Rule-Based Classification Systems Using Classification and Regression Tree (CART) Analysis

Rick L. Lawrence and Andrea Wright

Abstract

Incorporating ancillary data into image classification can increase classification accuracy and precision. Rule-based classification systems using expert systems or machine learning are a particularly useful means of incorporating ancillary data, but have been difficult to implement. We developed a means for creating a rule-based classification using classification and regression tree analysis (CART), a commonly available statistical method. The CART classification does not require expert knowledge, automatically selects useful spectral and ancillary data from data supplied by the analyst, and can be used with continuous and categorical ancillary data. We demonstrated the use of the CART classification at three increasingly detailed classification levels for a portion of the Greater Yellowstone Ecosystem. Overall accuracies ranged from 96 percent at level 1, to 79 percent at level 2, and 65 percent at level 3.

Introduction

Ancillary data, either in addition to or derived from remotely sensed data, has the potential for increasing classification accuracy. Incorporation of ancillary data into classification techniques, however, has been problematic. We developed a straightforward approach for creating a rule-based classification without expert knowledge by applying a commonly available statistical technique, classification and regression tree (CART) analysis, to multiple spectral and ancillary data layers.

Classification and Regression Tree Analysis—Background

CART is an increasingly popular form of statistical analysis available through widely used statistical packages, such as S-Plus (Venables and Ripley, 1997; MathSoft, 1998; Lawrence and Ripple, 2000). CART operates by recursively splitting the data until ending points, or terminal nodes, are achieved using preset criteria. CART therefore begins by analyzing all explanatory variables and determining which binary division of a single explanatory variable best reduces deviance in the response variable (Breiman *et al.*, 1984; Efron and Tibshirani, 1991; Venables and Ripley, 1997). In the case of image classification, explanatory variables consist of spectral and ancillary data, whether continuous or categorical, and the response variable is the land-cover/land-use class list.

For each portion of the data resulting from this first split, the process is repeated, continuing until homogeneous terminal nodes are reached in a hierarchical tree. In the S-Plus implementation of CART, terminal nodes are defined when either the total number of observations at the node is less than ten or the deviance at the node is less than 1 percent of the total deviance for the entire tree (Venables and Ripley, 1997).

CART usually will over-fit the model, creating a tree that explains substantially all of the deviance in the original data, but in a manner that is specific to the particular data used to fit the tree. It is necessary, therefore, to prune the tree back to a level where the tree can reasonably be expected to be robust. A common method used for pruning, and a method implemented in S-Plus, involves cross validation (Venables and Ripley, 1997). In this method, the original data are randomly divided into ten equal sets. Trees are generated for nine of the data sets and validated against the tenth, with the minimum average deviance indicating the best size tree. The analyst might select a smaller tree if the cross-validation method indicates that, although additional deviance can be reduced, the amount of reduction does not justify an overly complex tree.

The result of the CART analysis is a dichotomous decision or classification tree. Each path through the tree, defined by a series of dichotomous splits, specifies the conditions that lead to a most probable class. The tree, therefore, might be viewed as a series of rules that can be used for unknown observations to predict likely class membership. When used with remotely sensed and ancillary data, this naturally extends to a rule-based classification scheme.

Ancillary Data Incorporation in Classification—Background

Traditional methods of land-use/land-cover classification using satellite imagery have relied solely on the spectral information present in the images. With purely spectral approaches, the spectral and spatial resolutions of the imagery are the primary determinants of the level of classification detail that can be achieved. For example, given the spectral and spatial resolution of Landsat Thematic Mapper (TM) imagery, such images have generally been considered adequate for mapping USGS level II (Jensen and Cowen, 1999). By using ancillary data in addition to spectral responses, however, it might be possible to achieve either greater classification detail or greater classification accuracy for a given combination of spectral and spatial resolutions.

Classification techniques using ancillary data in addition to spectral data have demonstrated that, in many cases, the proper addition of ancillary data to spectral data can lead to greater class distinctions (e.g., Strahler *et al.*, 1978; Hutchenson, 1982; Trotter, 1991; Jensen, 1996). Ancillary data generally is derived from GIS layers, such as digital elevation models, but might also include information derived from the imagery, such as texture information or multi-date composites. Initially,

R.L. Lawrence is with the Mountain Research Center, Department of Land Resources and Environmental Sciences, P.O. Box 173490, Montana State University, Bozeman, MT 59717-3490 (rickl@montana.edu).

A. Wright is with the Center for the Environment, Cornell University, Ithaca, NY 14853 (awp9@cornell.edu).

Photogrammetric Engineering & Remote Sensing
Vol. 67, No. 10, October 2001, pp. 1137–1142.

0099-1112/01/6710–1137$3.00/0

methods incorporating ancillary data included pre-classification stratification, logical channel addition, and post-classification sorting (Jensen, 1996) and have met with considerable success. Each of these approaches builds on conventional supervised or unsupervised classification methods.

Using ancillary data for pre-classification stratification or post-classification sorting does not incorporate additional data into the actual classification algorithm, but might increase accuracy by segregating otherwise confused classes using such information as elevation or landscape position (Vogelmann *et al.*, 1998; Ricchetti, 2000). While these methods have been successful in increasing classification accuracies, failure to incorporate ancillary data into the classification algorithm might fail to fully exploit the range of information available. When ancillary data have been incorporated into traditional classification algorithms as logical channels (combining the ancillary data as an additional data layer with the spectral bands), the full range of information available in the ancillary data was used (e.g., Strahler *et al.*, 1978; Elumnoh and Shrestha, 2000; Ricchetti, 2000). With logical channel addition, ancillary data are given equal weight to single spectral bands unless weights are assigned in a maximum-likelihood classifier. This implicit weighting of ancillary data might not be appropriate or optimal, but proper weighting is usually unknown or must be sought out by trial and error. In addition, logical channel addition is not appropriate for categorical, as opposed to continuous, ancillary data.

More recently, ancillary data have been incorporated into modern classification methods such as expert systems (Goodenough *et al.*, 1987) and neural networks (Bruzzone *et al.*, 1997; Skidmore *et al.*, 1997). These approaches incorporate the ancillary data directly into the classification algorithms and are usually not dependent on *a priori* weights. Attempts to develop expert systems, however, are often hampered by the lack of requisite expert knowledge or difficulties in developing rules from such knowledge (Kontoes *et al.*, 1993; Huang and Jensen, 1997). For example, it might be known that certain tree species only exist above certain elevations, but spectral distinctions among tree species within an image generally are not known prior to the analysis. Neural network approaches are still largely in the developmental stage, are not easily implemented (Bruzzone *et al.*, 1997), and have had unpredictable results (Skidmore *et al.*, 1997).

Machine-learning approaches have been used to establish rule-based classification systems where expert knowledge was inadequate (Huang and Jensen, 1997). These methods use training data and machine-learning algorithms to develop a series of rules to define each class. Although machine-learning approaches overcome many of the limitations of other approaches to incorporating ancillary data, they have required sophisticated programming skills and have not been readily available to the wider remote sensing community. Our approach using CART has similarities to machine learning approaches, but the tools we used are readily available and easily implemented with commercially available software.

Study Area

The study area consisted of the northwest portion of the Greater Yellowstone Ecosystem (GYE). Covering approximately 20,900 km^2, the portion of the GYE included in this study spans two TM scenes, path 39, rows 28 and 29, and includes parts of Idaho, Montana, and Wyoming in the Rocky Mountain physiographic province. Most of the private land within the study area is found in broad valleys drained by active rivers (including the Gallatin, Madison, and Jefferson) and is surrounded by publicly owned, forested mountain ranges. The southeastern part of the study area contains the western-most boundary of Yellowstone National Park and the town of West Yellowstone, Montana. Remnants of the 1988 Yellowstone fires and large-scale clearcut logging are dominant disturbances in the southeast. The northern and western portions of the study area lie mainly within the Gallatin and Targhee National Forests and include the Gallatin Valley and the town of Bozeman, Montana, which is the largest urban center in the GYE. Vegetation in the study area consists of mixed species conifer forests, generally above 2000 m, intermixed with natural shrublands and grasslands. Less than 2 percent of the study area consists of exposed rock outcrops above treeline. At lower elevations, a mix of sagebrush, grasslands, and agricultural lands dominates the landscape. Hardwoods are mixed throughout the landscape, generally in linear patches following rivers (cottonwood and willow) and at the lower interfaces of forest and grassland (aspen). Water was excluded from the study area by thresholding TM band 5, and the limited areas of urban development were delineated by manual digitizing.

Methods

Two pairs of TM scenes were acquired for the study area from 28 June and 15 August 1994. These dates were chosen to encompass the growing season in the region and were the most cloud-free scenes available. The scenes were georeferenced, and paired scenes for each date were mosaicked. Radiometric correction between scenes was performed through dark-body subtraction to minimize artificial brightness and haze (Chavez, 1998). In addition to using TM bands 1-7 from each date, we performed a Tasseled Cap transformation (Crist and Cicone, 1984) for each date and used the first three components from the transformation (brightness, greenness, and wetness) for the analysis. Finally, for each of the three Tasseled Cap bands, we computed difference images between the two dates.

In addition to the data from the TM scenes, we used a 30-m USGS digital elevation model (DEM) to extract elevation, slope, and aspect for the study area. A cosine transformation was applied to the aspect layer to provide a continuous variable measuring angle from due north (Beers *et al.*, 1966), and the results were rescaled from 0 to 2000.

Reference data were collected through aerial photo interpretation using either 1:15,840- or 1:24,000-scale photos from the 1990s, as available. Within the study area, seven transects were located covering the known range of cover types (from U.S. Forest Service stand maps) and elevation (from the DEM). For each cover type, 30 to 100 sample sites were collected using a stratified random sampling design, with stratification by elevation, aspect, and cover type. A total of 500 reference sites were interpreted using this method and were evenly divided into training and accuracy assessment (verification) sites. An additional 129 training sites were interpreted to increase the sample sizes for underrepresented classes, resulting in 379 sites for classification training. In order to positively locate and interpret sample sites on the aerial photos, sample sites were designated as 2.25-ha areas, which was the smallest mapping unit practically sampled on the photos. Reference data were recorded at three levels of classification (Table 1).

Spectral and topographic data were extracted for each reference site. The ERDAS Imagine "Convert pixels to ASCII" utility was used to extract the data for each reference site from each data layer to a text file. Data extracted for each reference site were the mean value for a 5- by 5-pixel area to correspond to the size of the reference sites.

The ASCII text file, which included three levels of land-cover/land-use classification and 26 possible explanatory variables for each training site, was imported as an S-Plus data frame. CART analysis was performed to develop classification rules for the level 1 classification. Data classified as natural vegetation was subset and CART analysis was performed on this subset to develop classification rules for the level 2 classification. For the level 3 classification, four CART analyses were conducted, with the hardwood, herbaceous, agriculture, and

Table 1. Land-Use/Land-Cover Classification Scheme Used for the Study

Level 1	Level 2	Level 3
Urban	Urban	Urban
Agriculture	Agriculture	Perennial agriculture
		Annual agriculture
Natural vegetation	Conifer	Mixed conifer
		Douglas-fir
	Conifer/herbaceous mix	Conifer/herbaceous mix
	Burned	Burned
	Hardwood	Aspen
		Willow
		Cottonwood
	Hardwood/ herbaceous mix	Hardwood/ herbaceous mix
	Herbaceous	Sage-grassland
		Grassland

conifer classes from level 2 each being further classified with separate rule sets.

Using the results of the CART analyses as the basis for classification rules, the ERDAS Imagine Expert Classifier was used to classify the entire study area. At levels 1 and 3, the manually digitized urban areas were merged with the classified CART map for final presentation and accuracy assessment. At level 2, both the urban areas and the classified agricultural areas from level 1 were merged into the final maps. Verification sites were used to test the maps for accuracy, and error matrices were generated.

Results

CART analysis created 14 rules for classifying pixels at classification level 1, with each rule designated by a path through a dichotomous classification tree to a terminal node (Table 2). Data layers used in the classification rules included difference images for Tasseled Cap brightness, greenness, and wetness components; TM bands 1, 2, 4, 5, and 6 from the June TM image; TM bands 1 and 5 from the August image; and elevation. Aspect, slope, and the remaining spectral data were not used in the classification. Examining the CART results showed that a key factor used for distinguishing agriculture from natural vegetation was elevation. In areas of high brightness difference between June and August, no agriculture was classified above 1,754 m. Below 1,754 m, agriculture was distinguished from natural vegetation primarily by natural vegetation having intermediate values in band 5 (middle infrared) of the June TM image. In areas of lower brightness difference between June and August, no agriculture was classified above 1,682 m. Below 1,682 m, natural vegetation was distinguished from agriculture by natural vegetation having lower band 6 (thermal) values for the June TM image. These results correlate with the general patterns expected within the GYE: (1) forests tend to dominate higher elevations while agriculture is located almost exclusively in the lowlands; (2) forested canopies tend to have higher leaf water content than dry-land agriculture but lower than irrigated agriculture, resulting in intermediate values in the middle infrared; and (3) forested areas have greater canopy cooling and thus lower thermal values than agriculture.

At level 1, overall accuracy was 96 percent and the Kappa statistic was 0.92. For individual classes, accuracies ranged from 79 percent to 100 percent (Table 3). The primary source of confusion at level 1 was between the urban and agriculture classes, where several urban sites were misclassified as agriculture.

The level 2 classification CART analysis, which was conducted only on the natural vegetation areas from the level 1 classification, created 17 rules to classify natural vegetation

Table 2. Dichotomous Classification Tree for Level 1 Classification Based on CART Analysis

1. Tasseled Cap brightness difference < 22.5
1.1. Elevation < 1682 m
1.1.1. June TM band 6 < 139.5
1.1.1.1. June TM band 4 < 111, THEN Natural vegetation
1.1.1.2. June TM band 4 > 111, THEN Natural vegetation
1.1.2. June TM band 6 > 139.5
1.1.2.1. June TM band 2 < 32.5, THEN Agriculture
1.1.2.2. June TM band 2 > 32.5, THEN Agriculture
1.2. Elevation > 1682 m
1.2.1. June TM band 6 < 116.5, THEN Natural vegetation
1.2.2. June TM band 6 > 116.5, THEN Natural vegetation
2. Tasseled Cap brightness difference > 22.5
2.1. Elevation < 1754 m
2.1.1. Tasseled Cap greenness difference < 1
2.1.1.1. June TM band 4 < 116.5
2.1.1.1.1. Tasseled Cap wetness difference < −5.5
2.1.1.1.1.1. August TM band 5 < 138.5
2.1.1.1.1.1.1. June TM band 5 < 113.5
2.1.1.1.1.1.1.1. June TM band 5 < 93
2.1.1.1.1.1.1.1.1. June TM band 1 < 66.5, THEN Agriculture
2.1.1.1.1.1.1.1.2. June TM band 1 > 66.5, THEN Natural vegetation
2.1.1.1.1.1.1.2. June TM band 5 > 93, THEN Agriculture
2.1.1.1.1.1.2. June TM band 5 > 113.5, THEN Natural vegetation
2.1.1.1.1.2. August TM band 5 > 138.5, THEN Agriculture
2.1.1.1.2. Tasseled Cap wetness difference > −5.5, THEN Agriculture
2.1.1.2. June TM band 4 > 116.5, THEN Agriculture
2.1.2. Tasseled Cap greenness difference > 1
2.1.2.1. August TM band 1 < 77.5, THEN Agriculture
2.1.2.2. August TM band 1 > 77.5, THEN Non-vegetation
2.2. Elevation > 1754 m
2.2.1. June TM band 5 < 100.5, THEN Natural vegetation
2.2.2. June TM band 5 > 100.5, THEN Natural vegetation

into six sub-classes (Table 4). Layers used to make class distinctions at level 2 included difference images for Tasseled Cap brightness, greenness, and wetness components (as in level 1), as well as June TM bands 2, 3, and 4; Tasseled Cap brightness and greenness from the June TM image; and Tasseled Cap wetness from the August TM image. The rules for level 2 were harder to interpret than were the rules for level 1, primarily due to the more complex nature of the classification. Two patterns were evident, however. First, burned areas were distinguished from other classes by low values in band 4 (near infrared) of the June TM image. Low vegetation cover and dark burn residue in the burned areas resulted in low near-infrared reflectance and explained this distinction. Second, conifer and hardwood species were distinguished primarily through the use of difference images as a result of different phenological patterns throughout the growing season.

Overall accuracy for level 2 was 79 percent and the Kappa statistic was 0.74. Individual class accuracies ranged from 100 percent to 25 percent (Table 5). The greatest confusion at level 2

Table 3. Accuracy Assessment for Level 1 Classification. Columns Represent Reference Data and Rows Represent Classification Data

	Natural vegetation	Agriculture	Urban
Natural vegetation	**171**	1	1
Agriculture	3	**51**	4
Urban	0	0	19
Producer's accuracy	98%	98%	79%
User's accuracy	99%	88%	100%

Table 4. Dichotomous Classification Tree for Level 2 Classification Based on CART Analysis

1. August Tasseled Cap wetness < −17.5
  1.1. June Tasseled Cap greenness < −4.44
    1.1.1. June TM band 4 < 53.5, THEN Burned
    1.1.2. June TM band 4 > 53.5
      1.1.2.1. June Tasseled Cap brightness < 133.24, THEN Herbaceous
      1.1.2.2. June Tasseled Cap brightness > 133.24, THEN Burned
  1.2. June Tasseled Cap greenness > −4.44
    1.2.1. Tasseled Cap greenness difference < −13.5
      1.2.1.1. June TM band 3 < 28.5, THEN Herbaceous
      1.2.1.2. June TM band 3 > 28.5, THEN Herbaceous
    1.2.2. Tasseled Cap greenness difference > −13.5, THEN Conifer/herbaceous mix
2. August Tasseled Cap wetness > −17.5
  2.1. June Tasseled Cap brightness < 122.5
    2.1.1. Tasseled Cap brightness difference < −2.5
      2.1.1.1. Tasseled Cap wetness difference < 8.54, THEN Conifer
      2.1.1.2. Tasseled Cap wetness difference > 8.54, THEN Conifer
    2.1.2. Tasseled Cap brightness difference > 122.5, THEN Conifer/herbaceous mix
  2.2. June Tasseled Cap brightness > 122.5
    2.2.1. August TM band 6 < 140.5
      2.2.1.1. Tasseled Cap wetness difference < −3.5
        2.2.1.1.1. June TM band 2 < 26.5, THEN Hardwood
        2.2.1.1.2. June TM band 2 > 26.5, THEN Hardwood/herbaceous mix
      2.2.1.2. Tasseled Cap wetness difference > −3.5, THEN Hardwood
    2.2.2. August TM band 6 > 140.5, THEN Conifer

occurred with (1) conifer/herbaceous mix (defined as a mix of conifer and herbaceous with less than 70 percent conifer), which was primarily confused with the related classes of conifer and herbaceous, and (2) hardwood/herbaceous mix (defined as a mix of hardwood and herbaceous with less than 70 percent hardwood), which was primarily confused with the related class hardwood, as well as with agriculture. When conifer/herbaceous mix was combined with conifer, and hardwood/herbaceous mix was combined with hardwood, overall accuracy increased to 85 percent (Kappa statistic was 0.81) and individual class accuracies ranged from 62 percent to 100 percent.

The level 3 CART classification created four rules to subdivide hardwood classes into aspen, willow, and cottonwood; three rules to subdivide herbaceous classes in to sage-grassland and grassland; six rules to subdivide agriculture classes into perennial and annual agriculture; and four rules to subdivide conifer classes into mixed conifer species and Douglas-fir (Table 6). For hardwood classes, slope and band 1 (blue) from each of the June and August TM images were used to make class distinctions. Aspen was distinguished from other hardwood classes by occurring on steeper slopes, which was explained by the other hardwood classes—willow and cottonwood—occurring within stream bottoms. Willow was distinguished from cottonwood by lower reflectance in blue in the August TM image. Within the herbaceous classes, sage-grassland was distinguished from grassland by lower values in the Tasseled Cap wetness difference image, which resulted from August senescence of the grasses. Tasseled Cap brightness and wetness difference images, Tasseled Cap greenness and wetness from the June TM image, and aspect were used to distinguish annual from perennial agriculture. At lower values of brightness difference, annual agriculture had higher greenness difference values than did perennial agriculture. At higher values of brightness difference, perennial agriculture was distinguished by having higher greenness values in the June image. Finally, between conifer classes, Douglas-fir was distinguished from mixed conifer by having lower values in band 6 (thermal) of the August TM image, probably as a result of denser canopies in the Douglas-fir forests.

Table 6. Dichotomous Classification Trees for Level 3 Classifications Based on CART Analyses

1. Hardwood Classifications
  1.1. Slope gradient < 5°
    1.1.1. August TM band 1 < 59.66, THEN Willow
    1.1.2. August TM band 1 > 59.66
      1.1.2.1. June TM band 1 < 62.5, THEN Cottonwood
      1.1.2.2. June TM band 1 > 62.5, THEN Cottonwood
  1.2. Slope gradient > 5°, THEN Aspen
2. Herbaceous Classifications
  2.1. June Tasseled Cap brightness < 156.34
    2.1.1. Tasseled Cap wetness difference < −3.58, THEN Sage-grassland
    2.1.2. Tasseled Cap wetness difference > −3.58, THEN Grassland
  2.2. June Tasseled Cap brightness > 156.34, THEN Grassland
3. Agriculture Classifications
  3.1. Tasseled Cap brightness difference < 40.02
    3.1.1. Tasseled Cap greenness difference < −13.9
      3.1.1.1. Aspect < 1878.5, THEN Perennial agriculture
      3.1.1.2. Aspect > 1878.5, THEN Perennial agriculture
    3.1.2. Tasseled Cap greenness difference > −13.9, THEN Annual agriculture
  3.2. Tasseled Cap brightness difference > 40.02
    3.2.1. June Tasseled Cap wetness < −28.38
      3.2.1.1. June Tasseled Cap greenness < 6.22, THEN Annual agriculture
      3.2.1.2. June Tasseled Cap greenness > 6.22, THEN Perennial agriculture
    3.2.2. June Tasseled Cap wetness > −28.38, THEN Annual agriculture
4. Conifer Classifications
  4.1. June Tasseled Cap greenness < 8.06, THEN Mixed conifer
  4.2. June Tasseled Cap greenness > 8.06
    4.2.1. August TM band 6 < 133.34, THEN Douglas-fir
    4.2.2. August TM band 6 > 133.34
      4.2.2.1. Elevation < 1823.5 m, THEN Mixed conifer
      4.2.2.2. Elevation > 1823 m, THEN Mixed conifer

Table 5. Accuracy Assessment for Level 2 Classification. Columns Represent Reference Data and Rows Represent Classification Data

	Urban	Burned	Agriculture	Conifer	Hardwood	Herbaceous	Conifer/ herbaceous mix	Hardwood/ herbaceous mix
Urban	**19**	0	0	0	0	0	0	0
Burned	0	**31**	0	0	0	0	0	0
Agriculture	0	0	**51**	0	1	2	0	2
Conifer	1	2	1	**63**	9	2	4	1
Hardwood	0	0	0	0	**17**	0	2	3
Herbaceous	0	4	0	0	3	**18**	4	0
Conifer/herbaceous mix	0	0	0	5	1	2	**6**	0
Hardwood/herbaceous mix	0	0	0	0	3	0	0	**2**
Producer's accuracy	79%	84%	98%	93%	50%	75%	38%	25%
User's accuracy	100%	100%	85%	76%	77%	62%	43%	40%

Table 7. Accuracy Assessment for Level 3 Classification. Columns Represent Reference Data and Rows Represent Classification Data

	Urban	Burned	Annual agriculture	Perennial agriculture	Mixed conifer	Douglas-fir	Aspen	Cottonwood	Willow	Sage-grassland	Grass
Urban	**19**	0	0	0	0	0	0	0	0	0	0
Burned	0	**31**	0	0	0	0	0	0	0	0	0
Annual agriculture	0	0	**3**	6	0	0	0	0	0	0	0
Perennial agriculture	4	0	33	**9**	0	0	0	0	1	1	1
Mixed conifer	1	2	0	1	**57**	6	5	1	1	1	2
Douglas-fir	0	0	0	0	3	**2**	3	0	0	0	1
Aspen	0	0	0	0	0	0	**15**	0	0	0	0
Cottonwood	0	0	0	0	0	0	0	**8**	1	0	0
Willow	0	0	0	0	0	0	0	0	**7**	0	0
Sage-grassland	0	0	0	0	0	0	3	0	0	**5**	7
Grass	0	4	0	0	0	0	0	0	0	0	**7**
Producer's accuracy	79%	84%	8%	56%	95%	25%	58%	89%	70%	71%	41%
User's accuracy	100%	100%	33%	18%	74%	22%	100%	89%	100%	36%	64%

Overall accuracy of the level 3 classification was 65 percent and the Kappa statistic was 0.58. Individual class accuracies were highly variable, ranging from 8 percent to 100 percent (Table 7). Particular problems were encountered with distinguishing (1) annual from perennial agriculture, (2) Douglas-fir from other conifer forests, (3) aspen from conifer classes, and (4) sage-grassland from other grasslands without sage.

Discussion

A wide variety of classification options are available for image processing, and no single classification solution will always perform best. Incorporating ancillary data into rule-based classifications, however, has been shown to be an effective approach in certain circumstances. We developed and demonstrated a method for creating and executing such a classification system without extensive *a priori* expert knowledge. This method, based on CART analysis, was easily implemented using commonly available image processing and statistical software.

Overall and individual class accuracies at level 1 were excellent. While many classes were well classified at levels 2 and 3, certain class accuracies were clearly unacceptable. Primary reasons for unacceptable results might have included (1) inadequate spatial resolution for the desired level of classification, especially with respect to level 3 (Jensen and Cowen, 1999), and (2) irresolvable class overlap with the spectral and GIS layers available. Attempts to distinguish problem classes with other classification algorithms were not successful.

One of the strongest advantages of using CART to create classification rules was that a large array of potentially useful data could be entered into the analysis and CART automatically selected which layers were useful and which were not. This selection process distinguished CART from logical channel addition, expert systems, and neural networks. With logical channel addition, additional data must be selected before classification. With expert systems, *a priori* knowledge is necessary to select ancillary data. With neural networks, only useful layers will be used, but the selection might be hidden from the analyst, hindering interpretation of the results and application to other classification problems.

The automatic selection of useful data by CART should not be interpreted as a license to add layers to the analysis indiscriminately. As with any statistical analysis, the uncritical addition of potential explanatory variables increases the possibility of chance agreement between some explanatory variables and the response. Thus, an analyst should only include those data layers that are reasonably believed to have the potential to distinguish classes.

CART is sensitive to large discrepancies in the size (number of observations) of training samples among individual classes. Distinctions within the CART analysis are made based on minimizing total misclassifications for the entire training set. Thus, a class with a larger number of training pixels might have greater weight in the analysis because it potentially contributes a larger number of misclassified pixels. Reasonable efforts, therefore, should be taken to keep the number of training pixels per class roughly equivalent so that within class variations do not overwhelm the among class distinctions that are the primary interest of classification. Furthermore, in selecting training sites, the caveats that apply to all supervised classification methods apply to CART as well. For example, training sites should be taken from relatively homogeneous locations, include all class types known to be present within the area to be classified, and cover the range of conditions present for each class (Jensen, 1996).

In addition to providing predicted classes at terminal nodes, CART analysis reports for each terminal node the probability of misclassification and the probability of membership for each other class. This information can be used to assess the quality of the classification, assign fuzzy class memberships, or conduct Bayesian probability analysis.

Supervised classification with CART analysis is an effective and easily implemented means for creating a ruled-based classification when expert knowledge is insufficient. Applied in appropriate circumstances, it provides an alternative tool for the image analyst wishing to take advantage of ancillary data for classification.

Acknowledgments

The authors wish to thank Dr. Andrew Hansen, Ecology Department, Montana State University, for his advice and support throughout this research. This research was supported by NASA Land Cover Land Use Change Program.

References

Beers, T.W., P.E. Dress, and L.C. Wensel, 1966. Aspect transformation in site productivity research, *Journal of Forestry*, 64:691–692.

Breiman, L., J.H. Friedman, R.A. Olshen, and C.J. Stone, 1984. *Classification and Regression Trees*, Wadsworth International Group, Belmont, California, 358 p.

Bruzzone, L., C. Conese, F. Maselli, and F. Roli, 1997. Multisource classification of complex rural areas by statistical and neural-network approaches, *Photogrammetric Engineering & Remote Sensing*, 63(5):523–533.

Chavez, P.S., 1988. An improved dark-object subtraction technique for atmospheric scattering correction of multispectral data, *Remote Sensing of Environment*, 24:459–479.

Crist, E.P., and R.C. Cicone, 1984. Application of the tasseled cap concept to simulated Thematic Mapper data, *Photogrammetric Engineering & Remote Sensing*, 52:81–86.

Efron, B., and R. Tibshirani, 1991. Statistical data analysis in the computer age, *Science*, 253:390–395.

Elumnoh, A., and R.P. Shrestha, 2000. Application of DEM data to Landsat image classification: Evaluation in a tropical wet-dry landscape of Thailand, *Photogrammetric Engineering & Remote Sensing*, 66(3):297–304.

Goodenough, D.G., M. Goldberg, G. Pluckett, and J. Zelek, 1987. An expert system for remote sensing, *IEEE Transactions on Geoscience and Remote Sensing*, 25(3):349–359.

Huang, X., and J.R. Jensen, 1997. A machine-learning approach to automated knowledge-base building for remote sensing image analysis with GIS data, *Photogrammetric Engineering & Remote Sensing*, 63(10):1185–1194.

Hutchinson, C.F., 1982. Techniques for combining Landsat and ancillary data for digital classification improvement, *Photogrammetric Engineering & Remote Sensing*, 48:123–130.

Jensen, J.R., 1996. *Introductory Digital Image Processing: A Remote Sensing Perspective, Second Edition*, Prentice Hall, Upper Saddle River, New Jersey, 316 p.

Jensen, J.R., and D.C. Cowen, 1999. Remote sensing of urban/suburban infrastructure and socio-economic attributes, *Photogrammetric Engineering & Remote Sensing*, 65(5):611–622.

Kontoes, C., G. Wilkinson, A. Burrill, S. Goffredo, and J. Megier, 1993. An experimental system from the integration of GIS data in knowledge-based image analysis for remote sensing of agriculture, *International Journal of Geographic Information Systems*, 7(3):247–262.

Lawrence, R.L., and W.J. Ripple, 2000. Fifteen years of revegetation of Mount St. Helens: A landscape-scale analysis, *Ecology*, 81:2742–2752.

MathSoft, Inc., 1998. *S-PLUS for UNIX Users' Manual*, MathSoft, Inc., Seattle, Washington, 336 p.

Ricchetti, E., 2000. Multispectral satellite image and ancillary data integration for geological classification, *Photogrammetric Engineering & Remote Sensing*, 66(4):429–435.

Skidmore, A.K., B.J. Turner, W. Brinkhof, and E. Knowles, 1997. Performance of a neural network: Mapping forests using GIS and remotely sensed data, *Photogrammetric Engineering & Remote Sensing*, 63(5):501–514.

Strahler, A.H., T.L. Logan, and N.A. Bryant, 1978. Improving forest cover classification accuracy from Landsat by incorporating topographic information, *Proceedings of 12th International Symposium on Remote Sensing of Environment*, 20–26 April, Manila, Philippines, pp. 927–942.

Trotter, C.M., 1991. Remotely-sensed data as an information source for geographical information systems in natural resource management: A review, *International Journal of Geographic Information Systems*, 5:225–239.

Venables, W.N., and B.D. Ripley, 1997. *Modern Applied Statistics with S-Plus, Second Edition*, Springer-Verlag, New York, NY, 548 p.

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

(Received 19 July 2000; accepted 28 March 2001)

Forthcoming Articles

Peggy Agouris, *Anthony Stefanidis*, and *Sotirios Gyftakis*, Differential Snakes for Change Detection in Road Segments.

John C. Brock, *Asbury H. Sallenger*, *William B. Krabill*, *Robert N. Swift*, and *C. Wayne Wright*, Recognition of Fiducial Surfaces in Lidar Surveys of Coastal Topography.

R.P. Gauthier, *M. Maloley*, and *K.B. Fung*, Land-Cover Anomaly Detection along Pipeline Rights-of Way.

Zhilin Li, *Zhu Xu*, *Minyi Cen*, and *Xiaoli Ding*, Robust Surface Matching for Automated Detection of Local Deformations Using Least-Median-of-Squares Estimator.

Don Light, An Airborne Direct Digital Imaging System.

Mark A. Maloy and *Denis J. Dean*, An Accuracy Assessment of Various GIS-Based Viewshed Delineation Techniques.

Tazio Strozzi, *Urs Wegmüller*, *Luigi Tosi*, *Gabriele Bitelli*, and *Volker Spreckels*, Land Subsidence Monitoring with Differential SAR Interferometry.

C. Vincent Tao, Database-Guided Automatic Inspection of Vertically Structured Transportation Objects from Mobile Mapping Image Sequences.

Jan A.N. van Aardt and *Randolph H. Wynne*, Spectral Separability among Six Southern Tree Species.

Richard M. Westaway, *Stuart N. Lane*, and *D. Murray Hicks*, Remote Sensing of Clear-Water, Shallow, Gravel-Bed Rivers Using Digital Photogrammetry.

Yun Zhang, Texture-Integrated Classification of Urban Treed Areas in High-Resolution Color-Infrared Imagery.