SCIENTIA FORESTALIS

ISSN 2318-1222 (online)

ORIGINAL ARTICLE

Tree species classification using machine learning algorithms with OHS-2 hyperspectral image

Classificação de espécies de árvores usando algoritmos de aprendizado de máquina com imagem hiperespectral OHS-2

Nan Wang[1,2,3], Guisheng Wang[1,2,3]

[1]School of Spatial Information and Geomatics, Anhui University of Science and Technology, Huainan City, China
[2]Postdoctoral Station of Geological Resources and Geological Engineering, Anhui University of Science and Technology, Huainan City, China
[3]Postdoctoral Working Station, Anhui Huayin Mechanical and Electrical Co., Ltd., Huainan City, China

How to cite: Wang, N., & Wang, G. (2023). Tree species classification using machine learning algorithms with OHS-2 hyperspectral image. *Scientia Forestalis*, *51*, e3991. https://doi.org/10.18671/scifor.v51.18

Abstract

Considering the form diversity of tree species composition in the Bagong Mountain National Forest Park of China, we mapped tree species utilizing Machine Learning Algorithms (support vector machines (SVM) and random forest (RF) classifiers) based on the OHS-2 hyperspectral satellite image by different datasets which combined spectral information and hyperspectral-derived vegetation indices (VIs) for improving tree species classification and explored the best performance of them. To verify the improvement, the results of physically-based spectral classifiers (spectral angle mapper (SAM) and maximum likelihood (ML) classifiers) were applied to compare with the results of machine learning algorithms. The results indicated an overall accuracy of 94.01%, 96.08%, 82.9% and 79.3% for SVM, RF, SAM and ML classifiers of the best performance using different datasets. Highest accuracies resulted from two machine learning algorithms classifiers; SVM and RF compared to SAM and ML classifiers. Although SVM outperformed RF when using all hyperspectral bands and VIs, the overall accuracy of the RF classifier is higher when compared to the SVM classifier using VIs combined selected features. Meanwhile, the RF classifier performed better than SVM after removing the redundancy of spectral data in training samples. Moreover, the machine learning algorithms successfully classified a small number of tree species (*Cedrus deodara* and *Pterocarya stenoptera* C. DC.) in the study area, but the physical spectroscopy-based method failed to classify these species. Such integration strategy improved the effectiveness of enhancing the accuracy of tree species classification and mapping their distribution on broad spatial and temporal scales using machine learning algorithms and hyperspectral imagery.

Keywords: Physically-based spectral classification; Machine learning algorithms classification; Hyperspectral satellite data; Tree species; Broadleaf tree classification.

Resumo

Considerando a diversidade da composição de espécies arbóreas no Parque Nacional da Montanha Bagong, na China, mapeamos as espécies arbóreas utilizando Algoritmos de Aprendizado de Máquina (classificadores de máquinas de vetor de suporte (SVM) e floresta aleatória (RF)), com base na imagem hiperespectral do satélite OHS-2, por diferentes conjuntos de dados que combinam informações espectrais e índices de vegetação (VIs) derivados de hiperespectral para melhorar a classificação das espécies arbóreas e explorar o melhor desempenho desses algoritmos. Para verificar a melhoria, os resultados dos classificadores espectrais baseados em física (mapeador de ângulo espectral (SAM) e máxima verossimilhança (ML)) foram aplicados para comparar com os resultados dos algoritmos de aprendizado de máquina. Os resultados indicaram uma precisão geral de 94,01%, 96,08%, 82,9% e 79,3% para os classificadores SVM, RF, SAM e ML do

Financial support: This research was funded by the innovative project: National Key Research and Development Project of China under Grant no. 2018YFC0407703-1; National Natural Science Foundation of China under Grant Number 41501294; Excellent Young Talents Fund Program of Higher Education Institutions of Anhui Province of China 2016; University Natural Science Research Project of Anhui Province under Grant Number KJ2019A0136.
Conflict of interest: Nothing to declare.
Corresponding author: wnbadinine@163.com
Received: 2 February 2023.
Accepted: 18 May 2023.
Editor: Mauro Valdir Schumacher.

melhor desempenho usando diferentes conjuntos de dados, respectivamente. As precisões mais altas foram obtidas pelos dois classificadores de algoritmos de aprendizado de máquina, SVM e RF, em comparação com os classificadores SAM e ML. Embora o SVM tenha tido melhor desempenho que o RF ao utilizar todos os bandas hiperespectrais e VIs, a precisão geral do classificador RF é maior em comparação com o classificador SVM ao utilizar recursos selecionados de VIs combinados. Além disso, o classificador RF teve melhor desempenho que o SVM após a remoção da redundância de dados espectrais nas amostras de treinamento. Além disso, os algoritmos de aprendizado de máquina classificaram com sucesso um pequeno número de espécies arbóreas (*Cedrus deodara* e *Pterocarya stenoptera* C. DC.) na área de estudo, enquanto o método baseado em espectroscopia física falhou em classificar essas espécies. Essa estratégia de integração melhorou a eficácia de aprimorar a precisão da classificação de espécies arbóreas e mapear sua distribuição em escalas espaciais e temporais amplas usando algoritmos de aprendizado de máquina e imagens hiperespectrais.

Palavras-chave: Classificação espectral baseada em física; Classificação de algoritmos de aprendizado de máquina; Dados de satélite hiperespectral; Espécies de árvores; Classificação de árvores de folha larga.

INTRODUCTION

Forests contribute significantly to carbon fixation, regional ecological security, and oxygen release while supporting biodiversity. In forest management, forest resource information acquisition and monitoring, including tree species dominance, distribution, and composition, is one of the basic and core works (Dalponte et al., 2012; Heinzel & Koch, 2012; Schepaschenko et al., 2015; Rodríguez-González et al., 2017). Furthermore, the tree species were also required for landscape design and services. Thus, it is essential to recognize tree species accurately and map the tree species composition for sustainable forest management.

Owing to the fine-scale spatial variation, time-consuming nature of field survey, and high diversity of families, genera and species, an effective way to map tree species on large spatial and temporal scales is by combining remote sensing techniques with ground-based data. Remote sensing provides some efficiency to identify and map the tree species, data sources including multispectral, hyperspectral, and LiDAR. Since 2000, very high resolution (VHR) multispectral images. Images acquired by typical sensors of IKONOS and WorldView, were applied in several studies (Ferreira et al., 2019). They present detailed information regarding the spatial distribution of land cover or identifying forest type (broadleaf, conifer). In the past decades, forest classifications ranged from more general classifications of forest type (deciduous and coniferous trees) (Ghosh et al., 2014) to narrower focus classifications of single tree species (Zhang & Xie, 2012). Moreover, Light Detection and Ranging (LiDAR) instruments mounted on aircraft can present detailed vegetation structure measurements leading to accurate tree species information. However, compared to the hyperspectral images, there was poorer spectral information. Some of the forest types often yield similar spectral characteristics, which leads to the misclassification with wide-band remote sensing data. Furthermore, since the optical conditions orienting the optical remote sensing vary greatly, it would lead to significantly different spectral properties for the same types ("homologous" phenomena).

Because of the mixed forest comprising several different tree species, evaluating the spatial distribution of different tree species is one of the great challenges for forest managers. There are still disputes regarding the role of tree species classification at the individual species level in forest management, especially broad-leaved trees and homologous species. Compared with the multispectral images classifications in tree species, it was indicated that hyperspectral images classifications can enhance classification precision by providing more information from narrow bands to discriminate spectrally-similar targets. This is when multispectral classification fails to capture the slight spectral differences occurring between tree species (Dalponte et al., 2012; Heinzel & Koch, 2012; Fassnacht et al., 2014).

Tree species separability is restricted by high spectral intraspecies variability exceeding the interspecies variability. This is caused by phenological effects and differences in tree age and health, the openness of canopies, shadowing effects, and environmental variability (Waser et al., 2010; Richter et al., 2016). Thus, despite the abundance of information contained in hyperspectral images, the challenge remained to discriminate species within the same genus, often leading to misclassification (Heinzel & Koch, 2012; Peerbhay et al., 2013).

Therefore, we utilized dataset combining hyperspectral satellite data, collecting spectral data, and spectral vegetation indices to enhance the identification of tree species. Therefore, we compared the Physically-based Spectral and Machine Learning Algorithms Classification approach, providing a reference method for tree species identification in a complex and heterogeneous forest. The specific objectives are:

1) Classifying tree species utilizing a combination of hyperspectral satellite data, collecting spectral data and spectral vegetation indices, and evaluating the accuracy of the tree species classification;
2) Investigating the effect of the applied classifier and the influences of different spectral and spectral vegetation indices for tree species classification in complex environments.

MATERIAL AND METHODS

Study area: The study area, Bagong Mountain National Forest Park (32°33'-33°65'N, 116°30'-117°81'E), is located in the west of Huainan City in the middle of Anhui Province, China (Figure 1). It covers an area of approximately 2759 ha, which is a typical forest park and scenic site in China. It contains mountains at altitude ranges from 80 to 241.2 m with an average annual precipitation of 893.4 mm and annual sunshine hours of 1922.2 h. The forests of Bagong Mountain are mainly composed of manmade single forests, as well as deciduous forests and coniferous-broadleaved mixed forests dominated by *Platycladus orientalis* (L.) Franco (P), *Pinus massoniana* Lamb (P.L), *Vernicia fordii* (Hemsl.) Airy Shaw (V.S), *Castanea mollissima* Blume (C.B), and *Pistacia Chinensis* Bunge (P.B). Additionally, numerous tree species were successfully introduced and grew on bare land, including *Cedrus deodara* (Roxb. ex D. Don) G. Don (C.D), *Platanus orientalis* Linn. (P.O.L), *Broussonetia papyrifera* (B.P), *Robinia pseudoacacia* L. (R.L), *Pyrus sorotina* (P.S), *Prunus persica* L. (P.P.L), *Pterocarya stenoptera* C. DC. (P.DC) and *Magnolia Grandiflora* Linn (M.L) in the late 1990s.

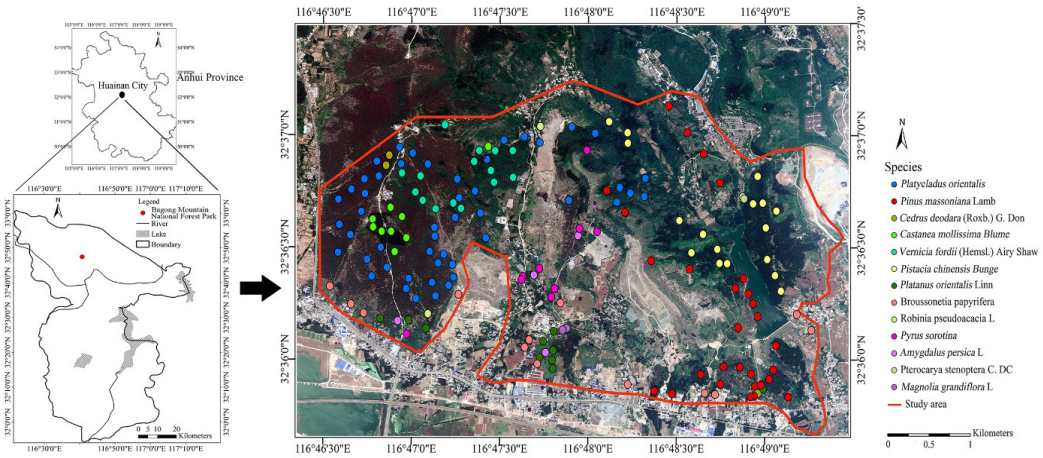

Figure 1. Location of the Bagong Mountain National Forest Park, Huainan City, the sample of trees species which were collected by field survey with GPS. The location map was cited from Google Earth, and the vegetation map was from the ArcGIS (GIS, Geographic Information System) database established in 2020.

Hyperspectral satellite (OHS-2) data: The Orbita hyperspectral satellites (OHS) were launched successfully on 26 April 2018. The OHS-2 satellite images present 32 spectral channels from 466 to 940 nm with a constant channel width of 2.5 nm and 10m spatial resolution. In this study, the hyperspectral image of the constellation, which was captured in the middle of the growing season on 6 October 2018 during good weather and clear skies, was used. The image contains 5056 × 5056 pixels, as well as the study site regions with spatial extents of 319 lines × 466 pixels with UTM WGS 84 Zone 50 North geometric projection.

Field survey data: The field surveys containing tree species identification and tree species-based spectrum collection were conducted under cloudless conditions in September and October 2018 during good weather and clear skies. The field spectrum data were collected with the ASD FieldSpec 4ground object spectrometer whose channel width was 1nm and the collected spectrum range from 350 to 2500 nm. The field survey data included location information of all land cover types and 13 tree species in the dominant layer. Each tree species consists of 10 measurement points while recording 30 spectra as a sampled spectrum. Considering the effect of water vapor in the atmosphere on the measurement, the spectral ranges of 1350 ~ 1410nm, 1800 ~ 1965 nm, and 2400 ~ 2500 nm were removed. Limited to wavelength range and spectral resolution of OHS-2 satellite data, we developed a reference spectral library with the wavelength range similar to OHS-2 satellite data based on field spectral. They were resampled to 10nm spectral resolution. The average reflectance curves of all the 13 species of canopy pixels at 466-940nm wavelengths along with the 32 hyperspectral bands are shown in Figure 2. Furthermore, to provide training and test samples of forest classification, a total of 180 field survey samples was collected with Hand-held GPS (GPS receiver is Garmin MAP 62CS; accuracy: ±3m). The data included location information of all land cover types and 13 tree species in the dominant layer. The data for each tree species were collected from more than 15 samples, of which 70% of these samples were used to classify, and 30% of the samples were utilized to verify the classification accuracy.

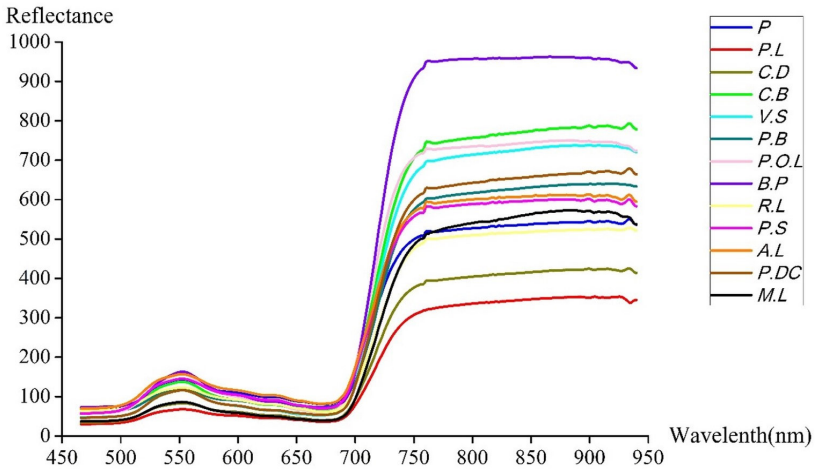

Figure 2. The mean reflectance value of all 13 tree species at 466-940 nm wavelengths.

Among the 13 species, the most obvious differences were concentrated in the green peak (550 nm), NIR shoulder, red-edge (667-755 nm), and plateau (755-940 nm). There were 2 pairs of classes with high similarity, including *Amygdalus persica* L. and *Pyrus sorotina*, as well as *Vernicia fordii* (Hemsl) Airy Shaw and *Platanus orientalis* L. However, *Broussonetia papyrifera* possessed the highest reflectance variation within the NIR wavelengths compared to the others. Considering the difference in green, red-edge, and NIR, we assessed the separability of species based on the average spectral data, and examined the multicollinearity. Therefore, we selected hyperspectral bands with significant reflectance differences of OHS image to constitute the dataset for tree species classification.

Classification datasets composition: A total of 4 classification datasets was shown in Table 1. Owing to the Hughes phenomenon which can lead to decreased classification accuracy (Hughes, 1968), the maximum noise fraction (MNF) transformation was used to obtain noise isolation, reduce the dimensionality, and pack the coherent information in a smaller set of features (Green et al., 1988; Boardman & Kruse, 1994). Transforming the hyperspectral with the MNF algorithm, the resulting MNF components were visually screened. The cumulative proportion of the MNF was 86.7% in 10 bands.

Table 1. Classification datasets composition with different features.

Classification dataset	Data source	Constituent Feature
Classification dataset 1	OHS-2 image	All the hyperspectral bands
Classification dataset 2	OHS-2 image	All the hyperspectral bands and VIs
Classification dataset 3	Selected hyperspectral bands and VIs	Band3-band7, band16-band18, band23-30, MNF components 1-10 and VIs
Classification dataset 4	MNF bands and VIs	MNF components 1-10 and VIs

In addition, we calculated 16 narrowband VIs and excluded all non-forest pixels with index values as the threshold (Table 2). Then, VIs were added to the classification datasets (see Table 1).

Table 2. Description of selected hyperspectral features.

Features used	Description	Description and Formula
Band3	Reflectance at 500 nm	-
Band4	Reflectance at 520 nm	-
Band5	Reflectance at 536 nm	-
Band6	Reflectance at 550 nm	Reflectance at 566 nm
Band7	Reflectance at 566 nm	Yellow edge position
Band16	Reflectance at 700nm	Red Edge Position
Band17	Reflectance at 716 nm	Red Edge Position
Band18	Reflectance at 730 nm	Red Edge Position
Band23-30	Reflectance at 750 -940 nm	Near Infrared (NIR) shoulder and plateau
VOG1	Vogelmann Red Edge Index 1	$VOG1 = \frac{\rho 740}{\rho 550}$
VOG2	Vogelmann Red Edge Index 2	$VOG2 = \frac{(\rho 734 - \rho 747)}{(\rho 715 + \rho 726)}$
ARI1	Anthocyanin Reflectance Index 1	$ARI1 = \frac{1}{\rho 550} - \frac{1}{\rho 700}$
ARI2	Anthocyanin Reflectance Index 2	$ARI2 = \rho 800\left(\frac{1}{\rho 550} - \frac{1}{\rho 700}\right)$
C1	Chlorophyll Index (C1)	$C1 = \frac{(\rho 850 - \rho 710)}{(\rho 850 + \rho 680)}$
C2	Chlorophyll Index (C2)	$C2 = \frac{\rho 750}{\rho 700}$
$NDVI_{705}$	Red Edge Normalized Difference Vegetation Index	$NDVI_{705} = \frac{\rho 750 - \rho 705}{\rho 750 + \rho 705}$
NVDI	Normalized Difference Vegetation Index	$NDVI = \frac{\rho 800 - \rho 670}{\rho 800 + \rho 670}$
REP_{mean}	The median reflectance between 690 nm to 740 nm	$REP_{mean} = \sum_{i=690}^{i=740} \frac{\rho_i}{N}$
RVI1	Ratio vegetation stress index1	$RVI1 = \frac{\rho 694}{\rho 760}$
RVI2	Ratio vegetation stress index2	$RVI2 = \frac{\rho 600}{\rho 760}$
RVI3	Ratio vegetation stress index3	$RVI3 = \frac{\rho 710}{\rho 760}$

Where: ρ_i is the reflectance at the i wavelength; and N is the band number.

Methodology

Pre-processing of the hyperspectral data: The pre-processing of the hyperspectral image occurred separately including radiometric calibration, atmospheric corrections and topographic correction. The radiometric calibration was implemented at first, which conveyed the digital numbers (DN) values of the OHS-2 images into the top of atmosphere (TOA) radiance with certain spectral response functions (Equation 1).

$$L_e = gain \times \frac{DN}{TDIStage} + offset \quad (1)$$

Where L_e is the apparent radiance; gain is the absolute radiation calibration gain coefficient, offset is the absolute radiation calibration deviation coefficient, and TDIStage is the integral series.

Then OHS-2 images were corrected atmospherically utilizing the fast line-of-sight atmospheric analysis of spectral hypercubes (ENVI-FLAASH) module based on MODTRAN 4+ radiation transmission model (Berk et al., 1987) converting the radiance to reflectance. In addition, the topographic correction (Equation 2) based on DEM (Digital Elevation Model) data, solar elevation, and sensor azimuth was applied, which reduced the influence of radiometric distortion caused by the mountainous terrain. The control points collected by hand-held GPS was used for geometric correction to solve the problems of squeezing, stretching, distorting and shifting of image pixels relative to the actual position of ground objects.

$$L_m = L - m \cdot \cos i - b + L_\alpha \quad (2)$$

Where L_m is the pixel brightness value after correction, L represents pixel brightness value before correction, L_α represents the average pixel brightness value before correction, and i is the incidence angle of the sun rays on the horizontal plane. The parameter of m and b represent the slope and intercept of the image brightness value and cosi linear regression equation, respectively.

Due to the influence of the natural environment, some noise appears in the spectral curve, which was indented in some range. Savitzky-Golay filter can smooth fine sawtooth noise while maintaining spectral characteristics unaffected (Savitzky & Golay, 1964). Therefore, it was utilized to filter the image and process the image more accurately and flexibly. We used the parameters which include the order of 0, length of 5 and degree of 3 after several experiments with different settings.

Classifier methods selection: In this study, we employed a comparative approach to assess the performance of physically-based spectral classifiers (spectral angle mapper and maximum likelihood classifier) and machine learning algorithms (support vector machine and random forest classifier) in identifying tree species. To this end, we used different datasets to train and test the classifiers.

As the typical representative of machine learning classifier, the random forest (RF) and support vector machine (SVM) classifier were widely used to solve problems related to hyperspectral data in both tree species classifications (Dalponte et al., 2009; Jones et al., 2010; Koetz et al., 2008) and land cover (forest type) (Tarabalka et al., 2010; Madroñal et al., 2017) classifications providing high reliability and classification accuracies.

RF classifier constructs decision trees using random resampling and node splitting techniques, which are then combined with Ensemble Learning. Input variables are chosen based on their ability to distinguish between target classes (Breiman, 2001; Belgiu & Drăguţ, 2016). In a random forest, each node is split using the best subset of variables randomly selected at that node. With robustness and compatibility for noise data and data with missing values in the RF algorithm, it was able to analyze the complex interactions among the classification features.

RF algorithm was chosen as a result of its non-parametric nature. RF classifier parameters include the number of features and the number of trees, the number of features used for training, and the number of trees used for constructing the composition of the classifier.

SVM classifier is adept at handling high-dimensional data without the need for an increase in training sample size (Mountrakis et al., 2011). Therefore, data reduction is unnecessary for classifiers such as SVM (Ghosh et al., 2014).The selected SVM classifier parameters include kernel function type, gamma, and penalty parameter. The kernel function controls dimension, the gamma controls the smoothness of the hyperplane, and the penalty parameter controls the error penalty.

SAM classifier matches pixel spectra of a satellite image to reference spectra using spectral angle as an index. It calculates the cosine of the angle between the spectra to determine their similarity in a space with dimensionality equal to the number of bands (Kruse et al., 1993).

ML classifier establishes a set of nonlinear discriminant functions based on Bayesian criteria for tree species classification (Richards & Jia, 1999). Unclassified pixels are assigned a belonging category based on the relative probability density functions of each category.

Accuracy assessment: After classification, we generated a total of 100 random samples and extracted their attributes based on classification result. Then we evaluated the classification assuracy using the random samples and field acquisition data of each tree species. The accuracy of the classifications were assessed based on the confusion matrix, overall accuracy, user's accuracy, producer's accuracy, and Kappa coefficient (Equations 3-5) (Congalton & Mead, 1983).

$$P_o = \sum_{i=1}^{n} P_i + P_j = \sum_{i=1}^{m} x_{ii} \Big/ N \tag{3}$$

$$P_e = \sum_{i=1}^{m} (x_{i+} x_{+i}) / N^2 \tag{4}$$

$$Kappa = \frac{P_o - P_e}{1 - P_e} \tag{5}$$

Where P_o is overall accuracy, P_e is the user's accuracy. P_i is the predicted probabilities, and P_j is the actual probabilities, n is the number of classes. x_{ii} is the number of pixels in row i or column i, x_{i+} is the total number of pixels in line i, x_{+i} is the total number of pixels in column i, and N is the total number of pixels used for accuracy evaluation.

RESULTS

Overall Classification Results: All classification datasets led to an overall accuracy of more than 80% (Table 3). For the datasets of spectral bands only in dataset 1, the RF exhibited obvious advantages over the others. Once the VIs was added to the dataset of 2, the SVM classifier performs even better, indicating that VIs became more significant for improving the overall accuracy of the SVM classifier. Based on the results of dataset 2 and 3, it is found that the RF classifier performs better than the SVM after removing the redundancy of spectral data. However, the dataset 3 and 4 with MNF components in the RF classifier exhibited advantages over the SVM classifier. Compared with the results of dataset 3 and 4, it is indicated that increasing the types of training data was helpful to improve the classification accuracy of both RF and SVM classifiers.

Table 3. The overall accuracy of the classification datasets based on three classifiers.

Classification datasets	Training sample	Classifier	Overall Accuracy	Kappa coefficient
Classification dataset 1	All the hyperspectral bands	RF	88.63%	0.892
		SVM	82.31%	0.813
		SAM	82.9%	0.787
		ML	79.3%	0.753
Classification dataset 2	All the hyperspectral bands and VIs	RF	91.6%	0.921
		SVM	94.01%	0.932
Classification dataset 3	Band3-band7, band16-band18, band23-30, MNF components 1-10 and VIs	RF	96.08%	0.951
		SVM	93.9%	0.912
Classification dataset 4	MNF components 1-10 and VIs	RF	90.36%	0.903
		SVM	87.14%	0.891

VIs were the vegetation indices, and the spectral features participating in the tree species classification were collected from OHS-2 data.

Comparison of four classifiers performance using different datasets: RF classifier was conducted to evaluate the performance of classifying tree species using 38 hyperspectral variables (classification dataset 3) (Figure 3). The parameters of the RF classifier had the feature number of 13 and the number of the trees of 200. Table 4 represents the confusion matrices, producers, and user's classification accuracies for each species utilizing the RF classifier. The classification yielded an overall accuracy of 96.08% and the Kappa value of 0.951. The producer's accuracies ranged from 75.0% to 97.96% (Table 4). Moreover, the classification of C.D and M.L represented the lowest and the highest producer's accuracies, respectively. The user's accuracies ranged from 75.56% to 98.25%, and the lowest and highest user's accuracies appeared in the classification of B.P and P. Regarding tree species classification utilizing dataset 1, some misclassified samples were localized, particularly at the junction of species distribution such as P and C.B, C.B, and P.B.

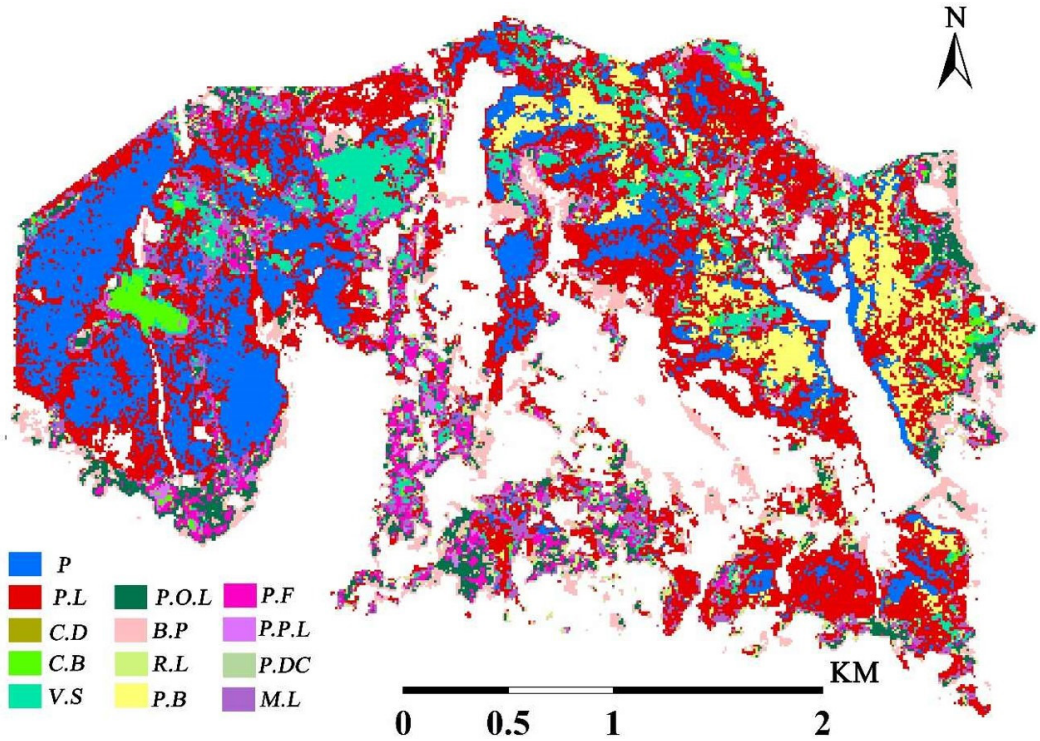

Figure 3. Tree species classification using RF classifier based on dataset 3.

Table 4. Confusion matrices for the RF classifier using Classification dataset 3

Class Name	P	P.L	C.D	C.B	V.S	P.B	P.O.L	B.P	R.L	P.S	P.P.L	P.DC	M.L	Total	User'sAccuracy (%)
P	336	3	0	4	0	0	0	0	0	0	0	0	0	343	97.96
P.L	4	170	2	0	0	0	0	0	0	0	0	0	0	176	96.59
C.D	2	3	27	0	0	0	0	0	0	0	0	0	0	32	84.38
C.B	5	0	0	87	1	5	0	0	0	0	0	1	0	99	87.88
V.S	0	0	0	1	85	2	0	0	0	0	0	0	0	88	96.59
P.B	0	0	0	8	0	89	0	0	2	0	0	0	0	99	89.90
P.O.L	0	0	0	0	5	0	36	3	0	0	0	0	0	44	81.82
B.P	0	0	0	0	2	0	3	34	2	2	2	0	0	45	75.56
R.L	0	0	0	0	0	2	0	2	18	0	0	0	0	22	81.82
P.S	0	0	0	0	0	0	0	0	0	32	3	0	0	35	91.43
P.P.L	0	0	0	0	0	0	0	0	0	6	21	0	0	27	77.78
P.DC	0	0	0	2	0	2	0	0	0	0	0	18	0	22	81.82
M.L	0	0	3	0	0	0	0	0	0	0	0	0	12	15	80.00
Total	347	176	36	98	93	100	39	39	22	40	26	19	16		
Producer's Accuracy (%)	96.82	96.59	75.00	88.78	91.40	89.00	92.31	87.18	81.82	80.00	80.77	94.74	75.00		
Overall Accuracy (%)	96.08														
Kappa coefficient	0.951														

Prior to using the SVM classifier for classification, normalization was applied to each spectral band and vegetation index using the logistic function to eliminate any negative effects caused by unusual sample data. SVM classifier utilized all hyperspectral bands and 12 vegetation indices (classification dataset 2) (Figure 4), with a gamma value of 0.021 and a penalty parameter of 100 using the radial basis function after data normalization. Incorporating vegetation indices (Table 5) improved the overall accuracies for tree species classification from 82.31% to 94.01% compared to the RF classifier, with a corresponding increase in kappa coefficient from 0.813 to 0.932. The SVM classifier resulted in a more generalized and uniform classification of species.

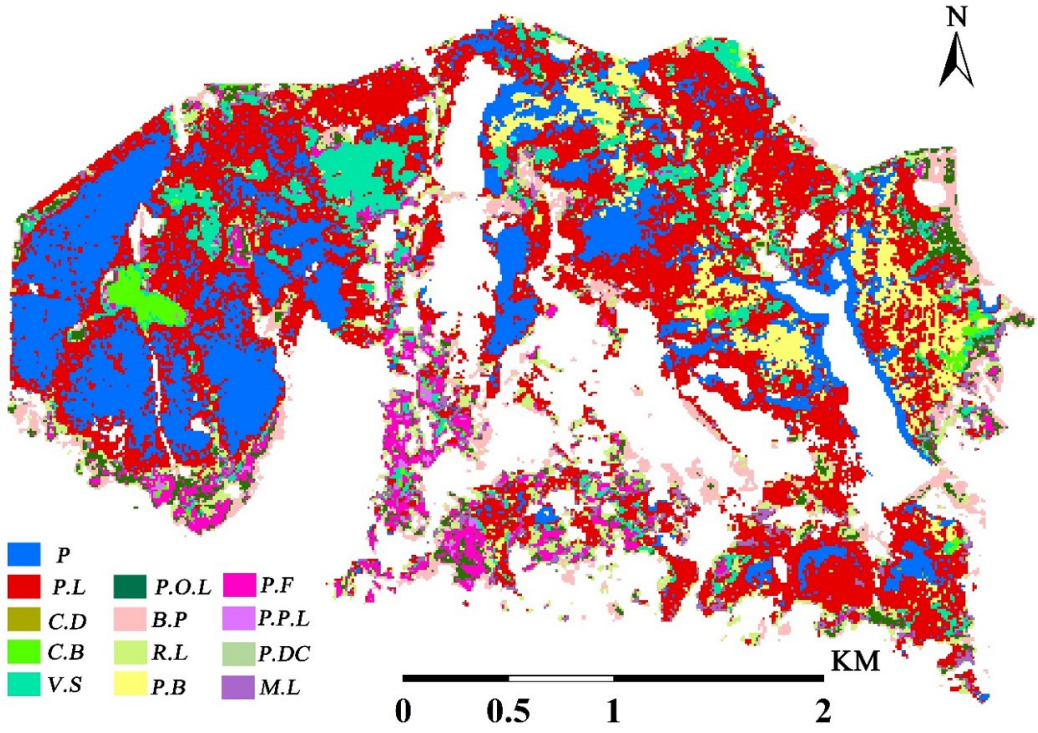

Figure 4. Tree species classification using SVM classifier based on dataset 2.

Table 5. Confusion matrices for the SVM classifier using classification dataset 2

Class Name	P	P.L	C.D	C.B	V.S	P.B	P.O.L	B.P	R.L	P.S	P.P.L	P.DC	M.L	Total	User's Accuracy (%)
P	349	1	4	3	0	0	0	0	0	0	0	0	0	357	97.76
P.L	3	185	1	0	0	4	0	0	3	0	0	0	0	196	94.39
C.D	0	0	25	0	0	0	0	0	0	0	0	0	2	27	92.59
C.B	2	0	0	78	2	0	0	0	0	0	0	0	0	82	95.12
V.S	0	0	0	4	85	0	0	0	0	1	2	3	0	95	89.47
P.B	0	3	0	3	2	34	0	0	0	0	0	0	0	42	80.95
P.O.L	0	0	0	0	0	0	20	2	0	0	0	0	0	22	90.91
B.P	0	0	0	0	0	0	1	27	0	0	0	0	0	28	96.43
R.L	0	0	0	0	0	0	0	6	16	0	0	0	0	22	72.73
P.S	0	0	0	0	0	0	0	0	0	32	3	0	0	35	91.43
P.P.L	0	0	0	0	0	0	0	0	0	2	29	0	0	31	93.55
P.DC	3	0	0	0	0	0	0	0	0	0	0	15	0	18	83.33
M.L	0	1	0	0	0	0	0	0	0	0	0	0	18	19	94.74
Total	357	190	30	88	89	38	21	35	19	35	34	18	20		
Producer's Accuracy (%)	97.76	97.37	83.33	88.64	95.51	89.47	95.24	77.14	84.21	91.43	85.29	83.33	90.00		
Overall Accuracy (%)	93.61														
Kappa coefficient	0.932														

The classification of P represented the highest producer's accuracy (97.36%), while the lowest producer's accuracies (83.33%) were represented by the classification of C.D and P.DC. The lowest and highest user accuracies appeared in the classification of P.B (80.95%) and P (97.76%), respectively. The user's accuracy of R.L is low, which may be caused by insufficient test samples. The classification accuracies slightly improved in most tree species by utilizing the combined classification dataset 2. Comparing the results of different variable combinations through visual assessment, it was consistently found that the SVM classifier led to more homogeneous and generalized results compared to others. Two areas tended to perform higher instances of misclassification standing along the east and north region. The stands along the east part caused a challenge for the other classification datasets, with P.B misclassified as C.B and P.L as P. The results indicated fragmentation in the northern part regardless of classification datasets and classifier. In this area, the forests are mainly broad-leaved mixed forests mainly comprising of C.B, P.B, and P.DC with larger sizes distributed in the upper layer and other broad-leaved trees in the understory.

We compared the angle between the end member spectral (ROI average spectral) collected from OHS-2 data and the field spectral data (Figure 5). The results revealed the best performance of the SAM classifier with overall accuracy at 82.9% and Kappa coefficient at 0.787 (Table 6). To summarize the classification accuracy of each tree species, the confusion matrix is used including overall accuracy, producer's, and user's classification accuracies. SAM classifier provided in Table 6 was used in this regard.

Table 6. Confusion matrices for the SAM classifier using classification dataset 1

Class Name	P	P.L	C.D	C.B	V.S	P.B	P.O.L	B.P	R.L	P.S	P.P.L	M.L	Total	User's Accuracy (%)
P	217	22	4	3	6	0	0	0	0	0	0	0	252	86.11
P.L	13	188	15	0	0	4	0	0	3	0	0	2	225	83.56
C.D	6	25	56	0	0	0	0	2	0	0	0	8	97	57.73
C.B	11	4	0	78	2	0	0	0	0	0	0	0	95	82.11
V.S	0	0	0	4	81	0	0	0	0	1	2	0	88	92.05
P.B	0	3	0	3	2	29	0	0	0	0	0	0	37	78.38
P.O.L	0	0	0	0	10	0	44	4	3	0	0	0	61	72.13
B.P	0	0	0	0	0	0	6	127	3	1	1	0	146	86.99
R.L	0	0	0	0	0	0	3	2	34	0	0	0	39	87.18
P.S	0	0	0	0	0	0	0	0	0	38	3	0	41	92.68
P.P.L	0	0	0	0	0	0	0	0	1	2	28	0	31	90.32
M.L	0	1	1	0	0	0	0	0	0	0	0	26	28	92.86
Total	247	243	76	88	101	33	53	135	44	42	34	44		
Producer's Accuracy(%)	87.85	77.37	73.68	88.64	80.20	87.88	83.02	94.07	77.27	87.50	82.35	59.09		
Overall Accuracy(%)	82.9													
Kappa coefficient	0.787													

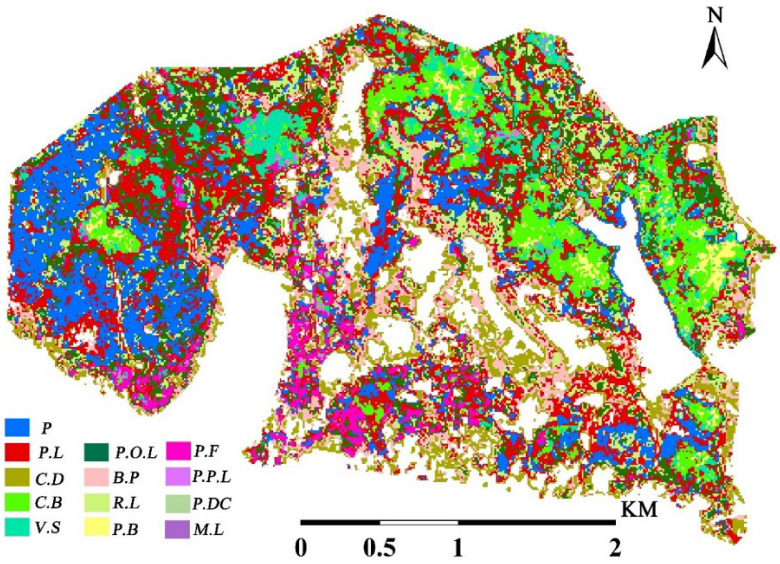

Figure 5. Tree species classification using SAM classifier based on dataset 1.

The producer's accuracies ranged from 58.33% to 94.07%, and the user's accuracy ranged from 52.87% to 95.49% (see Table 6). The classification of C.D and V.S represented the lowest and the highest producer's accuracies, respectively. The user's accuracies ranged from 52.87% to 95.49%, and the lowest and highest user's accuracies appeared in the classification of C.D and B.P. Based on the results above, the highest producer's and user's accuracies appeared in the classification of B.P (94.07% and 95.49%) respectively. C.D possessed the lowest accuracy whether it is producer's or user's accuracy. The results represented some areas of misclassified C.D in the P.L stand. Besides, the user's accuracy of P.O.L and the producer's accuracies of R.L and M.L were less than 75%.

ML classifier establishes the nonlinear discriminant function based on the spectral features reflected by the training samples in the remote sensing images (Figure 6). Then, the similarity between the area and training samples was calculated for classification. Thus, a total of 119 training samples for 13 tree species were established based on hand-held GPS data for tree species classification using the ML classifier. These training sample areas are evenly distributed in the whole remote sensing image according to the distribution of tree species.

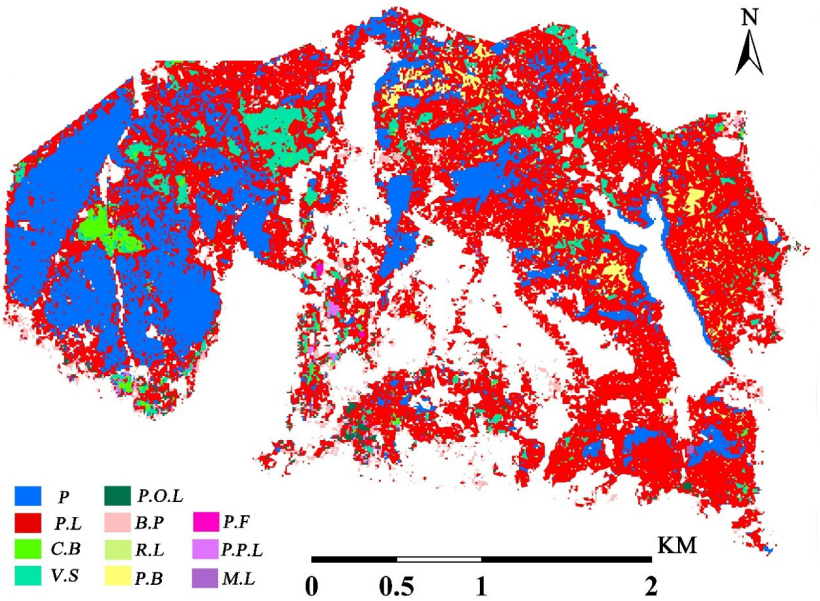

Figure 6. Tree species classification using ML classifier based on dataset 1.

Table 7 represented the result of the ML classifier, whose overall accuracy and the Kappa coefficient were 79.3% and 0.753, respectively. Only 11 tree species were successfully classified in 13 tree species when using ML classifier, excepting the C.D and P.DC. The producer's accuracies ranged from 65.36% to 88.81%, and the user's accuracy ranged from 62.08% to 90.52% (Table 7). The classification of P represented the highest producer's accuracy (88.81%), while the lowest value (65.36%) was observed in the classification P.L. The lowest and highest user's accuracies appeared in the classification of C.B (62.08%) and P (90.52%) respectively. Some pixels of other tree species such as C.B, V.S, P. B, and B.P were misclassified into P.L class, leading to the low producer's accuracy of P.L. Many pixels of C.B were misclassified into P, P.L, V.S, and P.B class. Similar to the results of the SVM classifier, the user accuracy of R.L was low, which may have been caused by insufficient test samples. Moreover, some pixels were misclassified as V.S, but belonged to C.B, P.F, and P.B.

Table 7. Confusion matrices for ML classifier using classification dataset 1

Class Name	P	P.L	C.B	V.S	P.B	P.O.L	B.P	R.L	P.S	P.P.L	M.L	Total	User's Accuracy (%)
P	611	51	13	0	0	0	0	0	0	0	0	675	90.52
P.L	63	349	16	0	12	26	36	3	0	0	0	505	69.11
C.B	12	38	167	21	31	0	0	0	0	0	0	269	62.08
V.S	0	30	12	297	0	0	0	0	1	2	0	342	86.84
P.B	0	32	13	15	226	0	0	0	0	0	0	286	79.02
P.O.L	0	0	0	22	0	155	12	0	0	0	0	189	82.01
B.P	0	14	0	0	0	8	153	0	0	0	0	175	87.43
R.L	2	6	0	0	0	3	2	22	0	0	0	35	62.86
P.S	0	6	0	10	0	0	0	0	81	16	0	113	71.68
P.P.L	0	0	0	0	0	12	0	0	16	93	6	127	73.23
M.L	0	8	0	0	0	0	0	0	0	0	29	37	78.38
Total	688	534	221	365	269	204	203	25	98	111	35		
Producer's Accuracy(%)	88.81	65.36	75.57	81.37	84.01	75.98	75.37	88.00	82.65	83.78	82.86		
Overall Accuracy(%)	79.3												
Kappa coefficient	0.753												

DISCUSSION

Comparison of different classifiers: Comparing the accuracy of different classifiers, the machine learning classifier was determined to have a classification-significant advantage over physically-based spectral classifiers. The results revealed that the combination of retrieved vegetation indices with selected spectral features (classification dataset 3) yielded the results of the highest accuracy utilizing the RF classifier. The RF classifier had an advantage over SVM when using the higher level of data dimensionality samples for classification. This result proved that the discrimination of tree species was aided by utilizing the selection of characteristic bands and incorporating vegetation indices.

Moreover, compared to using only hyperspectral bands, the combined datasets can enhance the discrimination of some tree species and efficiency. Although classification dataset 2 contains the most beneficial information, it also increases the dimension of the data. However, the increased data dimension does not lead to the decrease of the accuracy for the machine learning classifier. In contrast to the physically-based spectral classifier, RF and SVM classifiers can effectively avoid the curse of dimensionality ("Hughes" phenomenon) caused by increasing the use of variables (Oommen et al., 2008). Based on the confusion matrixes of RF and SVM classifiers, there are a large number of 0 values. A large quantity of 0 within the confusion matrix indicates that the test sample size is inadequate or the classification is highly successful. Considering the number of 0 values on the confusion matrix for some smaller classes (P.DC, R.L, B.P, P.O.L), the insufficient test samples were responsible for the high classification accuracies of some species.

We also examined the accuracy of physically-based spectral classifiers to identify tree species. In contrast, the physically-based spectral classifier resulted in the overall accuracy and Kappa coefficient dropping by at least 6% and 10% respectively from the previous two machine Learning algorithms classifications.The best performance of SAM classifier was obtained by classification dataset 1 only. However, SAM classifier is not sensitive to local features and radiation intensity. Thus, it is susceptible to noise and decreases classification accuracy. Although averaging was applied to the spectrum before classification, it still leads to the spectral overlap resulting in the misclassification of species through the SAM classifier. Maximum Likelihood classifier relies on statistical analysis of spectral features, which necessitates a larger number of training samples compared to machine learning classifier. Despite augmenting the training samples using ML classifier for tree species identification, it presents the lowest overall accuracy and kappa coefficient compared to the results of other classifiers. Limited spectral information obtained from a small number of patches may have contributed to the ML classifier inability to identify C.D and P.DC, as well as its misclassification of C.B as P, P.L, V.S, and P.B. Moreover, apart from dataset 1, the other classification datasets lack continuous spectral information and cannot be transformed into the same spectral scale space as the original hyperspectral data. Therefore, physically-based spectral classifiers are unable to utilize the other combined datasets except for dataset 1.

In terms of tree species classification, the P species exhibited the highest overall producer's and user's accuracies across all four classifiers for using each classification dataset. However, the ML classifier failed to identify C.D and P.DC, possibly due to limited spectral features available in the testing samples. Contrary, the limited training samples also led to very successful results such as M.L and R.L using machine learning classifiers. Moreover, the results showed that the misclassification rates within the same types were higher than those between the different types such as P and P.L. Some broadleaf tree species (C.B, V.S, P.B, P.O.L) were misclassified as other broadleaf species, which may be owing to the similar spectral signatures among the broadleaf species. These phenomena exist in the results which misclassified using a physically-based spectral classifiers.

In this study, the results demonstrated that machine learning classifiers have an advantage in tree species classification. The unique advantage of the SVM classifier is to cope with a small sample, nonlinear, and high-dimensional data problems (Pal & Mather, 2005). RF classifier can handle numerous input variables and achieve higher classification accuracy for sample sets with more categories. Therefore, the used vegetation indices were associated with the chlorophyll concentrations, anthocyanin, either directly or as a measure of the red edge of the vegetation reflectance spectrum. All the incorporation of these features can enhance the characteristics of tree species for classification.

CONCLUSION

In this study, we utilized satellite hyperspectral and field survey data to discriminate and map tree species in a northern subtropical semi-deciduous forest. Four classification datasets, which consisted of 32, 44, 38 and 22 variables, were adjusted including different spectral information, training sample types to examine classifier performance under these different conditions. The following results were achieved:

(1) Satellite hyperspectral images integrated with a small number of samples have strong application potential in classifying and identifying the forest tree species, mapping individual species in a complex and heterogeneous forest on challenging topography.

(2) The discrimination of tree species was enhanced by utilizing the selection of characteristic bands and the introduction of vegetation indices. The machine learning classifiers (SVM and RF classifiers) had a statistical advantage over physically-based spectral classifiers (SAM and ML classifier).

(3) In the context of a complex forest with varying class sizes, SVM classifier accuracy had a statistical advantage over the RF classifier by adding variables based on how each classifier handles small classes. RF classifier performed better than the SVM classifier after removing the redundancy of spectral data in training samples.

REFERENCES

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: a review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24-31. http://dx.doi.org/10.1016/j.isprsjprs.2016.01.011.

Berk, A., Bernstein, L. S., & Robertson, D. C. (1987). *MODTRAN: a moderate resolution model for LOWTRAN*. Burlington: Air Force Geophysics Laboratory/Air Force Systems Command/United Sates Air Force/Hanscom Air Force Base.

Boardman, J. W., & Kruse, F. A. (1994) Automated spectral analysis: a geological example using AVIRIS data, North Grapevine Mountains, Nevada. In *Proceedings, Tenth Thematic Conference on Geologic Remote Sensing* (pp. I-407-I-418). Ann Arbor, United States: Environmental Research Institute of Michigan.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. http://dx.doi.org/10.1023/A:1010933404324.

Congalton, R. G., & Mead, R. A. (1983). A quantitative method to test for consistency and correctness in photointerpretation. *Photogrammetric Engineering and Remote Sensing*, *49*(1), 69-74.

Dalponte, M., Bruzzone, L., & Gianelle, D. (2012). Tree species classification in the Southern Alps based on the fusion of very high geometrical resolution multispectral/hyperspectral images and LiDAR data. *Remote Sensing of Environment*, *123*, 258-270. http://dx.doi.org/10.1016/j.rse.2012.03.013.

Dalponte, M., Bruzzone, L., Vescovo, L., & Gianelle, D. (2009). The role of spectral resolution and classifier complexity in the analysis of hyperspectral images of forest areas. *Remote Sensing of Environment*, *113*(11), 2345-2355. http://dx.doi.org/10.1016/j.rse.2009.06.013.

Fassnacht, F. E., Neumann, C., Förster, M., Buddenbaum, H., Ghosh, A., Clasen, A., Joshi, P. K., & Koch, B. (2014). Comparison of feature reduction algorithms for classifying tree species with hyperspectral data on three central European test sites. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *7*(6), 2547-2561. http://dx.doi.org/10.1109/JSTARS.2014.2329390.

Ferreira, M. P., Wagner, F. H., Aragão, L. E., Shimabukuro, Y. E., & Souza Filho, C. R. (2019). Tree species classification in tropical forests using visible to shortwave infrared WorldView-3 images and texture analysis. *ISPRS Journal of Photogrammetry and Remote Sensing*, *149*, 119-131. http://dx.doi.org/10.1016/j.isprsjprs.2019.01.019.

Ghosh, A., Fassnacht, F. E., Joshi, P. K., & Koch, B. (2014). A framework for mapping tree species combining hyperspectral and LiDAR data: role of selected classifiers and sensor across three spatial scales. *International Journal of Applied Earth Observation and Geoinformation*, *26*, 49-63. http://dx.doi.org/10.1016/j.jag.2013.05.017.

Green, A. A., Berman, M., Switzer, P., & Craig, M. D. (1988). A transformation for ordering multispectral data in terms of image quality with implications for noise removal. *IEEE Transactions on Geoscience and Remote Sensing*, *26*(1), 65-74. http://dx.doi.org/10.1109/36.3001.

Heinzel, J., & Koch, B. (2012). Investigating multiple data sources for tree species classification in temperate forest and use for single tree delineation. *International Journal of Applied Earth Observation and Geoinformation*, *18*, 101-110. http://dx.doi.org/10.1016/j.jag.2012.01.025.

Hughes, G. (1968). On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, *14*(1), 55-63. http://dx.doi.org/10.1109/TIT.1968.1054102.

Jones, T. G., Coops, N. C., & Sharma, T. (2010). Assessing the utility of airborne hyperspectral and LiDAR data for species distribution mapping in the coastal Pacific Northwest, Canada. *Remote Sensing of Environment*, *114*(12), 2841-2852. http://dx.doi.org/10.1016/j.rse.2010.07.002.

Koetz, B., Morsdorf, F., Van der Linden, S., Curt, T., & Allgöwer, B. (2008). Multi-source land cover classification for forest fire management based on imaging spectrometry and LiDAR data. *Forest Ecology and Management*, *256*(3), 263-271. http://dx.doi.org/10.1016/j.foreco.2008.04.025.

Kruse, F. A., Lefkoff, A. B., Boardman, J. B., Heidebrecht, H. K. B., Shapiro, A. T., Barloon, P. J., & Goetz, A. F. H. (1993). The spectral image processing system (SIPS)-interactive visualization and analysis of imaging spectrometer data. *Remote Sensing of Environment*, *44*(2-3), 145-163. http://dx.doi.org/10.1016/0034-4257(93)90013-N.

Madroñal, D., Lazcano, R., Salvador, R., Fabelo, H., Ortega, S., Callico, G. M., Juarez, E., & Sanz, C. (2017). SVM-based real-time hyperspectral image classifier on a manycore architecture. *Journal of Systems Architecture*, *80*, 30-40. http://dx.doi.org/10.1016/j.sysarc.2017.08.002.

Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: a review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *66*(3), 247-259. http://dx.doi.org/10.1016/j.isprsjprs.2010.11.001.

Oommen, T., Misra, D., Twarakavi, N. K., Prakash, A., Sahoo, B., & Bandopadhyay, S. (2008). An objective analysis of support vector machine based classification for remote sensing. *Mathematical Geosciences*, *40*(4), 409-424. http://dx.doi.org/10.1007/s11004-008-9156-6.

Pal, M., & Mather, P. M. (2005). Support vector machines for classification in remote sensing. *International Journal of Remote Sensing*, *26*(5), 1007-1011. http://dx.doi.org/10.1080/01431160512331314083.

Peerbhay, K. Y., Mutanga, O., & Ismail, R. (2013). Commercial tree species discrimination using airborne AISA Eagle hyperspectral imagery and partial least squares discriminant analysis (PLS-DA) in KwaZulu–Natal, South Africa. *ISPRS Journal of Photogrammetry and Remote Sensing*, *79*, 19-28. http://dx.doi.org/10.1016/j.isprsjprs.2013.01.013.

Richards, J. A., & Jia, X. (1999). Feature Reduction. In J. A. Richards (Ed.), *Remote sensing digital image analysis* (pp. 238-140). Berlin: Springer. http://dx.doi.org/10.1007/978-3-662-03978-6_10

Richter, R., Reu, B., Wirth, C., Doktor, D., & Vohland, M. (2016). The use of airborne hyperspectral data for tree species classification in a species-rich Central European forest area. *International Journal of Applied Earth Observation and Geoinformation*, *52*, 464-474. http://dx.doi.org/10.1016/j.jag.2016.07.018.

Rodríguez-González, P. M., Albuquerque, A., Martínez-Almarza, M., & Díaz-Delgado, R. (2017). Long-term monitoring for conservation management: lessons from a case study integrating remote sensing and field approaches in flood plain forests. *Journal of Environmental Management*, *202*(Pt 2), 392-402. PMid:28190693. http://dx.doi.org/10.1016/j.jenvman.2017.01.067.

Savitzky, A., & Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, *36*(8), 1627-1639. http://dx.doi.org/10.1021/ac60214a047.

Schepaschenko, D., See, L., Lesiv, M., McCallum, I., Fritz, S., Salk, C., Moltchanova, E., Perger, C., Shchepashchenko, M., Shvidenko, A., Kovalevskyi, S., Gilitukha, D., Albrecht, F., Kraxner, F., Bun, A., Maksyutov, S., Sokolov, A., Dürauer, M., Obersteiner, M., Karminov, V., & Ontikov, P. (2015). Development of a global hybrid forest mask through the synergy of remote sensing, crowdsourcing and FAO statistics. *Remote Sensing of Environment*, *162*, 208-220. http://dx.doi.org/10.1016/j.rse.2015.02.011.

Tarabalka, Y., Chanussot, J., & Benediktsson, J. A. (2010). Segmentation and classification of hyperspectral images using watershed transformation. *Pattern Recognition*, *43*(7), 2367-2379. http://dx.doi.org/10.1016/j.patcog.2010.01.016.

Waser, L. T., Klonus, S., Ehlers, M., Küchler, M., & Jung, A. (2010). Potential of digital sensors for land cover and tree species classifications-a case study in the framework of the DGPF-project. *Photogrammetrie, Fernerkundung, Geoinformation*, *2010*(2), 141-156. http://dx.doi.org/10.1127/1432-8364/2010/0046.

Zhang, C., & Xie, Z. (2012). Combining object-based texture measures with a neural network for vegetation mapping in the Everglades from hyperspectral imagery. *Remote Sensing of Environment*, *124*, 310-320. http://dx.doi.org/10.1016/j.rse.2012.05.015.

Author contributions: NW: conceptualization, investigation, methodology, project administration; GW: supervision, writing – review & editing, data curation.