Journal of
Applied Remote Sensing

Exploiting machine learning algorithms for tree species classification in a semiarid woodland using RapidEye image

Samuel Adelabu
Onisimo Mutanga
Elhadi Adam
Moses Azong Cho

SPIE

Downloaded From: http://remotesensing.spiedigitallibrary.org/ on 02/10/2014 Terms of Use: http://spiedl.org/terms

Exploiting machine learning algorithms for tree species classification in a semiarid woodland using RapidEye image

Samuel Adelabu,[a] Onisimo Mutanga,[a] Elhadi Adam,[a] and Moses Azong Cho[b]

[a]University of KwaZulu-Natal, School of Agricultural, Earth & Environmental Sciences, Geography Department, P/Bag X01, Scottsville, Pietermaritzburg, 3209, South Africa
oyesams@gmail.com

[b]Natural Resources and Environment Council for Scientific and Industrial Research, P.O. Box 395, Pretoria 0001, South Africa

Abstract. Classification of different tree species in semiarid areas can be challenging as a result of the change in leaf structure and orientation due to soil moisture constraints. Tree species mapping is, however, a key parameter for forest management in semiarid environments. In this study, we examined the suitability of 5-band RapidEye satellite data for the classification of five tree species in mopane woodland of Botswana using machine leaning algorithms with limited training samples.We performed classification using random forest (RF) and support vector machines (SVM) based on EnMap box. The overall accuracies for classifying the five tree species was 88.75 and 85% for both SVM and RF, respectively. We also demonstrated that the new red-edge band in the RapidEye sensor has the potential for classifying tree species in semiarid environments when integrated with other standard bands. Similarly, we observed that where there are limited training samples, SVM is preferred over RF. Finally, we demonstrated that the two accuracy measures of quantity and allocation disagreement are simpler and more helpful for the vast majority of remote sensing classification process than the kappa coefficient. Overall, high species classification can be achieved using strategically located RapidEye bands integrated with advanced processing algorithms. © *2013 Society of Photo-Optical Instrumentation Engineers (SPIE)* [DOI: 10.1117/1.JRS.7.073480]

Keywords: random forest; support vector machines; tree species classification; semiarid environment; red-edge.

Paper 13239 received Jul. 4, 2013; revised manuscript received Oct. 17, 2013; accepted for publication Oct. 21, 2013; published online Nov. 19, 2013.

1 Introduction

In Southern Africa, natural forest and rangeland resources form an important resource base for food and medicinal products that form part of people's subsistence as well as their economic base and well-being. One such natural forest is the mopane woodland. Mopane trees provide varied products that include construction and fence poles; wood for tools, carvings, and utensils; firewood, rope, gum, tannin, medicines and resin, green manure, livestock browse, and edible caterpillars (commonly referred to as mopane worms).[1] The value of mopane woodland in Botswana alone has been estimated at US $3.3 million per annum, of which ~40% goes to producers who are primarily poor rural women.[2] Recent studies have, however, shown that the long-term sustainability of the woodland and its resources is under threat.[3,4] Mapping tree species in mopane woodland are extremely important for forest management purposes. Up until now, there are no category-specific maps of species distribution in mopane woodland. Moreover, forest managers need to understand the species diversity to suggest possible management practices that will enable efficient and sustainable use of the resources emanating from mopane woodland. This will further support scientific knowledge of environmental

0091-3286/2013/$25.00 © 2013 SPIE

Downloaded From: http://remotesensing.spiedigitallibrary.org/ on 02/10/2014 Terms of Use: http://spiedl.org/terms

management practices in Africa and other sites in the world more sensitive to global changes. However, it is nearly impossible to acquire detailed tree species information over large areas purely on the basis of field assessments. Therefore, enhanced methods are required to get explicit information on the tree species composition and distribution patterns.

Remote sensing has been a valuable source of information over the course of the past few decades in mapping and monitoring forests.[5] It provides a cost-effective tool to help forest managers better understand forest characteristics, such as forest area, locations, and species, even down to the level of characterizing individual trees. The application of remote sensing in forest management began with the manual interpretation of aerial photographs, but is increasingly reliant on new data and methods.[6] Over the last 20 years, the spectral and spatial resolution of satellite data has steadily increased. Medium-resolution satellite data such as Landsat and SPOT can obtain regional-scale forest variables.[7] Airborne hyperspectral sensors meet the requirements regarding spectral and spatial resolution for tree species mapping. However, due to the high costs and their limited availability, hyperspectral data have gained limited acceptance for operational use. As higher-resolution satellite imageries (e.g., RapidEye) become more available, there is an increasing potential to provide more detailed information. Unlike medium-resolution satellite imagery, which provides an aggregated response over a region, individual trees are visible in high-spatial resolution imagery. This provides opportunities to differentiate species and individual trees.

RapidEye has a relatively higher spatial resolution (5 m) and a new spectral band (red-edge) in addition to the four standard bands [blue, green, red, and near-infrared (NIR)]. The red-edge band is supposed to discriminate between healthy trees and trees that are affected by disease and to enhance the separation between different species and age classes. Recent studies have shown that with the inclusion of the red-edge channel, RapidEye has large capabilities for enhanced species mapping;[8,9] however, its applicability in species classification, especially in areas of limited training sites, is still evolving.

Many classification methods have been used for tree species mapping using remote sensing data. This includes maximum likelihood, minimum distance, discriminant analysis, and spectral angle mapper classifiers.[10–12] Maximum likelihood and minimum distance classifiers are commonly used supervised classification methods with conventional multispectral data. However, there is a limitation with the application of these classifiers for mapping areas with limited training samples.[11] Previous studies in semiarid environments have shown that collecting training samples is a mammoth task because the terrain and maps produced by vegetation canopy introduces noise to vegetation classification.[13] As a solution to this problem, powerful classification methods are essential for mapping species in semiarid environments.

We assessed two commonly used machine learning classification algorithms, namely, random forest (RF) and support vector machines (SVM). Previous studies have shown that the RF or SVM algorithm can be successfully used for species mapping and classification purposes.[14–17] To the best of our knowledge, so far no study has compared the main machine learning algorithms RF and SVM for tree species mapping in a semiarid environment and in particular, using the new-generation satellite sensors.

This study aims at separating *Colophospermum mopane* (CM) and its coexisting species in a semiarid forest using spectral information provided by the RapidEye sensor applying RF and SVM classification algorithms. We focus on the following research questions: (1) Can CM and its coexisting species be separated using the RapidEye image with strategically positioned spectral bands? (2) Do the additional red-edge band of RapidEye improve the classification accuracy (CA) significantly compared to the four standard bands? (3) Which of the classification algorithms (RF or SVM) is better for detailed tree species classification in areas with limited training samples based on CAs?

2 Study Area

The study was undertaken in and around the Palapye/Tswapong axis of Central Region of Botswana (longitude 27°00' E to 27°33' and latitudes 22°23' to 22°52' S). It is situated along the main north-south highway and is ~230 km north of Gaborone (Fig. 1), the capital

Downloaded From: http://remotesensing.spiedigitallibrary.org/ on 02/10/2014 Terms of Use: http://spiedl.org/terms

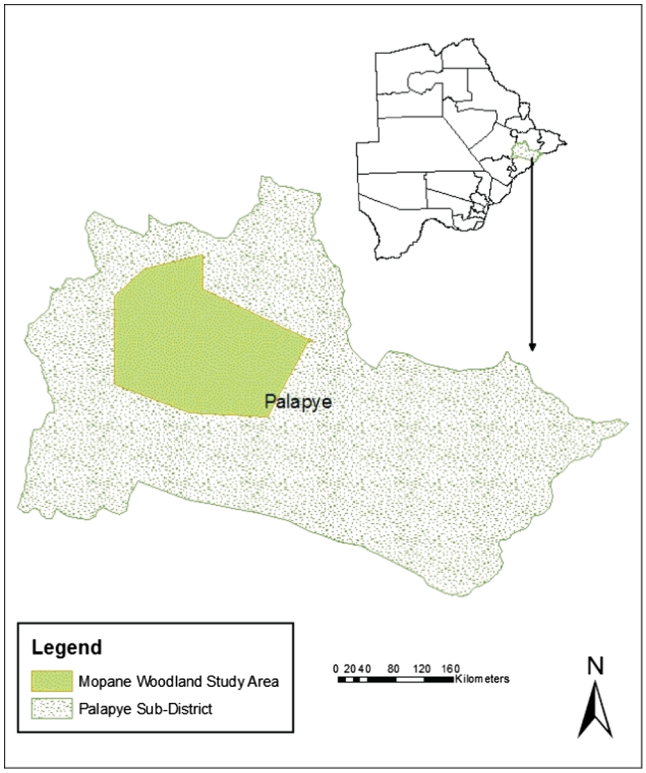

Fig. 1 Location of the study area in Central District of Botswana.

city of Botswana.[2] Soils around the area are well developed and variable, hence locally referred to as hardveld region. Temperature and rainfall regimes are highly variable both temporally and spatially, and are both characterized by seasonality in their occurrence.[18] The axis covered by this study is described as mopane bushveld because mopane tree is found in all its growth forms and is locally monospecific. However, several other tree species grow in association with mopane trees. Only the dominant tree species as identified by local ecologists have been used in this study (Table 1): *Grewia bicolar* (GB); *Dichrostachys cinerea* (DC); *Acacia erubiscens* (AE); *Acacia tortilis* (AT). GB and DC are deciduous species with short leafs and several flowering periods, while AT and AE have a dual flowering period during the rainy and dry seasons.[19] On the other hand, CM is a deciduous slow-growing species, with an erect narrow crown. The leaves are pinnate with two large leaflets that can vary considerably in size on the same tree[20] and within a growing season.[2] *C. mopane* drops its leaves in an irregular fashion from the onset of the dry season and is generally leafless from August to October. However, trees may retain their leaves between successive rainy seasons, depending on the amount and distribution of rainfall.[20]

Table 1 The number of sample plots, local names, and the type code for *Colophospermum mopane* and its coexisting species.

Species name	Local name	Type code	No. of plots
Grewia bicolar	Mogwana	GB	53
Dichrostachys cinerea	Moselesele	DC	51
Acacia erubiscens	Moloto	AE	53
Acacia tortilis	Mosu	AT	54
Colophospermum mopane	Mopane	CM	55

Downloaded From: http://remotesensing.spiedigitallibrary.org/ on 02/10/2014 Terms of Use: http://spiedl.org/terms

3 Field Data Acquisition

The field data acquisition was conducted during the summer month of January 2012 coinciding with the image acquisition date. Field measurements were done following random points that were generated from an existing land cover map of the study area using Hawth's analysis tool in ArcGIS 9.3.[21] A handheld Garmin eTrex30 GPS was then used to navigate to the respective points. Thereafter, vegetation polygons in which one or more of these species (GB, DC, AE, AT) coexist with CM was created around the centered point. The vegetation polygon was defined as covering 20 m × 20 m, where the target species were more homogenous and were representative of >80% of the target species in each plot. This method resulted in 51 to 55 vegetation polygons for each target species. Thereafter, the vegetation polygons were used to create training areas and then overlaid on the true color composite RapidEye image using Environment for Visualizing Images (ENVI) software.[22] The metadata, such as the site description (coordinates, altitude, and land cover class) and general weather conditions, were also recorded.

4 Image Acquisition and Data Preparation

High resolution multispectral imagery was acquired over the study area by the RapidEye sensor on January 25, 2012. The RapidEye dataset is composed of 5-band multispectral imagery with 5 m ground sampling distance. The five bands include blue (440 to 510 nm), green (520 to 590 nm), red (630 to 685 nm), red-edge (690 to 730 nm), and NIR (760 to 850 nm). For the purpose of machine learning optimization, spectra pixels from RapidEye image (5 m × 5 m) were extracted using ENVI software.[22] In this study, only pixels that fell completely within the measured polygons were used in the reference dataset so as to minimize variability and exclude mixed pixel effects of tree species.[23] Before field data were used for analysis, data for each polygon were then averaged to represent one sample.

The imagery over the study area contained 0% cloud cover, with a relatively clear atmosphere. All RapidEye's products are collected by a 12-bit imager. During on-ground processing, radiometric corrections are applied and all image data are scaled up to 16-bit dynamic range.[24] The scaling is done with a constant factor that converts the (relative) pixel digital number (DNs) from the sensor into values directly related to absolute radiances. The scaling factor was originally determined prelaunch. Previous experimentation performed by Naughton et al.[25] verified that the image registration was within a single pixel; hence further geometric processing was not applied. The image was atmospherically corrected using the quick atmospheric correction procedure in ENVI 4.7.[22]

5 Machine Learning Classifiers

5.1 *Random Forest Algorithm and Support Vector Machine*

RF is a machine learning algorithm that employs a bagging (bootstrap aggregation) operation where a number of trees (*ntree*) are constructed based on a random subset of samples derived from the training data. Each tree is independently grown to maximum size based on a bootstrap sample from the training dataset without any pruning, and each node is split using the best among a subset of input variables (*mtry*).[26] The multiple classification trees then vote by plurality on the correct classification.[26,27] The ensemble classifies the data that are not in the trees as out-of-bag (OOB) data, and by averaging the OOB error rates from all trees, the RF algorithm gives an error rate called the OOB classification error for each input variable.[26] Therefore, as part of the classification process, the RF algorithm produces a measure of importance of each input variable by comparing how much the OOB error increases when a variable is removed, while all others are left unchanged.[28] RF algorithm is easy to implement as only two parameters (*ntree* and *mtry*) need to be optimized.[26,29] A more detailed discussion on RF can be found in Ref. 26.

SVM, on the other hand, was originally introduced as a binary classifier[30] and is extensively described by Burges[31] and Hsu et al.[32] However, typical remote sensing problems usually involve identification of multiple classes (more than two). Adjustments are made to the simple

Downloaded From: http://remotesensing.spiedigitallibrary.org/ on 02/10/2014 Terms of Use: http://spiedl.org/terms

SVM binary classifier to operate as a multiclass classifier using methods such as one-against-all, one-against-others, and directed acyclic graph.[17] In its classical implementation, it uses two classes (e.g., presence/absence) of training samples within a multidimensional feature space to fit an optimal separating hyperplane (in each dimension, vector component is image gray-level). In this way, SVM tries to maximize the margin, that is, the distance between the closest training samples, or support vectors, and the hyperplane itself. SVM consists of projecting vectors into a high-dimensional feature space by means of a kernel trick and then fitting the optimal hyperplane that separates classes using an optimization function. Several kernels are used in the literature. According to Hsu et al.[32] and supported by many other authors,[16,17,30] the Gaussian radial basis function has both advantages of (1) being very successful since it works in an infinite-dimensional feature space and (2) having a single parameter contrary to the other well working kernels (e.g., polynomial). Noise in the data can be accounted for by defining a distance tolerating the data scattering, thus relaxing the decision constraint.

To demonstrate the effectiveness of RF and SVM for mapping CM and its coexisting species, classifications of the five species were trained on 70% occurrences and evaluated on the remaining 30%. The same number of pixels was used for the presence class and for the absence class in order to avoid discrepancies due to unbalanced training sets.[30,32] The EnMap box was used to carry out the RF and SVM analysis. EnMap is an interactive data language (IDL)-based tool for classification and regression analysis of remote sensing imagery. It can be fully integrated into the commercially available IDL/ENVI environment. It may also be run as an add-on freely available, license-free, and platform-independent processing environment for remote sensing imagery. EnMap uses generic image file formats with an ENVI-type text header for the image data as well as the continuous or categorical reference data and model outputs, a two-step image analysis consisting of separate model parameterization and application, and trained models are saved and may be applied several times, e.g., for transfer to other data sets and calculate variable importance and accuracy.

5.2 *Optimization*

Classifier parameters, namely, the number of trees and the number of variables to be randomly selected from the available set of variables for the RF and the regularization parameter C and the kernel parameter λ for the SVM, were selected.[32] The goal was to identify optimal parameters so that the classifier could accurately predict unknown data. This method is easy to use, quite fast, and can be more reliable than more advanced iterative techniques that do not always consider parameters as independent.

5.3 *Classification and Accuracy Assessment*

It has become customary in the remote sensing literature to report the kappa index of agreement for purposes of accuracy assessment since kappa also compares two maps that show a set of categories. However, recent studies have shown limitations of kappa in that it gives information that is redundant or misleading for practical decision making. Furthermore, Pontius Jr. and Millones[33] stated that the standard kappa and its variants are frequently complicated to compute, difficult to understand, and unhelpful to interpret. To solve this problem, Pontius Jr. and Millones[33] recommend that the use of kappa for the purposes of accuracy assessment and map comparison be abandoned, and instead summarize the cross-tabulation matrix with two much simpler summary parameters: quantity disagreement and allocation disagreement.

In this study, accuracy assessments were obtained using the kappa analysis and the quantity disagreement and allocation disagreement as proposed by Pontius Jr. and Millones.[33] A confusion matrix was constructed so as to compare the true class with the class assigned by the classifier and to calculate the overall accuracy (OA) as well as the CA.[34] We report on both statistics using the confusion matrix proposed by Pontius Jr. and Millones,[33] which is available online at http://www.clarku.edu/~rpontius/.

The two machine learning algorithms were compared on the basis of the accuracy they produce when trained on the same set of training data. Based on the accuracy metrics obtained for each classifier in each accuracy assessment method, a statistical analysis can be performed

Downloaded From: http://remotesensing.spiedigitallibrary.org/ on 02/10/2014 Terms of Use: http://spiedl.org/terms

to test if the difference is significantly equal or different. We compared the confusion matrix yielded by the RF and the SVM classifications by using the McNemar test with a 95% confidence interval. McNemar test is a nonparametric test that is simple to understand and execute. It has been shown to be sensitive and precise in comparing two or more accuracy assessments.[35,36] The test is based on chi square (z^2) statistics, computed from two error matrices, given as $(z^2) = (f_{12} - f_{21})^2/(f_{12} + f_{21})$,where f_{12} denotes the number of cases that are wrongly classified by classifier 1 but correctly classified by classifier 2 and f_{21} denotes the number of cases that are correctly classified by classifier 1 and wrongly classified by classifier 2.[36] The difference in accuracy between a pair of classifications is viewed as being statistically significant at a confidence of 95% if the calculated z score in McNemar test is >1.96.[36]

6 Results

6.1 *Optimization*

The results of optimizing RF parameters (*ntree* and *mtry*) at each split has shown that the setting of *mtry* value of 2 yielded the lowest OOB error rates. When examining the *ntree* parameter, results indicated that there were relatively stable OOB error rates after setting *ntree* value of 5000. Similarly, for the SVM optimization parameters C and λ, the best value was obtained at 0.01 and 10 for C and λ, respectively. The results show that the changes in *ntree* and *mtry* for RF and λ and C for SVM influence the accuracy of the classification results produced. Therefore, the settings of *mtry* (2), *ntree* (5000), C (0.01), and λ (10) were used for all further analyses.

6.2 *Classification Result*

The vegetation map of RF and SVM classifications are presented in Fig. 2. From the maps, it is very clear that using both methods of classification (RF and SVM), mapping of CM species and its coexisting species yielded good results from RapidEye images. The SVM classification produced higher overall accuracies of 88.75% compared to RF classification of 85%.

Generally, classification results were best in depicting CM and AE using both methods (Fig. 3). In fact, stands of CM were accurately classified with 100% class accuracy using both methods. Moreover, we tested the importance of the red-edge band as part of the RapidEye sensor for species classification in a semiarid environment by running the classification with and without the red-edge band. Results show that in both methods, when the red-edge band is excluded, the accuracies decreased by 8%. Furthermore, the relative importance of each of the RapidEye's bands ($n = 5$) in mapping the five species was measured using RF algorithm.

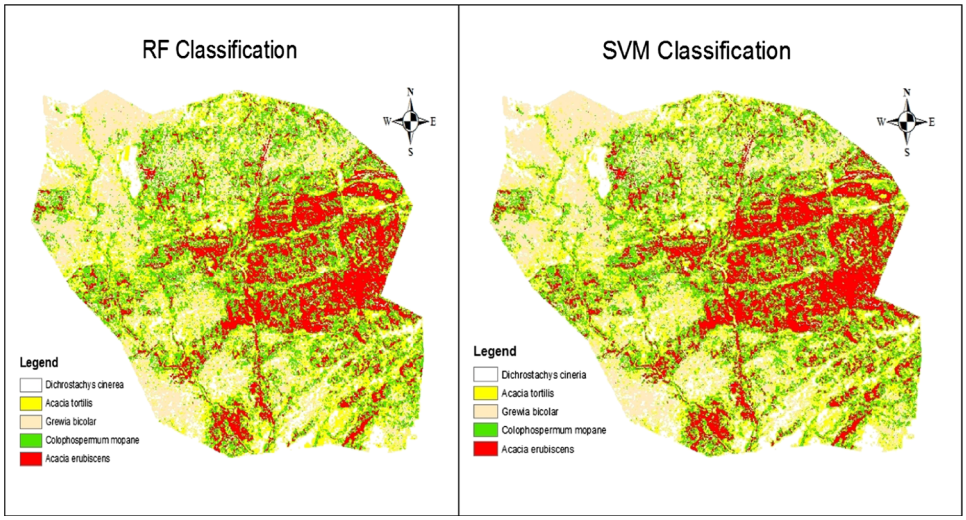

Fig. 2 Map showing the classification of the five tree species using random forest (RF) and support vector machine (SVM) learning algorithm.

Downloaded From: http://remotesensing.spiedigitallibrary.org/ on 02/10/2014 Terms of Use: http://spiedl.org/terms

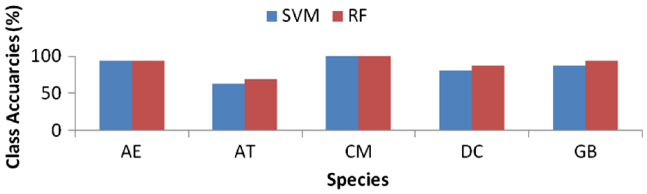

Fig. 3 Class accuracies for each species using the RF and SVM classification algorithm.

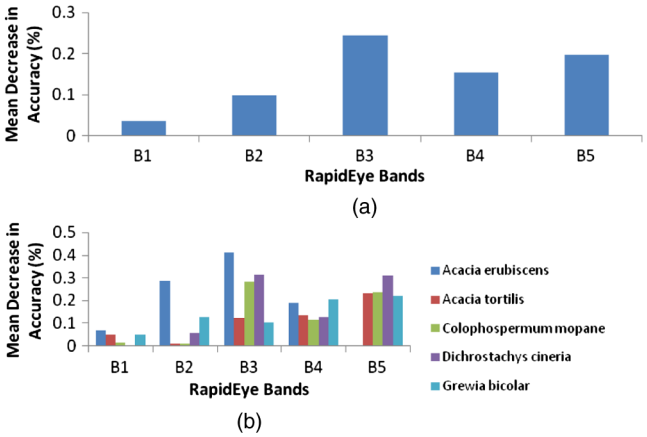

Fig. 4 (a) Comparing the relative importance of each RapidEye band in mapping the five species using the mean decrease in accuracy. The mean decrease in accuracy was estimated using the RF algorithm. (b) Comparing the relative importance of each RapidEye band in mapping individual species relative to other species using the mean decrease in accuracy. The mean decrease in accuracy was estimated using the RF algorithm.

The result further indicated that band 3 (red band; 630 to 690 nm) had the highest mean decreasing accuracy, with only the lowest mean decreasing accuracy obtained with band 1 (blue; 440 to 510 nm) [Fig. 4(a)].

Similarly, we also measured the importance of each band in discriminating individual species among the other species [Fig. 4(b)]. Results indicate that band 3 (red band; 630 to 690 nm) and band 5 (NIR; 760 to 880 nm) are more important in separating individual species from the other.

6.3 *Classification Accuracy and Assessment*

The confusion matrix derived from the test dataset for both RF and SVM are presented in Table 2. The matrix consists of CA and OA. When test data was used to test the RF classification result, the RF successfully mapped the five species (GB, DC, AE, AT, CM) with an OA of 85%. On the other hand, when the SVM was used for classification, the confusion matrix yielded an OA of 88.75%. Similarly, the accuracy assessment results for class pair indicates that AE and GB appear to be unique among the other species based on the highest CA of 93.75% (AE) and 93.75 and 100% (GB) for the SVM and RF classifications, respectively. Hence, it is easier to discriminate CM from AE and GB compared with DC, which yielded a class pair accuracy of 82.35 and 76.47% for both the SVM and RF classifications, respectively. Overall class accuracy of CM was 84.21% for SVM and 72.73% for RF.

The results of the performance of the two classification assessments (kappa and total disagreement) in mapping the five species are presented in Table 3 for both the SVM and RF classification methods. The total disagreement is separated into two components: quantity disagreement and allocation disagreement. Quantity disagreement is the amount of pixels of a class in the training data that is different from the quantity of pixels of the same class in the test data, while allocation disagreement is the location of a class of pixel in the training data that is different from the location of the same class in the test data. Kappa for both

Downloaded From: http://remotesensing.spiedigitallibrary.org/ on 02/10/2014 Terms of Use: http://spiedl.org/terms

Table 2 Comparison of confusion matrix obtained after the classification of *Colophospermum mopane* and its coexisting species from both the support vector machine SVM and random forest (RF). The confusion matrix includes overall accuracy (OA) and classification accuracy (CA).

	SVM					RF				
	AE	AT	CM	DC	GB	AE	AT	CM	DC	GB
AE	15	1	0	0	0	15	0	1	0	0
AT	0	11	1	3	1	0	10	3	3	0
CM	0	0	16	0	0	0	0	16	0	0
DC	1	0	1	14	0	1	0	2	13	0
GB	0	0	1	0	15	0	0	1	1	14
CA (%)	93.75	91.67	84.21	82.35	93.75	93.75	90.91	72.73	76.47	100
	OA = 88.75%					**OA = 85%**				

Table 3 Comparing kappa and total disagreement methods of classification assessments.

Parameters	SVM	RF
Kappa	0.86	0.81
Kappa total disagreement (%)	14	19
Allocation disagreement (%)	5	9
Quantity disagreement (%)	6	6
Total disagreement (%)	11	15

Note: All calculations were done using the confusion matrix proposed by Pontius Jr. and Millones[33] and sourced from http://www.clarku.edu/~rpontius/.

SVM (0.86) and RF (0.81) in this study represents a strong agreement between the training data and the test data (Table 3). Table 3 also shows that in both methods the total disagreement as proposed by Pontius Jr. and Millones[33] (11 and 15% for SVM and RF, respectively) is slightly lower than the disagreement from kappa (14 and 19% for SVM and RF, respectively).

We also used the confusion matrix to compare the accuracies of the two algorithm (RF and SVM) employed in this study using the McNemar. Table 4 shows that there are no significant differences in the accuracies obtained from both the RF and SVM classifications. The McNemar test z score was <1.96, which is required for the two algorithms to be statistically significantly different at 95% confidence level.

Table 4 Comparison of SVM and RF using McNemar test.

		SVM		
		Correctly classified	Misclassified	Total
RF	Correctly classified	67	4	71
	Misclassified	1	8	9
	Total	68	12	80

Note: McNemar z score = 1.342.

Downloaded From: http://remotesensing.spiedigitallibrary.org/ on 02/10/2014 Terms of Use: http://spiedl.org/terms

7 Discussion

This study highlights the effect of new generational multispectral image (RapidEye) in mapping tree species in semiarid environments where there are limited training samples. We also set out to compare two machine learning algorithms and two CA assessment methods for mapping tree species in semiarid environment.

7.1 *Spectral Classification of the Tree Species*

Classification of different tree species in semiarid areas, such as Palapye in Botswana, can be challenging because of the change in leaf structure and orientation as a result of limited soil moisture.[2,21] However, this study has shown that RapidEye satellite data are highly suitable for classifying tree species in mopane woodland. The achieved class accuracies of the various tree species ranged between 72.73 and 100% for RF and 82.35 and 100% for SVM. In both algorithms, the lowest class values were found for CM (72.73 and 82.35% for RF and SVM, respectively). The higher observed misclassifications of CM may be due to spectral overlaps between CM and the other tree species. Similarly, from our observations in the field, a great deal of the identified misclassifications could be explained by the very complex forest structure in the test area. Most of the leaves of the tree species investigated in this study look similar. This characteristic may have led to mixed pixels that may cause misclassifications.

In general, the present studies show an improvement to the works of Sebego et al.,[2] which was able to map mopane and its coexisting species with 74% accuracy. There are two reasons that may be responsible for this. First, the spatial resolution of the Landsat TM image used for their study is coarse (30 m) compared with that of the RapidEye (5 m) used in the present study. Second, the classification algorithms (RF and SVM) used in the present study has been proven to outperform the one used by the previous study (maximum likelihood).[37] We therefore conclude that spatial resolution and algorithms play an important role in classifying tree species in semiarid environments.

7.2 *Role of RapidEye Red-Edge Band in Tree Species Classification*

Although the five tree species could be separated accurately by using only the four standard bands (blue, green, red, and NIR), the use of the additional red-edge band led to a significant improvement of CA. For instance, the CAs increased from 78 and 80.25% to 85 and 88.75% for RF and SVM, respectively, when the red-edge band was added. The positive effect of the red-edge band can be explained by its relationship with the chlorophyll content of vegetation.[38] Several studies have reported the red-edge spectrum to be specifically sensitive to vegetation chlorophyll content.[38,39] Chlorophyll content can be regarded as an additional factor to explain particular sensitivities to the red-edge channel. The chlorophyll in green leaves absorbs light for photosynthesis in the red-edge region of the spectrum.[40] For this reason, the red-edge region is most useful for detecting the absorption of visible light by the chlorophyll pigments. Moreover, since at any time each of the tree species will be at different health states, the red-edge region will be efficient in separating the species based on their health status.

However, the present study did not show that the red-edge is the most important band in classifying the tree species. In fact, the red region and the NIR region outperformed the red-edge band in classifying the species. Other studies that have found the red and NIR region more important for classification of forest species using remote sensing data suggest that the presence of red-edge may only be important when incorporated with other standard bands.[8,41]

Overall, these results are consistent with the previous studies that have found that RapidEye bands, with its high spectral resolution, have great potential in terms of classifying and mapping vegetation species.[17,41–43]

7.3 *Machine Leaning Algorithms for Species Classification in Semiarid Environment*

The RF and SVM were compared on their ability to map CM and its four major coexisting species. McNemar test shows that there is no statistically significant difference between RF

Downloaded From: http://remotesensing.spiedigitallibrary.org/ on 02/10/2014 Terms of Use: http://spiedl.org/terms

and SVM. Nevertheless, SVM slightly outperformed RF by 3.75%. Previous studies have shown that RF and SVM are the best techniques for mapping tree species using high-spatial resolution imagery such as RapidEye.[44] These techniques are better than conventional classification methods such as maximum likelihood, minimum distance, etc., in that they handle voluminous, highly multicollinear, multispectral dataset.[45] Moreover, they are very computing time efficient compared to the conventional methods. To our knowledge, SVM and RF have never been compared for mapping tree species in areas with limited training samples like semiarid environments reminiscent of ours.

Other studies have shown that SVM generally outperforms RF, especially when the number of training pixels is small.[30,46] The main reason is most likely the result of the paradigm of SVM based on a small pixel sample (i.e., support vectors).[46] Consequently, SVM can be trained with a few meaningful pixels and is able to fit limited information. The results from this study show that in our context, SVM is able to map the tree species with higher accuracy from only small training pixels of each species (presence and absence) compared to RF. Based on the above, we therefore postulate that SVM should be used for mapping tree species in semiarid environments with small training pixels.

Furthermore, it is clear from our results that kappa, which incorporates an adjustment for random allocation agreement, is not a good statistic to determine the level of agreement between the training dataset and test dataset in classification process. Kappa has numerous conceptual flaws. Most importantly, it fails to distinguish quantity disagreement from allocation disagreement.[45] Therefore, it is far more important to report how much less than perfect the classification is, rather than how much better than random the classification is.[45,47] The quantity disagreement and allocation therefore present a better option to kappa.

8 Conclusion

This study has demonstrated the effectiveness of using new multispectral sensors with high spatial resolution and specific bands such as RapidEye when classifying tree species in a semiarid environment. Specifically, the study has highlighted three important conclusions.

1. CM and its coexisting species have a strong potential to be mapped using high-resolution data, such as RapidEye imagery, with high accuracy.
2. The presence of red-edge band in the RapidEye sensor has the potential for classifying tree species in semiarid environments when incorporated with other standard bands.
3. Considering the relative high accuracies, low cost, simplicity, and few parameters needed for operation, SVM will be preferred over RF in areas where there are limited training samples.

Acknowledgments

The authors would like to thank the University of KwaZulu-Natal for providing a research grant to conduct this study. We appreciate the Ministry of Environment in Botswana for granting a research permit to conduct the research.

References

1. W. Mojeremane and T. Kgathi, "Seed treatments for enhancing germination of Colophospermum mopane seeds: a multipurpose tree in Botswana," *J. Biol. Sci.* **5**(3), 309 (2005), http://dx.doi.org/10.3923/jbs.2005.309.311.
2. R. J. Sebego et al., "Mapping of Colophospermum mopane using Landsat TM in eastern Botswana," *S. Afr. Geogr. J.* **90**(1), 41–53 (2008), http://dx.doi.org/10.1080/03736245.2008.9725310.
3. M. Ditlhogo et al., "Interactions between the mopane caterpillar, Imbrasia belina, and its host, Colophospermum mopane in Botswana," in *Management of mopane in Southern Africa*, Ogongo Agricultural College, Namibia, C. Flower, G. Wardell-Johnson, and A. Jamieson, Eds., pp. 46–49 (1996).

Downloaded From: http://remotesensing.spiedigitallibrary.org/ on 02/10/2014 Terms of Use: http://spiedl.org/terms

4. S. Adelabu, O. Mutanga, and M. A. Cho, "A review of remote sensing of insect defoliation and its implications for the detection and mapping of Imbrasia belina defoliation of Mopane Woodland," *Afr. J. Plant Sci. Biotechnol.* **6**(1), 1–13 (2012).
5. L. J. Quackenbush and Y. Ke, "Investigating new advances in forest species classification," presented at *Proc. of 2007 ASPRS Annual Conf., 7-11 May 2007*, Reno, Navada (2007).
6. S. E. Franklin, *Remote Sensing for Sustainable Forest Management*, Lewis Publishers, Boca Raton, Florida (2001).
7. M. Wulder and S. Franklin, *Remote Sensing of Forest Environments: Concepts and Case Studies*, Kluwer Academic, Boston (2003).
8. C. Schuster, M. Forster, and B. Kleinschmit, "Testing the red edge channel for improving land-use classifications based on high-resolution multi-spectral satellite data," *Int. J. Remote Sens.* **33**(17), 5583–5599 (2012), http://dx.doi.org/10.1080/01431161.2012.666812.
9. J. Tigges, T. Lakes, and P. Hostert, "Urban vegetation classification: benefits of multitemporal RapidEye satellite data," *Remote Sens. Environ.* **136**, 66–75 (2013), http://dx.doi.org/10.1016/j.rse.2013.05.001.
10. A. Carleer and E. Wolff, "Exploitation of very high resolution satellite data for tree species identification," *Photogramm. Eng. Remote Sens.* **70**(1), 135–140 (2004).
11. M. A. Cho et al., "Improving discrimination of Savanna tree species through a multiple-endmember spectral angle mapper approach: canopy-level analysis," *IEEE Trans. Geosci. Remote Sens.* **48**(11), 4133–4142 (2010).
12. A. Lobo, "Image segmentation and discriminant analysis for the identification of land cover units in ecology," *IEEE Trans. Geosci. Remote Sens.* **35**(5), 1136–1145 (1997), http://dx.doi.org/10.1109/36.628781.
13. S. Ringrose et al., "Nature of the darkening effect in drought affected savannah woodland environments relative to soil reflectance," *Remote Sens. Environ.* **25**(25), 519–524 (1989).
14. J. C. W. Chan and D. Paelinckx, "Evaluation of random forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery," *Remote Sens. Environ.* **112**(6), 2999–3011 (2008), http://dx.doi.org/10.1016/j.rse.2008.02.011.
15. E. Adam and O. Mutanga, "Spectral discrimination of papyrus vegetation (Cyperus papyrus L.) in swamp wetlands using field spectrometry," *ISPRS J. Photogramm. Remote Sens.* **64**(6), 612–620 (2009), http://dx.doi.org/10.1016/j.isprsjprs.2009.04.004.
16. F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.* **42**(8), 1778–1790 (2004), http://dx.doi.org/10.1109/TGRS.2004.831865.
17. P. Krahwinkler and J. Rossman, "Using decision tree based multiclass support vector machines for forest mapping," presented at *IEEE Int. Geoscience and Remote Sensing Symp.*, Vancouver, Canada (2011).
18. Y. P. R. Bhalotra, "Climate of Botswana, Part 2: Elements of climate," Department of Meteorological Services, Ministry of Works, Transport and Communications: Botswana, B. M. Services, Ed. Gaborone, p. 104 (1987).
19. F. Do et al., "Stable annual pattern of water use by Acacia tortilis in Sahelian Africa," *Tree Physiol.* **28**(1), 95–104 (2008), http://dx.doi.org/10.1093/treephys/28.1.95.
20. I. Wessels, C. Waal, and W. D. Boer, "Induced chemical defences in Colophospermum mopane trees," *Afr. J. Range Forage Sci.* **24**(24), 141–147 (2007), http://dx.doi.org/10.2989/AJRFS.2007.24.3.4.297.
21. R. J. Sebego and W. Arnberg, "Interpretation of mopane woodlands using air photos with implication on satellite image classification," *Int. J. Appl. Earth Obs. Geoinf.* **4**(2), 119–135 (2002), http://dx.doi.org/10.1016/S0303-2434(02)00009-0.
22. ENVI, Environment for Visualising Images, ITT Industries Inc., Boulder, Colorado, ITT Industries Inc, USA (2006).
23. E. M. Adam et al., "Discriminating the papyrus vegetation (Cyperus papyrus L.) and its coexistent species using random forest and hyperspectral data resampled to HYMAP," *Int. J. Remote Sens.* **33**(2), 552–569 (2012), http://dx.doi.org/10.1080/01431161.2010.543182.
24. RapidEye, "Satellite imagery product specifications," 2011, http://www.RapidEye.de/upload/-RE_Product_Specifications_ENG.pdf (01 November 2011).

Downloaded From: http://remotesensing.spiedigitallibrary.org/ on 02/10/2014 Terms of Use: http://spiedl.org/terms

25. D. Naughton et al., "Absolute radiometric calibration of the RapidEye multispectral imager using the reflectance-based vicarious calibration method," *J. Appl. Remote Sens.* **5**(1), 053544 (2011), http://dx.doi.org/10.1117/1.3613950.
26. L. Breiman, "Random forests," *Mach. Learn.* **45**(1), 5–32 (2001), http://dx.doi.org/10.1023/A:1010933404324.
27. R. L. Lawrence, S. D. Wood, and R. L. Sheley, "Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (randomForest)," *Remote Sens. Environ.* **100**(3), 356–362 (2006), http://dx.doi.org/10.1016/j.rse.2005.10.014.
28. J. A. Benediktsson and J. R. Sveinsson, "Random forest classification of multisource remote sensing and geographic data," in *2004 IEEE Int. Geoscience and Remote Sensing Symp.*, Anchorage, Alaska, pp. 1049–1052 (**2004**).
29. A. Özçift, "Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis," *Comput. Biol. Med.* **41**(5), 265–271 (2011), http://dx.doi.org/10.1016/j.compbiomed.2011.03.001.
30. V. Vapnik, *Statistical Learning Theory. Support Vector Machines for Pattern Recognition*, John Wiley & Sons, New York (1998).
31. C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Min. Knowl. Discov.* **2**(2), 121–167 (1998), http://dx.doi.org/10.1023/A:1009715923555.
32. C. W. Hsu, C. C. Chang, and C. J. Lin, "A practical guide to support vector classification," 2009, http://www-personal.umich.edu/~yongjiaw/NLPproject/guide.pdf (22 October 2013).
33. R. Pontius, Jr. and M. Millones, "Death to kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment," *Int. J. Remote Sens.* **32**(15), 4407–4429 (2011), http://dx.doi.org/10.1080/01431161.2011.552923.
34. R. Ismail et al., "Determining the optimal resolution of remotely sensed data for the detection of Sirex noctilio infestations in Pinus patula plantations in KwaZulu-Natal, South Africa," *S. Afr. Geogr. J.* **90**(1), 196–204 (2008), http://dx.doi.org/10.1080/03736245.2008.9725308.
35. G. P. Petropoulos, C. Kalaitzidis, and K. P. Vadrevu, "Support vector machines and object-based classification for obtaining land-use/cover cartography from hyperion hyperspectral imagery," *Comput. Geosci.* **41**, 99–107 (2012), http://dx.doi.org/10.1016/j.cageo.2011.08.019.
36. R. Manandhar, I. O. A. Odeh, and T. Ancev, "Improving the accuracy of land use and land cover classification of Landsat data using post-classification enhancement," *Remote Sens.* **1**(3), 330–344 (2009), http://dx.doi.org/10.3390/rs1030330.
37. I. Nitze, U. Schulthess, and H. Asche, "Comparison of machine learning algorithms random forest, artificial neural network and support vector machine to maximum likelihood for supervised crop type classification," presented at *Proc. of the 4th GEOBIA, 7-9 May 2012*, Rio de Janeiro, Brazil (2012).
38. A. Gitelson and M. Merzlyak, "Quantitative estimation of chlorophyll-a using reflectance spectra: experiments with autumn chesnut and maple leaves," *J. Photochem. Photobiol. B* **22**(3), 247–252 (1994), http://dx.doi.org/10.1016/1011-1344(93)06963-4.
39. O. Mutanga and A. K. Skidmore, "Red-edge shift and biochemical content in grass canopies," *ISPRS J. Photogramm. Remote Sens.* **62**(1), 34–42 (2007), http://dx.doi.org/10.1016/j.isprsjprs.2007.02.001.
40. V. Thomas et al., "Canopy chlorophyll concentration estimation using hyperspectral and lidar data for a boreal mixedwood forest in northern Ontario, Canada," *Int. J. Remote Sens.* **29**(4), 1029–1052 (2008), http://dx.doi.org/10.1080/01431160701281023.
41. B. Tapsall, P. Milenov, and K. Tasdemir, "Analysis of RapidEye imagery for annual land-cover mapping as an aid to European Union (EU) common agricultural policy," in *ISPRS Technical Commission VII Symp.—100 Years*, Vienna, Austria, pp. 568–573 (2010).
42. S. Sah, "A multi-temporal fusion-based approach for land cover mapping in support of nuclear incident response," MS Thesis, Chester F. Carlson Center for Imaging Science of the College of Science Rochester Institute of Technology, Pune University (2013).
43. H. O. Kim, J. O. Yeom, and Y. S. Kim, "Agricultural land cover classification using RapidEye satellite imagery in South Korea—first result," *Proc. SPIE* **8174**, 817424 (2011), http://dx.doi.org/10.1117/12.897810.

Downloaded From: http://remotesensing.spiedigitallibrary.org/ on 02/10/2014 Terms of Use: http://spiedl.org/terms

44. R. Pouteau, A. Collin, and B. Stoll, "A comparison of machine learning algorithms for classification of tropical ecosystems observed by multiple sensors at multiple scales," 2011, http://www.isprs.org/proceedings/2011/isrse-34/211104015Final00913.pdf (22 October 2013).
45. D. J. Redo and A. C. Millington, "A hybrid approach to mapping land-use modification and land-cover transition from MODIS time-series: a case study from the Bolivian seasonal tropics," *Remote Sens. Environ.* **115**(2), 353–372 (2011), http://dx.doi.org/10.1016/j.rse.2010.09.007.
46. G. M. Foody and A. Mathur, "The use of small training sets containing mixed pixels for accurate hard image classification: training on mixed spectral responses for classification by a SVM," *Remote Sens. Environ.* **103**(2), 179–189 (2006), http://dx.doi.org/10.1016/j.rse.2006.04.001.
47. R. G. Pontius and M. Millones, "Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment," *Int. J. Remote Sens.* **32**(15), 4407–4429 (2011), http://dx.doi.org/10.1080/01431161.2011.552923.

Samuel Adelabu received the MSc degree in environmental science from University of Botswana. He is currently a doctoral candidate at the University of KwaZulu-Natal, South Africa. His main research lies in the use of remote sensing (RS) in monitoring forest health and climate change and is currently studying the remote sensing of insect defoliation in mopane woodland. He worked at the University of Botswana as a graduate assistant and is currently a graduate assistant at UKZN.

Onisimo Mutanga is a professor in the School of Agriculture, Earth and Environmental Science at the University of KwaZulu-Natal. He received a PhD in hyperspectral remote sensing of tropical grass quality and quantity from Wageningen University-ITC (Netherlands) in 2004. His expertise lies in ecological remote sensing. He has graduated 7 PhDs and 14 masters students. He has published more than 67 articles and has several conference proceedings and book chapters.

Elhadi Adam is senior lecturer in the Department of Geography at University of Limpopo. He has a PhD in hyperspectral remote sensing of wetland vegetation quality and quantity at University of KwaZulu-Natal in 2010. His research lies in applications of remote sensing (RS) and geographic information systems (GIS) in applied environmental science. He is currently supervising MSc and PhD students. He has published more than 15 articles in ISI journals.

Moses A. Cho is currently a principal research scientist with the earth observation (EO) group at the Council for Scientific and Industrial Research (CSIR), South Africa, and a research fellow with the University of KwaZulu-Natal South Africa. He obtained his PhD degree in 2007 from the Wageningen University-ITC (Netherlands). His current research interest involves the use of hyperspectral and new specialized spaceborne multispectral imagery for assessing floral diversity and vegetation productivity.

Downloaded From: http://remotesensing.spiedigitallibrary.org/ on 02/10/2014 Terms of Use: http://spiedl.org/terms