https://dspace.org/

# First validation of GEDI canopy heights in African savannas

**Xiaoxuan Li; K Wessels; John Armston; Steven Hancock; Laven Naidoo ...**

ELSEVIER SCIENCE INC
https://hdl.handle.net/10539/39509
WOS

Downloaded from a DSpace repository. https://dspace.org/

Applied Geomatics (2023) 15:209–225
https://doi.org/10.1007/s12518-023-00497-9

ORIGINAL PAPER

# Integrating random forest and synthetic aperture radar improves the estimation and monitoring of woody cover in indigenous forests of South Africa

**Mcebisi Qabaqaba[1] · Laven Naidoo[2] · Philemon Tsele[3] · Abel Ramoelo[1] · Moses Azong Cho[4,5]**

Received: 16 August 2022 / Accepted: 6 February 2023 / Published online: 21 February 2023
© The Author(s) 2023

## Abstract

Woody canopy cover (CC) is important for characterising terrestrial ecosystems and understanding vegetation dynamics. The lack of accurate calibration and validation datasets for reliable modelling of CC in the indigenous forests in South Africa contributes to uncertainties in carbon stock estimates and limits our understanding of how they might influence long-term climate change. The aim of this study was to develop a method for monitoring CC in the Dukuduku indigenous forest in South Africa. Advanced Land Observing Satellite (ALOS) Phased Arrayed L-band Synthetic Aperture Radar (PALSAR) global mosaics of 2008, 2015, and 2018, polarimetric features, and Grey Level Co-occurrence Matrix (GLCMs) were used. Machine learning models Random Forest (RF) vs Support Vector Machines (SVM) were developed and calibrated using Collect Earth Online (CEO) data, a free and open-access land monitoring tool developed by the Food and Agriculture Organisation (FAO). The addition of GLCMs produced the highest accuracy in 2008, $R^2$ (RMSE) = 0.39 (36.04%), and in 2015, $R^2$ (RMSE) = 0.51 (27.82%), and in 2018, only SAR variables gave the highest accuracy $R^2$ (RMSE) = 0.55 (29.50). The best-performing models for 2008, 2015, and 2018 were based on RF. During the ten-year study period, shrubland and wooded grassland had the highest transition, at 6% and 13%, respectively. The observed changes in the different canopies provide valuable insights into the vegetation dynamics of the Dukuduku indigenous forest. The modelling results suggest that the CEO calibration data can be improved by integrating airborne LiDAR data.

**Keywords** ALOS PALSAR · Woody canopy cover (CC) · Random Forest · Support Vector Machines · Collect earth online

✉ Mcebisi Qabaqaba
u19409738@tuks.co.za; mcebisiqabaqaba@gmail.com

Laven Naidoo
laven.naidoo@gcro.ac.za

Philemon Tsele
philemon.tsele@up.ac.za

Abel Ramoelo
abel.ramoelo@up.ac.za

Moses Azong Cho
MCho@csir.co.za

[1] Centre for Environmental Studies, Department of Geography, Geoinformatics and Meteorology, University of Pretoria, Private Bag X20, Hatfield 0028, South Africa

[2] Gauteng City-Region Observatory (GCRO), a Partnership of the University of Johannesburg, the University of the Witwatersrand, Johannesburg, the Gauteng Provincial Government and Organised Local Government in Gauteng (SALGA), Johannesburg, South Africa

[3] Department of Geography, Geoinformatics and Meteorology, University of Pretoria, Private Bag X20, Hatfield 0028, South Africa

[4] Precision Agriculture Unit, Advanced Agriculture, Food and Health Cluster, CSIR, Pretoria 0001, South Africa

[5] Department of Plant and Soil Sciences, University of Pretoria, Private Bag X20, Hatfield 0028, South Africa

## Introduction

Globally, indigenous forests are under threat (Thompson et al. 2009). In 1991, forests (including indigenous forests) covered about one-third of the global land surface; however, a declining trend has been observed over the last two decades (FAO 2015). Forests provide various ecosystem services such as erosion protection, carbon storage, and water recycling (DeFries, 2013; Gill et al. 2017). The decline in forest area is due to changes in woody vegetation resulting from deforestation (or degradation) and anthropogenic activities (DeFries, 2013; Gill et al. 2017). Quantifying changes in forests plays a critical role in understanding carbon cycles at regional, national, and global scales (Bonan 2008; Heckel et al. 2020). Such quantifications also support international protocols such as the United Nations Reducing Emissions from Deforestation and Forest Degradation (REDD +) (Mitchell et al. 2017) and the Kyoto Protocol (O'Neill and Oppenheimer 2002).

Forest structural parameters are presented by various metrics, such as woody canopy cover (CC), height, aboveground biomass, and canopy volume. To detect changes in woody cover and understand the effects of degradation and deforestation on forests, it is critical to conduct an accurate quantitative and spatially explicit assessment of forest cover. CC is defined as the percent area projected vertically onto a horizontal plane by woody plant canopies (Jennings et al. 1999). CC is an essential biophysical parameter important for estimating carbon content and vegetation dynamics (Pereira et al. 2013). In South Africa, there is a lack of locally calibrated and validated CC datasets for reliable detection of changes in time series. As a result, the lack of accurate information on CC in South Africa's indigenous forests contributes to uncertainties in current carbon stock estimates and limits scientific understanding of their potential contribution to long-term climate change.

Conventionally, woody structural parameters are typically measured using field-based methods such as plot inventories, horizontal point sampling, and line intersect sampling (e.g., Bester, 1999; Buitenwerf et al. 2012). Notwithstanding the advantages of field-based measurements, particularly in validating and calibrating models, their spatiotemporal representativeness is limited, for example, due to high cost, labour-intensive nature, and resource demands. Alternatively, satellite remote sensing provides data with a high spatiotemporal resolution that is used to assess structural parameters at a regional or global scale. This type of data is suitable for assessing woody plant structure at local, regional (Brandt et al. 2016; Gill et al. 2017), and global scales (Hansen et al. 2013; Sexton et al. 2013).

More recently, Synthetic Aperture Radar (SAR) remote sensing has emerged as a preferred tool for estimating woody plant structural parameters in savannas (Urbazaev et al. 2015; Naidoo et al. 2016; Skowno et al. 2017) as well as boreal forests (Saatchi and Moghaddam 2000) and temperate forests (Lucas et al. 2006). This rapid increase in the use of SAR is due to challenges associated with optical datasets, such as cloud or haze coverage (Anchang et al. 2020). In addition, some optical sensors have a problem with signal saturation, as they do not respond to changes in multi-layered dense forests, resulting in the inability of these sensors to distinguish between woody and non-woody cover (Zhao et al. 2016). The ability of SAR to determine the structural parameters of forests depends on the type of polarisation and wavelength used (Kellndorfer et al. 2019). SAR microwave pulses penetrate clouds, interact with vegetation cover (Santoro et al. 2007), and provide volumetric backscatter information based on canopy structure and environmental conditions (Lucas et al. 2010). Longer wavelengths such as L-band have a stronger and more universal relationship with woody structure (volume, biomass, and cover) than optical or short-wavelength sensors SAR, suggesting that L-band data may be suitable for modelling woody cover in the indigenous forests of South Africa (Mitchard et al. 2011).

The estimation of structural parameters of woody plants can be improved by adding textural features based on the Grey Level Co-Occurrence Matrix (GLCM) (Haralick et al. 1973) and polarimetric features (e.g., Pereira et al. 2018). Recent studies (Beguet et al. 2014; Wessels et al. 2019) have shown that there is a strong relationship between forest structure and texture extracted from remote sensing data. For example, Wessels et al. (2019) added textural features when estimating fractional woody cover in Namibia and obtained a root mean square error (RMSE) of 14% when GLCMs were added. Perreira et al. (2018) compared quad-polarised and dual-polarised L-band with the addition of polarimetric features in predicting woody biophysical parameters. The study concluded that the quad-polarised L-band yielded an RMSE of 10%, while the results of the dual-polarised L-band yielded slightly lower accuracies, i.e., RMSE = 13% (Perreira et al. 2018). Watanabe et al. (2018, 2020) demonstrated the use of the HH/HV band ratio, referred to here as the polarimetric feature, in mapping forested and non-forested areas in the Amazon, achieving an accuracy of over 80 percent.

Field data from CC are often correlated with optical, SAR, or Light Detection and Ranging (LiDAR) signals, using regression or machine learning models to extrapolate CC over large areas (Urbazaev et al. 2018). In South Africa, where reliable CC ground truth data or LiDAR is lacking, alternative methods of obtaining plot data for model calibration and validation need to be found. Anchang et al. (2020) used a tool called Collect Earth Online (CEO) to access

freely available high-resolution imagery (HRI) to generate field data for estimating CC in a savanna environment in Senegal. The CEO tool is key to developing land use change monitoring inventories, which is in line with the Intergovernmental Panel on Climate Change (IPCC) principles and has been used by several developing countries with limited inventories to meet their Nationally Determined Contributions (NDCs) (Tzamtzis et al. 2019).

The capabilities of machine learning algorithms such as Random Forest (RF) and Support Vector Machines (SVM) in estimating structural parameters of woody plants are well documented (Belgiu and Drăgut 2016; Lapini et al. 2020). Several studies have demonstrated the use of machine learning techniques and optical and SAR sensors to monitor different environments in southern Africa. Naidoo et al. (2014) evaluated different modelling algorithms for estimating CC in the savannah environments of South Africa and concluded that RF is a suitable algorithm. In addition, Urbazaev et al. (2015) used a multi-temporal approach, RF, and polarimetric L-band data to map woody cover in savannas. Ludwig et al. (2019) sought to monitor bush encroachment using RF, Sentinel 1, and Sentinel 2 data for modelling woody vegetation in semi-arid and arid environments of South Africa. The literature suggests that machine learning algorithms could improve accuracy in estimating woody structure metrics compared to parametric methods such as stepwise multiple linear regression (SMLR). Although there are a plethora of studies on the use of SAR and machine learning algorithms in estimating woody cover in southern Africa, the literature suggests that the majority of SAR-based research in southern Africa has primarily focused on savanna and semi-arid ecosystems. In particular, the potential and limitations of using the L-band to estimate CC in the indigenous forests of South Africa (with closed canopy) have not been extensively investigated.

Deforestation for agriculture, grazing, or urban development causes forest degradation in many parts of Africa (Van Wyk et al. 1996). Most often, the forest is fragmented into patches of different sizes and shapes, each surrounded by a variety of plants and/or land uses (Cho et al. 2015; Omer et al. 2016). Forest fragmentation has direct ecological impacts on soil erosion, loss of biodiversity, and invasion of alien species. Furthermore, it is difficult to find spatially precise data on how the above factors and forest fragmentation have affected carbon stocks at the local or regional level (Saunders et al. 1991; Cho et al. 2013, 2015).

The Dukuduku forest in Kwa-Zulu Natal, one of the most fragmented indigenous forests in South Africa, requires continuous monitoring, management, and conservation. (Omer et al. 2017). Periodic quantitative data on these forests is essential for decision-making, public management, and re-evaluation of environmental policies for conservation. Therefore, it is crucial to develop an effective remote sensing technique for monitoring CC. A detailed and reliable CC product for indigenous forests can therefore be useful for detecting changes in time series, assessing biodiversity, and planning conservation measures that are essential for managing the impacts of climate change at local, regional, and global scales.

In this study, we sought to develop a method for monitoring CC change (2008–2018) using SAR L-band data and machine learning algorithms in the Dukuduku forest in the Kwa-Zulu Natal province of South Africa with training and validation data from CEO. LiDAR data was used to evaluate SAR-derived CC products. To achieve this aim, the following research questions were addressed in this study.

1. Does combining L-band ALOS PALSAR backscatter with textural and polarimetric features using machine learning techniques improve CC estimation in the Dukuduku Indigenous closed canopy forest?
2. Can CC products derived from SAR L-band data be used to detect changes in woody vegetation (2008–2018) in the Dukuduku indigenous forest?

## Study area

The region under study is the Dukuduku Forest (Fig. 1) (28°25′ S, 32°17′ E), located between the towns of Mtubatuba and St Lucia in the northern part of KwaZulu Natal Province near iSimangaliso Wetland Park, South Africa (Ndlovu 2013; Omer et al. 2017). In the early 1950s, the Dukuduku forest comprised about 6000 hectares (ha) of indigenous tree species (Cho et al. 2015). However, due to deforestation, the forest was reduced to barely 3200 ha by 2011 (Cho et al. 2015; Sundnes, 2013). Most of the natural vegetation surrounding the forest was cleared for agricultural purposes, with sugar cane plantations in the south and eucalyptus plantations in the north (van Wyk et al. 1996; Cho et al. 2015). The climate is subtropical with warm, humid summers and mild winters with temperatures never below 10 °C and mean annual temperatures above 21 °C (Ndlovu et al. 2011). The area receives rainfall throughout the year, with an annual mean of about 1200 mm (Ndlovu et al. 2011; Mucina et al. 2003). The Dukuduku jungle is rich in biodiversity and is considered one of the last remnants of lowland forests along the South African coast (Ndlovu 2013). St Lucia, Africa's largest estuary, and the Dukuduku Forest are part of the iSimangaliso Wetland Park, a United Nations Educational, Scientific, and Cultural Organisation World Heritage Site (UNESCO).

The Dukuduku Forest is subject to two forest management protocols: (i) fragmented forests (Fig. 1A) (black boundary) and (ii) intact forests (Fig. 1A). The fragmented forests are managed and maintained by tribal representatives and local people. The intact forests are managed and maintained by

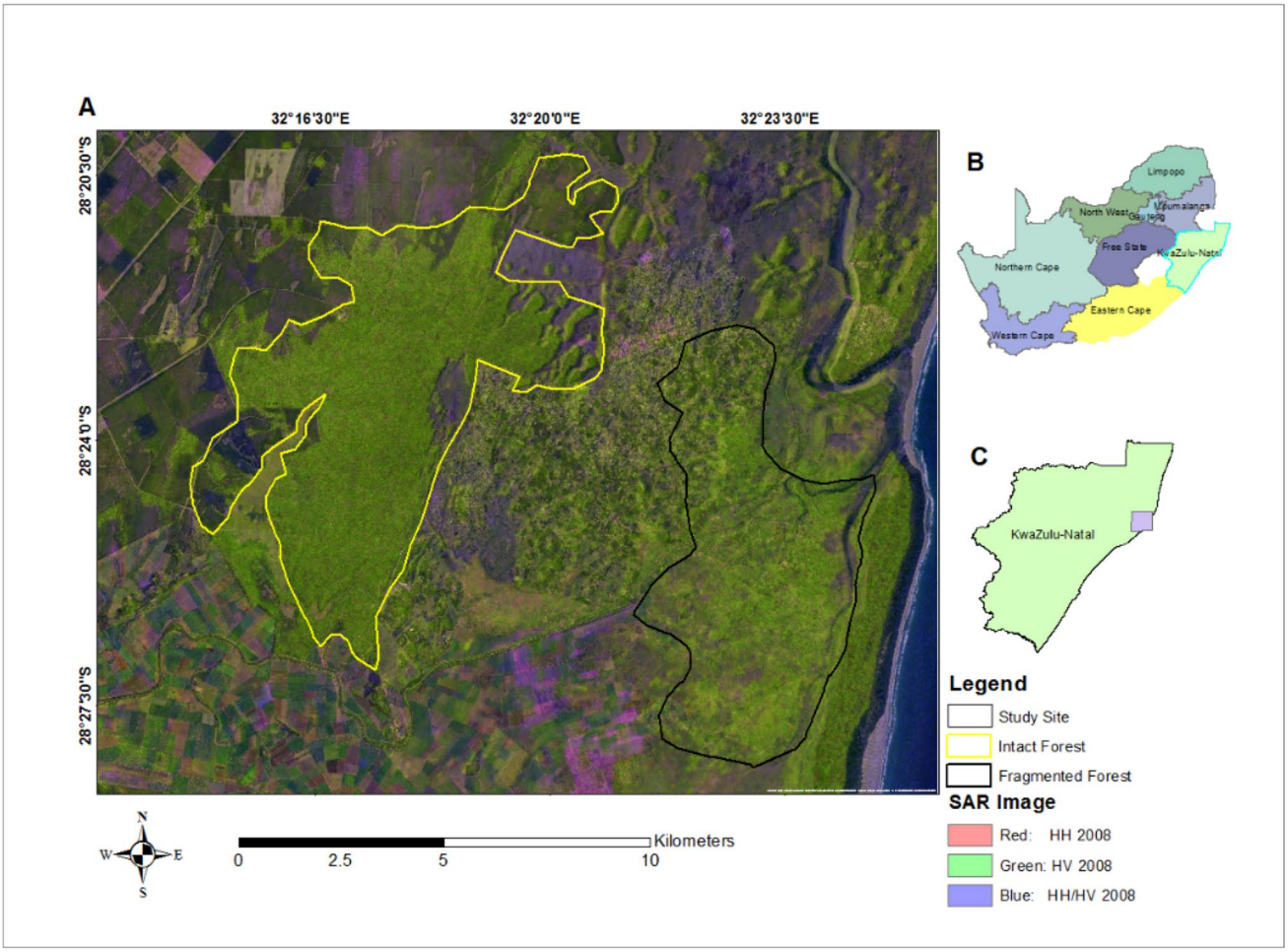


**Fig. 1** (**A**) Study area in KwaZulu Natal showing the Dukuduku indigenous (intact in yellow and fragmented in black) forest. (**B**) KwaZulu Natal Province relative to other provinces in South Africa. (**C**) Location of the study site in KwaZulu Natal Province

officials (e.g., Department of Environment, Forestry, and Fisheries) (Omer et al. 2016). Both the intact and fragmented forests were used to estimate canopy cover in the study area.

## Materials and methods

### Description of datasets used

#### ALOS PALSAR 1 and 2 Data

Global yearly, 25-metre ALOS PALSAR HH and HV dual-polarised mosaics, gamma-naught (γ0) backscatter coefficient were used as the primary predictor variables (SAR backscatter) for the prediction of canopy cover for 2008, 2015, and 2018. The ALOS 1 mosaics were derived using Fine Beam Dual (HH, HV) L-band SAR data collected by PALSAR on-board the Japanese ALOS platform launched by the Japan Aerospace Exploration Agency (JAXA). The ALOS 2 datasets were selected from Strip Map mode (SM1 with 3 m single/dual polarisation or SM3 with 10 m dual polarisation) and were chosen for each year and location based on visual assessment of browse mosaics available for that year. Strip data with the minimum response to surface moisture were chosen to reduce visible banding between adjacent strips (Shimada et al. 2014). The mosaic datasets for 2008, 2015, and 2018 were expressed as gamma (γ0) with backscatter normalised by illumination area under an assumption of scattering uniformity (Shimada & Ohtaki, 2010). The backscatter was also radiometrically and geometrically corrected for topography and standardised for incidence angle ($\theta_i$) ($\gamma 0 = \sigma 0/\cos\,\theta_i$) (Shimada & Ohtaki, 2010).

#### Texture and polarimetric features

Preliminary modelling results showed that textural features and polarimetric features improved modelling accuracy.

Four Haralick texture features, here referred to as GLCMs, namely variance, contrast, entropy, and homogeneity, were calculated over a 7 × 7 pixel window of data from SAR L-band data with a step size of 1 pixel. Bilinear weighting was applied within the sliding window so that pixels closer to the centre of the window were given a higher weight in the GLCM calculation. The GLCMs were calculated for both SAR polarisations HH and HV to obtain eight GLCMs. In this study, the HH/HV ratio is used as a polarimetric feature. The co-polarisation ratio HH/VV is commonly used to classify SAR images (Pereira et al. 2018; Sartori et al. 2011; Novo et al. 2010). In the absence of VV, we used HV instead: This provided eleven SAR input variables, i.e., γ0 HH and γ0 HV, eight texture features, and one polarimetric feature HH/HV ratio, which were used to generate CC maps for each year.

### Collect Earth Online data

Calibration and validation data of the study area for 2008, 2015, and 2018 were collected using CEO. The measurement data collected from CEO was used to train and validate the canopy prediction models. CEO is a free and open-access land monitoring tool (https://collect.earth/) developed by the Food and Agriculture Organisation of the United Nations (FAO). CEO is based on Google Desktop and cloud storage technology. Various satellite image data are freely available on the CEO platform. CEO includes archives of imagery with extremely high spatial resolution and exceptionally high temporal resolution (e.g., Google Earth, Bing Maps, and Mapbox imagery) (Anchang et al. 2020; Bey et al. 2016). Figure 2 illustrates how the plot data was acquired using CEO.

Stratified grid plots were created within the study area and assessed using the CEO tool. To ensure consistency with the SAR L-band dataset ALOS PALSAR, square plots of 25 m × 25 m with 1000 m spacing were created within the study area. The 1000 m spacing between plots was chosen to avoid spatial autocorrelation effects. Digital globe Bing Maps and Planet Norwegian International Climate and Forest Initiative (NICFI) images filtered for 2008 (winter), 2015 (winter), and 2018 (winter). Each plot was populated with spatially distributed gridded sample points at 5 m intervals, including plot edges (i.e., 6 × 6 = 36 samples points per plot). The percentage of CC in a plot was calculated by labelling each point in the sample as a tree or no tree and then adding them together to get the total for the whole plot (1 labelled point equals 1/36 or 2.77 per cent cover). The percentage of CC at the level of the 25 m × 25 m plot was subjected to the formula in Eq. (1):

$$\mathrm{CC} = \left(\sum Y/36\right) * 100 \tag{1}$$

Here, $Y$ stands for the presence of canopy cover treetops. The value 36 is the total number of sample points in the 25 m × 25 m plots. To ensure that the CEO analysis provides reliable results, the plots were assessed based on the highest visual quality in Mapbox and Planet NICFI in the respective years. However, additional images from Google Earth and RapidEye (https://www.planet.com/) were used in the plots

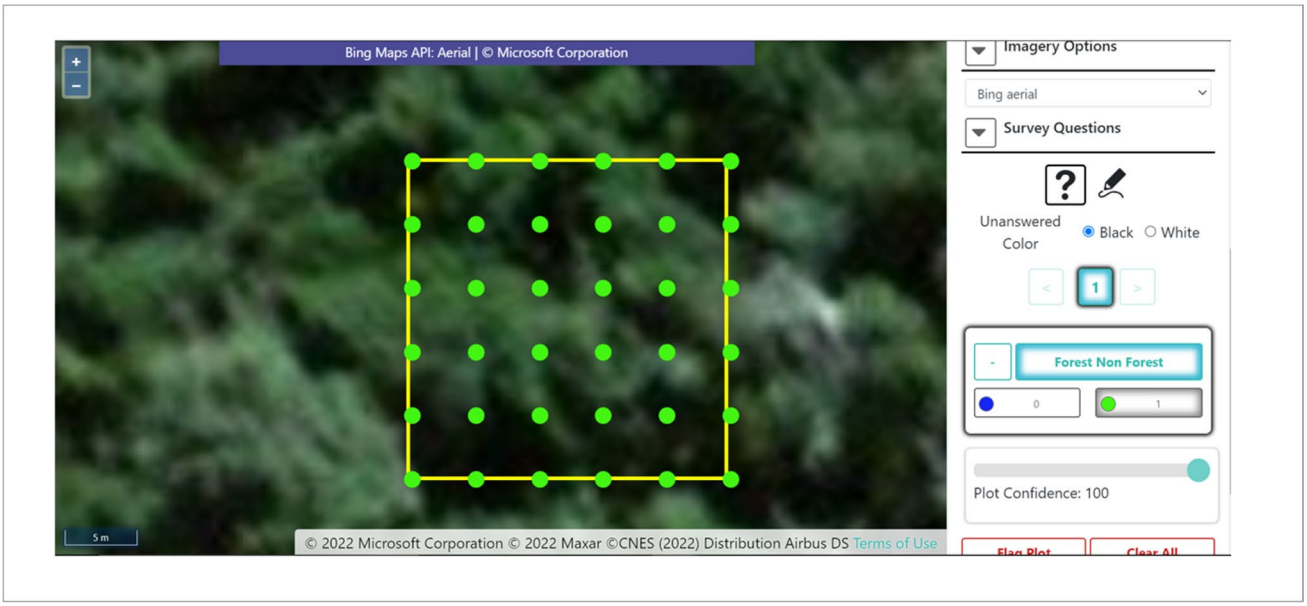


**Fig. 2** Example of plot level of CC using Collect Earth Online (https://collect.earth/). Plot dimension is 25 m by 25 m with 6 by 6 (36) sample points with 5 m spacing

where the CEO images did not have high visual quality. A total of 136 plot-level measurements for each year were used to calibrate and validate the CC predictive models for 2008, 2015, and 2018. The descriptive statistical summary of the CEO plot data for each year, i.e., 2008, 2015, and 2018, is shown in Table 1.

### LiDAR data processing and LiDAR-derived CC

Three Canopy Height Model (CHM) datasets were used to derive CC for the study site, one for 2008, 2015, and 2018. The 2018 CHM was created from a ~1 m Digital Elevation Model (DEM) and top-of-canopy surface models (CSM) generated by processing raw LiDAR point clouds following the processing steps described by Asner et al. (2012). The CHM (pixel size of 1.12 m) was calculated by subtracting the DEM from the CSM. The CC model was derived from the processed CHM. A DSM was obtained from ALOS JAXA for the 2008 and 2015 datasets (https://www.eorc.jaxa.jp/ALOS/). The DSM was used to create a digital terrain model (DTM) following the methods of Estoque et al. (2017). A CHM was then derived by subtracting the DTM from the DSM. CC for 2008 and 2015 was derived from this CHM. The model for 2008 and 2015 CC is referred to as ALOS-derived CC. The LiDAR-derived CC and ALOS-derived CC were created by first applying a data mask to the LiDAR image CHM to create a spatial array of 0 s (non-woody canopy) and 1 s (woody canopy). The percentage of the woody canopy was determined by summing all 1 s and diving by 625. The percentage was created at a spatial resolution of 25 m. LiDAR-derived CC and ALOS-derived CC were validated using CEO plots. The results for 2008, 2015, and 2018 gave RMSE = 29.97% ($R^2 = 0.15$), 30.80% ($R^2 = 0.09$), and 20.33% ($R^2 = 0.46$), respectively.

### SAR, LiDAR, and ALOS registration

Visual inspection of the LiDAR-derived CC product and ALOS PALSAR global mosaics revealed variable registration errors, i.e., SAR datasets did not align with the LiDAR-derived CC products. The ALOS PALSAR mosaics, LiDAR-derived CC, and ALOS-derived CC product were co-registered using the 15 m Landsat 8 panchromatic band to obtain accurate co-registration. During LiDAR CC and ALOS PALSAR mosaics co-registration, an RMSE of 0.5 m was maintained to minimise errors. The co-registration is also to ensure alignment between ALOS PALSAR data to allow change mapping with minimal misalignment anomalies. Furthermore, this allowed aligned error assessment between LiDAR CC and SAR CC. Settlements, main roads, and water bodies such as dams and rivers were masked and excluded from the analysis.

**Table 1** Woody canopy cover statistics of field data plots collected from CEO

| Year | No. of plots | Min | Mean | Max | SD |
|---|---|---|---|---|---|
| 2008 | 136 | 0 | 53.59 | 100 | 46.54 |
| 2015 | 136 | 0 | 67.87 | 100 | 42.93 |
| 2018 | 136 | 0 | 67.26 | 100 | 43.01 |

*Min*, minimum; *Max*, maximum; *SD*, standard deviation

## Data integration, modelling, and mapping

The 25 m by 25 m CEO plots were integrated with SAR backscatter intensities (Fig. 3) for 2008, 2015, and 2018, as well as LiDAR-derived CC products for 2018 and ALOS-derived CC for 2015 and 2018. CEO plots were used to extract mean values from the SAR input variables (HH, HV, HH-GLCMs, HV-GLCMs, and HH/HV ratio). Four modelling scenarios were selected for RF and SVM, namely (i) L-band only, (ii) L-band + GLCM, (iii) L-band + GLCM + HH/HV ratio, and (iv) L-band + HH/HV ratio, to investigate the capabilities of RF and SVM in mapping CC in the Dukuduku indigenous forest.

### Machine learning algorithms

The RF and SVM were used in this study. The RF was used because it accounts for non-linear relationships between variables and makes no assumptions about their statistical distribution (Breiman, 2001). RF is robust and efficient in terms of runtime and accuracy compared to other data mining, machine learning, and regression methods (Ismail et al. 2010). Unlike traditional and fast learning decision trees such as Classification and Regression Trees (CART), RF is insensitive to small changes in the training dataset and provides a lower probability of overfitting (Ismail et al. 2010). RF has become one of the most widely used methods for mapping forest structure parameters and carbon using various satellite and ancillary data (Wingate et al. 2018). RF requires two user-defined inputs: the number of trees or *ntree* formed in the "forest" and the number of possible partition variables for each node or *mtry* (Ismail et al. 2010).

The SVM algorithm is largely based on the principle of structural risk minimisation and statistical theory (Cortes and Vapnik, 1995; Cherkassky and Ma, 2004). The choice of a positive definite kernel function is very important in this method. In addition, the cost factor and gamma in the SVM algorithm affect the penalty imposed for misclassifying a sample and the complexity of the algorithm (Luo et al., 2020). SVMs are capable of handling complex and non-linear problems and a wide range of inputs and can achieve high accuracy even when little training data is available (Marabel and Taboada, 2013).

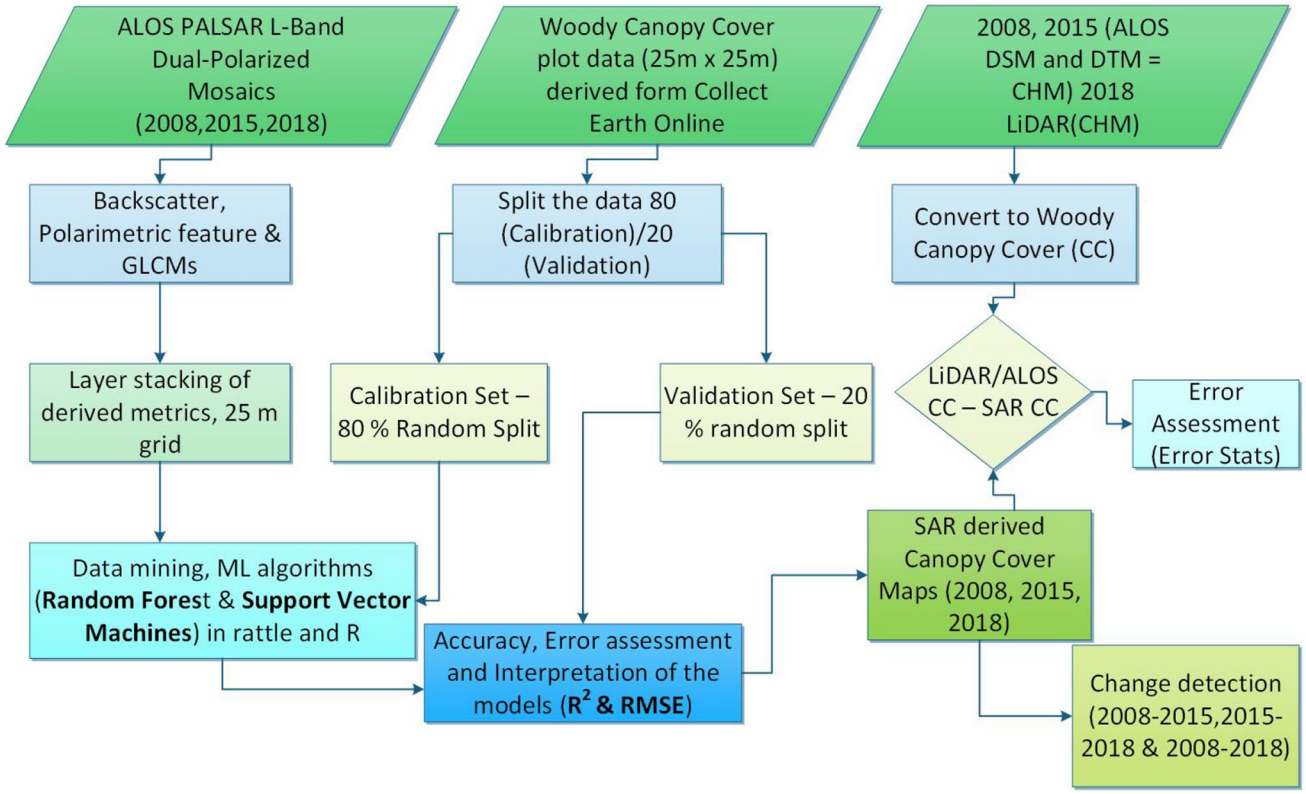


**Fig. 3** Flowchart describing data integration and modelling process

In this study, the radial kernel function with parameters sigma and c was chosen.

R software was used to implement RF and SVM models with the packages randomForest and ksvm. Rattle, a graphical user interface (GUI) for data mining with R (Togaware Pty Ltd., Copyright© 2006-2014), was used in this study. Rattle presents statistical and visual summaries of data, transforms data so that it can be easily modelled, builds both unsupervised and supervised machine learning models from the data, graphs the performance of the models, and evaluates new datasets for use in production. Rattle enables data partitioning, preprocessing, model tuning through resampling, and variable importance estimation. The root mean square error (RMSE), coefficient of determination ($R^2$), and mean absolute error (MAE) were calculated during parameter tuning, and the minimum RMSE value was used to select the optimal model. Both the RF and SVM models were implemented with default settings.

To evaluate the performance of the above two machine learning models, the split sample validation method was used, where 80% of the data was used for training and 20% for model validation. Separate models were created for 2008, 2015, and 2018 using the same RF and SVM parameters. RMSE, $R^2$, and MAE were determined to quantify the performance of the RF and SVM models under different modelling scenarios. A high $R^2$ value, a low RMSE value, and a low MAE value indicate good model performance.

The model with the best performance for each year was applied to the relevant predictor variables (modelling scenarios), which were overlaid and truncated with the boundary of the test area using a mapping script. The script was developed in the statistical software R (version 3.6.1, The R Foundation for Statistical Computing, Copyright© 2019). Map products from CC were imported into ArcMap 10.4 (ESRI, Copyright© 1999–2014) and presented in discrete class intervals to best illustrate the CC distribution representative of the entire modelled areas.

### Error assessment

Error assessment was carried out to examine the error caused by the different modelling scenarios and to determine the uncertainty of each modelling scenario. Error statistics and maps were produced by subtracting the LiDAR-derived CC, ALOS-derived CC, and SAR-derived CC (LiDAR-SAR) for both RF and SVM for all years and modelling scenarios. Both LiDAR-derived CC, ALOS-derived CC, and SAR-derived CC had spatial resolutions of 25 m. The error statistics from RF and SVM

for all modelling scenarios were presented in statistics. For interpretation purposes, the error statistics were divided into five classes using intervals that best covered the observed error range (based on the quadratic sum) in the different modelling scenarios. These classes were major overestimation, minor overestimation, negligible error, minor underestimation, and major underestimation. Furthermore, the CC product derived from 2018 SAR was correlated with the CC product derived from 2018 LiDAR, and the CC product derived from 2008 and 2015 ALOS was correlated with the CC product derived from 2008 and 2015 SAR to evaluate the performance of the modelling scenarios of the two machine learning algorithms in estimating CC in the Dukuduku indigenous forest. The evaluation was done using XY scatter plots of the SAR-derived CC and LiDAR-derived CC, with $R^2$ used to evaluate the relationship.

### CC change mapping

The CC change maps were created by subtracting the earlier CC map (2008) from the more recent CC map (2018), which were created using the best models for each year. The RMSE, which is a measure of the spread of the residuals (regression error) calculated for a LiDAR-derived CC product, was used to assess the uncertainty of a CC product for a single year. The uncertainty of the change product was calculated using the uncertainty of the CC product for each year. To determine the uncertainty, it was assumed that the uncertainties of the two products (the earlier CC SAR product and the later CC SAR product) were not related. The quadratic sum formula (Simard et al. 2006) was used to calculate the uncertainty of the change product under this assumption (Eq. (2)):

$$\sigma \mathrm{CC}\Delta \mathrm{year2–year1} = \sqrt{\sigma^2_{year2} + \sigma^2_{year1}} \quad (2)$$

where $\sigma$ is the RMSE. In addition, we assessed canopy cover change and rate of change using five discrete classes for 2008, 2015, and 2018 CC products derived from SAR. The five classes were based on the percentage cover of each 25 m pixel. The five classes created for the Dukuduku indigenous forest are illustrated in Table 2.

The annual rate of change was calculated for the CC classes; according to Teferi et al. (2013), the net change is the difference between gain and loss and is always an absolute value. The annual change rate of CC for 2008–2015, 2015–2018, and 2008–2018 was calculated according to the approach introduced by Puyravaud (2003), denoted in Eq. (3):

$$r = \left(\frac{1}{\mathrm{t2–t1}}\right) \mathrm{x}\ \ln\left(\frac{\mathrm{A2}}{\mathrm{A1}}\right) \quad (3)$$

**Table 2** Five discrete classes, their description, and percentage cover were created in assessing woody cover change in the Dukuduku indigenous forest, adopted from Song et al. (2014)

| Class name | Description | Percentage cover |
|---|---|---|
| Class 1 | Shrubland | <20% |
| Class 2 | Wooded grassland | 20–40% |
| Class 3 | Woodlands | 40–60% |
| Class 4 | Forest | 60–80% |
| Class 5 | Dense Forest | >80% |

where $r$ is the annual rate of change for each class per year, $A_2$ and $A_1$ are the class area (ha) at time two ($t2$) and time one ($t1$) (in years) between periods and $ln$ is the logarithm. The net change is the difference between the gain and the loss (Teferi et al. 2013). The gain and loss of CC during the study period were derived from the cross-tabulation of 2008–2015, 2015–2018, and 2008–2018.

## Results

### Performance of the models based on SAR and machine learning

The validation performance of the different machine learning algorithms under different modelling scenarios in predicting CC for the years 2008, 2015, and 2018 is shown in Table 3. The use of SAR variables (HH and HV) resulted in low accuracy for the RF model in 2008 ($R^2 = 0.24$) and 2015 ($R^2 = 0.15$) and low accuracy for the SVM model in 2008 ($R^2 = 0.36$), 2015 ($R^2 = 0.27$), and 2018 ($R^2 = 0.47$). However, for 2018 ($R^2 = 0.55$), the HH+HV modelling scenario resulted in high accuracy for the RF model. The (HH+HV+GLCM) modelling scenario resulted in higher accuracies than other modelling scenarios for 2008 ($R^2 = 0.37$) and 2015 ($R^2 = 0.51$) for RF. The modelling scenario (HH+HV+HH/HV_Ratio+GLCM) resulted in higher accuracies than other modelling scenarios for the SVM model for 2008 ($R^2 = 0.37$), 2015 ($R^2 = 0.46$), and 2018 ($R^2 = 0.49$). For 2018, the modelling scenario (HH+HV+HH/HV_Ratio+GLCM) resulted in lower accuracies than using only SAR backscatter and SAR backscatter plus HH/HV ratio for RF; however, the modelling scenario with all variables resulted in high accuracies for SVM (Table 3). RF machine learning algorithms produced the best model for 2008, 2015, and 2018.

### Overall model uncertainty of CC estimates

The model uncertainty of the SAR CC estimates was evaluated by correlating the LiDAR-derived CC and ALOS-derived CC estimates against the SAR-derived CC estimates.

**Table 3** Woody canopy cover (CC) modelling accuracy assessment (validation) results as obtained from the Random Forest and Support Vector Machines from four modelling scenarios

| | RF | | | SVM | | |
|---|---|---|---|---|---|---|
| | $R^2$ | RMSE | MAE | $R^2$ | RMSE | MAE |
| MS (2008) | | | | | | |
| $CC_{SAR}$ | 0.24 | 40.91 | 30.25 | 0.36 | 38.24 | 26.36 |
| $CC_{SAR+R}$ | 0.29 | 39.08 | 29.31 | 0.35 | 39.57 | 26.11 |
| $CC_{SAR+GLCMs}$ | **0.39** | **36.04** | **29.75** | 0.34 | 37.85 | 29.86 |
| $CC_{SAR+R+GLCMs}$ | 0.37 | 36.55 | 29.53 | **0.37** | **37.25** | **28.65** |
| MS (2015) | | | | | | |
| $CC_{SAR}$ | 0.15 | 39.53 | 33.62 | 0.27 | 41.91 | 30.18 |
| $CC_{SAR+R}$ | 0.19 | 38.05 | 33.48 | 0.24 | 42.04 | 30.69 |
| $CC_{SAR+GLCMs}$ | 0.48 | 32.01 | 28.14 | 0.45 | 33.14 | 25.60 |
| $CC_{SAR+R+GLCMs}$ | **0.51** | **31.36** | **27.82** | **0.46** | **31.77** | **27.91** |
| MS (2018) | | | | | | |
| $CC_{SAR}$ | **0.55** | **29.54** | **20.60** | 0.47 | 32.93 | 20.44 |
| $CC_{SAR+R}$ | 0.53 | 30.08 | 21.40 | 0.47 | 33.66 | 20.79 |
| $CC_{SAR+GLCMs}$ | 0.41 | 33.83 | 24.24 | 0.48 | 32.25 | 21.90 |
| $CC_{SAR+R+GLCMs}$ | 0.42 | 33.71 | 24.28 | **0.49** | **31.68** | **21.96** |

MS – modelling scenario's dataset split into 80% training and 20% for validation of models. SAR-HH&HV backscatter coefficient; R – HH/HV ratio; GLCMs – Grey Level Co-occurrence Matrix. Values in bold represent best performing modelling scenarios for each machine learning algorithm

Due to low $R^2$ and high RMSE, the estimation of CC was applied to all the modelling scenarios for RF and SVM. The 2008 and 2015 SVM and RF models produce low $R^2$ values; therefore, only the 2018 results are presented. The RF and SVM results modelling scenario (HH+HV+HH/HV_Ratio+GLCM) and (HH+HV+GLCM) modelling scenarios are presented in Fig. 4. The validation results of SAR CC versus LiDAR CC for the RF model under different modelling scenarios: $CC_{SAR}$ ($R^2 = 0.55$), $CC_{SAR+R}$ ($R^2 = 0.55$), $CC_{SAR+GLCMs}$ ($R^2 = 0.59$), $CC_{SAR+R+GLCMs}$ ($R^2 = 0.6$). The SVM modelling scenarios when compared with LiDAR CC for 2018 produced similar but slightly low $R^2$ relative to RF: $CC_{SAR}$ ($R^2 = 0.54$), $CC_{SAR+R}$ ($R^2 = 0.53$), $CC_{SAR+GLCMs}$ ($R^2 = 0.57$), $CC_{SAR+R+GLCMs}$ ($R^2 = 0.6$). Figure 4 illustrates the underestimations that the SAR-derived CC product yield when correlated with the LiDAR-derived CC. The SVM and RF model underestimation CC at lower coverage.

## SAR-derived CC maps and error statistics

CC was mapped for the study area using the best-performing models for each year ($CC_{SAR+GLCMs}$ for 2008, $CC_{SAR+R+GLCMs}$ for 2015, and $CC_{SAR}$ for 2018), but only the 2018 RF CC map is presented for the sake of brevity. Figure 5 illustrates the CC map derived using the best-performing model for 2018 SAR variables and RF for 2018. In the intact forest, there is a high CC with values between 60 and 100%, but low coverage is observed at the edges of the intact forest. The distribution of CC within the fragmented forest varies between 40 and 80%, with small patches of low and high cover.

Table 4 presents the LiDAR-SAR error statistics. For 2008 and 2015, both the RF and SVM models presented high major overestimations and high major underestimations. However, for 2018 estimates, major overestimations and major underestimations were significantly low, and negligible error was high for both RF and SVM over the four modelling scenarios. Key areas of discussion regarding overestimations and underestimations across LiDAR-SAR coverage for 2018 were selected and are presented in Fig. 5.

The addition of GLCMs and polarimetric features across the years for the estimates of RF and SVM CC did not reduce or increase the overestimations and underestimations of CC, implying that the overestimates and underestimates may be influenced by environmental characteristics. The use of the different CC products in determining uncertainty explains the significant differences in over- and underestimates between these years. The regions where the largest overestimates and underestimates observed from the 2018 RF modelling scenarios are shown in Fig. 6. Figure 6(ii) shows the largest underestimates and overestimates produced by the 2018 RF model when all predictor variables are used. These over- and underestimates correspond to the area of interest E from Fig. 5.

Majority of the overestimations and underestimations across all modelling scenarios for both RF and SVM models occurred in the fragmented forest and at the edges of the intact forest, as can be observed in Fig. 6 (i) areas of interest A–G (Fig. 5). A combination of major overestimations

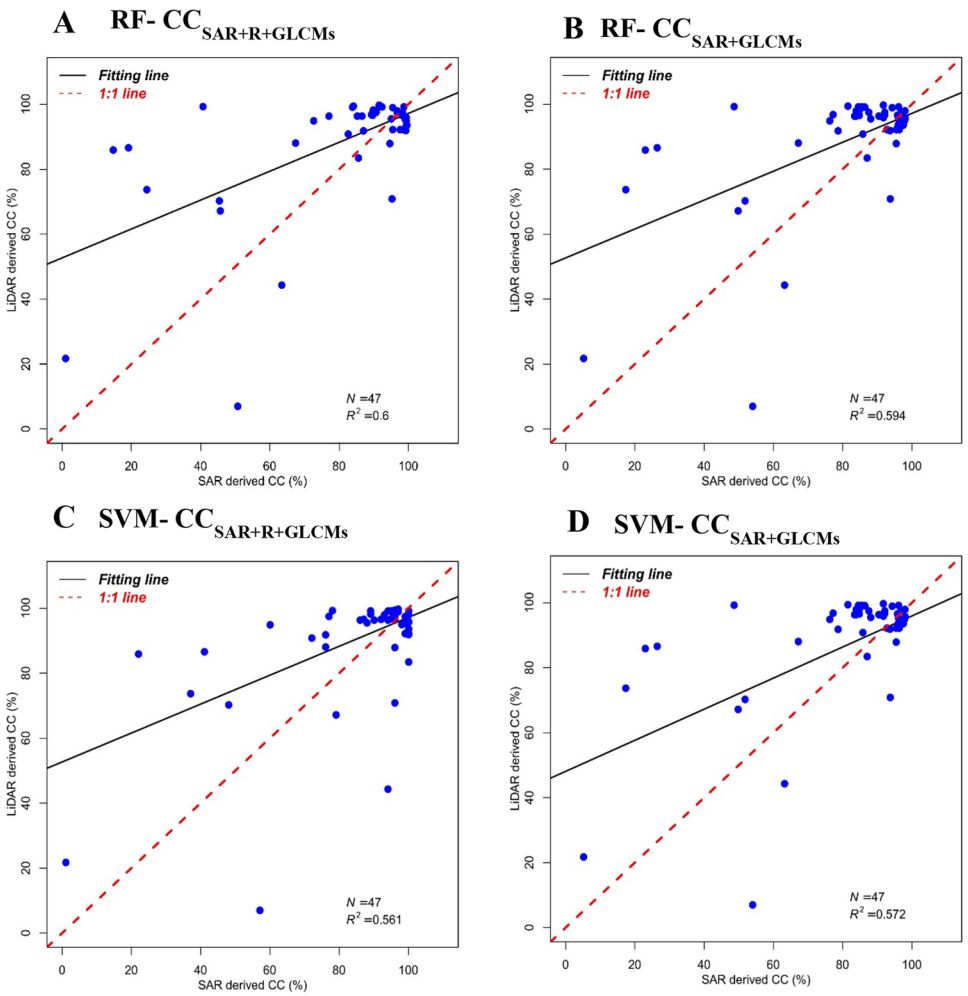


**Fig. 4** (**A**–**D**) SAR-derived CC vs LiDAR-derived CC for 2018 under different modelling scenarios

and negligible error plus minor overestimations in the $CC_{SAR+GLCMs}$ scenario was observed in the area of interest E on Fig. 5, which is further highlighted and illustrated in Fig. 6 (ii). Modelling scenarios (HH+HV+GLCMs) and (HH+HV) errors are also presented in Fig. 6 (iv, v), respectively.

## CC Change, error estimation, and rate of change

The change SAR-derived product was computed using the Random Forest-modelled SAR-derived estimates $CC_{SAR+R+GLCMs}$ for 2008 and $CC_{SAR}$ for 2018 because they were the best-performing models. The uncertainty of the change product was calculated from the annual CC maps. The change uncertainty of $\sigma CC\Delta_{2018–2008}$ of 38.41 represents the average uncertainty of all individual pixels. The $CC_{2018–2008}$ change was summarised into five classes (Table 3). Figure 7 illustrates the SAR-derived CC estimates when using all the predictor variables $CC_{SAR+R+GLCMs}$. The RF-modelled CC estimates using all predictor variables were presented to observe the area covered by each CC class (Table 5).

In 2008, dense forest, forest, woodlands, wooded grassland, and shrubland covered 37%, 17%, 20%, 23%, and 2%, respectively. The classes dense forest, woodlands, and shrubland increased from 37% (3365.31 ha), 17% (1564.13 ha), 20% (2028.19 ha), and 2% (205.63 ha) in 2008 to 40% (3605.44 ha), 21% (1866.31 ha), 22% (1938.38 ha), and 6% (678.5 ha) in 2018. On the contrary, wooded grassland drastically decreased from 23% (2028.19 ha) to 10% (901.63 ha). Figure 7 illustrates the quantitative differences in woody CC estimates area percentages for 2008 and 2018 using the RF algorithm.

The annual rate of change varied for each CC class across the study period (Table 5). Between 2008 and 2018, for dense forest, forest, woodlands, wooded grassland, and shrubland experienced an annual rate of change of 0.69% $ha^{-1}$(increase), 1.79% $ha^{-1}$ (increase), 0.59% $ha^{-1}$ (increase), 8.11% $ha^{-1}$ (decrease), and 11.93% $ha^{-1}$ (increase), respectively.

During the study period, wooded grasslands experienced the highest transition, with 80.64% of its total area in 2008 changing to other classes, with 24.90%, 21.45%, 18.16%, and 16.13% transitioning to woodlands, shrubland, forest, and dense forest, respectively. The other classes also

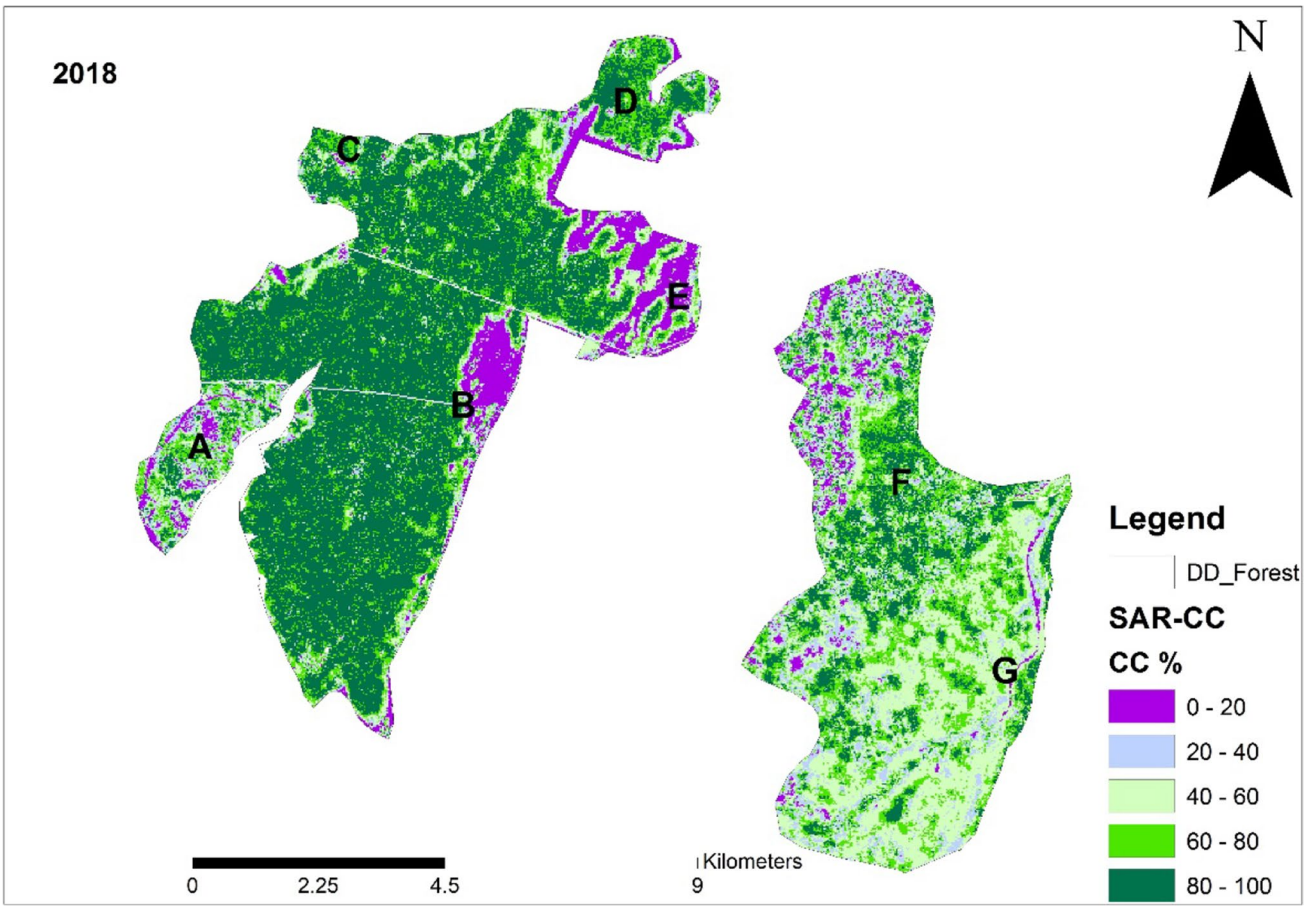


**Fig. 5** SAR-derived CC map (2018) using RF and all predictor variables. Letters A–G represent selected areas of interest used as examples for LiDAR-SAR error statistics. DD – Dukuduku

experienced change with dense forest (40.54% increase, 36.3 decrease), high cover (79.44% increase, 75.47% decrease), woodlands (75.53% increase, 74.04% decrease), and shrubland (90.64% increase, 69.48% decrease).

## Discussion

The results of the study show that RF provides the best model for 2008, 2015, and 2018. In 2008, it was RF which performed best with SAR variables and GLCMs. In 2015, it was the RF model that performed best with SAR variables, polarimetric features, and GLCMs. In 2018, it was the RF model with only SAR variables that performed best. These results are consistent with other studies that have used RF to estimate CC in different environments in Africa and southern Africa (Naidoo et al. 2014, 2015; Anchang et al. 2020). The use of CEO-derived field plots for model calibration and validation resulted in a low $R^2 < 0.5$ and a high RMSE (29.5–36.5%) for both RF and SVM. However, significant changes were observed within the different percentage coverages of CC across the study area, providing only information on changes in CC in the Dukuduku indigenous forest. Polarimetric and textural features slightly improved the RF and SVM models (Table 3 and Fig. 4). These results indicate the importance of textural and polarimetric features in estimating CC. Textural features are important for estimating the structural parameters of woody vegetation because they can examine pixels within a specific neighbourhood associated with tree clumps (Gonzalez and Woods, 1992). They also reveal underlying physical variations in the image and provide information about the structural arrangement of the surface and its relationships to the surrounding environment (Haralick et al. 1973). Several studies have used textural features to map and assess CC (Wood et al. 2012; Madonsela et al. 2017; Wessels et al. 2019). These studies used optical datasets or were conducted in other environments, such as the savannah. This study therefore extends the body of knowledge by using SAR L-band, textural, and polarimetric features with machine learning models to estimate CC in the indigenous closed canopy forests and provides an opportunity to test this approach in other environments of

**Table 4** CC percentage difference across the LiDAR-SAR coverage for the best-performing model four modelling scenarios and the two machine learning models

| | RF | | | | SVM | | | |
|---|---|---|---|---|---|---|---|---|
| | $CC_{SAR}$ | $CC_{SAR+R}$ | $CC_{SAR+GLCMs}$ | $CC_{SAR+R+GLCMs}$ | $CC_{SAR}$ | $CC_{SAR+R}$ | $CC_{SAR+GLCMs}$ | $CC_{SAR+R+GLCMs}$ |
| *2008 CC error classes* | | | | | | | | |
| Major overestimation (<−45%) | 40 | 38 | 44 | 40 | 39 | 39 | 43 | 43 |
| Minor overestimation (−15 to −45%) | 18 | 22 | 19 | 23 | 16 | 15 | 17 | 17 |
| Negligible error (−15 to 15%) | 25 | 23 | 15 | 17 | 31 | 30 | 21 | 21 |
| Minor underestimation (15–45%) | 6 | 6 | 14 | 10 | 6 | 7 | 11 | 11 |
| Major underestimation (>45 %) | 11 | 10 | 9 | 10 | 8 | 9 | 8 | 8 |
| *2008 CC error classes* | | | | | | | | |
| Major overestimation (<−45%) | 54 | 55 | 49 | 47 | 62 | 63 | 55 | 53 |
| Minor overestimation (−15 to −45) | 10 | 9 | 14 | 15 | 3 | 2 | 9 | 11 |
| Negligible error (−15 to 15%) | 16 | 13 | 9 | 12 | 25 | 26 | 24 | 22 |
| Minor underestimation (15–45%) | 14 | 17 | 21 | 18 | 10 | 9 | 8 | 10 |
| Major underestimation (> 45%) | 6 | 6 | 6 | 7 | - | - | 4 | 4 |
| *2018 CC error classes* | | | | | | | | |
| Major overestimation (<−45%) | 3 | 3 | 2 | 2 | 3 | 3 | 2 | 2 |
| Minor overestimation (−15 to −45%) | 7 | 5 | 5 | 5 | 8 | 8 | 7 | 7 |
| Negligible error (−15 to 15%) | 68 | 67 | 63 | 66 | 77 | 77 | 72 | 72 |
| Minor underestimation (15–45%) | 13 | 19 | 19 | 19 | 7 | 7 | 13 | 13 |
| Major underestimation (>45%) | 10 | 7 | 8 | 8 | 5 | 5 | 5 | 5 |

Southern Africa. In addition, the study used the estimated SAR products to obtain time series of changes in CC.

Machine learning techniques used to estimate woody structural parameters require calibration and validation data. The calibration and validation data may contain inaccuracies or inconsistencies that result in inaccurate models. The low accuracy of the RF and SVM models can be attributed to the inconsistencies between the CEO plot data and the ALOS PALSAR annual mosaics. This is evident when the relationship between the SAR-derived CC and the ALOS-derived CC and the LiDAR-derived CC is determined for all the years. There was a moderate positive relationship between the 2018 SAR-derived CC and the LiDAR-derived CC, $R^2 > 0.5$, and a poor positive relationship for 2008 and 2015, $R^2 < 0.4$. The difference in the dates of the ALOS-derived CC and SAR-derived CC products caused a poor correlation between the datasets. It should be noted that for the 2008 and 2015 error statistics when determining the uncertainty of the RF and SVM SAR-derived CC products, the CC product used for correlation was derived from DTM and DSM products that were not produced with LiDAR point cloud data. However, for the 2018 RF and SVM SAR-derived CC products, the uncertainty of the products was determined using the CC derived from LiDAR point cloud data. The structural parameters of woody plants are influenced by environmental conditions such as precipitation, soil type and moisture, and topography (Sankaran et al. 2005). Model accuracy could be improved by integrating environmental variables with remote sensing data (Wessels et al. 2019).

The SAR-derived CC change statistics and their respective interpretations have successfully addressed the goal of demonstrating the potential of SAR-derived CC maps in monitoring CC change, especially when using free calibration/validation data from CEO. Significant changes were observed between the different CC percentages during the study period, with shrubland gaining the most area during this period. Forest and wooded grassland each recorded the largest decreases. Looking at the changes in CC classes and the rate of change between 2008 and 2015 and 2015 and 2018, it is clear that the Dukuduku indigenous forest is facing threats that need to be investigated. However, these changes need to be interpreted with caution as we acknowledge that there are certain sources of error that have not been thoroughly investigated. Errors introduce uncertainty in the estimated products and are the result of inaccurate measurements of plot data, measurement biases, and inaccuracies in the geometric rectification of remote sensing images (Wang et al. 2005). Different independent validation datasets were used in the study, hence the observed errors. These uncertainties raise the question of whether the variables derived from the plots are representative of the landscape in question (Marvin et al. 2014). The CEO plot data used for calibration and validation in this study were validated using LiDAR-derived CC for 2018 and ALOS-derived CC for 2008 and 2015 to verify the accuracy of the plot data. Multi-temporal and multi-seasonal

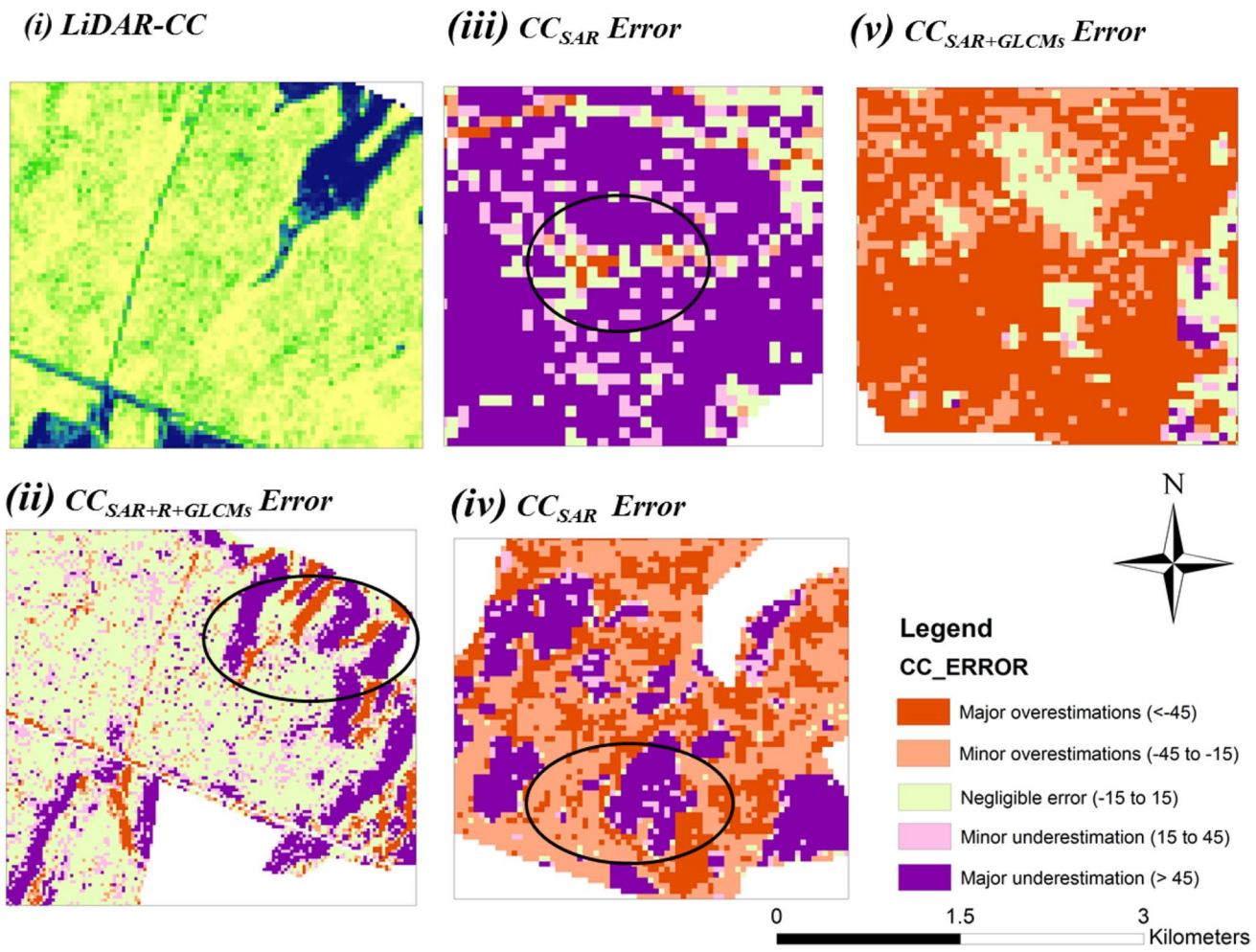


**Fig. 6** (**i**) 25 m LiDAR-derived CC maps. (**ii**–**v**) 2018 SAR-derived CC error maps of areas of interest, as illustrated in Fig. 5

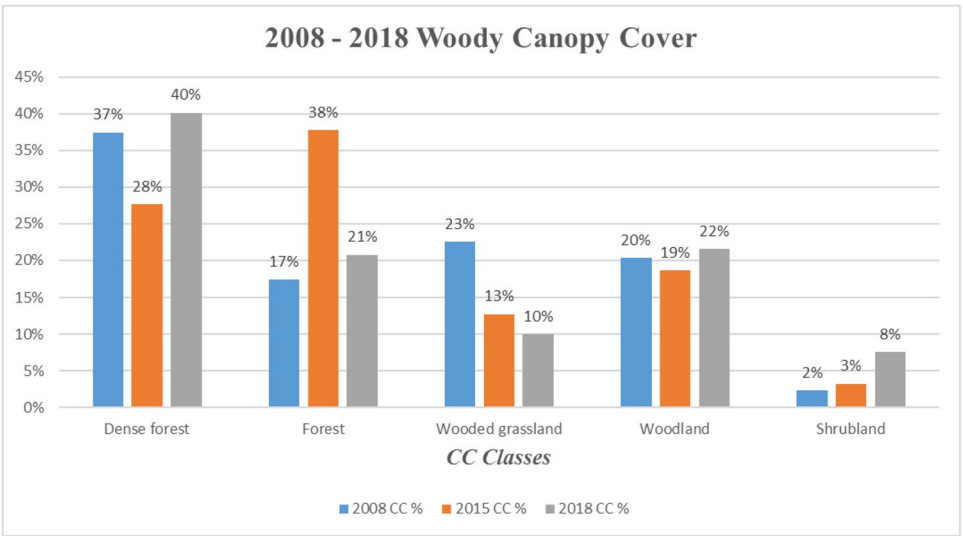


**Fig. 7** Quantitative differences in woody canopy cover estimates area percentages for 2008, 2015, and 2018 using the RF algorithm with all predictor variables ($CC_{SAR+R+GLCMs}$)

**Table 5** CC change trend and the annual rate of change across the study area

| CLASS | 2008 | | 2015 | | 2018 | | Change %[b] | | | Annual rate of change (%)[c] | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Ha | %[a] | Ha | %[a] | Ha | %[a] | 08–15 | 15–18 | 08–18 | 08–15 | 15–18 | 08–18 |
| DF | 3365.31 | 37 | 2485.81 | 28 | 3605.44 | 40% | −9 | 12 | 3 | −4.32 | 12.39 | 0.69 |
| F | 1564.13 | 17 | 3394.44 | 38 | 1866.31 | 21% | 21 | −17 | 4 | 11.08 | −17.94 | 1.76 |
| WG | 2028.19 | 23 | 1145.94 | 13 | 901.63 | 10% | −10 | −3 | −13 | −8.16 | −7.99 | −8.11 |
| Wd | 1827 | 20 | 1673.31 | 19 | 1938.38 | 22% | −1 | 3 | 2 | −1.26 | 4.90 | 0.59 |
| Shr | 205.62 | 2 | 290.75 | 3 | 678.5 | 8% | 1 | 5 | 6 | 4.95 | 28.24 | 11.93 |
| Total | **8990.25** | **100** | **8990.25** | **100** | **8990.25** | **100** | | | | | | |

[a]Percentage of each class out of the total area

[b]Percentage change in the class

[c]Percentage the annual rate of change in each class

*DF*, dense forest; *F*, forest; *WG*, wooded grassland; *Wd*, woodland; *Shr*, Shrubland. Values in bold represent total area in hactres and percentage

SAR data can improve modelling results (Naidoo et al. 2015); however, the ALOS PALSAR L-band data used in this study were obtained as annual mosaics. Therefore, the potential and limitations of using multi-temporal SAR data in the closed canopy forests of Southern Africa are unknown. Future missions such as BIOMASS, which will provide longer P-band SAR imagery (Ho Tong Minh et al. 2016) and NISAR, which will generate quad-polarimetric L-band and S-frequency (Rosen et al. 2017), promise to improve and extend the work presented in this study to other forests and woody vegetation types to provide accurate and improved woody canopy cover estimates. This is due to the fact that the SAR signals from these sensors are able to penetrate deep into a complex, multi-layered, closed forest (Rosen et al. 2017).

Both the RF and SVM models resulted in under- and overestimations of CC in certain areas of the study area under the different modelling scenarios. However, this was expected due to the saturation of L-band backscatter at biomass greater than 90–110 Mg $ha^{-1}$ associated with closed canopy forests (Yu and Saatchi, 2016). Several studies have reported this phenomenon (Urbazaev et al. 2018; Wessels et al. 2019; Naidoo et al. 2015). This phenomenon is also related to RF regressions due to the bias of ensemble trees towards the sample mean (Xu et al. 2016). Interestingly, the over- and underestimations in the SAR-derived CC estimates for 2018 were drastically reduced, fully supporting the use of LiDAR data for calibration and validation of SAR models through upscaling methods to reduce dependence on field plot data causing these errors and thus increase the accuracy of the SAR-derived CC estimates. Understanding the variables that cause some level of error is important for improving and building accurate models in the future. Most of the overestimations and underestimations, as shown in Fig. 5 and Fig. 6 (ii–iv), occurred in the fragmented forest and at the edges of the intact forest. Majority of over and underestimations occur at the edges of the intact forest (Fig. 5). The Dukuduku forest is categorised by dune ridges and is surrounded by floodplains to the east, south, and southwest, which explains the overestimations and underestimations that occur in these regions.

Observing the overestimation and underestimation error statistics over the study period and the fact that only the CEO plot data was used for calibration and validation of the machine learning algorithms. The uncertainty results obtained in the study confirm the recommendations of Naidoo et al. (2016) that the use of LiDAR for model calibration and validation is necessary to increase the accuracy of estimation of CC across different vegetation types. LiDAR data excels at delineating the understory of forests, which includes grasses, forbs, and shrubs distributed below the forest (Mahlangu et al. 2018). LiDAR metrics yield strong correlations with L-SAR data, as L-band signals interact with tree trunks and branches (Mitchard et al. 2011). However, in South Africa, there is limited LiDAR data due to the high costs associated with obtaining LiDAR data. Missions such as the Global Ecosystems Dynamics Investigations LiDAR (GEDI) data have the potential to address this problem. Data obtained from the GEDI mission can be used to extract other structural parameters of woody vegetation, such as canopy height and crown diameter, which are essential for ecological studies related to vegetation dynamics, biomass estimation, and spatial characterisation of vegetation (Ferraz et al. 2016).

## Concluding remarks

The aim of this study was to develop a method for monitoring CC in Dukuduku indigenous forest. This was done using SAR L-band mosaics of 2008, 2015, and 2018, polarimetric and GLCMs with machine learning models calibrated and validated with CEO data. This is the first attempt to map and monitor changes in CC in the indigenous closed canopy forests in South Africa using SAR data. Considering

the complexity of estimating CC in a closed canopy forest with highly fragmented forest and heterogeneous vegetation cover, we can conclude that the selected machine learning models and modelling scenarios have low $R^2$ ($0.2 < R^2 < 0.6$) and high RMSE (29.5–36.5%) values for estimating CC in the Dukuduku indigenous forest in South Africa.

In terms of machine learning algorithms suitable for monitoring CC in the Dukuduku indigenous forest, the results show that the machine learning algorithm RF provides the best model for estimating CC in 2008, 2015, and 2018; the addition of polarimetric and texture features improved the modelling accuracy. However, more accurate results for estimating CC could be obtained by using LiDAR data for calibration and validation. The methodology presented in this study can be used as an alternative for estimating forest cover when the availability of LiDAR data is limited. We also believe that the methodology can be adopted and implemented to estimate other structural parameters of woody vegetation if LiDAR data are used for calibration and training to improve modelling accuracy. The authors believe that CEO plot data is an alternative data source to LiDAR where LiDAR availability is limited or LiDAR data is not available for the study period. However, to confirm these results, further tests with more CEO plots integrated with environmental variables need to be conducted.

Beyond this study, multi-temporal SAR datasets can be used for long-term forest cover mapping and monitoring to assess the effects of seasonality and environmental conditions on woody canopy cover in different environments of Southern Africa. Finally, the results presented in this study can provide valuable information to the scientific community on the capabilities and limitations of SAR L-band and CEO data in estimating CC in indigenous forests of South Africa.

**Acknowledgements** The authors thank the Council for Scientific and Industrial Research (CSIR) for funding this study.

**Author contribution** **M.Q.** contributed to data processing, results and analysis, conceptualization, and paper writing. **L.N.** contributed supervision, guidance, paper draft editing, and provision of LiDAR datasets. **P.T.**: supervision, conceptualization, and paper writing. **A.R.**: supervision, conceptualization, and paper writing. **M.C.:** supervision, conceptualization, and paper writing.

**Funding** Open access funding provided by University of Pretoria. This work was supported by the Council for Scientific and Industrial Research (CSIR) – South Africa, the Southern Africa Science Service Centre for Climate and Adaptive Land Management (SASSCAL), and the National Research Foundation (NRF) – South Africa.

## Declarations

**Conflict of interest** The authors declare no competing interests.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

## References

Anchang JY, Prihodko L, Ji W et al (2020) Toward operational mapping of woody canopy cover in tropical savannas using Google Earth engine. Front Environ Sci Eng China 8. https://doi.org/10.3389/fenvs.2020.00004

Asner GP, Mascaro J, Muller-Landau HC, Vieilledent G, Vaudry R, Rasamoelina M, Hall JS, van Breugel M (2012) A universal airborne LiDAR approach for tropical forest carbon mapping. Oecologia 168(4):1147–1160. https://doi.org/10.1007/s00442-011-2165-z

Beguet B, Guyon D, Boukir S, Chehata N (2014) Automated retrieval of forest structure variables based on multi-scale texture analysis of VHR satellite imagery. ISPRS J Photogramm Remote Sens 96:164–178. https://doi.org/10.1016/j.isprsjprs.2014.07.008

Belgiu M, Drăguţ L (2016) Random Forest in remote sensing: a review of applications and future directions. ISPRS J Photogramm Remote Sens 114:24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Bester FV (1999) Major problem-bush species and densities in Namibia. Agricola 10:1–3

Bey A, Sánchez-Paus Díaz A, Maniatis D, Marchi G, Mollicone D, Ricci S, Miceli G (2016) Collect earth: land use and land cover assessment through augmented visual interpretation. Remote Sens 8(10):807

Bonan GB (2008) Forests and climate change: forcings, feedbacks, and the climate benefits of forests. Science 320:1444–1449. https://doi.org/10.1126/science.1155121

Brandt M, Hiernaux P, Tagesson T et al (2016) Woody plant cover estimation in drylands from Earth Observation based seasonal metrics. Remote Sens Environ 172:28–38. https://doi.org/10.1016/j.rse.2015.10.036

Breiman L (2001) Random Forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

Buitenwerf R, Bond WJ, Stevens N, Trollope WSW (2012) Increased tree densities in South African savannas: >50 years of data suggests $CO_2$ as a driver. Glob Chang Biol 18:675–684. https://doi.org/10.1111/j.1365-2486.2011.02561.x

Cherkassky V, Ma Y (2004) Practical selection of SVM parameters and noise estimation for SVM regression. Neural Netw 17:113–126. https://doi.org/10.1016/S0893-6080(03)00169-2

Cho MA, Ramoelo A, Debba P et al (2013) Assessing the effects of subtropical forest fragmentation on leaf nitrogen distribution using remote sensing data. Landsc Ecol 28:1479–1491. https://doi.org/10.1007/s10980-013-9908-7

Cho MA, Malahlela O, Ramoelo A (2015) Assessing the utility WorldView-2 imagery for tree species mapping in South African subtropical humid forest and the conservation implications: Dukuduku forest patch as case study. Int. J. Appl. Earth Obs. Geoinf 38:349–357. https://doi.org/10.1016/j.jag.2015.01.015

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297. https://doi.org/10.1007/BF00994018

DeFries, R. (2013) Why forest monitoring matters for people and the planet. Global forest monitoring from Earth observation 1-14
Estoque M (2017) Validating ALOS PRISM DSM-derived surface feature height: implications for urban volume estimation. Tsukuba Geoenviron. Sci 13:13–22
Ferraz A, Saatchi S, Mallet C et al (2016) Airborne LiDAR estimation of aboveground forest biomass in the absence of field inventory. Remote Sensing 8:653. https://doi.org/10.3390/rs8080653
Food and Agriculture Organization of the United Nations (2015) Global forest resources assessment 2015: how are the World's Forests Changing? Food and Agriculture Organization of the United Nations, Rome
Gill T, Johansen K, Phinn S et al (2017) A method for mapping Australian woody vegetation cover by linking continental-scale field data and long-term Landsat time series. Int J Remote Sens 38:679–705. https://doi.org/10.1080/01431161.2016.1266112
Gonzalez (1992) R. Woods digital image processing. Addison–Wesely Publishing Company
Hansen MC, Potapov PV, Moore R et al (2013) High-resolution global maps of 21st-century forest cover change. Science 342:850–853. https://doi.org/10.1126/science.1244693
Haralick RM, Shanmugam K, hak DI (1973) Textural features for image classification. IEEE Trans Syst Man Cybern SMC-3:610–621. https://doi.org/10.1109/TSMC.1973.4309314
Heckel K, Urban M, Schratz P et al (2020) Predicting forest cover in distinct ecosystems: the potential of multi-source sentinel-1 and -2 data fusion. Remote Sensing 12:302. https://doi.org/10.3390/rs12020302
Ho Tong Minh D, Le Toan T, Rocca F et al (2016) SAR tomography for the retrieval of forest biomass and height: cross-validation at two tropical forest sites in French Guiana. Remote Sens Environ 175:138–147. https://doi.org/10.1016/j.rse.2015.12.037
Ismail R, Mutanga O (2010) A comparison of regression tree ensembles: predicting Sirex noctilio induced water stress in Pinus patula forests of KwaZulu-Natal, South Africa. Int J Appl Earth Obs Geoinf 12:S45–S51. https://doi.org/10.1016/j.jag.2009.09.004
Jennings SB, Brown ND, Sheil D (1999) Assessing forest canopies and understorey illumination: canopy closure, canopy cover and other measures. Forestry 72:59–74. https://doi.org/10.1093/forestry/72.1.59
Kellndorfer F-A, Herndon (2019) Using SAR data for mapping deforestation and forest degradation. In: The SAR Handbook. Comprehensive Methodologies for Forest Monitoring and Biomass Estimation. Servir Global, Hunstville, AL, USA, pp 65–79
Lapini A, Pettinato S, Santi E et al (2020) Comparison of machine learning methods applied to SAR images for forest classification in Mediterranean areas. Remote Sens 12:369. https://doi.org/10.3390/rs12030369
Lucas RM, Cronin N, Lee A, Moghaddam M, Witte C, Tickle P (2006) Empirical relationships between AIRSAR backscatter and LiDAR-derived forest biomass, Queensland, Australia. Remote Sens Environ 100(3):407–425
Lucas R, Armston J, Fairfax R et al (2010) An evaluation of the ALOS PALSAR L-band backscatter—above ground biomass relationship Queensland, Australia: impacts of surface moisture condition and vegetation structure. IEEE J Sel Top Appl Earth Obs Remote Sens 3:576–593. https://doi.org/10.1109/JSTARS.2010.2086436
Ludwig M, Morgenthal T, Detsch F et al (2019) Machine learning and multi-sensor based modelling of woody vegetation in the Molopo Area, South Africa. Remote Sens Environ 222:195–203. https://doi.org/10.1016/j.rse.2018.12.019
Luo H-X, Dai S-P, Li M-F et al (2020) Comparison of machine learning algorithms for mapping mango plantations based on Gaofen-1 imagery. J Integr Agric 19:2815–2828. https://doi.org/10.1016/S2095-3119(20)63208-7
Madonsela S, Cho MA, Ramoelo A, Mutanga O (2017) Remote sensing of species diversity using Landsat 8 spectral variables. ISPRS J Photogramm Remote Sens 133:116–127. https://doi.org/10.1016/j.isprsjprs.2017.10.008
Mahlangu P, Mathieu R, Wessels K et al (2018) Indirect estimation of structural parameters in South African forests using MISR-HR and LiDAR remote sensing data. Remote Sens 10:1537. https://doi.org/10.3390/rs10101537
Marabel M, Alvarez-Taboada F (2013) Spectroscopic determination of aboveground biomass in grasslands using spectral transformations, support vector machine and partial least squares regression. Sensors 13:10027–10051. https://doi.org/10.3390/s130810027
Marvin DC, Asner GP, Knapp DE, Anderson CB, Martin RE, Sinca F, Tupayachi R (2014) Amazonian landscapes and the bias in field studies of forest structure and biomass. Proc Nat Acad Sci 111(48):E5224–E5232
Mitchard S, Lewis (2011) Measuring biomass changes due to woody encroachment and deforestation/degradation in a forest–savanna boundary region of central Africa using multi-temporal L-band radar backscatter. Remote Sens Environ 115(11):2861–2873
Mitchell AL, Rosenqvist A, Mora B (2017) Current remote sensing approaches to monitoring forest degradation in support of countries measurement, reporting and verification (MRV) systems for REDD+. Carbon Balance Manag 12:1–22
Mucina L, Geldenhuys C, Lawes M et al (2003) Classification system for South African indigenous forests. In: An objective classification for the department of water affairs and forestry. CSIR Environmentek, Pretoria, South Africa
Naidoo L, Mathieu R, Main R, et al (2014) The assessment of data mining algorithms for modelling Savannah Woody cover using multi-frequency (X-, C- and L-band) synthetic aperture radar (SAR) datasets. In: 2014 IEEE Geoscience and Remote Sensing Symposium. ieeexplore.ieee.org, pp 1049–1052
Naidoo L, Mathieu R, Main R et al (2015) Savannah woody structure modelling and mapping using multi-frequency (X-, C- and L-band) Synthetic Aperture Radar data. ISPRS J Photogramm Remote Sens 105:234–250. https://doi.org/10.1016/j.isprsjprs.2015.04.007
Naidoo L, Mathieu R, Main R et al (2016) L-band Synthetic Aperture Radar imagery performs better than optical datasets at retrieving woody fractional cover in deciduous, dry savannahs. Int J Appl Earth Obs Geoinf 52:54–64. https://doi.org/10.1016/j.jag.2016.05.006
Ndlovu NB (2013) Quantifying indigenous forest change in Dukuduku from 1960 to 2008 using GIS and remote sensing techniques to support sustainable forest management planning. Doctoral dissertation, Stellenbosch University, Stellenbosch
Ndlovu N, Luck-Vogel M, Schloms B, Cho M (2011) The quantification of human impact on the Dukuduku indigenous forest from 1960 to 2008 using GIS techniques as a basis for sustainable management. In: Fifth natural forest and wood land symposium. Department of Agriculture, Forestry and Fisheries, South Africa, KwaZulu Natal Richards Bay, South Africa
Novo EMLM, Costa MPF, Mantovani JE, Lima IBT (2010) Relationship between macrophyte stand variables and radar backscatter at L and C band Tucuruí reservoir Brazil. Int J Remote Sens 23(7):241–1260. https://doi.org/10.1080/01431160110092885
Omer G, Mutanga O, Abdel-Rahman EM, Adam E (2016) Empirical prediction of leaf area index (LAI) of endangered tree species in intact and fragmented indigenous forests ecosystems using WorldView-2 data and two robust machine learning algorithms. Remote Sens 8:324. https://doi.org/10.3390/rs8040324
Omer G, Mutanga O, Abdel-Rahman EM et al (2017) Mapping leaf nitrogen and carbon concentrations of intact and fragmented indigenous forest ecosystems using empirical modeling techniques

and WorldView-2 data. ISPRS J Photogramm Remote Sens 131:26–39. https://doi.org/10.1016/j.isprsjprs.2017.07.005
O'Neill BC, Oppenheimer M (2002) Dangerous climate impacts and the Kyoto protocol. Science 296:1971–1972. https://doi.org/10.1126/science.1071238
Pereira HM, Ferrier S, Walters M et al (2013) Essential biodiversity variables. Science 339:277–278. https://doi.org/10.1126/science.1229931
Pereira LO, Furtado LFA, Novo EMLM et al (2018) Multifrequency and full-polarimetric SAR assessment for estimating above ground biomass and leaf area index in the Amazon Várzea Wetlands. Remote Sens 10:1355. https://doi.org/10.3390/rs10091355
Puyravaud J-P (2003) Standardizing the calculation of the annual rate of deforestation. For Ecol Manag 177:593–596. https://doi.org/10.1016/S0378-1127(02)00335-3
Rosen PA, Kim Y, Kumar R, et al (2017) Global persistent SAR sampling with the NASA-ISRO SAR (NISAR) mission. In: 2017 IEEE Radar Conference (RadarConf). ieeexplore.ieee.org, pp 0410–0414
Saatchi SS, Moghaddam M (2000) Estimation of crown and stem water content and biomass of boreal forest using polarimetric SAR imagery. IEEE Trans Geosci Remote Sens 38:697–709. https://doi.org/10.1109/36.841999
Sankaran M, Hanan NP, Scholes RJ et al (2005) Determinants of woody cover in African savannas. Nature 438:846–849. https://doi.org/10.1038/nature04070
Santoro M, Shvidenko A, McCallum I et al (2007) Properties of ERS-1/2 coherence in the Siberian boreal forest and implications for stem volume retrieval. Remote Sens Environ 106:154–172. https://doi.org/10.1016/j.rse.2006.08.004
Sartori LR, Imai NN, Mura JC, Novo EMLM, Silva TSF (2011) Mapping macrophyte species in the Amazon floodplain wetlands using fully polarimetric ALOS/PALSAR data. IEEE Trans Geosci Remote Sens 49(12):4717–4728. https://doi.org/10.1109/TGRS.2011.2157972
Saunders DA, Hobbs RJ, Margules CR (1991) Biological consequences of ecosystem fragmentation: a review. Conserv Biol 5:18–32. https://doi.org/10.1111/j.1523-1739.1991.tb00384.x
Sexton JO, Song X-P, Feng M et al (2013) Global, 30-m resolution continuous fields of tree cover: Landsat-based rescaling of MODIS vegetation continuous fields with lidar-based estimates of error. Int J Digit Earth 6:427–448. https://doi.org/10.1080/17538947.2013.786146
Shimada M, Ohtaki T (2010) Generating large-scale high-quality SAR mosaic datasets: application to PALSAR data for global monitoring. IEEE J Sel Top Appl Earth Obs Remote Sens 3:637–656. https://doi.org/10.1109/JSTARS.2010.2077619
Shimada M, Itoh T, Motooka T et al (2014) New global forest/non-forest maps from ALOS PALSAR data (2007–2010). Remote Sens Environ 155:13–31. https://doi.org/10.1016/j.rse.2014.04.014
Simard M, Zhang K, Rivera-Monroy VH et al (2006) Mapping height and biomass of mangrove forests in Everglades National Park with SRTM elevation data. Photogramm Eng Remote Sens 72:299–311. https://doi.org/10.14358/PERS.72.3.299
Skowno AL, Thompson MW, Hiestermann J (2017) Woodland expansion in South African grassy biomes based on satellite observations (1990–2013): general patterns and potential drivers. Glob Chang Biol 23(6):2358–2369
Song X-P, Huang C, Feng M et al (2014) Integrating global land cover products for improved forest cover characterization: an application in North America. Int J Digit Earth 7:709–724. https://doi.org/10.1080/17538947.2013.856959
Sundnes F (2013) The past in the present: struggles over land and community in relation to the Dukuduku claim for land restitution, South Africa. Forum Dev Stud 40(1):69–86
Teferi E, Bewket W, Uhlenbrook S, Wenninger J (2013) Understanding recent land use and land cover dynamics in the source region of the Upper Blue Nile, Ethiopia: Spatially explicit statistical modelling of systematic transitions. Agric Ecosyst Environ 165:98–117. https://doi.org/10.1016/j.agee.2012.11.007
Thompson, Mackey, McNulty (2009) Forest resilience, biodiversity, and climate change. In: Secretariat of the Convention on Biological Diversity, Montreal. Technical Series no. 43. 1-67. (Vol. 43, pp. 1-67)
Tzamtzis I, Federici S, Hanle L (2019) A methodological approach for a consistent and accurate land representation using the FAO open foris collect earth tool for GHG inventories. CarbonManage 10:437–450. https://doi.org/10.1080/17583004.2019.1634934
Urbazaev M, Thiel C, Mathieu R et al (2015) Assessment of the mapping of fractional woody cover in southern African savannas using multi-temporal and polarimetric ALOS PALSAR L-band images. Remote Sens Environ 166:138–153. https://doi.org/10.1016/j.rse.2015.06.013
Urbazaev M, Thiel C, Cremer F, Dubayah R (2018) Estimation of forest aboveground biomass and uncertainties by integration of field measurements, airborne LiDAR, and SAR and optical satellite data in Mexico. Carbon Balance Manag 13(1):1–20
van Wyk GF, Everard DA, Midgley JJ, Gordon IG (1996) Classification and dynamics of a southern African subtropical coastal lowland forest. S Afr J Bot 62:133–142. https://doi.org/10.1016/S0254-6299(15)30612-8
Wang G, Gertner GZ, Fang S, Anderson AB (2005) A methodology for spatial uncertainty analysis of remote sensing and GIS products. Photogramm Eng Remote Sens 71:1423–1432. https://doi.org/10.14358/PERS.71.12.1423
Watanabe M, Koyama C, Hayashi M, et al (2018) Semi-automatic deforestation detection algorithm with PALSAR-2/ScanSAR HH/HV polarizations. In: IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium. ieeexplore.ieee.org, pp 4177–4180
Watanabe M, Koyama C, Hayashi M, et al (2020) Trial of deforestation detection by using 25m resolution PALSAR-2/ScanSAR data. In: IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium. ieeexplore.ieee.org, pp 3784–3787
Wessels K, Mathieu R, Knox N et al (2019) Mapping and monitoring fractional woody vegetation cover in the arid savannas of Namibia using LiDAR training data, machine learning, and ALOS PALSAR Data. Remote Sens 11:2633. https://doi.org/10.3390/rs11122633
Wingate VR, Phinn SR, Kuhn N, Scarth P (2018) Estimating aboveground woody biomass change in Kalahari woodland: combining field, radar, and optical data sets. Int J Remote Sens 39(2):577–606
Wood EM, Pidgeon AM, Radeloff VC, Keuler NS (2012) Image texture as a remotely sensed measure of vegetation structure. Remote Sens Environ 121:516–526. https://doi.org/10.1016/j.rse.2012.01.003
Xu L, Saatchi SS, Yang Y, Yu Y (2016) Performance of non-parametric algorithms for spatial mapping of tropical forest structure. Carbon Balance Manage 11(1):1–14
Yu Y, Saatchi S (2016) Sensitivity of L-band SAR backscatter to aboveground biomass of global forests. Remote Sens 8:522. https://doi.org/10.3390/rs8060522
Zhao P, Lu D, Wang G et al (2016) Examining spectral reflectance saturation in Landsat imagery and corresponding solutions to improve forest aboveground biomass estimation. Remote Sens 8:469. https://doi.org/10.3390/rs8060469