrieve the rainbow file created in step 25 and save the color scheme using the name "unclassed." This action will replace the color scheme created in step 25 with the one currently on display.

30) Using the **colormod** program option "t - trailer update of GIS file," place the color scheme into the master-raster trailer file.

<u>Recode</u>

Review the previous cluster labels, make changes to the previous labels as needed to edge-match the subset to the master-raster, and recode the subset to the classification values by the following steps.

31) Fill-out the "Master-Raster Recode" tracking form. This form is used to record the original cluster labels and the new labels within the master-raster. The columns for "old" labels and for the "change" should contain numbers from the classification numbering scheme (11 = Developed - high intensity, 12 = Developed - low intensity, etc). The only entries in the "change" column should be for those clusters whose label will be changed (including the "recluster" clusters whose new value will be 200). The column marked "new" labels will correspond to a numbering scheme that ranges from 200 to 219 (see step 35 below).

    Enter the Path, Row, and Subset at the top of the form. Retrieve the "Cluster Labels" tracking form from the folder and fill out the "old label" column of the "Master-Raster Recode" tracking form.

32) This step will create a color scheme that reflects the classes as originally labelled. Display the master-raster using the ERDAS **display** program. Take the default reduction factor that allows the entire master-raster to be displayed. Using the ERDAS program **colormod** option "c - color palette entry" set the cluster colors to correspond to the label classes from the "Master-Raster Recode" tracking form (step 31). Use the "o - open new palette file" option and use the file: "/drs1/rsches/defaults/codes.dat." Modify the colors of the original clusters according to the following table:

| Label value | Color name | Red | Green | Blue | Description |
|---|---|---|---|---|---|
| 0 | black | 174 | 171 | 128 | Areas outside the watershed |
| 11 | 11 | 175 | 0 | 0 | Developed - High Intensity |
| 12 | 12 | 255 | 0 | 0 | Developed - Low Intensity |
| 21 | 21 | 200 | 125 | 50 | Cultivated - Woody |
| 22 | 22 | 240 | 185 | 130 | Cultivated - Herbaceous |
| 31 | 31 | 255 | 255 | 0 | Herbaceous |
| 41 | 41 | 0 | 255 | 0 | Woody - Deciduous |
| 42 | 42 | 0 | 200 | 0 | Woody - Mixed |
| 43 | 43 | 0 | 150 | 0 | Woody - Evergreen |
| 51 | 51 | 200 | 200 | 200 | Exposed - Soil |
| 52 | 52 | 150 | 150 | 150 | Exposed - Sand |
| 53 | 53 | 100 | 100 | 100 | Exposed - Rock |
| 54 | 54 | 50 | 50 | 50 | Exposed - Evaporite Deposits |
| 61 | 61 | 255 | 255 | 255 | Snow & Ice |
| 71 | 71 | 0 | 255 | 255 | Woody Wetlands - Deciduous |
| 72 | 72 | 0 | 200 | 200 | Woody Wetlands - Mixed |
| 73 | 73 | 0 | 150 | 150 | Woody Wetlands - Evergreen |

33) Using the **colormod** program option "r - RAINBOW file I/O," retrieve the rainbow file created in step 25 and save the color scheme using the name "old."

34) The entire edge between the new subset and the previously added data must be visually inspected to determine if changes to the cluster labels must be made. Display the master-raster using the ERDAS program **display** with a magnification factor of 1. This step must be repeated until all portions of the subset boarder are visited.

During this step, determine whether a change in a cluster labels can be made in a way that minimizes the differences between subsets. Look for areas of homogenous cover type that straddle the edge and make sure that there is no difference between subsets. Any changes to a cluster label must be noted on the "Master-Raster Recode" tracking form. Changes to the cluster labels should be kept to a minimum. Remember that the original analyst used a variety of reference material and studied areas throughout the scene to determine the original cluster labels. Before making label changes, consider the original analyst's notes and the impact of the changes.

There are no step-by-step instructions for this process. The following is a list of routines, programs, and processes may be of some use.

- Use the **colormod** option "r - RAINBOW file I/O" to retrieve the rainbow file created in step 25 and updated in step 33. Toggling between the two color schemes may be useful for interpreting the image.
- If you decide on a label change, make a new color scheme reflecting that change and save it in a temporary rainbow file, along with the "old" scheme from the rainbow file list above. Use the same technique as above to toggle between the "old" and "new" color schemes. NOTE: A bug in the **colormod** program causes problems in saving the correct color scheme with the correct name. In

the past, this bug has resulted in loss of the entire file. Since you will be saving the rainbow file created in step 25, it is suggested that another temporary rainbow file be made for this technique.

- The ERDAS program **curses** can be used to find the cluster number that may need changing.
- Keep the color scheme from step 30 in the trailer of the master-raster. This action will make it easier to see the edge when a new section is displayed.
- If you are advancing the display along a relatively vertical edge of the subset, use the "Keyboard" option when entering file coordinates in **display** and simply add or subtract 1000 to the Y coordinate. Similarly, along relatively horizonal edges, add or subtract 1000 to the X coordinate.
- Look for notes on the "Cluster Labels" tracking form that may indicate which clusters the analyst had problems labelling. Sometimes these notes can indicate an alternative label that may be a better match. Discuss the changes with the original analyst to determine if the change is inappropriate.

This step must be repeated so that the entire edge of the subset is viewed at a magnification factor of 1.

35) Fill out the "new label" column of the "Master-Raster Recode" tracking form. This column will have an entry for every cluster in the file. Most of the entries should be a simple translation of the "old label" column values into the "new label" column values using the table below. The exceptions are those clusters with an entry in the "change" column. The following new values will be used as the master-raster class numbers:

| New value | Old value | Description |
|---|---|---|
| 0 | 0 | Areas outside of the watershed |
| 200 | 255 | Unclassed areas within the watershed |
| 201 | 11 | Developed - High Intensity |
| 202 | 12 | Developed - Low Intensity |
| 203 | 21 | Cultivated - Woody |
| 204 | 22 | Cultivated - Herbaceous |
| 205 | 31 | Herbaceous |
| 206 | 41 | Woody - Deciduous |
| 207 | 42 | Woody - Mixed |
| 208 | 43 | Woody - Evergreen |
| 209 | 51 | Exposed - Soil |
| 210 | 52 | Exposed - Sand |
| 211 | 53 | Exposed - Rock |
| 212 | 54 | Exposed - Evaporite Depcsits |
| 213 | 61 | Snow & Ice |
| 214 | 71 | Woody Wetland - Deciduous |
| 215 | 72 | Woody Wetland - Mixed |
| 216 | 73 | Woody Wetland - Evergreen |
| 217 | 81 | Herbaceous Wetland |
| 218 | 91 | Non-Vegetated Wetland |
| 219 | 100 | Water |

36) The program **recode2** will be used to change the pixel values from the cluster numbers to the master-raster classification numbers. Create a parameter file to be used for recoding the portion of the master-raster containing the subset. This file should be called "recode.pfile" and contain the following (correct entries on right):

Line 1: The input file name ....... /drs4/rsches/master-raster/master.gis

Line 2: A default recode value for pixels with values other then those in the recode table starting on line 4. Any negative value specifies that these pixels retain their original values .................... 200

Line 3: The following parameters must be separated by commas or spaces:
- element 1: line to begin processing .................. see step 25
- element 2: column to begin processing ............... see step 25
- element 3: line to end processing ................... see step 25
- element 4: column to end processing ................. see step 25

Line 4 - end of file: Starting in line 4, each line should have two integers separated by commas or spaces.
- element 1: old pixel value ......................... see step 28
- element 2: new pixel value ......................... see step 28

Your file should be longer, but similar to the following:

```
------------------------------------
/drs4/rsches/master-raster/master.gis
200
9369,20795,14607,23517
0  0
1  206
2  206
3  204
4  204
5  208
6  205
7  205
8  216
.  .
.  .
.  .
255 200
------------------------------------
```

Lines 1, 2 and 4 should always be the same, as above. The line 3 parameters can be taken from the "Combine Files Into Master Raster" tracking form (see step 25). The file coordinates in line 8 must be entered in reverse order then are listed on the tracking form (i.e. Y,X,Y,X not X,Y,X,Y). The file should have a line for every cluster number. All "new" values (the right column) must be between 200 and 219. The last line should be to change pixels with 255 (unclassed) to new class 200.

37) Obtain a printout of the above (step #) parameter file and compare it to the "Master-Raster Recode" tracking form. Make sure the contents of this file are correct before proceeding to the next step. File the hardcopy in the folder for this subset.

```
sunss03% lpr recode.pfile
```

38) Run the **recode2** program to change the values in the master-raster.

```
sunss03% recode2 recode.pfile
```

Archive

Rather then making separate backup 8-mm tapes for each subset, all important interim files will be kept on this system until all subsets are completed. The entire master-raster directory will be archived at once.

39) Remove the original files that were retrieved from archive in step 5. Also remove any temporary files that you may have generated during processing. Use the checklist on the "Combine Files Into Master Raster" tracking form to assure you delete the unnecessary files.

```
sunss03% rm sub1b*
```

40) Compress all files in the directory.

```
sunss03% compress *
```

41) Use the checklist on the "Combine Files Into Master Raster" tracking form to make sure all files and hardcopy outputs exist.

COMBINE FILES INTO MASTER RASTER

Path:______________ Row:___________ Subset:___________

Retrieve file

8-mm tape number:___________________________________________________

".gis" file name:___________________________________________________

".trl" file name:___________________________________________________

output file:___________________________________________________

Project to Albers

old UTM coordinates zone:_______

upper left: _____________________________,______________________________

lower right: _____________________________,______________________________

old file size

number of rows:______________ number of columns:______________

parameter file (file_______)

mapcon.log file (file_______)

new Albers coordinates

upper left: _____________________________,______________________________

lower right: _____________________________,______________________________

Albers coordinates rounded to multiple of 50

upper left: _____________________________,______________________________

lower right: _____________________________,______________________________ parameter file (file________)

output ".gis" file name:________________________________________________

<u>Subset</u>

Master file coordinates

upper left: ______________________________,______________________________

lower right: ______________________________,______________________________

<u>Recode</u>

re-code parameter file (file________)

File Checklist:

_____GEOPNTS.DAT
_____albers1.dig
_____albers1.pro
_____albers2.dig
_____geo.cfile
_____geo1.pfile
_____geo2.pfile
_____mapcon.log
_____recode.pfile
_____tm1533s1.gis
_____tm1533s1.rnb
_____tm1533s1.trl
_____utm1.dig
_____utm1.pro

Printout Check List:

_____geo1.pfile
_____mapcon.log
_____geo2.pfile
_____recode.pfile

## MASTER-RASTER RECODE

Path________ Row________ Subset________

| | new label | old label | change | notes |
|---|---|---|---|---|
| 1: | | | | |
| 2: | | | | |
| 3: | | | | |
| 4: | | | | |
| 5: | | | | |
| 6: | | | | |
| 7: | | | | |
| 8: | | | | |
| 9: | | | | |
| 10: | | | | |
| 11: | | | | |
| 12: | | | | |
| 13: | | | | |
| 14: | | | | |
| 15: | | | | |
| 16: | | | | |
| 17: | | | | |
| 18: | | | | |
| 19: | | | | |
| 20: | | | | |
| 21: | | | | |
| 22: | | | | |
| 23: | | | | |
| 24: | | | | |
| 25: | | | | |

| | new label | old label | change | notes |
|---|---|---|---|---|
| 26: | | | | |
| 27: | | | | |
| 28: | | | | |
| 29: | | | | |
| 30: | | | | |
| 31: | | | | |
| 32: | | | | |
| 33: | | | | |
| 34: | | | | |
| 35: | | | | |
| 36: | | | | |
| 37: | | | | |
| 38: | | | | |
| 39: | | | | |
| 40: | | | | |
| 41: | | | | |
| 42: | | | | |
| 43: | | | | |
| 44: | | | | |
| 45: | | | | |
| 46: | | | | |
| 47: | | | | |
| 48: | | | | |
| 49: | | | | |
| 50: | | | | |

| | new label | old label | change | notes |
|---|---|---|---|---|
| 51: | | | | |
| 52: | | | | |
| 53: | | | | |
| 54: | | | | |
| 55: | | | | |
| 56: | | | | |
| 57: | | | | |
| 58: | | | | |
| 59: | | | | |
| 60: | | | | |
| 61: | | | | |
| 62: | | | | |
| 63: | | | | |
| 64: | | | | |
| 65: | | | | |
| 66: | | | | |
| 67: | | | | |
| 68: | | | | |
| 69: | | | | |
| 70: | | | | |
| 71: | | | | |
| 72: | | | | |
| 73: | | | | |
| 74: | | | | |
| 75: | | | | |

| | new label | old label | change | notes |
|---|---|---|---|---|
| 76: | | | | |
| 77: | | | | |
| 78: | | | | |
| 79: | | | | |
| 80: | | | | |
| 81: | | | | |
| 82: | | | | |
| 83: | | | | |
| 84: | | | | |
| 85: | | | | |
| 86: | | | | |
| 87: | | | | |
| 88: | | | | |
| 89: | | | | |
| 90: | | | | |
| 91: | | | | |
| 92: | | | | |
| 93: | | | | |
| 94: | | | | |
| 95: | | | | |
| 96: | | | | |
| 97: | | | | |
| 98: | | | | |
| 99: | | | | |
| 100: | | | | |

| | new label | old label | change | notes |
|---|---|---|---|---|
| 101: | | | | |
| 102: | | | | |
| 103: | | | | |
| 104: | | | | |
| 105: | | | | |
| 106: | | | | |
| 107: | | | | |
| 108: | | | | |
| 109: | | | | |
| 110: | | | | |
| 111: | | | | |
| 112: | | | | |
| 113: | | | | |
| 114: | | | | |
| 115: | | | | |
| 116: | | | | |
| 117: | | | | |
| 118: | | | | |
| 119: | | | | |
| 120: | | | | |
| 121: | | | | |
| 122: | | | | |
| 123: | | | | |
| 124: | | | | |
| 125: | | | | |

| | new label | old label | change | notes |
|---|---|---|---|---|
| 126: | | | | |
| 127: | | | | |
| 128: | | | | |
| 129: | | | | |
| 130: | | | | |
| 131: | | | | |
| 132: | | | | |
| 133: | | | | |
| 134: | | | | |
| 135: | | | | |
| 136: | | | | |
| 137: | | | | |
| 138: | | | | |
| 139: | | | | |
| 140: | | | | |
| 141: | | | | |
| 142: | | | | |
| 143: | | | | |
| 144: | | | | |
| 145: | | | | |
| 146: | | | | |
| 147: | | | | |
| 148: | | | | |
| 149: | | | | |
| 150: | | | | |

**Photographic Accuracy Assessment Form**

Sample Site #: __________

Landsat Scene ID #: ________________________ Projection: __________

Classified Data Set: ______________________ Projection: __________

Classification System: ____________________

Classification Accuracy Table File: ________________________________

Photo Coverage .dig File: _________________ Projection: __________

Date of Photograph: ____________ Frame #: ________________

Stereo Coverage: _______________ Photo Media: ________________

1:100,000 Scale Map: _____________________________

Photo Quality: ______________________________________________________

Primary Class: _________ Secondary Class: _________

Sample Site Characteristics and Components:

______________________________________________________________________

______________________________________________________________________

______________________________________________________________________

______________________________________________________________________

______________________________________________________________________

Analyst: _____________________ Date of Interpretation: _________

## Sample Site Component Grid

Photograph Frame #: ________________ Sample Site #: __________

| | | |
|---|---|---|
| | | |
| | | |
| | | |

## CHESAPEAKE BAY WATERSHED METADATA

Data_set_identity: Chesapeake Bay Watershed Thematic Land Coverage/Land Use.
Theme_keywords: Chesapeake, thematic, land cover, land use, watershed.
Representation_model: Vector-topologic.
Spatial_object_types: Pixel/Grid
Native_data_set_size: 57 MB
Transfer_format: ARC Grid
Transfer_size: 57 MB
Transfer_format: ARC Grid
Transfer_size: 57 MB
Data_set_description: Landsat Thematic coverage of the Chesapeake Bay Watershed. Final data set includes ten thematic land cover categories at a 25 meter resolution. The final data set is in an ARC Grid format.
Intended_use: Data set to be used for the Chesapeake Bay Program Office's non-point pollution models. Also for use as a thematic map of the land cover within the Chesapeake Bay Watershed.
Data_set_extent: -257500,4507900,287400,3800500
Geographic_area: Chesapeake Bay Watershed
Intended_scale(s)_of_use: 24000,100000,250000
Resolution_of_data: 25 m
Projection_name: Albers Conical Equal Area
Horizontal_datum_or_ellipsoid: NAD83
Vertical_datum: NGVD
Projection_units: meters
Standard_parallel: 38.0
Standard_parallel: 42.0
Longitude_of_central_meridian: -77.5
Latitude_of_projection's_origin: 0
Coordinate_precision: Single
Contact_type: Source/Authority.
Contact_organization: U.S. Environmental Protection Agency, Environmental Monitoring and Assessment Program - Landscape Characterization.
Contact_person_title: Denice Shaw, Technical Coordinator.
Contact_mailing_address: U.S. Environmental Protection Agency, EMAP Center, Catawba Building, Research Triangle Park, NC 27711.
Contact_telephone: (919) 541-2698
Contact_email: denice.shaw@heart.epa.gov
Contact_instructions: contact for technical information via e-mail or regular mail
Contact_type: Distributor
Contact_organization: Customer Services, U.S. Geological Survey, EROS Data Center.
Contact_person: Customer Services
Contact_mailing_address: U.S. Geological Survey, EROS Data Center, Customer Services, Souix Falls, SD 57198
Contact_telephone: (605) 594-6511
Contact_instructions: Data are available on 8mm data tapes. Tape requests are filled at cost of duplication.
Transfer_mode: 8mm data tape

Transfer_instructions: Data is transferred in and ARC Grid format
Degree_of_digital_completion: Complete
Completion_status: Completed
Completion_date: 19940215
Percentage_complete: Complete.
Degree_of_availability: Complete.
Policy_status: Users may obtain these data at the cost of reproduction.
Copyright_status: Public domain.
Custodial_liability: Custodian does not assume liability.
Table_identity: chesgrid.vat
Table_definition: Polygon attribute table for land cover codes.
Table_definition_source: author
Attribute_identity: area
Attribute_definition: area measured in equal area meters.
Attribute_definition_source: software-defined
Attribute_table_identity: chesgrid.vat
Attribute_domain_value: positive real numbers.
Attribute_domain_value_definition: none
Attribute_format: real
Attribute_format_length: 12
Attribute_units_of_measure: square meters.
Attribute_authority: U.S. EPA, Environmental Systems Laboratory - Las Vegas
Source_name: Land Cover from Landsat Thematic Mapper imagery
Bibliographic_reference: U.S. Environmental Protection Agency, Environmental Monitoring Systems Laboratory - Las Vegas, 1994, EMAP Chesapeake Bay Watershed Pilot Project Final Report: U.S. Environmental Protection Agency, Office of Research and Development, Environmental Monitoring Systems Laboratory, Las Vegas, NV
Source_scale: 25 meter pixel units
Source_scale: N/A
Source_medium: Landsat Thematic Mapper digital data files in band sequential format
Creator_of_source: EOSAT Corporation, Lanham, Maryland
Date(s)_of_source_materials: 1988-1991
Source_projection: Universal Transverse Mercator (UTM), UTM Zones 17 and 18.
Final Projection: Albers Equal Area
Procedure: 1. The Landsat Thematic Mapper imagery was geocorrected by Hughes STX Corportation. The final data product was geocorrected to 25 meters. This data was then ship to the US. EPA's Environmental Systems Laboratory - Las Vegas for image processing.

2. The Landsat data was processed using a modified unsupervised image processing technique. The data was clustered using an unsupervised clustering algorithm. The data was then observed and the confusion clusters identified. These confusion clusters were then placed into the unsupervised lustering algorithm for reclustering of the spectral data.

3. After clustering the data was labelled, and the clusters recoded into the appropriate land cover category.

Procedure_date: 199404
Procedure_contact: Dorsey Worthy, Remote Sensing Program Manager, U.S. Environmental Protection Agency, 944 E. Harmon, Las Vegas, NV 89119, telephone (702) 798-2274
Positional_accuracy: +/- 15 meters

Positional_accuracy_method: Spatial Accuracy Test
Positional_accuracy_explanation: Landsat Thematic Mapper data met the spatial accuracy desired.
Attribute_accuracy: 80% overall within 85% confidence interval
Attribute_accuracy_method: Stratefied systematic random point photointerpretation with field validation.
Data_model_integrity: Data set contains thematic land cover categories for the Chesapeake Bay Watershed
Completeness: Complete.
Metadata_revision_date: 19940320
Metadata_contact: Dorsey Worthy, amdldw@vegas1.las.epa.gov