Journal of Applied Remote Sensing, Vol. 3, 033540 (11 August 2009)

Prediction and validation of foliage projective cover from Landsat-5 TM and Landsat-7 ETM+ imagery

John D. Armston [a], Robert J. Denham [a], Tim J. Danaher [b]
Peter F. Scarth [a], and Trevor N. Moffiet [c]

[a] Remote Sensing Center, Department of Environment and Resource Management, Climate Building, 80 Meiers Road, Indooroopilly, QLD, Australia, 4068
john.armston@qld.gov.au; robert.denham@qld.gov.au; peter.scarth@qld.gov.au

[b] Information Sciences Branch, Department of Environment and Climate Change, Suite 3, Alstonville Plaza, Main Street, Alstonville, NSW, Australia, 2477
tim.danaher@environment.nsw.gov.au

[c] Faculty of Science & Information Technology, University of Newcastle, University Drive, Callaghan, NSW, Australia, 2308
trevor.moffiet@newcastle.edu.au

Abstract. The detection of long term trends in woody vegetation in Queensland, Australia, from the Landsat-5 TM and Landsat-7 ETM+ sensors requires the automated prediction of overstorey foliage projective cover (FPC) from a large volume of Landsat imagery. This paper presents a comparison of parametric (Multiple Linear Regression, Generalized Linear Models) and machine learning (Random Forests, Support Vector Machines) regression models for predicting overstorey FPC from Landsat-5 TM and Landsat-7 ETM+ imagery. Estimates of overstorey FPC were derived from field measured stand basal area (RMSE 7.26%) for calibration of the regression models. Independent estimates of overstorey FPC were derived from field and airborne LiDAR (RMSE 5.34%) surveys for validation of model predictions. The airborne LiDAR-derived estimates of overstorey FPC enabled the *bias* and *variance* of model predictions to be quantified in regional areas. The results showed all the parametric and machine learning models had similar prediction errors (RMSE < 10%), but the machine learning models had less *bias* than the parametric models at greater than ~60% overstorey FPC. All models showed greater than 10% *bias* in plant communities with high herbaceous or understorey FPC. The results of this work indicate that use of overstorey FPC products derived from Landsat-5 TM or Landsat-7 ETM+ data in Queensland using any of the regression models requires the assumption of senescent or absent herbaceous foliage at the time of image acquisition.

Keywords: fractional cover, airborne LiDAR, stand basal area, regression, machine learning.

1 INTRODUCTION

Queensland is the second largest state in Australia and covers an area of 1.73 million km^2 therefore remote sensing is the only feasible approach to repeatable and cost-effective mapping of vegetation cover. The Statewide Landcover and Trees Study* (SLATS) has used Landsat-5 Thematic Mapper (TM) and Landsat-7 Enhanced Thematic Mapper Plus (ETM+) for mapping land cover change since 1988 [1], carbon accounting [2,3] and mapping the extent of regenerating woody vegetation [4]. These applications have assessed woody vegetation cover for single dates or changes between two dates (1 to 3 years apart), however the detection of long term trends in woody vegetation cover requires fractional estimates, which can be produced from the SLATS archive of near annual TM and ETM+ imagery.

* http://www.nrw.qld.gov.au/slats

 [DOI: 10.1117/1.3216031]
Received 23 Jul 2008; accepted 31 Jul 2009; published 11 Aug 2009 [CCC: 19313195/2009/$25.00]

The metric of vegetation cover adopted in many Australian vegetation classification frameworks is Foliage Projective Cover (FPC) [5]. FPC is defined as the vertically projected percentage cover of photosynthetic foliage of all strata [6], or equivalently, the fraction of the vertical view that is occluded by foliage. Overstorey FPC is defined as the vertically projected percentage cover of photosynthetic foliage from tree and shrub life forms greater than 2 m height and was the definition of woody vegetation cover adopted by SLATS [7]. Overstorey FPC is one minus the gap probability at a zenith angle of zero and therefore it has a logarithmic relationship with effective leaf area index [8]. Since Australian plant communities are dominated by trees and shrubs with sparse foliage and irregular crown shapes, overstorey FPC is a more suitable indicator of a plant community's radiation interception and transpiration than crown cover [9].

Operational mapping of overstorey FPC requires an efficient and automated method due to the large volume of Landsat data that require processing and interpretation. Numerous studies have evaluated different methods for estimating overstorey FPC or similar metrics over large areas for data from multi-spectral, high temporal resolution, coarse spatial resolution sensors such as AVHRR [10–12]. There has been far less work estimating overstorey FPC or similar metrics from multi-spectral medium spatial resolution sensors over large areas, especially in Australia [7,13]. This is likely due to limited ground truth data for calibration and radiometric, spatial and spectral uncertainties in Landsat imagery [14]. There have been several studies investigating different statistical regression techniques for estimating vegetation cover metrics across multiple Landsat scenes in Queensland [4,7,13,15]. Despite this, a controlled comparison of statistical regression techniques for predicting overstorey FPC over Queensland from the TM and ETM+ sensors was lacking.

Parametric techniques such as multiple linear regression (MLR) are commonly employed for the estimation of fractional vegetation cover [4,15]. MLR is similar to linear spectral unmixing when the calibration data is representative of the *a priori* distribution of the cover fraction [16]. Several studies have shown that MLR can be used to predict fractional vegetation cover provided the calibration data set is proximate to the validation data set [17–19]. Generalized linear models (GLM) are an extension of MLR that allow non-normal error distributions and bounded response variables, thus imposing additional constraints based on prior knowledge of the mechanism by which the data was generated. GLMs have been used for the prediction of vegetation cover from MODIS, a coarse spatial resolution sensor [20,21].

Machine learning techniques such as regression trees have often been used to build more advanced predictive models from remote sensing data [10,12,18]. These techniques are defined here as inductive algorithms that identify patterns and minimize prediction error through repeated, automated learning from a training dataset [22]. Numerous remote sensing studies have shown they handle non-linear relationships within high dimensional remotely sensed data sets, however a commonly cited limitation is that they over-fit the training data and do not predict accurately on independent data [13,18,22]. Random forests (RF) and support vector machines (SVM) are two recently popular machine learning algorithms that aim to minimize over-fitting and have previously been applied to classification problems with airborne hyperspectral data [23,24]. However there has been relatively little research evaluating them for regression problems with multispectral satellite data.

Validation of overstorey FPC predictions derived using any of these regression algorithms is required to: (i) select the best model to implement operationally; (ii) provide estimates of product accuracy and precision across a range of land cover types; and (iii) determine limitations of the products that need to be targeted in future research. Field acquisition of estimates of overstorey FPC can provide a highly accurate calibration or validation dataset, but the disadvantages are time constraints and the cost of acquisition which limit the number of observations that can be acquired over large areas. This makes accuracy assessment of predictions at local to regional scales impractical, particularly within remote and inaccessible areas of Queensland. Estimates of accuracy and precision of overstorey FPC predictions are

required by end users for ongoing studies attempting to detect processes such as woodland thickening and natural dieback using long-term time-series of Landsat data [3,25].

High to medium spatial resolution data are typically used as a tool for scaling up field measurements to the coarse spatial resolution of sensors such as MODIS [11,26] but have also been used for validation of products from medium spatial resolution sensors [4]. The validation approach taken in this work was to derive estimates of overstorey FPC from airborne Light Detection and Ranging (LiDAR). LiDAR records the time taken between the emission and receiving of a pulse of light and also records the intensity of received returns. A number of studies have spatially aggregated returns from airborne LiDAR sensors to simulate vertical profiles of fractional cover using either the proportion of returned energy or counts of returns [11,27–29]. Airborne LiDAR was chosen as it is able to derive spatially continuous estimates of FPC from these vertical profiles for different strata within the canopy. Previous studies have demonstrated comparable levels of accuracy and precision to field measurements of overstorey FPC for different environments in Queensland [27,28].

For any specific application it is often necessary to compare several candidate algorithms using multiple criteria other than just the accuracy of model fit to the training data [30]. The aim of this work was to conduct a controlled comparison of parametric (MLR, GLM) and machine learning regression techniques (RF, SVM) for predicting overstorey FPC across multiple Landsat scenes in Queensland, Australia. These techniques were selected for comparison because they are suitable for developing a predictive model using an extensive field dataset developed by SLATS [6]. The comparison of the regression techniques was done by assessment of the prediction error from: (i) the model fits; and (ii) validation using estimates of overstorey FPC derived from independent field and LiDAR surveys acquired over nineteen different land types in Queensland.

2 METHODS

2.1 Data acquisition and pre-processing

The 1.73 million km^2 land area of Queensland includes a wide variety of landscapes across temperate, wet and dry tropics and semi-arid to arid climatic zones (Fig. 1). Approximately 0.837 million km^2 (48%) of Queensland is covered by woody vegetation [1]. Figure 1 shows the location of field, image and airborne LiDAR survey data used in this work. The acquisition and pre-processing of these data are presented in this section.

2.1.1 Field data

The four field survey datasets sourced for this work are described in Table 1. The Kuhnell *et al.* [7], Hassett *et al.* [32] and present study field sites were located in mature undisturbed stands over a range of vegetation communities on various soil types and covering a range of structural formations from sparse low open woodland to tall closed forest communities. The Scarth *et al.* field sites were acquired on grasslands and land cleared to pasture. Estimates of stand basal area (SBA) were acquired using a calibrated optical wedge [33] or the Bitterlich method [32]. Floristic composition and Munsell soil colour were recorded at all sites.

Table 1. Description of the four field surveyed datasets used in this work.

Data Source	Observation Period	Nominal Plot Area	SBA		FPC		
			N	Counts	*N*	Point intercepts	Strata[+]
Kuhnell *et al.* [7]	1996–2005	1 ha	1397	5	221	100–200	O/U/H
Hassett *et al.* [32]	1994–1995	6 ha	51	5–12	51	200–600	O/U/H
Scarth *et al.* [15]	1999–2006	1 ha	514	5–7	514	200–300	H
Present study	2004–2005	1 ha	47	7	47	300	O/U/H

† O = overstorey (trees/shrubs > 2 m); U = understorey (trees/shrubs < 2 m); H = herbaceous.

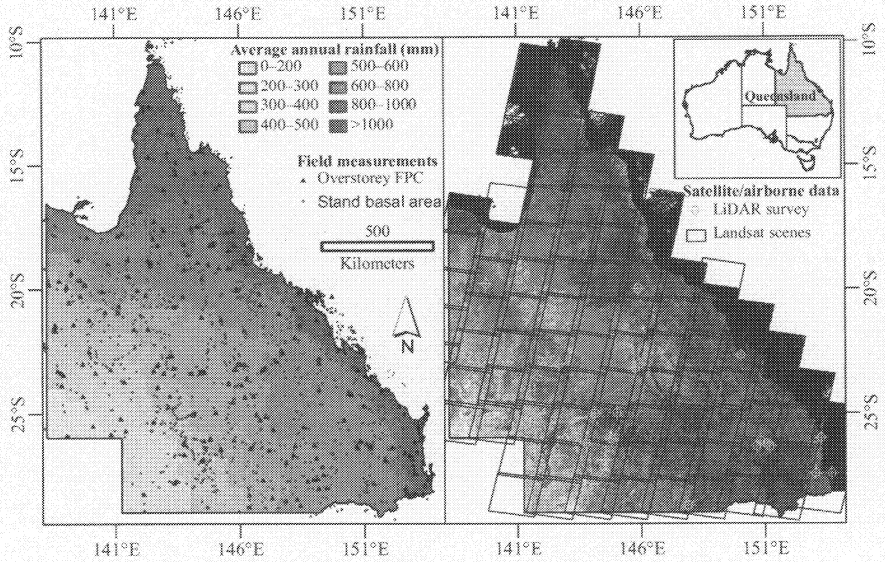

Fig. 1. Locations of field sites with stand basal area (SBA) measurements and both SBA and overstorey foliage projective cover (FPC) measurements used to develop a training dataset for the regression models in Sec. 2.2.1 (left). Locations of the airborne LiDAR survey centers and the footprints of the 88 Landsat scenes required to cover Queensland (right).

Further details of the Kuhnell *et al.* [7], Scarth *et al.* [15] and Hassett *et al.* [32] field site surveys are described therein and not repeated here. The 47 field sites surveyed for the present study were acquired using the sampling design illustrated in Fig. 2. For each 100 m by 100 m field plot, SBA estimate was an average of 7 optical wedge recordings and overstorey FPC was estimated from three 100 m point intercept transects. The advantage of this sampling design is that it is simple to implement in the field, allowing a larger number of field sites to be acquired with the resources available for field work. At 1 m intervals along each transect, overstorey (woody plants greater than or equal to 2 m height) and understorey (woody or herbaceous plants less than 2 m height) were recorded. The understorey herbaceous measurements were made with a laser pointer at a zenith of zero with intercepts classified as green leaf, dead leaf, bare, rock, cryptogam or litter by the observer. The overstorey and understorey woody plant intercepts were made using the vertical tube method [34] with intercepts classified as green leaf, dead leaf, branch or sky by the observer. The central position of each field site was measured using a sub-meter Fugro Omnistar Omnilite 132 differential global positioning system (GPS).

The Kuhnell *et al.* [7] and Hassett *et al.* [32] field sites with both overstorey FPC and SBA estimates were combined with the field sites surveyed for this study to develop the stand scale allometric relationship with SBA (Sec. 2.2.1). The Kuhnell *et al.* [7] and Scarth *et al.* [15] SBA estimates were combined for calibration of the regression models. These data were appended with 100 m × 100 m additional zero SBA sites acquired from interpretation of large scale aerial photography [4] or direct observation in the field [31] to represent spectrally dark and bright soils, respectively. These data were used for calibration of the regression models. The Hassett *et al.* [32] field data were not linked to the image data for calibration of the regression models due to low positional accuracy.

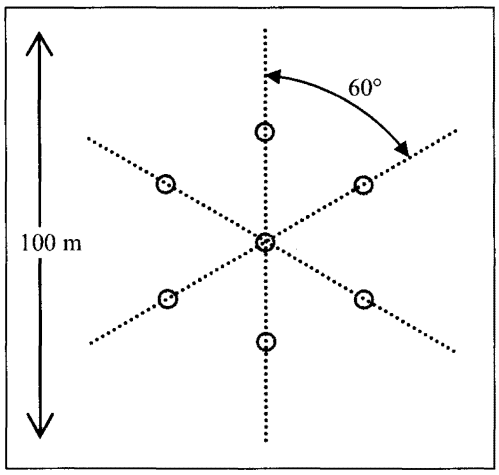

Fig. 2. Schematic of the sampling design used at the field sites coincident with the LiDAR surveys (Sec. 2.1.3). Estimates of overstorey foliage projective cover were derived from three 100 m point intercept (1 m spacing) transects oriented 0°, 60° and 120° from magnetic north (black dotted lines). Stand basal area was calculated as the average from seven optical wedge counts (green circles) located at 25 m and 75 m along each transect and the centre of the nominal 1 ha field plot.

The 47 field sites surveyed for the present study are near coincident with the LiDAR acquisitions described in Sec. 2.1.3. They were not directly used for calibration of the regression models, but reserved for derivation of overstorey FPC from the airborne LiDAR data (Sec. 3.1), and direct validation of the regression model predictions (Sec. 3.3.1). An empirical assessment of the impact of herbaceous FPC on the regression model predictions of overstorey FPC was done using a subset of 198 field sites from Scarth *et al.* [15] that provided estimates of herbaceous FPC in areas of zero SBA.

2.1.2 Image data

TM and ETM+ imagery were acquired with Australian Centre for Remote Sensing (ACRES) Level-5 processing[†]. The ACRES Level-5 processing includes systematic radiometric and geometric corrections and two dimensional resampling to fit a specific earth datum and map grid. It is similar to the USGS Level 1G processing. Scenes providing near-annual acquisitions between 1987 and 2005 were used in this work. The footprints of the 88 scenes required to cover Queensland are shown in Fig. 1. The selection of image dates was restricted to dry season months (May to October inclusive) in order to minimize cloud cover and photosynthetic herbaceous ground cover which reduces the spectral contrast between the foliage of woody and herbaceous plants.

All TM and ETM+ images were geometrically registered to an orthorectified ETM+ image mosaic based on differential GPS ground control points [35]. The onboard radiometric calibration for TM was removed and replaced with a vicarious calibration based on a model of the lifetime response of the sensor [31]. Pre-flight calibration was used for ETM+. An empirical radiometric calibration was applied to top-of-atmosphere (TOA) reflectance to remove combined surface and atmospheric bi-directional reflection distribution function (BRDF) effects [36]. In order to minimize scene to scene factors of uncertainty in the training

[†] http://www.ga.gov.au/acres/techdocs/techdoc.jsp

data, Band 1 was excluded due to remaining atmospheric effects from Rayleigh scattering [36]. Reflectance values affected by cloud, cloud shadow, land clearing, surface water or topographic shadow in any image date were removed using Landsat-derived masks [37].

Vapour pressure deficit (VPD) was included as a predictor as the evaporative power of the atmosphere in the boundary layer above the canopy has been shown to be strongly correlated with overstorey FPC [9,38]. VPD is defined as the difference between actual vapour pressure and vapour pressure under saturated atmospheric conditions therefore it is a direct measure of atmospheric demand for moisture. VPD was calculated as a long-term average of interpolated 5 km spatial resolution daily grids derived by Jeffrey *et al.* [39] from 1957 to 2005.

2.1.3 Airborne LiDAR data

A total of nineteen LiDAR flight paths were acquired by a commercial data provider using an Optech ALTM3025 laser scanner. Each of the transects were between 10 and 20 km long with approximately a 300 m swath (maximum scan angle of 10°), average sample spacing of 0.93 m (scan rate of 25,000Hz) and a beam divergence of 0.3 milliradians (0.23 m footprint). It is important to note that the data could only be acquired opportunistically due to the purchasing arrangement with the data provider and the final positions of transects were determined by vehicle access so the sampling design was not random. The aim was to sample the range of structural formations and remnant vegetation communities that are dominant in Queensland so the acquisitions were stratified as much as possible. Expert knowledge and qualitative observations from extensive field work [32], 1:100,000 scale maps of Regional Ecosystems [40] and Landsat overstorey FPC products (MLR model) were used to define the 19 regions within which LiDAR surveys were able to be acquired.

Descriptions of the LiDAR surveys are provided in Table 2 and their locations shown in Fig. 1. Field work was conducted at 16 of the 19 survey sites and, if possible, was completed within one month of the LiDAR acquisition. Section 2.1.1 describes the procedures followed at the 47 sites. Up to four sites were acquired for each survey, each following the sampling design shown in Fig. 2. Example photographs indicating the range and structure of plant communities sampled are shown in Fig. 3.

The Optech ALTM 3025 sensor records the range to maximum power of the first and last peaks in the time distribution of the return signal. A detection threshold determines the minimum power that can qualify as a peak in the return signal. A scaled value of any detected peaks is recorded by the sensor as the "intensity". The commercial data provider processed this information using Optech's REALM software to provide time sequential format ASCII files for each flight path. Each file contained fields of the easting, northing, elevation above sea level and intensity of first and last returns in order of time received at the sensor. Data were projected to the Geodetic Datum of Australia 1994. All subsequent processing routines were implemented in IDL® 6.2 [41].

The next step in analyzing these data was to classify returns originating from the ground. Due to reset delays in the circuitry of the Optech ALTM 3025 sensor, the last return has to be greater than 4.9 m after the first return otherwise both timing circuits will measure the same return [42]. Consequently, last returns of pulses with the first and last return elevation difference less than 4.9 m were discarded from the datasets. A progressive morphological filter of last returns based on the algorithm of Zhang *et al.* [43] was then implemented. Only the last returns were filtered because they have greater penetration through the canopy. Elevation of the ground at the position of non-ground returns was estimated by inverse distance weighted interpolation with an exponent of two. The height of all returns above the ground was then calculated by subtracting the ground elevation from the return elevation. Solitary pulses with unrealistic heights were discarded.

Table 2. Survey name (number of field sites), dates of the airborne LiDAR and field survey acquisitions, centre location and description of the plant community structure, dominant woody species and substrate.

Survey	Acquisition dates	Location	Description
adav01(4)	LiDAR: 2004/03/22 Field: 2004/04/23	144.56°E 26.02°S 291 m ASL	Low woodland (*Acacia anuera*). Red sandy clay loam soils.
adav02(2)	LiDAR: 2004/03/23 Field: 2004/04/24	144.76°E 24.98°S 364 m ASL	Woodland (*Acacia harpophylla, Acacia catenulata, Dodonaea sp.*). Red clay to sandy-clay loam with scattered silcrete stone.
adav03(3)	LiDAR: 2004/03/22 Field: 2004/04/25	143.89°E 24.99°S 188 m ASL	Low open woodland (*Acacia cambagei, A.anuera*). Red gravely clays and texture contrast soils.
aram01(0)	LiDAR: 2004/03/23 Field: -	145.53°E 23.17°S 277 m ASL	Low woodland (*A.cambagei*). Brown-reddish and grey cracking clay soils with light stone cover.
aram02(4)	LiDAR: 2004/03/23 Field: 2004/04/29	145.72°E 23.00°S 480 m ASL	Open woodland (*Eucalyptus spp., Melaleuca sp., A.shirleyi*). Loamy red/yellow soils on sand plains.
cade01(2)	LiDAR: 2004/03/23 Field: 2004/04/26	142.45°E 23.03°S 250 m ASL	Low open woodland (*A.aneura, A.shirleyi*). Stony lithosols with areas of weathered rock outcropping.
chat01(3)	LiDAR: 2004/03/25 Field: 2004/07/22	145.65°E 19.98°S 381 m ASL	Open woodland (*Eucalyptus spp.*). Alluvial plains with clays and texture contrast soils.
chin01(2)	LiDAR: 2005/06/23 Field: 2005/07/11	150.45°E 26.34°S 336 m ASL	Open forest (*Callitris sp.*). Duplex soils with sandy surfaces.
chin02(3)	LiDAR: 2005/06/23 Field: 2005/07/11	150.61°E 26.14°S 334 m ASL	Open forest (*E.microcarpa, Eucalyptus populnea, Callitris sp.*). Deep sandy soils.
chin03(1)	LiDAR: 2005/06/23 Field: 2005/07/12	150.82°E 26.24°S 446 m ASL	Tall closed forest (*Eucalyptus spp., A.harpophylla*). Deep sandy soils.
chin04(4)	LiDAR: 2005/06/23 Field: 2005/07/13	150.98°E 26.33°S 398 m ASL	Tall open forest (*Corymbia maculata, E.moluccana, Angophora costata*). Lateritic duricrust soils.
dirr01(0)	LiDAR: 2004/03/22 Field: -	147.65°E 28.70°S 153 m ASL	Woodland (*E. coolabah*). Clay or sometimes texture contrast soils.
gold01(2)	LiDAR: 2005/05/03 Field: 2005/08/19	152.83°E 27.43°S 393 m ASL	Tall open forest (*E.racemosa, Corymbia gummifera, Banksia aemula*). Coastal dunes with leached sandy soils.
gold02(2)	LiDAR: 2005/04/01 Field: 2005/07/29	153.49°E 27.44°S 114 m ASL	Open forest (*Eucalyptus spp.*). Soils consist of metamorphosed sedimentaries and interbedded volcanics.
gunp01(3)	LiDAR: 2004/03/23 Field: 2004/04/27	139.50°E 19.79°S 348 m ASL	Low woodland (*Corymbia papuana, A.shirleyi, Eucalyptus spp.*). Skeletal soils with iron stone scattered on surface.
quil01(4)	LiDAR: 2004/03/22 Field: 2004/04/21	144.27°E 26.88°S 185 m ASL	Open woodland (*A.cambagei, E.ochrophloia*). Grey and brown clays of light to medium textures.
quil02(4)	LiDAR: 2004/03/22 Field: 2004/04/22	143.53°E 26.21°S 160 m ASL	Low woodland (*A.cambagei, Corymbia terminalis*). Reddish-brown cracking clays or texture contrast soils with sandy surfaces.
stla01(0)	LiDAR: 2005/06/24 Field: -	149.81°E 22.65°S 49 m ASL	Woodland (*Eucalyptus spp., Allocasurina sp.*). Bleached sodic duplex soils.
suns01(4)	LiDAR: 2005/05/31 Field: 2005/08/11	153.05°E 26.01°S 70 m ASL	Closed forest to low open woodland (*E.racemosa, B.aemula, mixed species notophyll vine forest*). Deeply leached sandy soils on sand dunes/plains.

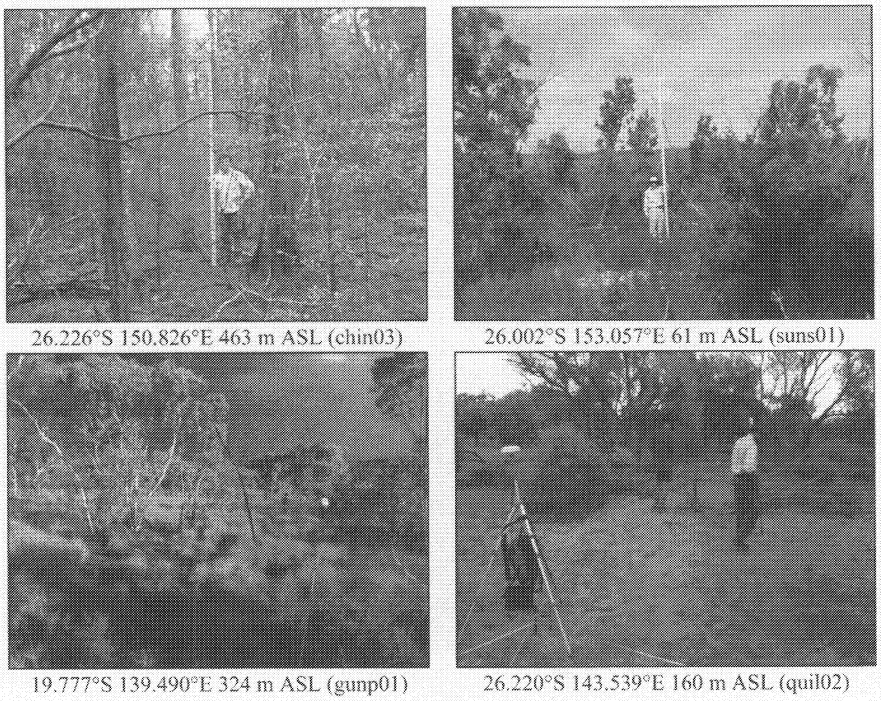

Fig. 3. Example photographs of four field sites coincident with the airborne LiDAR surveys.

2.2 Linking the field, airborne LiDAR and image data

2.2.1 Estimation of overstorey foliage projective cover from field measured stand basal area

All field sites with both optical wedge count and point intercept transect measurements were used to develop a stand scale allometric relationship between overstorey FPC and SBA for tree and shrub life forms. Specht and Specht [9] indicate this allometric relationship is valid for the arid to the prehumid zone in Queensland.

In order to minimize the impact of measurement error on the derived allometric relationship, only sites with a total point intercept transect length greater than 100 m were included. This resulted in a total of 126 field sites with coincident estimates of FPC derived using the point intercept transect method (FPC_T) and SBA. Percentage overstorey FPC were derived from the point intercept transects by

$$FPC_T = \frac{100 P_{OGL}}{1 - P_{OB}}, \quad (1)$$

where P_{OGL} and P_{OB} are the fraction of overstorey intercepts that were classified as green leaf and branch, respectively. Branch intercepts were removed from the sample as they are likely to occlude photosynthetic foliage from the viewpoint of the observer. SBA was calculated as the mean of all calibrated optical wedge counts at each site.

An exponential model was used for this bounded (0–100%) relationship

$$FPC_A = 100\left(1 - e^{\frac{SBA}{a+bSBA}}\right), \quad (2)$$

where a and b are parameters optimized from the field data and FPC_A is the allometric estimate of percentage overstorey FPC. In order to account for measurement error in the calibration and restrict all predictions to the bounds of the observation space, a Bayesian errors-in-variable model was used to estimate 'true' overstorey FPC. Residual error in each field observation of SBA and FPC_T were assumed to be from individual Gamma and Beta distributions respectively. Model fitting is by Markov Chain Monte Carlo (MCMC) methods implemented in the Bayesian modeling software WinBUGS[‡] and is described further by Moffiet [44].

2.2.2 Estimation of overstorey foliage projective cover from airborne LiDAR

LiDAR fractional cover is defined here as one minus the gap fraction probability P_{gap} at a zenith of zero. It was calculated from the proportion of first return counts by

$$1 - P_{gap}(z) = \frac{C_V(z)}{C_V(0) + C_G}, \quad (3)$$

where $C_V(z)$ is the number of first returns higher than z m above the ground and C_G is the number of first return counts from the ground. All pulse scan angles were assumed to be zero.

Images of the fractional cover estimates were calculated by aggregating all pulses into 25 m spatial resolution bins and applying Eq. (3). For estimates of fractional cover z was set to 0.5 m because overstorey individuals greater than 2 m height often have foliage lower than 2 m height above the ground. This value of z also reduces the impact of understorey and ground (e.g. litter, termite mounds) features, which are difficult to separate below 0.5 m.

Calibration of LiDAR fractional cover to estimates to overstorey FPC was performed using the 47 coincident FPC_T estimates. The fraction of LiDAR pulses intercepted by a canopy above height z is determined by FPC but calibration is required to account for two sources of error: (i) LiDAR directly estimates overstorey plant projected cover (PPC) rather than overstorey FPC because it cannot discriminate photosynthetic from non-photosynthetic foliage; and (ii) the effect of extrinsic (minimum return intensity required to register a return at the sensor; altitude and beam divergence) sampling properties of the LiDAR survey [42]. Calibrations of LiDAR fractional cover to overstorey FPC using counts of returns have in the past been linear [26,27], however these calibrations have been site and sensor specific and have not sampled high values of overstorey FPC. In this work a non-linear power function was used and had the property of being bounded between 0% and 100%,

$$FPC_L = 1 - P_{gap}^{\ a}, \quad (4)$$

where a is the exponent to be optimized using field estimates of FPC_T. Residual error in each observation of LiDAR P_{gap} and FPC_T were assumed to be from individual Beta distributions [44].

2.2.3 Relating the image data to the field and LiDAR data

For all TM and ETM+ images spatially coincident with the field sites used for calibration of the regression models (Sec. 2.1.1), DNs from bands 2 to 7 were extracted for a 3 by 3 pixel block centered on the field site location and then averaged. This provided the best match to the spatial extent of field measurements and also minimized the impact of geometric misregistration between image and field data.

[‡] http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml

For the airborne LiDAR, 3 by 3 pixel blocks were extracted from the nearest cloud free TM image in the SLATS image archive, averaged and then matched to 3 by 3 LiDAR bins of 25 m spatial resolution (Sec. 2.2.2). Again, this was to minimize the impact of geometric misregistration and spatial regularization of the TM data due to the sensor point spread function, adjacency effects and resampling of the data. LiDAR bins with overstorey FPC greater than zero and all returns below 2 m height above ground were excluded due to confusion in the separation of overstorey and understorey FPC for heathland vegetation.

The use of repeat Landsat observations spanning up to 18 years for each field site requires the assumption of negligible change in SBA over time. All field sites were located in mature undisturbed vegetation however the following sources of temporal change are present:

- Tropical savanna woody species are the predominant vegetation type in northern Queensland and exhibit a marked increase in litter fall and hence reduced overstorey FPC during the dry season. Williams *et al.* [45] showed the within crown overstorey FPC changed up to 20% between the wet and dry seasons for some evergreen species and up to 40% when averaged over 49 species including evergreen, partly and fully deciduous.
- Drought related tree dieback decreases the live SBA of a forest stand and hence the overstorey FPC. Extensive dieback in Eucalyptus dominated stands have been reported in northern Queensland during the 1990's drought [25].
- Woody thickening has occurred in places in response to climate change, altered fire regimes and introduction of livestock [3,46]. Averaged over 57 monitoring sites in Queensland, Burrows *et al.*[3] reported an increase of 1.06 m^2 ha^{-1} SBA over 14 years.

Substantial spectral variation can be expected between Landsat images acquired over tropical savannahs at different times of the year due to green leaf phenology; and between Landsat images acquired in different years due to long term changes in SBA. The effect of green leaf phenology was controlled for to some extent by selection of dry season image dates. To mitigate the effect of long term changes in overstorey FPC, each Landsat date for a field site was treated as an independent observation in the regression analysis and was given a weighting of the number of days between the field and image acquisition dates, expressed as a fraction of the maximum number of days for all observations.

2.3 Specification of the regression models

Table 3 shows the six different regression models compared in this work. Non-linear variants of the MLR models and the GLMs were created by natural spline transformations of the predictors (Sec. 2.3.1).

Table 3. The parametric (MLR, MLR-S, GLM, GLM-S) and machine learning (SVM, RF) regression models compared in this work. The model name acronyms are used throughout the text.

Model name	Regression technique
MLR	Multiple linear regression
MLR-S	Multiple linear regression with natural splines
QLR	Generalized linear model (quasi-likelihood)
QLR-S	Generalized linear model (quasi-likelihood) with natural splines
SVM	Support vector machines
RF	Random forests

All regression models were trained using the open source software R (Version 2.5.1) [47]. Weighting of individual observations was not available for the R implementation of the RF and SVM regression algorithms.

2.3.1 Transformation of the predictors

Transformations were performed in order to meet common assumptions of the GLM and MLR models. The frequency distribution of each predictor was inspected for normality and the most appropriate transformation for bands 2, 3, 4 and 7 was their natural logarithm. No transformation was required for band 5 and a reciprocal transformation was applied to VPD. SVM and RF can model highly dimensional data and non-linear relationships, therefore no transformations of the Landsat bands and VPD were necessary for these models.

The linearity assumption for the relationship between the predictors and the response variable can be relaxed through transformation of the predictors; the simplest way being the inclusion of higher powers of the explanatory variables (X) in the model. However polynomials do not fit logarithmic functions well and can produce spurious predictions at the peaks, valleys and edges of X. Restricted cubic splines, also called natural splines, overcome these problems by fitting third order polynomials within intervals of X; the join points of intervals are referred to as knots [48]. Natural splines are smooth at the knots, fit highly curved functions well and are constrained to be linear at the edges of X. Typically between 3 and 5 knots are used depending on the degree of non-linearity of the problem and the number of observations [48]. Three knots were used in this work to minimize the risk of over-fitting and were positioned at the 0.1, 0.5 and 0.9 quantiles of X (MLR-S and GLM-S models).

2.3.2 Multiple linear regression

The MLR model can be written as

$$Y = \sum_{i=1}^{p} X_i \beta_i + \varepsilon, \tag{5}$$

where β_i are the regression coefficients for the intercept and p-1 predictors (X) and ε is the unexplained variance with ε_i independent $N(0,\sigma^2)$. The predictors consist of transformed Landsat bands 2–7, cross products of these bands (e.g. X_2X_3) and VPD. The cross-products account for interactions between the Landsat bands where the relationship between the response variable (Y) and one band can change depending on the value of another band. In the case of predicting overstorey FPC from Landsat, Lucas *et al.* [4] found that incorporating interaction terms into a multiple linear regression model reduced the unexplained variance in the response variable.

Candidate MLR models were generated for all possible combinations of predictors. Separate MLR models were calibrated for the TM and ETM+ sensors due to differences in the spectral bandwidth and response, particularly in band 5 and the lack of an operational radiometric cross-calibration [31]. Including a 'dummy' variable in the MLR model for the sensor would allow the model intercept to be adjusted but not the interaction terms. Incorporating additional 'dummy' variables to model the effect of the sensor on the Landsat band interactions is valid, but was not implemented to avoid excessively complex models.

2.3.3 Generalized linear models

The GLM [49] is an extension of MLR and can be written as

$$E(Y) = \mu = g(\eta), \tag{6}$$

where $E(Y)$ is the expected value of Y, which is assumed to be from a distribution function (assumed to be normal for MLR). The mean value of each Y is equal to μ and is related to the linear predictor η, which is equal to the right hand side of Eq. (5), by a link function (g) that must be monotonic and differentiable. In this work FPC_A was not generated from binomial sampling therefore we cannot directly specify a binomial probability distribution to the data. Instead we used a quasi-likelihood approach [49].

The quasi-likelihood approach only requires specification of how the mean is related to the predictors (the link function) and how the variance is related to the mean (the variance function), not the full distribution. As the range of possible response values is bounded (0–100%) and was calibrated using measurements generated by binomial sampling, we assume the variance function resembles that of a binomial distribution. Consistent with logistic regression, a logit link function was used to relate FPC_A to the linear predictors and $V(\mu) = \mu(1-\mu)$ was used as the variance function where μ is the mean FPC_A.

Candidate GLMs were generated for all possible combinations of predictors. For the same reasons as the MLR models, separate GLMs were calibrated for the TM and ETM+ sensors. The predictors specified for the MLR models were also used for the GLMs.

2.3.4 Support vector machines

The objective of SVM regression is to find the least complex continuous function $f(X)$ where the deviation from the response is no larger than ε for all the training data. The SVM model only depends on a subset of the training data (termed support vectors), because the loss function ignores any training data that are close (within the threshold ε) to the model estimate. Non-linearity is handled by substituting a radial basis kernel function into the model optimization, which maps the input (X) to higher dimensional feature space. For a more detailed mathematical explanation of SVM regression see Smola and Scholkopf [50].

In addition to ε the training of the SVM model depends on the values of the regularization constant (C) and the Parzen window width for the radial basis kernel function (γ). C determines the trade-off between model complexity and the degree to which the deviations greater than ε are tolerated. In order to avoid a computationally intensive grid search to determine optimal values of γ, C and ε simultaneously, the approach of Cherkassky and Ma [51] was adopted. This analytical derivation of C and ε is based on a theoretical understanding of SVM. The outlier resistant calculation of C was

$$C = \max\left(\left|\bar{y} + 3\sigma_{\bar{y}}\right|, \left|\bar{y} - 3\sigma_{\bar{y}}\right|\right), \tag{7}$$

where $\bar{y}$ and $\sigma_{\bar{y}}$ are the mean and standard deviation of the response variable (FPC_A), respectively [51]. The parameter ε was calculated as

$$\varepsilon = 3\sigma\sqrt{\frac{\ln(n)}{n}}, \tag{8}$$

where σ is the standard deviation of the noise in the data, and n is the number of observations. Equation (8) is based on the knowledge that ε is proportional to σ [51]. We calculated σ by following the k–nearest-neighbour regression procedure of Cherkassky and Ma [51]. The value for γ was calculated by empirically tuning the SVM model using a grid search with a random subset of 10% of the calibration data. Other kernel functions (e.g. polynomial) and their parameters were tested but the radial basis function provided marginally better fits to the training data. These results are not presented here for brevity.

2.3.5 Random forests

Random Forests are an extension of the classification and regression tree (CART) algorithm [52]. A forest is a user determined number of decision trees. Each decision tree is grown by randomly sampling approximately two thirds of the training data with replacement (termed bagging). At each node of each tree, m variables are randomly selected from the set of predictors. The best split using these m variables is the split point and predictor that results in the greatest reduction in residual sums of squares between the sample of observations and the node mean. This process is used to perform recursive binary splits of the data.

The value of m is held constant and each decision tree is grown to the largest extent possible (no pruning). This process is repeated to build an ensemble of regression trees with each tree being a low bias, high variance regression model. The predictions are then averaged to calculate a final estimate of the response variable. The ensemble averaging results in a low bias, low variance regression model that is resistant to over-fitting [53].

The parameters m and n (the number of regression trees in the forest) need to be set to minimize the prediction error of the forest while minimizing computer processing time. The forest prediction error (PE_F) depends on: (i) the mean correlation between any two trees in the forest ($\bar{\rho}$); and (ii) the prediction error of a single tree in the forest (PE_T) [53],

$$PE_F \leq \bar{\rho}\, PE_T . \tag{9}$$

When m is equal to 1, $\bar{\rho}$ is low and PE_T high, and when m is raised to p, $\bar{\rho}$ is increased and PE_T reduced. Therefore the selection of a value for m is a trade-off between $\bar{\rho}$ and PE_T. The prediction error of random forests is largely insensitive to m [53] therefore it was calculated as the recommended $p/3$ to avoid a computationally intensive grid search.

2.4 Comparison of the regression models

2.4.1 Model selection

The error statistic used for the selection of models for each of the six regression techniques was the root mean squared error (RMSE),

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} E_i^2}\,, \tag{10}$$

where N is the number of observations and the error, $E_i = y_i - f(x_i)$, the difference between observed and predicted values of overstorey FPC, respectively.

The candidate models for each parametric regression technique were grouped by number of predictors and then sorted by RMSE. The five best fitting models in each group were retained. In order to ensure over-fitting was not occurring and to get an unbiased estimate of prediction error from the training data, a bootstrap estimate (25 samples with replacement) [48] of RMSE was then calculated for each model. The model with the lowest bootstrap RMSE in each group was then retained.

The optimum number of predictors was selected as the minimum number of terms of models with a percentage difference of the bootstrap RMSE relative to the minimum of all candidate models of < 0.1%. Although multi-collinearity makes it difficult to make inferences from the model coefficients as their standard errors are inflated [48] it was not considered problematic as the objective was simply to minimize the bootstrap RMSE.

The bootstrap RMSE of the RF and SVM regression models was also calculated using the same procedure. The R implementation of RF produces an out-of-bag error rate that is an unbiased estimate of prediction error, however for consistent comparison the same bootstrap approach used for the parametric models was followed. Candidate RF models were generated by increasing the number of trees (n) by 2^x, $x = 1, 2 \ldots 10$.

The comparison of the regression models then consisted of inspection of the model's prediction error (Sec. 3.2). It is important to note that visual inspection of overstorey FPC images derived from each of the regression models was also part of the model comparison.

2.4.2 Validation of the regression model predictions

Equation (10) squared is the mean squared error (MSE), which is decomposed into *bias* and *variance* error components. The *variance* error is used here as an indicator of the precision of

regression model predictions of overstorey FPC in comparison to the independent field and airborne LiDAR-derived estimates,

$$variance = \frac{1}{N}\sum_{i=1}^{N}\left(E_i - \overline{E}\right)^2, \tag{11}$$

where $\overline{E}$ is the mean error. The *bias* error is the average difference, an indicator of accuracy, between the regression model predictions of overstorey FPC and the independent field and airborne LiDAR-derived estimates,

$$bias = \frac{1}{N}\sum_{i=1}^{N}E_i. \tag{12}$$

The RMSE is then equal to $\sqrt{variance + bias^2}$. The validation of the regression models was done by assessment of the r^2, RMSE, *bias* and *variance* of model predictions in comparison with independent field and LiDAR survey estimates of overstorey FPC (Sec. 3.3).

3 RESULTS AND DISCUSSION

3.1 Evaluation of stand basal area and LiDAR-derived estimates of overstorey foliage projective cover

Figure 4 shows the field data and fitted model with a = -38.6 and b = 0.359. The residuals are larger at higher SBA (> 40 m^2 ha^{-1}) which is consistent with the Gamma error distribution. Eq. (2) assumes leaf angle is a function of SBA. There was no correspondence between the dominant genus and the magnitude of the residuals, suggesting differences in canopy structure between vegetation types do not contribute greatly to the observed scatter. Unfortunately, there were insufficient coincident field estimates of foliage clumping or leaf angle distribution to support further analysis of the effect of related canopy variables on the relationship at these sites. Despite these limitations the agreement is excellent (RMSE = 7.26% FPC_T) and the residuals are consistent with that expected from the joint Beta and Gamma probability distributions. The derived parameters a and b for Eq. (2) were used to calculate FPC_A using all field estimates of SBA.

There is excellent agreement between the field estimates of FPC_T and LiDAR-derived fractional cover and the residuals are consistent with a binomial sampling distribution (Fig. 5 left; RMSE = 5.34% FPC_T). The non-linear relationship (a = 0.4802), Eq. (4), indicates the sampling properties of the sensor are causing LiDAR fractional cover to systematically overestimate overstorey FPC. This is due to the footprint size of the LiDAR pulse being blind to small gaps in the foliage detected using the point intercept field technique. As overstorey FPC increases, the gap size distribution is increasingly dominated by gaps smaller than the LiDAR pulse footprint. It is assumed the gap size distribution changes as a function of overstorey FPC. This assumption is important because it is possible for the gap size distribution to change with inclination angle and clumping of leaves without any change in LiDAR fractional cover.

Published measurements of Australian woody species show the leaf size of many species is less than the LiDAR beam cross-sectional area used in this work and that there is considerable natural variation in leaf angle [54]. Therefore increased scatter in the FPC_T and LiDAR fractional cover relationship is possible due to variation in canopy structure and its interaction with the minimum intensity threshold required for a return to be detected by the sensor [29]. The intensity also depends on the reflectivity as well as the structural characteristics of the intercepted surface. Despite these sources of variation not being controlled for, the small residuals in Fig. 5 (left) suggest the assumption that the gap size distribution changes as a function of overstorey FPC is generally valid.

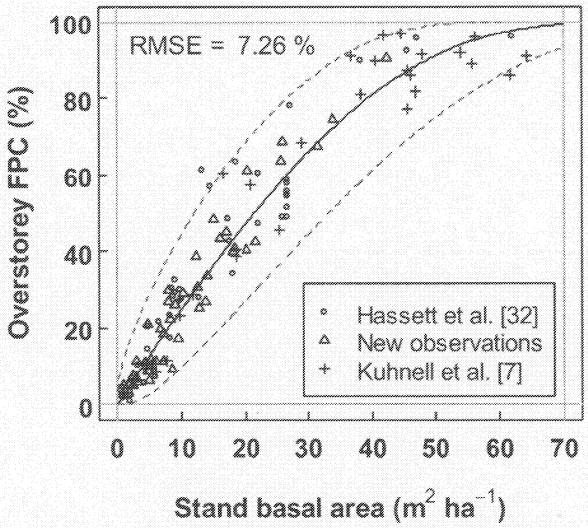

Fig. 4. The relationship between overstorey foliage projective cover (FPC) and stand basal area. Symbols correspond to the source of the field data (Sec. 2.1.1). The prediction intervals (dashed) and fitted (solid) lines are shown.

Overstorey FPC is always less than or equal to overstorey plant projective cover (PPC). The derivation of overstorey FPC using Eqs. (3) and (4) also assume the relative proportions of photosynthetic and non-photosynthetic plant material visible to the LiDAR sensor are constant or change as a function of overstorey FPC. Figure 5 (right) shows that there is greater variation in FPC_T around 50% PPC_T and far less towards 0% and 100% PPC_T, which is consistent with a binomial sampling distribution. Figure 5 (right) is suggested to represent the maximum envelop of overstorey FPC-PPC difference for remnant vegetation due to tree architecture. This is simply because field estimates are derived "looking up" while LiDAR estimates are "looking down" so photosynthetic foliage tends to occlude the stems and branches from the view of the LiDAR sensor.

The LiDAR estimates of fractional cover were calibrated to FPC_L using Eq. (4). This calibration will require further validation for acquisitions in new areas due to the limitations discussed above. Any changes in the intrinsic or extrinsic sampling properties of the LiDAR sensor will also require re-calibration. The return intensity is sensitive to the area of intercepted surfaces and be can be used to derive estimates of overstorey FPC using the Optech ALTM3025 sensor [11,29]. However its calibration is unknown and requires assumptions to be made about the relative reflectivity of the foliage and ground material and the intensity response of the sensor.

3.2 Comparison of the regression models' prediction error

The final models selected were the most parsimonious possible for each regression technique using the criterion outlined in Sec. 2.4. All the selected parametric models had nine predictors (5 bands, 3 interactions, VPD). The SVM parameters C, ε and γ were calculated as 75.78, 0.423 and 0.25 respectively. The final RF model had n set to 64, after which reductions in RMSE were < 0.1% FPC_A. All the selected regression models fitted the training data well

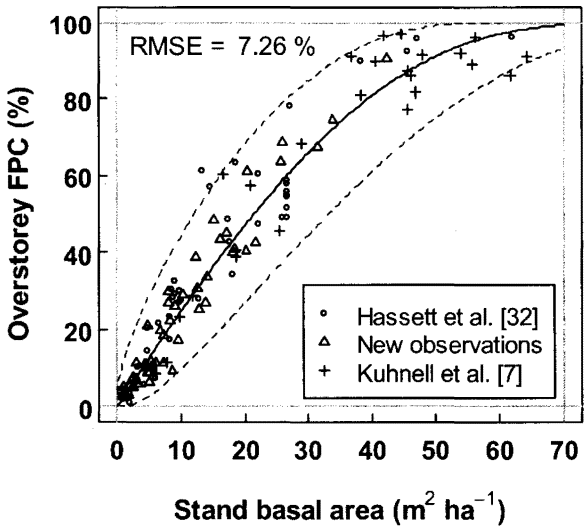

Fig. 4. The relationship between overstorey foliage projective cover (FPC) and stand basal area. Symbols correspond to the source of the field data (Sec. 2.1.1). The prediction intervals (dashed) and fitted (solid) lines are shown.

Overstorey FPC is always less than or equal to overstorey plant projective cover (PPC). The derivation of overstorey FPC using Eqs. (3) and (4) also assume the relative proportions of photosynthetic and non-photosynthetic plant material visible to the LiDAR sensor are constant or change as a function of overstorey FPC. Figure 5 (right) shows that there is greater variation in FPC_T around 50% PPC_T and far less towards 0% and 100% PPC_T, which is consistent with a binomial sampling distribution. Figure 5 (right) is suggested to represent the maximum envelop of overstorey FPC-PPC difference for remnant vegetation due to tree architecture. This is simply because field estimates are derived "looking up" while LiDAR estimates are "looking down" so photosynthetic foliage tends to occlude the stems and branches from the view of the LiDAR sensor.

The LiDAR estimates of fractional cover were calibrated to FPC_L using Eq. (4). This calibration will require further validation for acquisitions in new areas due to the limitations discussed above. Any changes in the intrinsic or extrinsic sampling properties of the LiDAR sensor will also require re-calibration. The return intensity is sensitive to the area of intercepted surfaces and be can be used to derive estimates of overstorey FPC using the Optech ALTM3025 sensor [11,29]. However its calibration is unknown and requires assumptions to be made about the relative reflectivity of the foliage and ground material and the intensity response of the sensor.

3.2 Comparison of the regression models' prediction error

The final models selected were the most parsimonious possible for each regression technique using the criterion outlined in Sec. 2.4. All the selected parametric models had nine predictors (5 bands, 3 interactions, VPD). The SVM parameters C, ε and γ were calculated as 75.78, 0.423 and 0.25 respectively. The final RF model had n set to 64, after which reductions in RMSE were < 0.1% FPC_A. All the selected regression models fitted the training data well

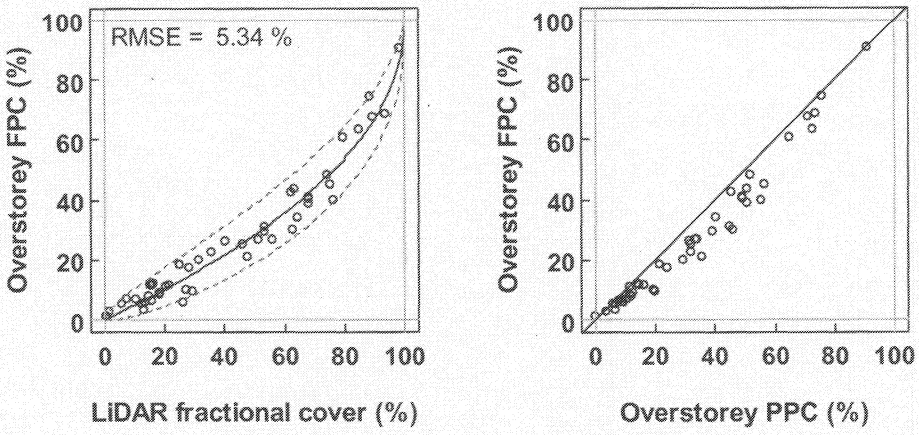

Fig. 5. The relationship between: (left) field estimated overstorey foliage projective cover (FPC) and LiDAR fractional cover; and (right) field estimated overstorey plant projective cover and overstorey FPC.

(Table 4) with all bootstrap RMSE estimates ≤ 10% FPC_A. The best results were for the RF model for both TM and ETM+. The SVM model also provided ~1% FPC_A improvement on the parametric models. There was less than 1% FPC_A difference between fits of the MLR and GLM models and their non-linear counterparts (MLR-S and GLM-S), indicating the unexplained variance in the MLR and GLM models are not due to simple non-linear relationships between the response and predictors. The differences between the models for RMSE, bootstrap RMSE, *bias* and *variance* were not significant between any of the models (< 3% FPC_A). These results are not due to over-fitting because of the large number of observations and the use of bootstrapping to calculate the error statistics.

The error statistics in Table 4 indicates all the models fit equally well however these 'global' statistics are weighted *a priori* to ranges of FPC_A where there were large proportions of the calibration data. Figure 6 is a check of the bootstrap-derived residuals against the fitted values and shows how well the models fit at different ranges of FPC_A. The RF and SVM techniques are the most appropriate models for the data because the median of the residuals were close to zero through the range of FPC_A compared to the other models. The MLR and GLM techniques show greatest difference at ≥ 40% FPC_A.

Table 4. Model and bootstrap error statistics calculated from the fitted models selected for each regression technique. Statistics calculated separately for the Landsat-5 TM and Landsat-7 ETM+ observation are shown in parentheses respectively. Sec 2.4 defines the error metrics used.

Model	RMSE	Bootstrap RMSE	Bootstrap *variance*	Bootstrap *bias*
MLR	9.90 (9.95, 9.69)	9.89 (9.96, 9.65)	97.82 (99.30, 93.05)	-0.01 (-0.01, -0.02)
MLR-S	9.27 (9.48, 8.99)	9.45 (9.53, 9.18)	89.32 (90.90, 84.24)	0.01 (0.02, -0.05)
GLM	10.18 (10.18, 10.11)	10.18 (10.21, 10.07)	103.65 (104.34, 101.43)	-0.01 (-0.02, 0.03)
GLM-S	9.25 (9.44, 9.01)	9.39 (9.45, 9.19)	88.13 (89.26, 84.48)	-0.04 (-0.05, -0.00)
SVM	8.13 (8.21, 7.86)	8.51 (8.56, 8.37)	72.45 (73.23, 69.90)	-0.23 (-0.20, -0.32)
RF	7.36 (7.35, 7.37)	7.47 (7.47, 7.48)	55.82 (55.79, 55.82)	-0.03 (0.05, -0.28)

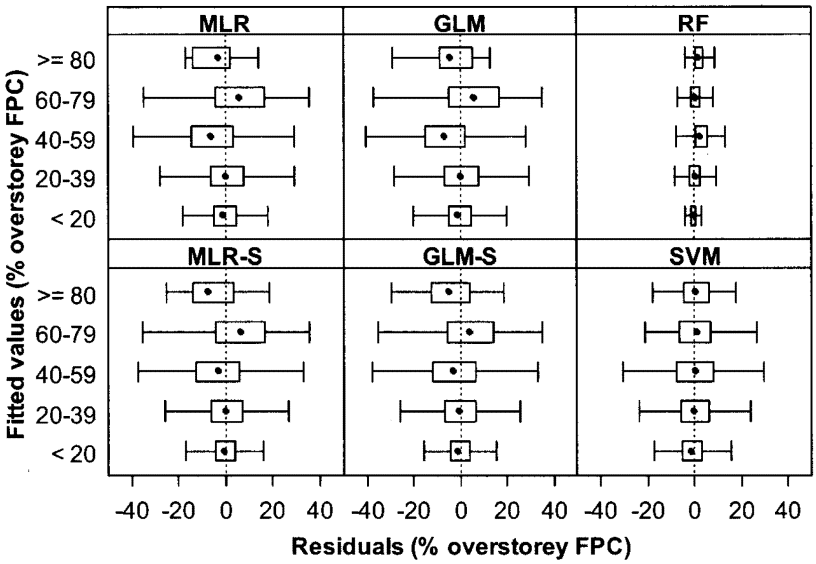

Fig. 6. Box-and-whisker plots of model bootstrap-derived residuals showing the median (dot), first and third quartiles (box), and the most extreme value which is no more than 1.5 times the length of the box away from the box (whiskers) at 20% intervals of overstorey foliage projective cover for each regression model.

These results are consistent with other published findings that have shown machine learning regression techniques provide more accurate predictions than parametric regression techniques [13,17,18]. RF is clearly the best model fit with a narrow distribution of residuals. Breiman [53] demonstrated that RF cannot over-fit the training data due to the "Law of Large Numbers". The bootstrap error results empirically support this however they are still dependent on how representative the training data is of the multidimensional spectral space of the predictors in Queensland, which has not been presented in this work. Therefore the regression models need to be tested against independent estimates of overstorey FPC at locations away from the training data. This is the focus of Sec. 3.3.

3.3 Validation of the regression model predictions

3.3.1 Comparison with field-derived estimates of foliage projective cover

The ETM+ and TM-derived predictions of overstorey FPC are strictly an estimate of FPC_A. Table 5 shows all the regression models are performing well in comparison to independent field estimates of FPC_A. The RMSE statistics are slightly greater than the model and bootstrap RMSE statistics shown in Table 4. Since there was only a relatively small number of independent observations available for comparison it was not possible to evaluate the *bias* and *variance* of the prediction error for different intervals of overstorey FPC and regional areas.

Table 5. A comparison of field-derived overstorey and evergreen foliage projective cover (FPC) and field stand basal area-derived estimates of estimates of overstorey FPC with predictions from near coincident Landsat-5 TM observations for each regression model ($n = 46$). Sec 2.4 defines the error metrics used.

Reference	Model	r^2	RMSE	*variance*	*bias*
Overstorey FPC_T	MLR	0.79	10.26	102.08	-2.31
	MLR-S	0.78	10.66	103.37	-3.54
	GLM	0.79	10.32	104.51	-2.08
	GLM-S	0.78	10.83	106.91	-3.57
	SVM	0.84	9.14	75.91	-3.04
	RF	0.83	9.27	80.72	-2.65
Evergreen FPC_T	MLR	0.89	8.16	68.07	0.08
	MLR-S	0.87	8.57	73.67	-1.15
	GLM	0.89	8.14	67.61	0.31
	GLM-S	0.86	8.60	74.18	-1.17
	SVM	0.85	8.83	79.24	-0.65
	RF	0.82	9.51	92.30	-0.25
Overstorey FPC_A	MLR	0.80	9.45	83.04	-2.83
	MLR-S	0.80	9.83	81.82	-4.06
	GLM	0.79	9.57	86.71	-2.60
	GLM-S	0.79	10.29	91.13	-4.08
	SVM	0.83	9.15	72.58	-3.56
	RF	0.82	9.09	74.22	-3.16

TM-derived predictions of overstorey FPC were also compared to independent field estimates of overstorey and evergreen FPC_T. Evergreen FPC was defined as the horizontally projected percentage cover of photosynthetic foliage from tree and shrub life forms of all heights. Table 4 shows the results are similar for all estimates of FPC however the RMSE and r^2 for evergreen FPC_T are slightly better for the parametric models. The biases for the evergreen FPC predictions are closer to zero for all models compared to the overstorey FPC predictions, however the *variance* of the residuals is worse for the SVM and the RF models. These differences in *bias* are sensible, since evergreen FPC in the understorey from small shrubs and saplings do not contribute to overstorey FPC but still partially determine the observed surface reflectance. It appears that the predictions from the parametric models correspond more closely with evergreen FPC rather than overstorey FPC. This doesn't hold to the same extent for the machine learning models.

The influence of herbaceous FPC on the models predictions is clarified by evaluating the ETM+ and TM-derived predictions in areas of known zero overstorey FPC. Figure 7 shows that over-prediction of overstorey FPC with increasing herbaceous FPC is the general trend for all models. The standard deviations of the residuals for each interval of herbaceous FPC are thought to be partially explained by the difference in date between the image and field data. Even though this is less than sixteen days, changes in green leaf phenology have been observed. The SVM and RF models exhibit less *bias* than the other models. One notable difference is the 85–95% interval for SVM which has a relatively small mean and standard deviation of the residuals, however there are only three observations in this interval so it is difficult to draw any conclusions.

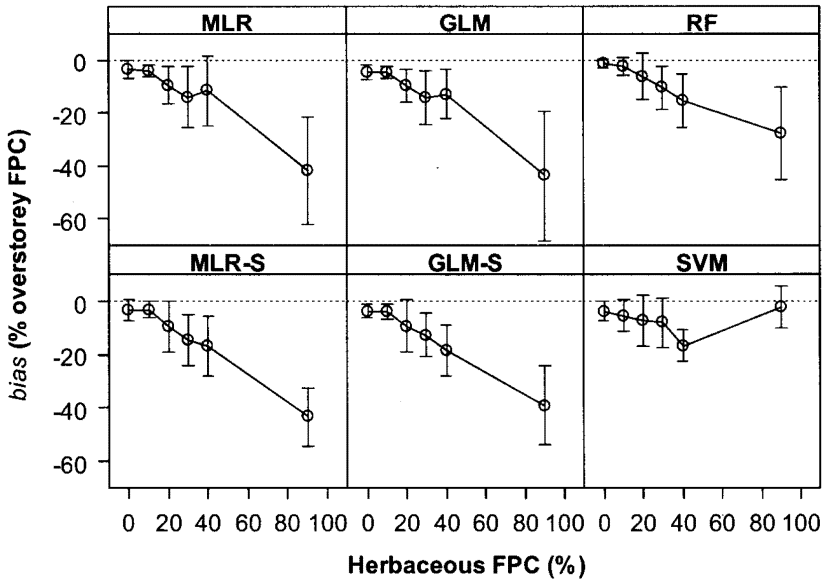

Fig. 7. The *bias* for 10% intervals of field estimated herbaceous FPC. The vertical bars show one standard deviation ($\sqrt{variance}$). Sec 2.4 defines *bias* and *variance*.

These results suggest that the overstorey and understorey components of FPC are not spectrally separable over multiple scenes in Queensland. The *bias* due to herbaceous understorey FPC is a far greater problem for monitoring long-term changes in overstorey FPC than *bias* due to other factors such as soil and vegetation type that are relatively invariant over time apart from perturbations from fire and anthropogenic change. A number of Australian studies using coarse spatial resolution but high temporal resolution sensors (AVHRR, MODIS) have demonstrated seasonal (usually herbaceous) and evergreen FPC can be separated using seasonal trend decomposition [11,55]. However the sparse temporal sampling of Landsat does not permit such analyses. Presentation of a method for separation of the overstorey and understorey components of FPC using simple temporal metrics [1] was beyond the scope of this paper because the objective was to develop regression models that could be applied to individual Landsat acquisitions rather than multi-temporal composites.

3.3.2 Comparison with LiDAR-derived estimates of overstorey foliage projective cover

A comparison of all the LiDAR and TM-derived estimates of overstorey FPC is shown in Fig. 8. The error statistics indicate little difference between the regression models, which is consistent with the bootstrap error (Table 4) and field validation (Table 5) results. The RF model performs best (RMSE = 8.6) and the GLM-S model worst (RMSE = 9.4). All models have an r^2 close to 0.8 and a RMSE less than 10%. The negligible difference in error statistics suggest all models are performing equally well. The clustered distribution of FPC_L values, which can be observed in Fig. 8, means that 0–20% and 40–50% FPC_L values are dominating the results.

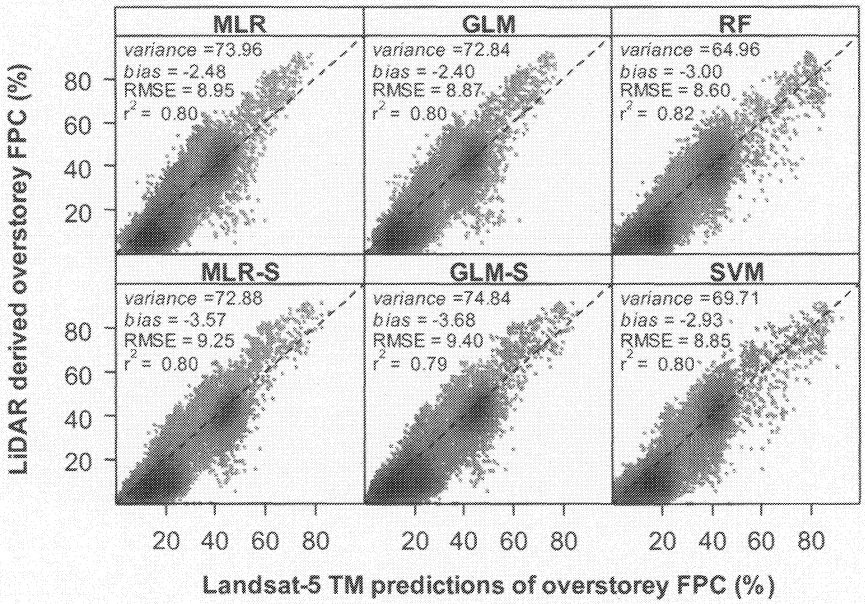

Fig. 8. Comparison of LiDAR-derived overstorey foliage projective cover (FPC) and the Landsat-5 TM-derived predictions for the six regression models. Dark regions of the scatter plots indicate a high density of points and bright regions indicate a relatively low density of points.

The absolute *bias* of prediction error above ~60% FPC_L is greater than 10% for the parametric models but less than 5% for the machine learning models. This is consistent with the pattern of model residuals presented in Fig. 6. In contrast, the *variance* of predictions larger than ~60% FPC_L are greater for the machine learning models than the parametric models. For example, the prediction error *variance* above 60% FPC_L is 65.93 for the RF model and 24.90 for the MLR model. The machine learning models are less stable predictors at high levels of overstorey FPC than the parametric models. It is important to note that the area of closed forest (70–100% overstorey FPC) in Australia is relatively small compared to other overstorey FPC classes [9], therefore the implications of high *bias* for overstorey FPC mapping using the parametric models are minor compared to *bias* in the 0–60% overstorey FPC range, where it is similar for all models.

The direction of the *bias* appears to be dependent on the environments sampled by the LiDAR surveys (Fig. 9). The regression models tend to exhibit negative *bias* for LiDAR survey sites where FPC_L is predominantly less than 20%. In most cases this is probably caused by herbaceous and/or low evergreen understorey FPC. For example, one site with a negative *bias* of approximately 10% for all models is the "gunp01" site (Fig. 9). Much of the region covered by this LiDAR transect has a sparse overstorey however the ground layer is dominated by *Triodia spp.* (spinifex; see "gunp01" in Fig. 3).

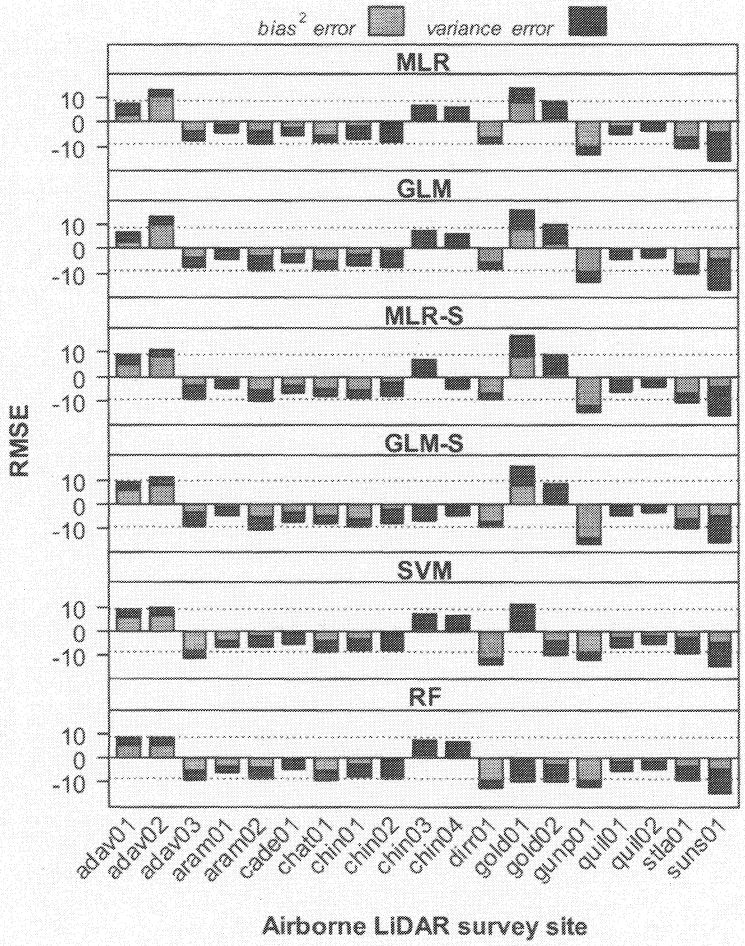

Fig. 9. The RMSE between the LiDAR-derived overstorey foliage projective cover (FPC) and Landsat-5 TM-derived predictions at each LiDAR survey site for the six regression models. The sign of the RMSE corresponds to the sign of the *bias*. The $bias^2$ and *variance* error proportions are relative to the mean squared error (MSE; Sec. 2.4). The dotted grey lines show the total RMSE for each model.

Triodia spp. dominate the vegetation over more than 20% of Australia, therefore introduce a significant limitation to mapping overstorey FPC in inland arid Queensland. *Triodia spp.* grow as a dense hummock with the green leaves on the outer surface. The outer surface of the plants tends to be perennially green unless perturbed by fire, therefore have similar temporal as well as spectral signatures to the photosynthetic foliage of woody species. The surface of the region is covered by scattered iron-stone and this may contribute to the negative *bias* observed, however the result is consistent with the *bias* caused by herbaceous FPC (Fig. 7).

Discrepancies between the LiDAR and TM-derived estimates of overstorey FPC occur due to differing measurement error. The "chin02", "gold02" and "suns01" LiDAR survey sites are unique because the *variance* proportion of the MSE is greater than the $bias^2$ for all the regression models (Fig. 9). These sites are characterized by varying quantities of evergreen

FPC present from 0 to 1 m above the ground (see "suns01" in Fig. 3 for an example). The LiDAR-derived estimates of overstorey FPC included foliage above 0.5 m causing confusion in the separation of overstorey and understorey FPC. Commission errors in the classification of LiDAR ground returns also occur when there is high understorey FPC. The *variance* proportion of the MSE for the "gold01" survey site is also high and visual comparison of the LiDAR and TM-derived images of overstorey FPC confirmed the regression model predictions were sensitive to topographic relief. The preprocessing of the TM and ETM+ imagery (Sec. 2.1.2) does not include a radiance based topographic correction, which is the subject of current research by SLATS.

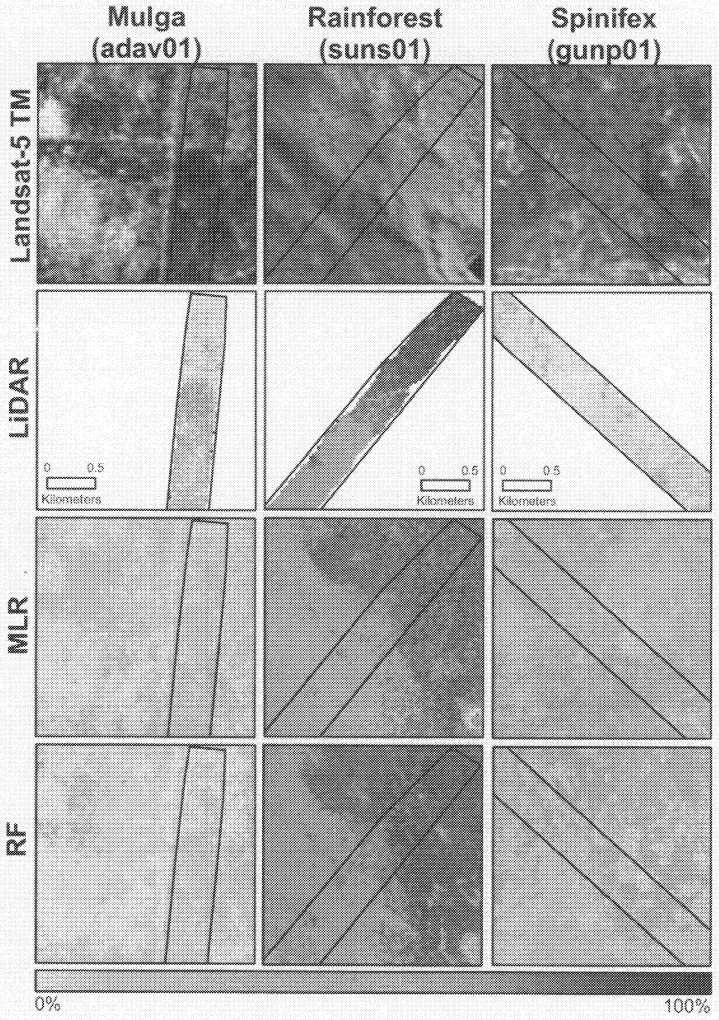

Fig. 10. Maps of LiDAR-derived overstorey foliage projective cover (FPC) and Landsat-5 TM-derived predictions for the multiple linear regression and random forests regression models. Images of areas with mulga ("adav01"), rainforest ("suns01") and spinifex ("gunp01") communities are shown. Note that overstorey FPC tends to be overestimated in spinifex communities.

FPC present from 0 to 1 m above the ground (see "suns01" in Fig. 3 for an example). The LiDAR-derived estimates of overstorey FPC included foliage above 0.5 m causing confusion in the separation of overstorey and understorey FPC. Commission errors in the classification of LiDAR ground returns also occur when there is high understorey FPC. The *variance* proportion of the MSE for the "gold01" survey site is also high and visual comparison of the LiDAR and TM-derived images of overstorey FPC confirmed the regression model predictions were sensitive to topographic relief. The preprocessing of the TM and ETM+ imagery (Sec. 2.1.2) does not include a radiance based topographic correction, which is the subject of current research by SLATS.

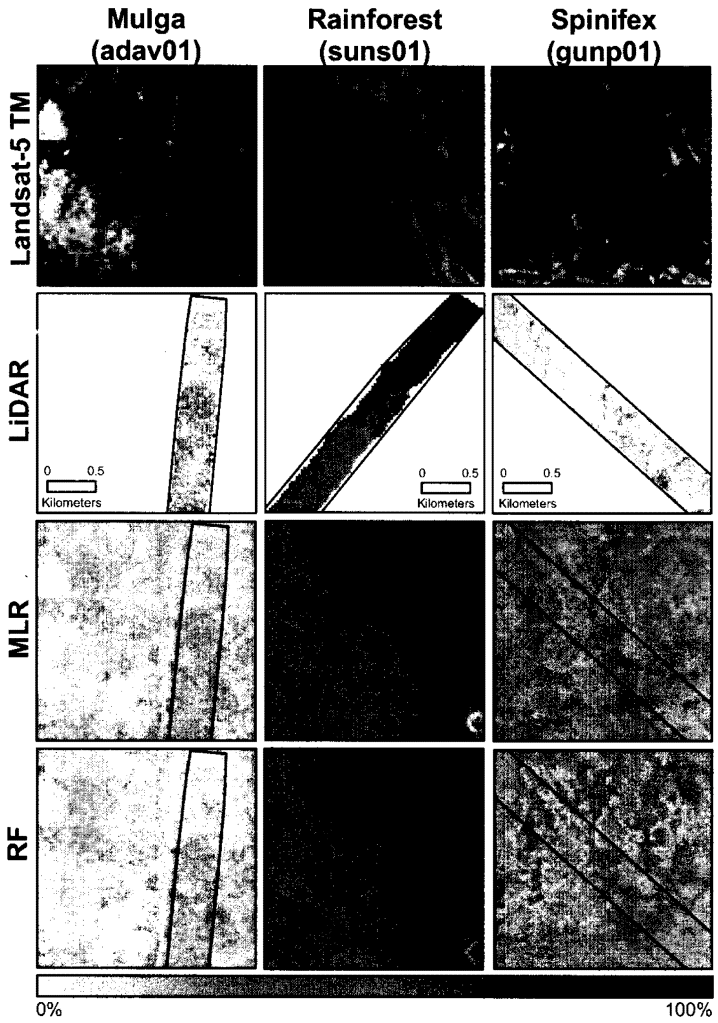

Fig. 10. Maps of LiDAR-derived overstorey foliage projective cover (FPC) and Landsat-5 TM-derived predictions for the multiple linear regression and random forests regression models. Images of areas with mulga ("adav01"), rainforest ("suns01") and spinifex ("gunp01") communities are shown. Note that overstorey FPC tends to be overestimated in spinifex communities.

The two LiDAR survey sites that on average have the greatest positive *bias*, apart from the three coastal sites, are dominated by *Acacia aneura* ("adav01" and "adav02"). The foliage of *A. aneura* is in contrast to other dominate species at other LiDAR sites as they had small sparsely distributed leaves within the crown and field observations indicated some individuals were defoliated due to drought and grazing. It is possible the FPC_L estimates are overestimated at these sites due to: (i) the vertically projected foliage cover within a LiDAR pulse cross-sectional area is less than that for other sites; and (ii) there is greater proportion of non-photosynthetic branches and stems visible to the sensor than at other sites. There may be inaccuracies in FPC_L due to using a single value of a in Eq. (4) for all sites.

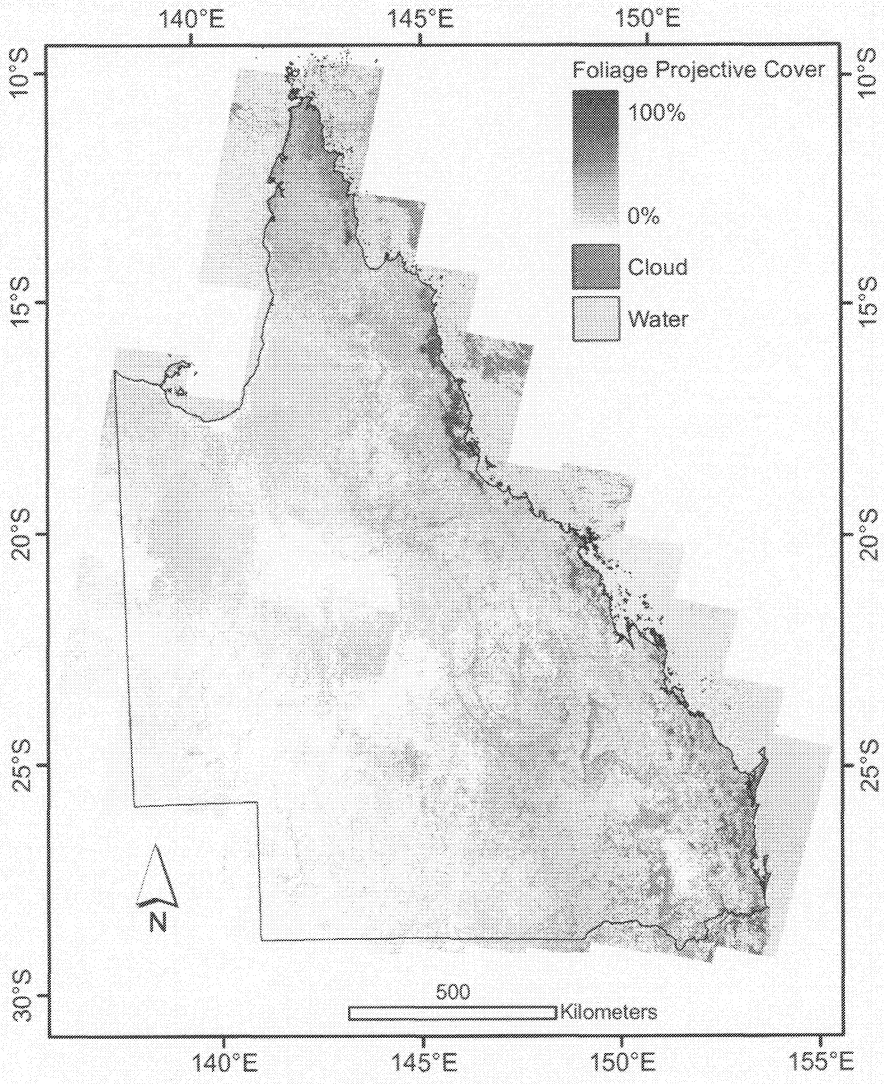

Fig. 11. A mosaic of Landsat-5 TM-derived predictions of overstorey FPC from the random forests regression model for Queensland, Australia. Images acquired during the 2005 dry season were used, but note the scene-to-scene differences due to phenological response to rainfall and other climate drivers.

Figure 10 shows regional examples of the differences between the LiDAR and TM-derived estimates of overstorey FPC. The left column shows the positive *bias* in areas of *A.aneura*. The middle column shows the larger positive *bias* of predictions from the parametric models relative to the machine learning models where FPC_L is greater than ~60%. Finally, the right column shows the negative *bias* in areas where there is high understorey herbaceous FPC (*Triodia spp.*). A 2005 dry season map of overstorey FPC predictions from the RF model for Queensland, Australia, is shown in Fig. 11 demonstrating Landsat imagery can be used for large area mapping of FPC. Some scene-to-scene differences can be observed and are indicative of changes in green leaf phenology in response to rainfall and other climate drivers, which are typical of mixed woody–herbaceous ecosystems. Careful selection of image dates is required to minimize the effects of herbaceous FPC, cloud and smoke cover on overstorey FPC predictions.

4 CONCLUSIONS

The aim of this work was to compare parametric (MLR, GLM) and machine learning algorithms (RF, SVM) for predicting overstorey FPC across multiple Landsat scenes in Queensland, Australia. For the application of estimating overstorey FPC, all of the models provided similar overall accuracy and precision for Queensland conditions. The machine learning models provided an improved fit to the data above ~60% overstorey FPC but otherwise had comparable accuracy and precision to the parametric models. The GLM provides similar quality predictions to MLR but is more appropriate for the error distribution of SBA or overstorey FPC and has the advantage of making no physically unrealistic negative predictions. If the objective is simply to maximize predictive accuracy of a single variable, the use of RF or SVM is recommended as both the model fits and validation showed these provided optimum results overall. The RF model has the advantage of being transparent and easily parallelizable for computational efficiency.

Highly accurate estimates of overstorey FPC (< 5% RMSE) were derived from airborne LiDAR over large range of plant communities in Queensland. These data allowed the accuracy of precision of Landsat-derived overstorey FPC to be assessed within regional areas, demonstrating the utility of airborne LiDAR for cost-effective sampling over large areas. This avoided the need to use stand scale allometric estimates of overstorey FPC derived from field estimates of SBA for calibration and its associated assumptions about canopy structure, which was a limitation of this work. Future development of airborne LiDAR for automated sampling of FPC at different heights in the canopy profile requires separation of photosynthetic and non-photosynthetic cover fractions, radiometric calibration of the backscattered intensity to a physical quantity and knowledge of the relative ground and foliage reflectivity. Future availability of the Carnegie Airborne Observatory [56] or a similar combination of a calibrated full-waveform LiDAR and hyperspectral scanner in Australia has potential to resolve these issues and improve automated sampling of canopy structure.

The validation of predictions from all the candidate models using independent field and LiDAR estimates of overstorey FPC has shown that overstorey FPC was predicted from TM multi-spectral imagery with less than 10% RMSE overall. The major limitation of predictions made from any of the regression models developed using the approach presented in this work was that overstorey and understorey FPC were not decoupled. This has serious implications for trend analysis of woody vegetation cover because photosynthetic herbaceous FPC will increase the variance of the sparse Landsat time-series. Several avenues of research at a range of spatial and temporal scales have since been followed to address this limitation [1,11]. Future research into the detection of long-term trends in woody vegetation cover is being directed at time-series analysis of the entire Landsat archive, which is becoming easier to access since the United States Geological Survey Landsat archive became freely downloadable from December 2008. The use of overstorey FPC products derived from TM or

ETM+ data using any of the regression models developed in this work requires the assumption of a senescent or absent herbaceous understorey at the time of image acquisition. Careful selection of dry season image dates is suggested for best results in Queensland.

Acknowledgments

The authors thank the following Queensland Department of Natural Resources and Water staff. Rob Hassett and Sel Counter for assistance in planning the airborne LiDAR surveys and acquiring the coincident field data, and Michael Schmidt and Dan Tindall for their constructive comments on an earlier version of this manuscript. We also thank the anonymous reviewers for their comments on the manuscript.

References

[1] *Land cover change in Queensland 2005–2006: A Statewide Landcover and Trees Study (SLATS) Report*, Feb. 2008, Dept. Natural Resources and Water, Brisbane, http://www.nrw.qld.gov.au/slats/report.html (2008).

[2] B. K. Henry, T. J. Danaher, G. M. McKeon, and W. H. Burrows, "A review of the potential role of greenhouse gas abatement in native vegetation management in Queensland's rangelands," *Rangeland J.* **24**(1), 112–132 (2002) [doi:10.1071/RJ02006].

[3] W. H. Burrows, B. K. Henry, P. V. Back, M. B. Hoffmann, L. J. Tait, E. R. Anderson, N. Menke, T. J. Danaher, J. O. Carter, and G. M. McKeon, "Growth and carbon stock change in eucalypt woodlands in northeast Australia: ecological and greenhouse sink implications," *Glob. Change Biol.* **8**, 769–784 (2002) [doi:10.1046/j.1365-2486.2002.00515.x].

[4] R. M. Lucas, N. Cronin, M. Moghaddam, A. Lee, J. Armston, P. Bunting, and C. Witte, "Integration of radar and Landsat-derived foliage projected cover for woody regrowth mapping, Queensland, Australia," *Rem. Sens. Environ.* **100**, 388–406 (2006) [doi:10.1016/j.rse.2005.09.020].

[5] D. Sun, R. J. Hantiuk, and V. J. Neldner, "Review of vegetation classification and mapping systems undertaken by major forested land management agencies in Australia," *Aust. J. Bot.* **45**, 929–948 (1997) [doi:10.1071/BT96121].

[6] R. L. Specht, "Foliage projective covers of overstorey and understorey strata of mature vegetation in Australia," *Austral Ecol.* **8**, 433–439 (1983) [doi:10.1111/j.1442-9993.1983.tb01340.x].

[7] C. A. Kuhnell, B. M. Goulevitch, T. J. Danaher, and D. P. Harris, "Mapping woody vegetation cover over the state of Queensland using Landsat TM imagery," presented at *9th Aust. Rem. Sens. Photogramm. Conf.*, Sydney (1998).

[8] J. M. Chen and J. Cihlar, "Plant canopy gap size analysis theory for improving optical measurements of leaf area index," *Appl. Opt.* **34**, 6211–6222 (1995) [doi:10.1364/AO.34.006211].

[9] R. L. Specht and A. Specht, *Australian Plant Communities: Dynamics of Structure, Growth and Biodiversity*, Oxford University Press (1999).

[10] M. C. Hansen, R. S. DeFries, J. R. G. Townshend, R. Sohlberg, C. Dimiceli, and M. Carroll, "Towards an operational MODIS continuous field of percent tree cover algorithm: examples using AVHRR and MODIS data," *Rem. Sens. Environ.* **83**, 303–319 (2002) [doi:10.1016/S0034-4257(02)00079-2].

[11] T. K. Gill, S. R. Phinn, J. D. Armston, and B. A. Pailthorpe, "Estimating tree foliage-cover change in Australia: challenges of using the MODIS 250m vegetation index product," *Int. J. Rem. Sens.* **30**(6), 1547–1565 (2009) [doi:10.1080/01431160802509066].

[12] R. S. DeFries, M. Hansen, M. Steininger., R. Dubayah, R. Sohlberg, and J. Townshend, "Subpixel forest cover in Central Africa from multisensor, multitemporal data," *Rem. Sens. Environ.* **60**, 228–246 (1997) [doi:10.1016/S0034-4257(96)00119-8].

[13] J. Chia, M. Zhu, P. Caccetta, and J. Wallace. "Derivation of a perennial vegetation map for the Australian continent," presented at *13th Aust. Rem. Sens. Photogramm. Conf.*, Canberra (2006).

[14] C. Song and C. E. Woodcock, "Monitoring forest succession with multi-temporal Landsat images: factors of uncertainty," *IEEE Trans. Geosci. Rem. Sens.* **41**(11), 2557–2567 (2003) [doi:10.1109/TGRS.2003.818367].

[15] P. Scarth, M. Byrne, T. Danaher, B. Henry, R. Hassett, J. Carter, and P. Timmers, "State of the paddock: monitoring condition and trend in groundcover across Queensland," presented at *13th Aust. Rem. Sens. Photogramm. Conf.*, Canberra (2006).

[16] J. Settle and N. Campbell, "On the errors of two estimators of sub-pixel fractional cover when mixing is linear," *IEEE Trans. Geosci. Rem. Sens* **36**(1), 163–170 (1998) [doi:10.1109/36.655326].

[17] I. Olthof and R. Fraser, "Mapping northern land cover fractions using Landsat ETM+," *Rem. Sens. Environ.* **107**(3), 496–509 (2007) [doi:10.1016/j.rse.2006.10.009].

[18] R. Fernandes, R. Fraser, R. Latifovic, J. Cihlar, J. Beaubien, and Y. Du, "Approaches to fractional land cover and continuous field mapping: A comparative assessment over the BOREAS study region," *Rem. Sens. Environ.* **89**, 234–251 (2004) [doi:10.1016/j.rse.2002.06.006].

[19] R. L. Iverson, E. A. Cook, and R. L. Graham, "Regional forest cover estimation via remote sensing: the calibration center concept," *Landscape Ecol.* **9**, 159–174 (1994) [doi:10.1007/BF00134745].

[20] M. Schwarz and N. E. Zimmermann, "A new GLM-based method for mapping tree cover continuous fields using regional MODIS reflectance data," *Rem. Sens. Environ.* **95**, 428–443 (2005) [doi:10.1016/j.rse.2004.12.010].

[21] J. Heiskanen and S. Kivinen, "Assessment of multispectral, -temporal and -angular MODIS data for tree cover mapping in the tundra–taiga transition zone," *Rem. Sens. Environ.* **112**(5), 2367–2380 (2008) [doi:10.1016/j.rse.2007.11.002].

[22] J. Rogan, J. Franklin, D. Stow, J. Miller, C. Woodcock and D. Roberts, "Mapping land-cover modifications over large areas: a comparison of machine learning algorithms," *Rem. Sens. Environ.* **112**(5), 2272–2283 (2008) [doi:10.1016/j.rse.2007.10.004].

[23] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Rem. Sens* **42**(8), 1778–1790 (2004) [doi:10.1109/TGRS.2004.831865].

[24] R. L. Lawrence, S. D. Wood, and R. L. Sheley, "Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (RandomForest)," *Rem. Sens. Environ.* **100**(3), 356–362 (2006) [doi:10.1016/j.rse.2005.10.014].

[25] R. J. Fensham and J. E. Holman, "Temporal and spatial patterns in drought-related tree dieback in Australian savanna," *J. Appl. Ecol.* **36**, 1035–1050 (1999) [doi:10.1046/j.1365-2664.1999.00460.x].

[26] J. T. Morisette, J. E. Nickeson, P. Davis, Y. J. Wang, Y. H. Tian, C. E. Woodcock, N. Shabanov, M. Hansen, W.B. Cohen, D. R. Oetter, and R. E. Kennedy, "High spatial resolution satellite observations for validation of MODIS land products: IKONOS observations acquired under the NASA Scientific Data purchase," *Rem. Sens. Environ.* **88**(1–2), 100–110 (2003) [doi:10.1016/j.rse.2003.04.003].

[27] D. Weller, R. Denham, C. Witte, C. Mackie, and D. Smith, "Assessment and monitoring of foliage projected cover and canopy height across native vegetation in Queensland, Australia, using laser profiler data," *Can. J. Rem. Sens.* **29**(5), 578–591 (2003).

[28] P. K. Tickle, K., A. Lee, R. M. Lucas, J. Austin, and C. Witte, "Quantifying Australian forest floristics and structure using small footprint LiDAR and large scale aerial photography," *Forest Ecol. Manag.* **223**(1–3), 379–394 (2006) [doi:10.1016/j.foreco.2005.11.021].

[29] J. L. Lovell, D. L. B. Jupp, D. S. Culvenor, and N. C. Coops, "Using airborne and ground-based ranging LiDAR to measure canopy structure in Australian forests," *Can. J. Rem. Sens.* **29**, 607–622 (2003).

[30] R. S. DeFries and J. C. Chan, "Multiple criteria for evaluating machine learning algorithms for land cover classification from satellite data," *Rem. Sens. Environ.* **74**, 503–515 (2000) [doi:10.1016/S0034-4257(00)00142-5].

[31] C. de Vries, T. Danaher, R. Denham, P. Scarth, and S. Phinn., "An operational calibration procedure for the Landsat sensors based on pseudo-invariant target sites," *Rem. Sens. Environ.* **107**, 414–429 (2007) [doi:10.1016/j.rse.2006.09.019].

[32] R. C. Hassett, H. L. Wood, J. O. Carter, and T. J. Danaher, "A field method for statewide ground-truthing of a spatial pasture growth model," *Aust. J. Exp. Agr.* **40**, 1069–1079 (2000) [doi:10.1071/EA00010].

[33] J. R. Dilworth and J. F. Bell, *Variable Probability Sampling — Variable Plot and Three–P*, OSU Books, Corvallis, OR (1971).

[34] T. Johansson, "Estimating canopy density by the vertical tube method," *Forest Ecol. Manag.* **11**, 139–144 (1985) [doi:10.1016/0378-1127(85)90063-5].

[35] J. Armston, T. Danaher, B. Goulevitch, and M. Byrne, "Geometric correction of Landsat TM and ETM+ imagery for mapping woody vegetation cover and change detection over Queensland," presented at *11th Aust. Rem. Sens. Photogramm. Conf.*, Brisbane (2002).

[36] T. Danaher, "An empirical BRDF correction for Landsat TM and ETM+ imagery," presented at *11th Aust. Rem. Sens. Photogramm. Conf.*, Brisbane (2002).

[37] J. K. Shaw and S. S. Gillingham, "Effects of cloud and smoke contamination on woody vegetation time-series trends," presented at *13th Aust. Rem. Sens. Photogramm. Conf.*, Canberra (2006).

[38] G. A. Duff, B. A. Myers, R. J. Williams, D. Eamus, A. O'Grady, and I. R. Fordyce, "Seasonal patterns in soil moisture, vapour pressure deficit, tree canopy cover and pre-dawn water potential in a northern Australian savanna," *Aust. J. Bot.* 45, 211–224 (1997) [doi:10.1071/BT96018].

[39] S. J. Jeffrey, J. O. Carter, K. B. Moodie, and A. R. Beswick, "Using spatial interpolation to construct a comprehensive archive of Australian climate data," *Environ. Modell. Softw.* **16**, 309–330 (2001) [doi:10.1016/S1364-8152(01)00008-1].

[40] A. Accad, V. J. Neldner, B. A. Wilson, and R. E. Niehus, *Remnant Vegetation in Queensland: Analysis of Remnant Vegetation 1997–1999–2000–2001–2003 including Regional Ecosystem Information*, Queensland Herbarium, Queensland Environmental Protection Agency, Brisbane (2006).

[41] ITTVIS, *IDL —Interactive Data Language*, ITT Visual Information Solutions Inc., Boulder, CO (2005).

[42] N. R. Goodwin, N. C. Coops, and D. S. Culvenor, "Assessment of forest structure with airborne LiDAR and the effects of platform altitude," *Rem. Sens. Environ.* **103**, 140–152 (2006) [doi:10.1016/j.rse.2006.03.003].

[43] K. Zhang, S. Chen, D. Whitman, M. Shyu, J. Yan, and C. Zhang, "A progressive morphological filter for removing nonground measurements from airborne LIDAR

data," *IEEE Trans. Geosci. Rem. Sens.* **41**, 872–882 (2003) [doi:10.1109/TGRS.2003.810682].

[44] T. Moffiet, "Bivariate relationship modelling on bounded spaces with application to the estimation of forest foliage cover by Landsat satellite," PhD Thesis, Univ. of Newcastle, Callaghan, NSW, Australia (2007).

[45] R. J. Williams, B. A. Myers, W. J. Muller, G. A. Duff, and D. Eamus, "Leaf phenology of woody species in a North Australian tropical savanna," *Ecology* **78**(8), 2542–2558 (1997) [doi:10.2307/2265913].

[46] R. J. Fensham, S. J. Low Choy, R. J. Fairfax, and P. C. Cavallaro, "Modelling trends in woody vegetation structure in semi-arid Australia as determined from aerial photography," *J. Environ. Manage.* **68**, 421–436 (2003) [doi:10.1016/S0301-4797(03)00111-7].

[47] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, http://www.R-project.org, Vienna (2007).

[48] F. E. Harrell, *Regression Modeling Strategies with Applications to Linear Models, Logistic Regression and Survival Analysis*, Springer Series in Statistics, Springer, New York (2001).

[49] P. McCullagh and J. A. Nelder, *Generalized Linear Models 2nd ed.*, Chapman and Hall, London (1989).

[50] A. J. Smola and B. Scholkopf, "A tutorial on support vector regression," *Stat. Comput.* **14**(3), 199–222 (2004) [doi: 10.1023/B:STCO.0000035301.49549.88].

[51] V. Cherkassky and Y. Ma, "Practical selection of SVM parameters and noise estimation for SVM regression," *Neural Networks* **17**(2), 113–126 (2004) [doi:10.1016/S0893-6080(03)00169-2].

[52] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, Wadsworth International Group, Belmont, CA (1984).

[53] L. Breiman, "Random forests," *Mach. Learn.* **45**, 5–32 (2001) [doi:10.1023/A:1010933404324].

[54] D. S. Falster and M. Westoby, "Leaf size and angle vary widely across species: what consequences for light interception?" *New Phytol.* **158**(3), 509–525 (2003) [doi:10.1046/j.1469-8137.2003.00765.x].

[55] H. Lu, M. R. Raupach, T. R. McVicar, and D. J. Barrett, "Decomposition of vegetation cover into woody and herbaceous components using AVHRR NDVI time series," *Rem. Sens. Environ.* **86**, 1–18 (2003) [doi:10.1016/S0034-4257(03)00054-3].

[56] G. P. Asner, D. E. Knapp, T. Kennedy-Bowdoin, M. O. Jones, R. E. Martin, J. Boardman, and C. B. Field, "Carnegie Airborne Observatory: in-flight fusion of hyperspectral imaging and waveform light detection and ranging (wLiDAR) for three-dimensional studies of ecosystems," *J. Appl. Rem. Sens.* **1**, 013536 (2007) [doi:10.1117/1.2794018].