Data mining and image segmentation approaches for classifying defoliation in aerial forest imagery

K. Fukuda[a] and P. A. Pearson[b]

[a] *Environmental Science Programme (Department of Mathematics and Statistics, Department of Computer Science and Software Engineering, and School of Forestry), University of Canterbury, Private Bag 4800, Christchurch, New Zealand*

[b] *Myelin Electronics, PO Box 20380, Bishopdale, Christchurch, New Zealand*

Abstract: Experimental data mining and image segmentation approaches are developed to add insight towards aerial image interpretation for defoliation survey procedures. A decision tree classifier generated with a data mining package, WEKA [Witten and Frank, 2005], based on the contents of a small number of training data points, identified from known classes, is used to predict the extents of regions containing different levels of tree mortality (severe, moderate, light and non attack) and land cover (vegetation and ground surface). This approach is applicable to low quality imagery without traditional image pre-processing (e.g., normalization or noise reduction). To generate the decision tree, the image is split into 20×20 pixel tiles and data points are created for each tile from peaks of smoothed histograms of red, green and blue colour channels, and their average. Colour channel peaks are examined to verify that histogram peaks effectively represent tree mortality, and to select an initial small training data set. Next, two small training data sets are selected randomly, to model the real-world training data selection process. Decision trees are generated using these training sets and tested on the remaining data. Stratified cross-validation is then performed on the full dataset, for comparison. The classification accuracy is 75% for cross validation and 31-49% for smaller training data sets. Assigning lower penalties for less severe errors gives a *weighted accuracy* of 79% for cross validation, 72% for manually selected and 48-65% for randomly selected training data. For comparison, the classification accuracy of the image segmentation method is 84%. Performance on small training sets still needs to be improved, although encouraging results were achievable with well identified heterogeneous training data.

Keywords: Classifications, Data mining, Defoliation, Image segmentation

1. INTRODUCTION

The increasing availability of remote sensing and geographic data helps monitoring and management for maintaining the health of forest ecosystems, which is important for the protection of natural resources and the economy. Satellite imagery, a remote sensing technique, is convenient for large-scale surveys, and has been used widely for land cover and habitat mapping using different applications [Friedl and Brodley, 1997; Kobler et al., 2006], but it has low resolution and it can be expensive to obtain timely imagery. Alternatively, aerial photography can provide higher resolution to allow monitoring of forest health and identification of tree species at an acceptable level of accuracy [Haara and Nevalainen, 2002]. White et al. [2005] investigated an automated interpretation method for detecting the *red attack* stage of trees attacked by the mountain pine beetle using satellite imagery, using aerial imagery for validation. However, not all studies are able to access such high quality data. In fact, environmental studies often deal with incomplete or poor quality data, as it is costly to obtain high quality data, and measurement relies on human observations that may be imprecise or uncertain. Hence, methods for processing poor quality data, perhaps involving statistics or knowledge discovery, can be advantageous.

The central interior region of British Columbia has suffered from increasing populations of mountain pine beetle (*Dendroctonus ponderosae*) since 1994 [White et al., 2005]. The British Columbia Ministry of Forests and Canadian Forest Service (BCMF and CFS) [2000] carries out annual defoliation surveys, where observers in small aircraft sketch infested regions on forest maps. Aerial surveying is said to be "not an exact science…as no matter what type of aircraft, the flying height, the weather, the survey map base, or the biological window, the survey is always going

to be less than perfect". The survey accuracy depends on the observers' knowledge of the local forest and pests. Usually only estimates of current tree mortality are indicated, but experienced personnel can estimate damage intensities fairly accurately with help of a multi-stage sampling procedure including aerial photography, GPS point and ground plot data, to ensure accuracy by enabling cross-validation.

The purpose of this study is to develop statistically and computationally driven methods using data mining and image segmentation, to add insight towards aerial imagery interpretation for the annual defoliation survey procedure. Data mining is used for knowledge discovery: the extents of infested regions and land cover are predicted using a decision tree classifier generated with a data mining package, WEKA [Witten and Frank, 2005], based on the contents of only a few known (training) data points that have been manually identified by an expert. Decision tree algorithms are considered suitable for remote sensing applications, since they are flexible and robust with respect to non-linear and noisy relations among input features and class labels, and prior assumptions regarding the *distribution* of input data are not required [Friedl and Brodley, 1997]. Typically, classification accuracy is tested in data mining projects by cross validation on as much data as possible, but this study takes a different approach. The classification tree is created from a small proportion of the data and tested on the rest of the data, to model the intended use of the system: for estimating tree mortality and land cover when complete ground truth is not available. The available image for this study has only low resolution (287 × 313 pixels), uneven lighting and varying scale, so the data mining approach is designed to be applicable to low quality imagery. It identifies patterns directly using the training data, thus traditional image preprocessing to normalize the image or remove noise is unnecessary. An image segmentation method is developed to compare with the data mining approach. This method uses manually-created pixel classification functions to detect attacked trees, then clusters pixels into regions, and estimates the tree mortality density in each region.

2. DEFOLIATION IMAGERY

Aerial imagery (Figure 1) was captured in Flathead Valley, Nelson Forest Region, in British Columbia, Canada, which suffers from mountain pine beetle attack. The studied aerial imagery is a low-resolution photo (287 × 313 pixels), downloaded from the source website*.

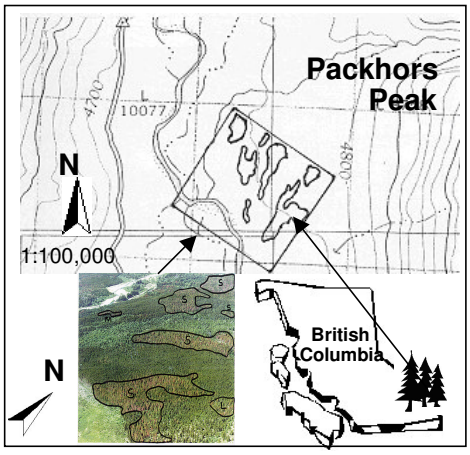

Figure 1. Location of the aerial image site, Flathead Valley, Nelson Forest Region, and original 70 mm photo [BCMF and CFS, 2000].

Over the mountain pine beetles' one-year life cycle, tree foliage becomes chlorotic, then yellow, and finally fades to red. The BCMF and CFS [2000] define three levels for tree mortality caused by defoliators and bark beetles: severe (S), moderate (M), and light (L). This study adds extra classes for land cover: vegetation (V), ground surface (Surface) and non attack (Non); all classes are shown in Table 1. Figure 1 shows one L, one M and five S regions, identified by BCMF and CFS.

Table 1. Tree mortality and land cover classification criterion.

Tree mortality Classification	Criterion	
	Bark beetles	Defoliators
Severe (S)	>30% of trees recently killed	Bare branch tips and completely defoliated tops. Most trees sustaining more than 50% total defoliation.
Moderate (M)	11-29% of trees recently killed	Pronounced discoloration. Noticeably thin foliage. Top third of many trees severely defoliated. Some completely stripped.
Light (L)	1-10% of trees recently killed	Discoloured foliage barely visible from the air. Some branch tip and upper crown defoliation.
Land cover classification	Criterion	
Vegetation (V)	Green regions that do not contain trees.	
Ground Surface (Surface)	Regions where the ground surface is exposed.	
Non attack (Non)	Regions that are not included in tree mortality classifications, assumed to be non-attack regions.	

3. DATA MINING APPROACH

To convert the image into a form suitable for analysis, it is divided into relatively large (20 × 20 pixel) tiles (Section 3.1). This tile size reduced noise in histograms and represented relevant region characteristics better than smaller (10 x 10 pixel) tiles. Next, training data points are created from the

* Aerial imagery used by permission of B.C. Ministry of Forests and Canadian Forest Service [2000] from: http://ilmbwww.gov.bc.ca/risc/pubs/teveg/foresthealth/assets/aerial-1.jpg

peak values of smoothed histograms of red (R), green (G) and blue (B) colour channels, and their average (A). The histograms are smoothed by Singular Spectrum Analysis (SSA) [Golyandina et al., 2001; Fukuda and Pearson, 2006], which was found to provide better results than a Fourier transform low-pass filter. The analysis is improved by adding the difference between each pair of colour peak values (e.g., R-G) to each training data point. Lastly, a univariate decision tree is generated via WEKA and tested using three different sets of training data points to predict the rest of the imagery, followed by stratified cross-validation on the entire image (Section 3.2). The predicted classes are then overlaid on the image, to provide visual feedback on the classification results.

3.1 Extraction of histograms

Let $L = \{(n, m), n = 1, \ldots, N, m = 1, \ldots, M\}$ be a 2D lattice of pixels for an image, *I,* where *n* and *m* represent columns and rows respectively. The image, *I,* is divided into $S = \{(n/p, m/p), 1 \le p \le n, 1 \le p \le m\}$ tiles of $p \times p$ pixels. Here, *I* is defined by N=313, M=287 with p=20 (20 × 20-pixel tiles) to give S = (15, 14), a total of 210 regions. Colour frequency histograms H_R, H_B, H_G and H_A are extracted from the four colour channels in each S_p tile. Now, SSA [Golyandina et al., 2001] is applied to smooth each histogram. Each *H* is treated as a 1D series of length Q=256, $H = (f_0, \ldots, f_{Q-1})$, and transferred into a set of *W-dimensional lagged vectors*, $X_i = (f_{i-1}, \ldots, f_{i+W-2})^T$, where $1 \le i \le K = Q - W + 1$ and *W* is the *window length* ($W \le Q/2$); for this analysis, W=32. This procedure turns the *H* series into the *W-trajectory matrix*, $X = [X_1:\ldots:X_K]$), which can be rewritten as

$$X = (x_{ij})_{i,j+1}^{W,K} = \begin{pmatrix} f_0 & f_1 & f_2 & \cdots & f_{K-1} \\ f_1 & f_2 & f_3 & \cdots & f_K \\ f_2 & f_3 & f_4 & \cdots & f_{K+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{W-1} & f_W & f_{W+1} & \cdots & f_{Q-1} \end{pmatrix} \quad (i+j = \text{const}). \qquad \textbf{(3.1)}$$

The obtained trajectory matrix (3.1) is decomposed by the singular value decomposition (SVD) to provide *i* eigentriples (consisting of eigenvalues, principal components and right singular vectors). The eigentriples are grouped and reconstructed to form the smoothed histograms for each tile in *S*. Here, the first three eigentriples were used, to provide >75% of the original variance.

Let *h* represent the smoothed histograms and h_c = (in turn) h_R, h_G, h_B, h_A for each colour channel. Maximum values of each h_c are then calculated for constructing training data points. The differences between each pair of values (e.g., $maxh_R - maxh_G$) are added to increase the number of attributes available for data mining.

3.2 WEKA classification

The J4.8 classifier from WEKA 3.4 [Witten and Frank, 2005], based on the C4.5 algorithm [Quinlan, 1993], is used to generate decision trees from a small number of training data points to predict tree mortality and land cover (class) for the rest of the image. Note that regions labelled as S are divided into five regions: S1 to S5, in decreasing order of size. Four experiments were performed, selecting training data with three different methods.

1) Manually selected training data: To test if the patterns of colour channel peaks effectively represent tree mortality, tiles were examined to find similar patterns of $maxh_c$ values (*colour patterns*) and verify the connection between these colour patterns and tree mortality/land cover classes, then up to four of the most representative tiles in each class were manually selected as training data points, to produce a decision tree which was tested on the remaining data. **2) Randomly selected training data:** To model the real-world training data selection process, first two, then three training data points were selected randomly from each class (S1 to S4, M, L, Surface, V and Non) to produce a decision tree, which was tested on the remaining data. **3) Stratified cross-validation:** To test the overall performance of the decision tree method, the entire dataset was tested using ten-fold stratified cross-validation, with S1-4 combined into a single S class.

Note that the S5 region is ignored as it only contains one tile, and L and M are also small, with four and two tiles respectively. The predicted class for each tile is overlaid on the imagery to allow visual interpretation of classification results (except for cross validation). To reduce the visual complexity of result images, the classes used internally are reduced to S, M, L, V, Surface and Non. Classification accuracy is presented as four numbers. *Overall accuracy* is the proportion of correct classifications when decision trees are tested on the entire dataset, including the training data used to create them. *Excluding training set* is the proportion of correct classifications when the training set is excluded from the test set. *Weighted* values weight different errors differently, giving a greater penalty for larger errors (e.g., Non misclassified as S) than for errors between adjacent classes (e.g., L misclassified as M).

4. IMAGE SEGMENTATION APPROACH

The image segmentation approach, in contrast to the single-pass tile-based data mining method, first attempts to detect whether individual pixels belong to attacked trees, then groups these *attack pixels*

into regions, and finally quantifies the severity of the attack in each region.

Let α represent the source image, such that $\alpha(x, y)$ represents the pixel at column x and row y of the image. Let $\alpha_H(x, y)$, $\alpha_S(x, y)$ and $\alpha_V(x, y)$ represent the hue, saturation and value attributes of the pixel $\alpha(x, y)$. Hues lie in the range [0°, 360°), while saturations and values lie in the range [0%, 100%].

4.1 Pixel classification

First, pixels are classified as to whether they are expected to correspond to attacked trees, using one of a number of manually-designed classifiers. The seven classifiers, with different hue, saturation and value criteria, are shown in Table 2.

Table 2. Pixel classification methods.

Method	Hue criterion	Saturation criterion	Value criterion
A	H < 24°	S > 20%	V > 50%
B	H < 54°	S > 10%	-
C	H < 54°	S > 20%	V > 50%
D	H < 24°	S > 20%	V > 50%
E	H < 24°	S > 20%	V > 39%
F	-	S < 10%	-
G	245° < H < 305°	S > 20%	V > 50%

Equation 4.1 shows how this step produces a *detection matrix*, D, for the A classifier:

$$D(x, y) = \begin{cases} 1 & \alpha_H(x, y) < 24^\circ \ \& \\ & \alpha_S(x, y) > 20\% \ \& \\ & \alpha_V(x, y) > 50\% \\ 0 & \text{otherwise} \end{cases} \quad \textbf{(4.1)}$$

4.2 Region detection and tree mortality quantification

Next, a new matrix, E, is created. Each cell in E contains the sum of all values within 10 cells of the corresponding value in D: $E(x,y) = \Sigma D(x',y')$, $\forall \sqrt{(x-x')^2+(y-y')^2} < 10$. A threshold, τ, is defined as 10% of the maximum value in E: $\tau = \max(E) / 10$. This threshold is then applied to E to produce R, which is equal to 1 where the corresponding element in E is greater than τ.

$$R(x, y) = \begin{cases} 1 & E(x, y) \geq \tau \\ 0 & E(x, y) < \tau \end{cases} \quad \textbf{(4.2)}$$

Now, a connected component analysis is performed to extract connected groups of nonzero elements in R, which represent possible infested regions. Let $C_T(n)$ represent the count of pixels in the n^{th} connected component, and $C_A(n)$ the count of detected *attack pixels* in the component (i.e., pixels for which $D(x, y) = 1$). Let $C_R(n)$ equal the proportion of attacked pixels in region n.

$$C_R(n) = \frac{C_A(n)}{C_T(n)}. \quad \textbf{(4.3)}$$

Regions are classified as severely attacked if $C_R(n) \geq 0.3$, moderately attacked if $C_R(n) \geq 0.2$, lightly attacked if $C_R(n) \geq 0.1$, and non-attack otherwise.

5. PATTERNS OF COLOUR CHANNEL HISTOGRAM PEAKS

Plotting colour channel histogram peaks for manually selected similar tiles (Figure 2) identified nine distinct colour patterns from the six classes (S, M, L, Non, V and Surface). The Non tiles contain three distinct patterns (Figure 2a, b, and g), visible in the image as yellow green, light green and grey. S tiles show two patterns: reddish yellow (S1 and S2, Figure 2e) and grey (S3 and S4, Figure 2f).

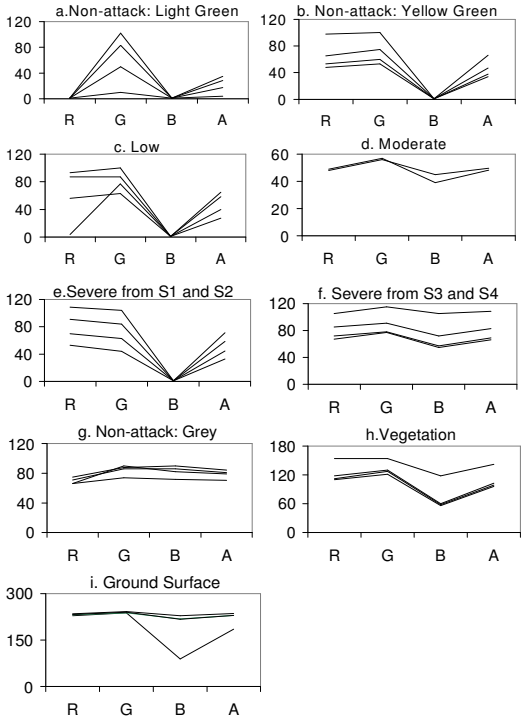

Figure 2. Colour channel histogram peaks, showing distinct patterns.

These results reflect the heterogeneous nature of Non and S regions. Interestingly, the use of only four values from each relatively large tile successfully describes the colour changes that take place over time during the process of tree *infestation* and *mortality*. The sequence may start from *light green* Non (Figure 2a), which has a high green peak, and red and blue peaks at zero. Next, red is added, as seen from *yellow green* Non (Figure 2b), which overlaps with L (Figure 2c). Then, blue is added as the class changes to M (Figure 2d). During the S stage, the blue peak drops to zero, and the red peak becomes higher than the green peak (Figure 2e). Red, green, blue, and average peaks are similar in S3/S4 (Figure 2f), as well as *grey* Non (Figure 2g), appearing flat when plotted, although S3/S4 tiles have a slightly

lower blue peak. These tiles only appear in the top quarter of the image, suggesting that the greyness may be due to light conditions and distance from the photographer, although another possibly is a high concentration of long-dead trees, which are known to appear grey, but marked Non because only recent attack is labelled [BCMF and CFS, 2000]. Structures of M (Figure 2d), V (Figure 2h), S3/S4 (Figure 2f) and Surface (Figure 2i) have similar patterns, but different ranges of peak intensities. There is an overlap between Surface and V, which may cause some misclassification. As S and Non tiles are heterogeneous due to poor lighting in the imagery, training data sets must contain samples from each different variant of S and Non for classification to be successful.

6. CLASSIFIED IMAGERY AND CONFUSION MATRICES

Classification figures of accuracies and confusion matrices are shown in Table 3, and images with overlaid prediction results are shown in Figure 3.

Table 3. Confusion matrices for test results.

Cross validation on full image

Detected	Actual value S	M	L	V	Surf	Non
S	**23**					7
M						
L						
V				**11**		
Surface					**4**	
Non	33	2	4	2		**154**

Training accuracy	80.0%
Weighted accuracy	82.2%
With cross-validation	**74.6%**
Weighted x-validation	**79.4%**

Colour pattern analysis

Detected	Actual value S	M	L	V	Surf	Non
S	**33**					20
M		**2**				6
L	15		**4**			49
V	1			**13**		9
Surface					**4**	
Non	7					**77**

Overall accuracy	55.4%
Weighted accuracy	75.7%
Excluding training set	**49.0%**
Weighted w/o train set	**72.0%**

Three training data points

Detected	Actual value S	M	L	V	Surf	Non
S	**35**		3			45
M	1	**2**				1
L	5		**3**			26
V				**6**		5
Surface				4	**4**	4
Non	15		1			**80**

Overall accuracy	54.2%
Weighted accuracy	68.2%
Excluding training set	**49.1%**
Weighted w/o train set	**64.8%**

Two training data points

Detected	Actual value S	M	L	V	Surf	Non
S	**38**					82
M		**2**				11
L	6		**2**			27
V				**9**		5
Surface				4	**4**	4
Non	12		2			**32**

Overall accuracy	36.3%
Weighted accuracy	52.1%
Excluding training set	**31.1%**
Weighted w/o train set	**48.2%**

The classification accuracy on the test set is 75% for cross validation and 31-49% when using smaller training data sets. Weighted accuracy figures are 79% for cross validation, 72% for manually selected training data and 52-68% for randomly selected training data. The best data mining results were obtained using the full image data, with all classes classified well except S (41% recall). The next best results were from the colour channel pattern analysis (Figure 3A), which detected V and Surface correctly (100%), although S (59%) was often misclassified as L or Non, and Non (48%) was often misclassified as L or S. This could be due to the similarity between S1/S2, L, and *yellow green* Non classes (the classifier selects the *middle ground* class L between the extremes of S and Non), or perhaps the gaps between large S1/S2 regions are bridged by L, as also seen from one space between an M and an S3/S4 region classified as M, but further investigation is required. As training tiles are added from the top of the image (especially *grey* Non, Figure 2g, and S3/S4, Figure 2f), misclassification of Non and S in that region is reduced. For the same reason, V and Surface tiles had 100% recall in this test.

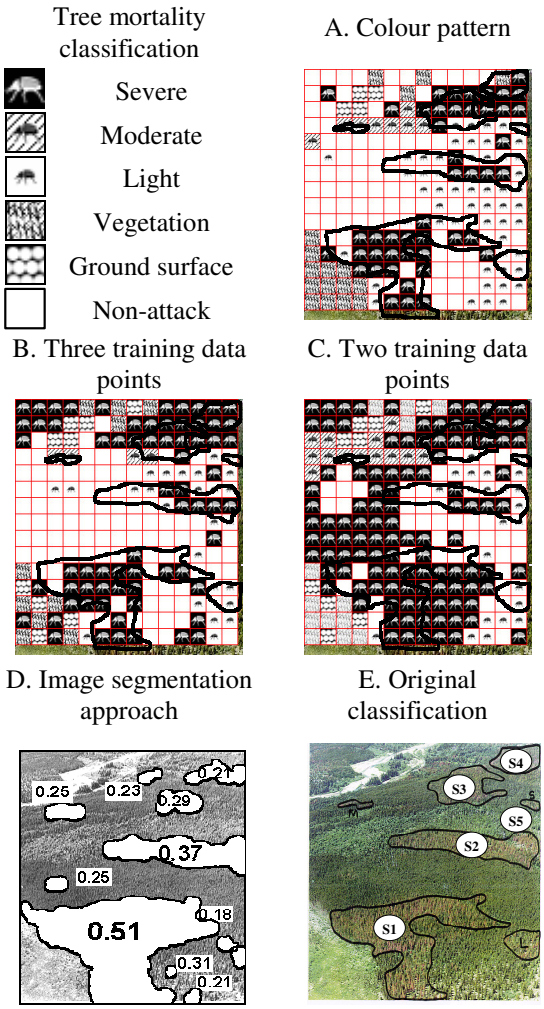

Figure 3. Prediction result overlay images. Manually classified regions of tree mortality from E appear as thick outlines in A-C.

The test using two training data points (Figure 3C) classified Surface (100%) and V (69%) tiles satisfactorily, though S (68%) tiles were often misclassified as Non and most Non tiles were misclassified. Classification improved when three training data points were used (Figure 3B), with much better separation of S and Non regions (Non recall improved from 20% to 50%). Both results classified both top corners, which appear grey in the original image, as S, and the three-point analysis confused several S3/S4, V and Surface tiles. This may be due to the similarities in their

colour patterns, as previously discussed. Performance on L and M tiles was poor, with only training data classified correctly.

Overall, all tests had significant misclassification between S and Non due to the heterogeneity of Non regions and aerial photography conditions, as previously discussed. Performance on small training sets still needs to be improved, although encouraging results were achievable with well identified heterogeneous training data.

Figure 4 shows the best decision tree, for colour pattern analysis, with 9 leaves and a size of 17. For classifying the aerial imagery, the blue histogram peak appears most important, followed by red and green. Grey (A) peaks were not found to be important alone, but R-A (red peak minus grey peak) was used to distinguish between Non, M and S.

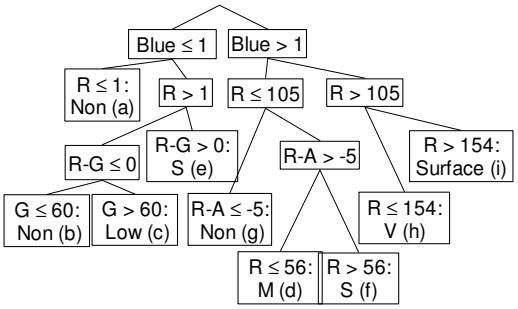

Figure 4. Decision tree for colour pattern analysis.

The image segmentation approach (Figure 3) detected regions marked by the BCMF and CFS [2000] as infested (S, M, L classes) with 84% accuracy, and identified three regions that had not been flagged by the human observer but appear infested on the image. The only serious misclassification occurred in the top right-hand corner of the image, where two large S regions were classified as M and only partly located.

7. CONCLUSIONS

The cross-validation test and image segmentation method provided the best classification accuracy. Results with small training data sets need to be improved, although the rate at which classification accuracy improves with the addition of well identified heterogeneous training data is encouraging for further investigation.

In future, more image features will be considered, and techniques such as hybrid decision trees, which Friedl and Brodley [1997] found to provide higher classification accuracy, will be investigated. Higher quality input data (e.g., higher resolution, overhead angle, even lighting) will improve results. The (generic) data mining approach can be applied to other image classification problems, and will be used to produce better initial pixel classification rules for the image segmentation method, while image analysis techniques will be used to provide richer data points for the data mining approach.

8. ACKNOWLEDGEMENTS

The authors thank two anonymous reviewers for their detailed comments. K. Fukuda acknowledges the University of Canterbury for providing a travel grant for iEMSs 2006 attendance.

9. REFERENCES

British Columbia Ministry of Forests and Canadian Forest Service (BCMF and CFS), Forest Health Aerial Overview Survey Standards for British Columbia, The B.C. Ministry of Forests adaptation of the Canadian Forest Service's FHN Report 97-1 "Overview Aerial Survey Standards for British Columbia and the Yukon", 23-24, 2000.

Friedl, M.A. and Brodley, C.E., Decision Tree Classification of Land Cover from Remotely Sensed Data, *Remote Sensing of Environment*, 61, 399-409, 1997.

Fukuda, K., and Pearson, P.A., Investigation of SSA and Machine Learning for Road Sign Location. In *Extended abstracts*, *7th Intl. Assoc. for Pattern Recognition workshop on Document Analysis Systems* (DAS 2006), Nelson, NZ, February 13-15, 29-32, 2006.

Golyandina, N., Nekrutkin, V., and Zhigljavsky, A., Analysis of Time Series Structure: SSA and Related Techniques, Chapman & Hall/CRC, Boca Raton, 2001.

Kobler, A., Džeroski, S., and Keramitsoglou, I., Habitat mapping using machine learning-extended kernel-based reclassification of an Ikonos satellite image, *Ecological Modelling*, 191, 83-95, 2006.

Haara, A., and Nevalanine, S., Detection of dead or defoliated spruces using digital aerial data, *Forest Ecology and Management*, 160, 97-107, 2002.

Quinlan, J.R., C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, 1993.

Witten, I.H. and Frank, E., *Data mining, practical machine learning tools and techniques*, 2nd ed., Morgan Kaufmann Publishers, San Francisco, 2005.

White, J.C., Wulder, M.A., Brooks, D., Reich, R., and Wheate, R.D., Detection of red attack stage mountain pine beetle infestation with high spatial resolution satellite imagery, *Remote Sensing of Environment*, 96, 340-351, 2005.