MAPPING IMPERVIOUS SURFACES AND FOREST CANOPY USING CLASSIFICATION AND REGRESSION TREE (CART) ANALYSIS

Nathaniel D. Herold, Applications Scientist
Gregory Koeln, Senior Vice President and Director of Envirnomental/GIS Services
David Cunnigham, Vice President of Environmental Services
Earth Satellite Corporation
Rockville, MD USA
nherold@earthsat.com
gkoeln@earthsat.com
dcunningham@earthsat.com

ABSTRACT

In recent years, research and applications utilizing classification and regression tree (CART) technology has increased and been successfully applied in ecology (De'ath and Fabricious, 2000; Prince and Steininger, 1999) and image processing or analysis (Friedl et al., 1999; Huang, et al., 1997; Lawrence, et al., 2001; Lloyd, 1990; Wylie et al., 2000; Yang, et al., 2002). EarthSat has developed software and a methodology that utilizes CART technology in order to map sub-pixel impervious surface and forest canopy surfaces at a 30-meter resolution. The complex interactions that exist between various input data sets, as they relate to the target impervious and canopy features, are learned and modeled through exhaustive examination. This allows the production of a "knowledge based" rule set classification through an inductive process of machine learning. Building of such a knowledge base typically requires a human expert to express his or her knowledge in a language easily understood, requiring a large amount time and understanding, which is often hampered by the lack of requisite knowledge or by difficulties in developing generalized rules based on theoretical knowledge (Huang et. al., 1997). CART technology, or machine learning, can provide a low-cost, high quality alternative, without such difficulties (Maniezzo, et. al., 1993). Results for estimating percent impervious surface and canopy cover, per ETM+, pixel were found to be very effective. The average error in percent estimation was within 8.4 percent and had a correlation coefficient of .90 to .93.

INTRODUCTION

A basic problem in economic planning, environmental studies and resource management is the availability of current, accurate information. Information characterizing urban landscapes, such as impervious surface and canopy cover, are often critical to urban planners and policy makers. Analysis of such location based problems without geographically coded data is analogous to analyzing a time series without knowing the chronological order of observations (Liverman, et al., 1998). Often, conventional survey and mapping methodologies are unable to deliver the necessary information in a timely and cost effective mode. Given their technological robustness, remote sensing technologies are increasingly affecting urban land-use change research (Geohegan, 1998). Many studies concerning the classification of medium-resolution satellite imagery, such as Landsat TM or SPOT data, in urban or urban-rural areas demonstrated that treed areas, except large wooded parks and suburban areas, could not be accurately extracted because of the lack of adequate spatial resolution (Zhang, 2001). This study reviews a process of sub-pixel percent impervious and percent canopy estimation, at this moderate resolution, utilizing multi-sensor and multi-source data and CART technology.

In recent years research has increased in the use of classification and regression tree (CART) technology, sometimes referred to as machine learning or expert systems. This technology has been successfully applied in ecology (De'ath, et al., 2000, Prince, et al., 1999) and image processing and analysis (Friedl, et al., 1999, Huang, et al., 1997, Lawrence, et al., 2001, Lloyd, 1990, Wylie, et al., 2000, Yang, et al., 2000). Earthsat has developed software and a methodology that utilizes CART technology in order to provide an accurate, cost-effective method of mapping sub-pixel impervious surface and canopy cover. The process utilizes medium resolution ETM+ imagery, samples of high-resolution images (ex. IKONOS or DOQQ's) within the ETM+ study area and ancillary data sources, such as soils information and/or digital elevation data. These information sources are sampled and examined by the software in an attempt to learn and model the complex relationships that exist between these various data layers and known values of a specified target feature. In the case of this study, we examined an area surrounding the eastern portion of the Chesapeake Bay watershed in order to map sub-pixel estimates of percent impervious or percent canopy cover surfaces.

The modeling and prediction within the CART analysis is accomplished through a recursive binary partitioning of "training" data, sampled from the various imagery sources, so that values are representative of the entire dataset. These samples are then used in the production of rule sets, based on the relationships modeled, essentially enabling the software to build a "knowledge base" with little interaction or required input from the user.

Building of a knowledge base typically requires a human expert to express his or her knowledge in a language easily understood by the designed classification software. This requires a large amount of user input, which is often difficult to understand or express. Attempts to develop such expert knowledge bases are often hampered by the lack of requisite knowledge or difficulties in developing rules based on general theoretical knowledge (Huang, et al., 1997). CART technology, or machine learning, has the potential to provide a low-cost, high quality knowledge base without such difficulties (Maniezzo, et al., 1993). Based upon the training samples, CART software is able to approximate the human learning process and make accurate generalizations concerning the relationships of input variables and the value of the target feature.

CART analysis allows for a large number of potential inputs to be analyzed, as not all input data sources must be used in characterizing each resulting relationship, or rule definition. Only inputs that are determined to provide meaningful information concerning a certain relationship are used in the application, or the resulting rule based on that defined relationship. This is because CART analyzes all input, explanatory or predictive, variables to determine the binary division of a single variable which best reduces the deviation in response, and therefore the two most homogenous stems. This process is repeated until homogeneous divisions, or terminal nodes, are found (Breiman, et al., 1984). CART analysis also allows for a wide range of possible input data types, because regression trees are based on nonparametric statistics. Categorical variables can, therefore, be incorporated and statistical assumptions about the distribution of the data are not required (De'ath, et al., 2000, Friedl, et al., 1999). As such, non-normally distributed data can readily be utilized in analysis without the difficulties such data normally provide to image processing and analysis.

While this allows a large array of potentially useful data to be entered into, or considered, for analysis, it is not to say that the inclusion of every possible data is useful. It would be better to select the potentially most useful inputs, while eliminating all other, unnecessary data in order to improve efficiency and reduce the occurrence of random noise, thereby increasing the accuracy and robustness of prediction. This selection of possible predictor variables is often a difficult process and one in which expertise, or expert knowledge is needed. For the purposes of this study, Spring, Summer and Fall dates of ETM+ imagery, ETM+ thermal information, various STATSGO soil layers, DEM and DEM derived data layers, as well as high resolution DOQQ images were available.

Each resulting production rule defines the conditions under which certain multivariate linear sub-models can appropriately be applied in predicting the target impervious or canopy value (Quinlan, 2002). The conditions that must be met represent the homogenous binary divisions, or individual stems, while the regression equation represents the relationship defined between the variables. Unlike traditional regression tree classifications, these linear models are not mutually exclusive, allowing overlap between several sub-models. Such piecewise-linear models can account for non-linear multiple relationships between predictive and target DN values with greater accuracy and predictability than that of simple linear models, including logistic regression. An example of one such sub-model, taken from a full rule set generated by Cubist, can be seen in Table 1.

IF (condition)	THEN (linear model)

`band03 > 42`
`band06 <= 191`
`band08 <= 110`
`band12 > 84`
	Dep = `-116.8 + 1.14 band06 + 0.86 band09`
	`+ 0.77 band07 - 0.54 band05`
	`- 0.53 band12 - 0.36 band08`
	`- 0.38 band04 - 0.46 band10`
	`+ 0.06 band01 + 0.05 band02`

Table 1. Example of linear sub-model rule developed through Cubist's CART analysis.

When the conditions within the IF column, on the left side of the Table, are met the linear model, on the right side of the table, is applied in order to predict the value of the desired dependent variable (Dep). As these sub-models are not mutually exclusive one pixel may meet the conditions of several individual rules. The multiple rules that may apply to any such pixel are dealt with in a way that the values resulting from each applicable linear equation may be averaged in order to produce a value that represents a combination of each rule that was applicable.

This allows a fuzzy classification decision to be made concerning these cases, providing a more robust classification in areas that would normally fall into more than one class, resulting in spectral confusion.

The final rule sets are then applied to the identified areas of cloud cover within each image layer of the target dataset. This allows the cloud free areas of the alternate date(s) of imagery to be used in predicting DN values that will replace the cloud obscured values. This preserves the useful values of the original dataset (unaffected by cloud coverage), while providing new DN values. Such a methodology is preferable to simple image substitution or matching, as it does not alter the original, cloud free areas of the target date, nor does it combine data acquired under diverse conditions, or achieved through separate analysis. This should provide the user with a data set and interpretation product of increased consistency and utility, as the resultant spectral information will be contained within one image and will be characteristic of the expected DN values of that image acquisition time frame.

METHODOLOGY AND PROCEDURES

Enhanced Thematic Mapper Plus (ETM+) images were acquired for three different seasons, for a study area centered at Washington, D.C. and surrounding the Chesepeake Bay (Figure 1), in order to capture vegetation dynamics over a growing season and to maximize land cover type separability (Yang, et al., 2001). These images were acquired within the time period between 1999 and 2001, and were selected to minimize the impact of cloud cover and atmospheric effects. The images were geometrically and radiometrically corrected using standard methods at the USGS EROS Data Center (Irish, 2000). Terrain correction using the USGS 1-arc second National Elevation Dataset was performed to improve geolocation accuracy. Raw digital numbers were converted to at-satellite reflectance for the 6 reflective bands, and to at-satellite temperature for the thermal band according to Markham and Barker (Markham, et al., 1986) and the Landsat 7 Science Data User's Handbook (Irish, 2000). All 7 bands were resampled to a 30 meter spatial resolution. DEM and DEM derived layers characterizing slope, aspect and topographic position were also available for this study area.

The methodology employed can be divided into three phases, which include: 1) the selection and sampling of the input predictive, or independent variables, for areas of cloud free target date imagery, 2) the building of the knowledge base and production rule set, and 3) the application of this production rule set in order to predict the outcome values for the areas covered by clouds within the target date.

The selection and sampling of the appropriate predictive, or training, variables are of key importance in obtaining an accurate prediction of target percent surfaces. For training purposes several high-resolution samples of digital ortho quarter quads (DOQQs) were selected and classified for target features, at one meter, based on spatial and spectral characteristics of the ETM+ imagery. This was accomplished with the Statistical Sample Selection tool within ERDAS Imagine's Frame Sampling tool set. This method identified areas that when sampled would capture the maximum amount of variability within the ETM+ data. By obtaining high-resolution imagery for these identified areas, we can increase the probability that training can accurately characterize the data set as a whole.

After these high resolution classifications were complete, they were then compared to the the corresponding ETM+ pixels in order to determine known percentages of impervious or canopy surfaces. These areas of known percentage are then used within the CART software in order to determine classification rules which characterize the relationships that exist between all the various data layers and these percent surface values. This is done utilizing software developed with the C Programmers Toolkit and ERDAS Macro Language (EML). Training data was sampled with a stratified random sampling methodology, while validation data was sampled in a purely random fashion. This sampling procedure was performed for each individual percent surface target dataset after areas of cloud, cloud shadow and haze were masked out and replaced with useful spectral information. Each of the target bands had 50,000 training pixels and 50,000 validation pixels sampled.

The training and validation samples were then input into the CART software. Rulequest Research's Cubist 1.11 software package was used as part of this study. This software package analyzed the relationships within the data and formulated an appropriate regression tree and production rule set. The output predicted values were evaluated against the training data itself, as well as the 50,000 separately sampled validation pixels. The final regression tree model, or rule set, was then applied directly to the full image data set in order to predict DN values for each of the target date bands in areas covered by clouds. This was implemented with a second GUI interface, within ERDAS Imagine, in order to achieve a full output dataset free of clouds and possessing spectral values characteristic of that date of imagery.

Earthsat developed this procedure as part of a USGS contract (USGS contract number 010112C0012) and has found results for generating and replacing cloud covered areas within Landsat TM data to be very effective.

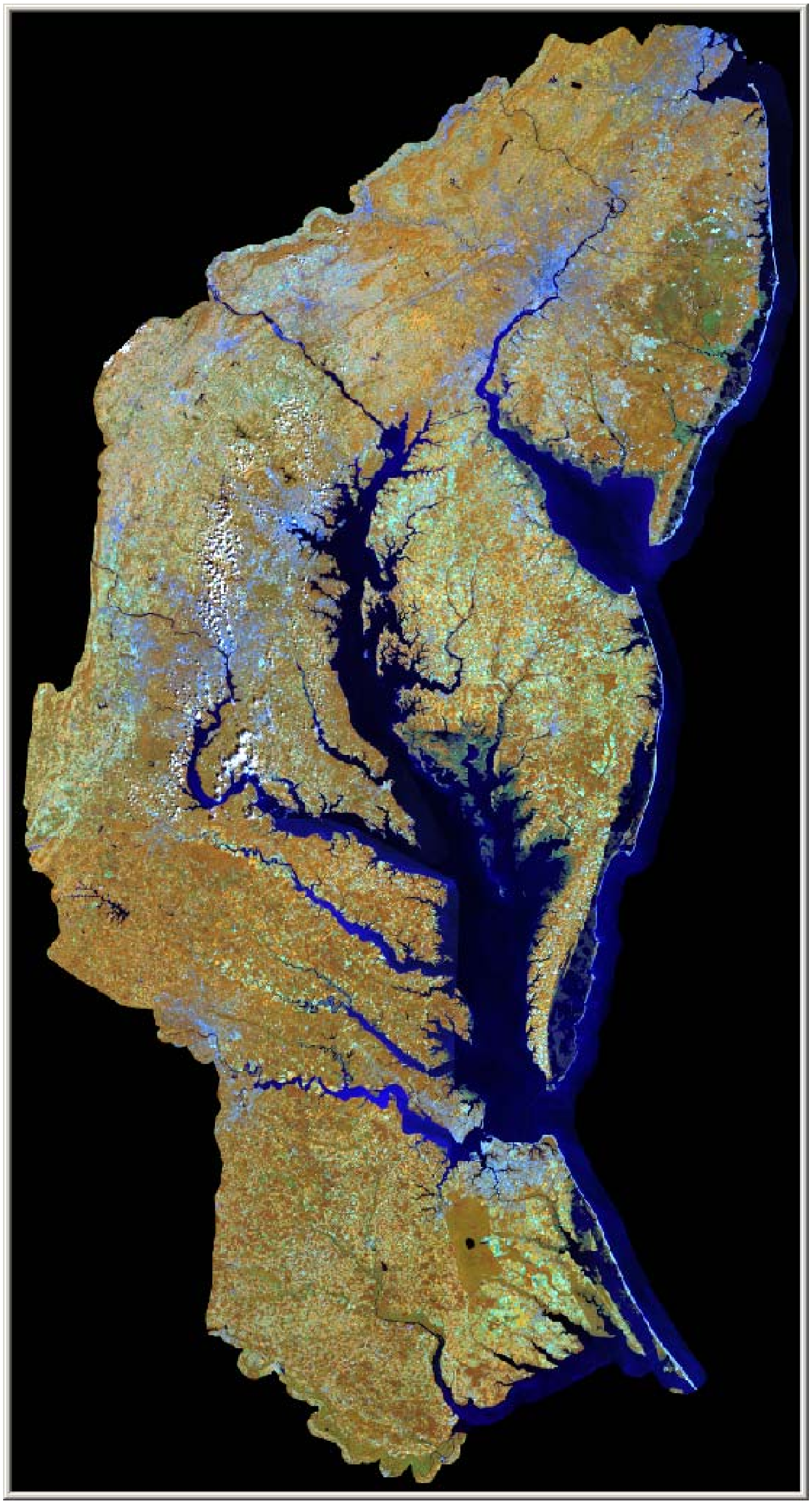

Figure 1. Fall ETM+ Imagery for NLCD Mapping Zone 60. Chesapeake Bay Area. Bands 453.

RESULTS

As can be seen in the following figures, the achieved results were visually appealling. Figure 2 is an example of the impervious. surface product for the Washington, D.C. area. The green space of the National Mall (downtown), can clearly be scene and is characterized with very low impervious percentage values (white), while obvious urban features, having high impervious values (red), are clearly visible

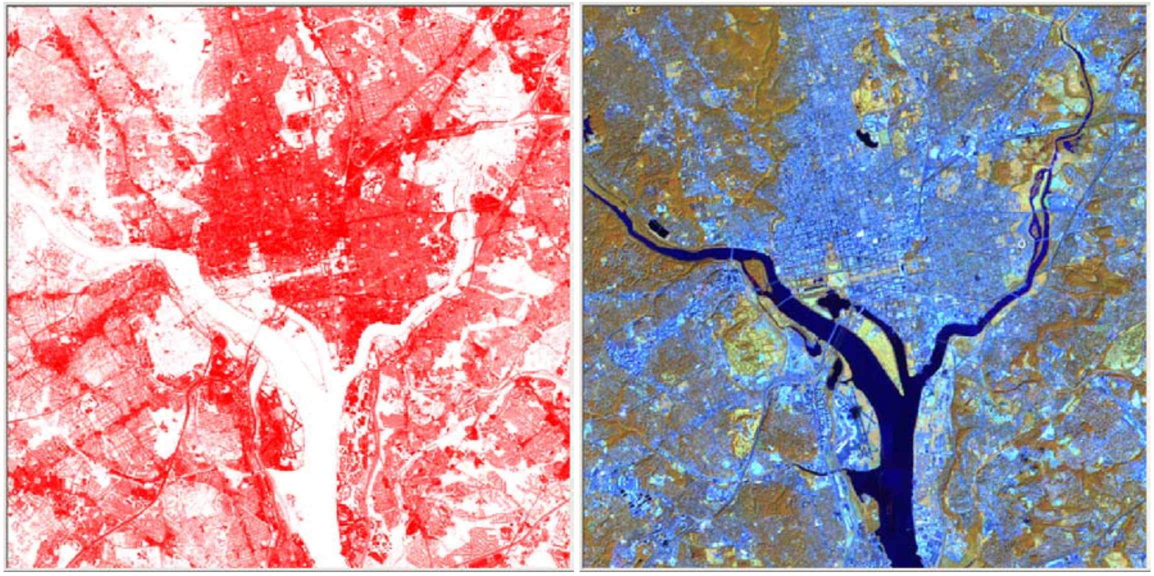

Figure 2. Impervious surface percentage (left) and corresponding ETM+ image for Washignton, D.C. area.

In general, discrimination of percent impervious values improved when all three dates of spectral information were utilized. This is reasonable as two dates of imagery would provide less information concerning the seasonal variability in spectral values for different land cover features and would therefore provide less discriminatory ability. The use of Tasseled Cap imagery over that of the original reflectance data also tended to be of greater value. Results achieved with these transformed images allowed the information from each date to be included in analysis, with fewer individual bands for Cubist to analyze, and performed with results approximating, if not surpassing, those achieved with the reflectance information. The addition of DEM values, DEM derived slope data, SATSGO soil quality data and one date of ETM+ thermal information were found to greatly increase the accuracy of the resulting classification, while the addition of the DEM derived aspect information was found to be somewhat less useful.

Cubist evaluates accuracy in several different ways. It generates measures of average error, as well as a Product-Moment correlation coefficient. Where the average error represents the average difference between the predicted and actual DN value and the correlation coefficient is a measure of fit between the actual and predicted values (Quinlan, 2002). In this study, we utilized 50,000 validation pixels sampled in a random manner, independent from those used in training.

The final percent impervious layer was produced using the Spring, Summer and Fall dates of ETM+ Tasseled Cap imagery, the Summer date of thermal imagery, the STATSGO soil quality layer, DEM and DEM derived slope values. This combination produced the lowest average error at 5.7 percent and had a correlation coefficient of 0.90.

Figure 3 is an example of the percent canopy estimation for the area of Beltsville, MD. Agricultural fields and areas of high intensity development can clearly be seen and are represented within the canopy layer with very low percentages of canopy (white), while established residential areas and areas of thicker forest growth are represented with increasing values for the canopy percentage (light to dark green).

The discrimination of percent canopy values also improved when all three dates of Tasseled Cap imagery and one date of thermal data were utilized. The addition of DEM values and DEM derived slope data improved the accuracy of the resulting classification, just as it did with the impervious surface study, but SATSGO soil quality information was found to be of little to no use for canopy identification.

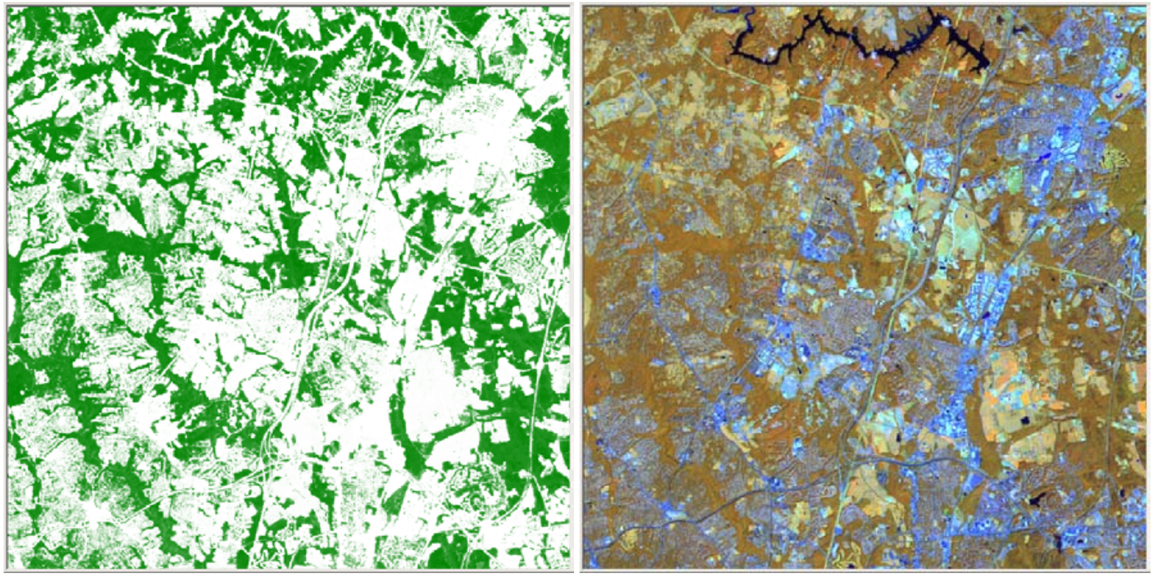

Figure 3. Percent canopy surfaceand corresponding ETM+ image for area of Beltsville, MD.

The final percent canopy layer was, therefore, produced using the Spring, Summer and Fall dates of ETM+ Tasseled Cap imagery, the summer date of thermal imagery, and the DEM and DEM derived slope layers. This average error for this classification was 8.4 percent and had a correlation coefficient of 0.93. Results for both products achieved reasonably accurate estimates of defined target, impervious and canopy, features by percent surface area without the difficult and costly effort of mapping the entire study area.

ACKNOWLEDGEMENTS

This work was performed in part under the Raytheon Corporation under U.S. Geological Survey Contract 1434-CR-97CN-40274.

REFERENCES

Breiman, L., Friedman, J., Olshen, R., and C. Stone. (1984). Classification and Regression Trees. Wadsworth International Group. Belmont, California.

De'ath, G. and K.E. Fabricius. (2000). Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*. 81: 3178–3198.

Friedl, M. A., Brodley, C. E., and A.H. Strahler. (1999). Maximizing land cover classification accuracies produced by decision trees at continental to global scales. *IEEE Trans. Geosci. and Remote Sens*. 37; 969 – 977.

Gerstl, S.A. (1990). Physics Concepts of Optical and Radar Reflectance Signatures. *International Journal of Remote Sensing*. 9(4):655-668.

Huang, X. and J. Jensen. (1997). A Machine-Learning Approach to Automated Knoweldge-Base Building for Remote Sensing Image Analysis with GIS Data. *Photogrammetric Engineering and Remote Sensing*. 63(10):1185-1194.

Irish, R. Landsat 7 science data user's handbook, Report 430-15-01-003-0, National Aeronautics and Space Administration, http://ltpwww.gsfc.nasa.gov/IAS/handbook/handbook_toc.html, 2000.

Lawrence, L. and A. Wright. (2001). Rule-Based Classification Systems Using Classification and Regression Tree (CART) Analysis. *Photogrammetric Engineering and Remote Sensing*. 67(10):1137-1142.

Liverman, D., Moran, E., Rindfuss, R. and P. Stern. (1998). People and Pixels: Linking Remote Sensing and Social Science. National Academy Press, Wahington, D.C.

Lloyd, D. (1990). A phonological classification of terrestrial vegetation cover using shortwave vegetation. *International Journal of Remote Sensing*. 11:2269–2279.

Maniezzo, V., Morpurgo, R. and S. Mussi. (1993). D-KAT: A Deep Knowledge Acquisition Tool. *Expert Systems*. 10(3):157-166.

Markham, B. and J. L. Barker. (1986). Landsat MSS and TM post-calibration dynamic ranges, exoatmospheric reflectances and at-satellite temperatures. *EOSAT Landsat Technical Notes.* Vol. 1, pp. 3-8, 1986.

Prince, S.D. and M.K. Steininger. (1999). Biophysical stratification of the Amazon basin. *Global Change Biology*. 5:1 – 22.

Quilan, R.. (2002). An Overview of Cubist. Retrieved July, 2002 from http://www.rulequest.com/cubist-win.html.

Verbyla, D.L. (1995). Satellite Remotes Sensing of Natural Resources. Lewis Publishers Inc., Boca Raton, Fl.

Wylie, B.K., D.J. Meyer, M.J. Choate, L. Vierling, and P.K. Kozak. 2000. Mapping Woody Vegetation and Eastern Red Cedar in the Nebraska Sand Hills Using AVIRIS. (http://popo.jpl.nasa.gov/html/aviris.biblios.html).

Yang, L., Homer, C., Hegge, K., Huang, C., and B. Wylie, "A Landsat 7 Scene Selection Strategy for a National Land Cover Database." In IEEE International Geoscience and Remote Sensing Symposium, Sydney, Australia, pp CD ROM, 1 disk, 2001.

Yang, L., Huang, C., Homer, C., Wylie, B. and M. Coan. (2002). An Approach for Mapping Large-Area Impervious Surfaces: Synergistic Use of Landsat 7 ETM+ and High Resolution Imagery. *Canadian Journal of Remote Sensing*. Submitted for review.