ADVANCES IN LAND COVER CLASSIFICATION FOR APPLICATIONS RESEARCH: A CASE STUDY FROM THE MID-ATLANTIC RESAC

Dmitry L. Varlyguin, Robb K. Wright, Scott J. Goetz, Stephen D. Prince
Mid-Atlantic Regional Earth Science Applications Center
Department of Geogr aphy
University of Maryland
College Park, MD 20742-8225
resac@geog.umd.edu

ABSTRACT

Regional land cover information is needed for a wide variety of applications, ranging from land use planning to water quality modeling. Multi-temporal Landsat-7 ETM+ imagery, extensive field measurements, historical air photos, and various other digital geographic data sets are being acquired by the Mid-Atlantic Regional Earth Science Applications Center (MA-RESAC), a NASA-sponsored activity centered at the University of Maryland (UMD). The data are being used to develop, among other products, a current land cover map of the Chesapeake Bay Watershed. Coordinated use of these data sets, combined with a decision tree classification approach, permits improved discrimination of land cover types of interest to a wide variety of applications, particularly nutrient transport models of relevance to Chesapeake Bay Watershed restoration efforts. This paper focuses on the development of the data sets and techniques required for the activity, and has relevance to other broad-area land cover mapping efforts.

INTRODUCTION AND BACKGROUND

The MA-RESAC was established to facilitate interactions between researchers and a broad base of end-users tasked with resource management in the mid-Atlantic region. In order to develop innovative applications of remotely sensed data and associated technologies, the MA-RESAC has advanced several key demonstration projects, and focused on a set of core projects that cut across a broad range of science applications (Goetz *et al*., 2000a). Specific applications being addressed by RESAC partners include forest management (resource inventory, stand quality and health, reforestation, harvesting), biological resources (habitat quality, fragmentation, wetlands), agricultural practices (cropping patterns, nutrient management, conservation easements, riparian zone buffers), and land use planning (suburban sprawl, growth trends, policy regulations and incentives). Each of these require reliable information on land use / land cover (LULC) patterns and trends, and associated process models useful for monitoring and planning. The ultimate goal of the RESAC is to facilitate the interaction between the users and developers of LULC information, and associated geospatial technologies, as well as to participate in product development to the extent possible with modest resources and a limited timeline.

The mid-Atlantic region (Figure 1) includes the entire 170,000 km^2 Chesapeake Bay Watershed (CBW), comprised of 37 physiographic provinces and a complex mosaic of land cover types, farming practices and land use management strategies. We have focused the CBW-wide component of the mapping activities being conducted by the MA-RESAC on the use of Landsat-7 Enhanced Thematic Mapper (ETM+) imagery, which became available in mid-1999. Among the advantages of ETM+ is that the data are available soon after acquisition at a reasonable cost to users ($450 to $600 per 183 km x 170 km scene). This has made it more feasible to use multi-temporal imagery, which has important implications for land cover mapping and makes it possible to work with more recent images when conducting fieldwork. Other data sets being used in this activity include digital elevation models (DEM), digital soil surveys, NASS crop statistics, digital orthophotos, county-level planimetric maps, and various other ancillary data in a geographic information system (GIS) framework. Use of these diverse data sets, in combination with extensive field measurements, are expected to be useful for LULC discrimination, particularly for spectrally similar classes such as croplands, grasslands, and pastures. These LULC types have previously been difficult to map accurately (Vogelmann *et al*., 1998).

Figure 1. The mid-Atlantic region depicted with unprocessed ETM+ images, showing field sites used in development of LULC map products.

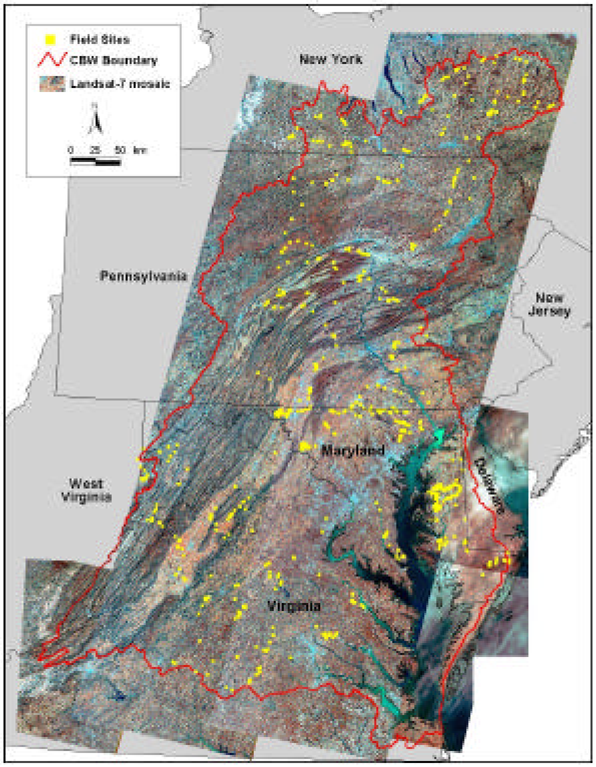


IMAGE PROCESSING AND CLASSIFICATION METHODOLOGY

As part of the LULC mapping effort the MA-RESAC has entered an agreement with the UMD Earth Science Information Partnership (ESIP), the UMD Landsat-7 Science Team, and the Landsat-7 Project Science Office to purchase all relatively cloud-free ETM+ imagery from July 1999 through December 2000. In addition, the RESAC has access to historical Landsat TM imagery of the region acquired by the Multi Resolution Land Cover Characterization (MRLC) project from the early 1990's, as well as EarthSat corporation's GeoCover imagery for the same time period. All of the ETM+ images have been or are being georeferenced to UTM coordinates at UMD using the orthocorrected GeoCover database as a reference.

Table 1. Landsat-7 ETM+ imagery acquired to date.

Stage	Period	#Scenes
1	1 July to 31 Dec 99	56
2	1 March to 31 July 00	32
3	1 August to 30 Sept 00	4
4	1 October to 30 Nov 00	25

The consortium has acquired 117 ETM+ images (Table 1) and the MA-RESAC is processing the data with the goal of having complete orthocorrected, radiometrically consistent, and mosaiced image coverage of the CBW for each of four seasons between Summer of 1999 and Autumn of 2000. Identifying the ETM+ scenes for purchase required visually reviewing over 300 browse images of the region. The scenes were selected relative to temporal distribution, cloud cover, and data quality. Whenever possible, scenes of similar dates were acquired for consistent phenological stage information. In some cases temporally consistent data were not available due to cloud cover, particularly in the Summer of 2000 which was anomalously cool and wet. Much of the remainder of this paper describes the processing of these data.

Image Geometric and Radiometric Corrections

We acquired ETM+ imagery as Level 0 data from the USGS EROS Data Center (EDC). The UMD Research Environment for Advanced Landsat Monitoring (REALM), a collective Landsat Science Team effort housed at UMD, processed the Level 0 data to Level 1G. The Level 1G output parameters for all 36 ETM+ scenes were set to the Space Oblique Mercator projection, with a satellite-up orientation and a pixel size of 28.5 meters for bands 1-5 and 7. This configuration of data parameters retains as much of the original geometric and radiometric properties as possible.

Topographic Correction

For project processing the ETM+ scenes were co-registered to the GeoCover data set. The GeoCover data set is a product of the NASA Scientific Data Buy program and produced by the EarthSat Corporation to create a global mosaic of orthocorrected Landsat imagery. The horizontal and vertical controls for this product were provided to EarthSat by the National Imagery and Mapping Agency (NIMA). As a global product, the GeoCover data set has a horizontal root mean square error (RMS) specification of 50 meters.

An elevation model incorporated in the rectification corrected for terrain displacement as viewed by the ETM+. The Appalachian Mountains run along the western boarder of the CBW, and significant terrain displacement can occur within an image. For consistency, all ETM+ images were orthocorrected. The National Elevation Database (NED), compiled by the USGS, was used as the DEM source for the orthocorrection of the ETM+ CBW images. This product has a resolution of one arc-second and is the most accurate DEM available to the public that encompasses the entire CBW. NIMA control points and the digital elevation model (DEM) are not yet available to the public.

Our analysis utilized a temporal series of images and required that co-registration be within 0.5 pixels. A larger pixel offset reduces the opportunity for land surface features to overlay correctly. ERDAS Imagine GIS software was used to co-register and orthorectify the ETM+ images. The Landsat orthocorrection model requires a DEM and a minimum of five control points. The RMS for all images was less than 0.5 pixels, and the nearest neighbor method was used for image resampling. Co-registration quality control was done through visual comparison of the ETM+ images with the corresponding GeoCover scene. All resulting planimetric errors were less than 0.5 pixels, representing an area of less than 15 meters on the ground. The highest error occurred in mountainous areas where differences in the DEM sources used by our orthocorrection and that of the GeoCover product occur.

Radiance Calibration

Pixel values in commercially available imagery represent the radiance of the surface in the form of Digital Numbers (DN), which are calibrated to fit a certain range of values. In the case of ETM+ imagery these radiance values are scaled to numbers between 0 and 255. Conversion of DN back into absolute radiance is a necessary procedure for comparative analysis of several images acquired at different times. In the case of Landsat-7, the DN values in overlapping regions of same-track and same-day acquisitions may not be equal due to a possible change in the gain used to scale an individual image (Irish, 2000). Conversion to radiance values allows for a more accurate comparison of images across rows, paths, and dates. ETM+ imagery was collected for 12 scenes over the CBW, in 4 neighboring paths.

The conversion from DN to spectral radiance (L_λ) uses the calibration coefficients included in each scene's metafile as:

$$L_\lambda = (L_{MAX} - L_{MIN}) / 255 \bullet DN + L_{MIN} \qquad \text{(Eq. 1)}$$

Where L_{MIN} and L_{MAX} are the minimum and maximum spectral radiance values (calibration coefficients) in the scene (Irish, 2000), which are given for each scene in the accompanying metadata.

Top of Atmosphere Reflectance

The reflectivity of the earth surface varies according to the positional relationship between the Earth and the Sun. The elliptical nature of the Earth's orbit varies the amount of solar irradiance. Solar irradiance at a point on the earth is also dependent upon the solar elevation, which varies according to the time of the year, the location of the point, and the time of image acquisition. Conversion to top of atmosphere or exoatmospheric reflectance was used to adjust the ETM+ reflectance values to include these additional parameters and is useful when combining imagery of varying dates and different sensors.

The conversion to exoatmospheric reflectance (ρ*) utilized the methods described in more detail by Goetz (1997). This conversion is calculated as:

$$\rho^* = (\pi \bullet L_\lambda \bullet d^2) / (\cos(\Theta) \bullet E_o) \quad \text{(Eq. 2)}$$

where d is the earth-sun distance in astronomical units and varies according to the Julian day, Θ is the sun elevation as given in each scene's metadata, and E_o is the mean ETM+ solar spectral irradiance value per band as given in the ETM+ handbook (Irish, 2000).

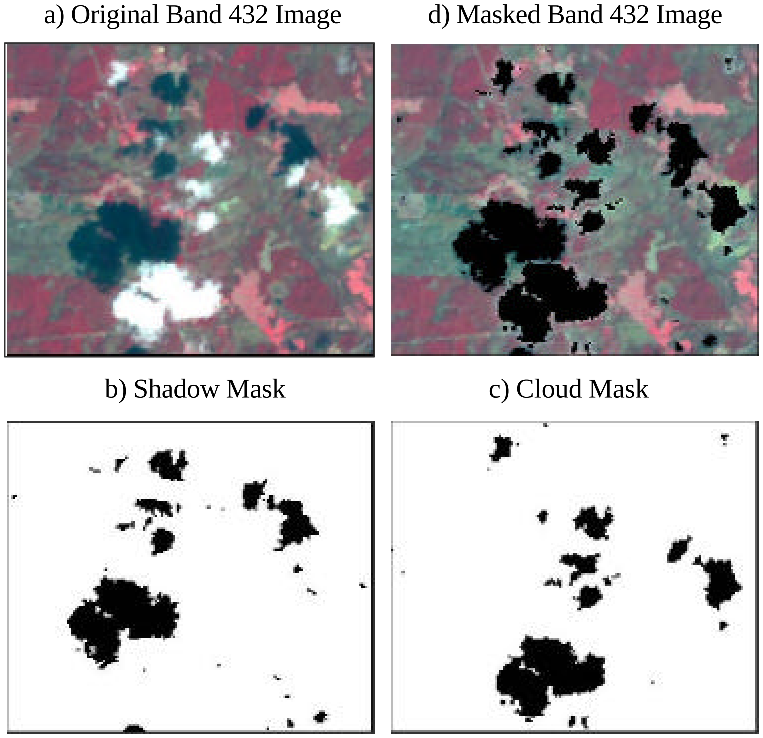
Scene Mosaicing

Land cover classification of an area larger than one scene can benefit from image mosaicing. While scenes of the same date (i.e. the same Landsat path) can be mosaiced, provided they are first radiometrically calibrated, joining scenes of different dates requires additional consideration. We used a linear regression technique for the overlap area between two adjacent Landsat rows, from which clouds and cloud shadows were masked, to adjust the individual band data to a common value among adjacent scenes in the path. In order to normalize imagery for time differences only scene elements representing the same land cover in both dates were sampled.

Cloud and Cloud Shadow Masks

For land cover classification algorithms to operate properly cloud and cloud shadows must be masked from the image statistics used to train the classifier. We used the first two components of a principal components analysis (PCA) developed from band 1, 2, 3 (PCA123) and 4, 5, 7 (PCA457) combinations. Cloud shadow masks were developed by differencing PCA457 and PCA123, assigning a zero value to all scene elements with positive values and a value of one to the remainder. Cloud masks were created in a similar manner, in which PCA123 values were numerically inverted and differenced with the PCA457 components. Binary image values were assigned as with the cloud shadow mask. The two masks were then combined and applied to the scene overlap areas used for band-by-band image mosaicing.

Figure 2. Cloud and shadow masks derived from and applied to a portion of an ETM+ image.


a) Original Band 432 Image

d) Masked Band 432 Image

b) Shadow Mask

c) Cloud Mask

Temporal Normalization

Within overlapping image areas it is reasonable to expect similar mean and variance statistics for corresponding bands on similar dates. Differences between adjacent images can be attributed to viewing and processing factors more than actual surface conditions. Thus for purposes of image mosaicing, linear regression can be used to estimate the gain and offset between images and to normalize scenes on an individual band-by-band basis. The intersection between images on two dates, with clouds and cloud shadows masked, was used to sample each scene element (pixel) for the regression. A single image was selected as the reference, and other images were radiometrically rectified to the reference (Eq. 3). Regressions were run for each corresponding pairwise set of bands. We found it beneficial to select the reference image on the basis of image data quality, within the center of the mosaic, and for a date when ground conditions were known (i.e. coincident with field visits). The coefficients of the linear regression were then used to rectify each image to the selected reference.

$$\text{Reference} = \text{Image} * \text{gain} + \text{offset} \qquad \text{(Eq. 3)}$$

Image Data Substitution

For our purposes, clouds and cloud shadows in the imagery represent areas of data loss. During the mosaicing process, however, areas of masked clouds and shadows in one date can be substituted with data from adjacent imagery. Because there is almost 30 percent overlap between adjacent ETM+ scenes along path, it is possible to compensate for image data gaps for approximately 60 percent of each scene. To date, we have employed this process to a limited extent but may expand the activity to compensate for the anomalously high cloud cover in the Summer of 2000.

Image Classification

Classification and regression tree approaches to ecological analysis and land cover mapping have become more widespread in recent years (Michaelson *et al*., 1987; Hansen *et al*., 1996; Friedl & Brodley, 1997; DeFries & Chan, 2000). More generally known as decision tree algorithms (Breiman *et al*., 1984; Venables & Ripley, 1994) they have several advantages over traditional classification techniques. In particular, decision trees are strictly nonparametric and do not require assumptions regarding the statistical properties of the input data. The decision tree classification can handle nonlinear relations between features and classes, allow for missing data values, and are capable of handling both numeric and categorical inputs. Decision trees also have significant intuitive appeal because the classification structure is explicit and easily interpretable. The method is flexible in design and can be tailored to specific data and classification requirements. Derived trees can be pruned, grafted, or forced to split based on user-defined criteria, allowing for improved accuracy in the final product.

We used the machine learning UNIX software package C5.0 (Quinlan, 1993) to generate classification trees from the ETM+ imagery, ancillary data sets, and field measurements. In the hierarchical tree structure, each split in the tree results in two branches. The algorithm searches for the dependent variable that, if used to split the population of pixels into two groups, explains the largest proportion of deviation of the independent variable. At each new split in the tree, the same exercise is conducted and the tree is grown until it reaches terminal nodes, or leaves, each leaf representing a unique set of pixels. Every leaf has a land cover class assignment.

Field Reconnaissance

High quality training data are essential for accurate land cover classification of any sort. The MA-RESAC conducted extensive field reconnaissance throughout the entire CBW in the Summers of 1999 and 2000 (Figure 1). As many training sites as possible were visited in each cover type as defined by a modified Anderson Level II categorization (Anderson *et al*., 1976; USGS, 1992). To insure high spatial and thematic precision of collected data, field sampling was conducted in a GIS framework using laptop computers and real-time differential GPS information and orthorectified, georeferenced, and resolution-enhanced ETM+ imagery. The ETM+ scenes were resolution enhanced by spectral merging with either 10m SPOT panchromatic imagery or the 15m ETM+ panchromatic band. Spectral merging improved visual interpretation, which enhanced our ability to identify sites for data collection. For each of 1118 sites extensive field information including species composition, level and type of management, soil characteristics, topography, forest layers and canopy characteristics were collected. Particular attention was given to agricultural and pasture lands. Digital photographs of each site were taken for documentation.

Post-visit evaluation of the data set was done to assess the quality of collected information. Locations of each site were checked in the ETM+ scenes and field notes, site descriptions, and site photographs were consulted to relate the site location to scene features. Through this process about 6 percent of the sites were omitted from analyses because they were not sufficiently representative of the adjacent landscape (Table 2). A comparable number of sites are being collected to characterize a greater variety of land cover types and to provide for independent validation of the land cover classification.

Table 2. Summary of the field reconnaissance data set.

Cover Type	Retained	Omitted
Cropland	758	20
Pasture	153	1
Fallow land	38	0
Grassland	13	0
Deciduous forest	41	11
Evergreen forest	12	1
Mixed forest	24	5
Wetland	3	1
Other	8	29
Total	1050	68

Early Classification Results

A decision tree LULC classification was developed and tested on Landsat path/row 15/33 centered on the Washington, DC metropolitan area. Landsat-5 scenes on six dates in 1998 (March 27, April 12, April 28, July 1, August 8, November 22) were used together with 30 meter DEM derivatives (elevation, slope, aspect) in the analysis. All Landsat-5 Thematic Mapper (TM) channels, except for the thermal band, were included. The decision tree was trained with the field measurements collected within the 31,000 km^2 area encompassed by the TM scene. The resulting land cover map, comparisons with other map results, accuracy assessments, and multi-temporal sensitivity analyses were reported by Goetz *et al.* (2000b).

To evaluate the contribution of multi-temporal information for LULC classification, three independent runs of decision tree were performed for: (i) a single peak growing season date, (ii) leaf on – leaf off dates, (iii) multi-temporal (all available) dates. When compared to the single date imagery alone, incorporation of the multi-temporal data into the analysis improved discrimination of specific LULC classes, particularly those dominated by vegetation. Differences among deciduous, evergreen, and mixed forest types as well as among croplands, pastures, and grasslands were improved over previous maps of the area. Discrimination of impervious surface areas, primarily in urban and suburban areas, did not significantly benefit from multi-temporal image acquisitions.

SUMMARY AND FUTURE DIRECTIONS

The Mid-Atlantic RESAC has advanced its LULC development activities to the point that the data sets and algorithms are sufficient to proceed with mapping the entire Chesapeake Bay Watershed. Early results suggest that the combined use of a decision tree classifier, extensive field measurements, and carefully rectified multi-temporal ETM+ imagery will permit more accurate LULC discrimination than has been available thus far.

The Chesapeake Bay Watershed LULC map will be just one of many products developed by the MA-RESAC, all of which will be widely available to the applications community, including the Chesapeake Bay Program and a wide variety of state and local government agencies, among others. The role of the MA-RESAC as a facilitator between the research and "end-user" communities ensures that the products (whether the CBW-wide map described here, high resolution maps of impervious surface area (Fisher & Goetz, 2001), or associated geospatial technologies used to interpret the LULC products) have been tailored to the needs of the users while considering the limitations and accuracies inherent in their development. Additional information is available at the MA-RESAC www site (www.geog.umd.edu/resac).

REFERENCES

Anderson, J.R., E. Hardy, J. Roach, and R. Witmer. (1976). A land use and land cover classification system for use with remote sensor data. *U.S. Geological Survey Profession Paper,* 964. Washington, DC. 24 p.

Breiman, L., J. Freidman, R. Olshend, and C. Stone. (1984). Classification and regression trees. Monterey, CA: Wadsworth.

DeFries, R.S., and J. Chan. (2000). Multiple criteria for evaluating machine learning algorithms for land cover classification from satellite data. *Remote Sensing of Environment*, 74: 503-515.

Fisher, J. and S.J. Goetz. (2001). Considerations in the use of high spatial resolution imagery: an applications research assessment. Available at www.geog.umd.edu/resac and on ASPRS CD-ROM *in* American Society for Photogrammetry and Remote Sensing (ASPRS) Conference Proceedings, St. Louis, MO.

Friedl, M.A. and C.E. Brodley. (1997). Decision tree classification of land cover from remotely sensed data. *Remote Sensing of Environment*, 61(3): 399-409.

Goetz, S.J. (1997). Multi-sensor analysis of NDVI, surface temperature and biophysical variables at a mixed grassland site. *International Journal of Remote Sensing*, 18(1): 71-94.

Goetz, S.J., S.P. Prince, M.M. Thawley, A.J. Smith, R. Wright, and M. Weiner. (2000a). Applications of multi-temporal land cover information in the Mid-Atlantic Region: a RESAC initiative. Available at www.geog.umd.edu/resac and on IGARSS CD-ROM *in* International Geoscience and Remote Sensing Symposium (IGARSS) Conference Proceedings, Honolulu, HI.

Goetz, S.J., S.P. Prince, M.M. Thawley, A.J. Smith, and R. Wright. (2000b). The Mid-Atlantic Regional Earth Science Applications Center (RESAC): An Overview. Available at www.geog.umd.edu/resac and on ASPRS CD-ROM *in* American Society for Photogrammetry and Remote Sensing (ASPRS) Conference Proceedings, Washington DC.

Hansen, M., R. Dubayah, and R.S. DeFries. (1996). Classification trees: An alternative to traditional land cover classifiers. *International Journal of Remote Sensing*, 17(5): 1075-1081.

Irish, R.R. (2000). ETM+ Science Data User's Handbook, NASA Document No. 430-15-01-003-0. (http://ltpwww.gsfc.nasa.gov/IAS/handbook/handbook_toc.html)

Michaelson, J., F. Davis, and M. Borchert. (1987). Non-parametric methods for analyzing hierarchical relationships in ecological data. *Coenoses*, 1: 97-106.

Quinlan, R. (1993). C4.5: Programs for Machine Learning. San Meteo, CA: Morgan Kaufmann.

USGS. (1992). Standards for digital line graphs for land use and land cover technical instructions. *Referral STO-1-2*. Washington, DC: US Government Printing Office. 60 p.

Venables, W.N. and B.D. Ripley. (1994). Modern Applied Statistics with S-plus. New York: Springer-Verlag.

Vogelmann, J.E., T. Sohl, and S.M. Howard. (1998). Regional characterization of land cover using multiple sources of data. *Photogrammetric Engineering and Remote Sensing*, 64(1): 45-57.

Acknowledgements

This work was supported by NASA grant NAG1399011, which provided for the establishment of the Mid-Atlantic RESAC. We thank Maria Eugenia Mazzocato, Erika Schaub, Carolina Santos, Michelle Thawley, and Jordan Borak for assistance with image processing and algorithm development.