

Identifying “Redtops”: Classification of Satellite Imagery for Tracking Mountain Pine Beetle Progression through a Pine Forest

Richard Cutler¹

Department of Mathematics and Statistics, Utah State University, Logan UT 84322-3900

Leslie Brown¹

Department of Mathematics and Statistics, Utah State University, Logan UT 84322-3900

James Powell¹

Department of Mathematics and Statistics, Utah State University, Logan UT 84322-3900

Barbara Bentz

USDA Forest Service, Rocky Mountain Research Station, Logan, UT 84321

Adele Cutler

Department of Mathematics and Statistics, Utah State University, Logan UT 84322-3900

Abstract

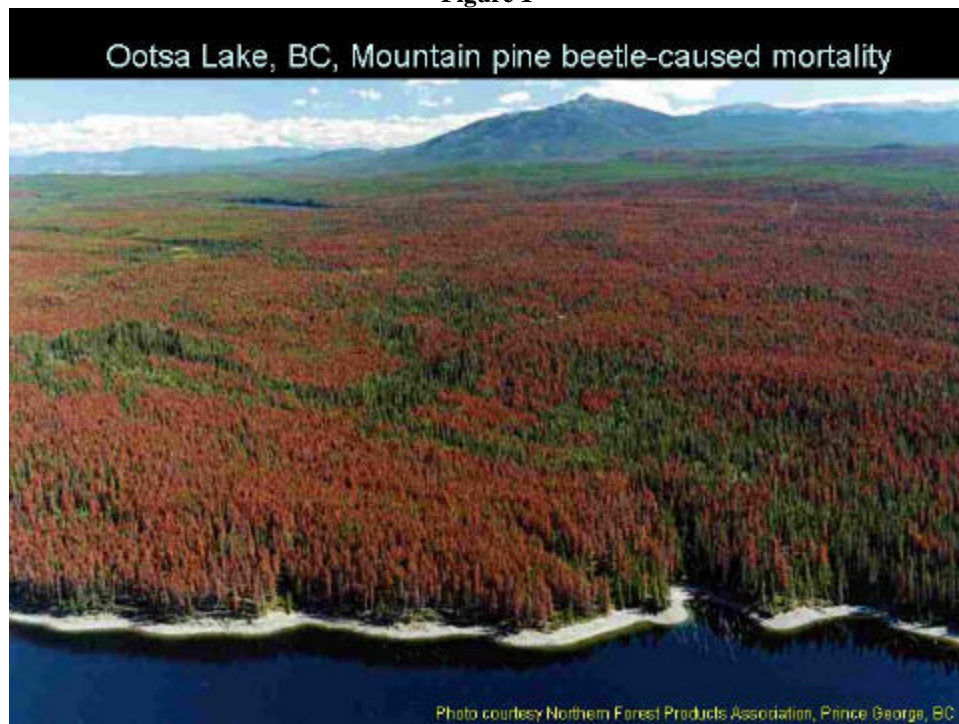
Mountain pine beetles are a pest indigenous to the pine forests of the western United States. Capable of exponential population growth, mountain pine beetles can destroy thousands of acres of trees in a short period of time. Much research has been carried out on the mathematical modeling of mountain pine beetle phenology. The research reported here is part of a larger project to demonstrate the application of, and evaluate, differential equation models for mountain pine beetle progression through pine forests. The study area is the Sawtooth National Recreation area in Idaho. To provide input parameter to the mathematical models, and to measure the changes in the pine forest in the study area, IKONOS satellite imagery was used to classify the vegetation of the study area. Five classifiers—linear discriminant analysis, quadratic discriminant analysis, k -nearest neighbor discriminant analysis, classification trees and random forests—were applied to raw and transformed multispectral and panchromatic satellite imagery, with and without an elevation variable. Quadratic discriminant analysis and random forests proved to be the best classifiers as measured by cross-validated error estimates, with overall classification rates of about 12% without elevation, and about 5% when elevation was included. Redtops were relatively easy to identify, with misclassification rates of about 5%—6%, but green lodgepole pine and Douglas fir were relatively difficult to discriminate between and had much higher misclassification rates.

¹ Research partially supported by NSF Grant No. DMS 0077663.

1. Introduction

Mountain pine beetles (*Dendroctonus ponderosae* Hopkins) are a pest indigenous to, and having co-evolved with, pine forests in the western United States. Outbreaks of the beetles may be quite devastating: mountain pine beetle populations are capable of exponential growth, and of killing thousands of acres of trees in a very short time. Figure 1 is picture of a lodgepole pine forest in British Columbia, Canada, in the early 1980s with substantial mountain pine beetle induced mortality.

Figure 1



Although most western pine species are suitable hosts for mountain pine beetles, their principal hosts are ponderosa pine, *Pinus ponderosae* Lawson, and lodgepole pine, *Pinus contorta* Douglas (Logan and Powell, 2001). The adult pine beetles burrow through the bark of the host and lay their eggs in the phloem of the host. The hatched larvae feed on the phloem, ultimately killing the host, and then, as adults, emerge to attack a new host. As the hosts die their needles turn red—hence the term “redtops” for newly killed trees—and then to gray over time.

Mountain pine beetles fulfill a very important ecological role in pine forests, such as those of lodgepole pine, that are rejuvenated by fire. The fallen needles from trees killed by the beetles are very combustible fuel for forest fires. The dead trees themselves provide a “ladder” for fire to reach the crowns. Fallen, rotting trees that are struck by lightning may smolder for days until the right conditions for a forest fire can occur. Without major forest fires, it is quite possible that lodgepole pine would be completely supplanted by Douglas fir and spruce that more vigorous than lodgepole pine when growing in shade.

The pines that are regularly afflicted by mountain pine beetle infestations have developed significant chemical defenses to attack, including the transportation of sap to attack sites. In order to overcome these defenses, mountain pine beetles must mount a synchronized, mass attack, at an appropriate time. Even a mass attack in the middle of winter will only result in lots of frozen mountain pine beetles. A substantial amount of research into mechanisms and mathematical models for timing and synchrony of attack has been carried out. See for example, White and Powell (1998), Powell *et al.* (2000a), Powell *et al.* (2000b), and Jenkins *et al.* (2001).

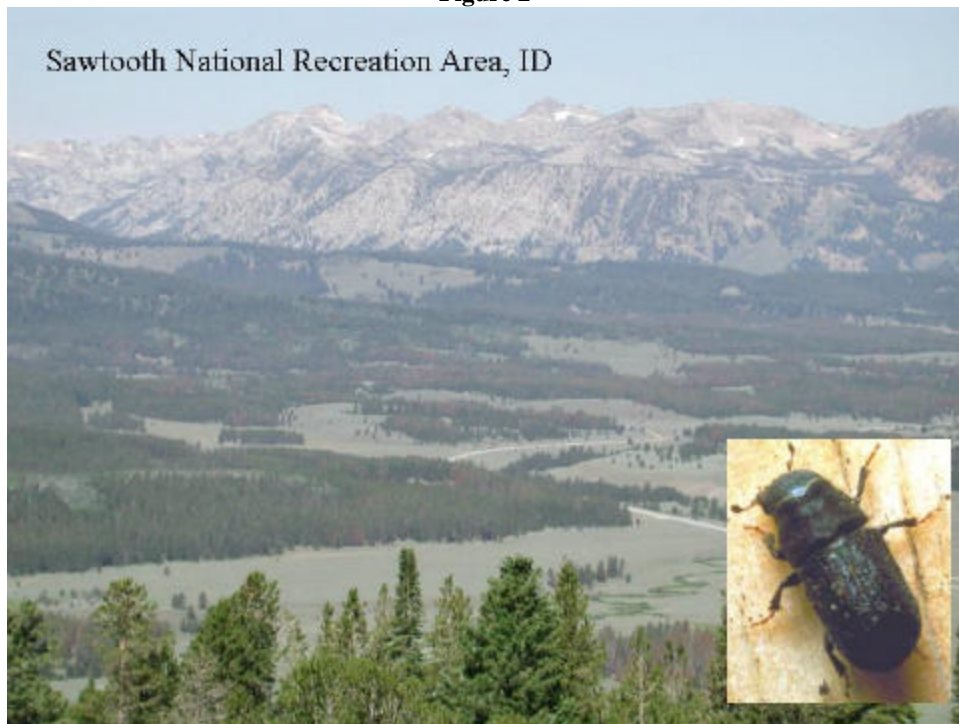
Pine species that grow at higher elevations, such as, such as bristlecone pine and white bark pines, have not been subject to mountain pine beetle infestation because the beetles cannot survive in the cold at the higher elevations. With the prospect of global warming comes the concern that these species will be devastated by mountain pine beetle attacks to which they have little or no resistance (see Logan *et al.*, 2003).

The research project that is the subject of this paper is part of an on-going collaboration between researchers in the USDA Forest Service Rocky Mountain Research Station unit, based in Logan Utah, and faculty in the Department of Mathematics and Statistics at Utah State University, to develop mathematical models for the phenology of mountain pine beetles. One aspect of the work involves applying the mathematical models derived in earlier work to a real landscape, the Sawtooth National Recreation Area (the SNRA) in Idaho. Figure 2 is a picture of the SNRA with an inset of an adult mountain pine beetle.

One component of the project involves classification of satellite imagery of the SNRA. The purpose of the classification is to estimate input parameters for the differential equation models, such as numbers of potential hosts and redtops, to measure the progression of the pine beetle infestation through the forest, and to evaluate the fit of the models.

The subject of this paper is the classification of the SNRA into different vegetation classes using satellite imagery from the first year of data collection. Of particular interest is the identification of the recently killed lodgepole pines, the so-called “redtops” in the title of this paper. Five different classifiers were evaluated, ranging from the simple linear (LDA) and quadratic discriminant analysis (QDA) procedures due to Fisher (1936, 1938), through classification trees (Breiman *et al.*, 1984) and random forests (Breiman, 2001). This classification exercise differs from many ecological applications of discriminant analysis in that the form of the classification model is of little intrinsic interest: only the classification accuracy is of interest to the authors.

Figure 2



2. Data

The IKONOS satellite was launched in September of 1999 and was the first commercial satellite to collect 1-meter black-and-white (“panchromatic”) and 4-meter multispectral data. The satellite orbits the earth at an altitude of about 680 kilometers approximately every 98 minutes in a sun-synchronous orbit. (Spectral data for visible light bands collected in the middle of the night are not particularly interesting!) The IKONOS satellite can produce 1-meter imagery of the same geography every 3 days.

The multispectral imagery has 4 bands: Blue ($0.45\mu\text{m}$ — $0.52\mu\text{m}$), Green ($0.52\mu\text{m}$ — $0.60\mu\text{m}$), Red ($0.63\mu\text{m}$ — $0.70\mu\text{m}$), and Near Infrared ($0.76\mu\text{m}$ — $0.85\mu\text{m}$). The black-and-white, 1-meter imagery extends from $0.45\mu\text{m}$ to $0.90\mu\text{m}$.

Through a process called *resolution merging* it is possible to construct a pseudo-1-meter resolution image using the four multispectral bands and the panchromatic band. Throughout this paper, we have consistently used the imagery at 4-meter resolution, and when we use the black-and-white band, we compute an average value for the sixteen $1\text{m} \times 1\text{m}$ pixels within each $4\text{m} \times 4\text{m}$ pixel.

Several derived variables from the multispectral data were also considered in our analyses. *Tasseled cap transformations*—linear transformations of the raw spectral data that effectively rotate the data onto new axes corresponding to physical characteristics of vegetation—were first proposed by Kauth and Thomas (1976) for use on Landsat data. The four transformed variables, or indices, are called Soil Brightness, Greenness, Yellowness, and Non-such. The first 3 transformations were initially determined from scatter plots to represent soil moisture content, vegetation moisture content

(“Greenness”), and the yellowness that some crops assume as they mature. Then the vectors were orthonormalized using the Gram-Schmidt process, starting with the soil brightness vector. Finally, the Non-such transformation is the 4-dimensional unit vector that is orthogonal to the first 3 transformations. The tasseled cap transformation coefficients are given in the table below.

Table 1
Tasseled Cap Transformations for IKONOS Satellite Imagery.

Tasseled Cap Indices	Raw Spectral Bands			
	<i>Blue</i>	<i>Green</i>	<i>Red</i>	<i>Near Infrared</i>
<i>Soil Brightness</i>	0.576	0.495	0.452	0.469
<i>Greenness</i>	-0.320	-0.190	-0.300	0.880
<i>Yellowness</i>	-0.310	0.830	-0.430	-0.090
<i>Non-Such</i>	-0.680	0.110	0.720	0.020

Tasseled cap transformations have some conceptual advantages over the raw multispectral data. First, by corresponding to physical characteristics of soil and vegetation they are easier to interpret when used in a model or other statistical method. Second, some of the best statistical classification procedures—notably classification trees—are *not* invariant to linear transformations, so the transformed data may yield better results than the raw data. Third, by construction, the tasseled cap transformed data are orthogonal, thus obviating any concerns of collinearity when modeling.

A non-linear transformation of the multispectral data that we also considered in the *Normalized Difference Vegetation Index (NDVI)*, defined by

$$NDVI = (Near\ Infrared - Red) / (Near\ Infrared + Red)$$

This index has been used for many years (by the US Geological Survey, and others) to measure and monitor plant growth, vegetation cover, and biomass production from multispectral satellite imagery. The idea behind the index is that chlorophyll causes high absorption of red light, whereas a plant’s mesophyll leaf structure causes considerable reflectance in the near infrared region of the electromagnetic spectrum. Thus, healthy vegetation (as opposed to water, road, dirt, and dead trees) should have large values of the NDVI. See, for example, Tucker (1979), Jackson *et al.* (1983), and Tucker *et al.* (1991).

Elevation was obtained from 30-meter resolution digital elevation maps and added to the data set. Elevation was considered to be an important discriminator between forest types in the SNRA. Sub-alpine fir only occurs above about 7000 feet and at about 6750 feet to 7250 feet the forest in the SNRA makes a transition from being mainly lodgepole pine to being mainly Douglas fir. Neither sub-alpine fir nor Douglas fir is a suitable host for the mountain pine beetle.

Creation of a training data set, comprising observations of vegetation on the ground and multispectral satellite imagery for the same points, proved to be a much more difficult task than originally anticipated. The fundamental problem is one of matching locations on the ground with pixels in the satellite image. In principle, this matching should be straightforward—the nominal accuracy of the GPS units used is less than 1

meter, but in practice, the combined error of the positioning on the ground and of the satellite imagery was of the order of tens of meters, not one or two meters. We speculate on why the error was so large.

1. The satellite was not quite vertical over the SNRA, and the tree crowns are 20m—40m above the ground. The IKONOS satellite is observing the reflectance from the tree crowns and projecting their position on the ground, but not quite orthogonally. Using simple trigonometric arguments, one can calculate that the displacement due to non-orthogonal projection is of the order of 3m—6m.
2. The satellite image is perfectly flat, whereas the SNRA is not. It is not clear exactly how the haphazard differences in aspect and elevation affect the coordinates of the points.
3. The number of different satellite signals the GPS units could pick up was smaller than recommended, perhaps increasing the error in the GPS measurements.
4. The positional accuracy of the IKONOS imagery is only guaranteed to be within 25 meters. Thus, the error—or a substantial portion of the error—may be coming from the imagery itself.

The net impact of the unexpected problems in matching ground observation to pixels on the satellite image was that the training data set comprised data values for sites on the ground that could easily be matched with the satellite image due to, for example, proximity to a road or stream, or because a tree was in an otherwise open area. Most of the training data sample points came from a small geographic subset of the SNRA that was easily accessible. As a consequence of the sampling methodology, and the geographic restriction, the training data is not like a representative sample of sites in the SNRA. The impact of this sample of convenience data collection on the predicted image will be evaluated by the use of validation samples whose collection is already underway.

The training data comprises 699 observations on 10 “vegetation” classes. The vegetation classes and the number of observations for each vegetation class are recorded in Table 2.

Table 2
Counts of observations in the training data set for all vegetation classes.

Vegetation Class	Number of Points	Vegetation Class	Number of Points
<i>Agriculture</i>	106	<i>Redtop</i>	68
<i>Dirt</i>	53	<i>Road</i>	68
<i>Douglas Fir</i>	66	<i>Sagebrush</i>	29
<i>Grass</i>	15	<i>Shadow</i>	84
<i>Green Lodgepole Pine</i>	55	<i>Water</i>	155

Not all the “vegetation” classes are actually types of vegetation: we included lakes and rivers as *Water*. *Roads* and *Dirt* were also vegetation classes. The *Shadow* class is forced upon us by the fact that it is very clear on the spectral images that some trees are

in the shadows of neighboring larger trees. In some sense, one could view the *Shadow* class as a “Don't know” class.

3. Statistical Methods

Five statistical classification procedures were applied to the SNRA multispectral data. A brief description of each procedure is given below.

Linear Discriminant Analysis

Due to Fisher (1936, 1938), linear discriminant analysis (LDA) effectively separates the groups using intersecting hyper-planes. For multivariate normal data, with equal covariance matrices for the different groups, linear discriminant analysis minimizes the expected misclassification rate (Johnson and Wichern, 2002).

Cutoff values for classification into the different vegetation types may be controlled by specifying prior probabilities of membership in the different vegetation classes. Because the numbers of observations in the training data set varied so widely for the different vegetation classes, we elected to set the prior probabilities proportional to the numbers of occurrences of each vegetation type in the training data set. This is one of the standard options within the SAS computer software used to fit the linear discriminant analysis, quadratic discriminant analysis, and k -Nearest Neighbor classification models.

Quadratic Discriminant Analysis

Quadratic discriminant analysis (QDA) generalizes linear discriminant analysis to the situation where the covariance matrices for the different classes may be different. The quadratic discriminant function takes the form of hyper quadratic functions. As with linear discriminant analysis, quadratic discriminant function may be derived as the minimizer of the expected misclassification rate, and hence is the optimal classifier in the case of multivariate normal data with different covariance matrices for the different groups.

k-Nearest Neighbor Classifier

Nearest neighbor classifiers (Fix and Hodges, 1951) were the first non-parametric alternative to parametric classifiers, such as linear and quadratic discriminant analysis, and logistic regression. No assumptions about the distribution of the data are made. Instead, each point is classified as a function of its nearest neighbors as measured by Euclidean or Mahalanobis distance. The number of neighboring observations that are used to classify each observation is denoted by k .

Classification Trees

Classification trees were developed by, and have been popularized as a classification tool by Breiman *et al.* (1984). Classification trees are a non-parametric method for discrimination. The basic methodology is one of recursive partitioning of the data. At each step, a variable and a cut point are chosen so that splitting the data into groups on

the basis of the chosen variable at the cut point leads to the largest decrease in an objective function, such as the residual deviance or the misclassification rate. The full tree, with every data point perfectly classified, over-fits the data in the sense that many of the splits at lower levels of the tree are modeling noise, so methods for “pruning” the tree or determining the appropriate number of splits are typically applied. In our analyses, the size of the appropriate classification tree was estimated by cross validation using the `cv.tree` and `prune.tree` functions in the `Splus@` package. A second tree of the estimated size was then fit to the data.

Random Forests

As the name suggests, random forests (Breiman, 2001) are collections of classification trees. Random forests is an *ensemble classifier*. Ensemble classifiers are discriminant procedures in which many classifiers are fit to the same data, and the predicted values for observations are obtained by majority vote of the predicted classes of the individual classifiers. Previous applications of ensemble classifiers to remotely sensed data include Steele (2000) and Steele and Patterson (2000 and 2002). The basic algorithm for random forests is as follows:

1. Many bootstrap samples of the original data are drawn.
2. On each bootstrap sample, a classification tree is fit.
3. At each node, in each classification tree, a randomly selected subset of the variables is made available for splitting.
4. The trees are fully grown and no pruning takes place.
5. For each tree, fitted on a single bootstrap sample, predictions are generated for all data values that were in the original data set, but which were not in the bootstrap sample. In the terminology of Breiman (1996), these predictions are for the “out-of-bag” data values.
6. For a given data value, the predicted class is the class with the highest count among out-of-bag predictions for that data point. Ties are split randomly.

Using only randomly chosen subsets of variables at each node in each of the classification trees ensures that the resulting trees will be quite different, and that is what is required to derive the full benefit of ensemble classifiers.

Random forests is one of the best, and arguably the best, of the classifiers currently available to statisticians and other researchers. A drawback of random forests is the difficulty in interpretation of the results: there is no simple formula for the predictions from random forests that may be interpreted.

Estimating Error Rates

For the LDA, QDA, *k*-NN and classification tree methods, two error rates were computed. The *resubstitution error rate* or *apparent error rate* (Johnson and Wichern, 2002) was computed by first fitting the appropriate model on all the data, and then calculating a predicted classification for each observation in the data set using the fitted

model. While intuitively appealing, resubstitution error rates tend to underestimate the actual error rate (also known as the prediction or generalization error rate).

The 10-fold cross-validation estimate of the error rate was also computed for each of LDA, QDA, k -NN and classification trees. One thousand random permutations of the data were generated. Each permutation was then split into two pieces, with the first 629 observations being assigned to be a training data set and the remaining 70 observations comprising a test data set. The four classifiers were then fit the training data, evaluated on the test data, and the predictive error rates averaged over all 1000 samples. The 10-fold cross-validation error estimates remain the gold standard for comparing different classifiers, and hence are the main focus of discussion in the results section.

For random forests, because only the predicted values for the “out-of-bag” observations are used in building the classifier, so the error rate is effectively a cross-validated error rate, and no resubstitution error rate is available.

4. Results

An initial classification was carried out with only the 4 multispectral bands (at 4-meter resolution). For the k -nearest neighbor classifier, the values of k that were considered were 4, 6, 8, 12, 16 and 20. Preliminary numerical and graphical summary analyses of the raw multispectral data indicated that it was approximately multivariate normal in distribution, but that the covariance matrices for the different groups were quite different. To illustrate this latter point, Table 3 contains the covariance matrices for the “Dirt” and “Green lodgepole pine” vegetation classifications. The variances and covariances for dirt are generally an order of magnitude, or more, larger than the corresponding variances and covariances for green lodgepole pine. Thus, one might expect a priori that quadratic discriminant analysis would perform well on this data set.

Table 3
Covariance matrices of four spectral bands for Dirt and Green lodgepole pine vegetation classes.

Vegetation Class	Spectral Band	Blue	Green	Red	Near Infrared
Dirt	<i>Blue</i>	111	140	122	74
	<i>Green</i>	140	221	222	147
	<i>Red</i>	122	222	253	178
	<i>Near Infrared</i>	74	147	178	160
Green Lodgepole Pine	<i>Blue</i>	4.4	6.3	8.3	4.8
	<i>Green</i>	6.3	12.8	14.4	15.9
	<i>Red</i>	8.3	14.4	19.1	15.6
	<i>Near Infrared</i>	4.8	15.9	15.6	73.1

Table 4 contains the overall resubstitution and cross-validated error rates for the 5 classification procedures. For the k -NN procedure only the results for $k=4$ and $k=8$ are included in Table 4. The misclassification rates were worse for the higher values of k that we considered. Note that in all cases, the cross-validated error rate is higher than the resubstitution error rate, and in most cases substantially higher.

Table 4.
Overall error rates for LDA, QDA, k-NN for k=4 and k=8, classification trees, and random forests.

Classification Method	Resubstitution Error Rate	Cross-validated Error Rate
<i>Linear Discriminant Analysis</i>	13.30%	14.10%
<i>Quadratic Discriminant Analysis</i>	8.58%	11.68%
<i>4-Nearest Neighbor</i>	9.87%	16.45%
<i>8-Nearest Neighbor</i>	11.16%	14.10%
<i>Classification Tree</i>	9.36%	13.20%
<i>Random Forests</i>	*	12.40%

As expected, quadratic discriminant analysis does well. Indeed, QDA has the lowest cross-validated error rate of any of the procedures, but the differences among the procedures are relatively small: the best classifier, QDA, has a cross-validated error rate of 11.68% and the worst classifier, the nearest neighbor method with $k=4$, had a cross-validated error rate of only 16.45% .

An overall error rate of less than 12% might well have been sufficient for the purposes of the study if the error rate was approximately constant across the different vegetation types. That was not the case. Table 5 contains the cross-validated error rates for QDA by vegetation type, and Table 6 is a portion of the cross-validated error matrix that includes the vegetation classes of most interest to the authors: redtops, green lodgepole pine, and Douglas fir.

Table 5
Misclassification rates by vegetation class, for QDA.

Vegetation Type	Cross-validated Error Rate
<i>Agriculture</i>	0.07%
<i>Dirt</i>	7.88%
<i>Douglas Fir</i>	25.19%
<i>Grass</i>	72.44%
<i>Green Lodgepole Pine</i>	43.90%
<i>Redtop</i>	4.20%
<i>Road</i>	17.68%
<i>Sagebrush</i>	17.77%
<i>Shadow</i>	7.13%
<i>Water</i>	0.02%

Table 6
Partial error matrix for QDA.

True Vegetation Type	Classified as Vegetation Type				
	<i>Grass</i>	<i>Green Lodgepole Pine</i>	<i>Douglas Fir</i>	<i>Redtop</i>	<i>All Other</i>
<i>Grass</i>	27.56%	0.28%	6.11%	4.84%	61.21%
<i>Green Lodgepole Pine</i>	3.90%	56.10%	40.00%	0.00%	0.00%
<i>Douglas Fir</i>	1.38%	23.82%	74.80%	0.00%	0.00%
<i>Redtop</i>	1.26%	0.01%	0.00%	95.80%	2.93%
<i>All Other</i>	0.40%	0.16%	0.20%	0.60%	98.64%

On the positive side, redtops, the main subject of this paper, seem to be rather easy to distinguish from other vegetation types. Over 95% of redtops were correctly classified as redtops (Table 5), and other vegetation types are generally not misclassified as redtops (Table 6). Other vegetation types that are well-classified are agriculture and water, both with error rates less than 0.1%, and even dirt and shadow, with error rates between 7% and 8%. Grass, on the other hand, is misclassified at a much higher rate. A pixel that was identified as being grass in the ground surveys was about 3 times as likely to be misclassified as correctly classified! However, most of the misclassification of grass pixels is into vegetation classifications that are not of interest in the project. Slightly over 6% of grass pixels were misclassified as Douglas fir, slightly less than 5% were misclassified as green lodgepole pine, but over 60% were misclassified as one of the other vegetation types (see Table 6). The misclassification of grass is, therefore, of little consequence to the mountain pine beetle project.

Some misclassifications that are of concern are Douglas fir being misclassified as green lodgepole pine, and green lodgepole pine being misclassified as Douglas fir. The green lodgepole pine are the potential hosts for the next round of mountain pine beetle attacks, and the numbers and densities of lodgepole pine are input parameters into the differential equation models for the mountain pine beetles. From Table 6 we see that 40% of green lodgepole pine is misclassified as Douglas fir, while almost ¼ of the Douglas fir pixels have been misclassified as green lodgepole pine.

Figure 3 is a classification of a portion of the SNRA in which much of the training data was collected. This image illustrates some of the positive aspects and the problems with the classifications. In the image, red denotes redtops. Both the placement and density of redtops are very consistent with the observations of the persons who collected the training data in that area.

Figure 3
Predictions of vegetation types for a portion of the SNRA using QDA.



Roads are clearly visible in white. Water is coded as blue, and both a river and lake are clearly delineated. Note that the center of the lake has been misclassified as shadow. Grass is yellow in the image. There is very little grass to be seen in Figure 3. This is at least in part due to the misclassification of grass as other vegetation types, notably sagebrush (which is cyan in the image), agriculture, and road.

The light green in Figure 3 is green lodgepole pine. There is a lot of light green in the image and that is consistent with the fact that this portion of the SNRA has a lot of lodgepole pine forest. Douglas fir is coded as dark green in Figure 3. It is a little hard to distinguish from the black of the shadow class, but one can see that there is a substantial amount of dark green in middle of Figure 3. This preponderance of dark green on the image is evidence of the high rate of misclassification of green lodgepole pine as Douglas fir because this is an area in the SNRA where there is very little Douglas fir. The misclassification of Douglas fir and green lodgepole pine is of some consequence to the project because green lodgepole pine are potential hosts for the mountain beetle, whereas Douglas fir are not and, as noted earlier, the numbers or densities of green lodgepole pine trees are input parameters to the differential equation models.

When thinking about the misclassifying the two tree species, elevation is a variable that comes to mind almost immediately. The forests in the SNRA change from lodgepole pine at lower elevations to Douglas fir at higher elevations, with the transition zone being at about 6800 to 7200. This strongly suggests that elevation should be included in the classification procedure.

The second set of classifications of the SNRA satellite data involved the 4 multispectral bands, the black-and-white band, the tasseled cap transformed bands, the NDVI index, and elevation. For linear and quadratic discriminant analysis, and for the nearest neighbor method, a certain amount of variable selection was required. Cross-validated error rates were used to decide between models with different variables as well as to compare different classification procedures. For LDA, QDA, and k-NN, the results reported are for the “best” combination of variables. For all 3 of those procedures, the best combination of variables was the 4 multispectral bands, blue, green, red, and near infrared, plus elevation.

Classification trees and random forests used all the variables, although to varying degrees. In all 5 procedures the red and near infrared spectral bands were by far the most important predictors of vegetation class, followed by elevation.

Overall misclassification rates for the 5 procedures are given in Table 7. As in Table 4, both the resubstitution and cross-validated error rates are reported.

Table 7.

Overall error rates for LDA, QDA, k-NN for k=4 and k=8, classification trees, and random forests, using all raw and transformed spectral variables and elevation.

Classification Method	Resubstitution Error Rate	Cross-validated Error Rate
<i>Linear Discriminant Analysis</i>	9.16%	9.68%
<i>Quadratic Discriminant Analysis</i>	4.29%	5.34%
<i>4-Nearest Neighbor</i>	5.15%	6.67%
<i>8-Nearest Neighbor</i>	6.29%	7.65%
<i>Classification Tree</i>	5.31%	6.32%
<i>Random Forests</i>	*	4.86%

The most striking aspect of Table 7 is that the error rates are substantially lower—in several cases more than 50% lower—than those reported in Table 4 using only the 4 multispectral bands. The improvement is almost entirely due to the inclusion of the elevation variable. As before, quadratic discriminant analysis performs well despite the marked skewness of the elevation measurements, with an overall cross-validated error rate of 5.34%, and is surpassed only by random forests, with a cross-validated error rate of 4.86%.

Much of the improvement in classification accuracy with the inclusion of elevation is reflected in the difficult-to-classify vegetation types: grass, green lodgepole pine, and Douglas fir. Table 8 contains the misclassification rates by vegetation class for QDA with the 4 raw multispectral variables and elevation. The results for random forests, which also had an overall cross-validated error rate of about 5%, were very similar. Table 8 shows how the improvement in classification accuracy is distributed among the vegetation types. The error rates in classifying Douglas fir and green lodgepole pine have been essentially halved with the inclusion of elevation in the classification models. The misclassification rates for grass, dirt, and sagebrush have been reduced by more than half, although these are not vegetation classes of particular interest to the mountain pine beetle project. Interestingly, the misclassification rate for redtops is actually slightly higher for QDA when elevation is included compared to when elevation is not included (5.86% versus 4.20%), but remains relatively low. The misclassification rate for pixels

identified as water on the ground is also higher for the model with elevation than without elevation, but still is very low, less than 1%.

Table 8
Misclassification rates by vegetation class, for QDA with 4 multispectral bands and elevation.

Vegetation Type	Cross-validated Error Rate
<i>Agriculture</i>	0.00%
<i>Dirt</i>	2.06%
<i>Douglas Fir</i>	13.82%
<i>Grass</i>	24.45%
<i>Green Lodgepole Pine</i>	20.15%
<i>Redtop</i>	5.86%
<i>Road</i>	2.78%
<i>Sagebrush</i>	7.43%
<i>Shadow</i>	4.06%
<i>Water</i>	0.63%

Table 9 is analogous to Table 6, and contains a partial error matrix for the QDA model including elevation. As in Table 6, it is clear that redtops are relatively easy to classify, with a misclassification rate of less than 6% and only grass is misclassified as redtop at a rate of more than a fraction of 1%.

The change in the misclassification rates of grass into other vegetation types when elevation is included in the model is quite remarkable. In Table 6 grass was misclassified as a vegetation class other than green lodgepole pine, Douglas fir or redtop over 60% of the time. When elevation is included, that figure is reduced to less than 7%.

Table 9
Partial error matrix for QDA with 4 multispectral bands and elevation.

True Vegetation Type	Classified as Vegetation Type				
	<i>Grass</i>	<i>Green Lodgepole Pine</i>	<i>Douglas Fir</i>	<i>Redtop</i>	<i>All Other</i>
<i>Grass</i>	75.55%	5.40%	7.10%	5.33%	6.62%
<i>Green Lodgepole Pine</i>	1.13%	79.85%	19.02%	0.00%	0.00%
<i>Douglas Fir</i>	1.69%	12.12%	86.18%	0.00%	0.00%
<i>Redtop</i>	2.45%	0.00%	1.74%	94.14%	1.67%
<i>All Other</i>	0.76%	0.21%	0.00%	0.38%	98.65%

The confusion between green lodgepole pine and Douglas fir remains as the largest problem with the classification of this satellite imagery. Although the misclassification rates for these two species are greatly reduced when elevation is used, they remain relatively high at about 20% for green lodgepole pine and 15% for Douglas fir.

5. Concluding Remarks

Our analyses suggest that classification of vegetation in the Sawtooth National Recreation area using elevation and IKONOS multispectral imagery at 4m resolution can yield usable results. These analyses should be regarded as very preliminary. There are many outstanding issues that remain unresolved at this point.

Difficulty in matching data points on the ground with pixels in the satellite imagery led to a very careful selection of training data from small areas that the researchers collecting the data were particularly familiar with, and in locations where the data type was “obvious” from the raw image. The impact of this highly non-random, and non-representative selection process on the classifications has yet to be fully determined. Data from another study spread over a wider range in the SNRA collected on 15m, 30m, and 60m square plots will be used for validation purposes. In addition, data collected in subsequent years sampling may be used for validation. Most vegetation types, such as water, Douglas fir, road, and dirt, will not change from year to year. Some green lodgepole pine will be afflicted by pine beetle attacks and will have turned into redtops, but the converse transformation is not possible: if a pixel is predicted by a classifier to be a redtop, and is subsequently sampled and found to be green lodgepole pine or some other vegetation class, then the classification of that pixel must have been incorrect.

Some vegetation types were not included in our analyses because very little or no data had been collected on them. Sub-alpine fir occurs in small amounts within the SNRA. Also, data was not collected on trees that have been killed in the past by mountain pine beetles, and which are still standing (snags). There are substantial areas of such trees in the SNRA, where mountain pine beetle outbreaks are a fact of life.

Even if a perfect classification were possible, some problems remain with regard to the use of the data in the mountain pine beetle project. The youngest, smallest lodgepole pine does not get attacked by mountain pine beetles, and distinguishing between small and large trees using satellite imagery would be extremely difficult, and likely quite unreliable.

Another problem concerns the age of the redtops. Only trees killed in the most recent growing season harbor pine beetles, yet trees that were killed 2, 3, even 4 years ago retain some degree of redness. Distinguishing between trees killed in consecutive growing seasons using satellite imagery is not feasible, and is even difficult to do on the ground. Thus, merely identifying redtops using the satellite imagery may give a misleading impression as to the numbers and whereabouts of pine beetles in the forest.

Two more years of imagery for the SNRA—perhaps from different sources and at a higher resolution—will be collected and analyzed. The classification work in this paper represents the first steps toward building classifiers for the entire SNRA. Already it is clear that two of the classifiers we tested, quadratic discriminant analysis and random forests, work quite well on this kind of data. We have also learned that the major classification problem will be one of separating Douglas fir from green lodgepole pine, and future ground data collection can emphasize these two vegetation types.

6. Acknowledgements

We would like to thank Dan Endreson and Jesse Logan of the USDA Forest Service for stimulating discussions and for help with data collection for the project. T.B. Murphy played an important role in the data collection and in the development of the training data set.

7. Bibliography

- Biesinger, Z., Powell, J., Bentz, B., and Logan, J. (2000). "Direct and indirect parametrization of a localized model for the mountain pine beetle—lodgepole pine system." *Ecological Modelling* **129** 273—296.
- Breiman, L. (2001). "Random Forests." *Machine Learning* **45** 1 5—32.
- Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). *Classification and Regression Trees*. Chapman & Hall, New York.
- Fisher, R.A. (1936). "The Use of Multiple Measurements in Taxonomic Problems." *Annals of Eugenics* **7** 179—188.
- Fisher, R.A. (1938). "The Statistical Utilization of Multiple Measurements." *Annals of Eugenics* **8** 376—386.
- Fix, E., and Hodges, J.L. (1959). "Discriminatory analysis: Nonparametric Discrimination: Consistency properties." *Report No. 4, Project No. 21-49-004, School of Aviation Medicine, Randolph Air Force Base, TX*.
- Jackson, R.D., Slater, P.N., and Pinter, P.J. (1983). "Discrimination of growth and water stress in wheat by various vegetation indices through clear and turbid atmosphere." *Remote Sensing of the Environment* **15** 187—208.
- Jenkins, J.L., Powell, J.A., Logan, J.A., and Bentz, B. (2001). "Low seasonal temperatures promote life-cycle synchronization." *Bulletin of Mathematical Biology* **63** 573—595.
- Johnson, R.A. and Wichern, D.W. (2002). *Applied Multivariate Statistical Analysis* (5th Edition). Prentice-Hall Inc., Upper Saddle River, NJ.
- Kauth, R.J. and Thomas, G.S. (1976). "The Tasseled Cap—A graphic description of the spectral-temporal development of agricultural crops as seen by Landsat." *Proceedings of the Symposium on Machine Processing of Remotely Sensed Data*, Purdue University, West Lafayette, Indiana, 4B41-4B51.
- Logan, J.A. and Powell, J.A. (2001). "Ghost Forests, Global Warming, and the Mountain Pine Beetle." *American Entomologist* **47** 3 160—172.
- Logan, J., White, P., Bentz, B., and Powell, J. (1998). "Model analysis of the temporal evolution of spatial patterns in mountain pine beetle outbreaks." *Theoretical Population Biology* **53** 236—255.

- Logan, J.A., Regniere, J., and Powell, J.A. (2003). "Assessing the impacts of global warming on forest pest dynamics." *Frontiers in Ecology and the Environment* **1** 3 130—137.
- Negron, J. (1997). "Estimating probabilities of infestation and extent of damage by the roundheaded pine beetle in ponderosa pine in the Sacramento Mountains, New Mexico." *Canadian Journal of Forest Research* **27** 1936—1945.
- Powell, J.A., Logan, J.A., and Bentz, B.J. (1996). "Local projections for a global model of mountain pine beetle attacks." *Journal of Theoretical Biology* **179** 243—260.
- Powell, J. and Rose, J. (1997). "Local consequences of a global model for mountain pine beetle mass attack." *Dynamics and Stability of Systems* **12** 1 3—24.
- Powell, J., McMillen, T., and White, P. (1998). "Connecting a chemotactic model for mass attack to a rapid integro-difference emulation strategy." *SIAM Journal of Applied Mathematics* **59** 2 547—572.
- Powell, J., Tams, J., Bentz, B., and Logan, J. (1998). "Theoretical analysis of 'switching' in a localized model for mountain pine beetle mass attack." *Journal of Theoretical Biology* **194** 49—63.
- Powell, J., Kennedy, B., White, P., Bentz, B., Logan, J., and Roberts, D. (2000). "Mathematical analysis of attack risk analysis for mountain pine beetles." *Journal of Theoretical Biology* **204** 601—620.
- Powell, J., Jenkins, J., Logan, J., and Bentz, B. (2000). "Seasonal temperatures alone can synchronize life cycles." *Bulletin of Mathematical Biology* **62** 977—998.
- Steele, B.M. (2000). "Combining multiple classifiers: An application using spatial and remotely sensed information for land cover type mapping." *Remote Sensing and the Environment* **74** 545—556.
- Steele, B.M. and Patterson, D.A. (2000). "Ideal bootstrap estimation of expected prediction error for k -nearest neighbor classifiers: Applications for classification and error assessment." *Statistics and Computing* **10** 349—355.
- Steele, B.M. and Patterson, D.A. (2002). "Land cover mapping using combination and ensemble classifiers." *Proceedings of the 33rd Symposium on the Interface*, Interface Foundation of North America, Fairfax Station, VA.
- White, P. and Powell, J. (1998). "Phase transition from environmental to dynamic determinism in mountain pine beetle attack." *Bulletin of Mathematical Biology* **59** 609—643.
- Tucker, C.J. (1979). "Red and photographic infrared linear combinations for monitoring vegetation." *Remote Sensing of the Environment* **8** 127—150.
- Tucker, C.J., Newcomb, W.W., Los, S.O., and Prince, S.D. (1991). "Mean and inter-year variation of the growing-season normalized difference vegetation index for the Sahel 1981-1989." *International Journal of Remote Sensing* **12** 1113—1115.

White, P. and Powell, J. (1998). "Spatial invasion of pine beetles into lodgepole forests: a numerical approach." *SIAM Journal of Science and Computing* **20** 1
164—184.