Environmental Conservation

cambridge.org/enc

A machine learning approach to mapping canopy gaps in an indigenous tropical submontane forest using WorldView-3 multispectral satellite imagery

Colbert M Jackson and Elhadi Adam

School of Geography, Archaeology and Environmental Studies, University of the Witwatersrand, Johannesburg 2050, South Africa

Summary

Selective logging in tropical forests may lead to deforestation and forest degradation, so accurate mapping of it will assist in forest restoration, among other ecological applications. This study aimed to track canopy tree loss due to illegal logging of the important hardwood tree *Ocotea usambarensis* in a closed-canopy submontane tropical forest by evaluating the mapping potential of the very-high-resolution WorldView-3 multispectral dataset using random forest (RF) and support vector machine (SVM) with radial basis function kernel classifiers. The results show average overall accuracies of 92.3 ± 2.6% and 94.0 ± 2.1% for the RF and SVM models, respectively. Average kappa coefficients were 0.88 ± 0.03 for RF and 0.90 ± 0.02 for SVM. The user's and producer's accuracies for both classifiers were in the range of 84–100%. This study further indicates that vegetation indices derived from bands 5 and 6 helped detect canopy gaps in the study area. Both variable importance measurement in the RF algorithm and pairwise feature selection proved useful in identifying the most pertinent variables in the classification of canopy gaps. These findings could allow forest managers to improve methods of detecting canopy gaps at larger scales using remote sensing data and relatively little additional fieldwork.

Research Paper

Cite this article: Jackson CM and Adam E (2022) A machine learning approach to mapping canopy gaps in an indigenous tropical submontane forest using WorldView-3 multispectral satellite imagery. *Environmental Conservation* **49**: 255–262. doi: 10.1017/S0376892922000339

Received: 8 February 2022
Revised: 4 August 2022
Accepted: 6 August 2022
First published online: 30 August 2022

Keywords:
canopy gaps; machine learning algorithm; *Ocotea usambarensis*; selective logging; tropical forests; WorldView-3

Author for correspondence:
Colbert M Jackson,
Email: mutisojackson@yahoo.com

Introduction

Tropical rainforests cover *c.* 7% of the globe, and they contain more than 53 000 tree species compared to *c.* 124 in temperate Europe (Slik et al. 2015). Vital environmental processes such as the water cycle, soil conservation, carbon sequestration and habitat protection are regulated by tropical tree species. Kenya's rainforest cover, mostly montane forests, is fragmented into patches that are being degraded (NEMA 2010). The Mount Kenya Forest Reserve (MKFR) sustains a variety of biodiversity and affords vital ecosystem services, as well as being a major forested water catchment area in Kenya (NEMA 2010). Intense growth of the human population around the MKFR and increasing poverty levels have led to forest degradation there due to illegal logging of important timber trees, especially *Ocotea usambarensis*, which is a hardwood tree with excellent decay and insect resistance. In 2000–2001, *O. usambarensis* was the most highly priced timber in Kenya. The State of the Environment report by the National Environment Management Authority (NEMA) listed *O. usambarensis* as endangered (NEMA 2010). Normally, *Ocotea* trees mature over 60–70 years and are relatively large with spreading crowns and stem diameters in the range 3.75–9.50 m. The tree has low seed viability and poor regeneration, produces seeds only every 10 years and seed germination is sporadic.

Tracking of selective logging (SL) in tropical forests is important due to its effects on biodiversity and other forest attributes, including ecosystem services, the microclimate and carbon pools (Dalagnol et al. 2019). Landscape-level spatial assessments of canopy gaps have primarily used ground-based methods, and, due to the amount of effort and expense needed to acquire the measurements, the areas covered by these surveys are small and not contiguous spatially. Remote sensing (RS) is a more cost-effective and less laborious method for modelling canopy gaps than using ground-based methods (Malahlela et al. 2014). However, SL can be challenging to detect and quantify because of its partial disturbance of the forest canopy and small scale of impact (Dalagnol et al. 2019). Three main types of very-high-resolution (VHR) earth observation data have been used to detect canopy gaps due to SL in tropical forests (i.e., optical, LiDAR (light detection and ranging) and radar (radio detection and ranging)). LiDAR technology has made the detection of small canopy gaps possible; for example, Asner et al. (2013) explored canopy height models from a single LiDAR data acquisition, while Andersen et al. (2014) used simple differencing of LiDAR canopy height models to detect disappearing tree crowns with exceptional accuracy. Ellis et al. (2016) used LiDAR data from a single acquisition to detect SL by estimating aboveground biomass, while Rex et al. (2020) used LiDAR data acquired before and after logging to estimate the change in aboveground biomass due to logging. However, LiDAR covers relatively small spatial extents and data acquisition costs are high. Therefore, researchers may opt to integrate data from different RS systems; for example, Dalagnol et al. (2019) combined airborne LiDAR and VHR satellite data to quantitatively assess and validate canopy gaps. Automated mapping using time-series approaches applied to calibrated synthetic aperture radar (SAR) data have been successful in detecting SL (Baldauf & Köhl 2009). Hethcoat et al. (2021) assessed the effectiveness of SAR data for monitoring tropical SL, but SAR-based biomass estimates have lower precision at the same resolution than optical data. Traditional aerial photography was successfully used for mapping canopy gaps before the introduction of high-resolution optical data (Malahlela et al. 2014); however, aerial photography data acquisition is cost-intensive, although technological advances have revitalized the use of aerial photography through unmanned aerial vehicles (UAVs). Spaias et al. (2016) used a hyperspectral camera on a UAV to detect and quantify small-scale canopy gaps in a tropical forest, although the amount of spatial and spectral data gained made the data processing computationally demanding, especially where cloud-computing resources were lacking. Ota et al. (2019) used digital aerial photographs acquired before and after logging to estimate the change in aboveground biomass linked to SL, while Kamarulzaman et al. (2022) used UAV data to detect forest canopy gaps attributed to SL. However, digital aerial photographs acquired using UAVs cover relatively small spatial extents.

Machine learning (ML) for the classification of RS data has been applied in mapping SL with increasing success. Dalagnol et al. (2019) used a random forest (RF) model to detect tree loss with an average precision of 64%. Hethcoat et al. (2019) reported a detection rate of logged pixels of *c.* 90% using RF. Kamarulzaman et al. (2022) compared conventional and ML classifiers. The support vector machine (SVM) and artificial neural network (ANN) classifiers attained higher overall accuracy of 85%. Using ordinary least squares regression and ML approaches (RF, *k*-nearest neighbour, SVM and ANN), Rex et al. (2020) monitored the change in aboveground biomass due to SL. Hethcoat et al. (2020, 2021) developed RF models and used logging records to detect SL. Therefore, RF, SVM and ANN approaches, which have superior image handling capabilities, have been mostly used in this area.

The development of VHR multispectral sensors such as WorldView-2 is critical for discriminating between tree canopies and vegetated gaps (Malahlela et al. 2014), because some of the inherent features of hyperspectral data, such as carotenoids and chlorophyll-sensitive bands, are preserved in WorldView-2/3 multispectral data (Mutanga et al. 2012). Visual interpretation of VHR multidate satellite imagery represents a promising way to detect canopy gaps with fairly low uncertainty (Dalagnol et al. 2019). Nonetheless, spatially accurate tree-scale validation data are not readily available, so automated approaches using VHR satellite data to accurately map canopy gaps over large and remote areas are not readily available (Dalagnol et al. 2019). The primary focus of this study is exploring the potential of WorldView-3 imagery to develop a SL monitoring system capable of detecting canopy gaps over large spatial extents in a closed-canopy submontane tropical forest using RF and SVM models.

Materials and methods

Study area

This study was conducted in a *c.* 264ha area in the MKFR (Fig. 1), which covers *c.* 213 083 ha and encircles the 7150-ha Mount Kenya National Park (MKNP), which begins at 3100 m and extends to the highest point, the Batiaan Peak, at 5199 m (Lange & Bussmann 1998). The phonolites from main volcanic events *c.* 2 million years ago form the bedrock of the study area, but the inorganic body of the soils originates from a later coverage of volcanic ashes and pyroclastic rocks (Lange & Bussmann 1998). The mountain lies on the equator (latitude 0°10´S, longitude 37°20´E) and forms one of the most pristine mountain ecosystems globally and a remarkable landscapes due to its peaks with rugged glacier-clad summits and diverse forests. The precipitation pattern consists of long rains from March to May and short rains from October to December (Supplementary Fig. S1, available online). The mountain shows a marked vegetational gradient dictated by altitude and rainfall amount. The lower tree line of the forest belt is due to agricultural and pastoral activities. *Ocotea usambarensis*, which never constitutes pure stands and prefers humid Nitisols and Acrisols, forms the evergreen submontane forests on the southern, south-eastern and eastern slopes of Mount Kenya between 1500 and 2500 m altitude (Lange & Bussmann 1998).

Acquisition and pre-processing of satellite data

WorldView-3 data were acquired on 15 September 2019 for detecting canopy gaps in selectively logged sites. WorldView-2 imagery acquired on 30 January 2014 and Google Earth were used for historical comparison. In order to cancel out the haze component caused by additive scattering from the RS data, the dark object subtraction method was applied (Chavez 1988). The WorldView-2 satellite captures panchromatic images (450–800 nm) with a spatial resolution of 0.46 m and multispectral images with eight visible–near-infrared (VNIR) bands (400–1040 nm) at 1.84m resolution, while WorldView-3 acquires panchromatic images with a spatial resolution of 0.30 m, multispectral imagery with eight VNIR bands at 1.2 m and eight shortwave-infrared (SWIR) bands (1195–2365 nm) at 3.7m spatial resolution. Additionally, there are 12 clouds, aerosols, vapours, ice and snow (CAVIS) bands with a spatial resolution of 30 m. Only WorldView-3 VNIR bands were used because SWIR bands covering the study area exhibited extensive cloud cover. The satellite data were pansharpened to obtain new bands with a spectral resolution of the multispectral bands and a spatial resolution of the panchromatic band.

Acquisition of field data

Ground truth points were collected in February 2020 using a handheld Global Positioning System (eTrex® 20 GPS Receiver; Garmin, Olathe, KS, USA) and a pansharpened WorldView-3 image (pixel = 0.30 m). Seventy vegetated gaps formed after illegal logging of *Ocotea* trees were located in the field (Fig. 1). In the WorldView-3 imagery, the canopy gaps were either partially/fully illuminated or not illuminated at all (Fig. S2b & c). Since the human-made canopy gaps shared similar reflectance characteristics with natural canopy gaps, GPS coordinates of 301 vegetated gaps and 301 shaded gaps were collected, including the 70 canopy gaps formed after illegal logging of *Ocotea* trees. A vegetated gap is a forest canopy gap with low vegetation inside it, which is the initial stage of vegetation recovery from forest disturbance. Coordinates of the approximate locations of gap centres were recorded and then overlaid on the pansharpened WorldView-3 image. Using a geographic information system (GIS; *ArcGIS*® v. 10.3; ESRI, Redlands, CA, USA), points were set on the vegetated and shaded canopy gap pixels on the WorldView-3 imagery, and by following the edges of the pixels, the points were made into polygons.

https://doi.org/10.1017/S0376892922000339 Published online by Cambridge University Press

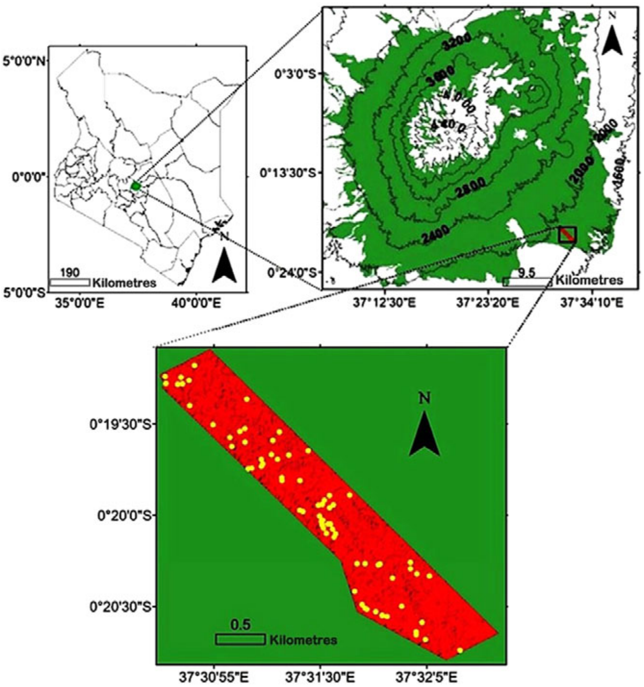

Fig. 1. Maps of the study area in the Mount Kenya Forest Reserve showing the locations of selectively logged *Ocotea* trees (dots).

Locations of closed forest canopy could be identified using the WorldView-3 imagery; thus, a total of 301 polygons were extracted. Canopy gaps formed after the logging of *Ocotea* trees had their dimensions collected in the field, such as their dripline measurements, maximum length and compass orientation, together with the maximum width perpendicular to the length. Using the GIS, a map of canopy gaps was generated from ground survey data. The accuracy of canopy gap delineation using RS was assessed by comparing them with the dimensions collected from the field. The ground reference data were then randomly split into train and test datasets of 70% and 30%, respectively.

Feature extraction and selection

In minimizing the effects of data saturation when mapping canopy gaps in dense forests, methods such as vegetation indices (VIs), image transform algorithms, texture measures and spectral mixture analysis have been utilized previously (Malahlela et al. 2014, Dalagnol et al. 2019). Table 1 presents 55 features (23 means, 23 standard deviations (SDs), 8 ratios and 1 brightness feature) extracted from pansharpened WorldView-3 imagery, and these features are subsequently referred to as variables in the analyses. The VIs were chosen from those sensitive to greenness and plant senescence (Malahlela et al. 2014). To accurately extract shaded gaps, the shadow detection index (SDI) of Shahi et al. (2014) was used in the modelling. Shade is associated with canopy gaps as nearby trees appear on the edges of gaps (Dalagnol et al. 2019). In reducing the redundancy and intercorrelation among the list of potential features, a subset of the best-performing features was extracted from the initial 55 features before classification; for these, approximate optimal thresholds were determined from the reference data using histograms and then adjusted to determine thresholds that resulted in the highest matching accuracy compared to the reference data.

Spectral separability

Spectral separability measures the distance between two signatures, and the separability between any combinations of variables can be used in the classification. Only the subset from the best-performing features was used in the analysis. The digitized vegetated gaps, shaded gaps and forest canopy samples (Fig. S3) were assigned spectral information using the mean pixel values within their polygons and then used to generate respective signature files. The transformed divergence (TD) index and Jeffries–Matusita (J–M) distance separability measures were used. Divergence (D) is calculated from the mean and variance–covariance matrices of the data representing feature classes (Kavzoglu & Mather 2000):

$$D_{ij} = \frac{1}{2} tr\left[\left(\sum_i - \sum_j\right)\left(\sum_j^{-1} - \sum_i^{-1}\right)\right] + \frac{1}{2} tr\left[\left(\sum_i^{-1} + \sum_j^{-1}\right)(\mu_i - \mu_j)(\mu_i - \mu_j)^T\right] \quad (1)$$

The TD is introduced to reduce the impact of well-separated classes that may raise the average divergence value and make the divergence measure misleading (Kavzoglu & Mather 2000):

$$TD_{ij} = c\left[1 - e^{\frac{-D_{ij}}{8}}\right], \quad (2)$$

where $tr[\cdot]$ is the trace of a matrix, which is the sum total of the diagonal elements of the matrix, and Σ_i and Σ_j are the variance–covariance matrices of classes i and j; μ_i and μ_j are the corresponding mean vectors; c is a constant value defining the range of TD values.

The J–M distance between distributions of two classes ω_i and ω_j has been defined as follows (Richards & Jia 1999):

$$J\text{–}M_{ij} = 2\left(1 - e^{-B_{ij}}\right), \quad (3)$$

where B_{ij} is the Bhattacharyya distance (Kailath 1967) computed as:

$$B_{ij} = \frac{1}{8}(\mu_i - \mu_j)^T \left(\frac{\sum_i + \sum_j}{2}\right)^{-1} (\mu_i - \mu_j) + \frac{1}{2} ln\left(\frac{1}{2}\frac{\left|\sum_i + \sum_j\right|}{\sqrt{\left|\sum_i\right|\left|\sum_j\right|}}\right), \quad (4)$$

where μ_i and μ_j are the mean reflectances of species i and j, Σ_i and Σ_j correspond to their covariance matrices, with $|\Sigma_i|$ and $|\Sigma_j|$ being the determinants of Σ_i and Σ_j, respectively, ln is the natural logarithm function and T is the transposition function. The J–M distance improves the Bhattacharya distance by normalizing it to between 0 and 2.

Pairwise feature comparison

A correlation matrix was computed from the reference samples (Table 2). Similarity scores were calculated between each pair of variables to determine whether or not two variables were co-referent. Pairwise comparisons are in form of a matrix: $\boldsymbol{C} = [c_{kp}]_{n \times n}$, where c_{kp} is the pairwise comparison rating for kth and pth criteria. The matrix $\boldsymbol{C}$ is reciprocal; that is, $c_{pk} = c_{kp}^{-1}$, and all of its diagonal elements are unity; that is, $c_{kp} = 1$, for $k = p$ (Malczewski 2016).

https://doi.org/10.1017/S0376892922000339 Published online by Cambridge University Press

Table 1. List of features used for detecting canopy gaps in the Mount Kenya Forest Reserve: 23 means (of the 15 vegetation indices and 8 visible–near-infrared bands), 23 standard deviations (of the 15 vegetation indices and 8 visible–near-infrared bands), 8 ratios (of the 8 visible–near-infrared bands) and 1 brightness feature (average of the means of bands 1–8).

Feature	Equation/description	Reference
Bright	Brightness, average of means of bands 1–8	–
Band 1–8	Means of pansharpened WorldView-3 bands 1–8	–
SD 1–8	Standard deviations of WorldView-3 bands 1–8	–
Ratio 1–8	*i*th band mean divided by sum of band 1–8 means	–
Red-edge position index (REPI)	Maximum first derivative: 680–750 nm	Horler et al. (1983)
Chlorophyll absorption ratio index (CARI)	$[(\rho\ 700 - \rho\ 672) - 0.2 \times (\rho\ 700 - \rho\ 553)]$	Kim (1994)
Structurally insensitive pigment index (SIPI)	$(\rho 802 - \rho 465)/(\rho 802 + \rho 681)$	Peñuelas et al. (1995)
Normalized pigment chlorophyll index (NPCI)	$(\rho 660 - \rho 425)/(\rho 660 + \rho 425)$	Peñuelas et al. (1995)
Red-edge normalized difference vegetation index (RENDVI)	$(\rho 753 - \rho 700)/(\rho 753 + \rho 700)$	Gitelson et al. (1996)
Photochemical reflectance index (PRI)	$(\rho\ 534 - \rho\ 572)/(\rho\ 534 + \rho\ 572)$	Gamon et al. (1997)
Pigment-sensitive normalized difference (PSND)	$(\rho 802 - \rho 676)/(\rho 800 + \rho 676)$	Blackburn (1998)
Plant senescence reflectance index (PSRI)	$(\rho 681 - \rho 502)/\rho 753$	Merzlyak et al. (1999)
Simple ratio red/green (SRred/green)	RED/GREEN (mean reflectance of red bands (600–699 nm) and green bands (500–599 nm), respectively)	Gamon and Surfus (1999)
Modified chlorophyll absorption ratio index (MCARI)	$[(\rho 700 - \rho 672) - 0.2 \times (\rho 700 - \rho 553)] \times (\rho 700/\rho 672)$	Daughtry et al. (2000)
Anthocyanin reflectance index (ARI)	$(1/553) - (1/\rho 700)$	Gitelson et al. (2001)
Visible atmospherically resistant index (VARI)	$(\rho 557 - \rho 643)/(\rho 557 + \rho 643 - \rho 465)$	Gitelson et al. (2002a, 2002b)
Carotenoid reflectance index 1 (CRI-1)	$(1/\rho 511) - (1/\rho 553)$	Gitelson et al. (2002b)
Carotenoid reflectance index 2 (CRI-2)	$(1/\rho 511) - (1/\rho 700)$	Gitelson et al. (2002b)
Shadow detection index (SDI)	$[(\rho 865 - \rho 480)/(\rho 865 + \rho 480)] - (\rho 835)$	Shahi et al. (2014)

RF and SVM models

The results in this study were obtained by training the RF and SVM models in *R* software (R Core Team, Vienna, Austria). In RF, to differentiate between predefined categories, decision trees recursively partition the source set into subsets with bagged samples by univariate splits at internal nodes (Breiman 2001). Before running the model, the number of decision trees (*ntree*) and the number of predictor variables (*mtry*) randomly selected at each node are defined. The RF model aggregates predictions from all decision trees, then the majority vote of all trees assigns a final class for unknown features (Breiman 2001). Using the grid search method, the mean decrease in accuracy (MDA) was used to extract a subset of the best-performing variables (Breiman 2001). The MDA shows how much accuracy the model losses by excluding each variable; therefore, the higher the MDA value, the more important the variable in the model.

The SVM assigns a class from one of the two possible labels when test data are introduced after the training phase (Vapnik 2000). The SVM separates the original data while maximizing the margin between classes and minimizing the misclassification error (Vapnik 2000). An advantage of ML classifiers such as SVM is that they are suited for extreme case binary classification. For any two distinguishable classes with k samples represented by $(x_1, y_1), \ldots, (x_k, y_k)$, where $x \in R^n$ is an n-dimensional space and y is a class label with values of $+1$ or -1, SVM will look for an optimal hyperplane defined by $w = (w_1, \ldots, w_n)$ and b, such that (Huang et al. 2008):

$$y_i = [w \cdot x_i + b]1i = 1, \ldots, k \quad (5)$$

The hyperplane can be located by minimizing the norm of w or the following function under the above inequality constraint (Huang et al. 2008):

$$F(w) = 0.5(w \cdot w) \quad (6)$$

Using kernel functions, SVM applies non-linear decision boundaries and introduces a cost parameter C and gamma parameter γ to quantify the penalty of misclassification errors and to give the curvature weight of the decision boundary, respectively. The robust radial basis function was selected as it has fewer parameter values to predefine. A parameter search must be done to select the best C and γ for a certain classification problem. Therefore, the γ parameter needs to be predefined (Huang et al. 2008):

$$K\left({}_i,\ x_j\right) = e^{-\gamma\left(x_i - x_j\right)^2} \quad (7)$$

The cost parameter C also needs to be predetermined for the canopy gap-mapping problem. A cross-validation quantitative analysis of pairs of values for the C and γ parameters was carried out. The combination of parameters with the lowest error was chosen to train the algorithm. Both RF and SVM classifiers were trained using 70% of the ground reference data, and for robust classification results the ten-fold cross-validation method was repeated ten times.

Measures of model performance

The performance of the RF and SVM models was evaluated using 30% of the ground truth data. Confusion matrices with overall accuracy, kappa coefficient and producer's and user's accuracies were computed and averaged over ten repetitions. Overall accuracy is computed by summing the number of pixels correctly classified divided by the total number of pixels, while producer's accuracy is the percentage of particular classes on the ground that are indicated as such on the classified map (Mutanga et al. 2012). The user's accuracy shows the probability that a pixel indicated as a specific feature is classified as such on the classification map (Mutanga et al. 2012). The kappa coefficient is the difference between the observed accuracy and the agreement that would have been

https://doi.org/10.1017/S0376892922000339 Published online by Cambridge University Press

Table 2. Pairwise correlations between the best-performing variables extracted from the WorldView-3 visible–near-infrared bands, computed from the reference samples. (See Table 1 for acronym definitions.)

	Brightness	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	SD
		B2	CARI	B1	CRI-2	B3	MCARI	B7	B8	NPCI	PSRI	B5	B6	REPI	SDI	B4	ARI
Brightness	–	0.611	0.332	0.531	0.355	0.839	0.004	0.985	0.988	0.320	0.203	0.599	0.992	−0.330	−0.002	0.724	−0.403
Mean B2		–	0.162	0.923	0.298	0.886	0.001	0.485	0.505	0.352	0.459	0.966	0.611	−0.163	−0.002	0.943	−0.437
Mean CARI			–	0.137	0.717	0.262	0.942	0.333	0.331	0.875	0.747	0.158	0.334	−1.000	0.940	0.213	0.564
Mean B1				–	0.290	0.810	0.001	0.405	0.426	0.227	0.446	0.900	0.532	−0.138	−0.002	0.864	−0.411
Mean CRI–2					–	0.378	0.649	0.321	0.327	0.852	0.872	0.400	0.361	−0.718	0.647	0.393	0.110
Mean B3					–	–	0.003	0.745	0.755	0.378	0.396	0.884	0.857	−0.260	−0.002	0.958	−0.460
Mean MCARI							–	0.004	0.004	0.832	0.752	0.002	0.005	−0.943	1.000	0.003	0.725
Mean B7								–	0.991	0.281	0.125	0.469	0.971	−0.330	−0.002	0.605	−0.357
Mean B8									–	0.287	0.138	0.489	0.971	−0.329	−0.002	0.622	−0.365
Mean NPCI										–	0.902	0.405	0.324	−0.877	0.830	0.402	0.350
Mean PSRI											–	0.545	0.208	−0.749	0.752	0.479	0.256
Mean B5												–	0.605	−0.159	−0.001	0.960	−0.464
Mean B6													–	−0.332	−0.002	0.743	−0.402
Mean REPI														–	−0.941	−0.213	−0.564
Mean SDI															–	−0.001	0.727
Mean B4																–	−0.465
SD ARI																	–

expected by chance. The McNemar test was used to compare and indicate the statistical significance of any differences between the two classifiers (Foody & Mathur 2004).

Results

Explanatory power of the variables extracted from the WorldView-3 bands

The most important variables as depicted by the highest values were the brightness feature, the means of the WorldView-3 VNIR bands, the chlorophyll absorption ratio index (CARI), the modified chlorophyll absorption ratio index (MCARI), the carotenoid reflectance index 2 (CRI-2), the normalized pigment chlorophyll index (NPCI), the plant senescence reflectance index (PSRI), the red-edge position index (REPI), the SDI and the SD of the anthocyanin reflectance index (ARI; Fig. 2).

Optimization of the RF and SVM models

The iteration closest to the model mean produced default *mtry* and *ntree* values of 14 and 500, respectively, with an out-of-bag error rate of 0.074 (Fig. S4). Using the same approach, the SVM model produced 0.1 and 10 for gamma and cost, respectively, yielding a cross-validation error of 0.060.

Spectral separability

The mean spectral reflectance curves of the training data were extracted from pixels of the 17 best-performing variables and plotted with their SDs (Fig. S5). Some of the variables, such as the means of bands 1–5, the NPCI, the PSRI and the SD of the ARI exhibited considerable spectral overlaps across the three classes. Only the means of bands 7 and 8 and the SDI helped separate the three classes beyond 1 SD of uncertainty. The brightness feature and the means of band 6, the CARI and the MCARI show considerable overlaps between forest canopy and vegetated gaps. The spectral separability (Table S1) between the forest canopy and vegetated gaps was low in both TD index and J–M distance by the means of the WorldView-3 VNIR bands, but the VIs indicate that the forest canopy and vegetated gaps were clearly separable. The RF model also evaluated the ability of each variable to detect vegetated gaps, shaded gaps and tree crowns, and the mean of band 4 was critical in the identification of the three classes compared to the other variables (Fig. S6). The mean of band 5 was crucial in detecting vegetated gaps, as were the means of the CARI, the CRI-2, the MCARI, the PSRI, the REPI and bands 1, 3 and 6. The mean of band 5 also helped detect shaded gaps. The means of the brightness feature, bands 2 and 7, the SDI, the NPCI and the SD of the ARI helped to detect the forest canopy.

Pairwise feature comparison

Generally, the means of the WorldView-3 VNIR bands were highly correlated, providing redundant information (Table 2). The same applied for the means of the VIs, such as the NPCI, the CARI, the MCARI, the CRI-2 and the PSRI. The highest negative correlations were between the means of the REPI and the CRI-2, the MCARI, the NPCI and the PSRI. High negative correlations were also recorded between the mean of the SDI and those of the CARI and the MCARI. The lowest positive correlation, 0.001, was between the MCARI and bands 1 and 2. The means of the SDI and the MCARI were perfectly positively correlated, while the means of the REPI and the CARI were perfectly

https://doi.org/10.1017/S0376892922000339 Published online by Cambridge University Press

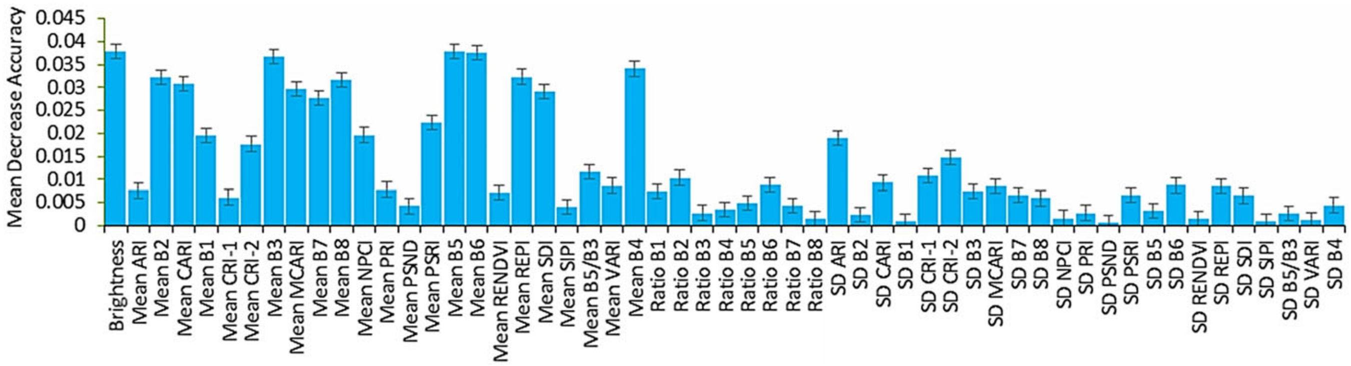

Fig. 2. The relative importance of the variables derived from WorldView-3 visible–near-infrared bands in discriminating vegetated and shaded gaps and forest canopy as measured by random forest classifiers using the mean decrease in accuracy. (See Table 1 for acronym definitions.)

negatively correlated. A high correlation was found between the SD of the ARI and the means of the MCARI and the SDI.

Logging feature detectability

In the WorldView-2 and Google Earth imagery covering the study area, 66 *Ocotea* trees could be identified, and with high confidence; their respective canopy gaps created after these trees were logged were identified in the WorldView-3 imagery. High confidence for gap identification meant a marked change in image pixels. Figure S7 shows the means of the CARI, the REPI and the MCARI extracted from WorldView-2 and WorldView-3. The circles in Fig. S7 compare the VI values before and after the SL events.

Model performance

The average overall accuracies for the RF and SVM models were similar (i.e., 92.3 ± 2.6% and 94.0 ± 2.1%, respectively; Table 3). The average kappa coefficients were 0.88 ± 0.03 for the RF model and 0.90 ± 0.02 for the SVM model. The user's accuracy ranges were 82–100% for the RF model and 86–100% for the SVM model. The producer's accuracy ranges were 83–100% for the RF model and 86–100% for the SVM model. Generally, the shaded gaps class showed the highest user's and producer's accuracies, primarily because the other two classes represent vegetation. Therefore, the non-vegetation class was spectrally distinguished from the vegetation classes. In general, forest canopy had lower user's and producer's accuracies, and this could be attributed to the range of reflectance characteristics of tree crowns in the WorldView-3 imagery.

Classification maps

In the post-processing stage, it was necessary to transform the classification maps so that they had only two classes, namely canopy gaps and forest canopy. As such, the shaded and vegetated gap classes were merged into one 'canopy gaps' class. In general, the classified maps indicated that SL of *Ocotea* trees caused mostly small-scale but spatially widespread disturbances in the MKFR (Fig. 3). The McNemar test returned a Z value of 0.88; thus, there were no significant differences ($Z \geq 1.96$) at the 95% confidence level amongst the confusion matrices of the two classifiers.

Discussion

Using VHR multispectral satellite data, field data and ML algorithms showed great potential for monitoring SL in this tropical forest. Generally, the index means outperformed the SD and ratio variables in terms of the detection of canopy gaps because spectral variability in areas of canopy gaps was only due to shadows from surrounding trees and/or low vegetation. Dalagnol et al. (2019) reported that the most important variables for tree loss detection were the SDs of the reflectance VNIR bands (especially the red band) and the shadow fraction. Therefore, marked increases in spectral variability in tree loss areas were due to shadows cast by nearby trees, the non-photosynthetic vegetation and exposed soil. In Malahlela et al. (2014), the observed improved results were associated with the use of the red-edge band of the WorldView-2 sensor. The current study has identified VIs derived from bands 5 and 6 as crucial in the detection of canopy gaps due to SL. They are therefore transferrable to other tropical, closed-canopy ecosystems with different species compositions where ground-truth data are not available. The WorldView-3 satellite also has a SWIR sensor, which provides rich data for precisely identifying and characterizing forest landscape features, further enhancing WorldView-3's capacity to monitor canopy gaps.

Although variable selection makes modelling simpler and faster to fit and predict, the MDA in the RF model is unable to detect false correlations; therefore, it may be biased because larger values are normally exaggerated and vice versa. Hur et al. (2017) developed an approach to overcome this based on using the Shapley value method on RF regression, but more experiments need to be conducted with other data types in order to confirm this.

Only correlation values >0.70 were significant (Table 2), thus providing redundant information, although ML algorithms can effectively handle this collinearity. In a negative correlation, one variable has the opposite effect compared to that of the other; therefore, the higher the absolute correlation coefficient, the more the variables might be critical during classification.

These results are an indication of the good agreement between the classification of logging of *Ocotea* trees and field data. However, error matrices only estimate the classification accuracy depending on the samples collected from the field; therefore, only biased conclusions can be drawn from such data (Foody & Mathur 2004). Future research will explore other metrics of model performance, such as balanced accuracy, bias score, precision, recall and F-score.

https://doi.org/10.1017/S0376892922000339 Published online by Cambridge University Press

Table 3. Confusion matrices for the random forest and support vector machine classifiers for the respective models whose overall accuracy was closest to the average overall accuracy.

Random forest algorithm						Support vector machine algorithm					
	FC	SG	VG	Total	UA (%)		FC	SG	VG	Total	UA (%)
FC	83	I	13	97	86	FC	81	1	6	88	92
SG	0	89	0	89	100	SG	0	89	0	89	100
VG	7	0	77	84	92	VG	9	0	84	93	90
Total	90	90	90	270		Total	90	90	90	270	
PA (%)	92	99	86			PA (%)	90	99	93		
Overall accuracy = 92.2%; kappa = 0.88						Overall accuracy = 94.1%; kappa = 0.91					

FC = forest canopy; PA = producer's accuracy; SG = shaded gap; UA = user's accuracy; VG = vegetated gap.

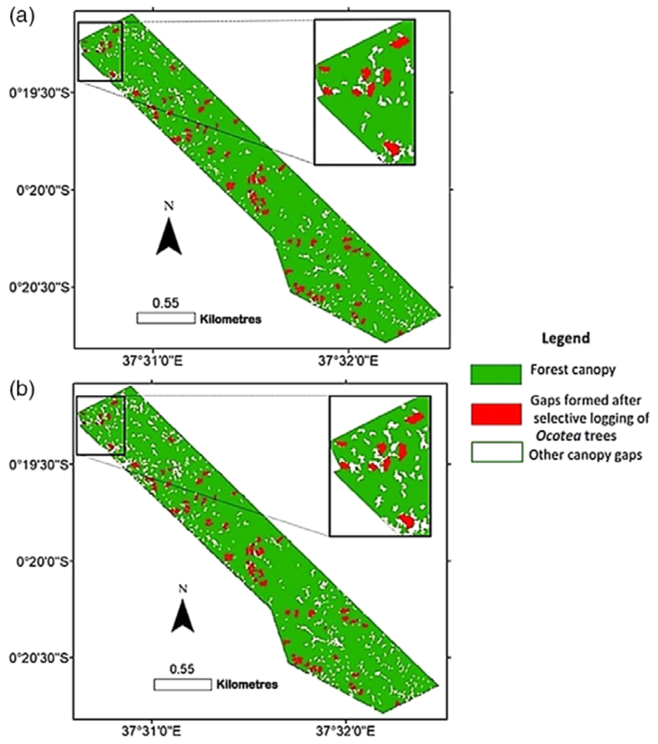

Fig. 3. Final classified maps showing the results of supervised pixel-based classification for (a) the random forest model and (b) the support vector machine model.

The logging of *Ocotea* trees has led to canopy gaps that have stimulated the regrowth of secondary forest dominated by fast-growing species, mostly *Macaranga kilimandscharica* and *Neoboutonia macrocalyx*.

Mapping canopy gaps in tropical forests using optical RS remains a challenge because of persistent cloud cover, further compounded by unreliable cloud and cloud shadow detection algorithms. Other data sources such as SAR, which can capture images of Earth's surface regardless of smoke, darkness or cloud cover, should be explored. Trees that had dropped their leaves or lianas on top of tree crowns that experience sudden dieback may wrongly be interpreted as canopy gaps. Therefore, the application of time series methods that use seasonal models should be explored. The method used in this study might not be applicable in sparser forests because the only notable changes detected here were the disappearances of the shadows of the logged trees. In addition, a tree can be felled in the direction of an existing gap, meaning that the existing gap may undergo only an insignificant increase in size. Researchers should aim to develop accurate methods to detect such canopy gaps.

Nevertheless, data integration, which is heavily dependent on the compatibility of multi-source RS data, specifically the consistency in spatial, spectral, temporal and radiometric resolutions, accompanied by more sophisticated data fusion techniques, is key to the measurement and mapping of forest attributes (Jackson & Adam 2020). Furthermore, LiDAR may serve as a sampling technique when trying to scale up the impacts of SL events to larger regions. In its absence, other sources of publicly available training data can be used. The development of advanced automated methods for processing LiDAR data would lower data-processing costs, allowing for data acquisition over extensive areas (Jackson & Adam 2020). Further studies should make use of UAVs to cover larger areas at reduced costs, with caution taken regarding the current challenges posed by UAVs. Furthermore, airborne/spaceborne hyperspectral imagery covering extensive geographical areas may be obtained for such research through the use of hyperspectral imaging satellites, which are to be launched in the next few years (e.g., NASA's Surface Biology and Geology (SBG) mission and the Carbon Mapper constellation of satellites).

Supplementary material. For supplementary material accompanying this paper visit https://doi.org/10.1017/S0376892922000339.

Acknowledgements. The authors are publishing this research with the help of the South African National Library Information Consortium (SANLiC), which has an agreement with Cambridge University Press (for open access publishing). The University of the Witwatersrand is a participating institution.

Financial support. This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

Competing interests. The authors declare none.

Ethical standards. None.

References

Andersen HE, Reutebuch SE, McGaughey RJ, D'Oliveira MV, Keller M (2014) Monitoring selective logging in western Amazonia with repeat lidar flights. *Remote Sensing of Environment* 151: 157–165.

Asner GP, Kellner JR, Kennedy-Bowdoin T, Knapp DE, Anderson C, Martin RE (2013) Forest canopy gap distributions in the southern Peruvian Amazon. *PLoS ONE* 8: e60875.

Baldauf T, Köhl M (2009) Use of TerraSAR-X for forest degradation mapping in the context of REDD: Presented at: *The World Forestry Congress XIII*, Buenos Aires, Argentina, 23 October.

Blackburn GA (1998) Quantifying chlorophylls and carotenoids from leaf to canopy scale: an evaluation of some hyperspectral approaches. *Remote Sensing of Environment* 66: 273–285.

Breiman L (2001) Random forests. *Machine Learning* 45: 5–32.

Chavez PS Jr (1988) An improved dark-object subtraction technique for atmospheric scattering correction of multispectral data. *Remote Sensing of Environment* 24: 459–479.

Dalagnol R, Phillips OL, Gloor E, Galvao LS, Wagner FH, Locks CJ et al. (2019) Quantifying canopy tree loss and gap recovery in tropical forests under low-intensity logging using VHR satellite imagery and airborne lidar. *Remote Sensing* 11: 1–20.

Daughtry CST, Walthall CL, Kim MS, de Colstoun EB, McMurtrey JE (2000) Estimating corn leaf chlorophyll concentration from leaf and canopy reflectance. *Remote Sensing of Environment* 74: 229–239.

Ellis P, Griscom B, Walker W, Gonçalves F, Cormier T (2016) Mapping selective logging impacts in Borneo with GPS and airborne lidar. *Forest Ecology and Management* 365: 184–196.

Foody GM, Mathur A (2004) Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification. *Remote Sensing of Environment* 93: 107–117.

Gamon JA, Serrano L, Surfus JS (1997) The photochemical reflectance index: an optical indicator of photosynthetic radiation-use efficiency across species, functional types, and nutrient levels. *Oecologia* 112: 492–501.

Gamon JA, Surfus JS (1999) Assessing leaf pigment content and activity with a reflectometer. *The New Phytologist* 143: 105–117.

Gitelson AA, Kaufman YJ, Merzlyak MN (1996) Use of a green channel in remote sensing of global vegetation from EOS-MODIS. *Remote Sensing of Environment* 58: 289–298.

Gitelson AA, Kaufman YJ, Stark R, Rundquist D (2002a) Novel algorithms for remote estimation of vegetation fraction. *Remote Sensing of Environment* 80: 76–87.

Gitelson AA, Merzlyak MN, Chivkunova OB (2001) Optical properties and non-destructive estimation of anthocyanin content in plant leaves. *Photochemistry and Photobiology* 74: 38–45.

Gitelson AA, Zur Y, Chivkunova, OB, Merzlyak MN (2002b) Assessing carotenoid content in plant leaves with reflectance spectroscopy. *Photochemistry and Photobiology* 75: 272–281.

Hethcoat MG, Carreiras JMB, Edwards DP, Bryant RG, Peres CA, Quegan S (2020) Mapping pervasive selective logging in the south-west Brazilian Amazon 2000–2019. *Environmental Research Letters* 15: 094057.

Hethcoat MG, Carreiras JMB, Edwards DP, Bryant RG, Quegan S (2021) Detecting tropical selective logging with C-band SAR data may require a time series approach. *Remote Sensing of Environment* 259: 112411.

Hethcoat MG, Edwards DP, Carreiras JMB, Bryant RG, França FM, Quegan S (2019) A machine learning approach to map tropical selective logging. *Remote Sensing of Environment* 221: 569–582.

Horler DNH, Dockray M, Barber J (1983) The red-edge of plant leaf reflectance. *International Journal of Remote Sensing* 4: 273–288.

Huang C, Song K, Kim S, Townshend JRG, Davis P, Masek JG et al. (2008) Use of a dark object concept and support vector machines to automate forest cover change analysis. *Remote Sensing of Environment* 112: 970–985.

Hur J-H, Ihm S-Y, Park Y-H (2017) A variable impacts measurement in random forest for mobile cloud computing. *Wireless Communications and Mobile Computing* 2007: 6817627.

Jackson CM, Adam E (2020) Remote sensing of selective logging in tropical forests: current state and future directions. *iForest – Biogeosciences and Forestry* 13: 286–300.

Kailath T (1967) The divergence and Bhattacharyya distance measures in signal selection. *IEEE Transactions on Communication Technology* 15: 52–60.

Kamarulzaman A, Wan Mohd Jaafar WS, Abdul Maulud KN, Saad SNM, Omar H, Mohan M (2022) Integrated segmentation approach with machine learning classifier in detecting and mapping post selective logging impacts using UAV imagery. *Forests* 13: 48.

Kavzoglu T, Mather PM (2000) The use of feature selection techniques in the context of artificial neural networks. Presented at: *26th Annual Conference of the Remote Sensing Society*, Leicester, UK, 12–14 September.

Kim MS (1994) The use of narrow spectral bands for improving remote sensing estimation of fractionally absorbed photosynthetically active radiation (fAPAR). MSc thesis. College Park, MD, USA: Department of Geography, University of Maryland.

Lange S, Bussmann R (1998). Ecology and regeneration of the subalpine *Hagenia abyssinica* forests of Mt. Kenya. *Botanica Acta* 110: 473–480.

Malahlela O, Cho MA, Mutanga O (2014) Mapping canopy gaps in an indigenous subtropical coastal forest using high-resolution WorldView-2 data. *International Journal of Remote Sensing* 35: 6397–6417.

Malczewski J (2016) Multicriteria analysis. In: B Huang, TJ Cova (eds), *Comprehensive Geographic Information Systems* (pp. 197–217). Oxford, UK: Elsevier.

Merzlyak MN, Gitelson AA, Chivkunova OB, Rakitin VY (1999) Non-destructive optical detection of pigment changes during leaf senescence and fruit ripening. *Physiologia Plantarum* 105: 135–141.

Mutanga O, Adam E, Cho MA (2012) High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. *International Journal of Applied Earth Observation and Geoinformation* 18: 399–406.

NEMA (2010) Kenya state of the environment and outlook [www document]. URL http://www.enviropulse.org/documents/Kenya_SOE.pdf

Ota T, Ahmed OS, Minn ST, Khai TC, Mizoue N, Yoshida S (2019) Estimating selective logging impacts on aboveground biomass in tropical forests using digital aerial photography obtained before and after a logging event from an unmanned aerial vehicle. *Forest Ecology and Management* 433: 162–169.

Peñuelas J, Baret F, Filella I (1995) Semi-empirical indexes to assess carotenoids chlorophyll-a ratio from leaf spectral reflectance. *Photosynthetica* 31: 221–230.

Rex F, Silva C, Paula A, Corte A, Klauberg C, Mohan M et al. (2020). Comparison of statistical modelling approaches for estimating tropical forest aboveground biomass stock and reporting their changes in low-intensity logging areas using multi-temporal LiDAR data. *Remote Sensing* 12: 1498.

Richards JA, Jia X (1999) *Remote Sensing Digital Image Analysis: An Introduction*, 3rd edition. Berlin/Heidelberg, Germany: Springer-Verlag.

Shahi K, Shafri HZM, Taherzadeh E (2014) A novel spectral index for automatic shadow detection in urban mapping based on WorldView-2 satellite imagery. *International Journal of Computer, Electrical, Automation, Control and Information Engineering* 8: 1769–1772.

Slik JWF, Arroyo-Rodrguez V, Aiba S-I, Alvarez-Loayza P, Alves LF, Ashton P et al. (2015) An estimate of the number of tropical tree species. *Proceedings of the National Academy of Sciences of the United States of America* 112: 7472–7477.

Spaias L, Suomlainen J, Tanago JGD (2016) Radiometric detection of selective logging in tropical forest using UAV-borne hyperspectral data and simulation of satellite imagery. Presented at: *2016 European Space Agency Living Planet Symposium*, Prague, Czech Republic, 9–13 May.

Vapnik V (2000) *The Nature of Statistical Learning Theory*, 2nd edition (pp. 138–167). New York, NY, USA: Springer-Verlag.

https://doi.org/10.1017/S0376892922000339 Published online by Cambridge University Press