Cheryl A. Brantley, Manager Production Services
Rob Robinson, Senior Applications Manager
Michael Bradley, Image Processing Specialist
Andrew S. Bury, Applications Analyst
ERDAS, Inc.
2801 Buford Highway, Suite 300
Atlanta, Georgia USA 30329
(404)248-9000
(404)248-9400 FAX

ISPRS Commission IV

USING ANCILLARY DIGITAL DATA TO IMPROVE LAND COVER CLASSIFICATION

ABSTRACT

The initial field work tasks associated with land cover classification efforts are usually expensive and time-consuming. By using available digital data in conjunction with the imagery to be classified, labor intensive ground truthing can be reduced or even eliminated. Techniques have been developed to integrate data bases such as USGS DLG (Digital Line Graph), Census TIGER, and USGS LUDA (Land Use/Land Cover) with imagery to target the selection of training sample sites. In this paper, several case studies will be used to illustrate these techniques and introduce advanced image processing and geographic information systems (GIS) software functions which increase classification accuracy rates while decreasing the overall time needed to conduct major land cover analyses.

INTRODUCTION

Today, GIS users are feeling the pressure of demands for more accurate and up-to-date data bases to support a growing variety of applications. For years the primary input source for GIS has been paper topographic and thematic maps (soils, land use, etc.). However, as the industry grows and software applications become more advanced, there is an increasing need for more specialized information. As GIS plays a greater role in planning and decision making, there is an emphasis on data bases to display actual current conditions. GIS coverages can no longer afford to be 10, five, or even one year out of date. Yet the cost of manually updating a large data base on an annual basis, makes it difficult, if not impossible, for many federal, state and private sectors to meet these demands. Cutbacks in personnel and spending, coupled with legislation mandating that correct geographic information be maintained, are placing many agencies in a dilemma. Factors such as soils makeup, wetlands, endangered species, crops and forests must be monitored. Therefore, many agencies, both public and private, are turning to the use of satellite data as a reliable source to calculate current conditions of vegetative and anthropogenic information.

The use of satellite mapping for land use/land cover is most effective when compiled for a large study area. Over the last several years, ERDAS, Inc., has participated in several large scale projects which have ranged in size from four to 12 scenes of satellite data. In our experience with projects of this size, we have found that the use of ancillary data is essential for breaking the data into logical analytical components. The use of this ancillary information increases the speed of the process time and helps refine the accuracy of the data. Inexpensive and widely available data sources make it possible to greatly increase the speed, ease and accuracy with which a satellite land use/land cover can be completed. This paper will focus on ancillary data as a tool to be used in the successful creation of large scale land use/land cover classifications. It will also examine the characteristics of data used in ERDAS Production Services projects, and discuss the positive uses and difficulties associated with each data type.

ANCILLARY DATA AS A TOOL IN REMOTE SENSING ANALYSIS

Ancillary data can be described as any supplemental data that might be used to enhance or become an additional part or section of the primary data base. Ancillary data can range from aerial photography to out-dated GIS coverages or maps. Usually several different types of ancillary data may be used to examine satellite data. Below are examples of readily available ancillary data:

- USGS Digital Line Graph Data (DLG)
- United States Bureau of Census TIGER Data
- USGS Topographic maps
- USGS Digital Elevation Model (DEM)
- 1:250,000 Land Use and Land Cover Data (LUDA)
- United States Department of Agriculture - ASCS 35mm color compliance photography
- National High Altitude Photography (NHAP-2)
- National Aerial Photography Program (NAPP)

The data listed above is available through federal agencies and, with the exception of aerial photography, can be purchased at a relatively low cost. Also, as state and local agencies begin to understand the need to coordinate activities, more spatial data bases will become available to the private and public sectors. The cost to compile and create computer based GIS is forcing states to begin to encourage or mandate the cooperation and sharing of spatial data base information. This, in effect, will set up a framework where more private and public sector businesses can move into the GIS user community. Therefore, it is important that the GIS user be cognizant of digital data bases available in their particular area of interest. In many cases, such data information can be used to reduce the amount of work and associated cost required to complete a project.

DATA INTEGRATION

The use of digital ancillary data helps to increase the speed and accuracy of satellite image analysis. However, much of the detailed digital information which would be considered helpful is typically stored in a vector format. Since satellite data is stored in a raster format, there has been the problem of how to store two completely different data sources on one system. Recently, driven by the need to transfer data from one format to another, many software companies have implemented software which allows the user to purchase data in one file structure and convert it to a different software environment. This implementation has created an open market for agencies to purchase ready made data bases to update rather than re-create.

Software which allows the complete integration of vector and raster image data also allows for the creation of more innovative approaches to satellite analysis. Computer firms such as ERDAS and ESRI, Inc., (Redlands, California) have linked forces to create recent software advances which allow vector and raster data bases to reside within the same software environment. This capability allows raster and vector coverages to be used interchangeably in a way that is transparent to the user. Advances include the ability to display spatial vector data on top of spatial/spectral image satellite data. Once integrated, these data can be used interchangeably; vector data can be used to enhance raster data for image processing, or raster data can be used to update and enhance vector data to reflect current land conditions.

PROCESSING SATELLITE IMAGERY USING ANCILLARY INFORMATION

The value of the information derived from satellite data is directly related to the spatial/spectral quality of that data and the scientific methodology used to extract the information. Regardless of the quality of the data, if the methodology used to analyze the data is weak, the information derived from the data will be weak. Therefore, it is very important that the image analyst use every resource available to extract information and then later verify it in accuracy assessments.

In working with large satellite coverage study areas, the methodology for image processing, classification, or data base development must be well thought out and focused on the project rationale. A project map or progress chart should be created by the project manager and image processing team. Once the final objectives of the image product have been established, a list of potential ancillary products and their potential uses should be listed. Each set of ancillary data should be discussed as to the relevancy of its use and how much additional work may be needed to bring the data into the system environment. Issues of concern include: How much will the data cost? How much time is needed to convert the data? Will it be necessary to purchase new software to complete the conversion? What level of personnel experience will it take to complete the data conversion? Each of these issues should be considered and factored into the overall time needed to complete the image analysis.

The following are two image processing case studies completed by ERDAS Production Services. These two studies were generated under completely different circumstances and required completely different types of ancillary data. The first case study was a typical classification study using standard types of ancillary information commonly used by today's image analyst. However, the ability to convert information from raster to vector formats speeded up the image processing stage. The second image processing study was unique in that 50% or more of the classification was completed using ancillary data, the time needed to complete the project was cut in half, and no ground truthing was completed.

Case Study One - Wetlands Classification of Georgia

The Georgia Department of Natural Resources (DNR) contracted with Production Services to produce a full land cover classification of the state for the purpose of identifying wetlands. The maps resulting from the classification effort were to be used as a reference tool in development offices throughout the state to identify and protect ecologically fragile areas.

Due to the size of the area to be mapped it became apparent that satellite imagery would allow the quickest, most cost-effective method of attaining wetland information. The study area consisted of 11 full and two one-quarter scenes of EOSAT Thematic Mapper data, which stretched across 10 distinct physiographic regions. Because limited aerial photography was available for training sample information or accuracy assessment, the Georgia DNR agreed to provide this information in the form of field investigation. Because of the size of the area it was decided that the following ancillary data would be used:

- 1:24,000 USGS Topographic Maps (1016 total)
- USGS DEM Data
- NHAP Aerial Photography (ordered for major waterways)
- LORAN Helicopter Coordinates
- Physiographic Map of Georgia, 1976
- ARC/INFO™ Map Grid

The state of Georgia is unique in that it is covered by several very distinct physiographic regions. These region differences, combined with the difference between each scene's date and path complicated the classification task further. As a result, it was decided that the state would be divided into regions based on a composite of the major physiographic regions and the Landsat satellite path. Division according to physiographic regions was done to simplify the classification by isolating some of the endemic geomorphic and vegetative differences which occur between regions.

To complete this task, scenes along the same path were mosaicked and then masked according to the region boundary lines. Each region was classified separately as a unique study area and assessed for accuracy by region. At the conclusion of the project the state of Georgia was completed in 14 separate classifications. Breaking the data into unique logical components which could be analyzed separately put the study in perspective for the project team. It also limited the overwhelming amount of image processing to be done and set up a cycle for work progression. The project was completed at an overall confidence level of greater then 85% for the entire state, and several regions reached accuracy levels of 90%.

Other ancillary data was used throughout the project. The mountainous areas of North Georgia required extensive use of topographic maps, GIS techniques to extract classification information from shadowed areas, and a limited amount DEM data. Final classification accuracy assessments were completed using GIS coordinates converted to LORAN helicopter coordinates.

Services to the Georgia DNR included the full image classification and production of raster GIS files of the 7 1/2 minute quad areas of Georgia (1,016 total), color electrostatic

hardcopy (three copies) of each quad, and conversion services of raster files into ARC/INFO vector format.

Case Study Two - Lake Michigan Ozone Study (LMOS)

To quantify the ozone source-receptor relationships in the area surrounding Lake Michigan, a non-profit corporation of the four states surrounding Lake Michigan jointly funded an emissions study for the Lake Michigan area. The study required development of a current land use/land cover data base of Illinois, Wisconsin, Indiana and Michigan, and was governed by the Lake Michigan Air Director's Consortium. This data provided source information to geographically locate ozone emissions and to estimate biogenic emissions from a variety of vegetative sources.

Due to the legal ramifications of this project, the Consortium required that the land use/cover classification be completed within a 10 month period. This time line allowed a maximum of four weeks, per scene, for all land use/cover processing and accuracy assessment. For this reason it was necessary to explore the use of ancillary data as a resource for breaking the data into logical components that could be analyzed independently.

It is commonly known that certain urbanized and non-urbanized classes reflect visible and infrared light in a very similar spectral signature. Previous experience had determined that separating these areas would minimize any possible confusion of these classes. For example, quarries, bare asphalt and beach usually reflect light similarly, therefore they were separated before automated classification techniques were implemented.

The ancillary data used for this project were grouped into three categories: aerial imagery, analog maps, and digital data sources. These digital data sources included four different types of digital data:

- •USGS 1:100,000 scale Digital Line Graph (DLG)
- •1:250,000 Land Use and Land Cover Data (LUDA)
- •1:250,000 scale Digital Elevation Model (DEM)
- •United States Bureau of Census 1990 Post-Census TIGER data

Aerial photography included the use of two types of aerial photography and one type of slide film. Aerial photography included photography from the National High Altitude Photography (NHAP-2) program and National Aerial Photography Program (NAPP). This photography was acquired to provide as much coverage as possible while adhering to budgetary constraints. The photography served a dual purpose: first, to provide a data source for initial classification accuracy, and second, to serve as a source of ground truth when computing the final land use and land cover classification accuracy assessment. The aerial coverage was divided into two data sets. The first set was used to extract signatures and perform initial accuracy assessments. The unused photography was set aside to perform the final accuracy assessments.

As an enhancement to the national photography acquired, United States Department of Agriculture - ASCS 35mm color compliance photography was ordered as well. This film proved valuable as a source of information to support the aerial photography. In addition, ASCS-578 and ASCS-156EZ crop report information was also available for this area. Individual county ASCS offices provided lithographed "photomaps" delineating each field and specific crop type for any desired year. Because of the dynamic nature of agricultural practices (i.e., crop rotation), this information provided an accurate representation of crop types for various areas. This information greatly enhanced the ability of the analyst to identify crop type and increase classification accuracy.

The LMOS project required that a significant amount of TM data be processed quickly and accurately. Therefore it was of great importance that the methodology for the classification be established and adhered to at the onset of the project. The Consortium requested that an initial test classification be completed which set the methodology for the complete 10 month study. This classification effort was unique in that a very complex set of pre-classification procedures were established to allow the image analyst to immediately begin processing. The data were processed by project teams with assigned tasks for completion of various phases of the project.

Pre-processing included tasks as simple as loading data onto the computer system and analyzing each band for data anomalies and cloud cover. Each data set was verified for acceptability prior to pre-processing. Once the data were verified, they were then clipped to data boundary limits and separated into urban and non-urban data sets using Census TIGER political boundaries or Place Boundaries. Although TIGER Place Boundary information is very general, it allows for the accurate separation of a majority of the data into urban versus non-urban data sets. The non-urban areas were also carefully checked for small urban areas which would not have been represented in the TIGER files. These small urban areas were separated with on-screen "heads up" digitizing using the ERDAS-ARC/INFO Live Link™.

Once the data were separated into urban and non-urban sets, the data were processed into principal components using a standard Principal Components Analysis (PCA) algorithm. The PCA was used primarily as a data reduction technique to eliminate redundant spectral information, thereby reducing the amount of data and speeding computer processing time. When the principal components analysis was completed, a varimax rotation was applied to the data. The varimax rotation manipulated the transformation coefficients to correlate more closely with specific bands in the original TM data. This allowed the analysts to more easily interpret what a specific component band represented.

After the data sets were separated and the principal components analysis completed, the data were processed through the ERDAS software program ISODATA (Iterative Self Organizing Data Analysis Technique). ISODATA generates a set of mean vectors and covariance matrices for each distinct spectral cluster. This unsupervised classification algorithm was used to derive a set of clusters which represented general features within the data set. The clusters were then used by the analyst to determine and refine manual signatures used in the final supervised classification.

This unique pre-processing classification methodology allowed for a great deal of repetitive process work to be completed prior to analysis by a professional. While other scenes were in the classification phase of the project, the pre-processing was completed. The image analysts were therefore able to move from one scene to the next with little distraction, establishing a tight progress cycle.

The final project classification utilized numerous levels of ancillary information. The addition of this information provided an excellent opportunity to utilize the latest technologies in GIS modeling. The GIS modeling package, GISMO™, was used to assure that various procedures were performed under consistent conditions each time the process was run. This allowed the project teams to more quickly complete multiple GIS processes in an efficient manner. In addition, use of this model assured that scenes matched in overlap areas.

Post-classification techniques were primarily GIS modeling functions using much of the ancillary digital information

previously defined. All ancillary digital information was converted to ARC/INFO and carefully verified with the ERDAS-ARC/INFO Live Link™. The ancillary data were then carefully modified as needed.

One example of the post-classification modeling was the reintroduction of linear DLG transportation information into the final classification because clustering often delineated the transportation network inconsistently. The clumping algorithms within supervised and unsupervised classifications have a tendency to classify roads bounded by heavy vegetation as non-urban. To correct this problem, roads and railroads were reintroduced as DLG's and gridded back into the final classification. DLG hydrology information was also used extensively to emphasize linear streams and to produce a lowland mask. Early in the project it became apparent that the signatures of lowland forest types and upland forest types were being confused in the classification. To correct this, a GISMO model was created to mask lowland forest and assign the information to the correct class. This allowed for the accurate delineation of the lowland rorest classes. The orchard class was problematic throughout the project study area as well. Because the spectral nature of orchard classes is similar to that of forest classes, orchard classes were extracted from USGS LUDA data and then added to the final classification.

Overall confidence levels were calculated yielding a lower limit of 85.01% and an upper limit of 87.40%. In addition to the overall percentage accuracy of 86.25%, a large number of test samples were taken to ensure a narrow range of uncertainty with a confidence range of only 2.39%. The lower limit of 85.01% for the 95% confidence range met the 85% accuracy criterion for the project, serving as a conservative estimate of the overall accuracy for the classification.

In using ancillary data as digital sources to aid the image processor in data analysis, it is important to fully understand the limitations of the data. Much of the data available today is generally outdated, very course in nature, or was created under lax digitizing standards. For instance, Digital Line Graph data (DLG) must often be rectified to the satellite image with which it will be used. Tiger Place boundaries provide only course urban boundary information and often do not take into account small rural town or urban fringe. For these reasons the user must give much thought as to how to use ancillary data and never rely on this information as the sole source of identification.

CONCLUSION

With careful planning and forethought at the beginning of a project, it is possible to use ancillary data to extract general information from satellite data. These general areas can then be analyzed in a more intense fashion, while other extraneous data are suppressed, (i.e., breaking out urban data from non-urban). The primary objective in the use of ancillary data is to increase the speed and accuracy of image analysis. It is also essential in a large project to reduce the data to managable pieces of information which can be processed and assessed for accuracy in a logical, timely progression.

The case studies discussed in this paper show but two ways ancillary data can be used to optimize image processing time and analysis. They are examples of the unique ways that raster and vector information can be integrated to enhance one another. With the integration of raster and vector sources available today, the creation of large scale satellite imagery classifications can be completed with more accuracy and faster than ever before.