

**Quantifying the structure of the woody element in Savannas using  
integrated optical and Synthetic Aperture Radar (SAR) approach: a stepping  
stone towards country wide monitoring in South Africa**

by

**Laven Naidoo**

Submitted in partial fulfilment of the requirements for the degree

**DOCTOR OF PHILOSOPHY (Geoinformatics)**

**in the Faculty of Natural and Agricultural Sciences, University of Pretoria**

**Supervisors:**

**Dr Renaud Mathieu (CSIR & University of Pretoria)**

**Dr Konrad Wessels (CSIR & University of Pretoria)**



**Date: 30 September 2017**

# Table of Contents

Table of Contents.....	2
List of Figures.....	6
List of Tables.....	9
Dedications.....	11
Acknowledgements.....	12
Plagiarism Declaration.....	13
List of Publications.....	14
List of Acronyms.....	15
<b>Thesis Abstract.....</b>	<b>17</b>
<b>Chapter 1: Introduction and literature review.....</b>	<b>20</b>
1.1 Savannahs and the importance of its woody component in ecosystem and monitoring processes.....	20
1.2 The current status of monitoring the woody component in South Africa.....	23
1.3 Woody structural metrics.....	24
1.3.1 Woody biomass.....	24
1.3.2 Woody canopy volume.....	25
1.3.3 Woody canopy cover.....	26
1.4 Remote sensing of woody structure.....	26
1.4.1 Passive remote sensing of woody structure.....	27
1.4.2 Active remote sensing of woody structure.....	28
1.4.3 Global forest remote sensing products.....	31
1.5 Multi-sensor and multi-temporal remote sensing data integration.....	31
1.6 Study aim.....	33
1.7 Study objectives and chapter breakdown.....	34
1.8 Primary and secondary research questions.....	35
Chapter 2.....	35
Chapter 3.....	35
Chapter 4.....	36
Chapter 5.....	36

<b>Chapter 2: Assessment of the Performance of Global Forest Cover Products in South Africa – Establishing the benchmark</b> .....	38
2.1 Abstract.....	38
2.2 Introduction .....	39
2.3 Study Area.....	42
2.4 Materials and Methodology .....	44
2.4.1 Global forest products .....	45
2.4.2 LiDAR validation datasets.....	45
2.4.3 Global Forest Product Pre-processing .....	46
2.4.4 Airborne LiDAR Data Pre-processing .....	47
2.4.5 Data extraction process .....	50
2.4.6 Global Product Accuracy Assessment.....	50
2.5 Results.....	53
2.5.1 ALOS PALSAR FNF validation results .....	53
2.5.2 Landsat VCF validation results .....	57
2.6 Discussion.....	62
2.7 Conclusions .....	66
<b>Chapter 3: Savannah woody structure modelling and mapping using multi-frequency (X-, C- and L-band) Synthetic Aperture Radar (SAR) data</b> .....	67
3.1 Abstract.....	67
3.2 Introduction - Background, Aims and Objectives .....	68
3.3 Study Area.....	72
3.4 Materials and Methodology .....	74
3.4.1 Remote sensing data.....	74
3.4.2 Field data.....	75
3.4.3 LiDAR data processing, woody structural metrics and validation .....	78
3.4.4 SAR data and processing.....	80
3.4.5 Data integration, modelling protocols and mapping.....	82
3.4.6 Error assessment.....	84
3.5 Results.....	86
3.5.1 Modelling Accuracy Assessment.....	86
3.5.2 Tree Structure Metric and Error Maps.....	88
3.6 Discussion.....	96

3.7 Concluding Remarks.....	100
<b>Chapter 4: Integration of Optical and L-band Synthetic Aperture Radar (SAR) datasets for the assessment of woody fractional cover in the Greater Kruger National Park region .....</b>	<b>102</b>
4.1 Abstract.....	102
4.2 Introduction .....	103
4.3 Study Area.....	107
4.4 Materials and Methodology .....	108
4.4.1 Remote Sensing Data .....	108
4.4.2 LiDAR Data Processing .....	109
4.4.3 SAR Data Processing.....	110
4.4.4 Landsat-5 Optical Data Processing and Derived Products.....	110
4.4.5 Data Analysis Grid .....	112
4.4.6 Modelling Algorithms, Modelling Scenarios, Model Validation and CC Mapping.....	112
4.5 Results.....	114
4.5.1 Individual and multi-seasonal Landsat-5 reflectance compared to SAR .....	114
4.5.2 Optical reflectance, textures and indices compared and integrated with SAR data results .....	115
4.6 Discussion.....	120
4.7 Conclusions .....	125
<b>Chapter 5: Scaling-up methods for national woody fractional cover mapping: Experiments and guideline on the amount of field plots and airborne LiDAR data required for training and validation .....</b>	<b>127</b>
5.1 Abstract.....	127
5.2 Introduction .....	128
5.3 Study Area.....	131
5.4 Materials and methodology.....	133
5.4.1 Airborne LiDAR datasets and processing.....	134
5.4.2 ALOS PALSAR FBD global image mosaic.....	134
5.4.3 Ancillary environmental parameters .....	135
5.4.4 Dataset integration and extraction process.....	135
5.4.5 Modelling process – algorithm, modelling scenarios and validation .....	139
5.4.6 Optimal amount of LiDAR simulated 1 hectare field plots .....	140
5.4.7 Optimal LiDAR training amount in terms of hectare coverage.....	141

5.5 Results .....	143
5.5.1 RF Modelling results (validation) according to modelling scenarios .....	143
5.5.2 Optimal field plot amount.....	145
5.5.3 Optimal amount of LiDAR data required .....	147
5.6 Discussion.....	152
5.7 Conclusions .....	158
<b>Chapter 6: Study conclusions, recommendations and ways forward .....</b>	<b>159</b>
<b>References .....</b>	<b>166</b>
<b>Appendices .....</b>	<b>188</b>
Appendix 2A: Example of poor VCF accuracy statistics according to stratified CC ranges and vegetation structural classes .....	189
Appendix 2B: FNF confusion matrix results based on new HV Forest class threshold of -19dB ....	189
Appendix 3A: Colgan et al. (2013) biomass allometric equation.....	189
Appendix 3B: Plot level above ground biomass up-scaling factors .....	189
Appendix 3C: The assessment of various data mining algorithms for modelling savannah woody cover .....	189
Appendices 4A and 4B: Random Forest optimisation attempts.....	199
Methodology for the creation of the EVI phenology graphs (Figures 4.4 and 4.5) .....	200

## List of Figures

### Chapter 2:

Figure 2.1: Study area with focus on the LiDAR dataset coverages (see table 1 for LiDAR specifications) .....	44
Figure 2.2: 25m ALOS PALSAR Forest/Non-Forest (left) and 30m Landsat Variable Continuous Field cover (right) products .....	47
Figure 2.3: Classification of vegetation types according to structure (canopy cover and height) (Willis, 2002). The LiDAR canopy cover and CHM products were used to reproduce this classification scheme for the extracted data.....	52
Figure 2.4: Summarised FNF validation results across various LiDAR-derived vegetation structural classes as outlined by (Willis, 2002) (% values refer to the FNF detection accuracy of that vegetation structural class where red cells = accuracies >90%, orange cells = accuracies between 15-30%, yellow cells = accuracies ≤5%).....	54
Figure 2.5: i) ALOS PALSAR FNF (left) versus LiDAR derived FNF (right) across the CAO LiDAR dataset; ii) ALOS PALSAR FNF (left) and L-band ALOS PALSAR FBD derived FNF (right), using LiDAR training, ( $R^2=0.81$ ; $RMSE=9.89\%$ - this product will be detailed in chapters 3 and 4) across the entire Kruger National Park extent [the red and blue encircled areas indicates areas of interest for discussion] ....	56
Figure 2.6: (i) Density scatterplot of LiDAR derived CC versus Landsat VCF CC across the complete extracted dataset [the dotted red line represents the 1:1 line while the solid black line represents the data trend line]. (ii) Landsat VCF product error (i.e. LiDAR CC – VCF CC) over a range of CC intervals [negative values indicate CC overestimation while positive values indicated CC underestimation by the MODIS VCF product; centre cross = mean value; box = standard error and whiskers = standard deviation].....	57
Figure 2.7: Summarised VCF validation results across various LiDAR-derived vegetation structural classes as outlined by (Willis, 2002) (% values refer to the VCF detection accuracy of that vegetation structural class where red cells = accuracies >20%, orange cells = accuracies between 10-20%, yellow cells = accuracies <10%). The red 5m height line indicates the limit of VCF product in which all classes coloured grey (below 5m height) was excluded.....	60
Figure 2.8: i) Landsat VCF CC (left) versus LiDAR derived CC (right) across the CAO LiDAR dataset; ii) Landsat VCF CC (left) and L-band ALOS PALSAR FBD derived CC (right), using LiDAR training, ( $R^2=0.81$ ; $RMSE=9.89\%$ ) across the entire Kruger National Park extent [the red and blue encircled areas indicates areas of interest for discussion] .....	61
Figure 2.9: LiDAR derived woody canopy cover versus ALOS PALSAR HV backscatter (dB) extracted over the complete LiDAR dataset coverage [the red line indicates the -15.6dB threshold value (Shimada et al., 2014) used to create the FNF over the continent of Africa while the orange box indicates the bulk of the LiDAR CC values captured by the FNF according to the CC values greater than and equal to the HV dB threshold] .....	64

### Chapter 3:

Figure 3.1: The Southern Kruger National Park region and the spatial coverage of all implemented remote sensing datasets. The solid red line indicates the coverage of the 2009 RADARSAT-2 scenes while the solid gold line indicates the two scenes of the 2010 ALOS dual-pol PALSAR imagery. The

dashed grey line indicates the five scenes of the 2012 TerraSAR-X StripMap imagery. The shaded black areas represent the coverage of the 2012 CAO LiDAR sensor tree cover product. The red squares indicate the 38 sample sites where field data collections took place..... 73

Figure 3.2: Ground sampling design including ground tree biomass and tree cover collection protocols (50m spacing between sample plots coincide with the auto-correlation distance – refer to data integration section)..... 77

Figure 3.3: Validation results of field-measured woody Canopy Cover (CC) versus LiDAR derived CC (above 0.5m height, Number of observations =37)..... 79

Figure 3.4: Validation results of field-measured Above Ground Biomass (AGB) versus LiDAR derived AGB (above 0.5m height, Number of observations =53)..... 80

Figure 3.5: Methodology schema describing the data integration and modelling process ..... 85

Figure 3.6: Observed versus Predicted Total woody Canopy Volume (TCV) scatter density plots (A-G) (dotted line is 1:1)..... 87

Figure 3.7: X+C+L SAR derived tree structural metric maps, for i) Above Ground Biomass (AGB), ii) Total woody Canopy Volume (TCV) and iii) woody Canopy Cover (CC), using random forest. Letters A-F represents key areas of interest for discussion (for all three metrics). The black boxes represent the rough extents of the LiDAR-SAR CC scenario difference maps for Area of Interests ‘A’ and ‘C’ ..... 89

Figure 3.8: Scatterplot of Above Ground Biomass (AGB), y-axis, versus woody Canopy Cover (CC), x-axis, under dense cover conditions (plotted from pixels extracted from the Area of Interest ‘A’)..... 90

Figure 3.9: LiDAR - SAR scenario difference (error) maps (i-iv) of Total woody Canopy Volume (TCV) for the Xanthia Forest Area of Interest (close to ‘A’); v) 25m LiDAR-derived TCV map ..... 92

Figure 3.10: LiDAR - SAR scenario difference (error) maps (i-iv) of Total woody Canopy Volume (TCV) for the Gabbro Intrusions Area of Interest ‘C’; v) 25m LiDAR-derived TCV map..... 93

Figure 3.11: Woody Canopy Cover (CC) Error box plots of: i) low LiDAR CC (<40%) and variable LiDAR vegetation height and ii) dense LiDAR CC (>70%) and variable LiDAR vegetation height (+’ve values = CC underestimation; -’ve values = CC overestimation; dashed line partitions the four main SAR scenarios across the x-axis classes, centre point = mean value, box = standard error and whiskers = standard deviation) (Number of pixels = 17559) ..... 95

**Chapter 4:**

Figure 4.1: The Southern Kruger National Park study area and coverage of remote sensing modelling datasets..... 107

Figure 4.2: Regional scale CC map of the study area using the best performing RF integrated L-band and single date Landsat-5 band reflectance model (2010 L-band & 2010 Autumn LT5 image; coverage excludes extensive cloud cover to the east)..... 118

Figure 4.3: Predicted CC versus Observed CC scatterplots for: i) 2008 Multi-seasonal Landsat-5 Reflectance-only, ii) 2008 SAR-only and iii) integrated 2008 Multi-seasonal Landsat-5 Reflectance and SAR modelled validation results ..... 119

Figure 4.4: Temporal fluctuations of mean EVI values (extracted from MODIS data) over a predominant grassland site (L1) and a predominant woodland site (L8) between the beginning of 2005 and end of 2012. Rainfall measurements between beginning of 2007 and end of 2011 have also been included..... 122

Figure 4.5: Temporal differences of mean grass and tree EVI values (extracted from MODIS data) over a predominant grassland site and a predominant woodland site between the beginning of 2005

and end of 2012. Red lines with numbers 1-12 indicate the multi-seasonal Landsat-5 image acquisition dates ..... 123

**Chapter 5:**

Figure 5.1: Study area of South Africa, including biome (Mucina and Rutherford, 2006) and airborne LiDAR acquisition coverage..... 133

Figure 5.2: Approximate hectare and percentage coverage of samples in all LiDAR datasets according to biome ..... 137

Figure 5.3: i) Samples according to vegetation structural classes (including hectare and percentage coverage statistics per class); ii) Vegetation structural classes (Willis, 2002) according to average height and projected woody plant canopy cover which was used to classify i)..... 138

Figure 5.4: Scatterplots of the mean predicted CC versus mean observed LiDAR derived CC resulting from models validated over five individual biomes (i-v) and the complete fixed validation dataset (vi) while using Savannah-only data for training. Error bars indicate confidence intervals of each point at the biome level. Black dotted line indicates the 1:1 trend line. .... 144

Figure 5.5: RF validation accuracies, including % change, across the different training sampling sizes obtained from the all-biome dataset. In the % change table, +’ve values indicate a percentage increase while –’ve values indicate a percentage decrease (No. = number; ha = hectares)..... 146

Figure 5.6: RF validation accuracies, including % change, across the different training sampling sizes obtained from the Savannah-only dataset. In the % change table, +’ve values indicate a percentage increase while –’ve values indicate a percentage decrease (No. = number; ha = hectares)..... 147

Figure 5.7: CC R<sup>2</sup> validation accuracies according to the number of simulated LiDAR acquisitions of different sizes used for RF modelling from the all-biome dataset (scenario (iv)). The **red solid line** indicates the mean R<sup>2</sup> value obtained from the 500 1ha field plot result while the **red dotted line** indicates the corresponding upper and lower Confidence Interval limits ..... 149

Figure 5.8: CC RMSE validation accuracies according to the number of simulated LiDAR acquisitions of different sizes used for RF modelling from the all-biome dataset (scenario iv). The **red solid line** indicates the mean RMSE value obtained from the 500 1ha field plot result while the **red dotted lines** indicate the corresponding upper and lower Confidence Interval limits. .... 150

Figure 5.9: Variability of woody fractional cover (i) and vegetation structure (ii) across Applebosch Ndwedwe LiDAR..... 156

**Chapter 6:**

Figure 6.1: Woody fractional canopy cover (CC) map of the South African Region (including parts of neighbouring countries)..... 165

**Appendices:**

Figure 2A: XY density scatterplot of Landsat (Hansen) VCF CC values versus LiDAR CC for: 1) the 40 to 80 CC range and 2) the Woodland structural class..... 194

Figure 1: Mean RF predicted CC versus mean observed CC for each multi-frequency scenario (The dotted line refers to the 1:1 line)..... 1944



Figure 4A: Root Mean Square Error (RMSE) variability box plot, derived from modelled 2008 L-band SAR and LT5 summer reflectance data results, for analysing the different RF tuning parameters (Unpruned = unpruned; nd = nodesize; mx = maxnodes; solid bar = mean RMSE; whiskers = max/min range; dots = outliers) ..... 1999

Figure 4B: Root Mean Square Error (RMSE) line graph, derived from modelled 2008 L-band SAR and LT5 summer reflectance data results, for analysing the different RF tuning parameters across varying number of trees in the forest (NTrees = number of trees; nd = nodesize; mx = maxnodes) ..... 200

## List of Tables

### Chapter 2:

Table 2.1: Summary of LiDAR datasets used, the year acquired, sensor specifications, coverage, environmental description and provider information ..... 49

Table 2.2: Summarised FNF validation results across stratified LiDAR-derived CC ranges ..... 53

Table 2.3: Summarised FNF validation results across various LiDAR-derived vegetation structural classes ..... 54

Table 2.4: Summarised VCF validation results across stratified LiDAR-derived CC ranges ..... 58

Table 2.5: Complete VCF CC versus LiDAR CC confusion matrix across fixed CC ranges ..... 58

Table 2.6: Summarised VCF validation results across various LiDAR-derived vegetation structural classes ..... 59

### Chapter 3:

Table 3.1: SAR and LiDAR datasets acquired and utilised for the modelling of woody structural metrics ..... 81

Table 3.2: Original, modified and final SAR pixel size changes during multi-looking and pre-processing steps ..... 82

Table 3.3: Woody Canopy Cover (CC), Total Canopy Volume (TCV) and Above Ground Biomass (AGB) parameter modelling accuracy assessment (validation) results obtained from the Random Forest algorithm according to seven SAR frequency scenarios ..... 86

Table 3.4: Total woody Canopy Cover (CC), Total Canopy Volume (TCV) and Above Ground Biomass (AGB) % error across the entire LiDAR-SAR coverage for the four main SAR frequency scenarios (Number of observations = 17559) ..... 91

### Chapter 4:

Table 4.1: Landsat-5, ALOS PALSAR and LiDAR data inventory ..... 109

Table 4.2: Reflectance, indices and textural optical products derived from Landsat-5 data ..... 112

Table 4.3: Individual seasonal Landsat-5, multi-seasonal Landsat-5 and individual SAR RF modelled CC validation results ..... 114

Table 4.4: Reflectance, indices and textural Landsat-5 (autumn 2010 image) product RF modelled CC validation results ..... 115

Table 4.5: Integrated SAR and best performing/multi-seasonal Landsat-5 reflectance RF modelled CC validation results (per year) .....	115
---	-----

**Chapter 5:**

Table 5.1: Accuracies of models including combinations of various predictive variables derived from L-band HH/HV backscatter, Digital Elevation Model and rainfall .....	143
Table 5.2: Accuracies of models based on training from Savannah-only versus all-biome data and validated with data from five different biomes.....	143
Table 5.3: Summarised CC RF validation accuracies according to the number, size and total hectares of simulated LiDAR acquisitions acquired from the all-biome dataset and Savannah-only dataset..	151
Table 5.4: Cost analysis of optimal LiDAR acquisition specifications of varying size and number (according to table 5.3).....	155
Table 5.5: Cost analysis of the optimal number of field plots versus the optimal LiDAR acquisition specification .....	157

**Appendices:**

Table 2B: Confusion matrix results of the new Forest/Non-Forest product using a HV threshold of -19dB .....	19488
Table 1: Validation accuracies for modelling CC across various SAR frequencies and algorithms (N= no. of observations) .....	1944

## **Dedications**

This thesis is dedicated to my family, friends and supervisors and colleagues at the CSIR, for their valued input, support, guidance and patience throughout this long PhD journey.

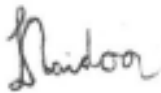
## Acknowledgements

The author would like to acknowledge the Council for Scientific and Industrial Research, Department of Science and Technology, South Africa (grant agreement DST/CON 0119/2010, Earth Observation Application Development in Support of SAEOS), the European Union's Seventh Framework Programme (FP7/2007-2013, grant agreement no. 282621, AGRICAB), and the Southern African Science Service Centre for Climate Change and Adaptive Land Management (SASSCAL) Task 205 ("Adaptation strategies for the South African, Namibian and Zambian dryland forests and timber plantations to climate change"). The X-band StripMap TerraSAR-X scenes were acquired under a proposal submitted to the TerraSAR-X Science Service of the German Aerospace Centre (DLR). The C-band Quad-Pol RADARSAT-2 scenes were provided by MacDonald Dettwiler and Associates Ltd. – Geospatial Services Inc. (MDA GSI), the Canadian Space Agency (CSA), and the Natural Resources Canada's Centre for Remote Sensing (CCRS) through the Science and Operational Applications Research (SOAR) programme. Major thanks go to Dr Waldo Kleynhans for processing the X-band TerraSAR-X dataset. The individual L-band ALOS PALSAR FBD scenes were acquired under a K&C Phase 3 agreement with the Japanese Aerospace Exploration Agency (JAXA). The ALOS PALSAR global mosaic product scenes were acquired free from JAXA. All Landsat-5 TM level-1 imagery were acquired free of charge from the United States Geological Survey (USGS). Special thanks go to Derick Swanepoel for the extraction of the multi-temporal EVI dataset from the WAMIS database and to the South African Weather Services (SAWS) for providing the monthly rainfall dataset. The Carnegie Airborne Observatory (CAO) LiDAR dataset was made possible by the Avatar Alliance Foundation, Margaret A. Cargill Foundation, John D. and Catherine T. MacArthur Foundation, Grantham Foundation for the Protection of the Environment, W.M. Keck Foundation, Gordon and Betty Moore Foundation, Mary Anne Nyburg Baker and G. Leonard Baker Jr., and William R. Hearst III. The KNP LiDAR dataset was acquired by AECOM (UK) and obtained via Dr Izak Smit of the scientific services in Kruger National Park. Very special thanks also go to ESKOM for the provision of the remaining 16 LiDAR datasets used in this thesis. Major thanks also go to Russell Main for the processing of the ESKOM LiDAR datasets and to the CAO team for processing the 2008, 2010 and 2012 CAO LiDAR datasets. The author would also like to acknowledge SAN parks (Dr Izak Smit), Sabi Sands Game Reserve (Michael Grover), WITS Rural facility (Rhian Twine and Simon Khosa), SAEON (Patrick Ndlovu and Mightyman Mashele) and Bushbuckridge local authorities and personnel for arranging land access, field work expertise and providing logistical support. Final thanks also go to Fiona West and Anteneh Sarbanes and Hannah Yang, of Princeton University, for assisting with the preliminary LiDAR dataset processing.

## Plagiarism Declaration

1. *I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.*
2. *Each contribution to, and quotation in, this thesis from the work(s) of other people has been attributed, and has been cited and referenced.*
3. *I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.*
4. *I declare that this thesis is my own work*

Thus, I, Laven Naidoo, declare that the dissertation/thesis, which I hereby submit for the degree PhD Geoinformatics at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.



\_\_\_\_\_  
SIGNATURE

20/04/2017

\_\_\_\_\_  
DATE

## List of Publications

- Naidoo, L., R. Mathieu, R. Main, W. Kleynhans, K. Wessels, G. Asner, B. Leblon. The assessment of data mining algorithms for modelling savannah woody cover using multi-frequency (X-, C- and L-band) Synthetic Aperture Radar (SAR) datasets. *IEEE International Geoscience and Remote Sensing Symposium*, Quebec, Canada, 13-18 July, 1-4 (2014)
- Naidoo, L., R. Mathieu, R. Main, W. Kleynhans, K. Wessels, G.P. Asner, B. Leblon. "Savannah woody structure modelling and mapping using multi-frequency (X-, C- and L-band) Synthetic Aperture Radar data". *ISPRS Journal of Photogrammetry and Remote Sensing*, **105**, 234-250 (2015)
- Naidoo, L., R. Mathieu, R. Main, K. Wessels, G.P. Asner. "L-band Synthetic Aperture Radar imagery performs better than optical datasets at retrieving woody fractional cover in deciduous, dry savannahs". *International Journal of Applied Earth Observation and Geoinformation*, **52**, 54-64 (2016)

## List of Acronyms

AGB	Above Ground Biomass
AOI	Area of Interest
CAO	Carnegie Airborne Observatory
CC	Woody Canopy Cover
CHM	Canopy Height Model
CI	Confidence Interval
dB	Sigma dB
DBH	Diameter-at-breast height
DEM	Digital Elevation Model
DN	Digital Number
EVI	Enhanced Vegetation Index
FAO	Food and Agriculture Organisation of the United Nations
FBD	Fine Beam Dual Polarisation
FNF	Forest/Non-Forest
GLCM	Grey-level Co-occurrence Matrices
JAXA	Japan Aerospace Exploration Agency
K&C	Kyoto and Carbon
KNP	Kruger National Park
LiDAR	Light Detection and Ranging
MGD	Multi-Look Ground Range Detected
NASA	National Aeronautics and Space Administration
NCAS-LCCP	Australian National Carbon Accounting System – Land Cover Change Program

REDD+	Reducing Emissions from Deforestation and Forest Degradation
RF	Random Forest
RMSE	Root Mean Squared Error
SAR	Synthetic Aperture Radar
SEP	Standard Error of Prediction
SLATS	Australian State-wide Landcover and Trees Study
SLC	Single Look Complex
TOC	Top of Canopy
TCV	Total Canopy Volume
VCF	Vegetation Continuous Field
VI	Vegetation Indices
WRC	Water Research Commission



## Thesis Abstract

Savannahs, which are defined as a heterogeneous mixture of herbaceous and woody plant components, occupy one fifth of the global land surface and is the largest biome in South Africa. The woody vegetation structure of savannahs is particularly important as it influences the fire regime, nutrient cycling and the water cycle of these environments and provides fuelwood to sustain the local human populace. Remote Sensing has been proven in numerous studies to be the preferred tool for quantifying and mapping this woody vegetation structure (in this study, defined as woody biomass, woody canopy volume and woody canopy cover metrics) over large areas, mainly due to its superior information gathering capabilities, wide spatial coverage and temporal repeatability. Active remote sensing sensors such as Light Detection and Ranging (LiDAR) and Synthetic Aperture Radar (SAR) are particularly useful in studying woody biomass and other canopy structural metrics, because of their capacity to image within-canopy properties. Passive optical imagery acquired over multiple seasons can also provide tree phenological information which can be used to ascertain the best period for monitoring tree structure, i.e. when tree canopies has sufficient leaves while the grasses are dry. The combined strength of these active (SAR and LiDAR) and passive (optical) sensor technologies, are yet to be applied to their full potential in the dynamic and heterogeneous savannah environment, with a special relevance in Southern African landscapes.

This PhD study aimed to evaluate various methods for estimating and upscaling woody structural metrics of South African savannahs using integrated SAR and optical remote sensing datasets and LiDAR datasets as training and validation. Before this aim could be tackled, two current global-scale remote sensing woody structural products (25m JAXA ALOS PALSAR Forest/Non-Forest or FNF and 30m Landsat-based Vegetation Continuous Field or VCF) were evaluated, within the South African context, with the help of high resolution airborne LiDAR datasets. These datasets were resampled to match the products' criteria and definition used to depict forests. It was found that the FNF product grossly under-represented the distribution of forests in savannah environments (20-80% CC ranges), due to the inadequate HV backscatter threshold chosen in its creation. The FNF product also showed a limited ability in detecting closed forest cover class (90-100%) and Natural Forest and Scrub Forest tree structural classes. The Landsat VCF product displayed strong CC underestimation with increasing variability and mean error from CC values of greater than 30%. The moderate accuracies at the 10-20% CC range (and in the Open Woodland tree structural class) suggests that the VCF product could be potentially applicable in low CC environments such as grasslands and sparse savannahs but can also marginally detect closed canopy environments (90-100% CC range).

These results provide the justification for developing new, locally calibrated woody structural products for South Africa. Next, the aim of this study was addressed, firstly, by developing methodologies for the estimation of key woody structural metrics (above ground biomass, woody canopy cover and woody canopy volume) for the Greater Southern Kruger National Park Region using multi-frequency SAR parameters (X-, C- and L-band backscatter and polarisations). Secondly, the most suitable SAR frequency was then tested against and in combination with various Landsat-5 TM optical features (textures, vegetation indices and multi-seasonal band reflectance) for improved regional modelling of woody canopy cover. In both cases, In-situ field measurements of woody vegetation structure were “scaled-up” to landscape and regional scales by using LiDAR, SAR and/or optical sensor products to produce reliable maps of woody structural metrics. A Random Forest modelling approach was predominantly used to meet the modelling challenges in this study and the LiDAR datasets were used for model calibration and validation.

For the multi-frequency SAR analysis, it was concluded that the L-band SAR frequency was more effective in the modelling of the CC ( $R^2$  of 0.77), TCV ( $R^2$  of 0.79) and AGB ( $R^2$  of 0.78) metrics in Southern African savannahs than the shorter wavelengths (X- and C-band) both as individual and combined (X+C-band) datasets. The addition of the shortest wavelengths also did not assist in the overall reduction of prediction error across sparse and dense vegetation conditions. Although the integration of all three frequencies (X+C+L-band) yielded the best overall results for all three metrics ( $R^2=0.83$  for CC and AGB and  $R^2=0.85$  for TCV), the improvements were noticeable but marginal in comparison to the L-band alone. The results, thus, do not warrant the acquisition of all three SAR frequency datasets for tree structure monitoring in this environment. For the integrated SAR and optical dataset analysis, results showed that Landsat-5 imagery acquired in the summer and autumn seasons yielded the highest single season modelling accuracies, depending on the year but the combination of multi-seasonal images yielded higher accuracies ( $R^2$  between  $\sim 0.6-0.7$ ). The derivation of spectral vegetation indices and image textures and their combinations with optical reflectance bands provided minimal improvement with no optical-only product combination yielding accuracies greater than winter SAR L-band backscatter alone ( $R^2$  of  $\sim 0.8$ ). However, there was significant, yet modest, improvement ( $R^2$  of  $\sim 0.08$ ,  $\sim 1.9\%$  of RMSE and  $\sim 7.5\%$  of SEP) in accuracy when 2010 multi-seasonal optical reflectance bands were combined with the L-band backscatter variables. These results showed that future monitoring of woody cover, in Southern African savannahs, will require priority access to L-band SAR imagery. Finally, in order to move towards upscaling woody canopy cover to the national scale, guidelines on the optimal quantity of field plots and LiDAR coverages, required for model training, were proposed for the country of South Africa.

The results have shown that the Savannah-only training dataset yielded high accuracies across Grasslands, moderate accuracies across Thickets but poorer accuracies in the Indigenous Forests and Fynbos biomes. Sampling the training data across all available biomes yielded higher accuracies. From the LiDAR-simulated field plot analysis, it was concluded that a minimum of 500, 1ha field plots would be sufficient for effective modelling of CC at the country-wide scale. Additional field plots, beyond this number (500) would improve the overall accuracies only slightly, but incurred significant increases in sampling efforts and costs. The most frugal LiDAR acquisition strategy was found to acquire only four separate 5000ha LiDAR acquisitions, distributed across the five vegetated biomes. The study found that much less LiDAR data were required to train the models than originally expected, provided that the acquisitions were sufficiently diverse in CC and vegetation type and could also be cheaper to acquire than collecting 500 1ha field plots. Following the lessons learnt from the various chapter results, a new and more accurate woody canopy cover map of South Africa was introduced which served as a first step towards the establishment of an operational monitoring system for the woody component.

## **Chapter 1: Introduction and literature review**

This PhD study aims to evaluate various methods for estimating and upscaling vegetation woody structural metrics of South African savannahs and forests by using a combination of SAR and optical remote sensing. This study feeds into the long term goal of developing a scientific foundation for the national mapping of the woody vegetation structure of South African savannahs and forests as only a limited knowledge base exists with no reliable, continuous and up-to-date geospatial data products being currently available (DAFF, 2015; Skowno et al., 2016). With the increase of tree cover at a rate of 5-6% per decade and the added threat of bush encroachment encroaching upon approximately 10-20 million hectares of land and alien invasive plants spreading at a rate of between 5 and 10% per year in South Africa, the creation of such map products is crucial (O'Connor et al., 2014; van Wilgen et al., 2012). In this chapter, Savannahs and the importance of its woody component in ecosystem processes and their monitoring will be introduced. The woody component will be broken down into woody structural metrics – woody biomass, woody canopy volume and woody canopy cover – for purposes of quantification. Methods which utilise remote sensing and the role of multi-sensor data integration will be reviewed as a primary means of monitoring and measuring these various woody structural metrics and compared to traditional field-based measurements. Finally the main research aim, objectives and specific research questions, which will be addressed in the subsequent analytical chapters, will be introduced. In this thesis, in order to eliminate any potential confusion between the terms savannahs and forests, these terms will be used without separation except when mentioned at the biome level. The reason for this, according to the FAO definition of forests (elaborated in Chapter 2), most savannahs systems can be potentially classified as forests but not all forests can be savannahs (e.g. Natural forests).

### **1.1 Savannahs and the importance of its woody component in ecosystem and monitoring processes**

Savannah woodlands cover half of the African continent and occupy one fifth of the global land surface (Scholes and Walker, 1993). Within the context of South Africa, the Savannah biome is the largest and makes up 35% of the country (Van Wilgen, 2009). Savannahs are broadly composed of herbaceous and woody components which are in a constant state of flux (Meyer et al., 2007). In this biome, total woody canopy cover values can range from dispersed trees in open-grasslands (~5%) to near-closed canopy woodlands (~60%) and more than 80% in riparian zones (Venter et al., 2003). Vegetation height can range between 1 and 20 metres (Low and Rebelo, 1996) and also possess an

above ground biomass range of less than 60 tonnes per hectare (Scholes and Walker, 1993). The Savannah biome contains six bioregions (the Central Bushveld, Mopane, Lowveld, Sub-Escarpment, Eastern Kalahari Bushveld and the Kalahari Duneveld bioregions) which vary according to their geographical locations, geology and soil types and dominant vegetation species (Rutherford et al., 2006). Savannahs, additionally, consist of Clay Thorn Bushveld, Mixed Bushveld, Sweet and Sour Lowveld Bushveld vegetation types (Mucina and Rutherford, 2006). Seasonally, savannahs experience phenological fluctuations in both herbaceous and woody components which influence their associated distribution in the landscape. In summer, both tree leaves and grasses are green while in autumn, grasses are dry with trees remaining green but beginning to lose leaves. In winter, most trees have lost leaves and grasses are dry while in spring, grasses are fairly dry while the trees first undergo a green flush of leaves (Archibald and Scholes, 2007). Consequently, savannahs are seen as highly complex in both vegetation structure and composition and are highly heterogeneous ecosystems.

Water availability and disturbance factors, such as fire and herbivory, mainly control the balance between the herbaceous and woody components in savannahs (Baudena et al., 2015; Sankaran et al., 2008). The distribution of the woody component, in savannahs, is constrained in areas which receive a mean annual precipitation of less than 650mm (Sankaran et al., 2005) and under the driest conditions (<200mm), savannahs do not occur as the herbaceous component outcompetes the tree saplings of the woody component (Baudena et al., 2015; Sankaran et al., 2004). As precipitation increases, however, the woody component can outcompete the grassy component with deeper and more established root systems than the herbaceous component (Jose and Montes, 1997). Above 650mm of mean annual precipitation the woody canopy has a cover above 80% unless disturbance factors are present. Fire is a driver that regulates the tree-grass balance in savannahs by preventing savannahs from becoming forested woodlands (Jose and Montes, 1997). Fire can promote the growth of the herbaceous layer by eliminating the woody component (including tree seedling recruitment) and also due to the quicker recovery ability displayed by the herbaceous layer (Baudena et al., 2015; Hanan et al., 2008). On the other hand, excessive herbivory of the herbaceous layer can promote increased growth of the woody component via bush encroachment (Ward, 2005; Wigley et al., 2009). This study will solely focus on the woody component which has a considerable impact on both natural and anthropogenic processes.

The savannah woody component impacts the fire regime, biomass production, nutrient cycling and the water cycle of these environments (Sankaran et al., 2008). From an anthropogenic point of view, the woody component provides numerous essential ecosystem services such as fuelwood (mostly firewood and self-produced charcoal derivatives), medicinal products, construction timber and edible fruits (Shackleton et al., 2007), which sustain the needs of the large rural populace in sub-Saharan Africa and regions of South Africa (Twine, 2005; Wessels et al., 2013, 2011). In South Africa, approximately 800 000 people of the rural populace heavily rely on this woody component as a source of income through the craft industry and as well as through the small scale trading of forest products (DAFF, 2015). Overall, this savannah woody component contributes approximately R17 billion to South Africa's annual Gross Domestic Product (GDP) (DAFF, 2015). Conversely, the densification of the savannah woody component, or bush encroachment, can also severely compromise the availability of grazing resources, that are essential to livestock populations and related human livelihoods (O'Connor et al., 2014; Wigley et al., 2009). Bush encroachment adversely affects agricultural productivity and biodiversity (e.g. loss of palatable grass species (Angassa, 2005)) of approximately 10-20 million hectares of South Africa (Ward, 2005). From an economic standpoint, neighbouring countries like Namibia, which rely heavily on livestock farming, have registered an annual loss in income of more than N\$700 million due to bush encroachment with approximately 12 million hectares of land already being severely encroached (De Klerk, 2004). Factors such as humans (via wood harvesting activities), African elephants and fire (less so than elephants and humans), in communal rangelands and protected areas, have also been found to alter the woody component by removing large trees which subsequently promotes an increase in shrub cover or encroachment due to reduced tree seedling survival rates caused by these factors (Asner and Levick, 2012; Asner et al., 2016; Mograbi et al., 2016).

Within the context of climate change, the sequestration of carbon by growing vegetation is understood as a significant mechanism for the removal of CO<sub>2</sub> from the atmosphere (Viergever et al., 2008b). With a mean net primary productivity of 7.2 tonnes Carbon per hectare per year, savannahs account for approximately 40% of the global carbon store (Collins et al., 2009; Grace et al., 2006). Understanding how carbon is stored as carbon sinks in vegetative biomass and quantifying this standing biomass is of paramount importance to understanding of the global carbon cycle. Initiatives such as REDD (Reduced Emissions from Deforestation and Degradation) and the Bonn Challenge provide incentives to developing countries by linking forest conservation to market-related monetary values of carbon stock. Adverse anthropogenic activities such as deforestation, from unsustainable harvesting, and the burning of biomass can turn carbon sinks into carbon

emission sources (Viergever et al., 2008b). These activities are especially prevalent in developing regions around the world such as the savannah woodlands of South America and Southern Africa. Conversely, with the increase in carbon dioxide (CO<sub>2</sub>) in the atmosphere over the past decade, vegetation growth in grasslands and savannahs has increased which trees are growing at a faster rate while utilising less resources to grow (Bond and Midgley, 2012; Stevens et al., 2015). As a result, increases in wooded savannahs, in terms of higher biomass and woody plant species presence, are predicted in the future according to the current climate condition trajectory (Higgins and Scheiter, 2012; Stevens et al., 2015). Given the importance of the woody component in global savannahs and the significant changes it undergoes on short and long-term time scales, it is essential to monitor the woody component effectively through time and space.

## **1.2 The current status of monitoring the woody component in South Africa**

Despite the environmental and anthropogenic importance of woody vegetation, particularly in savannahs, there is currently no monitoring programme available at the national level for South Africa to produce reliable and up-to-date products of the distribution and amount of the woody component (DAFF, 2015). This current inability to monitor the woody component has important legal implications as the South African government has a national requirement to report on the status of forests on a three year basis (Willis, 2002). The government also have additional legal obligations as signatories to various international treaties such as the Kyoto Protocol, United Nations Convention to Combat Desertification, United Nations Forum on Forests and the Food and Agriculture Organisation Forest Resource Assessment to map and monitor national carbon stocks (DAFF, 2015; DEA, 2010; Main et al., 2016). Consequently, insufficient spatial and quantitative information on the extent, the amount and possible changes in the South African woody component, especially in savannahs, has prevented management from sustainably managing, monitoring and utilising this woody resource. At the regional scale, various governmental initiatives such as Working for Water have been introduced in 1995 to monitor this woody component by reducing the density of established invasive alien plants (IAPs) via mechanical and chemical control especially along vital watershed catchments where IAPs restrict and limit water flow (Buch and Dixon, 2009; Richardson and Van Wilgen, 2004). Even after extensive clearing efforts and financial investment (approximately R1.8 billion in 2015), spreading rates are estimated to still be between 5 and 10% per year (van Wilgen et al., 2012). Similar challenges are being faced by the Working for Land project, established in 1997, to curtail bush encroachment via limited and small scale means such as invasive shrub removal, use of herbicides and the establishment of fire breaks. Ad-hoc and

sporadic studies, of variable temporal and spatial extents (mostly localised), have shown the rate of woody cover change to be between -0.131 and 1.275% per annum in Southern Africa with the majority reporting a net increase as a result of bush encroachment (O'Connor et al., 2014; Skowno et al., 2016). Across a variety of land use and management and rainfall gradients, only conservation areas with elephants seem not to be subjected to bush thickening. The exact spatial extent of such spread from both IAPs and bush encroachment is also currently unknown. From the perspective of global initiatives such as REDD+ and the Bonn Challenge, the identification of large, contiguous areas of degraded and fragment land is crucial before various forest restoration efforts can be made.

In order to take the necessary steps to create a national monitoring programme of the woody component, various challenges still currently remain unaddressed in the scientific literature. These challenges include determining which remote sensing datasets are most appropriate for mapping the woody component across Southern Africa, testing for the most effective modelling approaches to achieve the best possible accuracies and, finally, determining the optimal amount of training and validation data required to achieve the development of such a monitoring programme. In the absence of such a monitoring program, global forest products, derived from global modelled datasets, have been drawn upon, often erroneously. These global forest products are elaborated upon in section 1.4.3.

## **1.3 Woody structural metrics**

The woody component can be assessed via a variety of variables such as species composition, physiology (i.e. stress and productivity), phenology and structure. The structural variables of the woody component will be the focus of this study. The following main quantifiable variables were chosen as representative measurements or metrics which comprise of the savannah woody component: biomass, woody canopy volume and cover. These variables are by no means exhaustive but are capable of providing both two dimensional and three dimensional metrics of the woody component. Each of these woody component variables will be separately explained within the context of their definition, ecological importance and techniques used for measurement.

### **1.3.1 Woody biomass**



Biomass is defined as the mass of live or dead organic matter and is usually expressed in mass per unit area (Bombelli et al., 2009; Brown, 1997). The general term 'biomass' is made up of above-ground biomass (AGB), below-ground biomass (BGB) and dead mass and litter (Ghasemi et al., 2011; Lu, 2006). AGB is generally recognised as the main contributor of the total biomass and will be the focus of this study as BGB cannot be studied with any other means beside labour and time intensive in-situ sampling. In heterogeneous environments such as savannahs, AGB estimation is particularly challenging because of the complex stand structure as a result of the abundant species diversity of the vegetation (Lu, 2006). There are various methods for estimating AGB which varies depending on the spatial scale at which these estimates are predicted. The first method is an in-situ, destructive but direct biomass measurement which involves the manual harvesting of plants, drying them and then weighing the biomass. This is the most accurate and direct method, however it is extremely intensive in both labour and time and is usually limited within a small unit area such as at a single tree or plot level (Bombelli et al., 2009; Lu, 2006). The second is an in-situ, non-destructive measurement which does not involve the harvesting of plants but requires the collection of plant biometric measurements (e.g. height, diameter-at-breast height or DBH etc.) for input into allometric equations. These allometric equations are mathematical functions that relate tree dry mass to one or more tree dimensions, such as diameter or height, and can be used to extrapolate biomass to the unit ground area (Bombelli et al., 2009; Brown, 1997; Colgan et al., 2013; Nickless et al., 2011; Sawadogo et al., 2010). The final method entails the inference and mapping of regional level biomass from remote sensing data and related models. This particular woody structural variable is proven to be vital for governance in light of the REDD+ initiatives as it serves as a direct indicator of carbon (Global Forest Observations Initiative, 2016) and is also important for a number of applications such the sustainable assessment of fuelwood stocks in communal rangelands (Wessels et al., 2013) or the assessment of biomass resources for bioenergy projects (GOFC-GOLD, 2017).

### **1.3.2 Woody canopy volume**

Woody canopy volume, in its simplest form, can be derived from a simple product of canopy height and canopy cover which would indicate the cylindrical volume of vegetation. Other definitions of volume can be linked to various applications such as the forestry industry which relies on estimating stem volume or bole volume which represents the volume of the tree stems per unit area, including bark but excluding the branches and stump (Santoro et al., 2011). Woody canopy volume, usually derived from volume-based allometric equations using DBH and sometimes height measurements

(Abbot et al., 1997) at the in-situ level, can also be measured using remote sensing technologies. The one of the methodologies of deriving this woody structural parameter will be covered in greater detail in chapter 3. This variable serves as a valuable proxy of biomass density and distribution especially when biomass measurements are not possible due to the lack of available site specific allometry, for instance. It also provides the means of investigating in a synthetic indicator both woody cover and height variables which are highly variable across the savannah landscape. Apart from its importance in the forestry industry for wood volume yields and derived woody products (Foroughbakhch et al., 2012), higher tree volumes are associated with a wider ecological niche and correlates positively with both economic biodiversity value (EBV) and biodiversity indices (Hashemi, 2011; Merganic et al., 2013).

### **1.3.3 Woody canopy cover**

Woody canopy cover is a simple and widely used structural metric which define the area vertically projected on a horizontal plane by woody plant canopies (Jennings et al., 1999). Canopy cover is thus a two dimensional structural metric which indicates the spatial heterogeneity and possible fragmentation in the ecological landscape. When combined with canopy height, it can provide an informative indicator of volume and serve as an indirect proxy for biomass (Colgan et al., 2012). At the in-situ level, canopy cover can be measured with the use of various sampling strategies such as the vertical densitometer technique (Ko et al., 2009; Stumpf, 1993) which uses a point intercept sampling approach. The point intercept method is a small angle approach but a large angle approach called the “morphing” approach (Williams et al., 2003) has also been utilised. This approach morphs data from a circular fixed-area plot to a square one and then uses a torus edge correction technique to model the crowns of tree boles which fall outside a fixed plot but have their canopies partially covering portions of the plot (Williams et al., 2003). At the landscape scale, however, the canopy cover variable is adequately measured by remote sensing datasets. Measuring canopy cover in both levels will be elaborated upon in greater detail in chapter 3.

## **1.4 Remote sensing of woody structure**

Remote Sensing has been proven in numerous studies to be the preferred tool for the quantification and mapping of this woody component mainly due to its superior information gathering capabilities, wide spatial coverage and revisit capacity. In contrast to the limited spatial scope of ground based

techniques, remote sensing also has the ability to sense the high spatio-temporal variability of woody height, cover and biomass, as well as tree species diversity and plant phenological status – a defining but challenging set of characteristics typical of South African savannahs (Archibald and Scholes, 2007; Cho et al., 2012b; Mills et al., 2006). Additionally, remote sensing is more cost-effective, repeatable and most importantly, capable of effectively predicting environmental variables over large geographical areas. When predicting regional biomass and other woody structural parameters using remote sensing data, electromagnetic radiation (e.g. visible, infrared or microwave) interact at different spatial scales with the woody component via direct (e.g. sensed responses such as reflectance) or indirect (e.g. remote sensing derived products such as leaf area index or LAI) means (Bombelli et al., 2009; Lu, 2006). This is usually achieved with the use of models which can incorporate multi-scale (from in-situ field measurements to regional remote sensing derived parameters) and multi-sensor type (passive and active sensor) data in the analysis. It is important to note however, that the more open African savannah environments are relatively understudied in the field of remote sensing in comparison to other environment types such as dense forested environments (e.g. tropical and temperate forests) and other biomes (Gwenzi and Lefsky, 2014; Gwenzi, 2017). Though limited in the number of available studies, a variety of passive and active remote sensing sensor technologies have been employed to assess the savannah woody component at various spatial scales: Light Detection and Ranging (LiDAR) (Fisher et al., 2014; Mograbi et al., 2015), Synthetic Aperture Radar (SAR) (Mathieu et al., 2013; Mitchard et al., 2012; Ryan et al., 2012), optical and integrated sensor platforms (e.g. Carnegie Airborne Observatory or CAO system which integrates both hyperspectral and LiDAR sensors on the same platform, (Asner et al., 2007)).

#### **1.4.1 Passive remote sensing of woody structure**

AGB and other woody structural parameters have been successfully mapped using optical data from fine to coarse spatial scales (Boggs, 2010; Castillo-Santiago et al., 2010; Nichol and Sarker, 2011). This is made possible as forest structural characteristics (such as tree height, crown diameter etc.) can be measured from stereoscopic measurements, spectral and texture orientated modelling techniques (Lu, 2006). In terms of the electromagnetic spectrum, the red edge region has been proven to be related to woody structure, health, and leaf and canopy biophysical factors (Cho et al., 2012a, 2008; Delegido et al., 2011) and also played a role in estimating fresh and dry grass biomass (Cho et al., 2006). Image texture is defined as a function of the local variance in an image which is related to the spatial resolution and the size of target scene objects (e.g. tree canopies) (Nichol and

Sarker, 2011; Wood et al., 2012). For example, trees occurring over a bare soil background would increase the variance through sunlit and shaded pixels thus creating image texture. The major drawback of optical data, however, is the influence of high spectral variation and shadows at fine resolutions, resulting from canopy and topographic effects, and the issue of sensor signal saturation (e.g. MODIS sensor data) and mixed pixels, at the medium and coarser resolutions, on AGB model development and associated accuracies (Lu, 2006). Clouds and haze also detrimentally obscure optical data which, in African savannahs, are prevalent in summer (due to the rainy season) and winter (due to dry season veld fires). Another challenge is the effects of phenology on optical imagery in savannah environments which undergo distinct phenological seasonal changes during which the green fractional cover of grasses and woody plants varies considerably (Archibald and Scholes, 2007). These phenological seasonal changes could introduce noise especially during the wet or growing season when both woody plants and grasses are green. Thus, identifying the time period during the annual vegetation cycle at which a maximum contrast is achieved between green tree canopy and dry grass is important (Zeidler et al., 2012). These phenological changes also, however, experience noticeable inter-annual variability especially during years which experience periods of severe drought or high rainfall.

#### **1.4.2 Active remote sensing of woody structure**

Active remote sensing sensors such as LiDAR and SAR are particularly useful in studying woody biomass and other canopy related structural metrics, because of their capacity to image within-canopy properties. Airborne LiDAR systems provide high resolution geo-located measurements of the tree's vertical structure (upper and lower storey) and the ground elevations beneath dense canopies while SAR systems provide backscatter measurements which are sensitive to forest spatial structure and standing woody biomass due to its sensitivity to canopy density and geometry (Hall et al., 2011; Mitchard et al., 2011; Sun et al., 2011). Both sensors have an ability to penetrate vegetation canopies with SAR being unrestricted by challenging weather conditions such as dense cloud cover which would inhibit LiDAR and optical data acquisitions (Mitchard et al., 2011). SAR systems also operate at night, and altogether with all-weather capacity they can provide denser systematic "guaranteed" time series. However, unlike LiDAR sensors, the backscatter signal of SAR sensors can saturate (i.e. a reduction in the net backscatter due to the extinction of the signal – (Collins et al., 2009)) depending on factors mainly related to the frequency and polarisation of the sensor being used and density of vegetation structures being sensed. It was found that under these conditions of signal saturation, SAR backscatter correlated negatively with biomass as a result of

signal attenuation from denser forest canopies (Mermoz et al., 2014). This would result in a higher than expected under-estimation of biomass past the point of saturation. Given that South African savannahs possess a low to medium above ground biomass range of less than 60 tonnes per hectare (Mathieu et al., 2013; Scholes and Walker, 1993), it is expected that SAR signal saturation would not be an issue in this study. Another disadvantage of SAR, however, is that due to the side-looking nature of these sensors SAR backscatter is adversely affected by steep slope and topography in which the creation of artefacts such as foreshortening, shadowing and layover effects and backscatter calibration error are possible (Otukey et al., 2015; Van Zyl, 1992; van Zyl et al., 1993). These artefacts and calibration error would complicate the analysis of vegetation structure over such terrain.

Although the LiDAR technology is well established and the most suited remote sensing technology for mapping structure with high accuracies, airborne-based LiDAR systems are not well-suited to regional scale mapping as data acquisition is constrained by operational restrictions such as expensive flight campaigns, and access to sensors and data is dependent on the country. (Popescu et al., 2011) and (Lefsky et al., 1999), however, did successfully make use of canopy height metrics derived from satellite and small footprint airborne LiDAR to estimate forest AGB. Few studies have also utilised various LiDAR derived canopy metrics (e.g. plot-level and tree-level height and canopy cover metrics) to estimate AGB in the South African savannah environment (Colgan et al., 2013, 2012). Additionally, space-borne LiDAR missions (e.g. MOLI – multi-footprint observation LiDAR and Imager – to be launched by JAXA in late 2019 and GEDI – Global Ecosystem Dynamics Investigation LiDAR – to be installed by NASA on the International Space Station in late 2018) are coarse scale sensors with large gaps between samples which are inadequate for producing consistent maps across the landscape (the latter is also only a two year mission). Due to its precision and accuracies over a limited coverage, airborne LiDAR data can be extremely useful in creating a large representative ground truth dataset, once validated with collected field data, for regional scale modelling using coarser datasets (Mathieu et al., 2013; Naidoo et al., 2015). Compared to other high resolution optical imagery, however, airborne LiDAR, is the most expensive with a cost of approximately 1-5 US\$ per hectare depending on the total coverage, sensor specifications and location of deployment (Hummel et al., 2011; Kelly and Di Tommaso, 2015; Thompson et al., 2013; Wulder et al., 2008). Wall-to-Wall, repeat acquisitions of an entire country, particularly as large as South Africa (122.1 million ha), is currently not financially feasible, and thus there needs to be a trade-off between the area sampled with LiDAR and the total cost incurred (Ene et al., 2016; Wulder et al., 2008). With this in mind, it is thus imperative to establish a much needed guideline for the

quantity and distribution of LiDAR acquisitions required for training and validation of models in a national woody component monitoring system.

The concept of polarimetry, i.e. radiowave orientation, in SAR theory has played an important role in understanding ecosystem structure (Sagues et al., 2000). Polarimetric SAR systems emit and receive waves potentially in HH, HV, VH and/or VV polarisations with H referring to a horizontal wave orientation and V referring to a vertical wave orientation. This allows for a complete characterisation of the scattering properties of various ground targets which in turn, enables the extraction of greater structural information. Some SAR systems offer only single polarimetry – one polarization (e.g. ERS-1), dual polarimetry – two polarizations (e.g. Sentinel-1), or full polarimetry (e.g. RADARSAT-2) when all four polarizations are available. Additionally, when a system is fully polarimetric decomposition theorems (e.g. Freeman-Durden) can be applied to simulate and quantify dominant scattering mechanisms (volume, double bounce and single bounce) and relate these mechanisms to specific target properties such as volumetric scattering within tree canopies etc. (Touzi et al., 2004). (Le Toan et al., 2011) mapped biomass at a global scale (from 70°N to 56°S at 100-200m spatial resolution) by utilising P-band frequency fully polarimetric (HV) SAR backscatter data, modelled against in-situ biomass measurements, and interferometric SAR techniques. As an alternative to the modelling of SAR scattering and polarimetric variables, (Balzter et al., 2007) made use of polarimetric interferometric SAR (InSAR) techniques, in deciduous woodland, for the direct estimation of forest canopy height which allowed for the indirect prediction of AGB. Polarimetric InSAR principles involve the polarimetric separation of scattering phase centres in order to estimate tree canopy height (Balzter et al., 2007). Similar methods, involving X-band and C-band, have been explored in tropical savannah environments (Viergever et al., 2008a) but none have been attempted in South African savannahs with any reasonable success. Finally, (Mathieu et al., 2013) tested fully polarimetric RADARSAT-2 (C-band) in a Southern African savannah to assess various woody structural metrics. It was found that the HV band was the best single predictor over the other polarizations and that the polarimetric decomposition variables did not perform better than the simple intensity bands. This work also suggested that dual polarimetry SAR sensors may be more than suitable for assessing vegetation structure in open savannahs. Similar work conducted by (Urbazaev et al., 2015) with dual and fully polarimetric ALOS PALSAR (L-band) data also suggested the importance of co- and cross polarised backscatter channels (HH and HV) for woody cover assessment in South African savannahs, and confirmed the limited benefits of polarimetric decomposition for quantitative retrievals of forest parameters.

### 1.4.3 Global forest remote sensing products

With the increased availability of systematic and frequent acquisitions of high resolution remote sensing datasets and the development of integrated large scale processing platforms, global scale forest products were able to emerge to map the woody component. Well-known products include: high resolution (30m) global forest cover maps, derived from Landsat Data (Hansen et al., 2013); a 30m global continuous fields tree cover product, derived from Landsat-based rescaling of MODIS data (Sexton et al., 2013); a 25m global forest/non-forest (FNF) classification product derived from ALOS PALSAR L-band SAR backscatter intensity datasets (Shimada et al., 2014). These products were developed primarily as a means to highlight the extents of forest loss and gain at the global and possibly regional scales which can serve as a proxy for the impact on various ecosystem services such as biodiversity richness, carbon and nutrient storages and fluxes, water supply and exchange and also various climate implications (Hansen et al., 2013; Sexton et al., 2013). These global forest cover products (e.g. the (Hansen et al., 2013) product and ALOS PALSAR FNF), however, have mainly been validated against reference data collected in dense, homogeneous equatorial forested areas of Africa and other countries rather than in heterogeneous savannah and forested types with variable canopy cover and height profiles. As a result, most of these global forest products have yet to be accurately validated at the regional scale in South Africa. Due to the lack of available South African forest products, created from local training and validation datasets, these global forest products are temporally serving the need to monitor the woody component but with unknown local accuracies. Assessing whether these products are suitable for the monitoring of the South African woody component is thus of utmost importance.

### 1.5 Multi-sensor and multi-temporal remote sensing data integration

Remote Sensing techniques and derived models have steadily moved from the reliance on a single sensor type (e.g. SAR or LiDAR alone) to multi-sensor integration approaches. These data integration approaches amalgamate various sensors and derived features (e.g. optical-based texture, laser pulse return and microwave backscattering data), multi-temporal data (datasets acquired at different seasons), multi-frequency data (e.g. L- band and C-band SAR) and multi-polarised SAR data (HH, HV, VH and VV) in various modelled approaches. The frequency or wavelength of the SAR sensor can have a major influence on the structural features sensed in the ecosystem. For example, when sensing vegetation, the signal of shorter SAR wavelengths (e.g. X-band and C-band) interact with the fine leaf and branch elements of the vegetation resulting in canopy level backscattering with very

little signal penetration. The signal of longer SAR wavelengths (e.g. P-band and L-band), on the other hand, can penetrate deeper into the vegetation with backscatter resulting from signal interactions with larger vegetation elements such as major branches and trunks (Mitchard et al., 2009; Vollrath, 2010). Combining the properties of these different SAR frequencies in a multi-sensor approach can greatly enhance the sensing of the savannah woody component (Schmullius and Evans, 1997) which possesses a combination of fine and large woody elements within individual tree canopies and a heterogeneous distribution of large trees and smaller shrubs throughout the landscape.

The change in climate (rainy or dry) and vegetation phenology (green or senescent) throughout the seasons of a year can also have a dramatic impact on the scattering and reflecting characteristics of multi-sensor remote sensing datasets. Factors such as ground moisture and leaf-on and leaf-off vegetation conditions can either enhance or diminish SAR signal penetration and scattering and the reflectance of optical spectra (Global Forest Observations Initiative, 2016; Main et al., 2016; Naidoo et al., 2016; Urbazaev et al., 2015; Zeidler et al., 2012). Understanding these seasonal influences on these datasets will shed some light on which temporal frame would be best for sensing the savannah woody component.

(Sun et al., 2011) made use of LiDAR and SAR synergies for the mapping of forest biomass in which a comparable biomass map was generated using limited ground biomass data and SAR polarimetric and coherence variables derived from interferometric pairs. The advantages of the integrated approach was best illustrated by (Lucas et al., 2008) which made use of integrated Compact Airborne Spectrographic Imager (CASI) hyperspectral and LiDAR data to retrieve and map forest AGB and tree component biomass at the individual tree or tree cluster level and then scale-up to plot or stand level. This was made possible by utilising the optical CASI hyperspectral data as a means to delineate crowns and for species identification. The component biomass for the individual delineated trees was then estimated using LiDAR derived height and diameter measurements which were used as inputs into the species-specific allometric equations (Lucas et al, 2008). (Tsui et al., 2012), on the other hand, made use of multi-frequency SAR data (C-band and L-band data) for improved biomass estimations in coniferous temperate forests of Canada. (Collins et al., 2009) also made use of multi-frequency (P band, L band and C band data) fully polarimetric (HH, HV, VH, VV modes) SAR data to estimate AGB and carbon storage of Eucalypts in the open-forest savannahs of North Australia. Despite the success achieved in these various studies via combining different SAR wavelengths (Mougin et al., 1999; Tsui et al., 2012), the combined strength of both shorter (e.g. X-



and C-band) and longer SAR frequency (e.g. L-band) sensor technologies, however, have yet to be assessed in the heterogeneous and complex Southern African savannah environment.

Given the sensitivity of optical sensors to photosynthetically active vegetation and the sensitivity of SAR backscatter to vegetation structure, their possible integration may yield improved woody structure estimates due to complementary information which neither sensor type could provide alone. The integration of optical products has also proven useful in assisting the determination of shrub-based and coppicing tree cover (and possibly biomass) which is not easily identified in the LiDAR and SAR data products (Ghasemi et al., 2011). For example, the work by (Moghaddam et al., 2002) illustrated improve estimation of forest variables by the fusion of SAR (AIRSAR and TOPSAR) data and optical multispectral Landsat TM data which yielded higher modelled accuracies than the use of each dataset type alone. Other studies in dense forested environments, savannahs and plantations also integrated these two sensor technologies and yielded favourable results (Laurin et al., 2013; Lucas et al., 2006b; Moghaddam et al., 2002). None of these studies, however, have investigated the effects of vegetation phenology on optical imagery, especially in savannah environments with complex tree and grass phenological seasonal changes. Integrating optical and SAR imagery of the most appropriate phenological window (i.e. maximum contrast between green tree canopies and dry grass) could improve the modelling of the woody component in South African savannahs.

Despite the success achieved in these various studies, the combined strength of these active (SAR and LiDAR) and passive (optical) sensor technologies, however, have yet to be applied to a more heterogeneous and complex environment such as Southern African savannahs. This is evident from gaps in the literature for savannah environments in South Africa. The aim, objectives and specific research questions of the thesis will be detailed next.

## 1.6 Study aim

The overall aim of the thesis was to evaluate various methods of estimating and upscaling woody structural metrics of South African savannahs using integrated SAR and optical remote sensing datasets and LiDAR datasets as training and validation data.

Study areas ranging from the Greater Kruger National Park region to the eastern half of South Africa were chosen as the focus in this thesis.

## 1.7 Study objectives and chapter breakdown

The objectives of the thesis were:

- 1) To comprehensively validate current global-scale remote sensing woody structural products, within South Africa, using high resolution airborne LiDAR datasets. This task will serve as an important, quantitative benchmark for assessing the performance of these global products in South African forests and savannahs and thus providing the justification for the methodological development of new savannah-specific products in South Africa. This objective will be addressed in Chapter 2.
- 2) To develop and assess methodologies for the estimation of key woody structural metrics (biomass, woody canopy cover and woody canopy volume) for the Kruger National Park region using multi-frequency SAR parameters (backscatter and polarisations) and optical features derived from multiple remote sensing sensors. For this objective, In-situ field measurements of woody vegetation structure and biomass are “scaled-up” to landscape and regional scales by using LiDAR, SAR and optical sensor data to produce maps of woody structural metrics. As a prelude, various parametric and non-parametric modelling algorithms were tested in order to ascertain the best approach and these results are reported in detail in Appendix 3C. Two separate analytical chapters addressed this current objective. Chapter 3 focused on the woody structure modelling and mapping using multi-frequency SAR datasets (X-, C- and L-band). Chapter 4 investigated the benefits of combining optical data with L-band SAR datasets for estimating woody canopy (fractional) cover.
- 3) To investigate the scaling up of the woody structural mapping approach (developed in objective 2) to national scale while considering the challenges of predicting woody vegetation structure across diverse environments (different biomes, vegetation types, rainfall gradients and variable topography) and the LiDAR data requirements, (e.g. how much? And where?), for successful model training and validation over the entire country. This is particularly challenging for a country where the woody component is also dominantly present (and biased) across the savannah biome. The trade-off between the accuracy of model training and increasing LiDAR acquisition costs are considered. Optimal, but representative sampling with airborne LiDAR across the diverse vegetation types of South Africa is a vital challenge to address when developing a national scale monitoring system.

This research objective (to be addressed in Chapter 5), together with the lessons learnt from previous chapters, will help shape the requirements and specifications of a national woody structure monitoring system.

The final chapter, Chapter 6, includes a summary of the study's conclusions, recommendations and the ways forward.

## 1.8 Primary and secondary research questions

### Chapter 2

- Research Question 2.0: How accurate are two global forest products, the 30m Landsat-derived Vegetation Continuous Field (VCF) and the 25m JAXA ALOS PALSAR Forest/Non-Forest (FNF) global products, when validated against high resolution airborne LiDAR datasets across South African forests and savannahs?
- Research Question 2.1: Across which canopy cover ranges do the two products yield the highest and the lowest accuracies?
- Research Question 2.2: Across which vegetation structural type (e.g. grassland, woodland and natural forest (Willis, 2002)) do the two products yield the highest and lowest accuracies?

### Chapter 3

- Research Question 3.0: How do various SAR frequencies (X- or C- or L-band) perform in predicting woody canopy cover, woody canopy volume and above ground biomass in the Southern African savannahs of the Kruger National Park?
- Research Question 3.1: Does combining SAR backscatter of different frequency (X+C or X+L or C+L band or X+C+L-band) improve the predictions of the various woody structural metrics over the single SAR frequencies and by how much?
- Research Question 3.2: What does the examination of the patterns of error, from the different SAR frequency models, inform us on how the different SAR frequencies interact within South African savannah landscape?

## Chapter 4

- Research Question 4.0: Does the combination of SAR (ALOS PALSAR L-band) and multi-seasonal optical (Landsat-5) remote sensing datasets improve woody canopy cover estimation in comparison to the individual datasets alone?
- Research Question 4.1: Which season or seasons of Landsat-5 data is/are best for predicting woody canopy cover?
- Research Question 4.2: How does the accuracy of woody canopy cover predictions compare when using single and multi-seasonal Landsat versus L-band dual-polarised SAR datasets?
- Research Question 4.3: Does the integration of optical predictor parameters (e.g. textures, vegetation indices, and/or raw reflectance etc.) with L-band SAR data, improve the overall modelling accuracies? If so, how do these accuracies compare with the modelling results using only the SAR datasets?

## Chapter 5

- Research Question 5.0: What is the optimal representative sampling of airborne LiDAR data and LiDAR simulated field plots, across Savannah-only and all main biomes (Savannah, Grassland, Fynbos, Thicket and Indigenous Forest) for the training of models predicting woody canopy cover at the country level using ALOS PALSAR L-band SAR data? Secondary objectives also include the investigation of the inclusion of regional environmental variables (i.e. digital elevation-based and rainfall variables) for potential model improvements.
- Research Question 5.1: Does the inclusion of regional ancillary variables such as elevation, slope, and aspect and rainfall gradient improve the accuracy of modelling woody canopy cover when compared to using only the L-band HH and HV backscatter?
- Research Question 5.2: What is the impact on model accuracy of having LiDAR data that are limited to a single biome, i.e. the Savannah? More specifically, is LiDAR data which is limited to the Savannah biome (as specified in (Rutherford et al., 2006)) sufficient for training and validation for L-band SAR-based modelling and mapping of woody canopy cover for the whole country? Also, how do these results of using LiDAR from the Savannah only compare to those where diverse LiDAR datasets from Fynbos, Thicket, Grassland and Indigenous Forest biomes are used?

- Research Question 5.3: What is the optimal amount of field plots, as simulated from LiDAR datasets, required for modelling and mapping of woody canopy cover with L-band SAR across the country and in Savannahs only? The 'optimal amount', in this case, refers to the most favourable trade-off between modelling accuracies and sampling effort (i.e. number of field plots).
- Research Question 5.4: What is the optimal amount, in terms of area (hectares) and number of acquisitions of LiDAR data required for optimal L-band SAR-based modelling and mapping of woody canopy cover within (i) the Savannah and (ii) country-wide, in comparison with the accuracies achieved using an optimal number of field plots? The 'optimal amount', in this case, refers the most favourable trade-off between modelling accuracies and sampling effort (i.e. the number, size and total coverage of LiDAR acquisitions while taking into account the cost effectiveness of the various LiDAR acquisition specifications).

## Chapter 2: Assessment of the Performance of Global Forest Cover Products in South Africa – Establishing the benchmark

### 2.1 Abstract

There is a fervent debate on whether global forests are in the state of growth or loss. Global scale forest cover products have provided a means to measure where forest losses and forest gains are occurring. Most of these global forest cover products, however, have yet to be accurately validated at the local to regional scale especially within the savannah biome. This study aimed to assess the performance of two 2010 global forest cover products, the 30m Landsat derived Vegetation Continuous Field (VCF) and the 25m JAXA ALOS PALSAR Forest/Non-Forest (FNF) global products, against an extensive collection of airborne LiDAR data acquired during 2009 and 2013 across South Africa (SA), with special focus on detecting forest (as per the products' forest definition) in savannahs. The overall strategy was to 'resample' the LiDAR data to match the criteria used to create the VCF and FNF products. It was found that the FNF product grossly under-represented the distribution of forests in savannah environments (20-80% CC ranges), due to the inadequate HV backscatter threshold chosen in its creation. The FNF product also showed a limited ability in detecting closed forest cover class (90-100%) and Natural Forest and Scrub Forest tree structural classes. The Landsat VCF product displayed strong CC underestimation with increasing variability and mean error from CC values of greater than 30%. The moderate accuracies at the 10-20% CC range (and in the Open Woodland tree structural class) suggest that the VCF product could be potentially applicable in low CC environments such as grasslands and sparse savannahs. Limited detection accuracies (~30%) by the VCF, however, were also observed in closed canopy environments (90-100% CC range). Despite the lack of a completely balanced LiDAR acquisition coverage across the forested biomes of SA (most LiDAR acquisitions were biased to the Savannah biome with limited coverage over dense forests); these results give some insight into the inherent flaws of the global products especially over the savannah biome. These results provide the justification for developing new, locally calibrated woody structural products for South Africa.

**Keywords:** *Global forest cover, Landsat VCF, ALOS PALSAR FNF, LiDAR, validation*

## 2.2 Introduction

South African forests and savannahs are crucial ecosystems which provide a plethora of goods and services (food and energy) which benefit both natural and anthropogenic forces (Chidumayo and Gumbo, 2010; Shackleton and Shackleton, 2004; Twine, 2005; Wessels et al., 2011). There is a strong debate on whether these forests and savannahs are in the state of growth or loss. This state of flux is documented particularly in heterogeneous savannah environments, in which the woody resources are harvested for food or selectively logged to satisfy energy securities, by the local populaces, thus creating a perception of forest decline (Pereira et al., 2011; Ryan et al., 2012; Wessels et al., 2013). On the other hand, there is the issue of bush encroachment, which threatens the livestock grazing potential of Southern African rangelands (O'Connor et al., 2014; Ward, 2005; Wigley et al., 2009), or the occurrence of forest regeneration (Chazdon, 2008), either assisted or unassisted by humans, which thus creates a perception of forest growth. The emergence of global scale forest cover products have provided a means to confirm and measure where forest loss and forest gain is occurring at a global scale (Hansen et al., 2013) but whether these products are accurate enough to monitor forests in the Southern African region is left to be investigated.

The development of global scale forest cover products was made possible with the increasing availability of systematic and frequent acquisitions of high resolution remote sensing datasets (which are also ideal for regional monitoring efforts), and the development of integrated large scale processing platforms. Well-known products include: high resolution (30m) global forest cover maps, derived from Landsat 7 ETM+ data (Hansen et al., 2013); a 30m global continuous fields tree cover product, derived from Landsat-based rescaling of MODIS data (Sexton et al., 2013); a 25m global forest/non-forest (FNF) classification product derived from ALOS PALSAR L-band Synthetic Aperture Radar backscatter intensity datasets (Shimada et al., 2014). These products were developed primarily as a means to highlight the extents of forest loss and gain at the global and possibly regional scales which can serve as a proxy for the impact on various ecosystem services such as biodiversity richness, carbon and nutrient storages and fluxes, water supply and exchange and also various climate implications (Hansen et al., 2013; Sexton et al., 2013). Additionally, these global products play a major role in the greater scientific community as they contribute to global initiatives such as REDD+ (Reducing Emissions from Deforestation and forest Degradation) and greatly influence environmental management at the regional governance scale (Hansen et al., 2013; Sexton et al., 2013; Shimada et al., 2014). It is believed that such satellite-based global forest cover and

change products have actually help establish various environmental policy initiatives such as the Kyoto Protocol, REDD+ and the Aichi Biodiversity Targets (Sexton et al., 2015). Most of these global forest products, however, have yet to be accurately validated at the regional scale in South Africa, especially within the savannah biome. The global forest cover products mentioned earlier (Hansen et al., 2013; Sexton et al., 2013; Shimada et al., 2014) have mainly been validated against reference data collected in dense, homogeneous equatorial forested areas of Africa and other countries rather than in heterogeneous savannah and forested types with variable canopy cover and height profiles. As a result, when generalised at the continental scale, validation accuracies of these products are reasonable with validation sites biased to the dense forested areas (e.g. Figure 13a in (Shimada et al., 2014); Figure 2 in (Sexton et al., 2013)). (Kim et al., 2014) also confirmed that Landsat-based VCF global products have a relatively low certainty of forest and non-forest classification in semi-arid environments in which sparse and short trees persist such as the Miombo woodlands. Also, surprisingly, the Global Forest Watch web portal which is based on the Landsat and MODIS VCF products (Hansen et al., 2013; Townshend et al., 2011) (<http://www.globalforestwatch.org/>) does not acknowledge the presence of forest in the South Africa savannah Lowveld and is also limited to targeting trees greater than 5m in height.

What also compounds matters further, is that these products are derived according to different definitions of what constitutes a forest, with different definitions being introduced from various institutes and initiatives (e.g. United Nations Framework Convention on Climate Change, UNFCCC, versus Convention on Biological Diversity, CBD, versus Food and Agriculture Organization of the United Nations, FAO (Schoene et al., 2007)). The Forest Resources Assessment (FRA) of the FAO, for example, defines forest as land spanning more than 0.5 hectares with trees higher than 5m or trees able to reach these thresholds in situ and a canopy cover of more than 10% (FAO, 2015) while the UNFCCC defines forests more flexibly as a minimum area of land of 0.05-1 ha with crown tree cover (or equivalent stocking level) of more than 10 – 30% (UNFCCC, 2001). (Sexton et al., 2015) revealed that such an ambiguity in the definition of forests can potential result in a discrepancy of approximately  $19.3 \times 10^6$  km<sup>2</sup> in forest coverage (i.e. area of classified forest) at the global scale. Such a discrepancy can adversely affect forest area calculations in regions that have overall less dense tree cover such as savannah and shrubland environments (Rocchio, 2015). Since the savannah biome possess total woody fractional cover that can range from dispersed trees in open-grasslands (~5%) to near-closed canopy woodlands (~60%) and more than 80% in riparian zones (Venter et al., 2003), it is expected that forests should be present in such a system regardless of the definition of



forests implemented. Regardless of the definition utilised in these global forest products, a flexible and accurate validation data source is needed for such validation efforts. Light Detection and Ranging (LiDAR) is such a data source and is particularly well suited for woody structural measurements, because of its capacity to capture canopy geometry and structure (McGlinchy et al., 2014; Popescu et al., 2011; Sun et al., 2011). Additionally, in terms of the measurement of fractional tree cover, airborne LiDAR derived metrics have proven to be more accurate than field measured metrics derived from field laser, manual collection and hemi-spherical photography methods (Nickless et al., 2009). This accuracy together with the large geographical coverage managed by airborne LiDAR sensors, thus results in the availability of a large validation source for remote sensing product validation studies.

This study aimed to assess the performance of two 2010 global forest products, the 30m Landsat Vegetation Continuous Field (VCF) and the 25m JAXA ALOS PALSAR Forest/Non-Forest (FNF) global products, against an extensive collection of airborne LiDAR data collected over years 2009 and 2013 in South Africa, which served as the ground truth. These high resolution global forest products have yet to be assessed in South Africa - a country where no regionally derived forests products, from remote sensing data, are currently available despite being a national requirement for reporting on the state of the forests (Willis, 2002). The global 30m tree cover product created by (Sexton et al., 2013), however, was not assessed as the 2010 version of the product was not available. The primary focus of the study would be the assessment of both products for the accurate detection of forests, as per the products' forest definitions, in South African savannahs which are largely under-represented or excluded by such global products. Based on the validation results, and as a secondary objective, product error will be assessed over stratified canopy cover ranges and vegetation structural classes, (e.g. woodlands, natural forests and grasslands (Willis, 2002)), in order to ascertain the performance of these products according to vegetation type. Suggestions, also, were put forward to help improve these global forest products for the structurally variable South African environment. A variety of forest types (i.e. from savannah Lowveld vegetation to closed indigenous forests and plantations) were chosen in South Africa to cover the full expected range of canopy cover values and structure in the validation efforts. This study will ascertain whether these global forest products are applicable to the South African region or whether new regional forest products will needed to be developed.

## 2.3 Study Area

The eastern half of the country of South Africa, between latitudes 22° and 34° south and longitudes 25° and 33° east, where forests are dominant, is under investigation for the task of global forest product validation. Of approximately 120 million hectares in area, South Africa possesses a variety of biomes, topographic landscape features, climate and geological conditions. South Africa consists of nine main biomes (Mucina and Rutherford, 2006), each possessing a characteristic suite of plant and animal species which vary in distribution and according to environmental conditions. Of these biomes, forests are largely present in Savannah, Indian Ocean Coastal Belt (IOCB) and Forest biomes with the Savannah biome covering 35% of the South African land surface (Van Wilgen, 2009). Savannahs are characterised by a mixture of a grassy ground layer and a upper woody layer of plants which are in a constant state of flux depending on rainfall, fire and grazing pressures and occur mostly over the Lowveld and Kalahari regions of the country (Low and Rebelo, 1996). As mentioned earlier, Savannahs are of great importance as the woody layer is harvested by the local populace for energy provision while the grassy ground layer supports cattle ranging and grazing (Low and Rebelo, 1996; Ryan et al., 2012; Ward, 2005; Wessels et al., 2013). This could lead to threats of overharvesting of trees, overgrazing of the grass and subsequent emergence of bush encroachment. Structurally, Savannahs possess total woody fractional cover that can range from dispersed trees in open-grasslands (~5%) to near-closed canopy woodlands (~60%) and more than 80% in riparian zones, a general height range of 1-20m and a total biomass mostly less than 100 tonnes per hectare (t/ha) (Low and Rebelo, 1996; Mathieu et al., 2013; Venter et al., 2003). The Forest biome (including indigenous forest and the IOCB) are less prolific (<1% of SA land surface), occurring in patches rarely greater than 1km<sup>2</sup> in area and commonly occur along the South Coast, the Indian Ocean Coast and the Lowveld escarpment (Low and Rebelo, 1996). Due to high rainfall (>725mm) in such areas, these forests are less affected by fire (except under very dry conditions) (Low and Rebelo, 1996) but are susceptible to illegal logging activities of valuable timber, ring-barking resulting from the illegal extraction of medicinal bark by surrounding communities and the invasion of alien species (e.g. *Pinus* spp.) (Shackleton and Shackleton, 2004). Structurally, the vegetation are usually evergreen and multi-layered with high woody fractional cover (75-100%), high biomass (>100 t/ha) and tall heights (6-20m and greater) (Willis, 2002). Apart from naturally occurring indigenous forests, forests are also represented by commercial plantations with distributions most prevalent on South Africa's eastern escarpment and within the savannah, grassland and IOCB biomes (Scholes and Biggs, 2004). These commercial plantations support alien species cultivars for pole-wood and mulch production, for various commercial goods such as paper and furniture, and also for fruit production. Structurally,

depending on the age and plantation type (orchard versus woodlots), the vegetation is mostly continuous cover with high biomass yields and height measurements (similar to the Forest Biome). Finally, although not typically known to be containing forests, the Thicket Biome possesses evergreen, sclerophyllous vegetation that range from closed shrubland canopies to low forests with no discernible grassy ground layer (Low and Rebelo, 1996). The vegetation supported in this biome can possess high woody cover (75-100%), which can be impenetrable, with generally low height (1-2.5m) and biomass levels (Willis, 2002). One of the biggest threats to this biome is transformation of natural land into agriculture and ranching resulting in land degradation (Hoare et al., 2006). At the South African scale, average temperatures are generally mild but can vary according to location and proximity to the oceans. Annual average precipitation is about 450mm with a high-to-low rainfall gradient existing from east to west which mainly limits forest distribution. The map displayed below, in Figure 2.1, illustrates the study area and shows the LiDAR dataset coverages used for the validation of the global forest products.

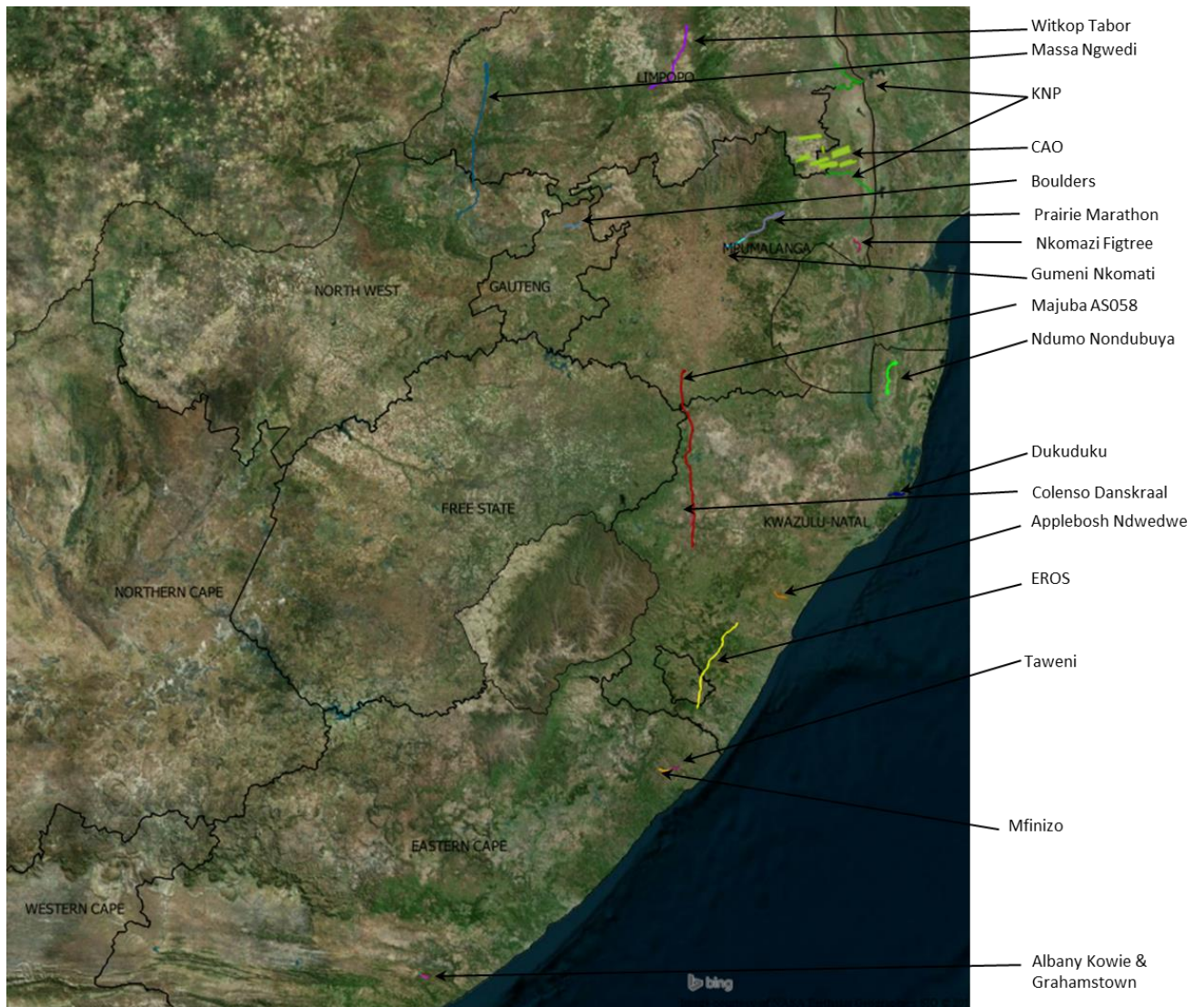


Figure 2.1: Study area with focus on the LiDAR dataset coverages (see table 2.1 for LiDAR specifications)

## 2.4 Materials and Methodology

Two well-known global forest earth observation products, 2010 30m Landsat VCF and 2010 25m ALOS PALSAR FNF (Figure 2.2), were validated at the country level against a geographically extensive dataset of airborne LiDAR. The strategy was fairly simple in that the airborne LiDAR derived data products, i.e. canopy height model (CHM), were processed and ‘degraded’ to fit the criteria and pixel size used to create the Landsat VCF and ALOS PALSAR FNF products. The LiDAR-based forest products were then compared with the global forest products. This takes into account the different definitions used to define and map tree cover in the various global forest products.

### 2.4.1 Global forest products

The 2010 Landsat VCF product was derived from monthly surface reflectance composites, composed from Landsat 7 Enhanced Thematic Mapper Plus (ETM+) imagery (particularly Landsat 7 ETM+ bands 3, 4, 5 and 7) taken throughout the year, derived NDVI, various band-ratios and low- and high-gain temperature bands (Hansen et al., 2013, 2011). The Landsat VCF product was derived from a similar methodology used to create the MODIS VCF product (Townshend et al., 2011). It is composed of three main components; percent tree cover, percent non-tree vegetation and percent bare ground; modelled with a non-parametric bagged decision tree approach (Hansen et al., 2014, 2011, 2003, 2002). The Landsat VCF percent tree cover component defines tree cover as any woody plant with a height greater than or equal to 5 metres (Hansen et al., 2011, 2003). In this study, percent tree cover of the Landsat VCF product was considered to be analogous to the woody canopy cover metric (CC). The ALOS PALSAR FNF was derived from dual-polarised (HH and HV) L-band Fine Beam Dual-polarised (FBD) imagery which were mainly acquired during dry conditions in South Africa (between June and September), according to the dual polarised data type Basic Observation Scenario (BOS) (Shimada et al., 2014). Unlike the continuous tree cover product of the Landsat VCF, the ALOS PALSAR FNF is purely categorical consisting of three classes: forest, non-forest and water. The product was created from country- and/or continent-specific HV backscatter (dB) thresholding for forest separation together with specific HH and HV backscatter (dB) thresholds for non-forested surfaces separation and utilised the FAO definition of a forest, which is all contiguous areas where the cover of woody vegetation is greater than 10% (in this case, within the 25m pixel resolution of the SAR backscatter imagery used for creating the FNF product) (FAO, 2000; Shimada et al., 2014). There was no vegetation height threshold used in the creation of the FNF product.

### 2.4.2 LiDAR validation datasets

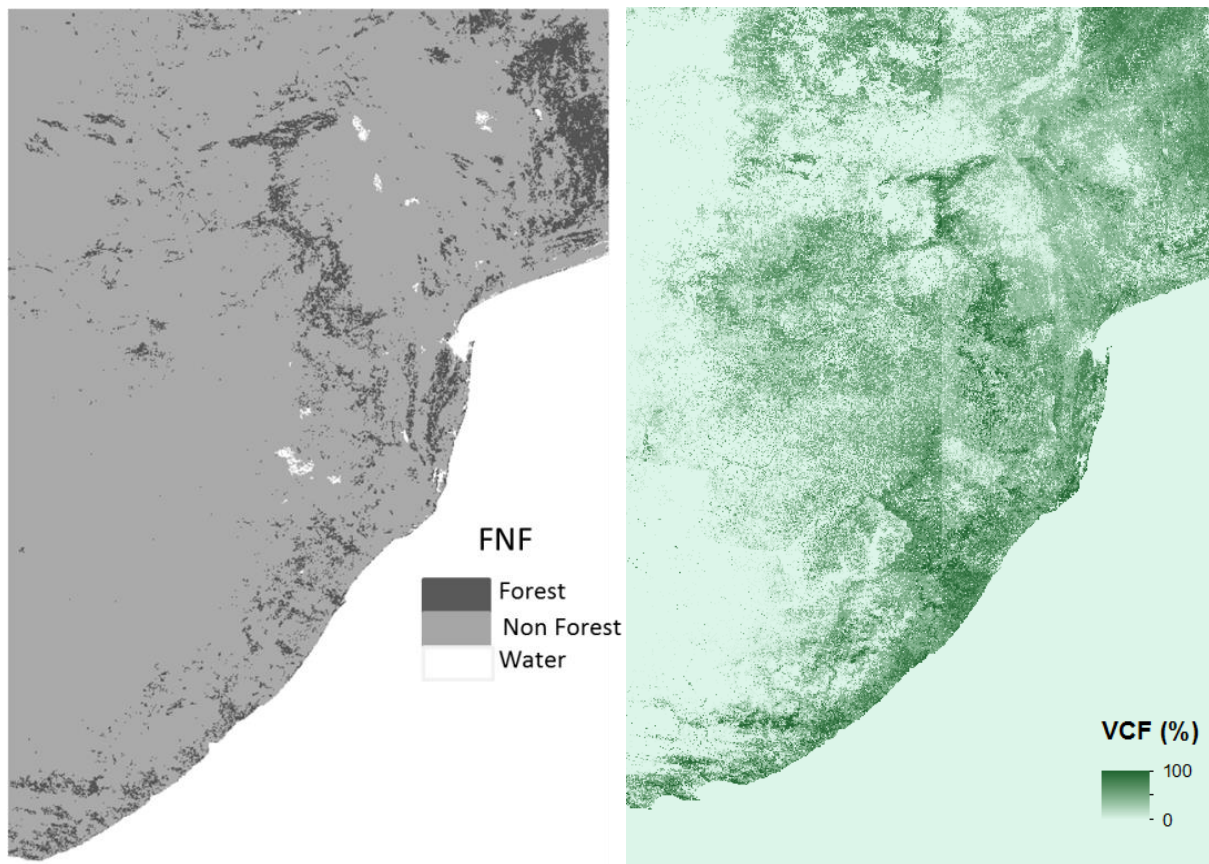
The airborne LiDAR validation datasets (totalling 122 052 hectares) were acquired from a variety of flight campaigns across the eastern part of South Africa between 2009 and 2013. These datasets were made available through scientific and collaborative agreements by the Carnegie Airborne Observatory (CAO), Southern Mapping Company, SANParks Scientific Services, AECOM (UK) and ESKOM. Due to the national scope of the study and the lack of available and extensive airborne LiDAR which matched exactly the global forest product acquisition year, a temporal difference between the global forest and LiDAR datasets was permitted. Most forest types do not change

extensively (indigenous forests are protected) but only gradually (savannahs through mostly selective logging or bush encroachment). Unfortunately, possible error associated with the temporal difference between LiDAR datasets and global forest products, however, could still be incurred during the validation process. The LiDAR datasets used in this study are outlined in Table 2.1.

### **2.4.3 Global Forest Product Pre-processing**

The Landsat VCF was obtained from the Global Land Cover Facility (<http://landcover.usgs.gov/glc/>) while the ALOS PALSAR FNF was publicly available from the Japan Aerospace and Exploration Agency (JAXA) portal ([http://www.eorc.jaxa.jp/ALOS/en/palsar\\_fnf/data/index.htm](http://www.eorc.jaxa.jp/ALOS/en/palsar_fnf/data/index.htm)). Both products were obtained with no further post-processing being conducted. The native projections of the products, geographic WGS84 and Sinusoidal projection for ALOS PALSAR FNF and Landsat VCF respectively, were re-projected to a common geographic projection with a WGS84 datum. It is important to note that preliminary work with the FNF and the ALOS PALSAR global HV mosaic, from which FNF was derived from, indicated that was a pixel misalignment between the ALOS products and the LiDAR datasets. This misalignment was apparent in heterogeneous vegetated areas with varying land use/cover types (forest patches distributed sporadically across sections of grassland and also along urban settlement outskirts). To address this discrepancy the FNF product was converted to an Albers Equal Area projection and shifted by a constant distance of 75m westwards and 50m northwards. After the shift, the FNF products were converted back to a WGS84 geographic projection. Changing to a common projection at the country scale was needed in order to eliminate any errors arising from mismatched projections and potential misalignment between LiDAR-derived extraction grids and the global forest products.





**Figure 2.2: 25m ALOS PALSAR Forest/Non-Forest (left) and 30m Landsat Variable Continuous Field cover (right) products**

#### 2.4.4 Airborne LiDAR Data Pre-processing

Although the LiDAR sensors and settings used varied, such as scan frequency, laser spot spacing and point density (outlined in Table 2.1), a common methodology was applied to all datasets to ensure consistency in the extraction of the canopy height models and associated woody fractional cover. Most of the raw LiDAR point cloud data were processed in TerraSolid LiDAR processing software in which a Digital Elevation Model (DEM) and top-of-canopy surface models (CSM) were created. DEM and CSM were generated at a pixel size varying from 1 to 5 m, depending on the dataset point densities. Canopy Height Models (CHMs), which varied in pixel size from 1 to 5m, were then computed by subtracting the DEM from the CSM. The 2012 CAO datasets were provided already processed by the CAO research team, see details, such as software, in Asner et al. 2012. The differences in LiDAR specifications would not be expected to be influential at the coarser resolutions of the global forest products. To match the criteria in which the ALOS PALSAR FNF product was created, the LiDAR CHM data were processed to generate forest versus non-forest products

considering a canopy cover with a threshold of greater than or equal to 10%, after the extraction process, which will be elaborated upon in the next section (2.4.5). To match the conditions in which the Landsat VCF product was created, the LiDAR CHM data was subjected to a tree height threshold of greater than 5m. All pixels which did not meet these specified thresholds were masked out and excluded from the rest of methodological workflow. The LiDAR datasets were kept in the pre-processed spatial resolution (ranging from 1m to 5m) as the datasets were indirectly resampled to match the global forest products during the data extraction process with the use of extraction grids (to be elaborated upon further in the next section).



**Table 2.1: Summary of LiDAR datasets used, the year acquired, sensor specifications, coverage, environmental description and provider information**

LiDAR dataset	Year	Scan frequency	Laser spot spacing (along/across track)	Point density	Final CHM Resolution	Area coverage	Province	Description	Provider(s)
CAO	2012	50 kHz	0.56m	6.4 points per m <sup>2</sup>	1m	63 000 ha	Mpumalanga	Majority savannah with communal rangelands	Carnegie Airborne Observatory
KNP	2012	70 kHz	0.5m	10 points per m <sup>2</sup>	1m	17 000 ha	Mpumalanga	Savannah riparian vegetation	AECOM (UK)
EROS	2013	150 kHz (max)	0.30m	12 points per m <sup>2</sup>	1m	6 700 ha	Kwa-Zulu Natal	Mixed consisting of azonal vegetation, forest plantations plus savannahs and grassland	CAD Mapping/ESKOM
Dukuduku	2013	300 kHz (max)	0.30m	10 points per m <sup>2</sup>	1m	2 100 ha	Kwa-Zulu Natal	Majority indigenous coastal forest	PROMAP/ESKOM
Boulders	2010	100 kHz	0.67m	2.26 points per m <sup>2</sup>	1m	900 ha	Gauteng	Highveld bushveld with urban cover	AOC/ESKOM
Gumeni Nkomati	2010	100 kHz	1.10m	0.83 points per m <sup>2</sup>	2m	742 ha	Mpumalanga	Mostly savannah with small tree patches	AOC/ESKOM
Nkomazi Figtree	2010	100 kHz	0.81m	1.53 points per m <sup>2</sup>	1m	659 ha	Mpumalanga	Lowveld shrubs with majority agriculture	AOC/ESKOM
Majuba AS058	2010	100 kHz	1.19m	0.71 points per m <sup>2</sup>	5m	7 085 ha	Kwa-Zulu Natal	Thornveld, shrub and grassland with small dense tree patches	Fugro/ESKOM
Albany Kowie	2010	100 kHz	1.56m	0.41 points per m <sup>2</sup>	2m	262 ha	Eastern Cape	Majority grassland and thicket	Southern Mapping Company/ESKOM
Applebosh Ndwedwe	2009	70 kHz	0.98m	1.05 points per m <sup>2</sup>	2m	650 ha	Kwa-Zulu Natal	Small patches of plantations with dense veld and sugar cane cropland	Southern Mapping Company/ESKOM
Colenso Danskraal	2011	100 kHz	1.27m	0.62 points per m <sup>2</sup>	5m	1 675 ha	Kwa-Zulu Natal	Majority thornveld	Southern Mapping Company/ESKOM
Grahamstown	2011	100 kHz	1.08m	0.86 points per m <sup>2</sup>	2m	400 ha	Eastern Cape	Majority grassland and thicket	Southern Mapping Company/ESKOM
Massa Ngwedi	2010	100 kHz	0.95m	1.11 points per m <sup>2</sup>	1m	6 981 ha	Limpopo	Combination of shrubby rangeland and savannah	Southern Mapping Company/ESKOM
Mfinizo	2010	100 kHz	1.41m	0.5 points per m <sup>2</sup>	2m	278 ha	Eastern Cape	Grassland with dense patches of bushveld	Southern Mapping Company/ESKOM
Ndumo Nondubuya	2011	70 kHz	0.87m	1.31 points per m <sup>2</sup>	2m	3 175 ha	Kwa-Zulu Natal	Bushveld and thicket with evergreen tree patch	Southern Mapping Company/ESKOM
Prairie Marathon	2009	70 kHz	1.20m	0.69 points per m <sup>2</sup>	2m	4 573 ha	Mpumalanga	Mixed consisting of dense bushveld, grassland and patches of plantations/orchards	Southern Mapping Company/ESKOM
Taweni	2010	100 kHz	1.34m	0.55 points per m <sup>2</sup>	2m	282 ha	Eastern Cape	Grassland with dense patches of bushveld	Southern Mapping Company/ESKOM
Witkop Tabor	2009	70 kHz	1.53m	0.43 points per m <sup>2</sup>	2m	5 590 ha	Limpopo	Mixed with agricultural fields and rangelands	Southern Mapping Company/ESKOM

### 2.4.5 Data extraction process

25 by 25m and 30 by 30m grid cells were aligned to the pixels of the ALOS PALSAR FNF and Landsat VCF products respectively. These grids were then clipped to the extent of the available LiDAR CHM datasets. Thus, these grids were used to extract the global forest product and the corresponding LiDAR data values for each cell. Any cells within the grid which fell within or overlapped with the LiDAR coverage edges, urban and informal settlements/built-up areas, major water bodies and other artefacts present within the LiDAR data (e.g. power lines) were manually identified from a Google Earth image backdrop and excluded from the validation process to minimise error caused by ‘mixed’ pixels. LiDAR woody canopy cover (CC), in percentage, was derived per cell, with the use of Equation 2.1 below, and considered woody vegetation above a height threshold of 0.5m (mainly for the FNF product rather than the VCF product) to avoid the influence of grass in the CC calculations.

$$\text{LiDAR CC (\%)} = \frac{\text{Number of Woody Vegetated (above 0.5m in height) LiDAR pixels in a grid cell}}{\text{Total number of LiDAR pixels in a grid cell}} \times 100$$

Equation 2.1

The total number of LiDAR pixels in a grid cell differs depending on the spatial resolution of the LiDAR CHM and the 25 by 25m (e.g. 625 1m LiDAR pixels) or 30 by 30m grid sizes (e.g. 900 1m LiDAR pixels) used for matching the corresponding LiDAR derived CC to the respective ALOS PALSAR FNF and Landsat VCF products. Finally for the ALOS PALSAR FNF product comparison, the CC forest threshold of  $\geq 10\%$  was applied to the LiDAR derived CC values, from the 25m by 25m grids, to create the LiDAR derived FNF values, i.e. a Forest ( $\text{CC} \geq 10\%$ ) and Non-Forest ( $\text{CC} < 10\%$ ) reclassification.

### 2.4.6 Global Product Accuracy Assessment

The data extracted from the individual LiDAR datasets and corresponding global product coverages have been combined for an overall assessment. Due to the nature of the different global forest products different validation techniques have been implemented to best assess these products’ accuracy using LiDAR-derived CC. For the categorical ALOS PALSAR FNF product, summarised confusion matrix statistics (particularly producer accuracies) together with overall accuracies, forest and non-forest accuracies have been derived. The continuous Landsat VCF CC product was correlated against LiDAR derived CC, from which the coefficient of determination ( $R^2$ ), root mean

square error (RMSE), bias and standard error of prediction (SEP) was derived. For a quantitative measure of the extent of overestimation and underestimation in the Landsat VCF product across the observed CC range, the LiDAR CC – Landsat VCF CC difference values were arranged into box plots over the 10% incremental CC classes ranging from 0-100%.

To evaluate the performance of both products considering vegetation types, the validation dataset was classified according to woody cover (CC) and structural classes. For the FNF product, the LiDAR CC data was reclassified or stratified into 10% incremental classes from the 0-100% range (i.e. 10 classes in total e.g. 0-10%, 10-20%, 20-30% etc.). The total number of correctly classified data records within each CC incremental class was divided against the total number of records in the particular classes and multiplied by 100 to ascertain the percentage accuracy of the FNF product within the different CC class increments. For the vegetation structural assessment of the FNF product, LiDAR CC and vegetation height record information was categorized according to structural classes proposed by (Willis, 2002) for categorizing forest structure in southern Africa, including dense tall natural forests, a range of open woodlands, and short thickets. This classification is presented in Figure 2.3. As with the CC incremental class assessment, the total number of correctly classified data records within each structural class (according to the LiDAR CC and height ranges) was divided against the total number of records in the particular classes, and multiplied by 100 to ascertain the percentage accuracy of the FNF product within the different vegetation structural classes. Due to the continuous nature of the Landsat VCF CC and LiDAR CC values, the LiDAR and VCF CC range were both reclassified into the 10% CC incremental classes for assessment, which followed the same methodology as described with the FNF product above. Since the VCF measures vegetation greater than or equal to 5m in height, for coherence this threshold was also applied to the vegetation structure class (Figure 2.3) thus resulting in fewer classes being represented than in the FNF product.

Average height in metres	>20m	Grassland/ herbland	Wooded Grassland	High (seldom occurs)												Natural Forest					
	6 -20m			Open (parkland, grassy woodland)				Tall woodland (Miombo)													
	2.5 – 6m			Low woodland (Bushveld)				Thicket				(scrub forest)									
	1 – 2.5m			Open Bushland								Bushland									
	1.			Grass-land	Open Shrubland				Shrubland								Closed Shrubland				
		5	10		15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95
Projected woody plant canopy cover (%)																					

**Figure 2.3: Classification of vegetation types according to structure (canopy cover and height) (Willis, 2002). The LiDAR canopy cover and CHM products were used to reproduce this classification scheme for the extracted data.**

As additional support to the product assessments via CC and vegetation structural classifications, product comparison maps (LiDAR vs FNF and LiDAR vs Landsat VCF) were also created to ascertain the visual distributions of error, i.e. extents of underestimation and overestimation, throughout the various landscape types. For a more regional assessment of the products, an ALOS PALSAR L-band derived CC map ( $R^2=0.81$ ;  $RMSE=9.89\%$ ) was also utilised for comparison purposes (SAR vs FNF and SAR vs Landsat VCF). This SAR product was derived according to the methodologies carried out in Chapters 3 (section 3.4.5) and 4 (section 4.4.6). The lower Kruger National Park region, in the Savannah Lowveld, was chosen as the area of focus.

## 2.5 Results

This section was divided into two sub-sections: the FNF and the VCF product validation results. In order to interpret these results, the LiDAR data is considered as the ground truth.

### 2.5.1 ALOS PALSAR FNF validation results

**Table 2.2: Summarised FNF validation results across stratified LiDAR-derived CC ranges**

CC Class	Classified from LiDAR (ground-truth)	Correctly detected by FNF as Forest (F) or Non-Forest (NF)	Grand Accuracy of FNF %
0-10% (NF)	738300	732321	99.19
10-20% (F)	278774	1570	0.56
20-30% (F)	276011	2398	0.87
30-40% (F)	231228	3339	1.44
40-50% (F)	225997	4454	1.97
50-60% (F)	176511	4594	2.60
60-70% (F)	153686	5500	3.58
70-80% (F)	100774	5282	5.24
80-90% (F)	92683	8354	9.01
90-100% (F)	164437	48622	29.57
<b>Total Forest (F)</b>	1700101	84113	4.95
<b>Total Non-Forest (NF)</b>	738300	732321	99.19
<b>Grand Total</b>	<b>2438401</b>	<b>816434</b>	<b>33.48</b>

From the summarised confusion matrix results (table 2.2), it was evident that the FNF product detected very well Non-Forest areas (99% for CC<10%) but performed poorly by detecting only 5% of actual forests (CC>10% according to FAO definition) across the LiDAR datasets. When analysing the results at stratified CC levels (table 2.2) and the detection of forest, it was clear that the FNF performed best at the 90-100% CC range, but still only yielded a marginal 30% forest detection accuracy. The product performed especially poorly throughout the 10-90% CC range with a less than 5% forest detection rate being obtained between the 10-70% CC ranges. The forest detection rate tended to increase with the CC values between the 10-100% CC ranges. In general, the high accuracy of the Non-Forest class and the large number of Non-Forest observations in the dataset (738300) resulted in pushing up the overall classification accuracy of the FNF product (33.48%). The results from table 2.3 and Figure 2.4, below, indicate the FNF product detection accuracies in various LiDAR-derived vegetation structural classes.

**Table 2.3: Summarised FNF validation results across various LiDAR-derived vegetation structural classes**

Structure Class	Classified from LiDAR (ground-truth)	Correctly detected by FNF as Forest (F) or Non-Forest (NF)	Grand Accuracy of FNF%
Bushland (F)	233019	11659	5.00
Closed Shrubland (F)	10725	486	4.53
Grassland (NF)	75501	74823	99.10
Grassland/herbland (NF)	579303	574063	99.10
High (F)	23	0	0.00
Natural Forest (F)	41031	11208	27.32
Open Bushland (F)	306811	2188	0.71
Open Shrubland (F)	253849	4502	1.77
Open Woodland (F)	225353	617	0.27
Scrub Forest (F)	121444	31562	25.99
Shrubland (F)	90199	1802	2.00
Thicket (F)	83920	13720	16.35
Wooded Grassland (NF)	83496	83435	99.93
Woodland (F)	333727	6369	1.91
<b>Total Forest (F)</b>	<b>1700101</b>	<b>84113</b>	<b>4.95</b>
<b>Total Non-Forest (NF)</b>	<b>738300</b>	<b>732321</b>	<b>99.19</b>
<b>Grand Total</b>	<b>2438401</b>	<b>816434</b>	<b>33.48</b>

Average height in metres	Canopy Cover (%)	Grassland / herbland (99.10%)		Wooded Grassland (99.93%)												Grassland (99.10%)												
		0	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100						
>20m		High (0%)																										
6-20m		Open Woodland (0.27%)												Woodland (1.91%)										Natural Forest (27.32%)				
2.5-6m		Open Bushland (0.71%)												Bushland (5%)										Scrub Forest (25.99%)				
1-2.5m		Open Shrubland (1.77%)												Shrubland (2%)										Closed Shrubland (4.53%)				
1m																												
		Projected woody plant canopy cover (%)																										

**Figure 2.4: Summarised FNF validation results across various LiDAR-derived vegetation structural classes as outlined by (Willis, 2002) (% values refer to the FNF detection accuracy of that vegetation structural class where red cells = accuracies >90%, orange cells = accuracies between 15-30%, yellow cells = accuracies ≤5%)**

According to the vegetation structural class results, table 2.3 and Figure 2.4, the FNF product achieved the highest detection accuracies (~99%) for all classes with CC below 10%, grassland, grassland/herbland and wooded grassland classes which were the Non-Forested classes according to the product's definition. Scrub Forest and Natural Forest structural classes, i.e. forested classes with medium to high vegetation height and high CC, obtained accuracies of 26% and 27% respectively while thickets, i.e. forested classes with low vegetation height and high CC, obtained accuracies of ~16%. The Closed Shrubland class, which is a forested class with high CC but very low vegetation height (<1m), yielded a very low detection accuracy of 4.5%. Other classes, which were structural classes within the 10-80% CC ranges, obtained very low detection accuracies of 5% and less, whatever the tree height profile. The High tree structural class yielded 0% detection accuracy but this structural class rarely occurs. Overall, the structural classification shows that forested class detection decreased with CC, and was possibly more affected by cover than height.

The spatial patterns of FNF product, and the corresponding ground truth product (i.e. LiDAR), are compared at the local and regional scale in figure 2.5.

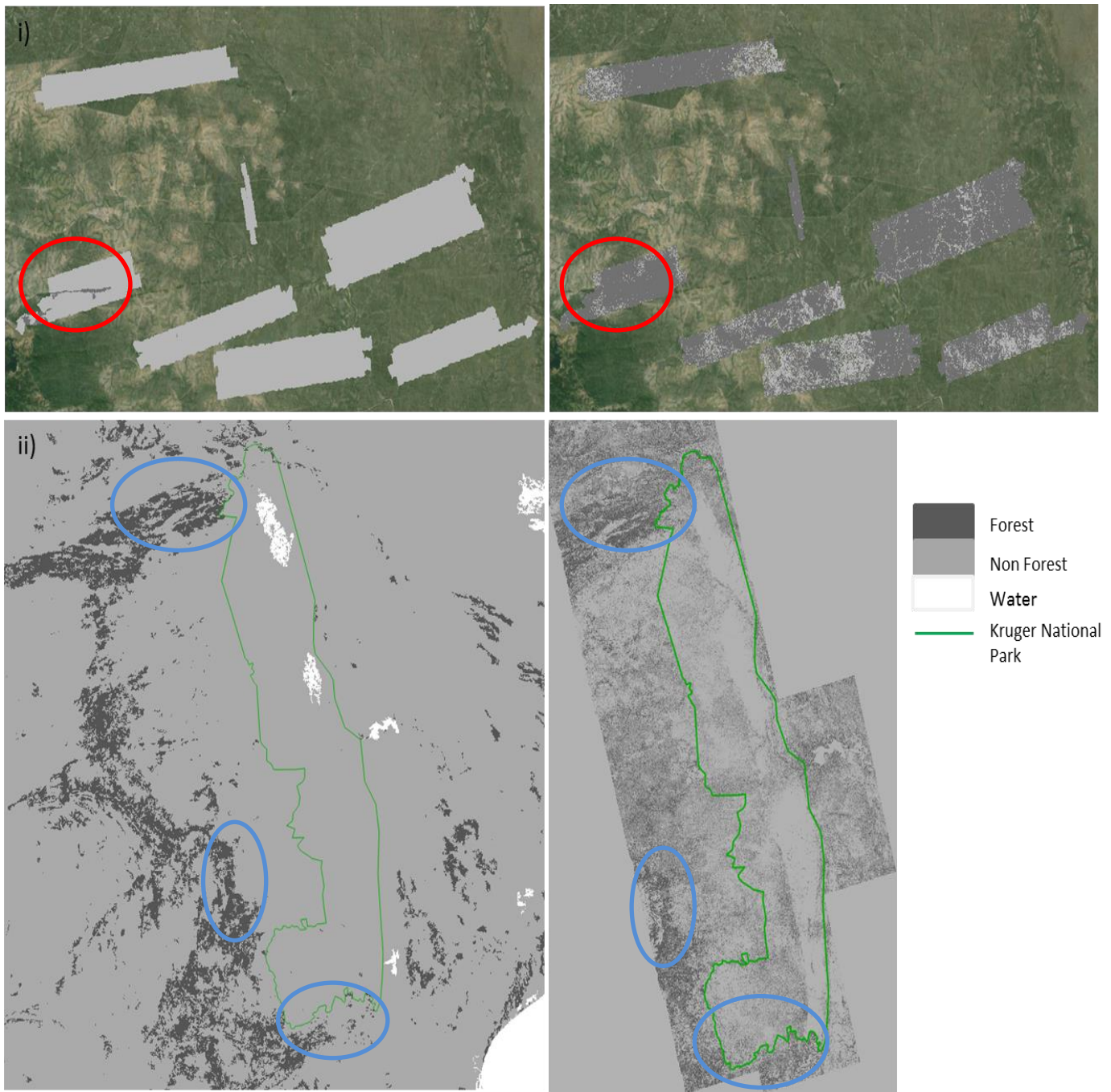


Figure 2.5: i) ALOS PALSAR FNF (left) versus LiDAR derived FNF (right) across the CAO LiDAR dataset; ii) ALOS PALSAR FNF (left) and L-band ALOS PALSAR FBD derived FNF (right), using LiDAR training, ( $R^2=0.81$ ;  $RMSE=9.89\%$  - this product will be detailed in chapters 3 and 4) across the entire Kruger National Park extent [the red and blue encircled areas indicates areas of interest for discussion]

At a larger scale (Figure 2.5ii), the FNF product showed a sensibly better agreement with actual forest class within the main densely forested zones along the South African escarpment (see blue encircled areas). Patches of zonal and intra-zonal indigenous forests and commercial plantations are present here (Mucina and Rutherford, 2006) but not to the extent represented by the FNF. A



potentially higher backscatter due to the topography of the escarpment may have boosted the detection of forests in such features. Outside the escarpment, however, most veld areas (i.e. areas within and along the Kruger National Park boundary) were classified as non-forest by the FNF. According to (Mucina and Rutherford, 2006), these veld areas consisted of mopane, sour bushveld, granite lowveld and sandy bushveld vegetation types, typical of the savannah biome, which were known to possess cover greater than 10%. These trends corroborate the previous results in which the FNF product lack the ability to detect the 10-80% CC range and the vegetation structural classes found in this range (e.g. Bushveld and Woodland classes etc.) while showing some detection potential in the high CC (80-100%) and dense structural classes (e.g. Scrub Forests and Natural Forests). The FNF product however, did also yield erroneous patches of water within the Kruger National Park extents as these areas were confused with areas of basaltic open grasslands. Locally as shown with the LiDAR tracks (Figure 2.5i), the FNF product displayed very little of the forest class compared to the amount of forest actually present in the CAO LiDAR maps which falls squarely in savannah biome. The red encircled area in the FNF product, a dense forested ridge, only showed limited evidence of forest which coincided, to some degree, with the LiDAR product (between 80-100% CC range along the ridge, according to the LiDAR).

## 2.5.2 Landsat VCF validation results

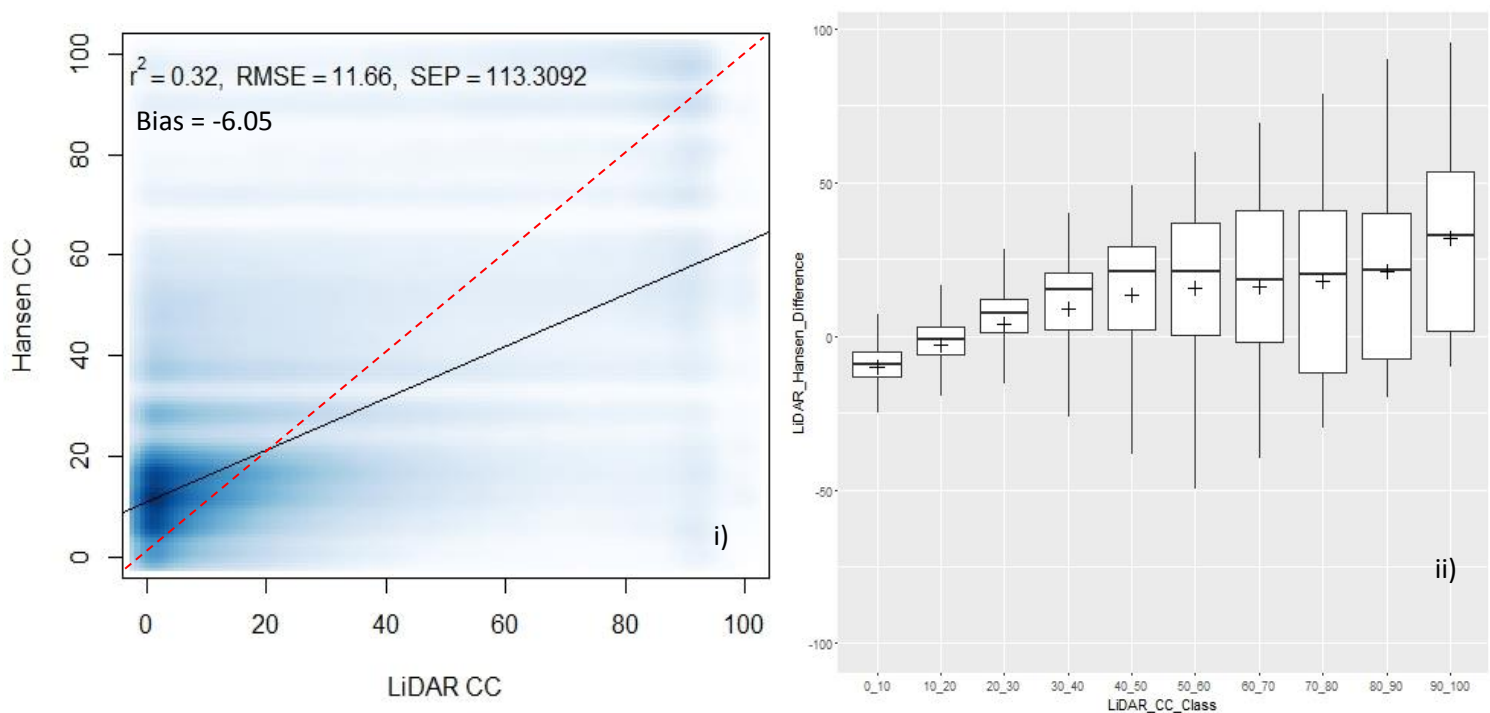


Figure 2.6: (i) Density scatterplot of LiDAR derived CC versus Landsat VCF CC across the complete extracted dataset [the dotted red line represents the 1:1 line while the solid black line represents the data trend line]. (ii) Landsat VCF product

error (i.e. LiDAR CC – VCF CC) over a range of CC intervals [negative values indicate CC overestimation while positive values indicated CC underestimation by the Landsat VCF product; centre cross = mean value; box = standard error and whiskers = standard deviation]

Figure 2.6i) and ii) illustrated the correlation between LiDAR derived CC and VCF CC as well as the level of over- and underestimation of the VCF product across the complete CC range. The density scatter plot of figure 2.6i) indicated a generally poor relationship with an  $R^2$  of 0.32, an RMSE of ~12 % and an SEP greater than 100%. The general trend was also hard to distinguish due to the large discrepancies between the corresponding LiDAR and VCF CC values especially at higher CC ranges. The CC difference box plot of figure 2.6ii), together with figure 2.6i), illustrated that the VCF product overestimated CC values (0-10%) slightly between the 0-20% range with general underestimation (10-40%) occurring past this point to higher CC values. The standard deviation and standard error values increased greatly towards higher CC values (30-100%).  $R^2$ , RMSE and SEP statistics were initially ascertained across the individual stratified CC ranges and vegetation structural classes (see Appendix 2A for example) but the results were poor (i.e. low  $R^2$  with high RMSE and SEP values) with no discernible patterns emerging. Thus a classification approach appeared to be more useful for detailed analyses of the VCF product at various CC and vegetation structural classes.

**Table 2.4: Summarised VCF validation results across stratified LiDAR-derived CC ranges**

CC Class	Classified from LiDAR (ground truth)	Correctly classified by VCF	Grand Accuracy of VCF %
0-10%	541181	188283	34.79
10-20%	136964	76674	55.98
20-30%	50130	10098	20.14
30-40%	22577	1454	6.44
40-50%	10595	988	9.33
50-60%	6517	905	13.89
60-70%	4508	168	3.73
70-80%	3693	219	5.93
80-90%	6945	353	5.08
90-100%	5988	1793	29.94
<b>Grand Total</b>	<b>789098</b>	<b>280935</b>	<b>35.60</b>

**Table 2.5: Complete VCF CC versus LiDAR CC confusion matrix across fixed CC ranges**

VCF CC	LiDAR CC										Grand Total
	0_10	10_20	20_30	30_40	40_50	50_60	60_70	70_80	80_90	90_100	
0_10	188283	29172	7675	2591	886	415	204	138	275	429	230068
10_20	288431	76674	25247	9247	3423	1531	648	361	447	486	406495
20_30	43850	19842	10098	5222	2458	1275	699	406	431	382	84663
30_40	7112	3777	2260	1454	889	581	385	227	274	198	17157
40_50	7649	3761	2191	1604	988	695	513	340	545	431	18717

50_60	4338	2643	1643	1443	957	905	774	707	1264	1093	15767
60_70	473	320	271	243	198	174	168	173	342	289	2651
70_80	368	332	292	289	250	250	220	219	562	439	3221
80_90	145	109	119	126	117	141	156	178	353	448	1892
90_100	532	334	334	358	429	550	741	944	2452	1793	8467
Grand Total	541181	136964	50130	22577	10595	6517	4508	3693	6945	5988	789098
Producer's Acc.	34.79	55.98	20.14	6.44	9.33	13.89	3.73	5.93	5.08	29.94	

The stratified CC results of the VCF, table 2.4, showed trends such as low detection accuracy at the 0-10% LiDAR CC range (~35%). Between this 0-10% LiDAR CC range, according to the confusion matrix (table 2.5), the bulk of the error of the VCF (~60% of the error) was evident between the 10-20% and 20-30% VCF CC classes but classes up to 50-60% class also contributed to this error. The VCF product, also, yielded moderate to low accuracies in the 10-30% CC range (56% and 20.14% for the 10-20% and 20-30% LiDAR CC classes respectively). For the VCF product, detection accuracies remained fairly low (<10%) across the 30-90% LiDAR CC range. Across this LiDAR CC range, the bulk of the VCF error (according to table 2.5) fell in much lower VCF CC classes (e.g. in the 10-20 and 20-30% VCF CC classes across 40-60% LiDAR CC range) which confirmed the general underestimation of the VCF product between 30-90% LiDAR CC range. The VCF product obtained an accuracy of 30% in the detection of vegetation with a 90-100% LiDAR CC range. Finally, the overall classification accuracy obtained by the VCF product was approximately 36%.

**Table 2.6: Summarised VCF validation results across various LiDAR-derived vegetation structural classes**

Structure Class	Classified from LiDAR (ground truth)	Correctly classified by VCF	Grand Accuracy of VCF %
Grassland/herbland	386280	146189	37.85
Wooded Grassland	154901	42094	27.17
Open Woodland	208877	88062	42.16
High	1180	181	15.34
Woodland	24927	2263	9.08
Scrub Forest	72	0	0.00
Natural Forest	12861	2146	16.69
<b>Total</b>	<b>789098</b>	<b>280935</b>	<b>35.60</b>

Average height in metres	>20m	Grassland / hermland (37.85%)	Wooded Grassland (27.17%)	High (15.34%)												Natural Forest (16.69%)					
	6-20m			(42.16%) Open Woodland						(9.08%) Woodland						Scrub Forest (0%)					
	5m	-----																			
	2.5-6m	Grassland	Open Bushland						Bushland						Thicket						
	1-2.5m		Open Shrubland						Shrubland						Closed Shrubland						
	1m		Open Shrubland						Shrubland						Closed Shrubland						
		0	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95
Projected woody plant canopy cover (%)																					

**Figure 2.7: Summarised VCF validation results across various LiDAR-derived vegetation structural classes as outlined by (Willis, 2002) (% values refer to the VCF detection accuracy of that vegetation structural class where red cells = accuracies >20%, orange cells = accuracies between 10-20%, yellow cells = accuracies <10%). The red 5m height line indicates the limit of VCF product in which all classes coloured grey (below 5m height) was excluded.**

Table 2.6 and Figure 2.7 illustrated the detection accuracy results across the various vegetation structural classes which were greater than or equal to 5m in vegetation height, as specified by the steps used to create the VCF. As with the stratified CC range results, the VCF yielded 38% and 27% accuracies for detecting grassland/hermland and wooded grassland vegetation structural classes which possessed low CC (<10%) and medium to high height ranges. The VCF product also yielded a moderate detection accuracy of 42% for the Open Woodland class (CC ranging from 10-40% and with a medium to high height range). On the high CC and height range, the VCF yielded 15% and 17% detection accuracy for the High and Natural Forests respectively while 0% accuracy was observed for the Scrub Forest class which mostly fell below the 5m height mark resulting in very few samples.

Local and regional scale Landsat VCF products (Figures 2.8i and 2.8ii) and their assessment against more accurate map products (i.e. LiDAR and SAR based CC maps) were introduced to understand the geographical distribution of this product error.



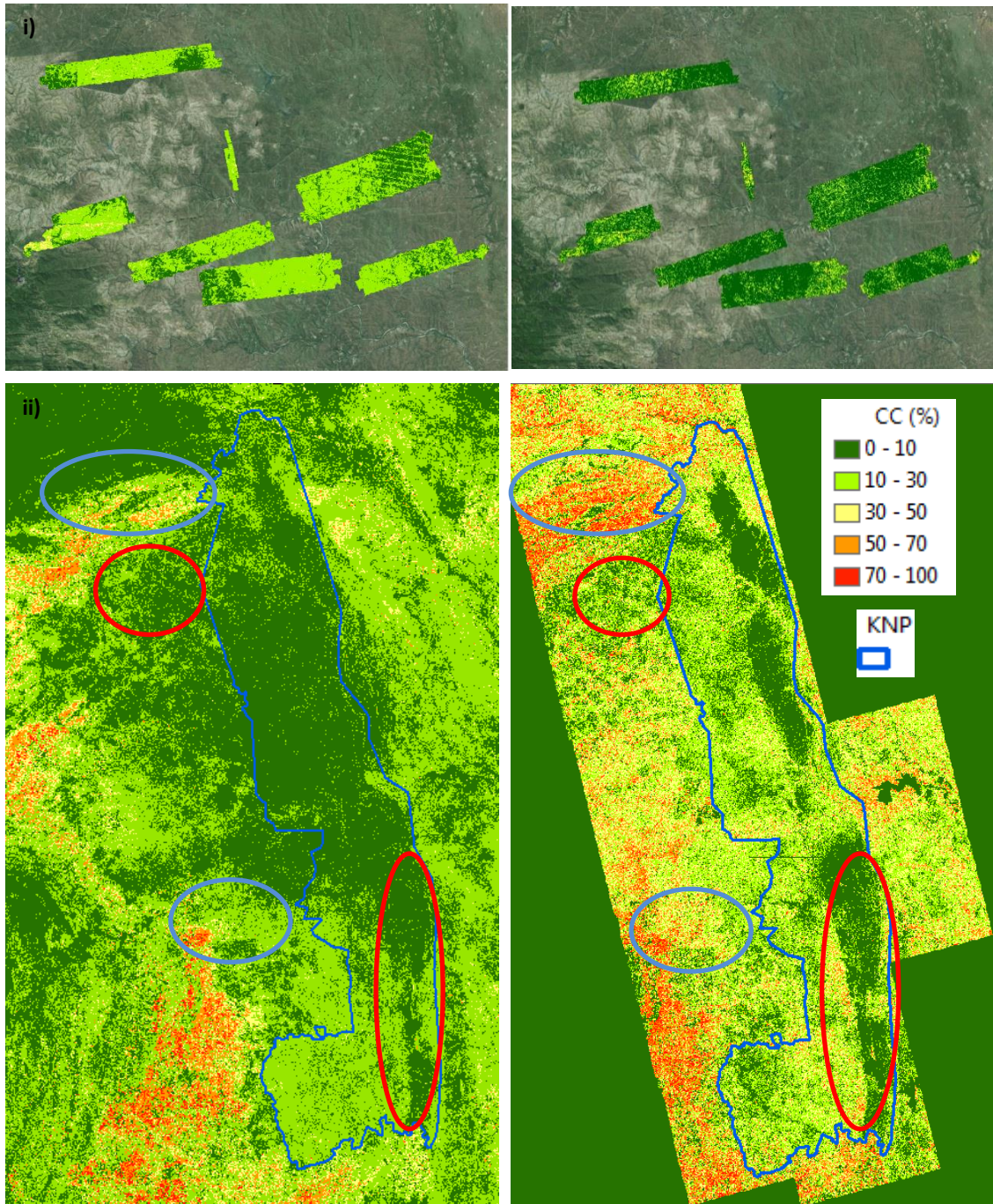


Figure 2.8: i) Landsat VCF CC (left) versus LiDAR derived CC (right) across the CAO LiDAR dataset; ii) Landsat VCF CC (left) and L-band ALOS PALSAR FBD derived CC (right), using LiDAR training, ( $R^2=0.81$ ;  $RMSE=9.89\%$ ) across the entire Kruger National Park extent [the red and blue encircled areas indicates areas of interest for discussion]

At the local scale (Figure 2.8i), a trend of VCF CC overestimation is clearly shown when compared to the observed LiDAR derived CC. The observed LiDAR derived CC product was created by incorporating the 5m height threshold used to create the Landsat VCF product. This corroborates the trends displayed in figures 2.6i) and ii) at the 0-20% CC range. The VCF product also lacks the

spatial detail of the low CC classes in comparison to the LiDAR CC product which has also been degraded to match the VCF conditions (vegetation height threshold of  $\geq 5\text{m}$ , 30m pixel size). Although the class difference of approximately 10% is noticeable between VCF CC and LiDAR derived CC, the VCF product does illustrate patterns of CC variability, though limited, across the landscape, as indicated by the LiDAR. Regarding Figure 2.8ii, it is important to note the SAR derived CC product was created without implementing the 5m height threshold which was used to create the VCF due to poor modelling results of the SAR datasets when modelled using LiDAR cal/val datasets with the 5m height threshold applied. Despite this discrepancy, the main trends observed between the products were fairly comparable. At the regional scale (Figure 2.8ii), VCF results illustrate some of the major patterns of high CC classes ( $> 70\%$ ) being represented along the South African escarpment in the modelled SAR CC product (see blue encircled areas). Additionally, the VCF product represents well some patches of low CC classes ( $< 30\%$ ), which correspond with the modelled SAR CC product, both along the grasslands of Kruger National Park and within rangeland patches outside the Kruger boundary (see red encircled areas). The southern portion of the regional VCF product resembled more of the patterns displayed in the corresponding portion of SAR derived CC map. This could be related to the wetter and greener vegetation conditions which were readily captured by the Landsat imagery while the drier conditions in the north led to a poorer representation of the vegetation signal. At the extreme northern tip of the Kruger National Park (both above and to the right of the highest positioned blue circle), however, there was a large CC difference between the VCF and the SAR product (10% CC in the VCF compared to between 50-70% CC in the SAR). According to Google Earth, the area to the right of the blue circle is a vegetated escarpment feature with a nearby riparian zone emerging while the area above the blue circle is dominated by mopane. A combination of topographic effects and the presence of dense vegetation may have led to higher CC classes while the phenological differences of the underlining Landsat imagery, used to create the VCF product, may have contributed to the low CC values in that area.

## 2.6 Discussion

This study sought to assess two global forest cover products, the 25m ALOS PALSAR FNF and 30m Landsat VCF, using environmentally diverse LiDAR dataset coverages across the forested regions of South Africa. The main focus was to quantify how well these products detect the presence of forests, according to the individual products' forest definitions, mostly within the savannah biome and other forest types present in South Africa. The products were also assessed across stratified CC

ranges and across particular vegetation structural classes (Willis, 2002) to ascertain if performance is consistent across vegetation types.

The FNF product only detected 5% of actual forests across the LiDAR datasets with majority contributions to the accuracy, though low, falling in the 90-100% CC range and Natural Forest and Scrub Forest structural class types. The fact that low lying vegetated areas with 90-100% canopy cover values (e.g. the Closed Shrublands class) were not detectable by the FNF product indicated that the product was not sensitive enough to classify vegetated or 'forested' areas lower than 1m in height. This poor result of the FNF, however, could be compounded by the reduced effectiveness of LiDAR sensors to capture vegetation less than 1m (Wessels et al., 2011). This can also be attributed to the large wavelength of the L-band SAR sensor (~23cm) which may have passed through these small vegetative elements such as leaves and stems (Naidoo et al., 2015; Vollrath, 2010). The FNF product yielded the poorest detection accuracy of 5% and less for the 10-80% CC range; together with the various associated woodland, shrubland and bushland structural classes; which illustrated that the forest within the savannah biome is not detected. This suggested that the FNF product largely under-represents the distribution of forests especially in savannah environments, which possess an average CC of 35% (Venter et al., 2003). Since savannahs cover roughly half of the African continent and occupy one fifth of the global land surface (Sankaran et al., 2005; Scholes and Walker, 1993; Venter et al., 2003), this result is not favourable especially for the applications of carbon assessment and change detection studies. The FNF product obtained 99% accuracy in detecting non-forested areas with the highest accuracy being observed in the 0-10% CC range and within grassland structural types. Though it has been considered that the contrasting backscatter responses between forested and non-forested surfaces could have contributed to this high detection accuracy of non-forest areas, it was the HV threshold used in the FNF product creation (Shimada et al., 2014), which was too high, that contributed mostly to this observation. An in-depth assessment of this threshold, involving Figure 2.9, will be conducted in the following paragraph. The poor ability of the FNF product to detect forests in savannahs, and the underperformance of the FNF product in the 80-100% CC ranges and in the Natural Forest and Scrub Forest classes, was also the result of the selection of the FNF threshold used to define forest versus non-forest in the African continent (Figure 2.9). Figure 2.9 correlated the LiDAR CC with the ALOS PALSAR HV backscatter (i.e. the global mosaic data used to create the FNF product), extracted over the complete LiDAR dataset coverage.

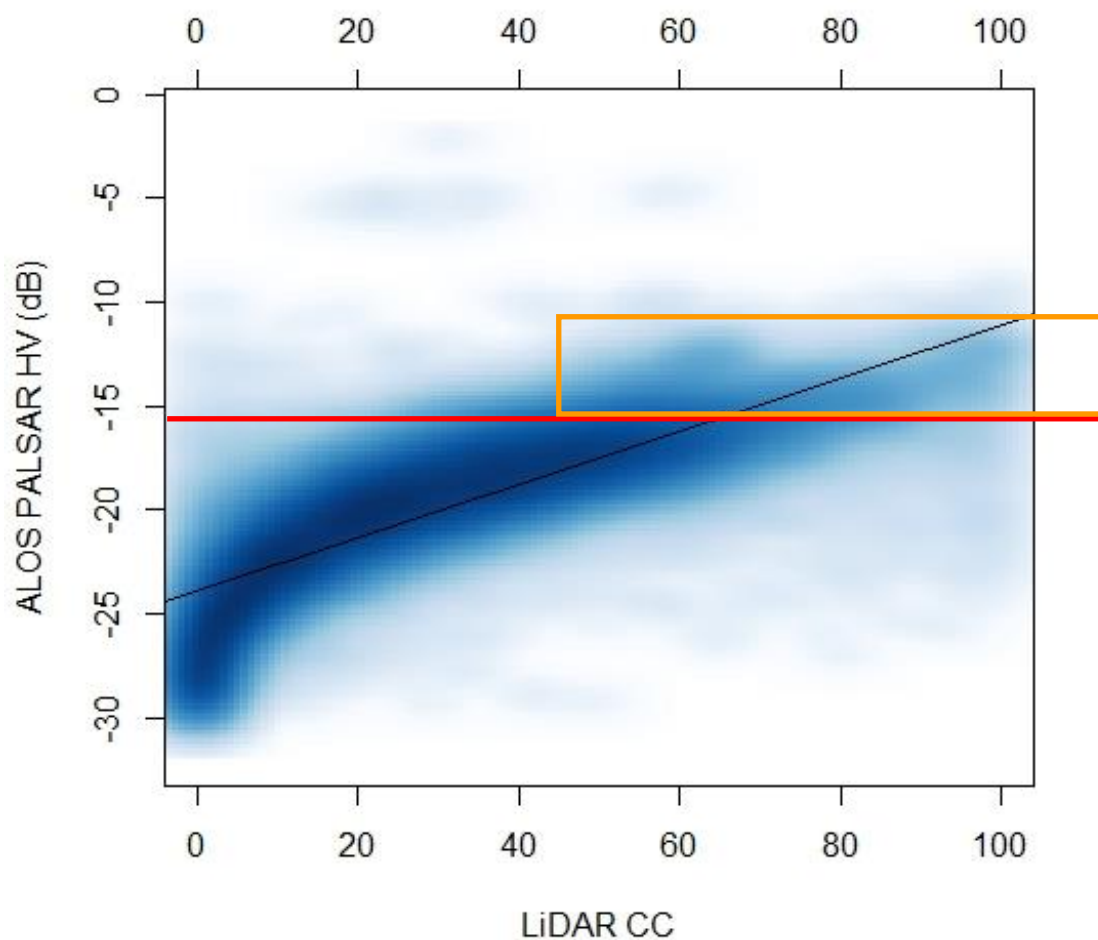


Figure 2.9: LiDAR derived woody canopy cover versus ALOS PALSAR HV backscatter (dB) extracted over the complete LiDAR dataset coverage [the red line indicates the -15.6dB threshold value (Shimada et al., 2014) used to create the FNF over the continent of Africa while the orange box indicates the bulk of the LiDAR CC values captured by the FNF according to the CC values greater than and equal to the HV dB threshold]

For the continent of Africa, a threshold range of -15.6dB HV backscatter, represented by the red line in Figure 2.9, was used for the FNF product creation (Shimada et al., 2014). This threshold was derived by ascertaining the cross-over point between forest and non-forest HV backscatter cumulative histograms collected across various regions of interest (Shimada et al., 2014). Backscatter values greater than and equal to threshold was classified as forest and the backscatter values less than the threshold was considered as non-forest (excluding the HH backscatter thresholding for urban class separation and waterbody classification outlined in (Shimada et al., 2014)). According to Figure 2.9, a small, limited portion of the upper observed CC values (50-100%) was captured by the FNF product which supported the limited representation of the major distribution of forested areas (Tables 2.2 and 2.3; Figure 2.5) and almost no representation of the 20-50% CC range. Figure 2.9, supported by table 2.2, also showed that the HV threshold of -15.6dB



contributed to the very high detection accuracy of the non-forest class (99% accuracy for CC values <10%). Obviously due to the inherent variability of the SAR signal with e.g. moisture, structural type, species, etc., and as shown by the point spread, it is clearly challenging to select a single HV dB threshold, especially across the continent of Africa, which covers the complete CC range in heterogeneous savannah environments. Hypothetically, and with the retrospective guidance of Figure 2.9, a single optimized HV threshold of -19dB can be recommended for improved forest detectability in savannah environments limited to the Southern African region. By adjusting the HV threshold to -19dB, and applying it to the 2010 ALOS PALSAR HV global mosaic data, FNF classification accuracies improved noticeably with an overall accuracy of 68.05%, a forest detection accuracy of 59.26% and a non-forest detection accuracy of 97.40% (see Appendix 2B).

The Landsat VCF product displayed underestimation past the 30% CC mark with increasing error margins towards the 90% CC mark. This increasing error margin was also documented in (Pengra et al., 2015). This trend of CC underestimation by the VCF (>30%) product was well documented in forested environments (Gao et al., 2014; Sexton et al., 2013; Song et al., 2013), in the MODIS VCF version, but not in great extents in African savannah environments. (Hansen et al., 2011) suggested that the lack of growing season imagery, in the Landsat archive over a particular area, could be one of the causes of the VCF's underestimation of forest cover. VCF product overestimation at the lower CC ranges (<30%), though minimal, was also corroborated by (Pengra et al., 2015). Within the context of South Africa, the signal noise related to grass present in open land types such as open woodlands and wooded grasslands etc., and the presence of trees less than 5m in height still captured by the Landsat imagery, could also have contributed to increased CC estimates from the VCF. This observation could be supported by the confusion matrix result (table 2.5) where the majority of the VCF error within the 0-10% LiDAR CC class fell in the higher neighbouring VCF CC classes (i.e. in the 20-30% VCF CC classes). In general, the Landsat-based VCF product did improve CC accuracies across agricultural areas, over the MODIS VCF derivative, but still experienced noted inaccuracies over woody cover areas which have a mixed tree-shrub gradient (Sexton et al., 2013). The author of this thesis recommends that more extensive ground-truth datasets, i.e. LiDAR-based metrics, especially over medium to dense forested areas and/or specific bioregions, would need to be incorporated to train the regression tree algorithm used to create the VCF product. Additionally, the characterisation of CC in the VCF product was successfully improved by integrating multi-source and multi-resolution map products (Song et al., 2013). The addition of a water mask, at the product development stage, will also help improve the VCF by distinguishing low CC values and water bodies (Montesano et al., 2009). The moderate accuracies at the 10-20% CC range and in the open

woodland tree structural class suggests that the VCF product could be potentially applicable in low CC environments such as grasslands and sparse savannahs but can also detect, to some extent, closed canopy environments (90-100% CC range).

In closing, regardless of the chosen definition of forests and product creation protocols, it is clear that forests, especially within the savannah biome, are severely underrepresented in both the VCF and FNF global forest cover products. The outcomes of (Sexton et al., 2015) illustrated the need for a standardisation of definition of forests as well as the movement towards quantitative CC products like (Hansen et al., 2013).

## 2.7 Conclusions

This study sought to validate the accuracies of two global forest cover products, the 30m Landsat Vegetation Continuous Field (VCF) and the recently introduced 25m JAXA ALOS PALSAR Forest/Non-Forest (FNF) global products, against an extensive collection of airborne LiDAR data. The primary focus of the study was to assess both products for the accurate detection of forests, as per the products' forest definitions, in Southern African savannahs which are not clearly presented or even excluded by such global products. It was found that the FNF product grossly under-represented the distribution of forests in savannah environments (20-80% CC ranges), due to the inadequate HV backscatter threshold chosen in its creation for the depiction of FNF across South Africa. With this HV threshold, however, the FNF product most accurately detected the Non-forest class (0-10% CC range), but this class also included wide tracks of forested lands. The FNF product also showed limited use in detecting closed forest cover class (90-100%) and Natural Forest and Scrub Forest tree structural classes. The Landsat VCF product displayed strong CC underestimation with increasing variability and mean error from CC values greater than 30%. The moderate accuracies at the 10-20% CC range and in the Open Woodland tree structural class suggest that the VCF product could be potentially applicable in low CC environments such as grasslands and sparse savannahs. There was, however, limited detection ability by the VCF in closed canopy environments (90-100% CC range). In the light of these results, a fixed definition of forests is necessary and a more accurate forest product, which has been specifically calibrated from locally collected datasets, will need to be developed to capture the full CC range found in the heterogeneous South African savannahs.

## Chapter 3: Savannah woody structure modelling and mapping using multi-frequency (X-, C- and L-band) Synthetic Aperture Radar (SAR) data

### 3.1 Abstract

Structural parameters of the woody component in African savannahs provide estimates of carbon stocks that are vital to the understanding of fuelwood reserves, which is the primary source of energy for 90% of households in South Africa (80% in Sub-Saharan Africa) and are at risk of over utilisation. The woody component can be characterized by various quantifiable woody structural parameters, such as tree cover, tree height, above ground biomass (AGB) or canopy volume, each been useful for different purposes. In contrast to the limited spatial coverage of ground-based approaches, remote sensing has the ability to sense the high spatio-temporal variability of e.g. woody canopy height, cover and biomass, as well as species diversity and phenological status – a defining but challenging set of characteristics typical of African savannahs. Active remote sensing systems (e.g. Light Detection and Ranging – LiDAR; Synthetic Aperture Radar - SAR), on the other hand, may be more effective in quantifying the savannah woody component because of their ability to sense within-canopy properties of the vegetation and its insensitivity to atmosphere and clouds and shadows. Additionally, the various components of a particular target's structure can be sensed differently with SAR depending on the frequency or wavelength of the sensor being utilised. This study sought to test and compare the accuracy of modelling, in a Random Forest machine learning environment, woody above ground biomass (AGB), canopy cover (CC) and total canopy volume (TCV) in South African savannahs using a combination of X-band (TerraSAR-X), C-band (RADARSAT-2) and L-band (ALOS PALSAR) radar datasets. Training and validation data were derived from airborne LiDAR data to evaluate the SAR modelling accuracies. It was concluded that the L-band SAR frequency was more effective in the modelling of the CC (coefficient of determination or  $R^2$  of 0.77), TCV ( $R^2$  of 0.79) and AGB ( $R^2$  of 0.78) metrics in Southern African savannahs than the shorter wavelengths (X- and C-band) both as individual and combined (X+C-band) datasets. The addition of the shortest wavelengths also did not assist in the overall reduction of prediction error across different vegetation conditions (e.g. dense forested conditions, the dense shrubby layer and sparsely vegetated conditions). Although the integration of all three frequencies (X+C+L-band) yielded the best overall results for all three metrics ( $R^2=0.83$  for CC and AGB and  $R^2=0.85$  for TCV), the improvements were noticeable but marginal in comparison to the L-band alone. The results, thus,

do not warrant the acquisition of all three SAR frequency datasets for tree structure monitoring in this environment.

**Keywords:** *Woody structure, Savannahs, SAR, Multi-frequency, LiDAR, Random Forest*

### 3.2 Introduction - Background, Aims and Objectives

Structural parameters of the woody component in African savannahs provide estimates of carbon stocks that are vital to the understanding of fuelwood reserves, which is the primary source of energy for 90% of households in South Africa (80% in Sub-Saharan Africa) and are at risk of over utilisation (Wessels et al., 2013, 2011). The woody component in African savannahs is an important physical attribute for many ecological processes and impacts the fire regime, vegetation production, nutrient and water cycles (Silva et al., 2001). The density of woody plants can also severely compromise the availability of grazing resources, valuable for livestock populations and related livelihoods, through bush encroachment (Wigley et al., 2009). Within the context of climate change, the sequestration of carbon by growing vegetation is a significant mechanism for the removal of CO<sub>2</sub> from the atmosphere (Falkowski et al., 2000; Viergever et al., 2008). Understanding how carbon is stored as carbon sinks in vegetative biomass and thus quantifying this standing biomass is central to the understanding of the global carbon cycle. Vegetation clearing (e.g. for cultivation) and degradation (e.g. for timber or fuelwood) and the burning of biomass, which are prevalent in developing regions and savannah woodlands of Southern Africa, can alter carbon stocks and emissions (Falkowski, 2000; Viergever et al., 2008b). Based on the important environmental implications revolving around woody vegetation, there are growing initiatives aiming at forest and woodland conservation that require its active inventorying, mapping and subsequent monitoring such as the Reducing Emissions from Deforestation and Forest Degradation programme (REDD+) (Asner et al., 2013; Corbera and Schroeder, 2011; Kanowski et al., 2011).

The woody component can be characterized by various quantifiable woody structural parameters, such as woody canopy cover (CC), tree height, above ground biomass (AGB) or total woody canopy volume (TCV), each been useful for different purposes. AGB is defined as the mass of live or dead organic matter above the ground surface (excluding roots etc.) and is usually expressed in tonnes per hectare or t/ha (Bombelli et al., 2009). Woody canopy cover (i.e. the percentage area occupied by woody canopy) is a key parameter used in monitoring vegetation change and can be combined with tree height to estimate approximate AGB (Colgan et al., 2012). Lastly, total woody canopy

volume indicates the volume of vegetation present within the vertical profile and serves as an alternative proxy for biomass density and distribution. Further, these metrics, both 2D (CC) or 3D (TCV and AGB) in nature can provide useful information regarding the prediction of density, habitat requirements and biodiversity assessments for conservation (Bradbury et al., 2005; Jung et al., 2012; Mueller et al., 2010).

Remote Sensing has been used in numerous studies as the preferred tool for quantifying and mapping woody structural features due mainly to its superior information gathering capabilities, wide spatial coverage, cost effectiveness and revisit capacity (Lu, 2006). In contrast to the limited spatial coverage of ground-based approaches, remote sensing also has the ability to sense the high spatio-temporal variability of e.g. woody canopy height, cover and biomass, as well as species diversity and phenological status – a defining but challenging set of characteristics typical of African savannahs (Archibald and Scholes, 2007; Cho et al., 2012b; Mills et al., 2006). Woody structural parameters have been successfully mapped using passive optical data at fine and coarse spatial scales (Boggs, 2010; Castillo-Santiago et al., 2010) by making use of textural (the local variance of an image related to its spatial resolution – (Nichol and Sarker, 2011)) and/or spectral (e.g. spectral vegetation indices related to vegetation structure – (Johansen and Phinn, 2006)) approaches. Passive optical data are, however, adversely affected by high spectral variation, which refers to the change in spectral properties or character of a target, due to seasonal dynamics, clouds and haze. These spectral variations are prevalent in the rainy season of African summers with veld fires in the dry winter, and in shadowed areas, which results from terrain topography and tree canopies, at fine resolutions and in mixed wood-grass pixels at the medium and coarser resolutions. Active remote sensing systems such as Light Detection and Ranging (LiDAR) and Synthetic Aperture Radar (SAR), on the other hand, may be more effective in quantifying the savannah woody component because of their ability to sense within-canopy properties of the vegetation and its insensitivity to atmosphere and clouds and shadows.

Airborne LiDAR systems provide high-resolution geo-located measurements of a tree's vertical structure (upper and lower storey) and the ground elevations beneath dense canopies. Although airborne LiDAR provides detailed tree structural products it relies on the availability of aircraft infrastructure, which is not always available in Africa. Satellite LiDAR is also currently not available. On the other hand, SAR systems provide backscatter measurements that are sensitive to forest spatial structure and standing woody biomass due to its sensitivity to canopy density and geometry

(Mitchard et al., 2011; Sun et al., 2011). A SAR-based approach offers an all-weather capacity, when using SAR intensity, to map relatively large extents of the woody component, which cannot be easily achieved with airborne LiDAR (Mitchard et al., 2011).

Polarization, which refers to the orientation of the emitted and received signal, and frequency of SAR data play important roles in sensing vegetation structure. Multi-polarized SAR systems emit and receive in HH, HV, VH and/or VV with H referring to a horizontal wave orientation and V referring to a vertical wave orientation. This allows the more complete characterisation of the scattering properties of ground targets which in turn, enables the extraction of greater structural information. For instance, HV or VH are better linked to canopy structure because of the volumetric water content in the canopies architecture (Schmullius and Evans, 1997) which brings about volumetric scattering within the canopy and its “random” scatterers, which tends to change the polarization of the emitted wave (e.g. H to V or V to H). The various components of a particular target’s structure can be sensed differently with SAR depending on the frequency or wavelength of the sensor being utilized. For example when sensing vegetation, the signal of shorter SAR wavelengths, such as X-band and C-band, interact with the fine leaf and branch elements of the vegetation resulting in canopy level backscattering with limited signal penetration. The signal of longer SAR wavelengths, such as P-band and L-band, on the other hand, can penetrate deeper into the vegetation with backscatter resulting from signal interactions with larger vegetation elements such as major branches and trunks (Mitchard et al., 2009; Vollrath, 2010). Consequently, the L-band frequency has been proven in numerous studies to be the most preferred (Carreiras et al., 2013; Mitchard et al., 2012; Ryan et al., 2012; Santos et al., 2002) and the most effective (Lucas et al., 2006a) in estimating woody structure, particularly AGB with a higher saturation level at 80-85 tonnes per hectare compared to the shorter wavelengths, in forested and savannah woodland environments. However, since woodlands and savannahs possess a sporadic combination of fine and large woody elements within individual tree canopies, and a heterogeneous distribution of large trees and smaller shrubs throughout the landscape, we hypothesized that combining the capabilities of these different SAR frequencies under a multi-sensor approach may enhance the sensing of the savannah woody element (Schmullius and Evans, 1997). Various studies have ‘fused’ or integrated multiple SAR frequency and polarimetric datasets for modelling and mapping of tree structural attributes across various environments from the coniferous temperate forests of North America to mangrove forests and to the open-forest woodlands of Australia (Collins et al., 2009; Mougin et al., 1999; Tsui et al., 2012). Despite the success achieved in these various studies via combining different SAR wavelengths (Mougin et al., 1999; Tsui et al., 2012), the combined strength of both shorter and

longer SAR frequency sensor technologies, however, have yet to be assessed in the heterogeneous and complex Southern African savannah environment.

This study sought to test and compare the accuracy of modelling woody above ground biomass (AGB), canopy cover (CC) and total canopy volume (TCV) in South African savannahs using a combination of X-band (TerraSAR-X), C-band (RADARSAT-2) and L-band (ALOS PALSAR) radar datasets. Training and validation data were derived from airborne LiDAR data to evaluate the SAR modelling accuracies. The research questions were:

- 1) How do various SAR frequencies (X- or C- or L-band) perform in predicting woody structural parameters (CC, TCV and AGB) in southern African savannahs?
- 2) Does combining SAR backscatter through different frequency combinations or scenarios (X+C or X+L or C+L band or X+C+L-band) improve the predictions of the various woody structural parameters and by how much?

We hypothesized that the combination of shorter wavelength, ~3cm X-band and ~5cm C-band, with longer wavelength, ~23cm L-band, SAR datasets, in a modelling approach, will yield an improved assessment of woody structure. This idea is based on the assumption that X- and C-band SAR signals interact with the finer woody structural constituents such as leaves and finer branchlets, typical of the shrubby/thicket layer, while the L-band SAR signal interact with the major tree structural components such as trunk and main branches which are typical of forested areas.

- 3) Finally, through the examination of the patterns of the prediction error, within the landscape for the different SAR frequency models, can the hypothesis, proposed above, be confirmed?

More specifically, the investigation of the interactions of the different SAR frequencies, and their possible combinations, across the different vegetation patterning and structural classes, such as grasslands, thickets and forests, will pin-point the effective application of the different SAR frequencies and their possible combinations in Southern African savannah landscapes.

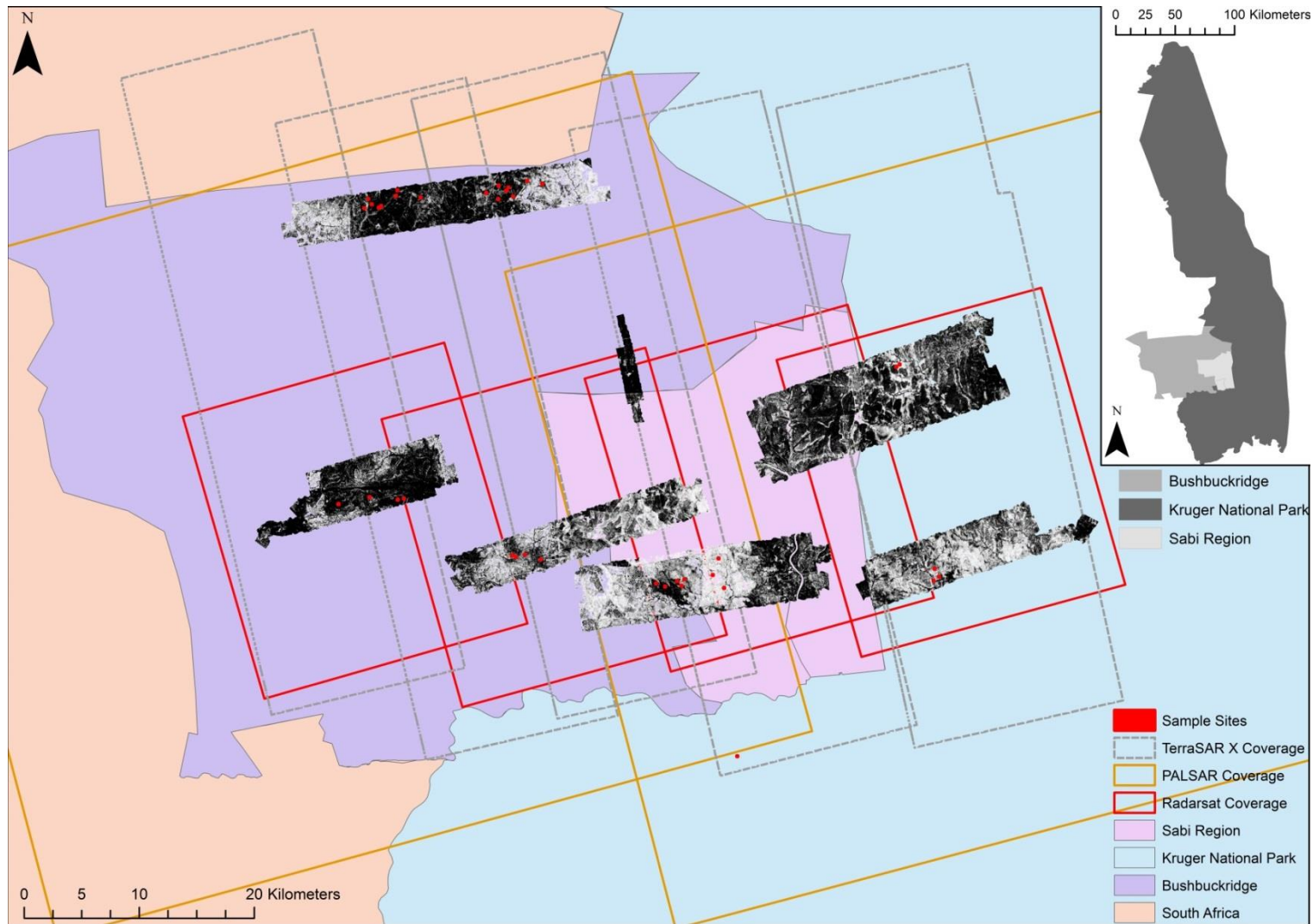
The study is broken down into various sections. Section 3.3 describes the study area under investigation. Section 3.4 and subsections focus on the material and methodology which outlines the remote sensing datasets used, field datasets collected, LiDAR and SAR pre-processing and metric generation, modelling protocols, mapping and finally validation and error assessment. Section 3.5

describes the modelling, mapping and error results while sections 3.6 and 3.7 discuss the main study outcomes and concluding remarks, respectively.

### 3.3 Study Area

The Kruger National Park regional study area is located in the Lowveld region of north-eastern South Africa, within the savannah biome (31°00' to 31°50' E longitude, 24°33' to 25°00' S latitude). The study area included portions of the southern Kruger National Park, the neighbouring Sabi Sands Private Game Reserve, and the densely populated Bushbuckridge Municipal District (BBR) (Figure 3.1). The area is characterised by short, dry winters and a wet summer with an annual precipitation varying from 235mm and 1000mm, and is representative of southern Africa savannahs. This rainfall range, together with grazing pressures, fire, geology, mega-herbivore activity and anthropogenic use (fuelwood collection and bush clearing for cultivation) govern the vegetation structure present in this biome. The vegetation comprise particularly of Clay Thornbush, Mixed Bushveld and Sweet and Sour Lowveld Bushveld (Mucina and Rutherford, 2006). The woody vegetation in the region is generally characterized as open forest with a canopy cover ranging from 20-60%, a predominant height range of 2 to 5m and biomass below 60 t/ha (Mathieu et al., 2013). The Sabi Sands Wildtuin consists of a group of private owners with a strong eco-tourism based approach to conservation with the Kruger National Park being more geared towards large-scale public conservation via the inclusion of large tracts of land for protection. The communal rangelands of BBR are primarily utilised for livestock ranching, fuelwood harvesting and various non-commercial farming practices (Wessels et al., 2013, 2011). This study region was selected to represent the differences in the woody structure (e.g. riparian zones, dense shrubs, sparse tall trees etc.) and spatial patterns of the different land management and disturbance regimes (communal rangeland management, private game reserve and national park management), varying vegetation types (lowveld savannah and mixed forest fringe species) and geological substrates (granite and gabbro).





**Figure 3.1: The Southern Kruger National Park region and the spatial coverage of all implemented remote sensing datasets. The solid red line indicates the coverage of the 2009 RADARSAT-2 scenes while the solid gold line indicates the two scenes of the 2010 ALOS PALSAR dual-pol imagery. The dashed grey line indicates the five scenes of the 2012 TerraSAR-X StripMap imagery. The shaded black areas represent the coverage of the 2012 CAO LiDAR sensor tree cover product. The red squares indicate the 38 sample sites where field data collections took place.**

## 3.4 Materials and Methodology

The general methodology sought to develop woody structural metric models between collected field data and airborne LiDAR data for detailed localised metric maps (25m spatial resolution to match the field data plots). These LiDAR derived metric products (CC, TCV and AGB) were then used as the ground truth for model up-scaling at the regional scale using multi-frequency SAR intensity backscatter datasets (X-, C- and L-band). This was achieved by integrating the LiDAR and SAR datasets with the use of a sampling grid and the extracted values were subjected to modelling using the Random Forest (RF) algorithm (Breiman, 2001). Different SAR frequencies were modelled in the form of various SAR frequency combination scenarios. The SAR-derived woody structural metrics were then validated using the LiDAR-derived woody structural metrics (CC, TCV and AGB) to ascertain error statistics and error distribution.

### 3.4.1 Remote sensing data

Five TerraSAR-X X-band dual-polarized (HH and HV), four RADARSAT-2 C-band quad-polarized (HH, VV, VH, and HV) and two ALOS PALSAR L-band dual-polarized (HH and HV) SAR intensity datasets (summarized in Table 3.1) were acquired to cover the study transect shown in Figure 3.1. Only dual polarized SAR data (HH and HV) was used because the HV polarization parameter is known to better model the structure of woody vegetation through volumetric backscatter interactions, while HH is also reported as been sensitive to structure although to a lesser extent than the cross-polarized band (Collins et al., 2009; Mathieu et al., 2013; Mitchard et al., 2009). Further, HH/HV was the common polarization configuration available for all three sensors. Winter seasonal SAR acquisitions were chosen because winter in the Lowveld is the dry season and exhibits the lowest level of moisture in the landscape. The tree leaves are off along with dry soil and dry grasses. This reduced the chance of interference of the SAR signal with variable moisture content while allowing a greater penetration of microwaves into the canopies. In the same region (Mathieu et al., 2013) reported the best retrieval of woody structural parameters with RADARSAT-2 data acquired in winter. An extensive airborne LiDAR dataset (total coverage of c.a. 63000 ha) were acquired for this study (Figure 3.1) by the Carnegie Airborne Observatory-2 AToMS sensor during April-May 2012. For our datasets, the LiDAR was operated at a pulse repetition frequency of 50 kHz with a 0.56m laser spot spacing and an average point density of 6.4 points per m<sup>2</sup> from a flying altitude of 1000m above ground level (Asner et al., 2012). In comparison with the LiDAR dataset, the SAR images were acquired during the winter 2009 (RADARSAT-2), 2010 (ALOS PALSAR), and 2012 (TerraSAR-X).

Unfortunately, the last ALOS PALSAR winter scenes were acquired during 2010 in the study area and no RADARSAT imagery were available closer to 2012.

### 3.4.2 Field data

Field data were collected in April – May, and November – December 2012 across 38 sampling sites (in Figure 3.1). These sites provided ground truth data to model and validate the LiDAR derived woody structural metric products to be used to model the SAR-based woody structural metrics. Ground sampling sites were located to represent the diversity in woody structure of the different vegetation types, management regimes, and geological substrates mentioned above. Each site covered a 100m X 100m area and vegetation measurements were taken from four clustered 25m X 25m sampling plots (with minimum distance > 50m, identified from geostatistic range assessments, (Wessels et al., 2011)), located at each of the four corners of the site (Figure 3.2). The 100m X 100m sites were positioned using high resolution imagery from Google Earth as well as earlier LiDAR datasets acquired in 2008 – 2010 to ensure that they are representative of the surrounding landscape.

Field AGB estimates were derived from height and stem diameter measurements using an allometric biomass estimation equation ((Colgan et al., 2013) – Equation 3.1 in Appendix 3A). The allometric equation was developed following destructive harvesting of 17 savannah tree species present in the study area (Number of trees sampled =707;  $R^2 = 0.98$ ; relative Root Square Error = 52%; ranging from 0.2 – 4531 kg per tree, (Colgan et al., 2013)). Tree height was measured using a height pole and Laser vertex/rangefinder, while stem diameter was measured using callipers and Diameter above Breast Height (DBH) tape. Stem diameter was measured at 10cm above the ground and for multi-stemmed plants every individual stem was measured as separate individuals (e.g. species such as *Dichrostachys cinerea*).

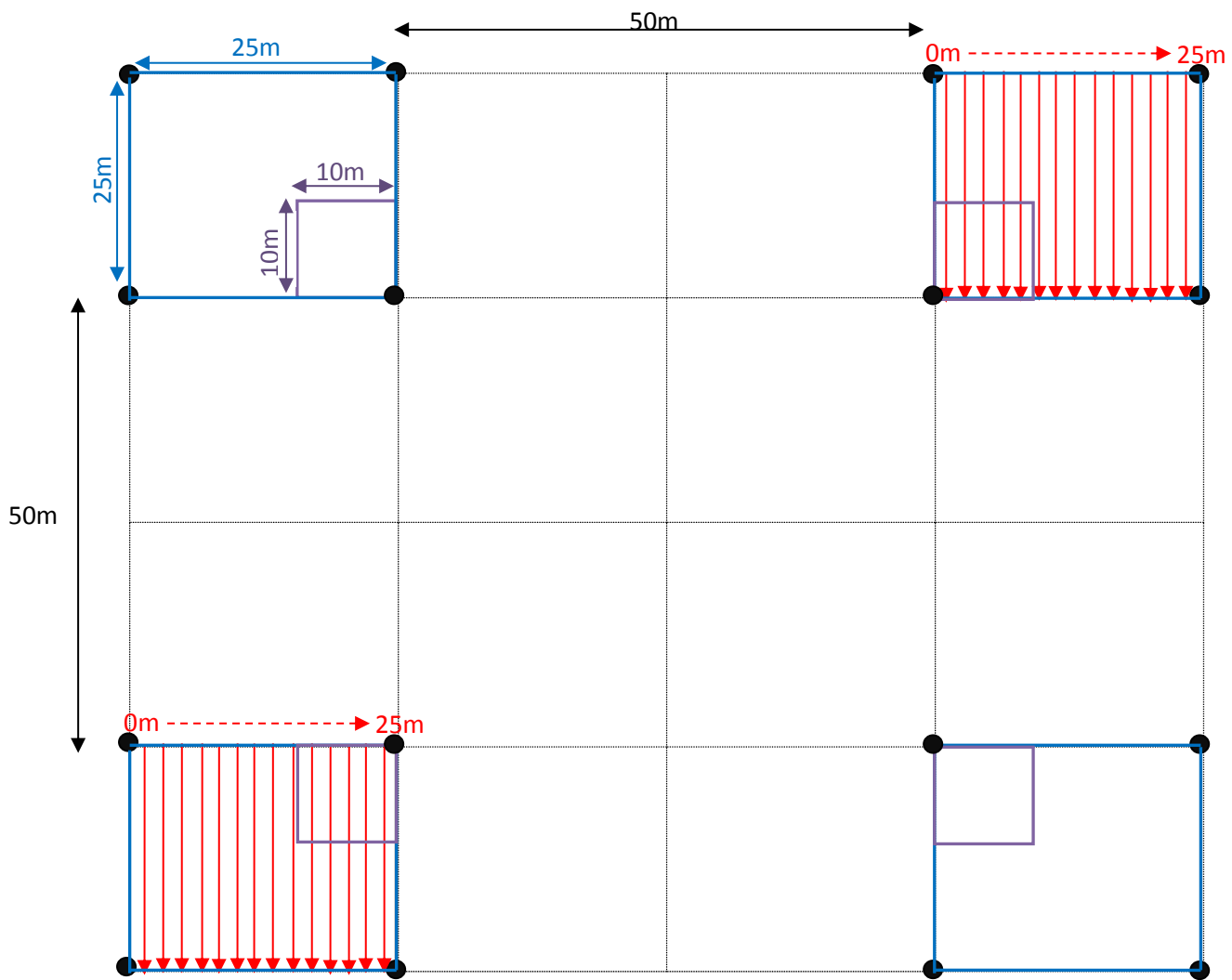
Due to logistical and time constrains associated with measuring every tree within the sample plot two main stem diameter ‘zones’ were identified inside the site to increase sampling efficiency while still yielding representative quantities of biomass estimates (Figure 3.2). The first diameter zone was the 25m X 25m plot where all trees with a stem diameter of 5cm and greater were recorded, provided that they had a height of 1.5m or greater, and the second diameter zone was a 10m X 10m area positioned at the inner corner of the 25m X 25m plot where all trees with a stem diameter

between 3 and 5cm and greater than 1.5m were also recorded. This allowed catering for a few sites, mostly in the communal lands, where most of the AGB consisted of dense stands of multi-stemmed plants (coppicing) with low DBH (Matsika et al., 2012). A total of 152 25m X 25m biomass plots were sampled. Individual tree level AGB was derived using Colgan's allometric equation (Colgan et al., 2013). AGB was then calculated for each diameter zone by summing the relevant tree level AGB values which was then subjected to particular AGB up-scaling factor (Equation 3.2 in Appendix 3B). The complete plot level AGB was calculated by summing all the corrected AGB subtotals for the stem diameter zones.

One or two sampling plots were chosen for most sites for CC data collection – the north east 25m X 25m plot and/or the south west 25m X 25m plot (DBH zone 2 – Figure 3.2). CC values were estimated following the vertical densitometer protocol (Ko et al., 2009; Stumpf, 1993), conceptually a point intercept sampling approach, and one of the most time-efficient techniques to implement. The point intercept method is a small angle approach well suited to measure the vertical canopy cover – i.e. vertical projection of canopy foliage onto a horizontal surface –, and as such is the most directly comparable with cover derived from remote sensing imagery such as LiDAR (Fiala et al., 2006). The sampling procedure involved laying down transects along a fixed 25m measuring tape orientated from north to south and moving from west to east within the subplot at 2m increments (Figure 3.2). Along these transects, the presence of canopy cover was determined using a 5m pole placed vertically above each sampled points every 2m along the transects. At each sampled point the presence of cover was coded as Y. For plot level canopy cover, in terms of percentage at the 25m X 25m scale, the CC presence and absence data were subjected to the formula below (Equation 3.3):

$$\text{Plot level CC (\%)} = (\Sigma Y / 169) \times 100 \qquad \text{Equation 3.3}$$

Where Y represents the presence of cover data. The value 169 represents the total number of sampling points in a 25m X 25m plot conducted at 2m sampling increments. A total of 37 (25m X 25m) plots of CC were recorded during the field campaign.



**Legend and Sampling Protocols for Tree Biomass and Cover:**

- Corner pole markers
- DBH\* Zone 1: 10m X 10m [Trees with DBH  $\geq$  3cm measured]
- DBH Zone 2: 25m X 25m [Trees with DBH  $\geq$  5cm measured]
- ▮ Line Transect and Vertical Densitometer methods (1m and/or 2m intervals)

\* Note: DBH refers to Diameter above Breast Height (DBH)

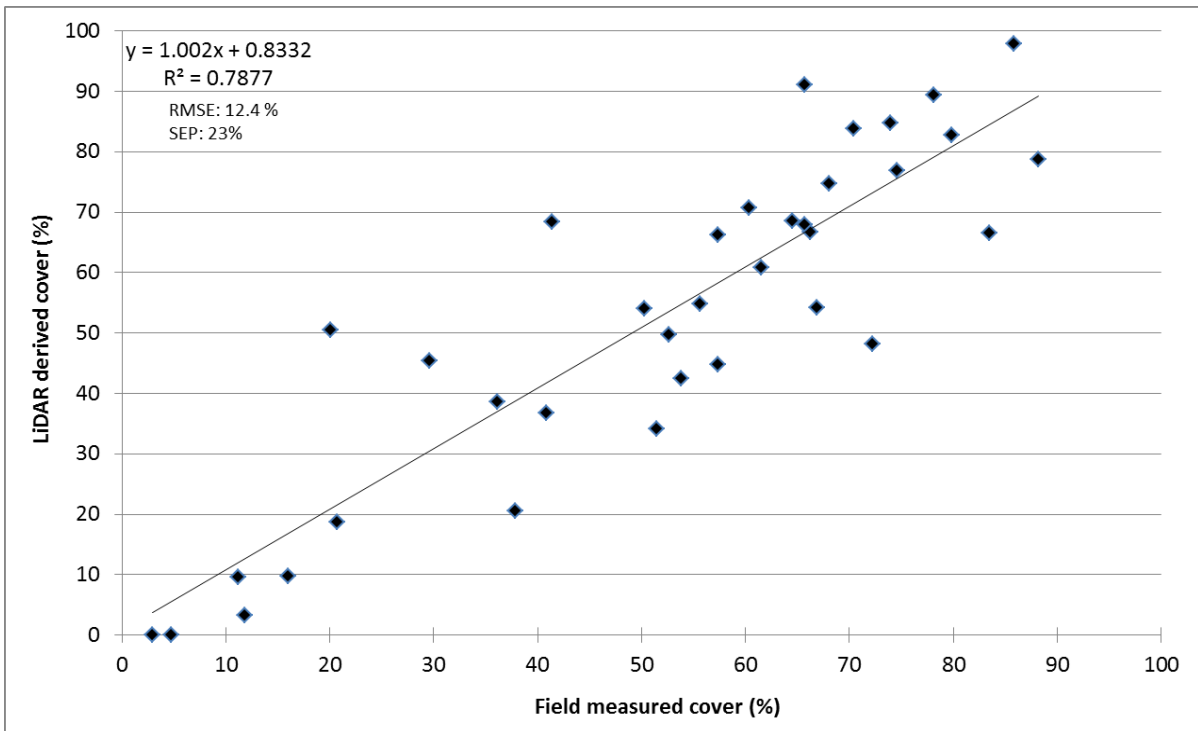
Figure 3.2: Ground sampling design including ground tree biomass and tree cover collection protocols (50m spacing between sample plots coincide with the auto-correlation distance – refer to data integration of section 3.4.5)

### 3.4.3 LiDAR data processing, woody structural metrics and validation

Two LiDAR datasets were utilised to derive the LiDAR tree structure metrics. For the first dataset, ~1m Digital Elevation Models (DEM) and top-of-canopy surface models (CSM) were created by processing the raw LiDAR point clouds according to the steps outlined in (Asner et al., 2012). Canopy height models (CHM, pixel size of 1.12m) were computed by subtracting the DEM from the CSM. For the second dataset, the raw point cloud data were further processed to pseudo waveforms, in which the LiDAR hits or returns falling within a cube placed above the ground were binned into volumetric pixels (voxels of 5m X 5m horizontal X 1m vertical) and weighted relative to the total number of hits within the vertical column (the result – LiDAR slicer data) (Asner et al., 2009).

Three woody structural metrics were derived from the processed LiDAR datasets. The derivation of the three metrics excluded all woody vegetation below a height threshold of 0.5m as to exclude the grassy savannah component. The Carnegie Airborne Observatory (CAO) LiDAR data were validated against field height measurements of approximately 800 trees. There was a strong relationship ( $R^2 = 0.93$ ,  $p$ -value  $< 0.001$ ) but a fraction of woody plants below 1.5-1.7m were not detected by the LiDAR (Wessels et al., 2011). This would introduce a source of error in the modelling process. However, since our objective was to investigate the potential contribution of short microwaves (X-band and/or C-band) in detecting the shrubby layer we still preferred to use a 0.5m height threshold over a higher height threshold at 1.5m. In addition, all metric products have been resampled and computed at the 25m spatial resolution to correspond with the ground data measurements (plot size of 25m X 25m) collected in the field for metric validation. These metrics are described in detail below:

- 1) Woody Canopy Cover (CC) is defined as the area vertically projected on a horizontal plane by woody plant canopies (Jennings et al., 1999). The metric was created by first applying a data mask to the LiDAR CHM image in order to create a spatial array of 0s (no woody canopy) and 1s (presence of a woody canopy). A percentage woody canopy cover distribution image (summing all the 1's and dividing by 625 and then percentage) was calculated at a spatial resolution of 25m. This metric was validated against the 37 25m X 25m CC ground truth plots (Figure 3.3). Results yielded a strong, positive, unbiased relationship ( $R^2=0.79$ ) with a low Root Mean Squared Error (RMSE) (12.4%) and Standard Error of Prediction (SEP) (23%).



**Figure 3.3: Validation results of field-measured woody Canopy Cover (CC) versus LiDAR derived CC (above 0.5m height, Number of observations =37)**

- 2) Total Canopy Volume (TCV) is a metric which approximates the area under the curve of the pseudo waveform (i.e. a plot displaying the LiDAR return frequency-by-height; (Muss et al., 2011)) and indicates the volume occupied by vegetation matter within the vertical profile. The metric was computed from the pseudo waveform LiDAR data (i.e. voxel) by the addition of the within-canopy LiDAR returns at different heights or slices (incrementally increasing by 1m) above 0.5m (Asner et al., 2009), and the value was converted to hectare. The TCV LiDAR metric was not validated with ground collected data as a suitable field sampling approach was yet to be defined for this type of savannah environment. However, in (Mathieu et al., 2013), the TCV metric, in comparison to all the other metrics, was best correlated with RADARSAT-2 backscatter and was thus considered a suitable metric in this study.
  
- 3) Above ground woody biomass (AGB) is defined as the mass of live organic matter present above the ground surface (Bombelli et al., 2009) and is expressed in this study as tonnes per hectare (t/ha). The AGB LiDAR derived metric was modelled using a linear regression, ground estimated AGB (within 25m field plots) and a simple HGT X CC LiDAR metric (where

HGT is the mean top-of-canopy height and CC is the canopy cover of a 25m pixel resolution) (Colgan et al., 2012). 65% of the 152 ground estimated AGB was used for model development while the remaining 35% was used for model validation. The validation results of ground versus LiDAR AGB (Figure 3.4) indicate a moderate positive correlation ( $R^2=0.63$ ). With the use of allometric equations from (Colgan et al., 2013) for ground AGB estimation, the RMSE (19.2 t/ha) and SEP (63.8%) is, however, high with underestimation at high biomass levels by the LiDAR. Due to the intensive and time consuming nature of sampling these very high biomass plots, an insufficient number of these plots may have been sampled to suitably train the model which thus led to such a deviation from the 1:1 line at the high biomass levels in Figure 3.4. In the absence of better biomass estimates, the LiDAR derived AGB metric was deemed sufficient for the modelling and validation.

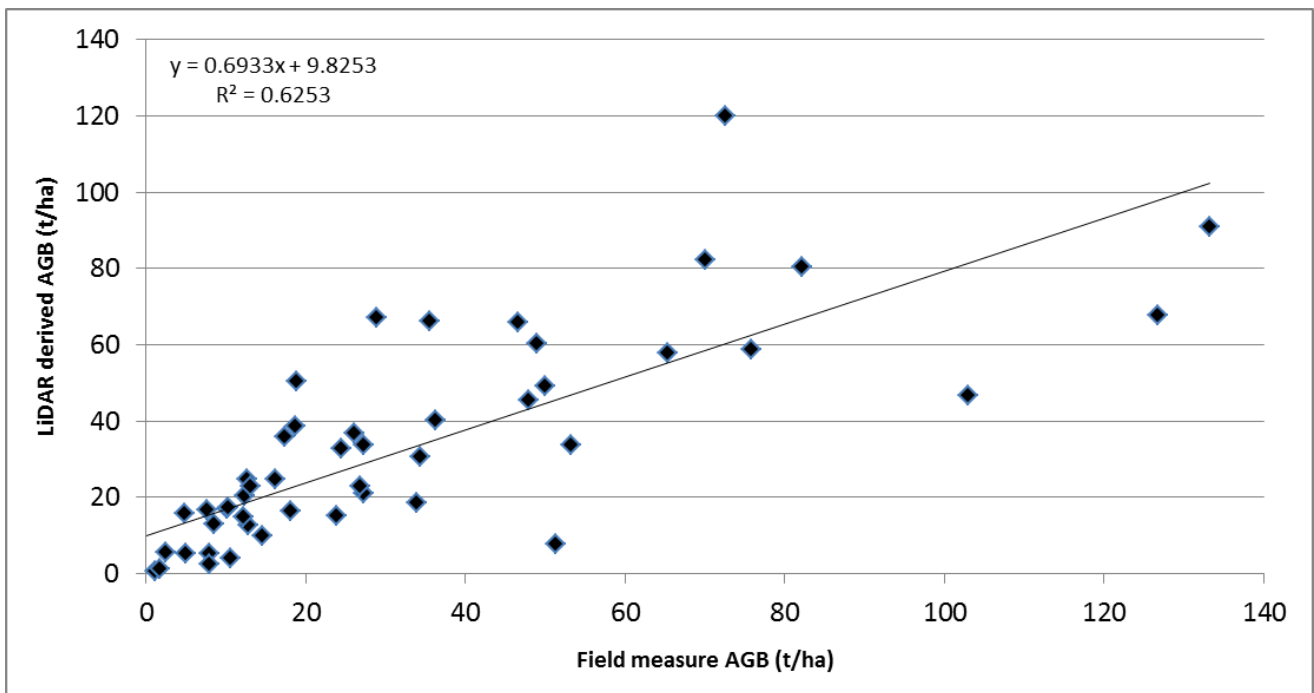


Figure 3.4: Validation results of field-measured Above Ground Biomass (AGB) versus LiDAR derived AGB (above 0.5m height, Number of observations =53)

### 3.4.4 SAR data and processing

The SAR intensity images (X-, C- and L-band) were pre-processed according to the following steps: multi-looking, radiometric calibration (conversion of raw digital numbers into sigma naught ( $\sigma^0$ ) backscatter values), geocoding, topographic normalization of the backscatter and filtering. These steps were compiled in the form of scripts in GAMMA<sup>TM</sup> radar processing software (Gamma Remote Sensing, Copyright © 2000-2011) for the Dual Polarised TerraSAR-X X-band (StripMap, Level 1b,



Multi Look Ground Range Detected), Fine Quad Polarised RADARSAT-2 C-band (Single Look Complex) and Dual Polarised ALOS PALSAR L-band (Level 1.1) data. A 20m Digital Elevation Model (DEM) and a 90m Shuttle Radar Topography Mission (STRM) DEM were both used for the geocoding and orthorectification of the X-, C- and L-band SAR imagery. The 20m DEM was computed from South African 1:50 000 scale topographic maps (20m digital contours, spot-heights, coastline and inland water area data – ComputaMaps; [www.computamaps.com](http://www.computamaps.com)) with Root Mean Square (RMS) planimetric error of 15.24m and a total vertical RMS error of 6.8m. The 90m (3 arc sec) STRM DEM was gap-filled using Aster Global Digital Elevation Map data and was derived from 20m interval contour lines extracted from 1:50 000 topographical maps. An automated hydrological correction was applied to correct inaccuracies along river lines and tributaries (Weepener et al., 2011). The multi-looking factors and filtering were chosen to best minimize the effect of speckle while not deteriorating the spatial detail captured by the sensors. 4:4, 1:5 and 2:8 range and azimuth multi-looking factors were implemented for the X-, C- and L-band datasets respectively. All datasets were resampled, using a bicubic-log spline interpolation function, to their final map geometry resolutions. This was achieved by applying a DEM oversampling factor (DEM resolution / Final image resolution) to the multi-looked SAR datasets which was set in the “gc\_map” module under the GAMMA Differential Interferometry and Geocoding package. The original pixel size, multi-looking factors used in the pre-processing, modified pixel size (after multi-looking) and the final pixel size (i.e. map geometry) of the different SAR datasets were summarised in Table 3.2. Finally, a Lee filter (3 pixel X 3 pixel filtering window) (Lee, 1980) was applied to the images. It is important to note that the full extents varied for the different SAR datasets due to sensor coverage programming and specifications (Figure 3.1).

**Table 3.1: SAR and LiDAR datasets acquired and utilised for the modelling of woody structural metrics**

Imagery	Sensor	Mode	Incidence angle	Acquisition time	Season
1 2 3 4 5	TerraSAR-X X-band $\beta$	StripMap Dual Polarized (HH & HV)	38.1-39.3° 21.3-22.8° 37.2-38.4° 36.2-37.4° 39.1-40.2°	08/09/2012 23/08/2012 28/08/2012 19/09/2012 30/09/2012	Late Winter 2012
1 2 3 4	RADARSAT-2 C-band $\yen$	Quad Polarized (HH, HV, VH, VV) but only HH and HV used	34.4 - 36.0° 39.3 - 40.1° 32.4 - 34.0° 37.4 - 38.9°	13/08/2009 06/08/2009 06/09/2009 30/08/2009	Winter 2009
1 2	ALOS PALSAR L-band $\sigma$	Dual Polarized (HH & HV)	34.3°	14/08/2010 31/08/2010	Winter 2010
AGB (kg) Product CC (%) Product TCV Product	CAO LiDAR $\phi$	Discrete Footprint	Nadir	1/04/2012- 24/05/2012	End summer 2012

$\beta$ : <http://www.geoimage.com.au/satellite/TerraSar> ;  $\yen$ : <http://www.asc-csa.gc.ca/eng/satellites/radarsat/radarsat-tableau.asp> ;  $\sigma$ : <http://www.eorc.jaxa.jp/ALOS/en/about/palsar.htm> ;  $\phi$ : Asner et al., (2012)

**Table 3.2: Original, modified and final SAR pixel size changes during multi-looking and pre-processing steps**

SAR Dataset	Original Pixel Size [m] (Range X Azimuth)	Multi-Looking factors (no. Looks for Range X Azimuth)	Modified Pixel Size [m] (after multi-looking)	Final Pixel Size [m] (map geometry) $\Phi$
ALOS PALSAR FBD	9.37 X 3.23	2 X 8	18.74 X 25.84	12.5 X 12.5
RADARSAT-2 SLC	4.70 X 5.10	1 X 1	4.70 X 5.10	5 X 5
TerraSAR-X StripMap MGD	2.75 X 2.75	4 X 4	11 X 11	12.5 X 12.5

$\Phi$  Resolutions used in the modelling stage but all were resampled to 12.5m for mapping

### 3.4.5 Data integration, modelling protocols and mapping

Before modelling could be conducted the different datasets had to be processed to a common spatial grid. A sampling grid strategy was implemented as the relationship between dependent (LiDAR) and independent (SAR backscatter intensity) datasets were not evident on a pixel-by-pixel basis mainly due to issues of SAR speckle and pixel-level inaccuracy of co-registration between datasets. This strategy also served as a means of extracting information from various remote sensing datasets of varying spatial resolutions (see Table 3.1 and Table 3.2) without the need for pixel level fusion procedures. A regular spatial grid made up of 105m resolution cells at 50m distance spacing was created in QGIS 2.2 (Quantum GIS, Copyright © 2004-2014) and applied over the datasets. The choice of the cell size was informed by (Mathieu et al., 2013), who tested various grid sizes ranging from 15m and 495m with RADARSAT-2 C-band data, and reported the 105m grid size as the resolution which provided the best trade-off between the finest spatial resolution/mapping scale and strongest correlation with the LiDAR woody structure parameters. Similar results (50-125m grid size) were reported with ALOS PALSAR L-band data in the region (Urbazaev et al., 2015). The 50m distance spacing between the grid cells was chosen to avoid autocorrelation effects arising from the inherent distribution of the vegetation structural parameters across the landscape (Wessels et al., 2011). Informal settlements, the main roads and water surfaces such as rivers and dams were masked and excluded from the analysis. Mean values within each cell were extracted for the SAR (X-HH, X-HV, C-HH, C-HV, L-HH and L-HV) and LiDAR metric datasets (CC, TCV and AGB). Due to the differences in spatial coverage of the multi-frequency SAR datasets in relation to the LiDAR coverage (Figure 3.1), a varying number of data records (21170 records for X-band, 17980 records for C-band and 21467 records for L-band) were obtained during aggregation to the 105m grid. Various data mining, regression and machine learning algorithms (linear regression, support vector machines, REP decision trees, artificial neural network and random forest) were tested in (Naidoo et al., 2014) and Random Forest (Breiman, 2001) was found to be the

most robust and efficient, in terms of running time and accuracies (Ismail et al., 2010; Prasad et al., 2006). The article of (Naidoo et al., 2014) is available in Appendix 3C in its entirety. Unlike other traditional and fast learning decision trees (e.g. Classification And Regression Trees or CART), RF is insensitive to small changes in the training datasets and are not prone to overfitting (Ismail et al., 2010; Prasad et al., 2006). Additionally, RF is less complex and less computer intensive in comparison to the high levels of customisation required for Artificial Neural Networks (ANN) and the long 'learning' or training times for Support Vector Machines (SVM) (Anguita et al., 2010). RF requires two main user-defined inputs – the number of trees built in the 'forest' or 'ntree' and the number of possible splitting variables for each node or 'mtry' (Ismail et al., 2010; Prasad et al., 2006).

RF was applied, using R rattle data mining software (Togaware Pty Ltd., Copyright © 2006-2014), to the data with 35% of the data being used for model training and the remaining 65% being used for model validation. For the modelling process, the SAR frequency datasets were selected as the input (independent) variables while the LiDAR derived metrics were selected as the target (dependent) variables. The random forest models were built using the default setting parameters ('ntrees' = 500 and 'mtry' =  $\sqrt{\#}$  SAR predictors) and the trees were allowed to grow without pruning. Predicted versus observed scatterplots and validation scores were outputted to calculate the model accuracy statistics. The coefficient of determination ( $R^2$ ), Root Mean Square Error (RMSE) and Standard Error of Prediction (SEP in % which also known as the Relative RMSE) were computed and the modelling algorithm accuracies were compared for the individual SAR scenarios. RMSE and SEP are considered to be more informative in assessing model performance than  $R^2$  and its derivatives (e.g. adjusted  $R^2$ ). Seven modelling SAR scenarios (X-band only, C-band only, L-band only, X+C-band, X+L-band, C+L-band and X+C+L-band) were chosen to investigate the relationships between the individual SAR frequencies alone and different multi-frequency SAR combinations correlated against the three LiDAR metrics.

The best performing RF model, for each woody structural metric, was applied to the relevant SAR imagery, which were all clipped to a common coverage, resampled (pixel aggregate) to a common resolution of 12.5m to match the coarsest L-band and stacked, by using a mapping script. This script was developed in the R statistical software (Version 2.15.2, The R Foundation for Statistical Computing, Copyright © 2012) which utilised the combination of the 'ModelMap', 'Random Forest' and Geospatial Data Abstraction Library (GDAL) modules. The map products were imported into ArcMap 10.1 (ESRI, Copyright© 1995-2014) and displayed in discrete class intervals (total of 6

classes) to best illustrate the tree structural metric distribution representative of the entire modelled ranges.

### 3.4.6 Error assessment

The purpose of this section was to investigate the error produced by the different SAR models under varying tree structural scenarios, and to ascertain whether spatial patterns in error were associated with specific vegetation structural cohort types (e.g. grassland versus woodland conditions etc.). Error statistics and maps were created by subtracting the LiDAR-derived and SAR-derived woody (LiDAR – SAR) structural metric maps for TCV, AGB and CC. The SAR derived metric maps were resampled to 25m, via pixel aggregate, to match the LiDAR metric spatial resolution first before the subtraction. The error statistics for all metrics were documented but the TCV error maps were chosen for presentation over CC due to the metric's three dimensional properties which would best capture the SAR backscatter interactions. AGB error maps, however, were not displayed due to the high error in the dense forest canopies (plots not displayed but supported by the error observed between the ground AGB and LiDAR derived AGB in Figure 3.4, before AGB up-scaling to the SAR). For ease of interpretation of the error statistics and maps, the error values were grouped into 5 main groups using intervals which best covered the error range observed in the different metrics. These groups were major overestimation, minor overestimation, negligible error, minor underestimation and major underestimation.

Additionally, we assessed the following main vegetation structural cohort types typical of savannah landscapes: low cover and variable tree height (e.g. sparse veld), high cover and high tree height (e.g. forests) and high cover and low tree height (e.g. bush encroaching shrubs). The combined use of CC and vegetation height metrics best described these structural cohorts than the use of AGB and/or TCV metrics. Box and whisker plots were created from the mean LiDAR-SAR difference values (i.e. prediction error), which were extracted from the same sampling (105m) grid used in the predictor variable extraction process, and interpreted. A total of 17559 difference pixel values were used to generate the boxplots with the outlier values being removed. Similar error assessment analyses were conducted over different landscape geologies (e.g. granite versus gabbro) and topographic features (e.g. crest, slope and valleys) but the error distribution patterns were fairly similar without any distinct patterns to comment on. The complete methodology have been summarized and compiled in the form a methodological schema (Figure 3.5).

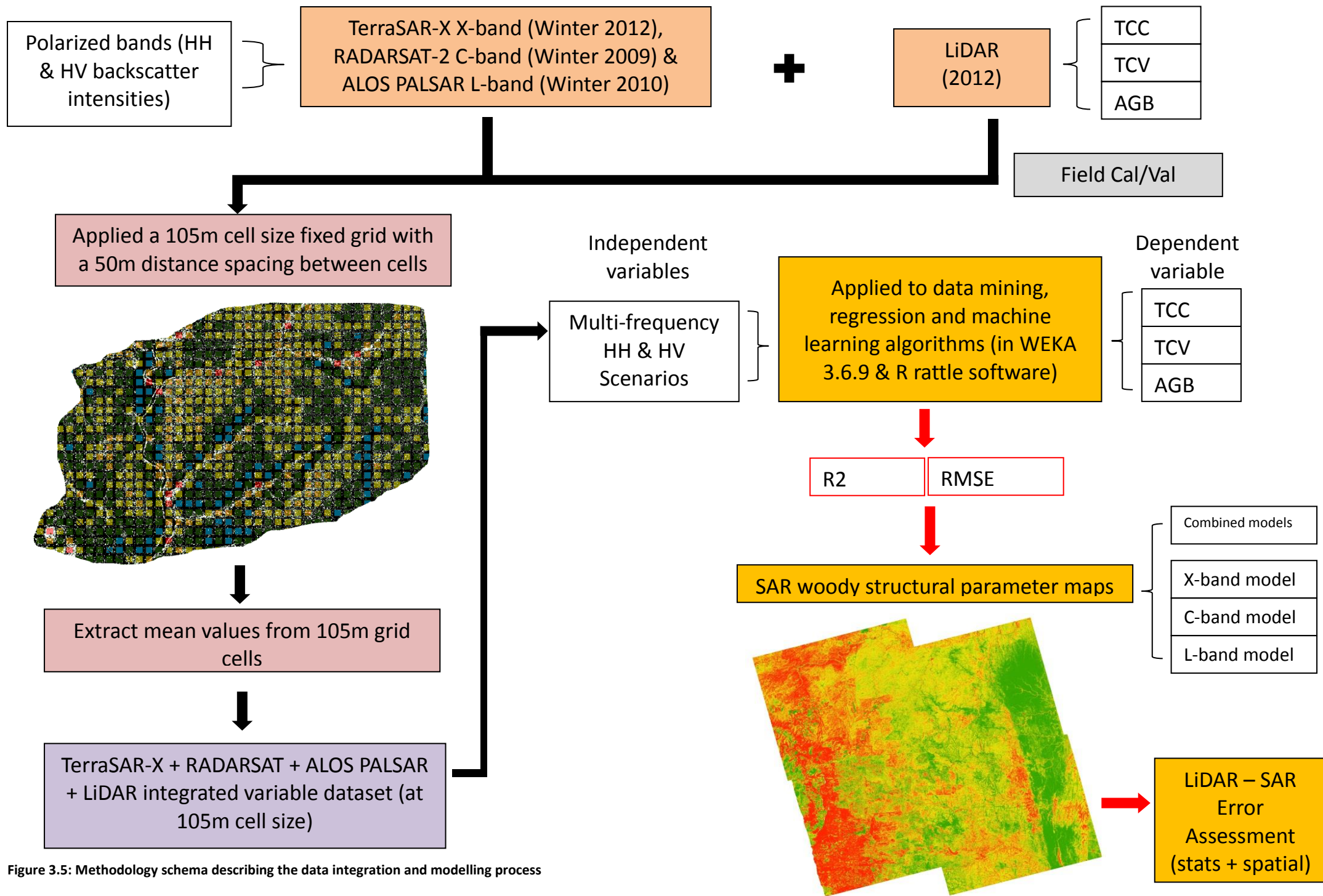


Figure 3.5: Methodology schema describing the data integration and modelling process

## 3.5 Results

### 3.5.1 Modelling Accuracy Assessment

**Table 3.3: Woody Canopy Cover (CC), Total Canopy Volume (TCV) and Above Ground Biomass (AGB) parameter modelling accuracy assessment (validation) results obtained from the Random Forest algorithm according to seven SAR frequency scenarios**

SAR Frequency	CC (%)		TCV (unitless per hectare)		AGB (tonnes per hectare)	
	R <sup>2</sup>	RMSE (SEP %)	R <sup>2</sup>	RMSE (SEP %)	R <sup>2</sup>	RMSE (SEP %)
<i>X-band only</i>	0.34	18.12 (50.87)	0.35	35534.50 (33.79)	0.32	10.88 (59.82)
<i>C-band only</i>	0.61	13.20 (38.50)	0.66	24731.06 (24.07)	0.60	7.81 (43.66)
<i>L-band only</i>	0.77	10.59 (29.64)	0.79	19902.79 (18.88)	0.78	6.05 (32.90)
<i>X+C-band</i>	0.69	11.71 (33.94)	0.72	22243.64 (21.59)	0.67	7.19 (40.33)
<i>X+L-band</i>	0.80	9.90 (27.78)	0.82	18609.04 (17.70)	0.81	5.70 (31.35)
<i>C+L-band</i>	0.81	9.23 (26.94)	0.83	17236.50 (16.77)	0.81	5.45 (30.44)
<i>X+C+L-band</i>	<b>0.83</b>	<b>8.76 (25.40)</b>	<b>0.85</b>	<b>16443.57 (15.96)</b>	<b>0.83</b>	<b>5.20 (29.18)</b>

*Datasets split into 35% Training and 65% Validation for modelling*

Table 3.3 illustrates the validation performances of the different SAR predictors, under various multi-frequency SAR scenarios, in predicting the three woody structural LiDAR metrics (CC, TCV and AGB). When examining the individual SAR frequency performances for modelling all three metrics, the longer wavelength L-band PALSAR predictors consistently yielded higher accuracies in comparison to the shorter wavelength predictors of both X-band TerraSAR-X and C-band Radarsat-2. The X-band TerraSAR-X predictors by far consistently produced the lowest modelling accuracies. The combination of the short wavelength SAR datasets (X- and C-band) improved the tree structural modelling over the individual dataset accuracies results but never produced accuracies greater than the use of the L-band dataset alone. The combined use of all three SAR frequencies (X-, C- and L-band) data in the modelling process consistently yielded the highest accuracies for modelling all three structural metrics (refer to the highlighted results for each metric in Table 3.3). In comparison to the results for L-band alone, there was a relative improvement of 10% or greater for all three structural metrics in modelling accuracies when the shorter wavelength datasets (X- and C-band) were added. However, the inclusion of the L-band frequency contributed the most to the overall accuracies. Overall, the three metrics were modelled at high accuracies under the multi-frequency scenario (X-, C- and L-band) and with similar patterns when considering the various individual scenarios.

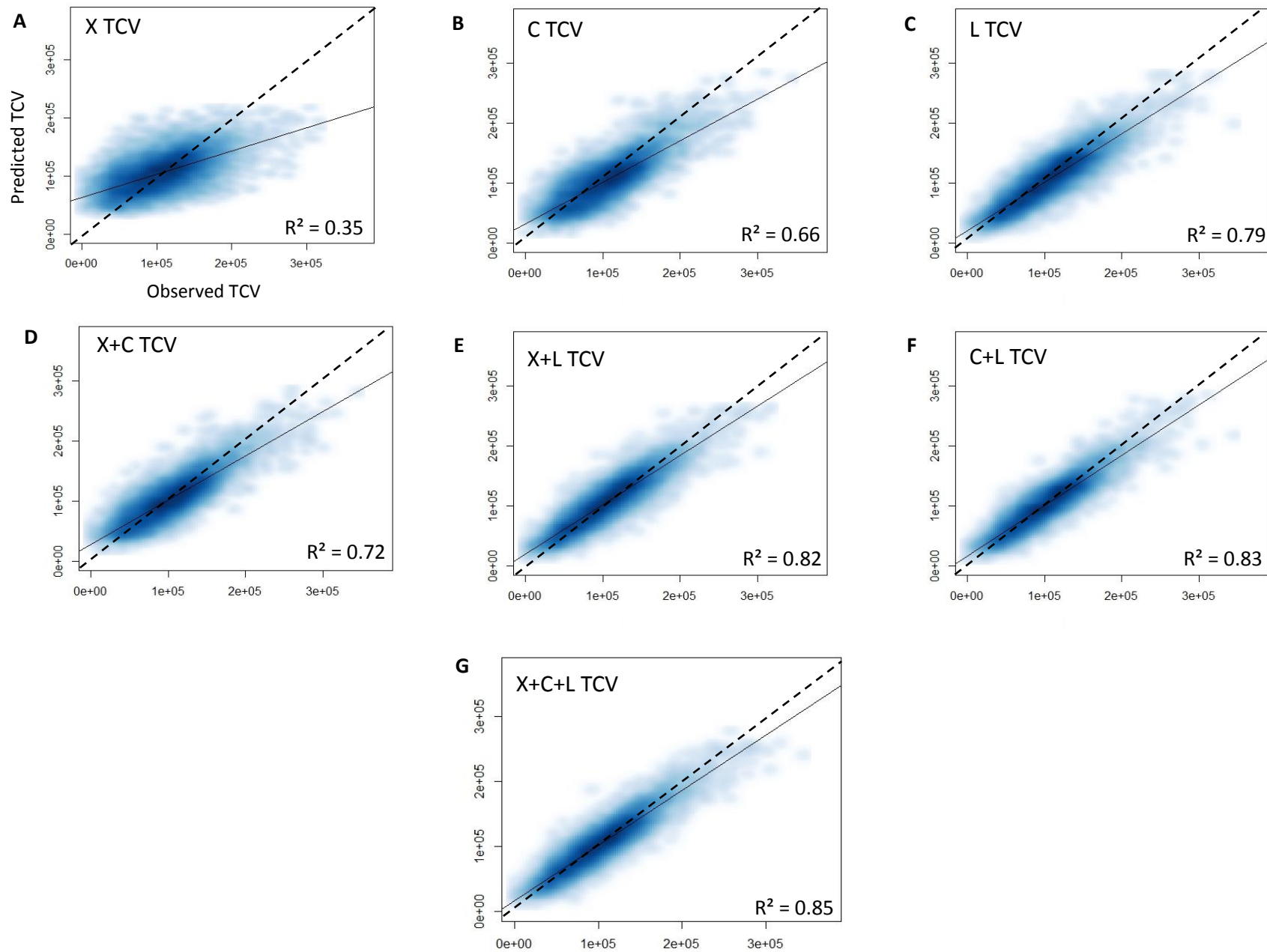


Figure 3.6: Observed versus Predicted Total woody Canopy Volume (TCV) scatter density plots (A-G) (dotted line is 1:1)



Figures 3.6A-G illustrates, by way of the 1:1 line, the extent of over-prediction and under-prediction by the models which is gradually reduced towards the multi-frequency scenarios. The TCV results were chosen for representation in Figures 3.6A-G as the metric yielded the highest overall modelled accuracies and the remaining metrics (CC and AGB) displayed similar trends throughout the different SAR frequency combinations. For TCV (Figures 3.6A-G), general over-prediction is observed at values less than  $\pm 100000$  (no unit) TCV while general under-prediction is observed at values greater than this threshold.

### 3.5.2 Tree Structure Metric and Error Maps

All three metrics were mapped for the study area (Figure 3.7i-iii) using the multi-frequency SAR models (X+C+L-band). Figures 3.7(i-iii) illustrate the spatial distributions of AGB (Figure 3.7i), TCV (Figure 3.7ii) and CC (Figure 3.7iii) which overall were very similar with high and low AGB and TCV regions coinciding with high and low CC. The spatial distribution of these metrics, coupled with the authors' knowledge and observations, will be elaborated upon in detail in the discussion section (3.6). Figure 3.8 shows the AGB vs. CC scatterplot for AOI 'A' (Figure 3.7), a dense forested site. The point cloud generally displays a high correlation between the 2D (CC) and 3D (AGB) variable, but also a triangular shape with an increasing base as the CC increases up to 75% (highlight by the white labels in figure 3.8). Hence, dense cover conditions (CC>70%) are characterized by AGB values varying from moderate (35-40 t/ha) to high (>60 t/ha), corresponding to a range of tree sizes from coppicing thicket and medium sized tree bush encroachment to taller tree forests.



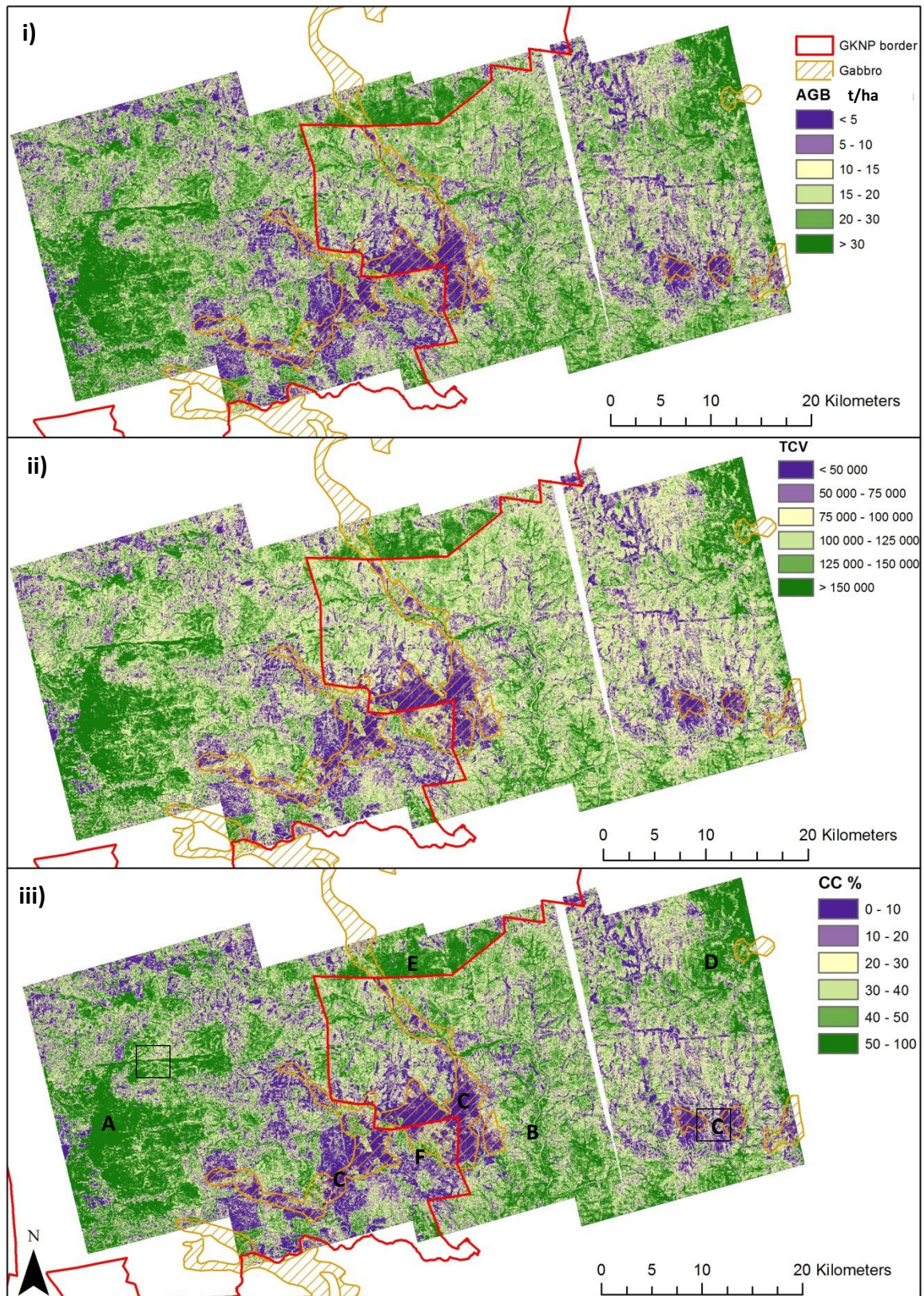


Figure 3.7: X+C+L SAR derived tree structural metric maps, for i) Above Ground Biomass (AGB), ii) Total woody Canopy Volume (TCV) and iii) woody Canopy Cover (CC), using random forest. Letters A-F represents key areas of interest for discussion (for all three metrics). The black boxes represent the rough extents of the LiDAR-SAR CC scenario difference maps for Area of Interests 'A' and 'C'.

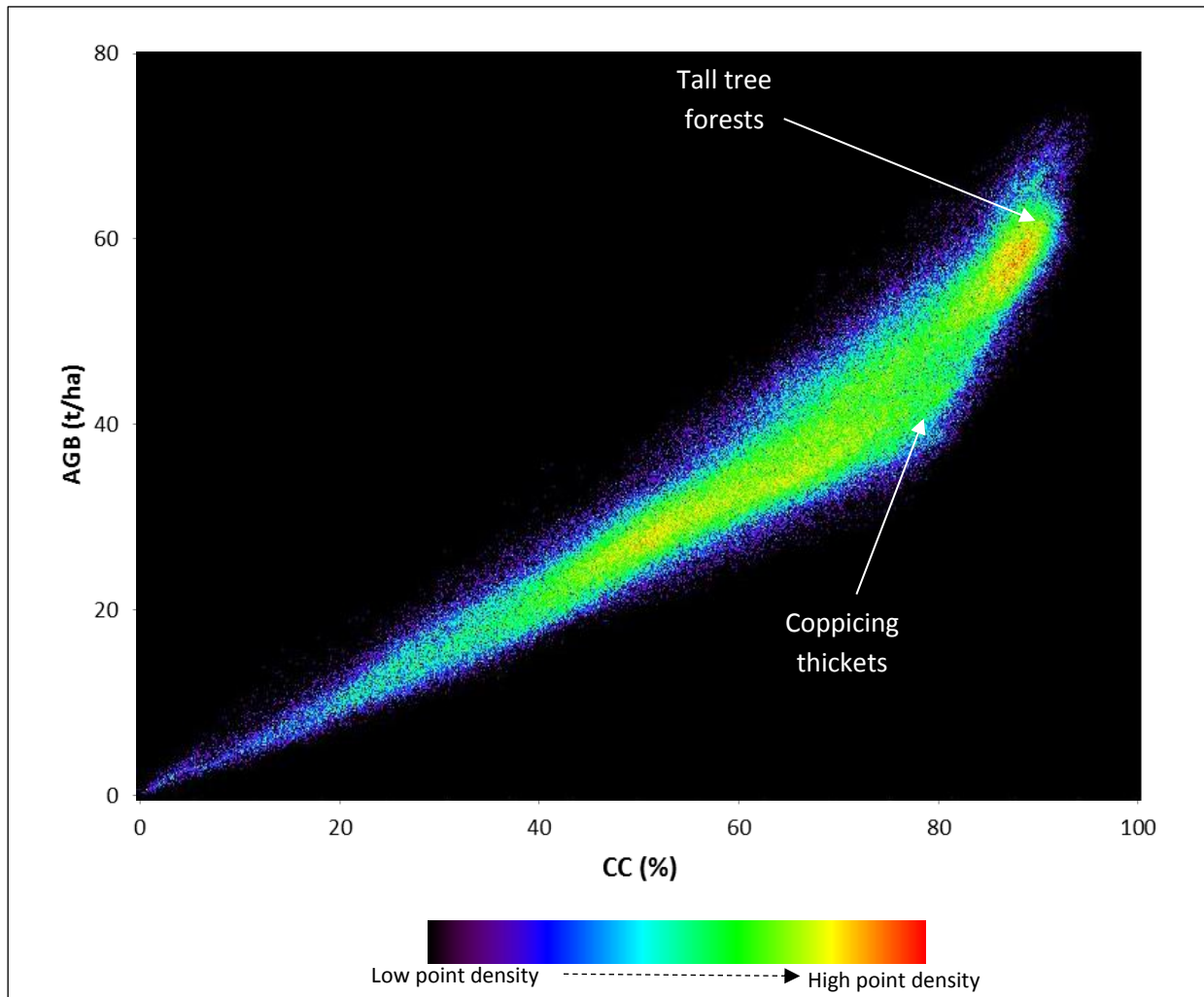


Figure 3.8: Scatterplot of Above Ground Biomass (AGB), y-axis, versus woody Canopy Cover (CC), x-axis, under dense cover conditions (plotted from pixels extracted from the Area of Interest 'A')

Examples of TCV error maps for dense forested (black box near 'A' in Figure 3.7iii) and sparse gabbro (black box over 'C' in Figure 3.7iii) sites were presented in Figures 3.9 and 3.10, respectively. Total CC, TCV and AGB error statistics were calculated to investigate the contributions of the four main SAR frequencies scenarios (X-band, C-band, L-band and X+C+L-band) to the modelling and mapping error (Table 3.4).

**Table 3.4: Total woody Canopy Cover (CC), Total Canopy Volume (TCV) and Above Ground Biomass (AGB) % error across the entire LiDAR-SAR coverage for the four main SAR frequency scenarios (Number of observations = 17559)**

<b>CC Error Classes</b>	<b>X-band Error</b>	<b>C-band Error</b>	<b>L-band Error</b>	<b>X+C+L-band Error</b>
Major overestimation (<-15%)	21.02	13.87	12.78	9.43
Minor overestimation (-15% to -5%)	17.30	16.38	16.74	16.85
<b>Negligible error (-5% to 5%)</b>	<b>19.52</b>	<b>24.58</b>	<b>31.34</b>	<b>31.84</b>
Minor underestimation (5% to 15%)	13.87	16.95	19.27	20.08
Major underestimation (>15%)	28.29	28.21	19.87	21.80
<b>TCV Error Classes</b>	<b>X-band Error</b>	<b>C-band Error</b>	<b>L-band Error</b>	<b>X+C+L-band Error</b>
Major overestimation (<-50k)	7.54	1.69	0.40	0.35
Minor overestimation (-50k to -10k)	28.58	22.96	22.32	18.57
<b>Negligible error (-10k to 10k)</b>	<b>4.64</b>	<b>8.26</b>	<b>15.56</b>	<b>16.62</b>
Minor underestimation (10k to 50k)	32.41	58.43	57.12	60.31
Major underestimation (>50k)	26.82	8.66	4.60	4.14
<b>AGB Error Classes</b>	<b>X-band Error</b>	<b>C-band Error</b>	<b>L-band Error</b>	<b>X+C+L-band Error</b>
Major overestimation (<-15t/ha)	4.53	1.95	0.79	0.65
Minor overestimation (-15t/ha to -5t/ha)	27.46	18.85	15.47	13.16
<b>Negligible error (-5t/ha to 5t/ha)</b>	<b>13.29</b>	<b>22.05</b>	<b>36.42</b>	<b>36.05</b>
Minor underestimation (5t/ha to 15t/ha)	25.07	41.00	37.24	39.70
Major underestimation (>15t/ha)	29.65	16.15	10.08	10.43



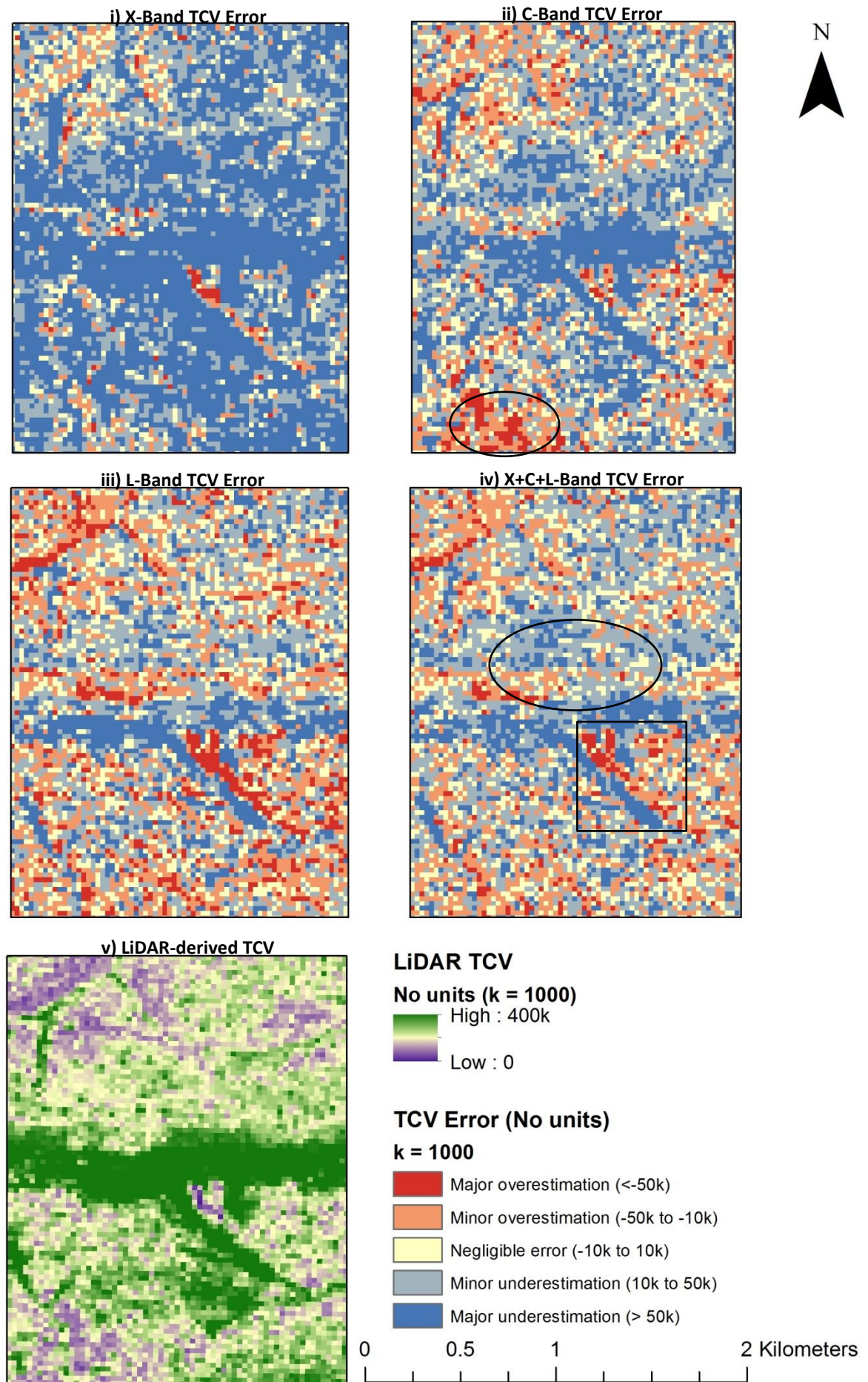


Figure 3.9: LiDAR - SAR scenario difference (error) maps (i-iv) of Total woody Canopy Volume (TCV) for the Xanthia Forest Area of Interest (close to 'A'); v) 25m LiDAR-derived TCV map



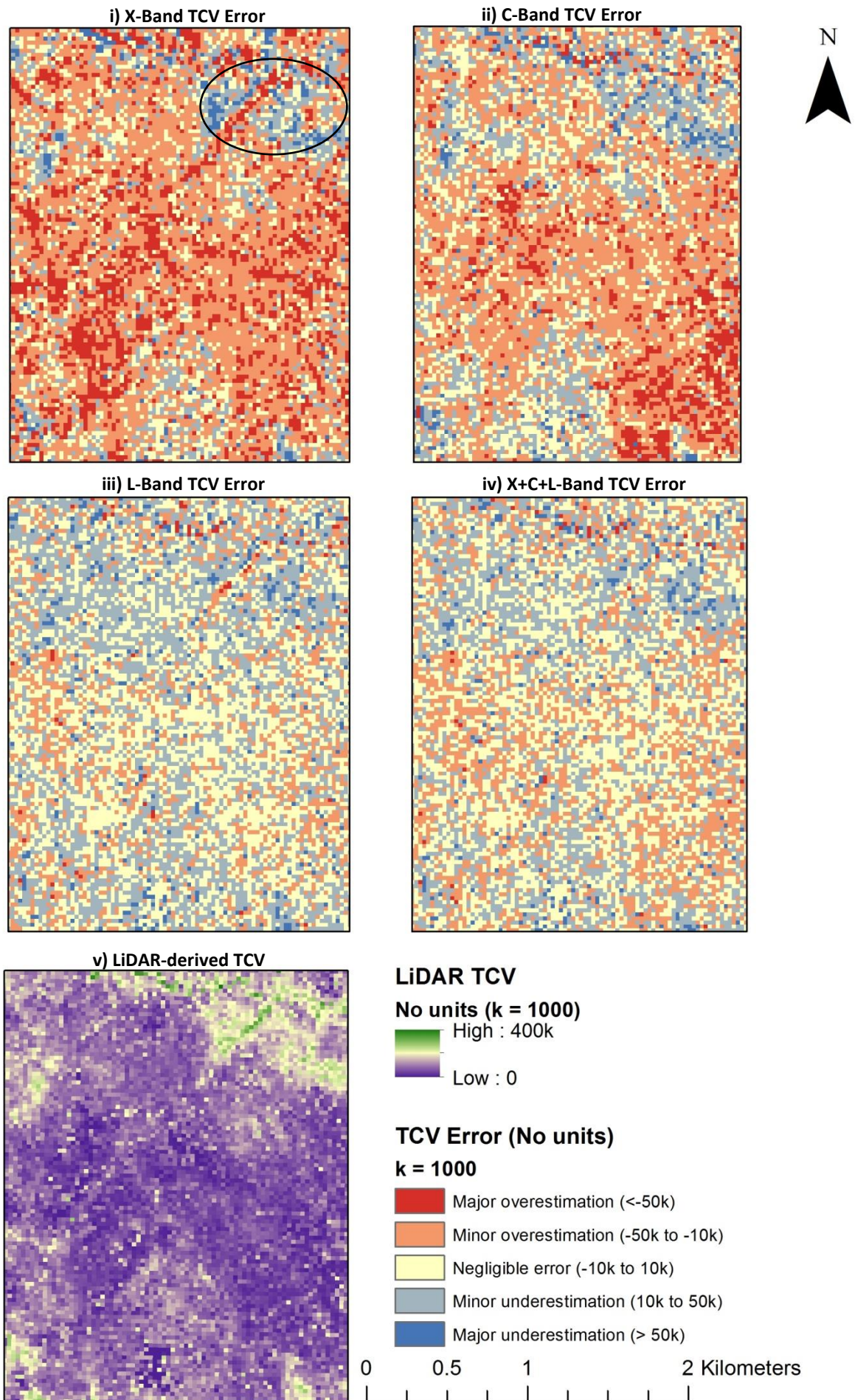


Figure 3.10: LiDAR - SAR scenario difference (error) maps (i-iv) of Total woody Canopy Volume (TCV) for the Gabbro Intrusions Area of Interest 'C'; v) 25m LiDAR-derived TCV map

In Table 3.4, there is a noticeable decline in major overestimation and major underestimation with an increase in negligible error for all three metrics from shorter wavelengths (X-band to C-band) to the longer wavelength (L-band). For all metrics, the X+C+L-band combined scenario further reduced major overestimation and marginally increased negligible error but at the cost of an increase in major underestimation in comparison to the L-band results. The TCV metric, under L-band and X+C+L-band scenarios, illustrated the most noticeable reduction in major overestimation and underestimation, in comparison to the other metrics, but at the cost of a higher percentage of minor underestimation (~60% between 10 000 to 50 000 TCV units). The greatest percentage increase in negligible error (-5t/ha to 5t/ha) was noticed in AGB metric for the L-band and X+C+L-band combined scenarios. More specifically for the TCV metric, under dense forested conditions (Figures 3.9i-v), the X-band scenario (Figure 3.9i) illustrate major TCV underestimation. C-band results (Figure 3.9ii) indicate an overall decrease of patches of major TCV underestimation but some of these have been replaced with major TCV overestimation across less dense patches of large trees (see encircled area in Figure 3.9ii). Further improvement is visible for the L-band scenario (Figure 3.9iii) with a noticeable increase in the minor TCV underestimation (10 000 to 50 000 TCV units) and negligible TCV error (evident in Table 3.4). Finally, the X+C+L scenario in Figure 3.9iv illustrated noticeable increases in the negligible TCV error coverage, especially over the dense green ridge visible in the LiDAR TCV of Figure 3.9v, but also indicated an increase in major TCV underestimation over dense vegetation patches north of the ridge (see encircle area in Figure 3.9iv). Patches of major TCV overestimation, however, still persist across riparian zones of minor tributaries (rectangle area in Figure 3.9iv). Under sparse vegetated conditions across gabbro intrusions (Figures 3.10i-v), however, X-band and C-band scenarios (Figures 3.10i and 3.10ii) indicate vast extents of major TCV overestimation for the sparse vegetation areas and major TCV underestimation for the dense forested patches (see encircled area in Figure 3.10i). The L-band scenario (Figure 3.10iii) illustrates a drastic improvement with an extensive increase in negligible TCV error across the Area of Interest (AOI). Across patches of dense vegetation, major TCV underestimation still persists (similar to the trend in Figure 3.9). The X+C+L-band scenario (Figure 3.10iv) also yields favourable results similar to the L-band scenario with no visible improvement. More quantitative results (box-plots, Figures 3.11i-ii) were introduced next to further assess the individual SAR frequency error contributions under different sparse and dense vegetation conditions.

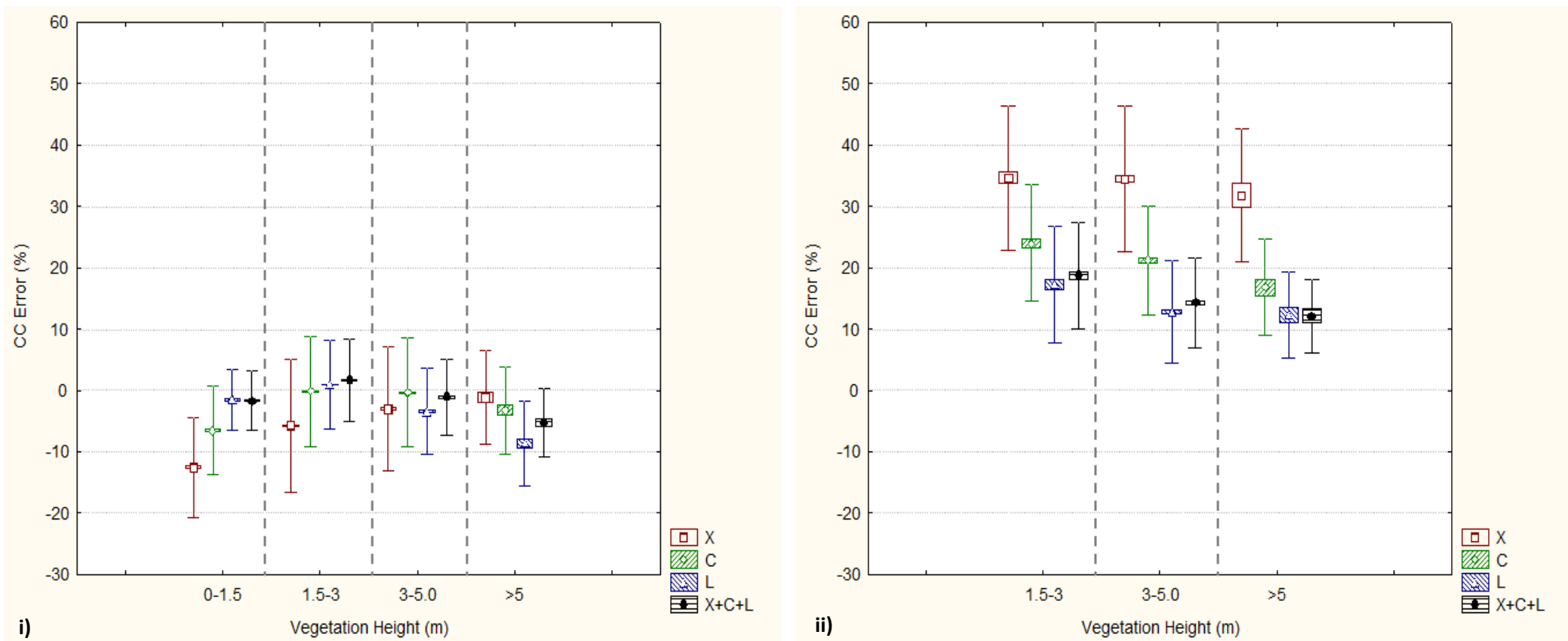


Figure 3.11: Woody Canopy Cover (CC) Error box plots of: i) low LiDAR CC (<40%) and variable LiDAR vegetation height and ii) dense LiDAR CC (>70%) and variable LiDAR vegetation height (+ve values = CC underestimation; -ve values = CC overestimation; dashed line partitions the four main SAR scenarios across the x-axis classes, centre point = mean value, box = standard error and whiskers = standard deviation) (Number of pixels = 17559)

CC error boxplots of the four main SAR frequency scenarios, Figure 3.11, were chosen to investigate error across vegetation structural types, classified from the LiDAR CHM, and including sparse shrubs (CC <40% and height <3m) or trees (CC <40% and height >3m) (Figure 3.11i), and dense forested (CC >70% and height >3m) or bush encroached (CC >70% and height <3m) conditions (Figure 3.11ii). In general, SAR derived CC is mostly overestimated across sparse vegetation but is underestimated across conditions of dense cover which coincides with the main trends of Figures 3.9i-v and 3.10i-v. The L-band scenario yielded the lowest overall CC errors (in terms of mean error or variance, or both) across both low levels of CC (<40%) and low height (<3m), and dense CC (>70%) across all height (<3m to >5m) in comparison to the X-band (highest variability and mean CC error) and C-band. Thus under sparse and low vegetation and bush encroaching conditions, it is the L-band which yields the lower levels of CC error and not the shorter wavelengths (X-band or C-band) as we may have expected. Also, the inclusion of the shorter wavelength datasets (X-band and C-band) with the L-band dataset led to minor improvements in the overall variability and mean of CC error across most sparse vegetation structural conditions (except regarding vegetation conditions with CC <40% and height >5m which is inconclusive) and across tall dense vegetation conditions (CC >70% and height >5m). Most significant improvement of the addition of the high frequency data occurred for the sparse and tallest trees (CC <40% and >3m) conditions.

### 3.6 Discussion

The modelling results indicated that it was the longer wavelength L-band dataset which contributed the most to the successful estimates of all three woody structural metrics. This finding agrees with other studies in the literature across a variety of ecosystem types such as coniferous forests (Dobson et al., 1992), boreal forests (Saatchi and Moghaddam, 2000) and temperate forests (Lucas et al., 2006a). The results obtained for the L-band can be attributed to its ability to penetrate deeper into the canopy, allowing the signal to interact the most with the larger tree constituents such as the trunk and branches (Mitchard et al., 2009), and thus produces stronger correlations with the LiDAR metrics. Despite the leaf-off conditions of most trees in winter, the shorter wavelengths (X- and C-band), 5.6cm for RADARSAT-2 and 3.1cm for TerraSAR-X, may have had a limited penetration of the canopy, and generally produced higher errors than the L-band for dense tree canopy (Figure 3.11ii). In the case of open woodlands (CC<40, Figure 3.11i), results suggest that some penetration did occur through the larger gaps with some good performance of C- and X-band compared to L-band (see tree height >3 m). However, C-band may have also been more sensitive to variability of surface roughness features (e.g. dense to sparse grass cover, fire scars etc.) which were too small to affect



the coarser L-band (Bourgeau-chavez et al., 2002; Menges et al., 2004; Wang et al., 2013). This interaction of the smaller wavelengths with these surface features may have introduced noise, which could have weakened correlations between the SAR signal and the LiDAR metrics.

The integration of the shorter wavelengths (e.g. X-band, C-band and X+C band), with L-band, yielded relatively small improvements in comparison to the L-band result alone (a reduction in SEP by ~3% and less for some metrics). The combination of all three frequencies yielded the highest overall accuracies for all metrics than each SAR frequency dataset alone. This result implies that the combination of short wavelength and long wavelength SAR datasets (X+C+L-band) does provide improved estimation in the modelling of the complete vegetation structure in terms of CC, TCV and AGB. As an aside to the modelling results, CC and AGB field data were initially investigated as a LiDAR-substitute for SAR model calibration and validation but preliminary results showed poorer modelling accuracies ( $R^2 < 0.60$ ) in comparison to the LiDAR derived results. This demonstrated the importance of extensive LiDAR coverage as the preferred source for modelling.

The three metric total percentage error statistics (Table 3.4), the TCV error AOI maps (Figures 3.9-10i-v) and the CC error box plots (Figures 3.11i-ii) reaffirmed the modelling accuracy observations but provided greater insight into the specific SAR frequency contributions to the overall prediction error under a variety of woody structural conditions. The use of L-band alone and its integration with the shorter wavelengths reduced the overall metric overestimation error (mean error and variability) under sparse vegetation conditions while reducing overall metric underestimation under dense vegetated conditions, in comparison to the shorter wavelengths alone and their combinations. These observations thus go against the first part of the main hypothesis made in this study which hypothesised the importance of shorter wavelengths for interaction with the finer woody structural elements and shrubby vegetation cohorts as L-band appears to be more effective in this regard. The incorporation of the shorter wavelengths with the L-band improved the overall metric error budget by reducing the overall mean error and the overall variability of the error under most vegetation structural conditions. Additionally, L-band and X+C+L-band were more suited for assessing the 3D metrics (TCV and AGB) than the single 2D metric (CC) with the highest percentage of negligible AGB error and lowest percentages of major TCV under- and overestimation being observed. These results can be supported by the fact that the L-band was expected to penetrate deeper and interact more with the lower levels of vegetation structure than the X- and C-band but the shorter wavelengths may have provided minor assistance to the L-band by interacting with the

smaller canopy elements (Rosenqvist et al., 2003). Further investigation will be needed to ascertain the exact cause of these trends but the overall results, however, advocate the suitability of the L-band over C- and X-band for analysing dense forested environments (>70% CC with an expected error ranging from ~7% to ~18%) and thus confirms the second part of the main hypothesis which stated that the L-band SAR signal interacts with the major tree structural components (e.g. trunk and main branches typical of forested areas) (Carreiras et al., 2013; Lucas et al., 2006a; Mitchard et al., 2012). In the absence of L-band data, C-band has proven to be effective in sparser cover, i.e. less than 40% CC, savannah environments which coincided with the recommendations made by (Mathieu et al., 2013).

Among the three structural metrics, TCV was consistently modelled with higher accuracies, amongst all seven SAR scenarios (Table 3.3). This result concurs with that of (Mathieu et al., 2013). TCV is a metric which indicates the volume of vegetation present within the vertical structure and its higher modelled accuracies could be attributed to the leaf-off conditions typical of the dry winter season which allowed for greater wave penetration into the canopy for all wavelengths, even the shorter wavelengths. CC and AGB metrics yielded similar  $R^2$  values with higher SEP values observed for AGB which may be due to the associated error propagated through the allometric equation and the LiDAR model (results of Figure 3.4). Since SAR is a system which utilises penetrating radio waves, the SAR signals will be expected to be more related to 3D structural metrics such as TCV and AGB rather than to the 2D CC metric (which achieved marginally poorer modelled results). This is due to the fact that CC is a metric for which the 2D horizontal coverage fluctuates seasonally depending on the phenological state of the vegetation, at least in comparison to TCV and AGB, which relies on the 3D nature of the woody structure which includes height and is thus more consistent across seasons (in the absence of disturbance).

The multi-frequency (X+C+L-band) model maps created for AGB (Figure 3.7i), TCV (Figure 3.7ii) and CC (Figure 3.7iii) illustrate patterns and distributions resulting from influence of numerous biotic (mega-herbivore herbivory and anthropogenic pressures such as fuelwood extraction and cattle ranching) and abiotic factors (fire regimes, geology and topographic features) relevant to the study area. In order to discuss the common patterns in CC, TCV and AGB in these maps, it will be collectively referred to as “woody vegetation”. Dense woody vegetation patterns are observed in the protected forested woodlands (Bushbuckridge Nature Reserve) and in the exotic pine plantations within the vicinity of A. Generally, the riparian zones of major rivers and tributaries (e.g. B, the Sabie

River catchment) have high values of CC, TCV and AGB compared to lower levels on the hill crests. In contrast to the vegetation occurring on granitic soils, the intrusions of the Timbavati gabbro geology group (Figure 3.7 C) have very low woody CC, TCV and AGB. These geological substrates naturally support more open landscapes than the more densely vegetated granite soils. Rangeland areas in and within the vicinity of informal settlements, such as Justicea (F), also showed lower levels of CC, TCV and AGB which could be linked to the heavy reliance of the local populace on fuelwood collection for energy requirements (Shackleton et al., 1994; Wessels et al., 2013, 2011). The area of interest E (Athole area which consisted of historical rotational grazing camps which are currently inactive – Barend Erasmus, personal communication, 27/02/2013) possesses a sharp fence line contrast in tree structure between the dense woody vegetation evident in the northern extents of Athole (i.e. north of fence) and the sparse woody vegetation in Sabi Sands Private Game Reserve (i.e. south of fence). The extended absence of grazing and browsing pressures in the old pasture and paddock enclosures in the northern reaches of the Athole fence line boundary (Figure 3.7 E) caused dense woody vegetation which contrasted sharply with the sparser woody vegetation in the more open and highly accessed areas south of the fence boundary. Additionally, the dense woody vegetation associated with the *Acacia welwitschii* thicket which dominates the ecca shales geological group of Southern Kruger National Park (outside map extents) was clearly visible at D (Mathieu et al., 2013). In conclusion, the accuracy and credibility of these maps and their trends have been supported by the various observations made during field visits and by the authors' general knowledge of the study area. The general range of these tree structural metric values also agreed with the ranges reported in other related studies conducted in this savannah region (Colgan et al., 2012; Mathieu et al., 2013).

Although overall modelling and mapping results yielded favourable accuracies, it is, however, important to acknowledge the different sources of error which were introduced in this study. The first error source was the temporal difference between the acquisition of the SAR predictor datasets and the reference datasets such as collected field data and/or LiDAR datasets. This was unavoidable due to sensor failure (e.g. ALOS PALSAR in early 2011) and logistical restrictions to the current research project (e.g. specific RADARSAT-2 datasets available from collaborations). Although there has been documented evidence of big tree loss in the study region (Asner and Levick, 2012), no major error was observed in the modelling results, especially when the 2010 L-band model was trained and validated using 2012 LiDAR data which produced expected results for this environment (Colgan et al., 2012; Mathieu et al., 2013). This loss in trees which occurred during the different SAR dataset acquisitions times (between 2009 and 2012) may have also introduced a certain margin of

error in the modelling results. It was expected, however, that the main structure of the remaining vegetation would not have changed prominently enough to extensively vary backscatter target interactions between the different acquisition times. A final source of error was introduced by the fact that the LiDAR reference dataset, which was set to target woody canopies with complete foliage, was acquired during the wet-dry transition season where the senescence process had just started. This may have resulted in a distorted representation of the woody structural metrics expected on the ground. Understanding these sources of error will help improve future studies by promoting the creation of more accurate models.

### 3.7 Concluding Remarks

This study investigated the accuracy of modelling and mapping above ground biomass (AGB), woody canopy cover (CC) and total canopy volume (TCV) in heterogeneous South African savannahs using multi-frequency SAR datasets (X-band, C-band and L-band including their combinations). Various studies have implemented L-band SAR data for tree structural assessment in a savannah type environment (Carreiras et al., 2013; Mitchard et al., 2012) but the use of shorter wavelengths, such as C-band, have also been proven to perform relatively well (Mathieu et al., 2013). This study also served to compare the three SAR frequency datasets (X-, C- and L-band) in the same study region of (Mathieu et al., 2013) and is the first attempt in an African Savannah context. It was hypothesized that the shorter SAR wavelengths (e.g. X-band, C-band), since interacting with the finer woody plant elements (e.g. branchlets) would be useful for mapping the shrubby/thicket layer while the longer SAR wavelengths (e.g. L-band) would interact with larger vegetation elements such as major branches and trunks typical of forested areas (Mitchard et al., 2009; Vollrath, 2010). It was thus proposed that the combination of these different SAR frequencies would provide a better assessment of the savannah woody element than the individual SAR frequencies (Schmullius and Evans, 1997).

After reviewing all the modelling and error assessment results, it can be concluded the L-band SAR frequency was more effective in the modelling of the CC, TCV and AGB metrics in Southern African savannahs than the shorter wavelengths (X- and C-band) both as individual and combined (X+C-band) datasets. Although the integration of all three frequencies (X+C+L-band) yielded the best overall results for all three metrics, the improvements were noticeable but marginal in comparison

to the L-band alone. The results do not warrant the acquisition of all three SAR frequency datasets for tree structure monitoring. Further the addition of the shortest wavelengths did not assist in the overall reduction of prediction error specifically of the shrubby layer as hypothesized. With the recent launch of the ALOS PALSAR-2 L-band sensor, the use of such L-band based models will be critical for future accurate tree structure modelling and monitoring at the regional and provincial scale. The modelling results obtained from the C-band SAR frequency alone, however, does yield promising results which would make the implementation of similar models to the free data obtained from the recently launched Sentinel-1 C-band sensor (launched in April 2014) viable when L-band datasets are not available. Sentinel-1 data are as far as we know the only upcoming operational, free and open access SAR dataset available in the near future, especially in Southern Africa. Building up of seasonal / annual time series may also improve on the performance of single date C-band imagery. The inclusion of seasonal optical datasets (e.g. reflectance bands, vegetation indices and textures derived from Landsat platforms), which can provide more woody structural information, may also augment the modelling results.

As a way forward beyond this study, in order to reduce the error experienced in the AGB results (at field collection, LiDAR and SAR levels), new and more robust savannah tree allometric equations, with a greater range of representative tree stem and height sizes, will need to be produced but such efforts will require extensive ground level harvesting campaigns. Due to the success of this study, particularly the positive results using L-band SAR data, future work will seek to up-scale these results to greater regional and provincial areas using more extensive LiDAR calibration and validation datasets.

## Chapter 4: Integration of Optical and L-band Synthetic Aperture Radar (SAR) datasets for the assessment of woody fractional cover in the Greater Kruger National Park region

### 4.1 Abstract

Savannahs consist of mixed tree-grass communities and can be best described as an ecosystem possessing a continuous herbaceous and a discontinuous woody layer. The woody component has considerable impact on natural and anthropogenic processes; for instance it impacts the fire regime, biomass production, nutrient cycling, soil erosion and the water cycle of these environments while providing numerous ecosystem resources, such as fuelwood, building material and non-timber products, such as fruit and bark and roots which are used for medicinal purposes. Woody canopy cover or CC is the simplest two dimensional metric for assessing the presence of the woody component. Synthetic Aperture Radar (SAR) sensors are particularly well suited and extensively used for woody structural measurements, because it senses the canopy geometry to retrieve structural information while optical sensors, which have been used successfully in national CC monitoring programmes outside South Africa, relies mostly on an optimum contrast between the “greenness” of tree canopies and the grass or bare background for CC assessment. The objective of this study was to evaluate the accuracy of modelling CC using multi-temporal datasets of SAR (L-band ALOS PALSAR) and optical (Landsat-5 TM) sensor data, both independently and in combination, in a Random Forest modelling environment. This research was based on the assumption that the integration of optical and SAR sensor data will yield improved results by allowing for the extraction of more detailed structural information and reducing associated uncertainty than the individual datasets. Additional objectives saw the testing of Landsat-5 image seasonality for the preferred acquisition season and the inclusion of spectral vegetation indices and image textures, as possible optical enhanced predictors, for improved CC modelling. Due to its accuracy, extensive airborne Light Detection and Ranging (LiDAR) data was used for model training and validation. Results showed that Landsat-5 imagery acquired in the summer and autumn seasons yielded the highest single season modelling accuracies using RF, depending on the year but the combination of multi-seasonal images yielded higher accuracies ( $R^2$  between  $\sim 0.6-0.7$ ). The derivation of spectral vegetation indices and image textures and their combinations with optical reflectance bands provided minimal improvement with no optical-only product combination yielding accuracies greater than winter SAR L-band backscatter alone ( $R^2$  of  $\sim 0.8$ ). However, there was significant, yet modest, improvement ( $R^2$  of  $\sim 0.08$ ,  $\sim 1.9\%$  of RMSE and  $\sim 7.5\%$  of SEP) in accuracy when 2010 multi-

seasonal optical reflectance bands were combined with the L-band backscatter variables. This research shows that considering the importance of savannahs in the region, future monitoring of woody canopy cover will require priority access to L-band SAR imagery from planned missions such as SAOCOM, TerraSAR-L, and NISAR. However, it is recommended by the authors that these results be verified in other bioregions, especially those dominated by evergreen canopies such as indigenous forest, thickets, and plantations. Finally, the integration of seasonally appropriate and cloud-free Landsat-5 image reflectance and L-band HH and HV backscatter data does provide a significant improvement for CC modelling at the higher end of the model performance.

**Keywords:** *Woody canopy cover, SAR, LiDAR, Landsat-5, textures, spectral vegetation indices, Random Forest*

## 4.2 Introduction

Savannahs consist of mixed tree-grass communities and can be best described as an ecosystem possessing a continuous herbaceous and a discontinuous woody layer (Sankaran et al., 2008). Savannahs cover half of the African continent and occupy one fifth of the global land surface (Scholes and Walker, 1993). The woody component has considerable impact on natural and anthropogenic processes, for instance it impacts the fire regime, biomass production, nutrient cycling, soil erosion and the water cycle of these environments (Sankaran et al., 2008) while providing numerous ecosystem resources, such as fuelwood, building material and non-timber products, such as fruit and bark and roots which are used for medicinal purposes (Shackleton et al., 2007; Twine, 2005). At the regional scale, the quantification of carbon captured in woody plants also plays an important role in understanding the global carbon cycle and fluxes between carbon sinks and sources (Valentini et al., 2014; Viergever et al., 2008b). Monitoring regional woody resources is essential to its sustainable management, which is threatened by adverse activities, such as deforestation, excessive fuelwood extraction and charcoal production (Shackleton et al., 1994; Wessels et al., 2013).

The woody component can be represented by a variety of woody structural parameters such as vegetation height, fractional cover, above ground biomass, basal area and canopy volume. Woody canopy cover is the simplest two dimensional metric for assessing the presence of the woody component and can be defined as the area vertically projected on a horizontal plane by woody plant canopies (Jennings et al., 1999). When expressed as a percentage per unit area, this metric is

referred to as fractional canopy cover or CC. CC can be combined with canopy height, to provide an informative indicator of volume and serve as a direct proxy for biomass (Colgan et al., 2012). In complex environments such as the heterogeneous savannahs of Southern Africa, CC also varies considerably across a variety of structural classes (e.g. from tall closed forests to short closed, bush encroached shrubs to sparsely distributed tall trees with a short shrub understory – (Edwards, 1983)). In South Africa and southern Africa there is no locally calibrated and validated national maps of CC, despite it being recognised as an Essential Biodiversity Variable by the international research community (Pereira et al., 2013).

In contrast to the limited spatial scope of ground based techniques, remote sensing is considered as the most appropriate tool for assessing woody structure across large areas. This is due to its ability to sense the high spatio-temporal variability, species diversity and phenological status, over large geographical scales – a defining but challenging set of characteristics typical of African Savannahs (Archibald and Scholes, 2007; Cho et al., 2012a). Synthetic Aperture Radar (SAR) sensors are particularly well suited and extensively used for woody structural measurements, because of their capacity to capture within-canopy properties (Collins et al., 2009; Le Toan et al., 2011; Santoro et al., 2007; Sun et al., 2011). SAR sensors are useful to regional scale studies due to their all-weather capabilities and lack of sensitivity to dense cloud cover and hazy conditions (e.g. fire smoke) which limit optical data acquisitions (Mitchard et al., 2011). Among the different available SAR frequencies, the L-band (a longer wavelength between 15 and 30cm) has been proven to be the preferred wavelength ((Carreiras et al., 2013; Mitchard et al., 2012; Ryan et al., 2011; Santos et al., 2002) and most effective in estimating woody structure in forests and savannahs (Lucas et al., 2006a; Naidoo et al., 2015). This is due to the fact that the signal of longer SAR wavelengths (e.g. P-band and L-band) can penetrate deeper into the vegetation and can interact with the major constituents of vegetation such as the main branches and trunks (Mitchard et al., 2009). Recent research in southern African savannahs showed that SAR can also provide a good performance to retrieve CC, especially L-band imagery (Mathieu et al., 2013; Naidoo et al., 2015). SAR backscatter signal, on the other hand, can be influenced by the variability in soil and canopy moisture, and by the variability in surface roughness, which may hamper woody canopy assessment in a particular environment (Bucini et al., 2009).

Although not known to be adept in sensing three dimensional vegetation structure (e.g. biomass where reflectance saturates readily), multi-spectral optical sensors (with visible, near- and mid-



infrared spectral coverage) are well suited for mapping two dimensional structure such as canopy cover at various spatial scales, and in dense tropical forests (Foody et al., 1997; Hansen and Loveland, 2012; Hansen et al., 2008), savannahs (Armston et al., 2009; Boggs, 2010; Lehmann et al., 2013) and finally shrublands and grasslands (Purevdorj et al., 1998; Ramsey et al., 2004). In contrast with the SAR technology which senses the canopy geometry to retrieve structural information, the mapping of canopy cover with optical sensors relies mostly on an optimum contrast between the “greenness” of tree canopies and the grass or bare background. Thus, the investigation and use of the time period at which a maximum contrast is achieved between green tree canopy and dry grass during the annual vegetation cycle is important (Zeidler et al., 2012). Textural image products, which provide information regarding the local variance, can be used as a measure of the canopy roughness, gaps, and associated shadow. In addition, non-parametric classification algorithms and spectral unmixing have been implemented for extracting fractional canopy cover at the regional scale (Chen et al., 2004; Foody et al., 1997; Lu, 2006; Nichol and Sarker, 2011). Optical sensor technologies with especially medium to coarse spatial resolutions of  $\geq 30\text{m}$ , however, can be limited in that they are highly influenced by spectral variation in time and space, mixed pixels and are obscured by cloud and shadow (Lu, 2006). Nevertheless, these optical sensor technologies have been adopted into successful national programmes for monitoring temporal woody canopy cover changes. These include the Australian Statewide Landcover and Trees Study (SLATS) (Armston et al., 2009) and the Australian National Carbon Accounting System – Land Cover Change Program (NCAS-LCCP) (Lehmann et al., 2013) which utilised Landsat TM and ETM+ data. Another programme also included the Landsat Ecosystem Disturbance Adaptive Processing System (LEDAPS) for monitoring North American forest disturbance using Landsat and ASTER datasets (Hansen et al., 2013; Ju et al., 2012). Finally, the Amazon Deforestation Monitoring Project (PRODES) which maps deforestation in the Amazon using Landsat datasets (Hansen and Loveland, 2012). Unfortunately, such national programmes are not in place for the savannahs of Southern Africa, despite a very large reliance on their ecosystem services (Scholes and Biggs, 2004; Wessels et al., 2013). The ultimate purpose of this research is to identify the possible contribution of Landsat to develop a national system for monitoring CC in South African savannahs.

Given the sensitivity of optical sensors to photosynthetically active vegetation and the sensitivity of SAR backscatter to vegetation structure, their possible integration could yield improved woody structure estimates via the provision of complementary information which neither sensor type could provide in isolation. The integration of SAR and optical technologies for woody structure assessment have been successfully applied in previous studies (Lucas et al., 2006b; Miles et al., 2003;

Moghaddam et al., 2002), which included dense forested environments, savannahs and plantations (Bucini et al., 2009; Shimabukuro et al., 2007; Wang and Qi, 2008), with reasonable accuracies ( $R^2 > 0.60$ ). Unfortunately, none of these studies have taken into account the effects of phenology on optical imagery, especially in savannah environments with complex tree and grass phenological seasonal changes. With this in mind, the objective of this study was to evaluate the accuracy of modelling CC, at the 30m spatial resolution, using multi-temporal datasets of SAR (L-band ALOS PALSAR) and optical (Landsat-5 TM) sensor data, both independently and in combination. Airborne LiDAR data recorded using the Carnegie Airborne Observatory (CAO) Alpha system (Asner et al., 2007) was used as a training and validation dataset. This research was based on the premise that the integration of optical and SAR sensor data will yield improved results by allowing for the extraction of more detailed structural information and reducing associated uncertainty than the individual datasets (Roberts et al., 2007). There were two main sets of research questions in our study. The first set of questions focused on how the accuracy of CC predictions compared when using Landsat versus L-band dual-polarised SAR input data, whether the integration of additional optical predictor features (e.g. textures and vegetation indices) improved modelling accuracies in comparison to the L-band SAR-based CC results and, finally, whether the integration of optical Landsat and L-band SAR data yielded any noticeable improvements in CC modelled predictions. The second research question sought to ascertain the season or seasons in which Landsat-5 data predicted CC with the highest accuracies. This question is related to the fact that savannah vegetation undergoes distinct seasonal phenological changes during which the green fractional cover of grasses and woody plants varies considerably (Fuller et al., 1997; Scholes and Archer, 1997). We hypothesized that the season when trees are completely covered in green foliage, while grasses are dry, should be the best period to retrieve CC, since there is limited interference by green grass (Archibald and Scholes, 2007). The identification of phenologically optimised optical imagery may improve CC estimation, when integrated with SAR data, in these heterogeneous savannahs where there is general dearth of such studies.

This paper is structured into four main sections. The first section (4.3) outlined the study area and associated landscape features and climate. The second (4.4) outlined the main methodological steps taken which included the outlining and pre-processing of the different remote sensing datasets utilised, the integration of these datasets and modelling scenarios, the modelling algorithm used and accuracy assessment and CC mapping. The third section (4.5) displayed the study's main findings while the fourth and final section (4.6) discussed these findings within context of multi-temporal changes in phenology, Landsat acquisition times and reliable regional monitoring applicability.

### 4.3 Study Area

The region under study includes the southern portion of the Greater Kruger National Park Region, South Africa, which falls between approximately 23° 39'S to 25° 19'S and 30° 57'E to 32° 11'E. This region consists of the mixture of communal rangelands (Bushbuckridge Municipality District), private game reserves (Sabi Sands) and national or provincial parks (southern Kruger National Park, Andover) (figure 4.1). The region covers an extensive range of geologies (e.g. granite, basalt, gabbro, tonalite, shale etc.), vegetation types (plantations to Clay Thorn Bushveld, Mixed Bushveld, Sweet Lowveld Bushveld and Open Grassland - (Mucina and Rutherford, 2006)), rainfall (mean annual precipitation of 1200mm in the west to 550mm in the east- (Shackleton, 2000)), management regimes (communal and protected) and disturbance regimes (fire, elephant damage, grazing and browsing patterns of herbivores and fuelwood harvesting).

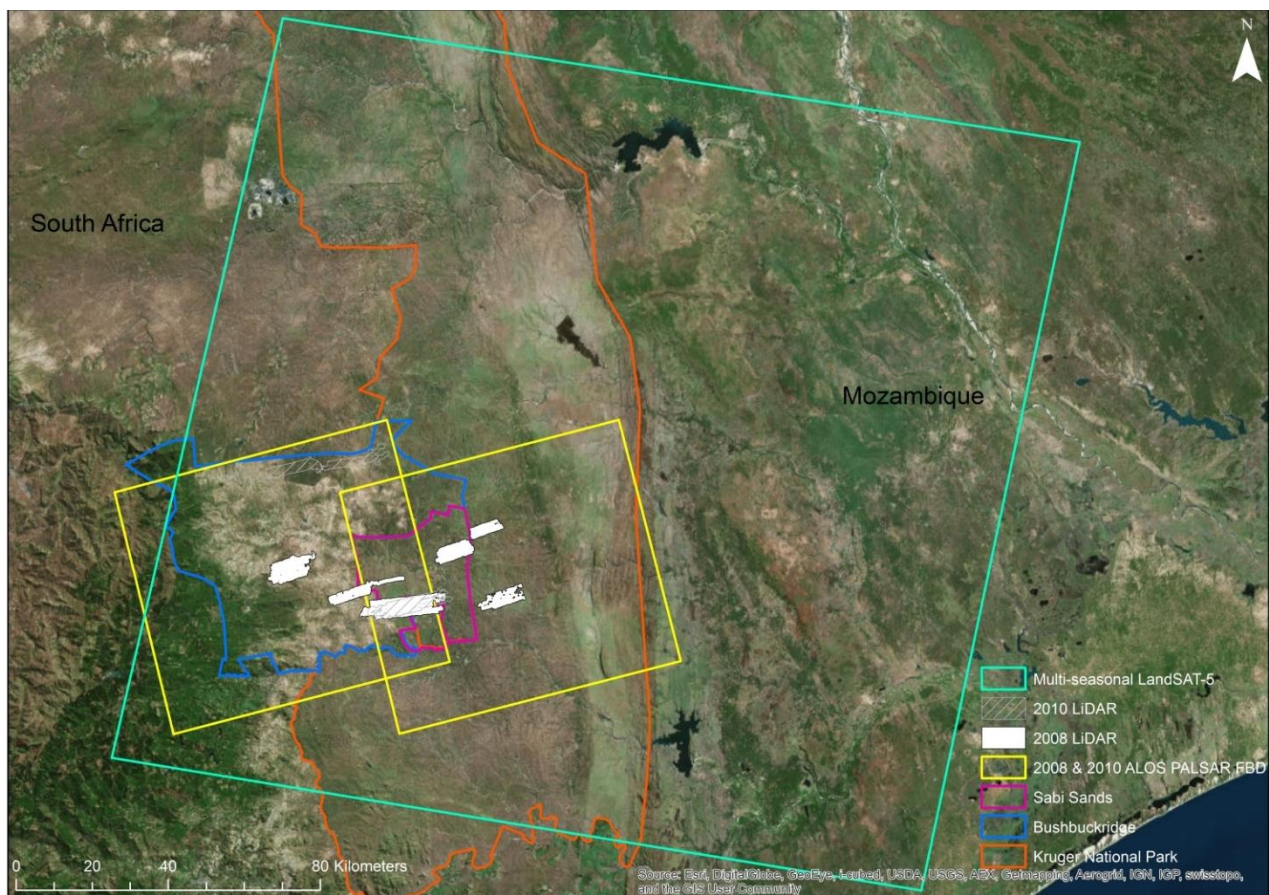


Figure 4.1: The Southern Kruger National Park study area and coverage of remote sensing modelling datasets

## 4.4 Materials and Methodology

Various scenarios were used to predict CC to determine the respective contribution of the Landsat and SAR-based variables. CC derived from very high resolution airborne LiDAR data were used as training and validation of the models. Firstly, models were developed to predict CC using reflectance data extracted from Landsat-5 images acquired at different seasons. These Landsat-based modelled results were compared to L-band SAR-derived models using ALOS PALSAR dual polarised (HH and HV) as input data. Second, the best performing Landsat-5 reflectance model was then expanded to include combinations of additional input variables including image texture features and vegetation indices. Finally, the integration of both multi-temporal optical Landsat reflectance and L-band SAR datasets were assessed for the possible improvement in CC prediction. All modelling scenarios were implemented using a Random Forest (RF) non-parametric machine learning algorithm (Breiman, 2001).

### 4.4.1 Remote Sensing Data

A collection of 2008 and 2010 dual polarised (HH, HV) ALOS PALSAR L-band intensity scenes and multi-seasonal Landsat-5 (bands 1-7, excluding the thermal band 6) scenes were collected over the study region (table 4.1). The L-band imagery (2 images for each year) was acquired in winter (25<sup>th</sup> of August and 23<sup>rd</sup> September (very early spring while landscape is dry and leaf-off) 2008; 14<sup>th</sup> and 31<sup>st</sup> August 2010) when the environment was dry and the trees devoid of leaves. These were shown to be the best conditions to extract CC with RADARSAT-2 C-band data in the same region (Mathieu et al., 2013). Landsat-5 scenes were inventoried from 2007 to 2011 (to match the LiDAR dataset available in 2008 and 2010, with an acceptable difference of plus and minus one year) and acquired in various seasons to assess the potential effects of differential phenology between trees and grasses. Specifically, Landsat-5 imagery were acquired for spring (September- November), summer (December - March), autumn (April - May) and winter (June - August), where available, of 2007, 2008, 2009 and 2010 from U.S Geology Survey Landsat Earth Explorer portal (along path 168 and row 77). In summer, both tree leaves and grasses are green while in autumn, grasses are dry with trees remaining green but beginning to lose leaves. In winter, most trees have lost leaves and grasses are dry while in spring, grasses are fairly dry while the trees first undergo a green flush of leaves (Archibald and Scholes, 2007). Only Landsat-5 imagery with an overall scene cloud cover of  $\leq 6\%$  was considered. Due to cloud occurrence one image was available at each season only in 2008; three seasons were achieved in 2007 and two in 2009 and 2010 (Table 4.1). No suitable Landsat-5

imagery was available for the year 2011 and was thus not included in the analyses. Several years were considered to assess the possible model inconsistencies which may result from a high inter-annual variability of rainfall, and associated variability of greenness and phenology. Extensive airborne 2008 and 2010 LiDAR dataset (total coverage of c.a. 35000 ha and 10000 ha respectively) were acquired for this study (figure 4.1) by the Carnegie Airborne Observatory (CAO) Alpha sensor (Asner et al., 2007) during April-May of 2008 and 2010.

**Table 4.1: Landsat-5, ALOS PALSAR and LiDAR data inventory**

Sensor	scene ID	Season	Date of Acquisition
Landsat-5 TM	LT51680772007047JSA00	Summer	16/02/2007
Landsat-5 TM	LT51680772007143JSA00	Autumn	23/05/2007
Landsat-5 TM	LT51680772007175JSA00	Winter	24/06/2007
Landsat-5 TM	LT51680772007223JSA00	Winter	11/08/2007
Landsat-5 TM	LT51680772008034JSA01	Summer	03/02/2008
Landsat-5 TM	LT51680772008098JSA01	Autumn	07/04/2008
Landsat-5 TM	LT51680772008242JSA00	Winter	29/08/2008
Landsat-5 TM	LT51680772008274JSA02	Spring	30/09/2008
Landsat-5 TM	LT51680772009084JSA00	Summer	25/03/2009
Landsat-5 TM	LT51680772009132JSA00	Autumn	12/05/2009
Landsat-5 TM	LT51680772010023JSA00	Summer	23/01/2010
Landsat-5 TM	LT51680772010119JSA00	Autumn	29/04/2010
ALOS PALSAR	ALPSRP137816680	Winter	25/08/2008
ALOS PALSAR	ALPSRP142046680	Spring	23/09/2008
ALOS PALSAR	ALPSRP242696680	Winter	14/08/2010
ALOS PALSAR	ALPSRP245176680	Winter	31/08/2010
CAO LiDAR	CAO 2008	Autumn	April-May 2008
CAO LiDAR	CAO 2010	Autumn	April-May 2010

#### 4.4.2 LiDAR Data Processing

1.1m Digital Elevation Models (DEM) and top-of-canopy surface models (CSM) were created by processing the raw 2008 and 2010 LiDAR point clouds using REALM (Optech Inc., Vaughn, Canada) and TerraScan/TerraMatch (Terrasolid Ltd., Jyväskylä, Finland) LiDAR software. Canopy height models (CHM, pixel size of 1.12m) were computed by subtracting the DEM from the CSM. The 2008 and 2010 LiDAR fractional woody canopy cover metric were then created by first applying a data mask to the LiDAR CHM image in order to create a spatial array of 0s (no woody canopy) and 1s (presence of a woody canopy). Fractional woody canopy cover distribution products were calculated at 25m spatial resolution using equation 4.1:

$$LiDAR\ CC\ (\%) = \frac{\sum(1's)}{625} \times 100 \quad \text{Equation 4.1}$$

Where 625 is the area (in m<sup>2</sup>) of a 25m X 25m pixel. A height threshold of 0.5m was applied to the CHM in order to avoid the inclusion of the grass layer in final product. The 2008 CAO LiDAR data were validated against field height measurements of approximately 800 trees. There was a strong relationship ( $r^2 = 0.93$ ,  $p < 0.001$ ), and only a fraction of woody plants below 1.5-1.7m were not detected by the LiDAR (Wessels et al., 2011). Additionally, a LiDAR derived woody canopy cover product obtained from a new LiDAR campaign done in 2012 correlated well with ground CC data collected from 37 25m X 25m sites in May/April 2012 ( $R^2=0.79$ ; Root Mean Square Error=12.4%) and thus this CAO LiDAR sensor technology was considered adequate for calibration and validation dataset extraction for this study.

#### 4.4.3 SAR Data Processing

The 2008 and 2010 level 1.1 ALOS PALSAR L-band intensity datasets (HH, HV) were processed in GAMMA™ SAR remote sensing software in which a script was developed to achieve the following steps: multi-looking, radiometric calibration (from raw digital numbers to sigma nought backscatter), geocoding and topographic normalization. Multi-looking factors of 2 and 8 was applied to the range and azimuth, respectively, to best remove unwanted speckle and distortions. This was sufficient to have the majority of the speckle removed, while preserving image detail, and hence no filtering was applied. A 20m DEM was used for the geocoding and topographic normalization. It was computed from 1:50 000 South African topographic maps (20m digital contours, spot-heights, coastline and inland water area data – ComputaMaps; [www.computamaps.com](http://www.computamaps.com)) with Root Mean Square planimetric error of 15.24m and a total vertical RMS error of 6.8m. As a final step the imagery was resampled, via bicubic-log spline interpolation function, by using a DEM oversampling factor of 1.6, to achieve a fixed spatial resolution of 12.5m to create images with a finer spatial detail.

#### 4.4.4 Landsat-5 Optical Data Processing and Derived Products

The Landsat imagery, in raw digital number format, underwent atmospheric correction with the use of ATCOR 2 (Multi-spectral sensor atmospheric correction for flat terrain) which converted the raw digital number data to top of canopy (TOC) reflectance using a Modtran®-5 radiative transfer code. The necessary information (e.g. Min and Max radiance values) from the default post May 2003 calibration file was used. Dry rural, fall (spring) rural, mid-latitude summer and winter rural



atmospheric models were also utilised with the visibility distance set between 9.0km and 59km depending on the season and year (historical Skukuza visibility data obtained from <http://weatherspark.com> were used if no values were automatically recommended by ATCOR).

The TOC reflectance of the individual images was used as the main model input variables to be tested. However, additional vegetation indices and image textures were derived from the best performing Landsat seasonal image for further analyses. This included a number of grey-level co-occurrence matrices (GLCM) and spectral vegetation indices (e.g. Enhanced Vegetation Index or EVI and Soil Adjusted Vegetation Index or SAVI) which have been known to be sensitive to vegetation structure (table 4.2). The selected vegetation indices which use the red, near-infrared and mid-infrared bands were also effectively correlated with the vegetation structure of various forested and woodland environments (Cohen et al., 2003; Freitas et al., 2005; Zheng et al., 2004). The soil-adjusted vegetation index (SAVI) was included over other more common indices (e.g. NDVI and single ratios) as it includes a soil adjustment factor which reduces sensitivity to soil and moisture conditions in the environment (Huete and Jackson, 1988; Jiang et al., 2008). As a more advanced vegetation index, the enhanced vegetation index (EVI) optimises the vegetation signal (especially in high biomass environments) by reducing the influence of atmospheric effects and the canopy background signal (Jiang et al., 2008). EVI is also known to be more linearly correlated to leaf area index (LAI), a major vegetation structural parameter derived from optical data, than other spectral indices. The non-linear vegetation index (NLI) was developed to account for the possible non-linear relationship between indices and biophysical parameters (Gong et al., 2003). Finally, the moisture vegetation index (MVI) was chosen as it possesses a higher signal saturation threshold especially in dense, high biomass environments (Freitas et al., 2005).

GLCM texture parameters, such as variance and entropy, were also selected as they were reported to be strongly correlated with vegetation structure (Asner et al., 2002; Nichol and Sarker, 2011) and in some case even better correlated than spectral indices (Lu, 2005). Preliminary results illustrated that variance, entropy, dissimilarity and contrast textures, derived from the bands 1 to 5 and 7, were particularly correlated with CC (results not presented). The combination of these selected indices and textures could provide more detailed structural CC information than the optical reflectance bands alone.

**Table 4.2: Reflectance, indices and textural optical products derived from Landsat-5 data**

Type	Product	Formulae or description if not applicable	Reference
Reflectance	Raw TOC reflectance	Band 1 (450-520nm) – Blue Band 2 (520-600nm) – Green Band 3 (630-690nm) – Red Band 4 (760-900nm) – NIR Band 5 (1550-1750nm) – MIR5 Band 7 (2080-2350nm) – MIR7	
Vegetation Index	Enhanced Vegetation Index (EVI)	$2.5 \times \frac{(NIR - Red)}{(NIR + (6 \times Red) - (7.5 \times Blue) + 1)}$	(Huete et al., 1997)
Vegetation Index	Modified Simple Ratio (MSR)	$\frac{\left(\frac{NIR}{Red}\right) - 1}{\sqrt{\frac{NIR}{Red} + 1}}$	(Sims and Gamon, 2002)
Vegetation Index	Non-linear Vegetation Index (NLI)	$\frac{NIR^2 - Red}{NIR^2 + Red}$	(Goel and Qin, 1994)
Vegetation Index	Soil-Adjusted Vegetation Index (SAVI)	$\frac{NIR - Red}{NIR + Red + 0.5} \times (1 + 0.5)$	(Huete and Jackson, 1988)
Vegetation Index	Simple Ratio (SR)	$\frac{NIR}{Red}$	(Jordan, 1969)
Vegetation Index	Normalised Difference Vegetation Index (NDVI)	$\frac{NIR - Red}{NIR + Red}$	(Rouse et al., 1973)
Vegetation Index	Moisture Vegetation Index (MVI band 7)	$\frac{NIR - MIR7}{NIR + MIR7}$	(Sousa and Ponzoni, 1998)
GLCM Textures	Variance, Entropy, Dissimilarity & Contrast (3 X 3 window)	Applied to bands 1-7	(Haralick et al., 1973)

TOC= Top of Canopy; NIR = Near Infrared; MIR = Middle Infrared

#### 4.4.5 Data Analysis Grid

To analyse the data of different resolutions, a fixed grid of 105m X 105m cells, with a 50m spacing to avoid spatial autocorrelation of CC, was used to extract SAR, optical and LiDAR CC products. The grid was created to match the extent of the LiDAR CC product coverage (i.e. the calibration/validation dataset for CC) and exclude any cells occupying water bodies, main roads, rivers and informal settlements and especially clouds (in the Landsat imagery). The resolution of the grid cells was supported by (Mathieu et al., 2013) and (Urbazaev et al., 2015) as the resolution which provided the best trade-off between the finest mapping resolution and strongest correlation with the LiDAR CC metrics. The extraction process was conducted in ENVI 4.8 where mean values for each cell in the grid were extracted. Due to the varying conditions of the different Landsat imagery (i.e. by way of cloud cover) and the differences in LiDAR coverage between 2008 and 2010, the total number of observations included in the modelling also varied and ranged between 1174 and 8804.

#### 4.4.6 Modelling Algorithms, Modelling Scenarios, Model Validation and CC Mapping

A random forest (RF) non-parametric machine learning algorithm (Breiman, 2001) was applied in the R rattle modelling software with 35% of the data being used for model training and the remaining



65% being used for model validation. Other well-known parametric algorithms, such as linear regression, and non-parametric algorithms, such as Support Vector machines (SVM), REP Tree decision tree and Artificial Neural Network, were also tested but preliminary results showed that RF consistently obtained higher modelling accuracies. Due to its use of multiple decision trees, bagging and internal cross-validation mechanisms, RF is seen as a major improvement over other traditional decision tree types and when compared to the other non-parametric algorithms. The algorithm is easy to implement and is robust as it only requires two main user-defined inputs (number of trees built in the 'forest' and the number of possible splitting variables for each node - (Ismail et al., 2010; Prasad et al., 2006)).

Before the final implementation of RF, efforts were made to test the generalisation of RF modelling by introducing an additional independent test dataset for model tuning before validation. During the tuning phase, the total number of trees ('ntree') in the forest and the RF tree complexity were varied to test their influence on accuracy whilst trying to limit the complexity of the RF model. RF tree complexity included the minimum number of terminal nodes ('nodesize') and the maximum number of terminal nodes that the trees can have in the forest ('maxnodes') (Breiman, 2001). After repeating the process three times, results showed that an 'unpruned' (i.e. no limitation on a tree's depth and number of terminal nodes) tree architecture with 200 trees within the forest, yielded the optimum results (refer to Appendix section; Figures 4A and 4B). In the light of these preliminary results the RF models was created based on the following parameters: 'ntrees' = 200 and 'mtry' =  $\sqrt{\#}$  SAR predictors (a rule of thumb for 'mtry' which was supported by (Liaw and Wiener, 2002)) with the trees being allowed to grow unpruned.

For the modelling process, several scenarios were assessed. The optical reflectance bands served as input variables which were tested individually (12 individual Landsat images) in order to ascertain the best season for predicting woody fractional cover. All available seasonal images were also combined for each year (four years in total) in order to investigate any improvements using multi-seasonal datasets. Seven additional scenarios using reflectance, texture and vegetation indices were also proposed in order to test the benefits of more advanced optical metrics. This was only performed for the best performing optical reflectance bands-only scenario mentioned above. 2008 and 2010 L-band SAR dataset-only scenarios served as the scenario of comparison for the optical-only tests. Due to the large number of vegetation indices and textures used in this study, which may display high degrees of co-linearity, a RF variable importance measure called the permutation

accuracy or %IncMSE (percentage increase in mean squared error) was considered to select the top three indices and texture variables for inclusion in the RF ‘Textures’ and ‘Indices’ modelling scenarios. %IncMSE records the percentage increase in the mean squared errors in the model when a particular variable is assigned random values while the remaining variables are left unchanged (Liaw and Wiener, 2002). The higher the resultant error, the more important that particular variable is to the model.

Finally, the SAR datasets were integrated with the five best performing seasonal Landsat-5 images and the combined multi-seasonal Landsat-5 datasets for each year to quantify the benefits of combining SAR and optical data for the modelling of CC. The RF validation results of the different scenarios were expressed in the form of coefficient of determination ( $R^2$ ), root mean square error (RMSE) and Standard error of prediction (SEP). SEP refers to the standard deviation of the prediction errors and is a measure of the unexplained variation of a model. The most accurate model, together with the most relevant independent variables, was implemented to produce a CC map. The ALOS PALSAR images were resampled to 30m spatial resolution (using pixel aggregated resampling) and clipped to fit the Landsat-5 image and stacked for mapping. The CC RF mapping was conducted using the Model-Map module of the R statistical software.

## 4.5 Results

### 4.5.1 Individual and multi-seasonal Landsat-5 reflectance compared to SAR

**Table 4.3: Individual seasonal Landsat-5, multi-seasonal Landsat-5 and individual SAR RF modelled CC validation results**

Dataset	Acquisition Date	Season of Imagery	$R^2$	RMSE (%)	SEP (%)	Total No. Obs*
Individual Landsat-5 TM	16/02/2007 <sup>1</sup>	Summer	0.47	12.64	52.02	8804
	23/05/2007 <sup>1</sup>	Autumn	0.34	13.96	58.46	8804
	24/06/2007 <sup>1</sup>	Winter	0.32	14.25	58.76	8804
	11/08/2007 <sup>1</sup>	Winter	0.32	14.10	58.69	8733
	03/02/2008 <sup>1</sup>	Summer	0.53	11.84	49.24	8804
	07/04/2008 <sup>1</sup>	Autumn	0.46	12.89	52.64	8010
	29/08/2008 <sup>1</sup>	Winter	0.37	13.60	56.73	8804
	30/09/2008 <sup>1</sup>	Spring	0.40	13.19	53.2	8339
	25/03/2009 <sup>1</sup>	Summer	0.44	12.76	52.86	8804
	12/05/2009 <sup>1</sup>	Autumn	0.50	12.04	49.6	8697
	23/01/2010 <sup>2</sup>	Summer	0.64	14.77	46	2098
	<b>29/04/2010<sup>2</sup></b>	<b>Autumn</b>	<b>0.65</b>	<b>13.55</b>	<b>44.43</b>	<b>3201</b>
Multi-seasonal Landsat-5 TM	2007 <sup>1</sup>	All available images	0.58	11.27	47.23	8733
	2008 <sup>1</sup>	All available images	0.64	10.53	43.31	8010
	2009 <sup>1</sup>	All available images	0.57	11.36	46.92	8697

	<b>2010<sup>2</sup></b>	<b>All available images</b>	<b>0.72</b>	<b>12.84</b>	<b>39.75</b>	<b>2098</b>
SAR	<b>25/08/2008<sup>1</sup></b>	<b>Winter</b>	<b>0.80</b>	<b>7.88</b>	<b>32.08</b>	<b>8804</b>
	<b>14/08/2010<sup>2</sup></b>	<b>Winter</b>	<b>0.81</b>	<b>10.17</b>	<b>33.16</b>	<b>3201</b>

\* Variable depending on LiDAR coverage per year (35% training; 65% validation) and LT cloud cover; <sup>1</sup> 2008 LiDAR dataset for the reference dataset; <sup>2</sup> 2010 LiDAR dataset for the reference dataset

When examining the individual seasonal Landsat-5 reflectance accuracies (table 4.3), the season which yielded the highest model accuracies varied between years; summer was best in 2007 and 2008, and autumn the best in 2009 and 2010. Amongst all the individual datasets, the April 2010 Landsat-5 reflectance (autumn) dataset yielded the highest model accuracies in comparison to the other individual images (according to R<sup>2</sup> and SEP values). The winter datasets that were available in 2007 and 2008 yielded the poorest modelled CC results. Overall the performance of single Landsat datasets was poor with a SEP varying between 44 and 58%. Combining all the multi-seasonal images for each year improved the accuracies by an RMSE of ~1-2% and SEP of ~4-6% compared to the best individual seasonal image for that year. However, both individual seasonal and combined multi-seasonal image yielded significantly lower accuracies than those of the individual SAR images. For instance, the SAR models produced in 2008 and 2010 had a SEP of 15 and 10% lower, compared to the best Landsat season of that specific year. Moreover, both SAR models produced consistent results, with a similar R<sup>2</sup> and SEP.

#### 4.5.2 Optical reflectance, textures and indices compared and integrated with SAR data results

**Table 4.4: Reflectance, indices and textural Landsat-5 (autumn 2010 image) product RF modelled CC validation results**

2010 Optical Product(s) <sup>1</sup>	R <sup>2</sup>	RMSE (%)	SEP (%)	Total No. Obs
Reflectance only	0.65	13.55	44.43	3201
Textures only*	0.03	23.66	77.96	3201
Indices only*	0.45	17.22	57.16	3201
Reflectance + Textures*	0.67	13.30	43.74	3201
Reflectance + Indices*	0.66	13.52	44.93	3201
Indices* + Textures*	0.47	17.06	55.87	3201
Reflectance + Textures* + Indices*	0.68	12.98	43.53	3201
2010 SAR only <sup>1</sup>	<b>0.81</b>	<b>10.17</b>	<b>33.16</b>	<b>3201</b>

<sup>1</sup> Utilized the 2010 LiDAR dataset as the reference dataset; \* Top 3 indices/textures used based on %IncMSE

**Table 4.5: Integrated SAR and best performing/multi-seasonal Landsat-5 reflectance RF modelled CC validation results (per year)**

Dataset	Acquisition Year	Season of Imagery	R <sup>2</sup>	RMSE (%)	SEP (%)	Total No. Obs
SAR + Best Landsat-5 TM	2007 <sup>1</sup>	SAR + Summer	0.84	6.89	28.73	8733
	2008 <sup>1</sup>	SAR + Summer	0.85	6.84	28.24	8010
	2009 <sup>1</sup>	SAR + Autumn	0.83	7.09	29.82	8697

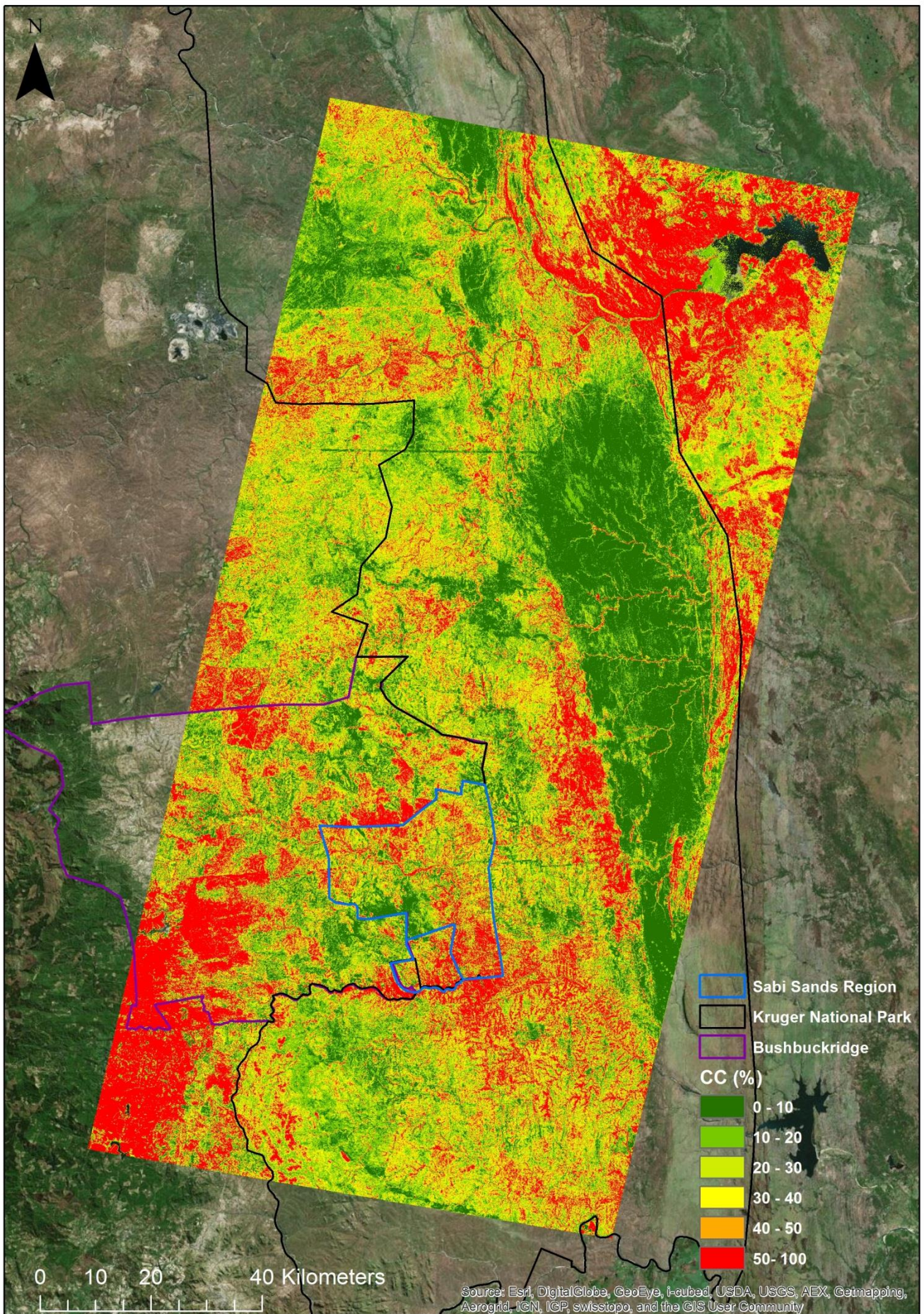
	<b>2010<sup>2</sup></b>	<b>SAR + Autumn</b>	<b>0.88</b>	<b>8.51</b>	<b>26.15</b>	<b>3201</b>
SAR + Optical* Products	<b>2010<sup>2</sup></b>	<b>SAR + Autumn</b>	<b>0.88</b>	<b>8.15</b>	<b>26.90</b>	<b>3201</b>
SAR + Multi-seasonal Landsat-5 TM	2007 <sup>1</sup>	All available images	0.85	6.75	28.37	8733
	2008 <sup>1</sup>	All available images	0.85	6.67	27.34	8010
	2009 <sup>1</sup>	All available images	0.84	6.91	28.79	8697
	<b>2010<sup>2</sup></b>	<b>All available images</b>	<b>0.89</b>	<b>8.32</b>	<b>25.64</b>	<b>2098</b>
SAR	<b>2008<sup>1</sup></b>	<b>Winter</b>	<b>0.80</b>	<b>7.88</b>	<b>32.08</b>	<b>8804</b>
	<b>2010<sup>2</sup></b>	<b>Winter</b>	<b>0.81</b>	<b>10.17</b>	<b>33.16</b>	<b>3201</b>

<sup>1</sup>Utilized the 2008 LiDAR dataset as the reference dataset and 2008 SAR dataset as one of input variables; <sup>2</sup>Utilized the 2010 LiDAR dataset as the reference dataset and 2010 SAR dataset as one of input variables; \* Optical Products refers to the Reflectance + Textures + Indices scenario in Table 4.4

Image textures and spectral vegetation indices (top 3 of each parameter selected according to the highest %IncMSE) were added as additional features to the best performing Landsat-5 reflectance dataset (April 2010 according to table 4.3) in order to determine if these improve the prediction of CC (table 4.4). The optical reflectance-only scenario yielded the best results, followed by the derived vegetation indices, and the textures-only produced by far the poorest results. However, the combination of reflectance and textures yielded marginally better results than the reflectance and indices combination which suggested that image textures do provide more additional information in comparison to the indices. Combining all three datasets (reflectance, textures and indices) provided the highest overall accuracy, however improvement was marginal compared to the optical reflectance-only scenario. Although not presented here, in the interest of brevity, these results were consistent for other years (2007, 2008 and 2009). Combining the best seasonal Landsat-5 reflectance dataset per year with SAR data brought about modest, but significant improvements (improved SEP of ~4-5%) in the modelled CC accuracies for the individual years in comparison to SAR-only scenarios (table 4.5). Also, the difference in accuracy between the best seasonal reflectance and combined multi-seasonal images, integrated with SAR datasets, were minimal (improved SEP of 0.5-1%). The year 2010 obtained the highest accuracies, ( $R^2=0.89$ ;  $RMSE=8.32\%$ ;  $SEP=25.64\%$  for the integrated SAR and multi-seasonal dataset). The combination of 2010 SAR data with 2010 Autumn Landsat-5 reflectance and the three most important vegetation indices and textures did not improve the combined 2010 SAR and 2010 Autumn Landsat-5 reflectance results. The best trade-off between accuracy and complexity were given by the 2010 integrated SAR and autumn season reflectance model ( $R^2=0.88$ ;  $RMSE=8.51\%$ ;  $SEP=26.15\%$ ), as it used a single SAR and single Landsat-5 image. This model was therefore used to create the regional CC map (figure 4.2).

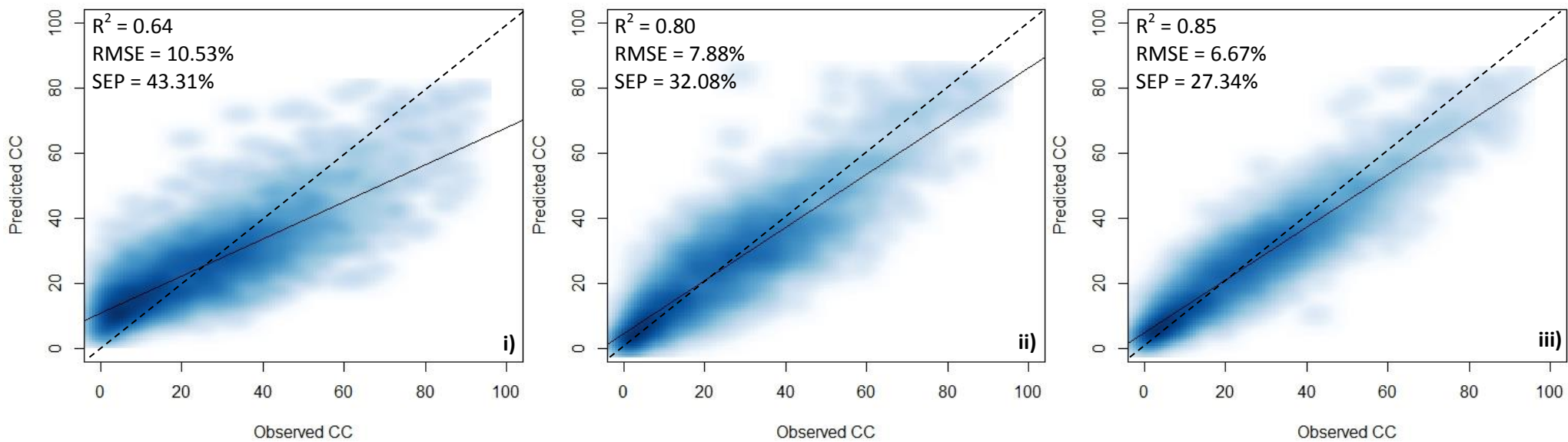
The observed CC versus predicted CC XY scatterplots (figures 4.3i-iii) supported the main findings from Landsat-5 reflectance-only, SAR-only and integrated SAR backscatter and Landsat-5 reflectance analyses. The 2008 multi-seasonal Landsat-5 reflectance only scatterplot (figure 4.3i) illustrated noticeable overestimation below 25% observed CC mark with major underestimation beyond this point, according to the 1:1 line. In comparison, the 2008 SAR-only scatterplot (figure 4.3ii) illustrated drastic improvements in reducing the severity of CC overestimation and underestimation. The integration of the SAR and multi-seasonal reflectance scatterplot (figure 4.3iii) however, yielded a similar trend to the SAR-only scatterplot with a slightly tighter clustering of points around the 1:1 line.





**Figure 4.2: Regional scale CC map of the study area using the best performing RF integrated L-band and single date Landsat-5 band reflectance model (2010 L-band & 2010 Autumn LT5 image; coverage excludes extensive cloud cover to the east)**





**Figure 4.3: Predicted CC versus Observed CC scatterplots for: i) 2008 Multi-seasonal Landsat-5 Reflectance-only, ii) 2008 SAR-only and iii) integrated 2008 Multi-seasonal Landsat-5 Reflectance and SAR modelled validation results**

## 4.6 Discussion

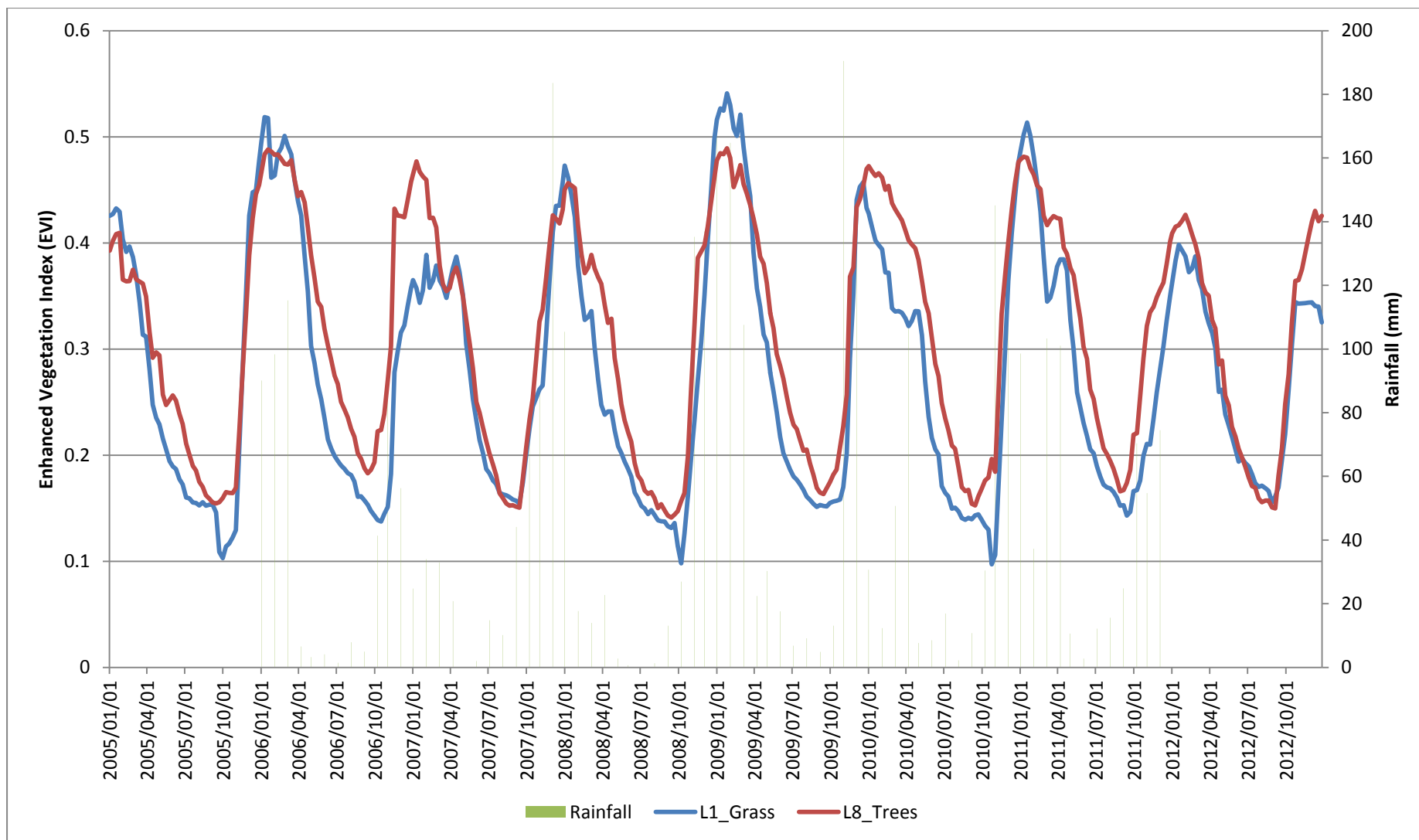
This study was carried out as a step towards addressing the need for a long term (requiring data and sensor continuity), accurate and repeatable regional woody canopy cover mapping in southern African forests and savannahs. A previous study demonstrated that L-band ALOS PALSAR data has the potential for accurate tree cover mapping in South African savannahs, and outperformed C-band RADARSAT-2 and X-band TerraSAR data (Naidoo et al., 2015). Since Landsat data are freely available and routinely used for regional forest monitoring in tropical forests and Australian savannahs (Armston et al., 2009; Hansen et al., 2013; Lehmann et al., 2013), this study sought to compare and integrate optical imagery (Landsat-5 reflectance, vegetation indices and textures) and SAR data (ALOS PALSAR L-band) across various seasons in order to determine if multi-temporal datasets and the combination of sensor technologies improves the accuracy of woody canopy cover mapping.

In this study, it was hypothesized that the season when trees are covered in green foliage, while grasses are dry, should be the best period to retrieve CC, since there is limited interference by green grass (Archibald and Scholes, 2007; Fuller and Prince, 1996; Fuller, 1998; Justice et al., 1985). RF modelling results of the individual Landsat-5 seasonal images (table 4.3) indicated that summer and autumn seasons yielded the highest accuracies, for particular years, with the winter imagery consistently yielding the poorest results. A detailed examination of the temporal phenological fluctuations in tree and grass greenness by way of MODIS EVI time series (Jin et al., 2013) was explored to explain these seasonal results (figure 4.4). This EVI information was linked to precipitation data which is one of the main drivers of phenological cycles and interannual fluctuations in greenness from the plot to regional scales (Scanlon et al., 2005). Monthly EVI values (aggregated from 8 day image composites), extracted from the pixels of the 500m MODIS MCD43 BRDF-corrected surface reflectance data (Schaaf et al., 2002), were selected across two types of landscapes: grass dominated gabbro and tree dominated granite, for the years 2005 to 2013 (a total of 368 pixels extracted for tree and grass dominated landscapes – refer to the Appendix 4 section for exact methodology) (Venter et al., 2003). Generally, EVI values follow a distinct cyclical but variable pattern for trees and grasses, with EVI values peaking during summer (January-February) but falling noticeably to the lowest point in each year during the late winter and early spring (July-September). Generally in savannahs, trees green up earlier than the grasses which only starts greening up after the first rains of the growing season but senesce more rapidly ((Archibald and Scholes, 2007; Chidumayo, 2001) also visible in figure 4.4). Hence, trees have a longer green period compared to grasses ((Higgins et al., 2011); figure 4.4). (Archibald and Scholes, 2007) illustrated, at the landscape



scale, that trees and grasses have different seasonal patterns of leaf display. Savannah trees have a less variable inter-annual phenological cycle since trees use their long-term, accumulated water reserves and are constrained by its architecture (i.e. limited by root system distribution) in contrast to the more variable phenological cycle of grasses that rely on short term resources such as summer rainfall. The fluctuations in rainfall therefore caused more variable grass phenological cycles which would make the separation between tree and grass patterns even more difficult. For instance, in 2009 (wet year) the grass dominated landscape reached a higher maximum EVI value compared to the tree dominated landscape, while the opposite was observed during a typically dry year in 2007.

In order to ascertain the periods throughout the year where the difference between tree and grass greenness were the greatest, we plotted this difference through time (grass EVI minus tree EVI) (Figure 4.5). The periods when the difference in EVI was the most pronounced were brief moments in late spring or during some autumn periods, or in some cases brief summer periods in a dry year (as the case in 2007 and 2010), throughout the time series, with these peak differences varying greatly between years and even in some cases where these differences are small (e.g. years 2009 and 2010). The higher accuracies obtained from the summer Landsat images in years 2007 were most likely caused by dry conditions which resulted in larger differences in the spectral characteristics of grasses and trees. This was not the case when conditions were wetter, with greener grasses (e.g. in 2009). It is important to note that the above patterns were not observed in every year but there was a significant trend between the corresponding difference in EVI and modelled SEP values of the seasonal Landsat-5 images ( $R^2 = 0.37$ ;  $p < 0.05$ ). The poor results obtained with the only spring image available (year 2008) is linked to the image timing which was acquired too early during the spring season while trees had not started to flush leaves (Figure 4.4-4.5). In winter, since most of the trees are deciduous and shed their leaves when grasses are dry, the EVI contrast is consistently the smallest and produced the poorest results. In contrast, the dominance of evergreen tree canopies with dry grass, during the prolonged winter periods, supports the successful use of Landsat for mapping tree cover in the Australian landscapes of the SLATS and NCAS-LCCP programmes (Armston et al., 2009; Lehmann et al., 2013). The brief transitional periods experienced in our South African landscapes during which the contrast between green trees (high tree EVI) and dry grass (low grass EVI) is high are difficult to target, as none of the historic Landsat image acquisition dates actually fell within the period of biggest EVI difference (Figure 4.5), and thus cannot be reliably used to take advantage of this difference. In addition, unavoidable presence of clouds, which at times occur irrespective of season, confounds matters further.



**Figure 4.4: Temporal fluctuations of mean EVI values (extracted from MODIS data) over a predominant grassland site (L1) and a predominant woodland site (L8) between the beginning of 2005 and end of 2012. Rainfall measurements between beginning of 2007 and end of 2011 have also been included**

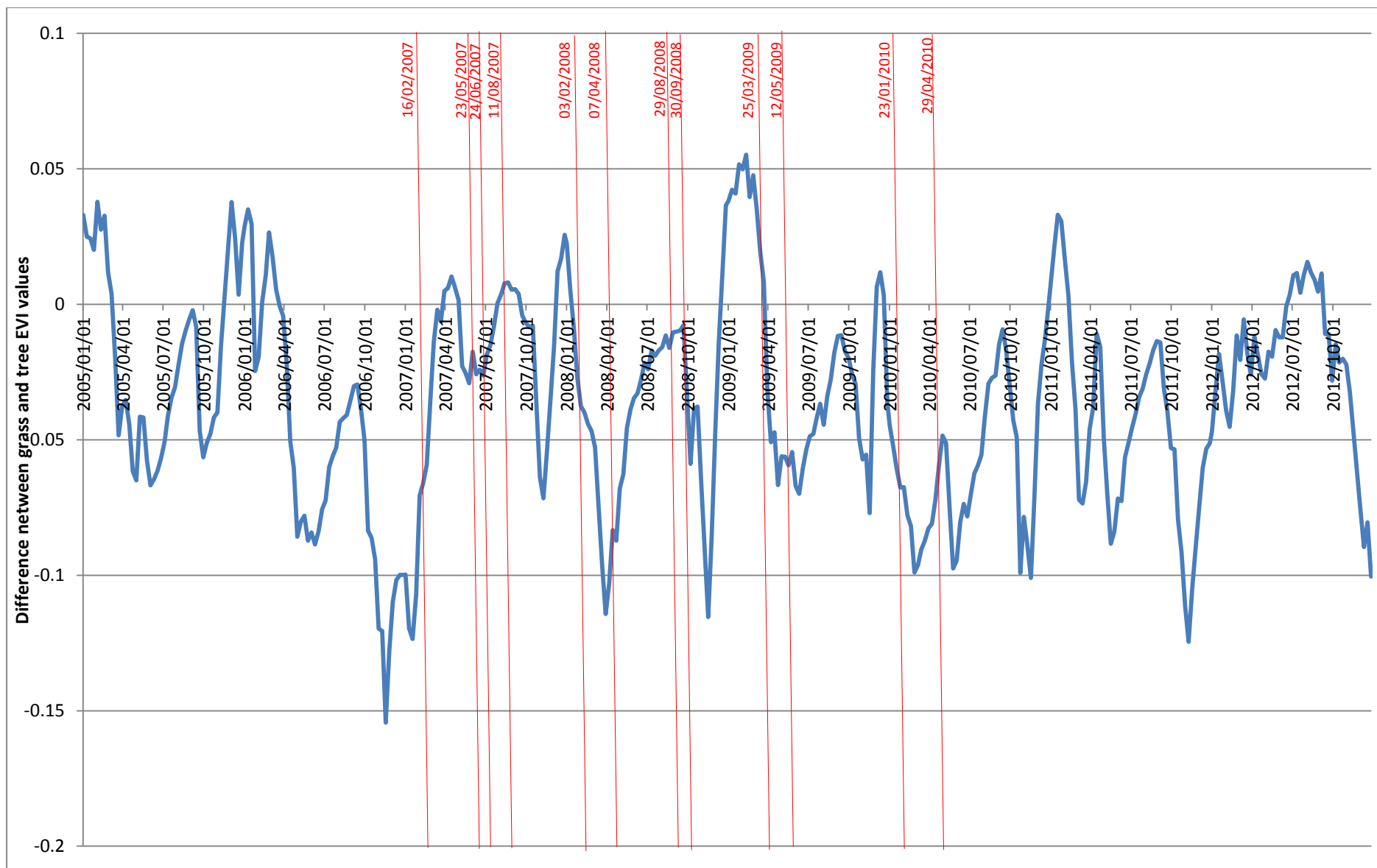


Figure 4.5: Temporal differences of mean grass and tree EVI values (extracted from MODIS data) over a predominant grassland site and a predominant woodland site between the beginning of 2005 and end of 2012. Red lines with numbers indicate the multi-seasonal Landsat-5 image acquisition dates

Attempts were made to improve on the single date Landsat-5 modelling results by using multi-seasonal Landsat images per year and various vegetation indices and image textures derived from the best seasonal image (i.e. the autumn 2010 image). These additions yielded slightly improved model results over the best single date Landsat-5 images results (improved SEP of ~1-5%; table 4.3 and 4.4). The incorporation of multi-seasonal images for each year provided complimentary spectral information that is not present in a single season (e.g. the dry grass signal not present in summer seasons). The addition of image textures with the spectral reflectance contributed more towards improving modelling accuracies (improved SEP of 1%) than the incorporation of vegetation indices, though marginally, which may have contained more redundant spectral information. Since image textures are sensitive to the local variations in brightness arising from the biophysical properties of the tree canopy e.g. shadow (Asner et al., 2002), the textures may have contributed more to the distinguishing of the tree and grass components for CC modelling in the autumn 2010 image than the spectral indices.

Despite the limited improvements provided by the inclusion of indices, textures and multi-seasonal reflectance data, none of the best performing Landsat-5 only models had accuracies that measured up to those obtained with a single winter L-band SAR image ( $R^2 = 0.72$  versus  $R^2 = 0.81$ ). These results are in contrast with those of (Laurin et al., 2013) where optical Landsat textures contributed the most to improving accuracy, surpassing the contribution of SAR backscatter. The highest modelled CC accuracies achieved by Landsat derived optical products ( $R^2=0.72$ ;  $RMSE\sim 12\%$ ) in our study was significantly lower than the CC accuracies achieved in the Australian Statewide Landcover and Trees Study (SLATS) Program (Armston et al., 2009) and the Australian National Carbon Accounting System – Land Cover Change Program (NCAS-LCCP) (Lehmann et al., 2013) which mapped savannah and forested landscapes ( $R^2>0.79$  and  $RMSE<10\%$ ). The limited accuracies achieved by Landsat-5 only RF models (table 4.3 and 4.4) clearly indicated that the implementation of a CC monitoring system based solely on Landsat-derived data will not be adequate in the South African Savannahs. L-band SAR data prove to be a much more effective alternative for reliable and consistent CC mapping and monitoring in this open forest environment.

The integration of the SAR dataset with the best single season and multi-seasonal Landsat-5 reflectance yielded models with the highest accuracies, which corroborates findings in previous studies (Laurin et al., 2013; Lucas et al., 2006b; Moghaddam et al., 2002; Rosenqvist et al., 2003; Townsend, 2002). For instance, the SAR-only and the multi-seasonal Landsat only models explained

81% and 72% of the CC variance, respectively, while a model combining the two explained 89% of the variance. The significant increase in accuracy (7.5% improvement of SEP) at the high end of the model performance demonstrated that optical reflectance data provided additional information which is complementary to that captured by the SAR backscatter. Significant complementarity between SAR and Landsat data was demonstrated by (Lehmann et al., 2015) with the combined datasets yielding highly accurate results within the Australian NCAS-LCCP (global classification accuracy of 90%). Additionally, in this study the differences in accuracy between the best single seasonal and combined multi-seasonal reflectance datasets, when both integrated with SAR datasets, were minimal (e.g. for 2010,  $R^2 = 0.88$  versus  $R^2 = 0.89$ ). In closing, this study provides important insights for the development of a national woody vegetation monitoring programme for South African savannahs where extensive L-band SAR and LiDAR calibration and validation datasets need to be prioritised for acquisition. The combination of winter SAR scenes with summer/autumn Landsat-5 scenes would be the optimal data inputs in an operational CC mapping programme. The recent launch of the ALOS PALSAR-2 (L-band) sensor ensures long-term provision of L-band SAR data on which an operational woody vegetation monitoring system could be based, although the high cost may be a limiting factor.

## 4.7 Conclusions

This study aimed to map regional-scale woody fractional cover (CC) at the highest possible accuracies using SAR (L-band ALOS PALSAR) and multi-seasonal optical (Landsat-5 TM) data. Landsat-5 imagery acquired in the summer and autumn seasons yielded the highest single season modelling accuracies, depending on the year, but the combination of multi-seasonal Landsat-5 images yielded higher accuracies. The addition of vegetation indices and image textures and their combinations to the spectral reflectance bands provided minimal improvements, with none of the optical-only combinations yielding accuracies greater than those achieved using any single winter SAR L-band image. Due to the unpredictability of the narrow temporal 'window' during which trees and grass may differ sufficiently in phenological greenness, CC mapping and monitoring in savannahs based solely on Landsat data, is not recommended. Extensive cloud cover during the summer or even autumn seasons further compounds this problem. However, there was significant, yet modest, improvement ( $R^2$  of  $\sim 0.08$ ,  $\sim 1.9\%$  of RMSE and  $\sim 7.5\%$  of SEP) in accuracy when 2010 multi-seasonal optical reflectance bands were combined with the L-band backscatter variables (i.e. the best performing SAR and optical integrated dataset scenario). The best trade-off, however, between accuracy and complexity was given by a model using 2010 winter SAR and autumn season Landsat-5

reflectance as input variables, as the model utilised a single SAR and single Landsat-5 image. The authors recommend that further testing of the performance of Landsat imagery, alone and in combination with winter SAR data, be conducted in other southern African vegetation types where tree canopies are evergreen, such as in commercial plantations, indigenous forests and thickets, and where Landsat may produce a better performance. It is also recommended that a system based on L-band SAR datasets, with supporting airborne LiDAR data for model calibration and validation, should be applied to other bioregions (e.g. afro-montane and coastal indigenous forests) before a national CC monitoring programme can be established in the future.

# Chapter 5: Scaling-up methods for national woody fractional cover mapping: Experiments and guideline on the amount of field plots and airborne LiDAR data required for training and validation

## 5.1 Abstract

Accurate mapping of woody fractional canopy cover (CC) at the country-wide scale remains challenging due to the large image data volumes of sufficiently high resolution. Both field plots and LiDAR datasets serve as representative training and validation datasets for country-wide CC mapping using SAR data. This study sought to establish the optimal quantity of field plots and LiDAR coverages required to train a Random Forest model to map CC at a country-wide scale using ALOS PALSAR HH and HV backscatter and DEM ancillary variables. 35% of randomly selected training data, from the five main biomes (Fynbos and Thicket, Indigenous Forest, Savannah and Grassland) and the Savannah biome alone were used to train RF models and validated against a fixed dataset of each of the biomes. Field plots were simulated from the high resolution LiDAR data. This approach assessed the representativeness of the samples and the optimal number of field plots and quantity (size and number) of LiDAR coverage. Optimal number of field plots and the quantity of LiDAR coverage were selected where the modelling results showed the highest accuracy, i.e. the lowest Root Mean Square Error (RMSE), with respect to sampling effort. The results have shown that the Savannah-only training dataset yielded high accuracies across Grasslands, moderate accuracies across Thickets but poorer accuracies in the Indigenous Forests and Fynbos biomes. Sampling the training data across all available biomes yielded higher accuracies. From the LiDAR-simulated field plot analysis, it was concluded that a minimum of 500, 1ha field plots, i.e. 125 1ha field plots equally sampled within 0-20%, 20-40%, 40-60% and >60% CC ranges, would be sufficient for effective modelling of CC at the country-wide scale. Additional field plots, beyond this number (500) would improve the overall accuracies only slightly, but incurred significant increases in sampling efforts and costs. The analyses also suggest that the most economical LiDAR acquisition strategy would include only four separate 5000ha LiDAR acquisitions, distributed across the five vegetated biomes. Thus, an optimal sampling strategy would require only 20000ha of the total number of 122052ha of LiDAR at our disposal. The study found that much less LiDAR data were required to train the models than originally expected, provided that the acquisitions were sufficiently diverse in CC and vegetation type and could also be cheaper to acquire than collecting 500 1ha field plots.

**Keywords:** *Woody canopy cover, SAR, LiDAR, training, validation, Random Forest*

## 5.2 Introduction

A variety of national, regional and global woody fractional cover products are available around the world. Such regional or country-wide initiatives include the Australian State-wide Land cover and Trees Study (SLATS) (Armston et al., 2009), the Australian National Carbon Accounting System – Land Cover Change Program (NCAS-LCCP, (Lehmann et al., 2013)), the Landsat Ecosystem Disturbance Adaptive Processing System (LEDAPS) of North America (Ju et al., 2012) and the Amazon Deforestation Monitoring Project (PRODES, (Hansen and Loveland, 2012)) which have been derived from locally calibrated datasets. Global initiatives, such as the MODIS Vegetation Continuous Field (VCF, (Townshend et al., 2011)), the JAXA Forest/Non-Forest (FNF, (Shimada et al., 2014)), Hansen’s global maps of forest cover change (Hansen et al., 2013) and Sexton’s continuous fields of tree cover (Sexton et al., 2013), provide a significant means for large-scale woody fractional cover (CC) monitoring. These programmes focused on mapping woody fractional cover or CC, and global forest change in some cases, (i.e. the area vertically projected on a horizontal plane by plant canopies – (Jennings et al., 1999)) as it is one of the simplest metrics for monitoring the woody vegetation component. Within South Africa, the Savannah biome is the largest and makes up 35% of the country (Van Wilgen, 2009). In this biome, total CC values can range from dispersed trees in open-grasslands (~5%) to near-closed canopy woodlands (~60%) and more than 80% in riparian zones (Venter et al., 2003). In this country, recent work has shown that woody plants may have increased at a rate of 5-6% per decade, suggesting the occurrence of bush encroachment which can adversely affect the environment via the reduction of land productivity (e.g. reduced livestock grazing), the alteration of species composition and reduction of water availability (O’Connor et al., 2014; Shackleton et al., 1994). Bush encroachment can also provide positive environmental impacts such as carbon sequestration (Mitchard et al., 2011; Sankaran et al., 2008) and the provision of anthropogenic ecosystem services, such as fuelwood (Shackleton et al., 2007). However, competitive claims over natural resources (grazing, fuelwood provision and biodiversity) can only be managed if information on spatial patterns, and most importantly temporal dynamics, are available. The regular and accurate monitoring of CC is thus warranted. Unfortunately, such dedicated woody vegetation monitoring programmes are currently not in place for South Africa, but current research may make it a reality in the near future.

The majority of such programmes make use of optical datasets (e.g. Landsat and MODIS) due to their regular, long-term coverage, medium spatial resolution and easy, free data access. Under



particular circumstances, these optical sensors can take advantage of the differences in seasonal tree and grass phenologies to map CC successfully (as the case in the Australian SLATS programme). However, within the context of South Africa, (Naidoo et al., 2016) demonstrated that an L-band SAR system, particularly winter acquisitions, would be more accurate in modelling CC than optical-based sensors given the specificities of the phenological cycle of woody versus grass components of the South African savannah vegetation. Given that SAR is also unaffected by cloud cover, haze and smoke, L-band SAR data is particularly useful for woody vegetation monitoring in South Africa.

Upscaling from local pilot studies to the national scale present a number of challenges. The first challenge is in the mass-downloading, storage, processing, and analyses of large volumes of medium resolution satellite data, e.g. Landsat. This challenge is currently being overcome by advances in information technology, specifically high performing computing (Hansen and Loveland, 2012), and with the delivery to the community of global medium resolution mosaics such as for the sensor, ALOS PALSAR (JAXA). The second challenge is that upscaling typically decreases modelling performances because of large environmental conditions prevailing over large areas, such as topography, rainfall and vegetation type, which increases the variability within the remote sensing sensor signal (Wu and Li, 2009). Within the context of vegetation cover, for instance, (Sankaran et al., 2005) demonstrated that maximum woody cover can be constrained in Savannahs which receive a mean annual precipitation less than 650mm. Vegetation cover is also influenced by topographical parameters (such as elevation, slope and aspect) which control the direction and speed of flow and the accumulation of water in the landscape as a result of gravitational forces (Florinsky and Kuryakova, 1996). A solution to this challenge would involve the incorporation of such environmental conditions, in the form of ancillary modelling input variables, into the modelling process. Finally, the biggest challenge, however, is the availability of representative training and validation datasets which serve as the backbone of remote sensing-based vegetation monitoring programmes. Field measurements, in most of these cases, are typically used to train and validate other higher resolution satellite datasets via model upscaling. Due to the labour and time intensive nature of field measurements, lack of standard methods, and the poor spatial distribution of field data, a more spatially representative and reliable alternative is required (Bombelli et al., 2009).

Airborne LiDAR data has served as an accurate and reliable source of training and validation for model upscaling and mapping (Englhart et al., 2011; Montesano et al., 2016; Naidoo et al., 2016,

2015). Compared to other high resolution optical imagery, airborne LiDAR, is the most expensive with a cost of approximately 1-5 US\$ per hectare depending on the total coverage, sensor specifications and location of deployment (Hummel et al., 2011; Kelly and Di Tommaso, 2015; Thompson et al., 2013; Wulder et al., 2008). This fee, however, excludes airport logistical costs and aircraft fees (e.g. rental and fuel costs) and manoeuvring costs which vary on the number of turns the aircraft has to perform during the acquisition. In most situations, wall-to-wall acquisitions of an entire country, particular as large as South Africa (122.1 million ha), is to date not financially feasible as a base line cover, and hence from a monitoring perspective. Thus there needs to be a trade-off between the area sampled with LiDAR and the total cost incurred (Ene et al., 2016; Wulder et al., 2008) for using these datasets nationally for calibration and validation of satellite remote sensing models. LiDAR can sample much larger, representative areas at higher detail than possible through extensive field sampling. In fact (Hummel et al., 2011) found that LiDAR data acquisition and processing costs were comparable with field data collection across the same area. Moreover, LiDAR data have been found to be on par with corresponding field measurements (Næsset and Økland, 2002; Nickless et al., 2009; Wulder et al., 2008). However, testing the suitability (via modelling requirements and financial costs) of both LiDAR and field plot measurements for upscale modelling efforts would be beneficial especially when LiDAR datasets are not available. A limited number of studies have successfully utilised or tested field plots, of varying number and size of plots, for upscaling modelling efforts of vegetation structure (Saatchi et al., 2007; Urbazaev et al., 2015).

Even though it is expected that the cost of LiDAR data will reduce in the near future (Asner, 2009), costs will remain high enough to warrant the design of optimal LiDAR sampling schemes at the national scale. It is thus important to establish a guideline for the quantity and distribution of LiDAR acquisitions, and associated field plots, required for training and validation of models in a national CC monitoring system. Thanks to a combination of national (ESKOM and SANParks) and international (Carnegie Airborne Observatory or CAO) collaborations, a total of roughly 122 052 hectares (ha) of airborne LiDAR coverage, acquired between 2009 and 2013 across South Africa, was assembled for this study. This large collection of LiDAR datasets were used as the main source for model training and validation, and served as a source from which 1ha field plots were simulated. The main aim of this study was to ascertain the optimal representative sampling of airborne LiDAR data and LiDAR simulated field plots, across Savannah-only and all main biomes, for the up-scaled modelling of woody fractional cover (CC) at the country level using ALOS PALSAR L-band SAR data. The Savannah biome was chosen as the point of comparison for other biomes as it is the biggest

vegetated biome in South Africa and possesses one of the most representative CC range while also serving as a cost effective alternative to sampling across all biomes. Secondary objectives include the investigation of the inclusion of regional, environmental variables (i.e. elevation-based and rainfall variables) for potential modelling improvements. To achieve this aim and secondary objectives, the research questions were:

- 1) Does the inclusion of regionally stable ancillary variables such as elevation, slope, and aspect and rainfall gradient information assist L-band HH and HV backscatter in modelling CC at the country wide scale?
- 2) What is the impact of having LiDAR data that are limited to a single biome, i.e. the Savannah? More specifically, is LiDAR data which is limited to the Savannah biome (as specified in (Rutherford et al., 2006)) sufficient for training and validation for L-band SAR-based modelling and mapping of CC for the whole country? Also, how do these results of using LiDAR from the Savannah only compare to those where diverse LiDAR datasets from Fynbos, Thicket, Grassland and Indigenous Forest biomes are used?
- 3) What is the optimal amount of field plots, as simulated from LiDAR datasets, required for modelling and mapping of CC with L-band SAR across the country and in Savannahs only? The 'optimal amount', in this case, refers to the point of the most favourable trade-off between modelling accuracies and sampling effort (i.e. number of field plots).
- 4) What is the optimal amount, in terms of area (hectares) and number and size of acquisitions of LiDAR data required for optimal L-band SAR-based modelling and mapping of CC within (i) the Savannah and (ii) country-wide, in comparison with the accuracies achieved using an optimal number of field plots? The 'optimal amount', in this case, refers to the point of the most favourable trade-off between modelling accuracies and sampling effort (i.e. the number, size and total coverage of LiDAR acquisitions while taking into account the cost effectiveness of the various LiDAR acquisition specifications).

### 5.3 Study Area

The study area chosen for this study is the entire country of South Africa (SA) and mainly echoes the information provided in the Study Area section of Chapter 2. The information provided here, in this chapter, briefly describes the key geological, climate, general topography and ecosystem related vegetation feature types found in SA. At the country level, average temperatures are generally mild but can vary according to location and proximity to the oceans. Annual average precipitation is

about 450mm with a high-to-low rainfall gradient existing from east to west which mainly limits forest distribution. The country possesses a diverse range of more than 60 different geological substrates. This ranges from granite, basalt and gabbro derivatives which dominate lowveld Savannah to mudstone and shale derivatives of the Karoo and central portions of the country to name a few (<http://waterresourceswr2012.co.za/>). Geologies are more mixed along the South African east and west coasts. With respect to the general topography of the country, areas of low elevation such as the Lowveld of Mpumalanga and coastal regions (with concentrated mountain ranges in the Western Cape region) and areas of high elevation, e.g. the Highveld in the central interior, are separated by the presence of a prominent ridged escarpment running across the country's southern extents (Weepener et al., 2011). Many of the major rivers and tributaries cut through the escarpment from the high lying interior and mountainous areas towards the coastal areas and finally, out to sea. This diverse terrain, together with the various underlying geological substrates, supports an equal diversity of vegetation types across the country. These range from a variety of widespread thorned, mixed and sweet bushveld types in the Savannah biome, to subtropical thickets of Eastern Cape, to concentrated patches of natural Afromontane, coastal and mistbelt forest types around South Africa, and to the highly endemic, sclerophyllous patches of fynbos in the Western Cape region (Mucina and Rutherford, 2006). Distributed within these biomes, particularly over higher rainfall regions within the eastern and southern parts of the country (e.g. within parts of the Savannah, Fynbos and Forest biomes), are patches of commercial plantations (Dye and Versfeld, 2007). These plantations are roughly made up of 57% Pine, 35% Eucalyptus and 8% *Acacia mearnsii* (Black Wattle) and are grown for timber pole and pulp production (Dye and Versfeld, 2007). The study area, including biome and airborne LiDAR coverage, is displayed below in figure 5.1.

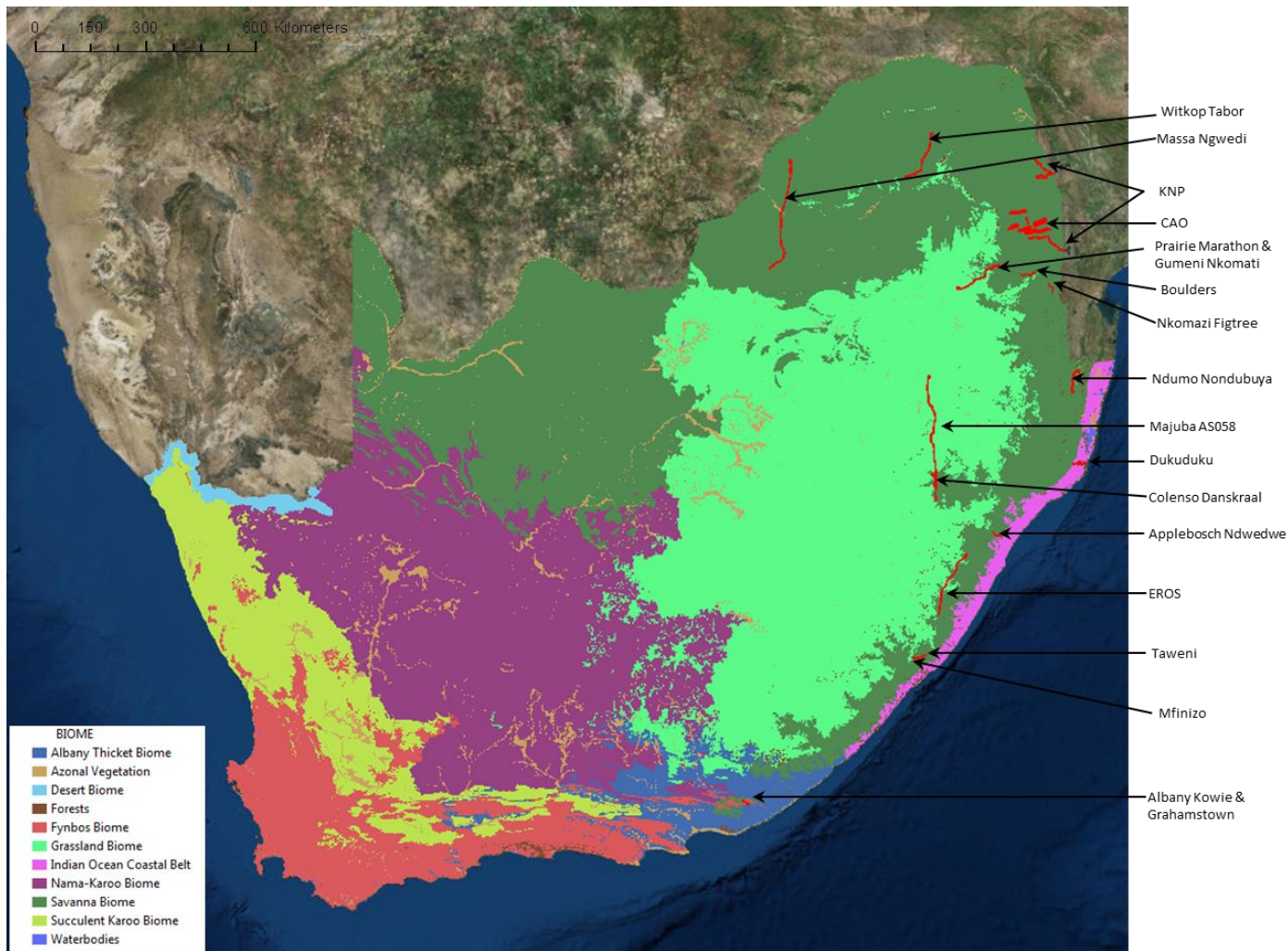


Figure 5.1: Study area of South Africa, including biome (Mucina and Rutherford, 2006) and airborne LiDAR acquisition coverage

## 5.4 Materials and methodology

The proposed methodology in this chapter draws upon the major findings, by way of recommended remote sensing datasets and analytical techniques, of chapters 2, 3 and 4. The extensive LiDAR datasets, used for the global forest product validation assessment in chapter 2, was used as the main training and validation source while L-band ALOS PALSAR FBD Global image mosaic datasets (25m) were used as the main input variable for the modelling procedures. A bootstrapped Random Forest (RF) machine learning algorithm was used in the modelling approach. The modelling dataset, consisting of LiDAR CC metric as the dependent and ALOS PALSAR HH and HV dB backscatter as independent variables, was split into various modelling scenarios to test the robustness of the RF model sampling and address the research questions. Regional environmental indicators, such as

elevation, aspect and slope, and rainfall information were also utilised as additional input variables in the RF modelling.

#### **5.4.1 Airborne LiDAR datasets and processing**

The extensive LiDAR datasets, utilised in chapter 2 (Section 2.4.4), was used in its entirety for this chapter with the identical processing steps being followed. For the sake of brevity, this information will not be repeated. These LiDAR datasets, which covered approximately 122 052 hectares (ha), were flown between 2009 and 2013 and across different ecosystems (ranging from coastal forests to savannahs, bushveld, thornveld, thickets and plantations) of South Africa (see Table 2.1). In terms of coverage, these LiDAR datasets were mostly acquired across the eastern half of SA where the majority of forested vegetation occurs. This distribution is the result of the 400mm “rainfall line” or isohyet which separates forested areas of mean annual rainfall greater than 400mm, in the eastern regions of South Africa, and non-forested areas of less than 400mm, in the western regions of South Africa. Though the processing of these datasets remain unchanged from chapter 2, the way in which the data were extracted differed and will be elaborated in a later section (5.4.4). As outlined in the previous chapters (2, 3 and 4), the woody canopy cover (CC) was derived from each LiDAR CHM dataset by extracting the total number of woody vegetation pixels divided by the total number of LiDAR pixels (can vary depending on the resolution of the processed CHM) within a 25m X 25m pixel resolution (corresponding to the pixel size of the ALOS PALSAR global mosaic), then multiplied by 100 to get a percentage.

#### **5.4.2 ALOS PALSAR FBD global image mosaic**

25m 2010 ALOS PALSAR HH and HV global image mosaics were utilised as it was the only consistent L-band SAR dataset available for the South African country-wide coverage and were publicly available for download from the JAXA dataset portal ([http://www.eorc.jaxa.jp/ALOS/en/palsar\\_fnf/data/index.htm](http://www.eorc.jaxa.jp/ALOS/en/palsar_fnf/data/index.htm)). This global mosaic has been both radiometrically and geometrically calibrated using particular processing and mosaicking steps described in (Shimada and Ohtaki, 2010), therefore only limited additional SAR image processing steps were required. After downloading the imagery for the whole country, the subsequent product required converting the raw Digital Number (DN) to Sigma dB (dB), using a calibration formula

(equation 5.1) provided by JAXA, and mosaicking into easy to handle image ‘chunks’ (consisting of roughly three scenes) in ENVI 4.8. This procedure was conducted for both HH and HV polarisations.

$$\text{Backscatter values (dB)} = 10 \times \log_{10}(\text{DN}^2) - 83.0 \quad \text{Equation 5.1}$$

The image chunks were projected from their native geographic WGS84 projection to an Albers Equal Area projection. As mentioned in Chapter 2, this was done to fix an alignment issue between the ALOS PALSAR mosaic and the various LiDAR datasets which was evident during preliminary analysis runs. The solution, through trial-and-error, was to shift the ALOS PALSAR mosaic datasets by a constant configuration of 75m westwards and 50m northwards.

### 5.4.3 Ancillary environmental parameters

Because of the regional scale of the analysis, the benefit of the inclusion of additional regional environmental variables to model CC was tested rather than the use of direct measurement variables. Aspect (i.e. the direction that a surface faces in degrees clockwise from North; 0-360°) and Slope (i.e. the percentage or degree change in elevation over distance) variables were derived from SRTM30m digital elevation model data (i.e. Elevation or the height above sea level in metres) ([https://remotepixel.ca/projects/srtm\\_leaflet.html](https://remotepixel.ca/projects/srtm_leaflet.html)) using the raster surface toolbox (in 3D Analyst Tools) in ArcMap 10.1. Finally a 2005 country-wide climate rainfall map (200-1000mm and greater) was obtained from the South African Water Research Commission (WRC) to add general rainfall class information (200-400mm, 400-600mm, 600-800mm, 800-1000mm and >1000mm) to the extracted data. These variables were chosen as they do play a role in influencing the distribution of CC across South Africa and was expected to help capture some of the variability of CC across the landscape during the modelling approach. (Sankaran et al., 2005) demonstrated that maximum woody cover can be constrained in Savannahs which receive a mean annual precipitation less than 650mm. Vegetation cover is also known to depend on topographical parameters (such as elevation, slope and aspect) which control the direction and speed of flow and the accumulation of water in the landscape as a result of gravitational forces (Florinsky and Kuryakova, 1996).

### 5.4.4 Dataset integration and extraction process

A fixed grid of 105m X 105m cells, with a 50m distance to avoid spatial autocorrelation, was implemented to extract corresponding LiDAR, SAR and ancillary data values for creation of the

training dataset. This method was described in the data integration sections of chapters 3 and 4 (Naidoo et al., 2016, 2015). The grids were created in QGIS with an Albers Equal Area projection to define the grid size and spacing distance but were re-projected to a WGS84 geographic projection to allow for overlaying over the different LiDAR and SAR datasets which possessed different projections. These grids were clipped to the different LiDAR flight extents and cells which fell partly or fully across water bodies, seasonal agriculture centre-pivot plots, powerlines, urban settlements and infrastructure were removed to allow for uncontaminated SAR to LiDAR correlations. The cells were removed via visual interpretation in which Google Earth was used as a backdrop. Zonal statistics in the Spatial Analyst Toolbox (ArcMap 10.1) was used to extract the aggregated mean cell values of the different dataset types. Variables including mean LiDAR CC, SAR HH and HV backscatter and ancillary DEM parameters and rainfall information were extracted for a total of 48 007 cells, which were treated as individual samples.

85.46% of the samples (i.e. the extracted cells) fell within the Savannah biome, 13.26% within the Grassland biome, and 0.87% within the Indigenous Forest biome (Figure 5.2). Thicket and Fynbos biome coverage was very small (0.08% and 0.26% respectively). The biome layer was derived from a 2006 national vegetation map produced by (Mucina and Rutherford, 2006). Biomes are described as a land grouping, governed by climate, which possesses plants and animals living together with the same degree of permanence and demonstrate large-scale patterns in global plant cover (Mucina and Rutherford, 2006). The LiDAR samples were also classified according to Willis's vegetation structural classification scheme (Willis, 2002) for southern Africa to ascertain the structural variability of the samples (Figure 5.3i), captured at the individual LiDAR acquisition scale, and thus at a higher resolution than the coarser biome scale. Willis's classification scheme considers the mean canopy cover and tree height and was implemented by classifying the LiDAR derived CC and height metrics according to the scheme (Figure 5.3ii).



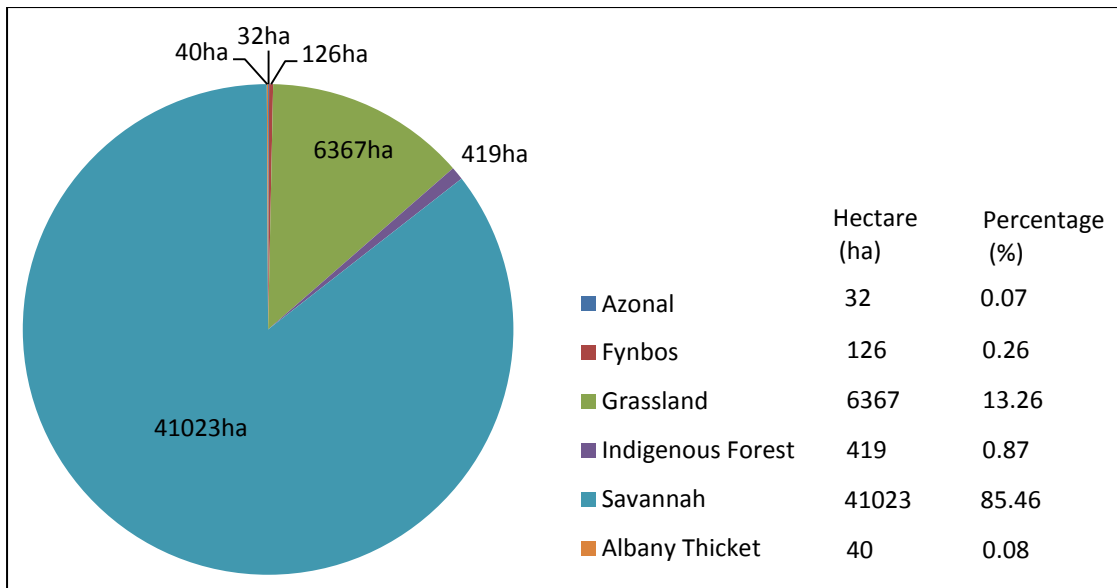


Figure 5.2: Approximate hectare and percentage coverage of samples in all LiDAR datasets according to biome

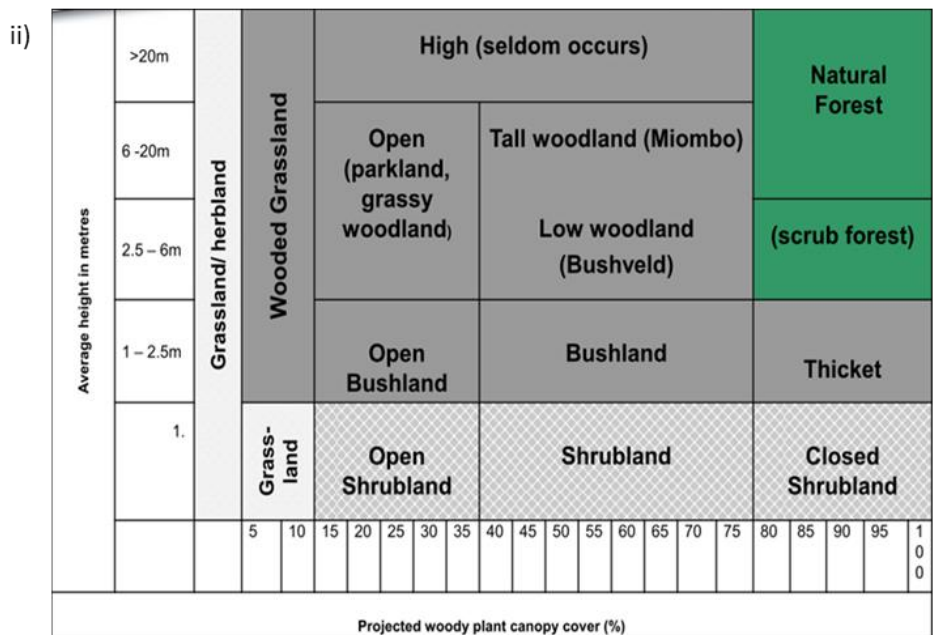
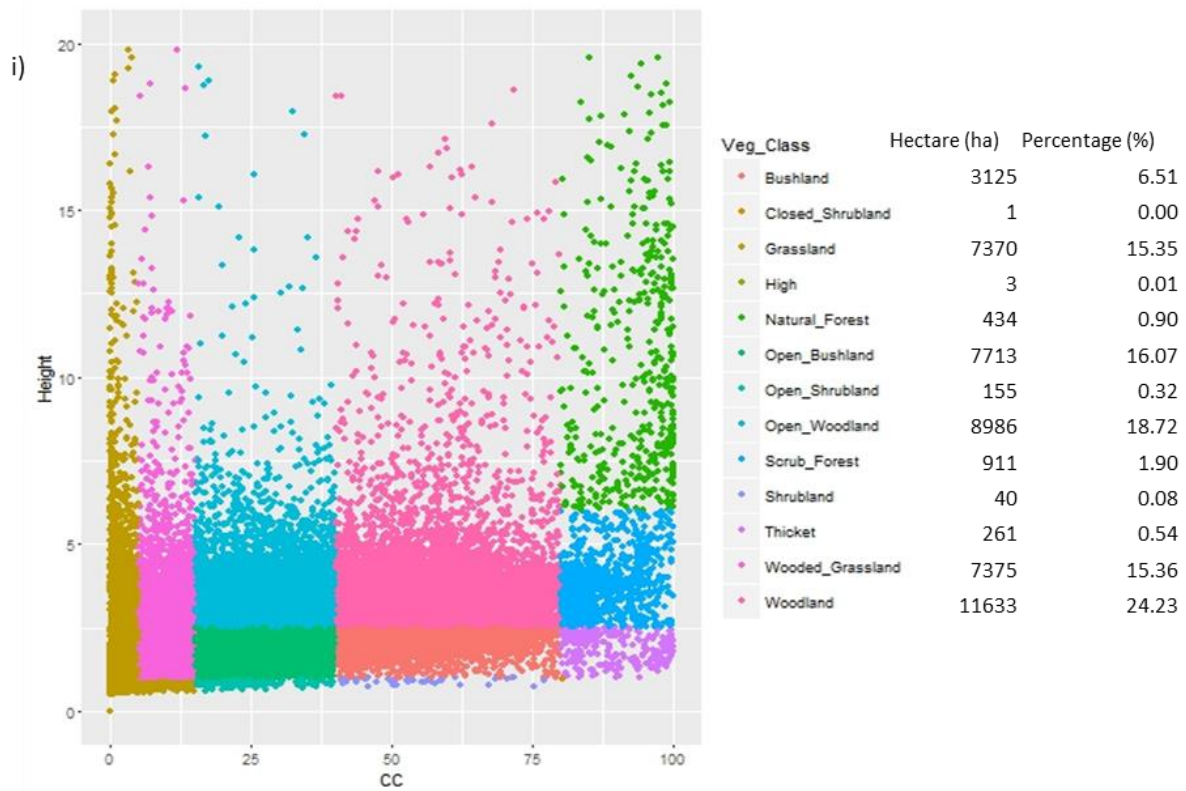


Figure 5.3: i) Samples according to vegetation structural classes (including hectare and percentage coverage statistics per class); ii) Vegetation structural classes (Willis, 2002) according to average height and projected woody plant canopy cover which was used to classify i)

Despite the bias of the LiDAR dataset collected predominantly over the Savannah biome, a large number of the vegetation structural classes were well represented in our sample (Figure 5.3i), from grasslands to open or close woodlands, and dense natural forests or thickets. This illustrated a

higher degree of vegetation structural variability being captured by the extracted data than one would have expected from only examining earlier biome distribution results (Figure 5.2). In Figure 5.3i), however, the vegetation structural classes less than 1m in height (i.e. grassland and shrubland classes) were not well presented. This was due to the combination of the 0.5m height threshold being applied to the LiDAR CHMs, to remove the influence of the grass layer on CC estimates, and due to the reduced capacity of LiDAR to detect vegetation less than 1m in height (Wessels et al., 2011).

#### **5.4.5 Modelling process – algorithm, modelling scenarios and validation**

A Random Forest (RF) (Breiman, 2001) machine learning algorithm was implemented for the modelling approach as it performed well in related studies, within the context of CC prediction, in South Africa (Naidoo et al., 2016, 2015, 2014). Recommended by these previous studies and others (Ismail et al., 2010; Prasad et al., 2006), the default number of trees in the forest, or *n*tree, (i.e. 500) and the default number of possible splitting variables for each node, or *m*try, (i.e. square root of the number of input parameters used) were implemented.

For the first research question, HH and HV backscatter variables were assessed alone and together with the addition of DEM ancillary and rainfall information variables for CC modelling. The resultant four modelling scenarios were: (i) HH and HV backscatter only; (ii) backscatter and DEM parameters (elevation, slope and aspect); (iii) backscatter and rainfall information and (iv) backscatter, DEM parameters and rainfall information combined. 35% of the total extracted dataset (here referred to, throughout the chapter, as the all-biome dataset) was randomly selected for RF training and the model was validated against a fixed validation dataset. In order to mitigate the sampling bias towards the Savannah biome, a fixed validation dataset was selected from the extracted dataset which was excluded from the model training process. 40 samples were randomly selected from each biome category of the all-biome dataset (Savannah, Grassland, Indigenous Forest, Thicket and Fynbos) for inclusion into the fixed validation dataset, thus resulting in a total number of 200 samples. The 32 samples from the azonal vegetation class were excluded from the creation of the fixed validation dataset. The number of 40 samples was chosen as it was the maximum number of observations present in the Thicket biome which was the least represented biome in the extracted dataset (Figure 5.2). Unfortunately, this meant that no training samples from the Thicket biome could have been utilised in the modelling approach as all samples were incorporated in the fixed

validation dataset. This fixed validation dataset, which was considered to be both appropriate in terms of the RF modelling validation and representative in terms of vegetation structure and CC, was used for validating all the scenarios assessed to address the different research questions. This also facilitated the comparison of the various scenarios. The RF algorithm was performed with cross-validation (K-fold of 10) for added model stability. The accuracy assessment statistics, which included the mean coefficient of determination ( $R^2$ ), the mean Root Mean Squared Error (RMSE), the mean Standard Error of Prediction (SEP) and their 95% confidence interval (CI) counterparts, were recorded.

The best performing combination of variables (i.e. backscatter and/or ancillary variables) was used in the modelling processes of the remaining research questions. For the second research question, the samples extracted from Savannahs only were used for RF training (35% randomly selected) and the resultant models were validated against the five individual biomes in the fixed validation dataset: Savannahs (scenario v), Grasslands (scenario vi), Indigenous Forests (scenario vii), Thickets (scenario viii) and Fynbos (scenario ix) and total validation accuracy (scenario x). These scenarios (v-x) were repeated in which the same number of training data (approximately 14 344 observations used in the Savannah-only case) was randomly selected from the all-biome dataset to keep the comparison of validation results consistent. The final two research questions were discussed in the upcoming sections (5.4.6 and 5.4.7) which dealt with ascertaining the optimal amount of training data (i.e. LiDAR simulated field plots and LiDAR acquisitions specifications) required for CC modelling and mapping at the country scale. Ascertaining the optimal training data amount is vital for reduced sampling efforts, optimised modelling accuracies and for reducing training data acquisition costs.

#### **5.4.6 Optimal amount of LiDAR simulated 1 hectare field plots**

This section served to create a sampling guideline for users who do not have possession of LiDAR datasets and will be relying on the collection of field plots for CC upscaling at the country scale. To test the optimal amount of field plots required for country-wide CC modelling and mapping, we simulated field plots from the LiDAR data for both the complete and the Savannah-only extracted datasets. 48 007 extracted LiDAR cells approximate 48 007 1ha simulated field plots. For the analysis, discrete amounts of training data, which increased gradually according to increasing number of 1ha “field plots” (i.e. 12, 24, 48, 100, 248, 500, 1000, 2000, 3000, 5000, 10000, 15000 and

20000 X 1ha field plots), were selected in a stratified random manner, according to four broad CC classes (0-20%, 20-40%, 40-60% and >60%), and used as training data for RF models. The selection process for the stratified 1ha random sampling was forced to extract an equal number of “field plots” within each of these four CC classes. These four broad CC classes were also chosen as it was practical enough for users to replicate from a variety of reference remote sensing sources (e.g. Google Earth or aerial photographs) during the “field plot” pre-selection process. This CC class stratification would prevent potential sampling bias of “field plots” across the CC range (e.g. sampling more “field plots” in the lower and medium CC ranges than the higher CC ranges etc.). This process is standard practice when selecting field plots. The fixed validation dataset, used in section 5.4.5, was also used for RF model validation to prevent model validation bias. A cross-validation (K-fold of 10) approach was again implemented. The validation-based mean  $R^2$ , RMSE and their CI counterparts (upper and lower limits) were plotted against the corresponding number of 1ha “field plots”, used for model training. The optimal number of 1ha “field plots” was determined by examining the percentage change in both  $R^2$  and RMSE as the number of “field plots” was increased, while considering sampling effort/size versus the gains in accuracy.

#### **5.4.7 Optimal LiDAR training amount in terms of hectare coverage**

This section served to create a sampling guideline for users who intend to acquire airborne LiDAR for CC upscaling at the country scale. Factors controlling the optimal amount of LiDAR are complex as individual acquisition/track size; the number of acquisitions/tracks; the distribution of acquisitions/tracks and total acquisition coverage all influence the training data available for modelling, the achievable modelling accuracy and also subsequent LiDAR costs. For instance, airborne LiDAR can be economical if acquired in tracks instead of large areas but a larger sample area will provide more training sample data but in turn incur higher costs (such as aircraft deployment, manoeuvring and other related logistical costs). However, increasing the number of acquired LiDAR tracks will also drive up costs. Also for country wide mapping of CC, sampling across diverse CC ranges and vegetated biomes are also necessary. Balancing these factors along with the users’ requirements and available budget is challenging but essential.

This section followed a procedure fairly similar to the previous optimal field plot section (section 5.4.6) but instead of dealing with individual simulated 1ha “field plots”, the entire LiDAR dataset

(48007 cells) was split up into a number of contiguous data chunks, of varying sizes in hectares, in order to simulate typical airborne LiDAR acquisitions (i.e. the number, size and subsequent total acquisition area in ha) for country-wide CC modelling and mapping. Simulated LiDAR chunks or acquisition sizes (of contiguous data) of 250ha, 500ha, 1000ha, 2500ha, 5000ha and 10000ha and the number of these acquisition sizes were tested and compared to the modelling performance of the optimal number of simulated field plots. The analysis was conducted for both the complete and the Savannah-only extracted dataset. To create the chunks of contiguous data, a unique identifier number was assigned to all the data entries which made up a specific chunk size (varied depending on the chunk size). This unique identifier then followed a sequential numeric sequence to segment the rest of the total LiDAR dataset. These unique chunk identifiers were used for the random selection process in which the specific number of chunks was selected and used in the modelling process. Data (i.e. extracted cells of section 5.4.4) which did not fit exactly in a chunk was not included in the analysis. This strategy constrained the sampling of training data to a prescribed population size (i.e. LiDAR acquisition(s)). Data contiguity was assumed based on the manner in which the LiDAR was extracted which was geographically (i.e. spatially connected sequence of cell extraction) for the individual datasets (Table 2.1). Within each simulated LiDAR population (e.g. LiDAR acquisition size and number of acquisitions), 35% of data were selected for model training and the fixed validation dataset, used in the above sections, were again used for model testing. A cross-validation (K-fold of 10) approach was again implemented. The validation-based mean  $R^2$  and RMSE, including their confidence intervals, were plotted against the corresponding number of simulated LiDAR acquisitions for each particular LiDAR acquisition size. The optimal number of LiDAR acquisitions for each LiDAR acquisition size, was determined as the point where both the  $R^2$  and RMSE values closely matched the modelling performance of the optimal number of simulated field plots (above). This was done as a point of comparison between the optimal number of “field plots” and the equivalent, optimal LiDAR acquisition specifications.

## 5.5 Results

### 5.5.1 RF Modelling results (validation) according to modelling scenarios

**Table 5.1: Accuracies of models including combinations of various predictive variables derived from L-band HH/HV backscatter, Digital Elevation Model and rainfall**

Modelling scenarios	10 iterations						N (training)	N (validation)
	R <sup>2</sup>	R <sup>2</sup> CI	RMSE (%)	RMSE CI	SEP (%)	SEP CI		
<i>Random Selection with fixed validation and 35% training</i>								
Scenario (i) HH and HV backscatter only	0.65	0.007	23.48	0.205	42.90	0.375	16732	200
Scenario (ii) HH and HV backscatter & DEM parameters (elevation, slope & aspect)	0.74	0.004	19.19	0.163	35.07	0.297	16732	200
Scenario (iii) HH and HV backscatter & Rainfall classes (200-400, 400-600, 600-800, 800-1000 and >1000mm)	0.68	0.005	22.35	0.185	40.84	0.338	16732	200
Scenario (iv) HH and HV backscatter & DEM parameters (elevation, slope & aspect) & Rainfall classes (200-400, 400-600, 600-800, 800-1000 and >1000mm)	0.77	0.005	17.49	0.237	31.95	0.432	16732	200

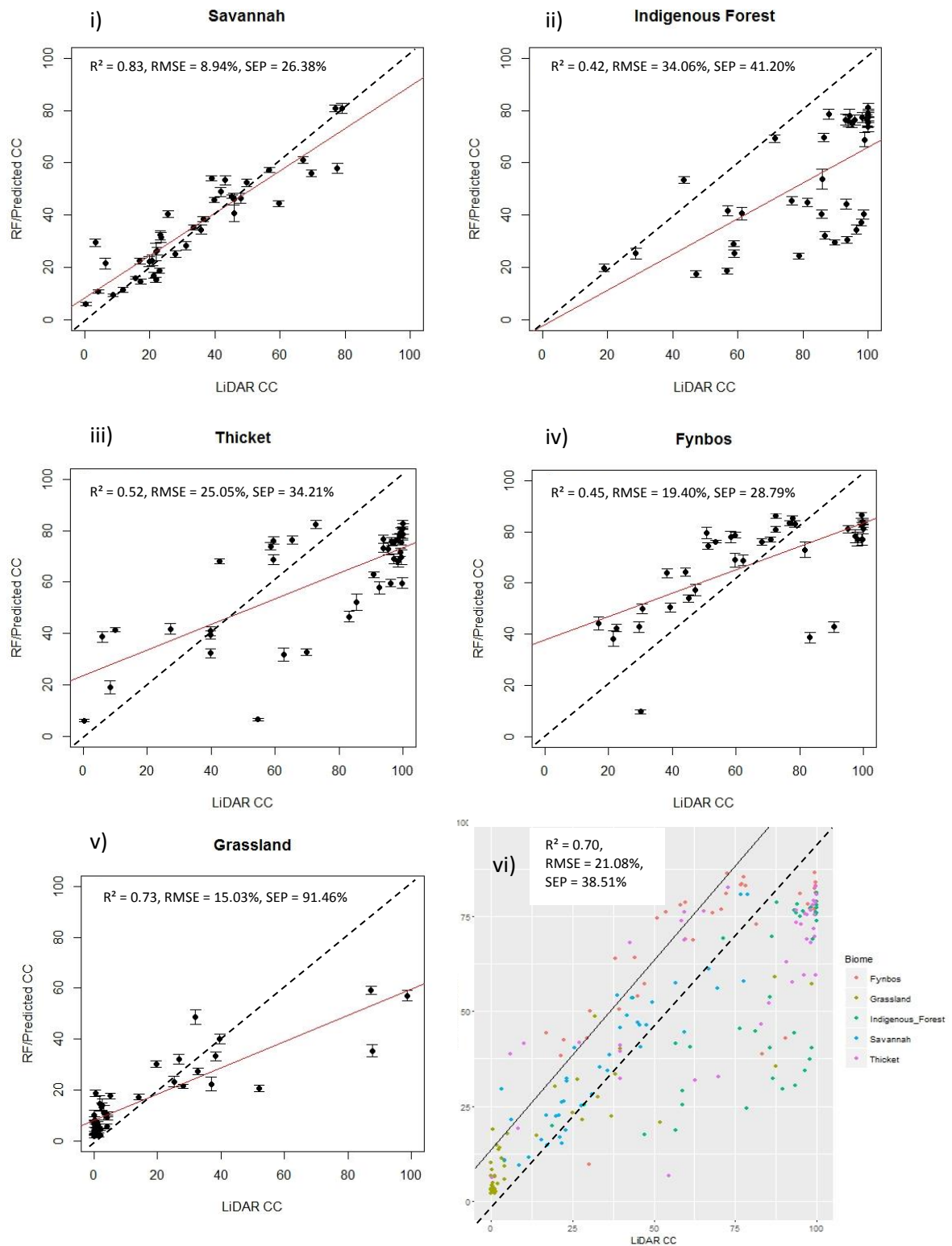
CI = confidence interval, DEM = digital elevation model, R<sup>2</sup> = coefficient of determination, RMSE = Root Mean Square Error; SEP = standard error of prediction; N = total number of observations

**Table 5.2: Accuracies of models based on training from Savannah-only versus all-biome data and validated with data from five different biomes**

Modelling scenarios (10 iterations)	Savannah-only Dataset						All-Biome Dataset						N (training)	N (validation)
	R <sup>2</sup>	R <sup>2</sup> CI	RMSE (%)	RMSE CI	SEP (%)	SEP CI	R <sup>2</sup>	R <sup>2</sup> CI	RMSE (%)	RMSE CI	SEP (%)	SEP CI		
<i>Random Selection with fixed validation and 35% training</i>														
Scenario (v) - Validated on Savannahs*	0.83	0.015	8.94	0.361	26.38	1.065	0.82	0.012	9.15	0.322	27.01	0.950	14344	40
Scenario (vi) - Validated on Grasslands*	0.73	0.016	15.03	0.276	91.46	1.680	0.82	0.016	11.59	0.465	70.51	2.832	14344	40
Scenario (vii) - Validated on Indigenous Forests*	0.42	0.015	34.06	0.857	41.20	1.036	0.55	0.021	18.81	0.758	22.75	0.917	14344	40
Scenario (viii) - Validated on Thickets*	0.52	0.011	25.05	0.317	34.21	0.433	0.53	0.008	24.61	0.334	33.60	0.456	14344	40
Scenario (ix) - Validated on Fynbos*	0.45	0.015	19.40	0.262	28.79	0.389	0.45	0.030	19.64	0.561	29.14	0.832	14344	40
Scenario (x) - Validated on complete validation dataset*	0.70	0.006	21.08	0.281	38.51	0.513	0.77	0.006	17.85	0.251	32.62	0.458	14344	200

\* Variables from scenario (iv) was used

CI = confidence interval, R<sup>2</sup> = coefficient of determination, RMSE = Root Mean Square Error; SEP = standard error of prediction; N = total number of observations



**Figure 5.4: Scatterplots of the mean predicted CC versus mean observed LiDAR derived CC resulting from models validated over five individual biomes (i-v) and the complete fixed validation dataset (vi) while using Savannah-only data for training. Error bars indicate confidence intervals of each point at the biome level. Black dotted line indicates the 1:1 trend line.**

The addition of ancillary parameters to the HH and HV SAR backscatter improved the accuracy of predicted CC (3-5% improvement in RMSE) than when using the SAR backscatter alone (Table 5.1). The DEM ancillary variables contributed the most to this improvement ( $R^2=0.74$ ;  $RMSE=19.19\%$ ;  $SEP=35.07\%$ ). However, when HH and HV SAR backscatter was combined with the DEM variables

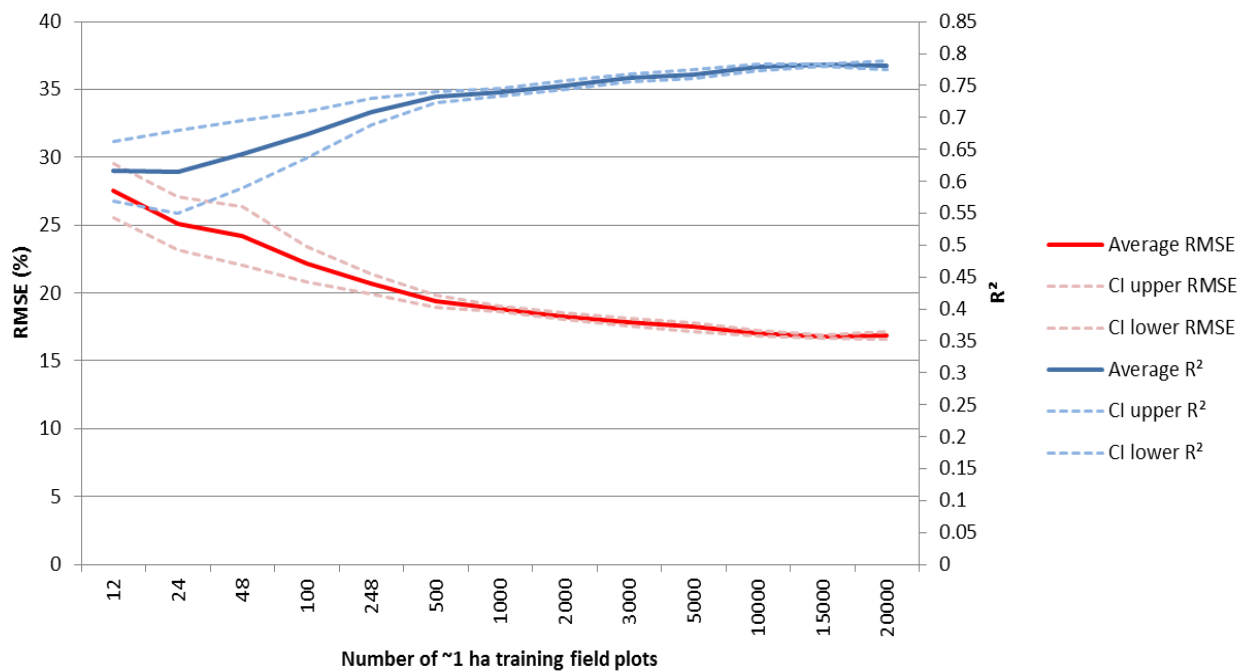


and rainfall, the highest accuracy was achieved ( $R^2=0.77$ ;  $RMSE=17.49\%$ ;  $SEP=31.95\%$ ). The confidence intervals (CI) were particularly low for the scenarios (i-iv) which indicated consistent RF modelled results. The highest accuracies were achieved in scenario (iv) and therefore this combination of variables was used for all the remaining scenarios (v-x) and analyses. Table 5.2 indicates the model accuracies achieved from the use of Savannah-only training data (including Figure 5.4) versus the complete dataset over five separate biome regions. With training data limited to the Savannahs, the modelling performance was suited to biomes with low to intermediate CC ranges as high accuracies were documented across the Savannah ( $R^2 = 0.83$ ,  $RMSE = 8.94\%$ ) and Grassland biomes ( $R^2 = 0.73$ ,  $RMSE = 15.03\%$ ). Accuracies, however, were lower for the Indigenous Forest and Thicket biomes with a trend of gross under-prediction of CC, at the high CC classes (Figures 5.4ii and iii). When training data were incorporated from the all-biome dataset, results generally improved across all biomes especially across the Indigenous Forest and Grassland biomes with an improvement of an RMSE of 11.31% and 3.44% respectively. Improvements in accuracy across the Thicket biome, however, were negligible. However, in the case of the Fynbos biome, no improvements were observed with poor accuracies being achieved when using both the all-biome and Savannah-only training datasets ( $R^2=0.45$ ;  $RMSE\sim 19.50\%$ ;  $SEP\sim 29\%$ ). The Savannah-only training dataset obtained favourable results ( $R^2=0.70$ ;  $RMSE=21.08\%$ ;  $SEP=38.51\%$ ) with an improvement achieved from including training data from all the other biomes (increases in  $R^2 = 0.07$ ,  $RMSE = 3.23\%$  and  $SEP = 5.89\%$ ).

### 5.5.2 Optimal field plot amount

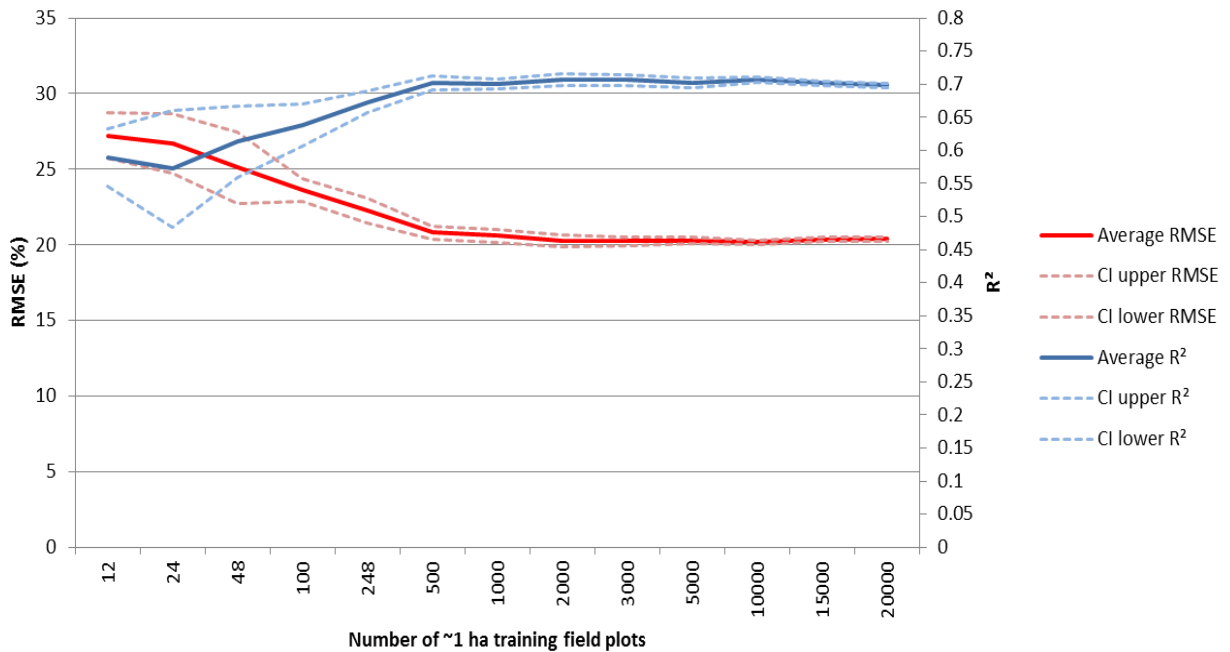
Figures 5.5 and 5.6 illustrated both the mean  $R^2$  and RMSE values, and their incremental percentage change with increasing number of simulated 1ha field plots, used as training from both the all-biome dataset and Savannah-only dataset. Regardless of the training dataset used and as expected, both the  $R^2$  and RMSE values improved with increasing number of 1ha field plots until it reached a plateau beyond which the accuracies did not improve noticeably with the increase in sampling effort. The optimal trade-off point was reached at 500 1ha field plots for the all-biome dataset, achieving a mean  $R^2$  and RMSE of 0.73 and 19.39% respectively. Although there were significant percentage improvements in  $R^2$  and RMSE of 3.25% and 6.13%, respectively, between 248 and 500 field plots, the percentage increase reduced to 0.96% and 2.96% (for  $R^2$  and RMSE, respectively) between 500 and 1000 plots (Figure 5.5). This total of 500 1ha field plots amounts to only approximately 1% of the total LiDAR dataset. According to the stratified sampling scheme, this total implies that 125 1ha

field plot sites would be required as training within each of the 0-20%, 20-40%, 40-60% and >60% CC ranges across the entire South Africa. Increasing the number of training field plots beyond 500 did however yield improvements in accuracies (not greater than 3% in percentage change in  $R^2$  and RMSE and with signs of plateauing between 10000 and 20000 1ha field plots) but it required 2-4 times the number of field plots to achieve this. The upper and lower  $R^2$  and RMSE confidence intervals (CI), also, was large at the lower number of field plots but quickly reduced until negligible at the 500 field plots (Figure 5.5). Although 500 1ha plot number was considered optimal due to the high accuracies achieved with a reasonable sampling effort, the recommended/optimal amount is ultimately informed by a user's required accuracy and budget available for field sampling.



No. 1ha training field plots	12	24	48	100	248	500	1000	2000	3000	5000	10000	15000	20000
% change $R^2$	0	-0.08	4.40	4.74	5.34	3.25	0.96	1.56	1.49	0.81	1.38	0.43	-0.02
% change RMSE	0	-8.68	-3.57	-8.63	-6.67	-6.13	-2.96	-2.78	-2.60	-1.90	-2.77	-1.27	0.38

**Figure 5.5: RF validation accuracies, including % change, across the different training sampling sizes obtained from the all-biome dataset. In the % change table, +ve values indicate a percentage increase while -ve values indicate a percentage decrease (No. = number; ha = hectares)**



No. 1ha training field plots	12	24	48	100	248	500	1000	2000	3000	5000	10000	15000	20000
% change R <sup>2</sup>	0	-2.83	7.10	4.11	5.53	4.29	-0.18	0.88	-0.13	-0.54	0.67	-0.74	-0.53
% change RMSE	0	-1.88	-5.98	-5.87	-5.80	-6.55	-1.03	-1.64	-0.14	0.25	-0.52	0.82	0.25

**Figure 5.6: RF validation accuracies, including % change, across the different training sampling sizes obtained from the Savannah-only dataset. In the % change table, +’ve values indicate a percentage increase while -’ve values indicate a percentage decrease (No. = number; ha = hectares)**

When the field plots for training was limited to the Savannah-only dataset (Scenario x), the optimal trade-off point was also reached at 500 1ha field plots (Figure 5.6). The last major improvement in R<sup>2</sup> and RMSE according to the percentage change statistics (improvements of 4.29% and 6.55% for R<sup>2</sup> and RMSE, respectively) was observed between 248 and 500 plots. The performance was slightly poorer but comparable to the results where the all-biome dataset was sampled (Figure 5.5) with mean R<sup>2</sup>, RMSE and SEP values of 0.70, 20.81% and 38.03% respectively. Similarly to Figure 5.5, the modelling performance improved marginally with the increasing number of training field plots past the 500 1ha field plot mark (less than a 1.7% improvement in R<sup>2</sup> and RMSE) but, unlike Figure 5.5, the modelling performance plateaued at the 2000 1ha field plot mark with no further, noticeable improvements in R<sup>2</sup> and RMSE. The R<sup>2</sup> and RMSE CI intervals, observed in Figure 5.6, also tapered to and remained negligible beyond the 500 1ha field plot mark.

### 5.5.3 Optimal amount of LiDAR data required

Results from the all-biome dataset, were given in Figures 5.7 and 5.8. As the results from the Savannah-only dataset yielded similar patterns to the all-biome dataset, although with slightly lower accuracies, it was briefly summarised in Table 5.3.

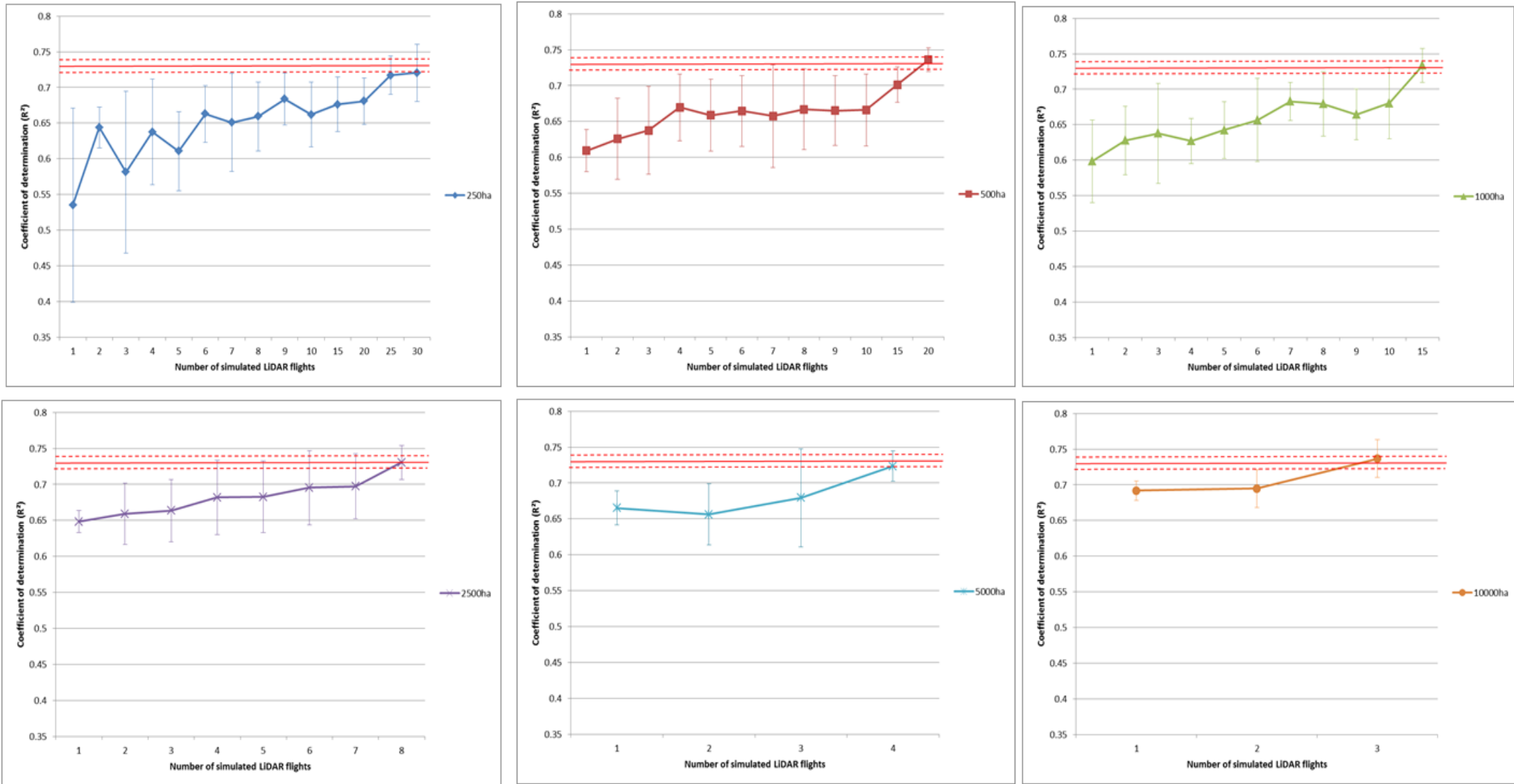


Figure 5.7: CC  $R^2$  validation accuracies according to the number of simulated LiDAR acquisitions of different sizes used for RF modelling from the all-biome dataset (scenario (iv)). The red solid line indicates the mean  $R^2$  value obtained from the 500 1ha field plot result while the red dotted line indicates the corresponding upper and lower Confidence Interval limits

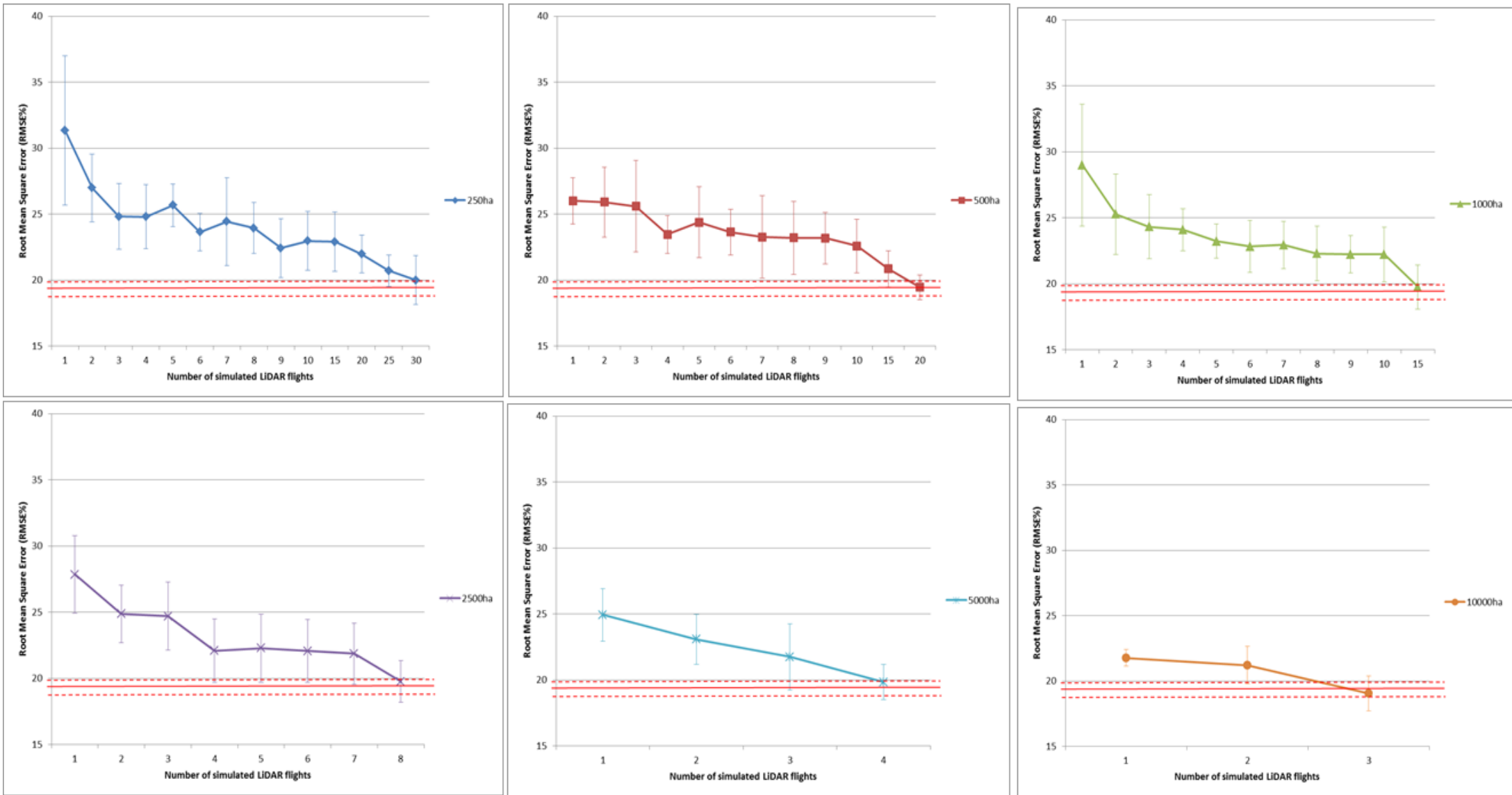


Figure 5.8: CC RMSE validation accuracies according to the number of simulated LiDAR acquisitions of different sizes used for RF modelling from the all-biome dataset (scenario iv). The red solid line indicates the mean RMSE value obtained from the 500 1ha field plot result while the red dotted lines indicate the corresponding upper and lower Confidence Interval limits.

**Table 5.3: Summarised CC RF validation accuracies according to the number, size and total hectares of simulated LiDAR acquisitions acquired from the all-biome dataset and Savannah-only dataset**

LiDAR acquisition size	No. LiDAR acquisitions	Total acquired LiDAR coverage	Complete Dataset		Savannah-only Dataset	
			R <sup>2</sup> (CI)	RMSE (%) (CI)	R <sup>2</sup> (CI)	RMSE (%) (CI)
250ha	30	7500ha	0.72 (0.040)	20.00 (1.867)	0.70 (0.008)	21.13 (0.305)
500ha	20	10000ha	0.74 (0.017)	19.45 (0.937)	0.70 (0.009)	21.08 (0.348)
1000ha	15	15000ha	0.73 (0.024)	19.75 (1.663)	0.70 (0.009)	21.09 (0.288)
2500ha	8	20000ha	0.73 (0.024)	19.76 (1.587)	0.70 (0.009)	21.11 (0.356)
5000ha	4	20000ha	0.72 (0.021)	19.84 (1.322)	0.70 (0.005)	21.11 (0.229)
10000ha	3	30000ha	0.74 (0.027)	19.05 (1.329)	0.70 (0.008)	20.99 (0.292)
<b>500 1ha plots</b>		<b>500ha</b>	<b>0.73 (0.009)</b>	<b>19.40 (0.427)</b>	<b>0.70 (0.011)</b>	<b>20.81 (0.450)</b>

Figures 5.7 and 5.8 illustrated the RF model accuracy ( $R^2$  and RMSE) with simulated LiDAR acquisitions, obtained from the all-biome dataset, of varying sizes (in terms of hectares) and increasing number of acquisitions in comparison to that of the optimal 500 1ha field plot benchmark (also see Table 5.3). It was found that both the optimal number of LiDAR acquisitions and the size or coverage was the same for both the all-biome and Savannah-only datasets in reaching the field plot benchmark but with varying obtained accuracies. The all-biome dataset, however, always obtained higher modelling accuracies than the Savannah-only dataset for all LiDAR acquisition sizes. These graphs and table illustrated a distinct trend where a larger number (e.g. 20 and 30) of smaller simulated LiDAR flight acquisitions (e.g. 250ha and 500ha) yielded a smaller overall area acquired (e.g. 7500ha and 10000ha) than the fewer number of larger acquisitions (e.g. four 5000ha LiDAR flights which equated to a total of 20000ha flown). The confidence intervals were particularly high when acquiring a limited number (1-3) of the small acquisitions (250ha-1000ha) (Figures 5.7 and 5.8). These confidence intervals remained variable as the number of acquisitions increased but eventually stabilised and reduced towards the 500 1ha benchmark. The LiDAR acquisition size of 10 000ha did show the most stable results with generally small confidence intervals being observed with the increasing number of acquisitions but suggested the highest total LiDAR acquisition coverage (30 000ha) for matching the field plot benchmark. For the sake of a more reliable recommendation on the LiDAR acquisition specifications, the mean validation accuracy result (both  $R^2$  and RMSE) which firmly occurred within the confidence range of the 500 1ha benchmark was used to select the optimal number of LiDAR acquisitions of the various sizes. Though anyone of these LiDAR specification results (i.e. number, size and total area acquired) in table 5.3 could be selected by a potential user, depending on the available budget, the four 5000ha LiDAR acquisition specification was recommended by the author. This choice was elaborated upon in the discussion section next.

## 5.6 Discussion

With the increasing financial and logistical challenges associated with acquiring sufficient training and validation datasets at the country wide scale, this chapter sought to establish a guideline for the LiDAR acquisition in terms of acquisition size, number of acquisitions, representativeness and total coverage, as well as and the quantity and representativeness of field plots, simulated from LiDAR data, within South Africa. In addition, this study also sought to establish whether sampling within the Savannah biome only compared to sampling five biomes (in five vegetation biomes), was sufficient for Random Forest (RF) model training and subsequent country-wide CC modelling. Various ancillary input variables (derived DEM parameters and rainfall information) were also tested in combination with L-band HH and HV ALOS PALSAR backscatter, and compared to SAR backscatter alone, for the RF modelling of CC.

The addition of DEM parameters (elevation, slope and aspect) to HH and HV backscatter data yielded significantly improved accuracies over the use of backscatter alone and the combination of backscatter with rainfall class information (table 5.1). It was previously demonstrated that an RF model can use the DEM parameters as a topographic landscape separator for associating particular CC ranges with particular physical landscape conditions (Li and Chen, 2005). For example, high CC values would be associated with steep slopes and riverine areas while low CC values would be associated with mid-slope areas and flat, high altitude grasslands. An example of such an observation can be seen from the transition from the Drakensberg escarpment to the Lowveld areas of Mpumalanga and is supported by (Florinsky and Kuryakova, 1996) who noted that topography controls the accumulation and availability of water in the landscape (via gravitational forces). This in turn controls vegetation distribution and cover. The final addition of rainfall class information, together with DEM parameters and backscatter data, improved RF modelling results further as areas of high rainfall generally correlated well with higher values of CC (e.g. Indigenous Forests along the east coast of SA) and vice versa for the low rainfall, low CC areas (e.g. wooded grassland, Savannah). Although rainfall and CC is positively correlated at the biome scale, examples exist where high CC values (> 70%) are found in more arid environments which have less than 500mm of rainfall (e.g. low height Albany Thickets of Eastern Cape, (Mucina and Rutherford, 2006)). Within the Savannah context, however, (Sankaran et al., 2005) demonstrated that maximum CC is constrained by rainfall, particularly areas receiving a mean annual precipitation less than 650mm, but CC is also influenced



at the landscape scale by the geology and locally by disturbance forces such as fire, human activities, and herbivory which tend to decrease the woody component (Staver et al., 2011).

In terms of the suitability of Savannah-only data for RF model training, it was evident that the accuracies were relatively poor across biomes of high CC such as Thickets and Indigenous Forest, which grossly underestimated CC (Figure 5.4, table 5.2), but high accuracies were observed for medium to low CC biomes such as Savannahs and Grasslands. When utilising the all-biome dataset (LiDAR across all five biomes) for the selection of model training data, improvements in the high CC biomes (particularly Indigenous Forests) were observed. These results were considered fair (i.e. without bias), despite an over representation of data from Savannahs, as the same number of samples were randomly selected for each biome during the validation process. This in turn implied that a Savannah-only training dataset would not be sufficient to achieve high accuracies in the high CC value range, for successful modelling of CC at the country-wide scale. Despite this apparent outcome, these results serve as a benchmark for limited sampling within a single, dominant biome for the purpose of country-wide CC modelling. The plateauing of the simulated field plot results (figure 5.6) and lower simulated LiDAR acquisition accuracy results (table 5.3) also supported the fact that the modelling of CC, at the country-scale, was limited in terms of achievable accuracies when RF training data was restricted to the Savannah biome only. Also, there were limited possibilities for further improvements in the modelling accuracy even with the increase of training data used for RF training. The simulated field plot and LiDAR acquisition results will be discussed at a later stage. An anomaly, however, was observed within the Fynbos biome modelling result in which similarly poor results were observed between the Savannah-only and the all-biome dataset. Of the total 126 samples collected over the Fynbos biome, 86 samples were available for potential selection in the training process of the all-biome dataset but no improvement was observed. A number of factors may explain this poor performance such as relief-induced variations on SAR backscatter due to the high slope angles, though this is partially corrected but not fully eliminated at the image processing phase, and the inability of the L-band to detect a range of cover of low-lying shrublands and needle-like leaved (sclerophyllous) vegetation types due to its large wavelength (~23cm) (Bayer et al., 1991; Mitchard et al., 2009).

From the optimal field plot analysis, the results adhered to the 'Law of Diminishing Returns' (Collins English Dictionary, 2012) where after reaching a particular training amount (500 1ha plots or 125

1ha plots collected within the four main CC classes), there was marginal improvement (<3%; Figures 5.5 and 5.6) in CC modelling accuracies, despite increasing training samples. In other words, past this point, the gains in RF accuracy were small with the increase in sampling effort (i.e. the number of field plots and associated logistical costs). Acquiring more field plots past this point, in order to obtain higher overall modelling accuracies, are entirely at the discretion of the data users if budget is not an issue. However, this is only recommended to sample across the five biomes rather than in savannahs only as improvements in modelling accuracy were more tangible. It is, however, very difficult to compare the 500 1ha plot amount, and the optimal LiDAR acquisition amount, with other studies as the author believes that this study is the first of its kind at a country scale to achieve this goal but the recommended optimal field plot number is not unrealistic to achieve at the country scale with a similar number of plots being acquired in (Saatchi et al., 2007). In the southern Kruger National Park region, however, (Urbazaev et al., 2015) found that a minimum of 180 samples (50m by 50m in size) was optimal (i.e. the point where the  $R^2$  and RMSE accuracies stabilised) for CC model training and validation but these results were limited to a small geographical area and a CC range predominately less than 50%. The fact that the 500 1ha field plot amount can be stratified into the collection of 125 1ha field plots randomly within broad 0-20%, 20-40%, 40-60% and >60% CC ranges means that potential users can plan field campaigns more effective and efficiently by avoiding sampling bias especially in less frequent CC classes (e.g. greater than 60% CC range in Savannahs). The use of these four broad CC classes for sampling stratification also means that users can use a variety of remote sensing services (e.g. Google Earth) and reference datasets (e.g. aerial photographs) to help with field campaign planning.

Regarding the optimal LiDAR amount results, it was clear that LiDAR data contiguity was a major factor in influencing modelling accuracies. In other words, acquiring many small sized LiDAR acquisitions (e.g. 250ha each in size) across the Savannah or all-biome dataset, yielded accuracies similar to the respective field plot benchmark with significantly lower total acquired LiDAR coverage than much larger contiguous LiDAR acquisitions (e.g. 10 000ha each in size). The number and size of the LiDAR acquisitions drive the total cost incurred so a brief cost analysis was conducted to determine the most cost effective acquisition specification. For this analysis, the conservative amount of 3 US\$ per ha cost and a fixed plane mobilisation fee (pilot rates etc.) of R40 000 per flight (roughly expected according to local providers, e.g. the Southern Mapping Company or SMC) was considered. The travelling cost of the plane, to and from the acquisition sites, was excluded as this

fee can be waived if the plane is present in the acquisition vicinity as a result of another contract (SMC, personal communication – 20 February 2017).

**Table 5.4: Cost analysis of optimal LiDAR acquisition specifications of varying size and number (according to table 5.3)**

<b>Cost Items</b>	<b>30 X 250ha</b>	<b>20 X 500ha</b>	<b>15 X 1000ha</b>	<b>8 X 2500ha</b>	<b>4 X 5000ha</b>	<b>3 X 10000ha</b>
Cost per ha (3\$ per ha; 1\$ = R12.95)	R 291 375	R 388 500	R 582 750	R 777 000	R 777 000	R 1 165 500
Mobilisation costs (R40 000 X # acquisitions)*	R 1 200 000	R 800 000	R 600 000	R 320 000	R 160 000	R 120 000
<b>Total</b>	<b>R 1 491 375</b>	<b>R 1 188 500</b>	<b>R 1 182 750</b>	<b>R 1 097 000</b>	<b>R 937 000</b>	<b>R 1 285 500</b>

\*Airplane travelling fees to and from acquisitions excluded

The acquisition of many small (e.g. 250ha) LiDAR flights is not financially feasible from an airborne LiDAR campaign perspective (~R1.5 million; table 5.4). The same can be said for very few but large LiDAR acquisition sizes such as 10 000ha which can also be expensive to acquire using an airborne platform due to acquisition duration and logistics etc. (~R1.2 million; table 5.4). Based on the number of acquisitions, not too many to incur extensive costs and not too few to lack potential CC representativeness within the sampling biome(s), and the size of these acquisition, including the total acquisition area, the four 5000ha LiDAR acquisition specification, based on the results of table 5.4 above, can be considered the most appropriate from a cost (total cost of R937 000) and statistical point of view and closely matched the accuracies obtained from the 500 1ha simulated field plots (table 5.3). Additionally, if the user is interested in capturing the CC variability across all five vegetated biomes (Savannah, Grassland, Indigenous Forests, Fynbos and Thickets) for maximum modelling accuracy potential, this LiDAR configuration would be the most feasible in achieving this. Although the eight 2500ha LiDAR acquisition specification did cover a total area of 20 000ha which was similar to that of the four 5000ha LiDAR acquisition specification, the doubling of the number of LiDAR acquisitions would be more expensive (table 5.4) especially when needing to be acquired across all five vegetated biomes. LiDAR has often been reported as expensive to acquire (Hummel et al., 2011; Kelly and Di Tommaso, 2015). However, this study illustrated that not so much, in terms of total hectare coverage (coupled with a cost effective acquisition specification), was actually needed, provided that the acquisition(s) were representative of the CC range, for accurate CC mapping at the country wide scale. This was possible as sufficient CC, and vegetation structural, variability could be present within the individual LiDAR acquisition(s) to sufficiently train the RF model. One of many examples of this was the Applebosch Ndwedwe LiDAR acquisition (approximately 650ha in size) in Figure 5.9 below which shows the wide range of CC and vegetation structural types present in a relatively small dataset. There has also been an increase in the amount of airborne LiDAR being collected over recent years for utility installations and infrastructure monitoring (e.g. powerlines,

roads, pipelines etc.). These can be sourced and made available to users at no cost, in a number of instances. With this increase in popularity and regularity of use, it is expected that the cost of LiDAR data will reduce in the near future (Asner, 2009).

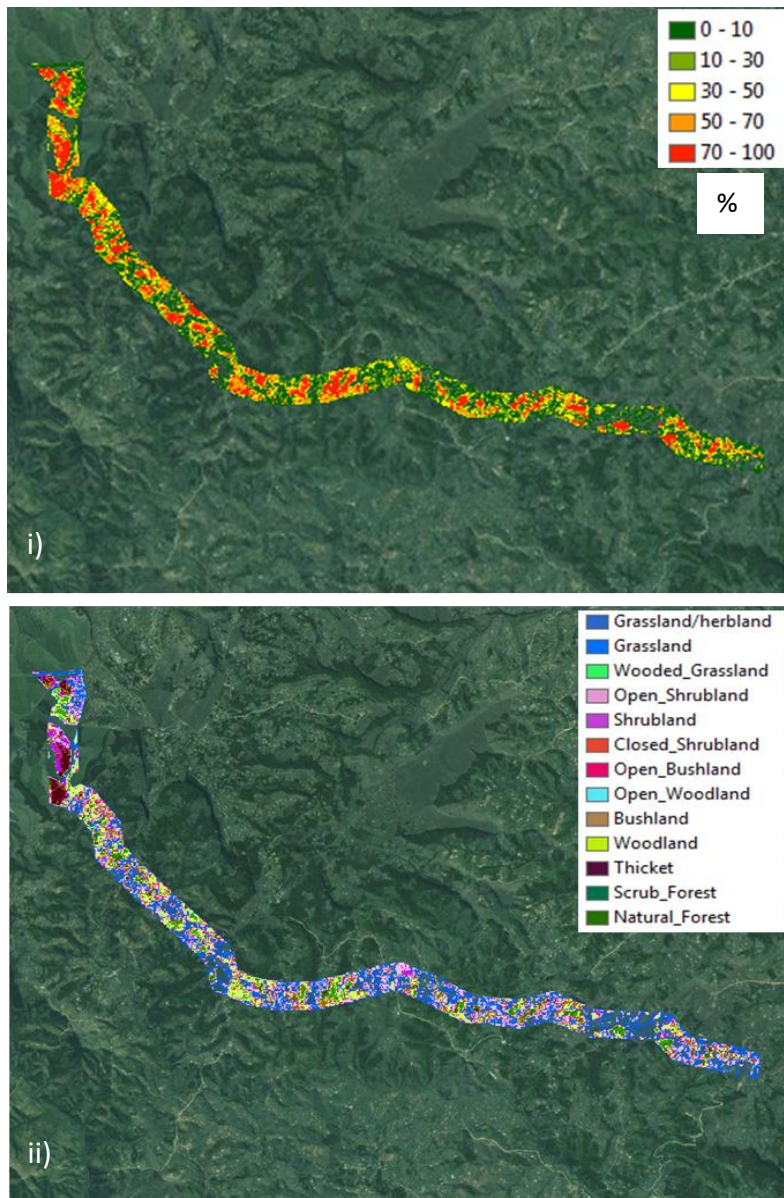


Figure 5.9: Variability of woody fractional cover (i) and vegetation structure (ii) across Applebosch Ndwedwe LiDAR

Ultimately the preferred LiDAR acquisition specification, based on the guidelines presented in Table 5.3, is dependent on the data user's preferences and budget. Overall, this outcome also supported the results from (Ene et al., 2016; Wulder et al., 2012) which discouraged the acquisition of 'wall-to-wall' LiDAR coverage and promoted for more cost effective partial coverages. To briefly test which option is the most cost effective – the optimal LiDAR acquisition specification (four 5000ha

acquisitions) or the optimal number of field plots (500 1ha field plots) – another cost analysis was conducted. Similar to the LiDAR cost assessment, field costs for a field campaign aiming to acquire CC data from 500 1ha field plots were estimated. We considered a team of three labourers and one specialist. Assuming that two field plots could be realistically collected per day by the team, a variety of per day cost items (charge out rates, transport hire, accommodation, subsistence, equipment and fuel) were considered. These figures were based on the author’s experience in organising such field campaigns in the past.

**Table 5.5: Cost analysis of the optimal number of field plots versus the optimal LiDAR acquisition specification**

<b>Items</b>	<b>500 1ha field plots</b>	<b>Items</b>	<b>4 X 5000ha LiDAR acquisitions</b>
Field team of three workers rates per day	(R300) X 3 pp. = R900	Cost per ha (3\$ per ha; 1\$ @ R12.95)	R 777 000
Team specialist rate per day	R 650	Mobilisation costs (R40 000 X 4 acquisitions)*	R 160 000
Transport hire costs per day	R 1 000		
Accommodation per day	(R600) X 4pp. = R2400		
Subsistence costs per day	(R200) X 4pp. = R800		
Fuel costs per day	R 500		
Number of days (2 sites per day)	250 days		
<i>Sub-total</i>	<i>R 1 562 500</i>		
Equipment costs (DGPS, tapes and poles)	R 80 000		
<b>Total</b>	<b>R 1 642 500</b>	<b>Total</b>	<b>R 937 000</b>

\*Airplane travelling fees to and from acquisitions excluded

Taking the cost of the variety of items outlined in table 5.5, it was evident that the 20 000ha (i.e. four 5000ha acquisitions) of acquired LiDAR data costed much less (R705 500 cheaper) than collecting data from 500 1ha plots at the country-wide scale. Acquiring LiDAR data, under the scenario considered in table 5.5, instead of field plots would yield a 43% cost saving to a potential user. Another issue considering field plot CC acquisition would be determining the amount of time and optimal size of the work force needed to sample 500 1ha plots across various CC and biome ranges. LiDAR, as a source of training and validation, appears to be both the cost effective, convenient and time-saving choice for CC modelling at the SA scale.

## 5.7 Conclusions

This study sought to establish guidelines for the quantity of LiDAR and LiDAR simulated field plots recommended as reliable sources of training and validation in SA national CC modelling efforts using HH and HV polarised L-band ALOS PALSAR global mosaic and ancillary variable data. In term of the model input variable results, it was evident that the inclusion of ancillary DEM variables (slope, aspect and elevation) and rainfall classes (200-400mm, 400-600mm, 600-800mm, 800-1000mm and >1000mm), together with HH and HV backscatter, yielded the highest RF modelling accuracies of all other input variable combinations. The sampling of RF training data from the Savannah biome-only yielded high accuracies across Grasslands and Savannahs, moderate accuracies across Thickets but poor accuracies across Indigenous Forests and Fynbos biomes. Sampling the training data across all five vegetated biomes yielded higher accuracies. From the LiDAR simulated field plot analysis, it was concluded that an optimal number of 500 1ha field plots, i.e. 125 1ha field plots equally sampled within 0-20%, 20-40%, 40-60% and >60% CC ranges, would be required for effective modelling of CC at the South African country-wide scale. Collecting additional field plots, past this point, would provide a limited increase in the overall accuracies (when using the complete dataset only for training data selection) but require significant increases in sampling efforts and costs. Concluding on the recommended optimal LiDAR amount, which matched the accuracies obtained from the 500 1ha field plots, was more challenging as a variety of LiDAR acquisition specifications (i.e. size of acquisition, number of acquisitions and total hectares acquired) could achieve this result. Choosing the best LiDAR acquisition would depend solely on the data user's available budget. By balancing the LiDAR acquisition coverage, number of acquisitions and expectant costs, the authors recommend an acquisition of four separate 5000ha LiDAR acquisitions (i.e. 20 000ha of total acquired coverage) across the five vegetated biomes. Overall, this study found that much less LiDAR data were required to train the RF models than originally expected, provided that the acquisitions were sufficiently diverse in CC and vegetation type and was also cheaper to acquire than collecting 500 1ha field plots. In conclusion, the recommendations put forward served merely as guidelines which could help reduce overall sampling effort while maximizing modelling accuracies of CC at the country-wide scale.

## Chapter 6: Study conclusions, recommendations and ways forward

This study sought to evaluate various methods for estimating and upscaling woody structural metrics of South African savannahs using integrated SAR and optical remote sensing datasets and LiDAR datasets as training and validation data sources. This study came about due to the urgent need for an active monitoring system for monitoring the status of the woody component across South Africa, especially over the savannah biome which has received less attention in the international literature. Since there are very limited research on the use of remote sensing for modelling and mapping the woody component in South Africa, this study built the foundation of fundamental knowledge and identified the initial steps for developing such a monitoring system. It is believed that this study is the first of its kind to establish accurate modelling approaches for the mapping of the woody component in the complex and heterogeneous savannah biome of South Africa, using integrated SAR and optical data. These steps involved firstly assessing existing global forest products across South Africa with a special focus on product accuracies in savannahs. The next steps involved establishing accurate SAR and integrated optical models to predict and map the woody component at the local to regional scale within the savannah environment. In the process, the most suitable SAR frequency and the best seasonal optical data source for increasing modelling accuracies, was established. The final step involved establishing some guidelines for best practices to upscale these methods to national woody component mapping while optimising the amount of training and validation data required to reliably achieve this. The current chapter serves as a summary of all the key conclusions emerging from the research conducted in the previous chapters. Recommendations and ways forward are also suggested to take this research further in both the conceptual and scientific contexts. This section will be structured on a per chapter basis.

Before attempts could be made to improve current woody structural products, the detailed validation of current global forest products was necessary to justify the overall research theme. **Chapter 2's** study sought to validate the accuracies of two global forest products, the 30m Landsat Vegetation Continuous Field (VCF) and the recently introduced 25m JAXA ALOS PALSAR Forest/Non-Forest (FNF) global products, against an extensive collection of airborne LiDAR data. Special focus was given to savannahs which are largely under-represented and even excluded by such global products. It was found that the FNF product grossly under-represented the distribution of forests in savannah environments (20-80% CC ranges), due to the inadequate HV backscatter threshold chosen in its creation (Shimada et al., 2014). The FNF product, however, most accurately detected the Non-

forest class (0-10% CC range) and also showed limited use in detecting closed forest cover class (90-100%) and Natural Forest and Shrub Forest tree structural classes. The Landsat VCF product displayed strong CC underestimation with increasing variability and mean error for CC values greater than 30%. The moderate accuracies at the 10-20% CC range and in the open woodland tree structural class suggests that the VCF product could be potentially applicable in low CC environments such as grasslands and sparse savannahs but can also marginally detect closed canopy environments (90-100% CC range).

It is recommended that more extensive ground-truth datasets, especially over medium to dense forested areas and/or specific bioregions, would need to be incorporated to train the regression tree algorithm which was used to create the Landsat VCF product. Additionally, the characterisation of CC in the Landsat VCF product could be successfully improved by integrating multi-source and multi-resolution map products as achieved for the MODIS VCF product in specific studies (Montesano et al., 2009; Song et al., 2013). For the FNF product, it is recommended that a lower HV dB threshold (such as -19 dB), used for the product derivation, be implemented to improve forest detectability in savannah environments of South Africa. This chapter provides a detailed understanding of the potential strengths and weaknesses of these two popular global forest products which is vital in allowing potential data users to make informed decisions when choosing to use these products or not. In the light of these results, a fixed definition of forests is necessary and a more accurate forest product, which has been specifically calibrated from locally collected datasets, will need to be developed to capture the full CC range found in the heterogeneous South African savannahs. Overall, there is a major need for improvement in the derivation and accuracy of such products within the context of South Africa and its savannahs. Chapters 3 to 5, in this thesis, sought to achieve this goal.

In **chapter 3**, the study aimed to test and compare the accuracy of modelling woody above ground biomass (AGB), canopy cover (CC) and total canopy volume (TCV) in South African savannahs using a combination of X-band (TerraSAR-X), C-band (RADARSAT-2) and L-band (ALOS PALSAR) radar datasets. The assessment of these three SAR frequencies (separately and in combination) has not been conducted before in a savannah environment. It was hypothesized that the combination of shorter wavelength with longer wavelength SAR datasets, in a modelling approach, will yield an improved assessment of woody structure based on the assumption that X- and C-band SAR signals



interact with the finer woody structural constituents while the L-band SAR signal interacts with the major tree structural components. In accordance with literature in other environments the L-band SAR frequency was conclusively found to be more effective in the modelling of the CC, TCV and AGB metrics in South African savannahs than the shorter wavelengths (X- and C-band) both as individual and combined (X+C-band) datasets. Although the integration of all three frequencies (X+C+L-band) yielded the best overall results for all three metrics, the improvements were noticeable but marginal in the light of the L-band results alone. C-band, however, was found to yield promising results, especially across open savannah environments, which would make the implementation of similar woody structure models which use regular, free, data obtained from the Sentinel-1 C-band sensor viable when L-band datasets are not available. The results, thus, do not warrant the acquisition of all three SAR frequency datasets for tree structure monitoring. Furthermore, the addition of the shortest wavelengths (X-band and C-band) did not assist in the overall reduction of prediction error specifically of the shrubby layer as hypothesized in the chapter study. In chapter 4 it was proposed that the inclusion of seasonal optical datasets (e.g. reflectance bands, vegetation indices and textures derived from Landsat platforms), which can provide more woody structural information, may also augment the modelling results. The inclusion of these optical predictor variables, together with the positive results of the L-band SAR in modelling CC, were tested in the next chapter.

As a way forward, in order to reduce the error experienced in the AGB estimation (at the field collection, LiDAR and SAR levels), new and more robust savannah tree allometric equations, with a greater range of representative tree stem and height sizes, will need to be produced. To elaborate this study solely implemented the Colgan (2013) AGB allometric equation for tree level AGB estimations (accumulated at plot level) for AGB upscaling efforts but this equation was limited in that it was built on trees which were sampled over a relatively small geographical area (a mining site of less than 1km<sup>2</sup> in area) in the Savannah Lowveld with a limited number of large trees harvested (i.e. DBH > 30cm and dry mass > 4 tonnes). Other equation limitations were observed in more regional allometric equations such as (Nickless et al., 2011) in which the equation was only applicable to trees with a DBH less than or equal to 33cm. Temperate or hardwood generic allometric equations are usually applied to trees greater than this maximum DBH which would introduce error at the ground level of AGB calculation. More robust allometric equations, derived from a more regionally representative sampling range which especially encompasses larger tree sizes (greater than 30cm in DBH), needs to be developed. Such efforts, however, would require extensive destructive harvesting campaigns which could be costly and conflict with ecological

management objectives, such as the preservation of biodiversity. Additionally when upscaling ground AGB to LiDAR, further research needs to be conducted on the implementation of other LiDAR-derived predictor variables (other than the H x CC metric used in this study) such as maximum height and crown area metrics. In this study, ground AGB was aggregated and upscaled to the 25m spatial resolution using the LiDAR dataset but the implementation of more advanced techniques such as individual tree crown object-oriented AGB modelling should be investigated further for improved LiDAR-derived AGB.

The research in **chapter 4**, was based on the premise that the integration of optical and SAR sensor data will yield improved results by allowing for the extraction of more detailed structural information and reducing associated uncertainty than when using the individual datasets. We mainly tested how the accuracy of woody canopy cover (CC) predictions compared when using Landsat versus L-band dual-polarised SAR input data, whether the integration of additional optical predictor features (e.g. textures and vegetation indices) improved modelling accuracies in comparison to the L-band SAR-based CC accuracies and finally, whether the integration of optical Landsat and L-band SAR data yielded any noticeable improvements. This study also sought to ascertain the season or seasons in which Landsat-5 data predicted CC with the highest accuracies. It was found that Landsat-5 imagery acquired in the summer and autumn seasons yielded the highest single season modelling accuracies, depending on the year, but the combination of multi-seasonal Landsat-5 images yielded higher accuracies. The addition of vegetation indices and image textures and their combinations to the spectral reflectance bands provided minimal improvements, with none of the optical-only combinations yielding accuracies greater than those achieved using any single winter SAR L-band image. Also due to the unpredictability of the narrow temporal 'window' during which trees and grass may differ sufficiently in phenological greenness, CC mapping and monitoring in savannahs based solely on Landsat data, is not recommended. The finding that Landsat data alone achieves significantly lower accuracies than the L-band SAR based estimates contrast with other studies in Australia and is a significant contribution to the research topic. There was significant, yet modest, improvement in accuracy when 2010 multi-seasonal optical reflectance were combined with the L-band backscatter variables. The best trade-off, however, between accuracy and complexity was given by a model using 2010 winter SAR and autumn season Landsat-5 reflectance as input variables. Extensive cloud cover, however, during the summer or even autumn seasons may adversely affect modelling accuracies by reducing the amount of available training data. It is recommended that further testing of the performance of Landsat imagery, alone and in

combination with winter SAR data, be conducted in other southern African vegetation types where tree canopies are evergreen, such as in commercial plantations, indigenous forests and thickets, and where Landsat may produce better performance. It is also recommended that a system based on L-band SAR datasets, with supporting airborne LiDAR data for model calibration and validation, should be applied to other bioregions (e.g. afro-montane and coastal indigenous forests) before a national CC monitoring programme can be established in the future. Looking at the future research opportunities, with the reduced revisit time of approximately five days and the recent launch of Sentinel 2B (7<sup>th</sup> March 2017), the Sentinel 2 series of satellites could help monitor phenological changes between tree and grasses in savannahs at a greater temporal interval than Landsat thus increasing the chance of acquiring optical imagery at the ideal phenological window which could assist L-band SAR for improved CC mapping.

Finally, **chapter 5** sought to establish guidelines for the optimal, representative sampling of airborne LiDAR data and LiDAR simulated field plots, across Savannah-only and all main biomes, for the up-scaled modelling of woody fractional cover (CC) at the country level using ALOS PALSAR L-band SAR data. The inclusion of regional environmental variables (i.e. elevation-based and rainfall variables) were also investigated for potential modelling improvements. It was found that the inclusion of ancillary DEM variables (slope, aspect and elevation) and rainfall classes (200-400mm, 400-600mm, 600-800mm, 800-1000mm and >1000mm), together with HH and HV backscatter, yielded the highest RF modelling accuracies of all other input variable combinations. Additionally, the sampling of RF training data from across the Savannah biome-only yielded high accuracies across grasslands and savannahs, moderate accuracies across thickets, but poorer accuracies across indigenous forests and fynbos biomes. Sampling the training data across all five vegetated biomes yielded higher accuracies. In terms of the LiDAR simulated field plot analysis, it was concluded that an optimum number of 500 1ha field plots, i.e. 125 1ha field plots equally sampled within 0-20%, 20-40%, 40-60% and >60% CC ranges, would be required for effective modelling of CC at the South African country-wide scale. Collecting additional field plots, past this point, would provide an added boost to the overall accuracies (when using the complete all-biome dataset only for training data selection) but at a significant increase of sampling efforts and costs which might not be warranted (cost versus accuracy paradigm). Drawing conclusions on the recommended optimum LiDAR amount, which matched the accuracies obtained from the 500 1ha field plots, was more challenging, because a variety of LiDAR acquisition specifications (i.e. size of acquisition, number of acquisitions and total hectares acquired) could achieve this result. Choosing the best LiDAR acquisition would depend

solely on the available budget. By balancing the LiDAR acquisition coverage, number of acquisitions, the authors recommend an acquisition of four separate 5000ha LiDAR acquisitions (i.e. 20 000ha of total acquired coverage) across the five vegetated biomes. According to a brief cost analysis, this LiDAR acquisition configuration was also considered to be the most cost effective. Overall, this study found that much less LiDAR data is required to train the RF models than originally expected, provided that the acquisitions were sufficiently diverse in CC and vegetation type and was also cheaper to acquire than collecting 500 1ha field plots.

In closing, this study brought forward various scientifically sound methodologies, which made use of a suite of LiDAR, SAR and optical remote sensing datasets, for estimating vegetation woody structural attributes of South African savannahs. Using the lessons learnt from the key findings above, a new woody canopy cover map product, for the country of South Africa, can be created which can be more accurate than other available global forest products. The creation of such a product can serve as an essential stepping stone towards the establishment of an operational monitoring system for South African ecosystems. In accordance with this goal, a SAR-derived CC map, using just HH and HV backscatter ( $R^2 = 0.65$ ; RMSE = 23.48%; SEP = 42.90%) as input variables, was created which is believed to be the first, locally calibrated and validated CC map created for the entire country of South Africa at the 25m spatial resolution (Figure 6.1). The next step would entail the creation of such detailed, national scale products (Figure 6.1) in a multi-temporal fashion (yearly) in order to document woody vegetation change across the years and in the process, highlight areas of vegetation loss (e.g. deforestation) and gain (e.g. bush encroachment). An additional challenge associated with multi-temporal change detection, however, would involve the appropriate management of the product error across the multiple product years and also the propagated error through the upscaling process for each year. Addressing this challenge is crucial for the understanding of actual vegetation change in the national landscape. Once established, such a monitoring system will ultimately help to address the numerous environmental issues that plague South African savannahs. These include assisting in curbing threats of bush encroachment mainly due to rising global CO<sub>2</sub> levels, deforestation through subsistence fuel wood removal, big tree loss in reserves due to elephant and fire forces and finally the spread of invasive alien species (IAPs) choking vital riparian zones as well as watersheds. Finally, the establishment of such a monitoring system will also allow the country of South Africa to meet its policy and legal obligations in terms of the regular monitoring of the status of forests and carbon stock levels at the national scale.

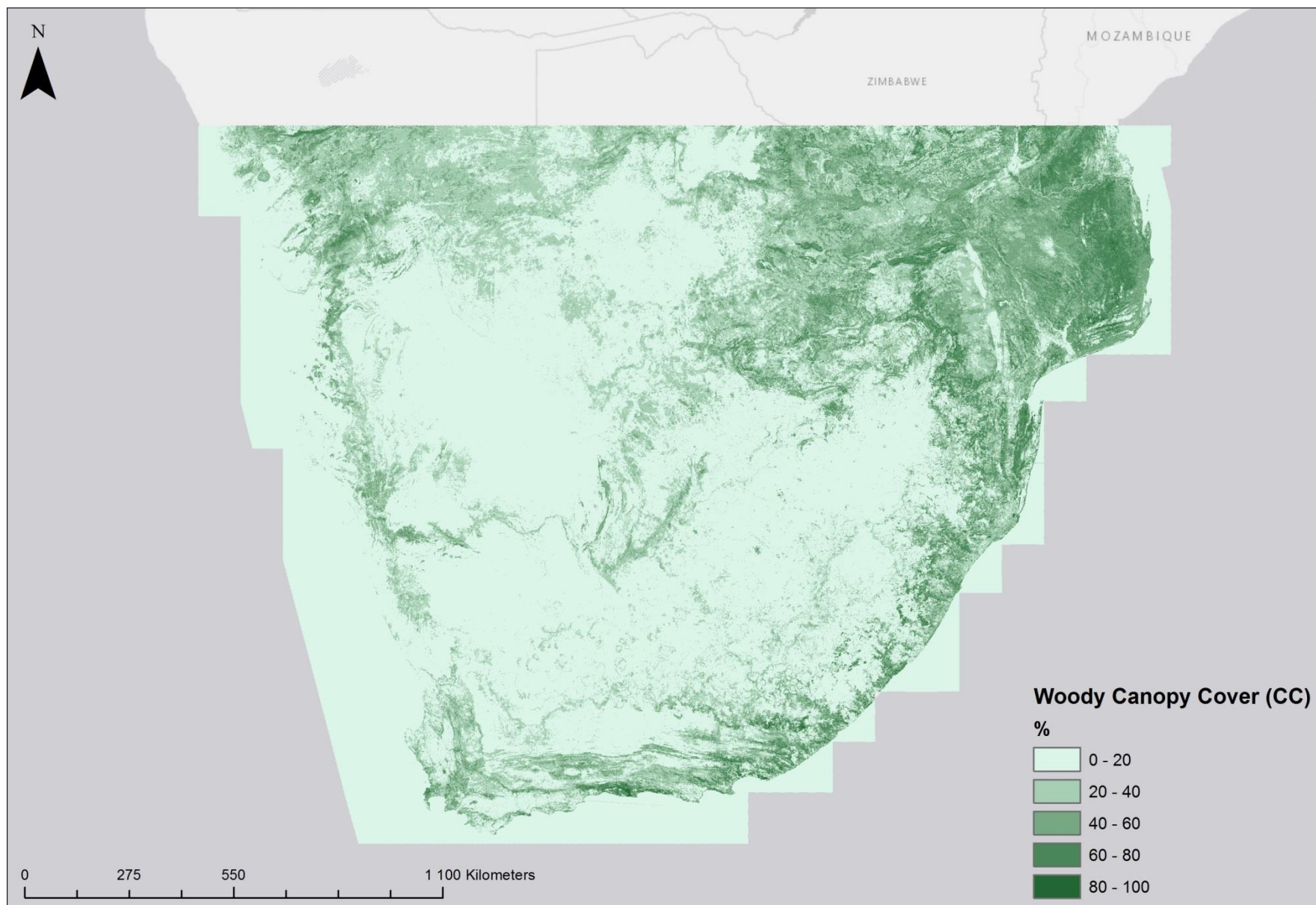


Figure 6.1: Woody fractional canopy cover (CC) map of the South African Region (including parts of neighbouring countries)

## References

- Abbot, P., Lowore, J., Werren, M., 1997. Models for the estimation of single tree volume in four Miombo woodland types. *For. Ecol. Manage.* 97, 25–37. doi:10.1016/S0378-1127(97)00036-4
- Angassa, A., 2005. The ecological impact of bush encroachment on the yield of grasses in Borana rangeland ecosystem. *Afr. J. Ecol.* 43, 14–20. doi:10.1111/j.1365-2028.2005.00429.x
- Anguita, D., Ghio, A., Greco, N., Oneto, L., Ridella, S., 2010. Model Selection for Support Vector Machines: Advantages and Disadvantages of the Machine Learning Theory, in: *IEEE Transactions on Geoscience and Remote Sensing (IGARSS 2010)*. pp. 1–8.
- Archibald, S., Scholes, R., 2007. Leaf green-up in a semi-arid African savanna – separating tree and grass responses to environmental cues. *J. Veg. Sci.* 18, 583–594.
- Armston, J.D., Denham, R.J., Danaher, T.J., Scarth, P.F., Moffiet, T.N., 2009. Prediction and validation of foliage projective cover from Landsat-5 TM and Landsat-7 ETM+ imagery. *J. Appl. Remote Sens.* 3, 033540. doi:10.1117/1.3216031
- Asner, G., Knapp, D., Kennedy-Bowdoin, T., Jones, M., Martin, R., Boardman, J., Field, C., 2007. Carnegie Airborne Observatory: in-flight fusion of hyperspectral imaging and waveform LiDAR for 3D studies of ecosystems. *J. Appl. Remote Sens.* 1, 1–27.
- Asner, G., Levick, S., Kennedy-Bowdoin, T., Knapp, D., Emerson, R., Jacobson, J., 2009. Large-scale impacts of herbivores on the structural diversity of African savannas, in: *Proceedings of the National Academy of Sciences of the United States of America*. pp. 4947–4952.
- Asner, G.P., 2009. Tropical forest carbon assessment: integrating satellite and airborne mapping approaches. *Environ. Res. Lett.* 4, 034009. doi:10.1088/1748-9326/4/3/034009
- Asner, G.P., Keller, M., Pereira, R., Zweede, J.C., 2002. Remote sensing of selective logging in Amazonia: Assessing limitations based on detailed field observations, Landsat ETM+, and textural analysis. *Remote Sens. Environ.* 80, 483–496.
- Asner, G.P., Knapp, D.E., Boardman, J., Green, R.O., Kennedy-Bowdoin, T., Eastwood, M., Martin, R.E., Anderson, C., Field, C.B., 2012. Carnegie Airborne Observatory-2: Increasing science data dimensionality via high-fidelity multi-sensor fusion. *Remote Sens. Environ.* 124, 454–465. doi:10.1016/j.rse.2012.06.012

- Asner, G.P., Levick, S.R., 2012. Landscape-scale effects of herbivores on treefall in African savannas. *Ecol. Lett.* 15, 1211–7. doi:10.1111/j.1461-0248.2012.01842.x
- Asner, G.P., Mascaro, J., Anderson, C., Knapp, D.E., Martin, R.E., Kennedy-Bowdoin, T., van Breugel, M., Davies, S., Hall, J.S., Muller-Landau, H.C., Potvin, C., Sousa, W., Wright, J., Bermingham, E., 2013. High-fidelity national carbon mapping for resource management and REDD+. *Carbon Balance Manag.* 8, 7. doi:10.1186/1750-0680-8-7
- Asner, G.P., Vaughn, N., Smit, I.P.J., Levick, S., 2016. Ecosystem-scale effects of megafauna in African savannas. *Ecography (Cop.)*. 39, 240–252. doi:10.1111/ecog.01640
- Balzter, H., Rowland, C.S., Saich, P., 2007. Forest canopy height and carbon estimation at Monks Wood National Nature Reserve, UK, using dual-wavelength SAR interferometry. *Remote Sens. Environ.* 108, 224–239. doi:10.1016/j.rse.2006.11.014
- Baudena, M., Dekker, S.C., Van Bodegom, P.M., Cuesta, B., Higgins, S.I., Lehsten, V., Reick, C.H., Rietkerk, M., Scheiter, S., Yin, Z., Zavala, M.A., Brovkin, V., 2015. Forests, savannas, and grasslands: Bridging the knowledge gap between ecology and Dynamic Global Vegetation Models. *Biogeosciences* 12, 1833–1848. doi:10.5194/bg-12-1833-2015
- Bayer, T., Winter, R., Schreier, G., 1991. Terrain influences in SAR backscatter and attempts to their correction. *IEEE Trans. Geosci. Remote Sens.* 29, 451–462. doi:10.1109/36.79436
- Boggs, G., 2010. Assessment of SPOT 5 and QuickBird remotely sensed imagery for mapping tree cover in savannas. *Int. J. Appl. Earth Obs. Geoinf.* 12, 217–224.
- Bombelli, A., Avitabile, V., Balzter, H., Beileli, L., 2009. Essential Climate Variables GTOS 67: T12 Assessment of the status of the development of the standards for terrestrial essential climate variables - Biomass, Global Terrestrial Observation System.
- Bond, W.J., Midgley, G.F., 2012. Carbon dioxide and the uneasy interactions of trees and savannah grasses. *Philos. Trans. R. Soc. B Biol. Sci.* 367, 601–612. doi:10.1098/rstb.2011.0182
- Bourgeau-chavez, L., Kasischke, E.S., Brunzell, S., Mudd, J.P., 2002. Mapping fire scars in global boreal forests using imaging radar data. *Int. J. Remote Sens.* 33, 4211–4234.
- Bradbury, R., Hill, R., Mason, D., Hinsley, S., Wilson, J., Balzter, H., 2005. Modelling relationships between birds and vegetation structure using airborne LiDAR data: A review with case studies from agricultural and woodland environments. *Ibis (Lond. 1859)*. 147, 443–452.

- Breiman, L., 2001. Random Forests. *Mach. Learn.* 45, 5–32.
- Brown, S., 1997. Estimating Biomass and Biomass Change of Tropical Forests: a Primer. (FAO Forestry Paper - 134), FAO FOREST. ed. FAO - Food and Agriculture Organization of the United Nations Rome, Rome.
- Buch, A., Dixon, A.B., 2009. South Africa's working for water programme: Searching for win-win outcomes for people and the environment. *Sustain. Dev.* 17, 129–141. doi:10.1002/sd.370
- Bucini, G., Saatchi, S., Hanan, N., Boone, R.B., Smit, I., 2009. WOODY COVER AND HETEROGENEITY IN THE SAVANNAS OF THE KRUGER NATIONAL PARK , SOUTH AFRICA, in: *IEEE Transactions on Geoscience and Remote Sensing (IGARSS)*. pp. 334–337.
- Carreiras, J., Melo, J., Vasconcelos, M., 2013. Estimating the Above-Ground Biomass in Miombo Savanna Woodlands (Mozambique, East Africa) Using L-Band Synthetic Aperture Radar Data. *Remote Sens.* 5, 1524–1548. doi:10.3390/rs5041524
- Castillo-Santiago, M., Ricker, M., de Jong, B., 2010. Estimation of tropical forest structure from SPOT-5 satellite images. *Int. J. Remote Sens.* 31, 2767–2782.
- Chazdon, R.L., 2008. Beyond Deforestation : Restoring Degraded Lands. *Communities* 1458, 1458–1460. doi:10.1126/science.1155365
- Chen, X., Vierling, L., Rowell, E., DeFelice, T., 2004. Using lidar and effective LAI data to evaluate IKONOS and Landsat 7 ETM+ vegetation cover estimates in a ponderosa pine forest. *Remote Sens. Environ.* 91, 14–26. doi:10.1016/j.rse.2003.11.003
- Chidumayo, E., Gumbo, D.J., 2010. The Dry Forests and Woodlands of Africa - Managing for Products and Services, *The dry forests and woodlands of Africa: Managing for products and services*. Earthscan, London and Washington, DC. doi:10.4324/9781849776547
- Chidumayo, E.N., 2001. Climate and phenology of savanna vegetation in southern Africa. *J. Veg. Sci.* 12, 347–354. doi:10.2307/3236848
- Cho, M., Debba, P., Mutanga, O., Duden-Tlhone, N., Magadla, T., Khuluse, S., 2012a. Potential utility of the spectral red-edge region of SumbandilaSat imagery for assessing indigenous forest structure and health. *Int. J. Appl. Earth Obs. Geoinf.* 16, 85–93.
- Cho, M., Mathieu, R., Asner, G., Naidoo, L., van Aardt, J., Ramoelo, A., Debba, P., Wessels, K., Main, R., Smit, I., Erasmus, B., 2012b. Mapping tree species composition in South African savannas



- using an integrated airborne spectral and LiDAR system. *Remote Sens. Environ.* 125, 214–226.
- Cho, M., Skidmore, A., Atzberger, C., 2008. Towards red-edge positions less sensitive to canopy biophysical parameters for leaf chlorophyll estimation using properties optiques spectrales des feuilles (PROSPECTS) and scattering by arbitrarily inclined leaves (SAILH) simulated data. *Int. J. Remote Sens.* 29, 2241–2255.
- Cho, M., Sobhan, I., Skidmore, A., 2006. Estimating fresh grass/herb biomass from HYMAP data using the red-edge position, in: *Remote Sensing and Modelling of Ecosystems for Sustainability III. Proceedings of the Society of Phot-Optical Instrumentation Engineering.* pp. 29805–29805.
- Cohen, W.B., Maersperger, T.K., Gower, S.T., Turner, D.P., 2003. An improved strategy for regression of biophysical variables and Landsat ETM+ data. *Remote Sens. Environ.* 84, 561–571.
- Colgan, M.S., Asner, G.P., Levick, S.R., Martin, R.E., Chadwick, O. a., 2012. Topo-edaphic controls over woody plant biomass in South African savannas. *Biogeosciences* 9, 1809–1821.  
doi:10.5194/bg-9-1809-2012
- Colgan, M.S., Asner, G.P., Swemmer, T., 2013. Harvesting tree biomass at the stand level to assess the accuracy of field and airborne biomass estimation in savannas. *Ecol. Appl.* 23, 1170–84.
- Collins English Dictionary, 2012. Dictionary.com “diminishing returns,” in *Collins English Dictionary - Complete & Unabridged 10th Edition.*
- Collins, J.N., Hutley, L.B., Williams, R.J., Boggs, G., Bell, D., Bartolo, R., 2009. Estimating landscape-scale vegetation carbon stocks using airborne multi-frequency polarimetric synthetic aperture radar (SAR) in the savannas of north Australia. *Int. J. Remote Sens.* 30, 1141–1159.
- Corbera, E., Schroeder, H., 2011. Governing and implementing REDD+. *Environ. Sci. Policy* 14, 89–99.  
doi:10.1016/j.envsci.2010.11.002
- DAFF, 2015. *State of the forests report 2010-2012 - Department of Agriculture, forestry and fisheries.*
- De Klerk, J.N., 2004. *Bush Encroachment in Namibia. Report on Phase 1 of the Bush Encroachment Research, Monitoring and Management Project.*
- DEA, 2010. *National Climate Change Response White Paper - Department of Environment and Natural Resources.*
- Delegido, J., Verrelst, J., Alonso, L., Moreno, J., 2011. Evaluation of sentinel-2 red-edge bands for

- empirical estimation of green LAI and chlorophyll content. *Sensors* 11, 7063–7081.  
doi:10.3390/s110707063
- Dobson, M., Ulaby, F., Le Toan, T., Beaudoin, A., Kasischke, E., Christensen, N., 1992. Dependence of radar backscatter on coniferous forest biomass. *IEEE Trans. Geosci. Remote Sens.* 30, 412–415.
- Dye, P., Versfeld, D., 2007. Managing the hydrological impacts of South African plantation forests: An overview. *For. Ecol. Manage.* 251, 121–128. doi:10.1016/j.foreco.2007.06.013
- Edwards, D., 1983. A broad-scale structural classification of vegetation for practical purposes. *Bothalia* 14, 705–712.
- Ene, L.T., Næsset, E., Gobakken, T., 2016. Simulation-based assessment of sampling strategies for large-area biomass estimation using wall-to-wall and partial coverage airborne laser scanning surveys. *Remote Sens. Environ.* 176, 328–340. doi:10.1016/j.rse.2016.01.025
- Englhart, S., Keuck, V., Siegert, F., 2011. Aboveground biomass retrieval in tropical forests — The potential of combined X- and L-band SAR data use. *Remote Sens. Environ.* 115, 1260–1271. doi:10.1016/j.rse.2011.01.008
- Falkowski, P., 2000. The Global Carbon Cycle: A Test of Our Knowledge of Earth as a System. *Science* (80-. ). 290, 291–296. doi:10.1126/science.290.5490.291
- FAO, 2015. Forest Resources Assessment Working Paper 2015 Terms and Definitions.
- FAO, 2000. FRA 2000 On Definitions of Forest and Forest Change, Food and Agricultural Organization of the United Nations.
- Fiala, A.C.S., Garman, S.L., Gray, A.N., 2006. Comparison of five canopy cover estimation techniques in the western Oregon Cascades. *For. Ecol. Manage.* 232, 188–197. doi:10.1016/j.foreco.2006.05.069
- Fisher, J.T., Erasmus, B.F.N., Witkowski, E.T.F., van Aardt, J., Wessels, K.J., Asner, G.P., 2014. Savanna woody vegetation classification - now in 3-D. *Appl. Veg. Sci.* 17, 172–184. doi:10.1111/avsc.12048
- Florinsky, I. V., Kuryakova, G.A., 1996. Influence of topography on some vegetation cover properties. *Catena* 27, 123–141. doi:10.1016/0341-8162(96)00005-7
- Foody, G.M., Lucas, R.M., Curran, P.J., Honzak, M., 1997. Mapping tropical forest fractional cover from coarse spatial resolution remote sensing imagery. *Plant Ecol.* 131, 143–154.

- Foroughbakhch, R., Carrillo Parra, A., Hernández Piñero, J.L., Alvarado Vázquez, M.A., Rocha Estrada, A., Cardenas, M.L., 2012. Wood Volume Production and Use of 10 Woody Species in Semiarid Zones of Northeastern Mexico. *Int. J. For. Res.* 2012, 1–7. doi:10.1155/2012/529829
- Freitas, S.R., Mello, M.C.S., Cruz, C.B.M., 2005. Relationships between forest structure and vegetation indices in Atlantic Rainforest. *For. Ecol. Manage.* 218, 353–362. doi:10.1016/j.foreco.2005.08.036
- Fuller, D.O., 1998. Trends in NDVI time series and their relation to rangeland and crop production in Senegal, 1987-1993. *Int. J. Remote Sens.* 19, 2013–2018. doi:10.1080/014311698215135
- Fuller, D.O., Prince, S.D., 1996. Rainfall and foliar dynamics in tropical Southern Africa: Potential impacts of global climatic change on savanna vegetation. *Clim. Change* 33, 69–96. doi:10.1007/BF00140514
- Fuller, D.O., Prince, S.D., Astle, W.L., 1997. The influence of canopy strata on remotely sensed observations of savanna-woodlands. *Int. J. Remote Sens.* 18, 2985–3009.
- Gao, Y., Mas, J.F., Paneque-gaivez, J., Skutsch, M., Ghilardi, A., Antonio, J., Paniagua, I., 2014. Validation of MODIS vegetation continuous fields in two areas in Mexico, in: *International Workshop on Earth Observation and Remote Sensing Applications*. pp. 1–5.
- Ghasemi, N., Sahebi, M.R., Mohammadzadeh, A., 2011. A review on biomass estimation methods using synthetic aperture radar data. *Int. J. Geomatics Geosci.* 1, 776–788.
- Global Forest Observations Initiative, 2016. *Integrating remote-sensing and ground-based observations for estimation of emissions and removals of greenhouse gases in forests*, 2nd ed. Food and Agriculture Organisation, Rome.
- Goel, N.S., Qin, W., 1994. Influence of canopy architecture on various vegetation indices and LAI and FPAR: a computer simulation. *Remote Sens. Rev.* 10, 309–347.
- GOFC-GOLD, 2017. *A Sourcebook of Methods and Procedures for Monitoring Essential Biodiversity Variables in Tropical Forest with Remote Sensing*. Report version UNCBD COP-13, GOFC-GOLD Land Cover Project Office, Wageningen University, The Netherlands.
- Gong, P., Pu, R., Biging, G.S., Larrieu, M.R., 2003. Estimation of forest leaf area index using vegetation indices derived from Hyperion hyperspectral data. *IEEE Trans. Geosci. Remote Sens.* 41, 1355–1362.

- Grace, J., Jose, J.S., Meir, P., Miranda, H.S., Montes, R. a., 2006. Productivity and carbon fluxes of tropical savannas. *J. Biogeogr.* 33, 387–400. doi:10.1111/j.1365-2699.2005.01448.x
- Gwenzi, D., 2017. Lidar remote sensing of savanna biophysical attributes: opportunities, progress, and challenges. *Int. J. Remote Sens.* 38, 235–257. doi:10.1080/01431161.2016.1259683
- Gwenzi, D., Lefsky, M.A., 2014. Modeling canopy height in a savanna ecosystem using spaceborne lidar waveforms. *Remote Sens. Environ.* 154, 338–344. doi:10.1016/j.rse.2013.11.024
- Hall, F.G., Bergen, K., Blair, J.B., Dubayah, R., Houghton, R., Hurtt, G., Kellndorfer, J., Lefsky, M., Ranson, J., Saatchi, S., Shugart, H.H., Wickland, D., 2011. Characterizing 3D vegetation structure from space: Mission requirements. *Remote Sens. Environ.* 115, 2753–2775. doi:10.1016/j.rse.2011.01.024
- Hanan, N.P., Sea, W.B., Dangelmayr, G., Govender, N., 2008. Do fires in savannas consume woody biomass? A comment on approaches to modeling savanna dynamics. *Am. Nat.* 171, 851–856. doi:10.1086/587527
- Hansen, M.C., DeFries, R.S., Townshend, J.R.G., Carroll, M., Dimiceli, C., Sohlberg, R. a., 2003. Global Percent Tree Cover at a Spatial Resolution of 500 Meters: First Results of the MODIS Vegetation Continuous Fields Algorithm. *Earth Interact.* 7, 1–15. doi:10.1175/1087-3562(2003)007<0001:GPTCAA>2.0.CO;2
- Hansen, M.C., DeFries, R.S., Townshend, J.R.G., Sohlberg, R., Dimiceli, C., Carroll, M., 2002. Towards an operational MODIS continuous field of percent tree cover algorithm: Examples using AVHRR and MODIS data. *Remote Sens. Environ.* 83, 303–319. doi:10.1016/S0034-4257(02)00079-2
- Hansen, M.C., Egorov, A., Potapov, P. V., Stehman, S. V., Tyukavina, A., Turubanova, S.A., Roy, D.P., Goetz, S.J., Loveland, T.R., Ju, J., Kommareddy, A., Kovalsky, V., Forsyth, C., Bents, T., 2014. Monitoring conterminous United States (CONUS) land cover change with Web-Enabled Landsat Data (WELD). *Remote Sens. Environ.* 140, 466–484. doi:10.1016/j.rse.2013.08.014
- Hansen, M.C., Egorov, A., Roy, D.P., Potapov, P., Ju, J., Turubanova, S., Kommareddy, I., Loveland, T.R., 2011. Continuous fields of land cover for the conterminous United States using Landsat data: first results from the Web-Enabled Landsat Data (WELD) project. *Remote Sens. Lett.* 2, 279–288. doi:10.1080/01431161.2010.519002
- Hansen, M.C., Loveland, T.R., 2012. A review of large area monitoring of land cover change using Landsat data. *Remote Sens. Environ.* 122, 66–74. doi:10.1016/j.rse.2011.08.024

- Hansen, M.C., Potapov, P. V, Moore, R., Hancher, M., Turubanova, S. a, Tyukavina, A., Thau, D., Stehman, S. V, Goetz, S.J., Loveland, T.R., Kommareddy, A., Egorov, A., Chini, L., Justice, C.O., Townshend, J.R.G., 2013. High-resolution global maps of 21st-century forest cover change. *Science* (80-. ). 342, 850–3. doi:10.1126/science.1244693
- Hansen, M.C., Roy, D.P., Lindquist, E., Adusei, B., Justice, C.O., Altstatt, A., 2008. A method for integrating MODIS and Landsat data for systematic monitoring of forest cover and change in the Congo Basin. *Remote Sens. Environ.* 112, 2495–2513. doi:10.1016/j.rse.2007.11.012
- Haralick, R.M., Shanmugam, K., Dinstein, I., 1973. Textural Features for Image Classification. *IEEE Trans. Syst. Man Cybern.*
- Hashemi, S.A., 2011. Biodiversity indices of natural hornbeam stands in relation to stand volume in management area. *Adv. Environ. Biol.* 5, 2527–2531.
- Higgins, S.I., Delgado-Cartay, M.D., February, E.C., Combrink, H.J., 2011. Is there a temporal niche separation in the leaf phenology of savanna trees and grasses? *J. Biogeogr.* 38, 2165–2175. doi:10.1111/j.1365-2699.2011.02549.x
- Higgins, S.I., Scheiter, S., 2012. Atmospheric CO<sub>2</sub> forces abrupt vegetation shifts locally, but not globally. *Nature* 488, 209–212. doi:10.1038/nature11238
- Hoare, D.B., Mucina, L., Rutherford, M.C., Vlok, J.H., 2006. Chapter 10: Albany Thicket Biome, in: *The Vegetation of South Africa Lesotho and Sawziland*. pp. 541–567.
- Huete, A.R., Jackson, R.D., 1988. Soil and atmosphere influences on the spectra of partial canopies. *Remote Sens. Environ.* 25, 89–105.
- Huete, A.R., Liu, H.Q., Batchilyl, K., van Leeuwen, W., 1997. A comparison of vegetation indices global set of TM images for EOS-MODIS. *Remote Sens. Environ.* 59, 440–451.
- Hummel, S., Hudak, a T., Uebler, E.H., Falkowski, M.J., Megown, K. a, 2011. A Comparison of Accuracy and Cost of LiDAR versus Stand Exam Data for Landscape Management on the Malheur National Forest. *J. For.* 109, 267–273.
- Ismail, R., Mutanga, O., Kumar, L., 2010. Modelling the potential distribution of pine forests susceptible to *Sirex Noctilo* infestations in Mpumalanga, South Africa. *Trans. GIS* 14, 709–726.
- Jennings, S., Brown, N., Sheil, D., 1999. Assessing forest canopies and understorey illumination: canopy closure, canopy cover and other measures. *Forestry* 72, 59–73.

- Jiang, Z., Huete, A.R., Didan, K., Miura, T., 2008. Development of a two-band enhanced vegetation index without a blue band. *Remote Sens. Environ.* 112, 3833–3845.  
doi:10.1016/j.rse.2008.06.006
- Jin, C., Xiao, X., Merbold, L., Arneith, A., Veenendaal, E., Kutsch, W.L., 2013. Phenology and gross primary production of two dominant savanna woodland ecosystems in Southern Africa. *Remote Sens. Environ.* 135, 189–201. doi:10.1016/j.rse.2013.03.033
- Johansen, K., Phinn, S., 2006. Mapping Structural Parameters and Species Composition of Riparian Vegetation Using IKONOS and Landsat ETM+ Data in Australian Tropical Savannas. *Photogramm. Eng. Remote Sens.* 72, 71–80.
- Jordan, C.F., 1969. Derivation of leaf area index from quality of light on the forest floor. *Ecology* 50, 663–666.
- Jose, J.S., Montes, R. a., 1997. FIRE EFFECT ON THE COEXISTENCE OF TREES AND GRASSES IN SAVANNAS AND THE RESULTING OUTCOME ON ORGANIC MATTER BUDGET. *Interciencia* 22, 289–298.
- Ju, J., Roy, D.P., Vermote, E., Masek, J., Kovalskyy, V., 2012. Continental-scale validation of MODIS-based and LEDAPS Landsat ETM+ atmospheric correction methods. *Remote Sens. Environ.* 122, 175–184. doi:10.1016/j.rse.2011.12.025
- Jung, K., Kaiser, S., Boehm, S., Nieschulze, J., Kalko, E., 2012. Moving in three dimensions: effects of structural complexity on occurrence and activity of insectivorous bats in managed forest stands. *J. Appl. Ecol.* 49, 523–531.
- Justice, C.O., Townshend, J.R.G., Holben, B.N., Tucker, C.J., 1985. Analysis of the phenology of global vegetation using meteorological satellite data. *Int. J. Remote Sens.* 6, 1271–1318.  
doi:10.1080/01431168508948281
- Kanowski, P.J., McDermott, C.L., Cashore, B.W., 2011. Implementing REDD+: lessons from analysis of forest governance. *Environ. Sci. Policy* 14, 111–117. doi:10.1016/j.envsci.2010.11.007
- Kelly, M., Di Tommaso, S., 2015. Mapping forests with Lidar provides flexible, accurate data with many uses. *Calif. Agric.* 69, 14–20. doi:10.3733/ca.v069n01p14
- Kim, D.H., Sexton, J.O., Noojipady, P., Huang, C., Anand, A., Channan, S., Feng, M., Townshend, J.R., 2014. Global, Landsat-based forest-cover change from 1990 to 2000. *Remote Sens. Environ.* 155, 178–193. doi:10.1016/j.rse.2014.08.017

- Ko, D., Bristow, N., Greenwood, D., Weisberg, P., 2009. Canopy cover estimation in semiarid woodlands: comparison of field-based and remote sensing methods. *For. Sci.* 55, 132–141.
- Laurin, G.V., Liesenberg, V., Chen, Q., Guerriero, L., Del Frate, F., Bartolini, A., Coomes, D., Wilebore, B., Lindsell, J., Valentini, R., 2013. Optical and SAR sensor synergies for forest and land cover mapping in a tropical site in West Africa. *Int. J. Appl. Earth Obs. Geoinf.* 21, 7–16.
- Le Toan, T., Quegan, S., Davidson, M.W.J., Balzter, H., Paillou, P., Papathanassiou, K., Plummer, S., Rocca, F., Saatchi, S., Shugart, H., Ulander, L., 2011. The BIOMASS mission: Mapping global forest biomass to better understand the terrestrial carbon cycle. *Remote Sens. Environ.* 115, 2850–2860. doi:10.1016/j.rse.2011.03.020
- Lee, J., 1980. Digital image enhancement and noise filtering by use of local statistics. *IEEE Trans. Pattern Anal. Mach. Intell.* 2, 165–168.
- Lefsky, M., Harding, D., Cohen, W., Parker, G., Shugart, H., 1999. Surface lidar remote sensing of basal area and biomass in deciduous forests of eastern Maryland, USA. *Remote Sens. Environ.* 67, 83–98.
- Lehmann, E. a, Caccetta, P., Lowell, K., Mitchell, A., Zhou, Z., Held, A., Milne, T., Tapley, I., 2015. Remote Sensing of Environment SAR and optical remote sensing : Assessment of complementarity and interoperability in the context of a large-scale operational forest monitoring system. *Remote Sens. Environ.* 156, 335–348. doi:10.1016/j.rse.2014.09.034
- Lehmann, E. a., Wallace, J.F., Caccetta, P. a., Furby, S.L., Zdunic, K., 2013. Forest cover trends from time series Landsat data for the Australian continent. *Int. J. Appl. Earth Obs. Geoinf.* 21, 453–462. doi:10.1016/j.jag.2012.06.005
- Li, J., Chen, W., 2005. A rule-based method for mapping Canada’s wetlands using optical, radar and DEM data. *Int. J. Remote Sens.* 26, 5051–5069. doi:10.1080/01431160500166516
- Liaw, a, Wiener, M., 2002. Classification and Regression by randomForest. *R news* 2, 18–22.
- Low, A.B., Rebelo, A.G., 1996. Vegetation of South Africa, Lesotho and Swaziland. DEAT, Pretoria.
- Lu, D., 2006. The potential and challenge of remote sensing-based biomass estimation. *Int. J. Remote Sens.* 27, 1297–1328. doi:10.1080/01431160500486732
- Lu, D., 2005. Aboveground biomass estimation using Landsat TM data in the Brazilian Amazon. *Int. J. Remote Sens.* 26, 2509–2525. doi:10.1080/01431160500142145

- Lucas, R.M., Cronin, N., Lee, A., Moghaddam, M., Witte, C., Tickle, P., 2006a. Empirical relationships between AIRSAR backscatter and LiDAR-derived forest biomass, Queensland, Australia. *Remote Sens. Environ.* 100, 407–425. doi:10.1016/j.rse.2005.10.019
- Lucas, R.M., Cronin, N., Moghaddam, M., Lee, A., Armston, J., Bunting, P., Witte, C., 2006b. Integration of radar and Landsat-derived foliage projected cover for woody regrowth mapping, Queensland, Australia. *Remote Sens. Environ.* 100, 388–406. doi:10.1016/j.rse.2005.09.020
- Lucas, R.M., Lee, a. C., Bunting, P.J., 2008. Retrieving forest biomass through integration of CASI and LiDAR data. *Int. J. Remote Sens.* 29, 1553–1577. doi:10.1080/01431160701736497
- Main, R., Mathieu, R., Kleynhans, W., Wessels, K., Naidoo, L., Asner, G.P., 2016. Hyper-temporal C-band SAR for baseline woody structural assessments in deciduous savannas. *Remote Sens.* 8, 1–19. doi:10.3390/rs8080661
- Mathieu, R., Naidoo, L., Cho, M.A., Leblon, B., Main, R., Wessels, K., Asner, G.P., Buckley, J., Van Aardt, J., Erasmus, B.F.N., Smit, I.P.J., 2013. Toward structural assessment of semi-arid African savannahs and woodlands: The potential of multitemporal polarimetric RADARSAT-2 fine beam images. *Remote Sens. Environ.* 138, 215–231.
- Matsika, R., Erasmus, B.F.N., Twine, W., 2012. A tale of two villages: assessing the dynamics of fuelwood supply in communal landscapes in South Africa. *Environ. Conserv.* 40, 71–83.
- McGlinchy, J., Aardt, J.A.N. Van, Erasmus, B., Asner, G.P., Mathieu, R., Wessels, K., Knapp, D., Kennedy-bowdoin, T., Rhody, H., Kerekes, J.P., Member, S., Ientilucci, E.J., Wu, J., Sarrazin, D., Cause-nicholson, K., 2014. Small-Footprint Waveform Lidar for Biomass Estimation in Savanna Ecosystems. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 7, 480–490.
- Menges, C.H., Bartolo, R.E., Bell, D., Hill, G.J.E., 2004. The effect of savanna fires on SAR backscatter in northern Australia. *Int. J. Remote Sens.* 25, 4857–4871. doi:10.1080/01431160410001712945
- Merganic, J., Marusak, R., Merganicova, K., Stolarikova, R., Tipmann, L., 2013. Relationship between biodiversity indicators and its economic value - Case study. *Period. Biol.* 115, 391–397.
- Mermoz, S., Rejou-Mechain, M., Villard, L., Le Toan, T., 2014. Biomass of dense forests related to L-Band SAR backscatter?, in: *International Geoscience and Remote Sensing Symposium (IGARSS 2014)*. IEEE New York, Quebec, Canada, pp. 1037–1040.
- Meyer, K.M., Wiegand, K., Ward, D., Moustakas, A., 2007. The rhythm of savanna patch dynamics. *J.*



Ecol. 95, 1306–1315. doi:10.1111/j.1365-2745.2007.01289.x

Miles, V. V., Bobylev\*, L.P., Maximov, S. V., Johannessen, O.M., Pitulko, V.M., 2003. An approach for assessing boreal forest conditions based on combined use of satellite SAR and multi-spectral data. *Int. J. Remote Sens.* 24, 4447–4466. doi:10.1080/0143116031000070436

Mills, A.J., Rogers, K.H., Stalmans, M., Witkowski, E.D.T.F., 2006. A Framework for Exploring the Determinants of Savanna and Grassland Distribution. *Bioscience* 56, 579–590.

Mitchard, E.T. a, Saatchi, S.S., Lewis, S.L., Feldpausch, T.R., Woodhouse, I.H., Sonk??, B., Rowland, C., Meir, P., 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, 2861–2873. doi:10.1016/j.rse.2010.02.022

Mitchard, E.T. a, Saatchi, S.S., White, L.J.T., Abernethy, K. a., Jeffery, K.J., Lewis, S.L., Collins, M., Lefsky, M. a., Leal, M.E., Woodhouse, I.H., Meir, P., 2012. Mapping tropical forest biomass with radar and spaceborne LiDAR in Lop?? National Park, Gabon: Overcoming problems of high biomass and persistent cloud. *Biogeosciences* 9, 179–191. doi:10.5194/bg-9-179-2012

Mitchard, E.T. a, Saatchi, S.S., Woodhouse, I.H., Nangendo, G., Ribeiro, N.S., Williams, M., Ryan, C.M., Lewis, S.L., Feldpausch, T.R., Meir, P., 2009. Using satellite radar backscatter to predict above-ground woody biomass: A consistent relationship across four different African landscapes. *Geophys. Res. Lett.* 36, L23401. doi:10.1029/2009GL040692

Moghaddam, M., Dungan, J.L., Acker, S., 2002. Forest variable estimation from fusion of SAR and multispectral optical data. *IEEE Trans. Geosci. Remote Sens.* 40, 2176–2187. doi:10.1109/TGRS.2002.804725

Mograbi, P., Asner, G., Witkowski, E., Erasmus, B., Wessels, K., Mathieu, R., Vaughn, N., 2016. Humans and elephants as treefall drivers in African savannas. *Ecography (Cop.)*. 1–11. doi:10.1111/ecog.02549

Mograbi, P.J., Erasmus, B.F.N., Witkowski, E.T.F., Asner, G.P., Wessels, K.J., Mathieu, R., Knapp, D.E., Martin, R.E., Main, R., 2015. Biomass increases go under cover: Woody vegetation dynamics in South African rangelands. *PLoS One* 10, 1–21. doi:10.1371/journal.pone.0127093

Montesano, P.M., Neigh, C.S.R., Sexton, J., Feng, M., Channan, S., Ranson, K.J., Townshend, J.R.G., 2016. Calibration and Validation of Landsat Tree Cover in the Taiga ´ Tundra Ecotone. *Remote*

Sens. 8, 5–7. doi:10.3390/rs8070551

Montesano, P.M., Nelson, R., Sun, G., Margolis, H., Kerber, a., Ranson, K.J., 2009. MODIS tree cover validation for the circumpolar taiga-tundra transition zone. *Remote Sens. Environ.* 113, 2130–2141. doi:10.1016/j.rse.2009.05.021

Mougin, E., Proisy, C., Marty, G., Fromard, F., Puig, H., Betouille, J.L., Rudant, J.P., 1999. Multifrequency and multipolarization radar backscattering from mangrove forests. *IEEE Trans. Geosci. Remote Sens.* 37, 94–102.

Mucina, L., Rutherford, M., 2006. *The vegetation of South Africa, Lesotho and Swaziland*, 1st ed. South African National Biodiversity Institute, Pretoria.

Mueller, J., Stadler, J., Brandl, R., 2010. Composition versus physiognomy of vegetation as predictors of bird assemblages: The role of lidar. *Remote Sens. Environ.* 114, 490–495.

Muss, J.D., Mladenoff, D.J., Townsend, P. a., 2011. A pseudo-waveform technique to assess forest structure using discrete lidar data. *Remote Sens. Environ.* 115, 824–835. doi:10.1016/j.rse.2010.11.008

Næsset, E., Økland, T., 2002. Estimating tree height and tree crown properties using airborne scanning laser in a boreal nature reserve. *Remote Sens. Environ.* 79, 105–115. doi:10.1016/S0034-4257(01)00243-7

Naidoo, L., Mathieu, R., Main, R., Kleynhans, W., Wessels, K., Asner, G., Leblon, B., 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. doi:10.1016/j.isprsjprs.2015.04.007

Naidoo, L., Mathieu, R., Main, R., Kleynhans, W., Wessels, K., Asner, G., Leblon, B., 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: *IEEE International Geoscience and Remote Sensing Symposium*. pp. 1–4.

Naidoo, L., Mathieu, R., Main, R., Wessels, K., Asner, G.P., 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. doi:http://dx.doi.org/10.1016/j.jag.2016.05.006

Nichol, J.E., Sarker, L.R., 2011. Improved Biomass Estimation Using the Texture Parameters of Two

- High-Resolution Optical Sensors. *IEEE Trans. Geosci. Remote Sens.* 49, 930–948.
- Nickless, A., Scholes, B., Verstraete, M., Archibald, S., Mennell, K., 2009. DETAILED STRUCTURAL CHARACTERISATION OF THE SAVANNA FLUX SITE AT SKUKUZA , SOUTH AFRICA 1 . *Ecosystem Processes and Dynamics , Natural Resources and the Environment , CSIR , PO BOX 395 , Pretoria , 0001 , South Africa* 2 . *Global Environment Monitoring , I*, in: *IEEE International Geoscience and Remote Sensing Symposium*. pp. 186–189.
- Nickless, A., Scholes, R.J., Archibald, S., 2011. A method for calculating the variance and confidence intervals for tree biomass estimates obtained from allometric equations. *S. Afr. J. Sci.* 107, 1–10. doi:10.4102/sajs.v107i5/6.356
- O'Connor, T.G., Puttick, J.R., Hoffman, M.T., 2014. Bush encroachment in southern Africa : changes and causes. *African J. Range Forage Sci.* 31, 67–88. doi:10.2989/10220119.2014.939996
- Otukey, J.R., Blaschke, T., Collins, M., 2015. Fusion of TerraSAR-x and Landsat ETM+ data for protected area mapping in Uganda. *Int. J. Appl. Earth Obs. Geoinf.* 38, 99–104. doi:10.1016/j.jag.2014.12.012
- Pengra, B., Long, J., Dahal, D., Stehman, S. V., Loveland, T.R., 2015. A global reference database from very high resolution commercial satellite data and methodology for application to Landsat derived 30m continuous field tree cover data. *Remote Sens. Environ.* 165, 234–248. doi:10.1016/j.rse.2015.01.018
- Pereira, H.M., Scharlemann, J.P.W., Al, E., 2013. Essential biodiversity variables. *Science* (80-. ). 339, 277–278. doi:10.1126/science.1229931
- Pereira, M.G., Sena, J.A., Freitas, M.A.V., Silva, N.F. Da, 2011. Evaluation of the impact of access to electricity: A comparative analysis of South Africa, China, India and Brazil. *Renew. Sustain. Energy Rev.* 15, 1427–1441. doi:10.1016/j.rser.2010.11.005
- Popescu, S.C., Zhao, K., Neuenschwander, A., Lin, C., 2011. Satellite lidar vs. small footprint airborne lidar: Comparing the accuracy of aboveground biomass estimates and forest structure metrics at footprint level. *Remote Sens. Environ.* 115, 2786–2797. doi:10.1016/j.rse.2011.01.026
- Prasad, A., Iverson, L., Liaw, A., 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems* 9, 181–199.
- Purevdorj, T., Tateishi, R., Ishiyama, T., Honda, Y., 1998. Relationships between percent vegetation cover and vegetation indices. *Int. J. Remote Sens.* 19, 3519–3535.

doi:10.1080/014311698213795

Ramsey, R.D., Wright, D.L., McGinty, C., 2004. Evaluating the Use of Landsat 30m Enhanced Thematic Mapper to Monitor Vegetation Cover in Shrub-Steppe Environments. *Geocarto Int.* 19, 39–47. doi:10.1080/10106040408542305

Richardson, D.M., Van Wilgen, B.W., 2004. Invasive alien plants in South Africa: How well do we understand the ecological impacts? *S. Afr. J. Sci.* 100, 45–52. doi:10.2307/2405025

Roberts, J., Tesfamichael, S., Gebreslasie, M., van Aardt, J., Ahmed, F., 2007. Forest structural assessment using remote sensing technologies: an overview of the current state of the art. *South. Hemisph. For. J.* 69, 183–203. doi:10.2989/SHFJ.2007.69.3.8.358

Rocchio, L.E.P., 2015. NASA/USGS Mission helps answer: What is a forest. *TerraDaily* 1.

Rosenqvist, Å., Milne, a, Lucas, R., Imhoff, M., Dobson, C., 2003. A review of remote sensing technology in support of the Kyoto Protocol. *Environ. Sci. Policy* 6, 441–455. doi:10.1016/S1462-9011(03)00070-4

Rouse, J.W., Haas, R.H., Schell, J.A., Deering, D.W., 1973. Monitoring vegetation systems in the Great Plains with ERTS, in: In: S.C. Fraden, E.P. Marcanti, M.A.B. (eds. . (Ed.), *Third ERTS-1 Symposium*. NASA SP-351, Washington D.C. NASA, pp. 309–317.

Rutherford, M., Mucina, L., Powrie, L., 2006. Chapter 3: Biomes and bioregions of southern Africa, in: *The Vegetation of South Africa Lesotho and Sawziland*. SANBI, pp. 31–51.

Ryan, C.M., Hill, T., Woollen, E., Ghee, C., Mitchard, E., Cassells, G., Grace, J., Woodhouse, I., Williams, M., 2011. Quantifying small-scale deforestation and forest degradation in African woodlands using radar imagery. *Glob. Chang. Biol.* 18, 243–257.

Ryan, C.M., Hill, T., Woollen, E., Ghee, C., Mitchard, E., Cassells, G., Grace, J., Woodhouse, I.H., Williams, M., 2012. Quantifying small-scale deforestation and forest degradation in African woodlands using radar imagery. *Glob. Chang. Biol.* 18, 243–257. doi:10.1111/j.1365-2486.2011.02551.x

Saatchi, S., Houghton, R.A., Dos Santos Alvala, R.C., Soares, J. V., Yu, Y., 2007. Distribution of aboveground live biomass in the Amazon basin. *Glob. Chang. Biol.* 13, 816–837. doi:10.1111/j.1365-2486.2007.01323.x

Saatchi, S., 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.
- Sagues, L., Fabregas, F., Broquetas, A., 2000. Extraction of vegetation structure and surface parameters using polarimetric and interferometric SAR techniques. *IEEE Int. Geosci. Remote Sens. Symp.* 141–143. doi:10.1109/IGARSS.2000.860448
- Sankaran, M., Hanan, N.P., Scholes, R.J., Ratnam, J., Augustine, D.J., Cade, B.S., Gignoux, J., Higgins, S.I., Le Roux, X., Ludwig, F., Ardo, J., Banyikwa, F., Bronn, A., Bucini, G., Caylor, K.K., Coughenour, M.B., Diouf, A., Ekaya, W., Feral, C.J., February, E.C., Frost, P.G.H., Hiernaux, P., Hrabar, H., Metzger, K.L., Prins, H.H.T., Ringrose, S., Sea, W., Tews, J., Worden, J., Zambatis, N., 2005. Determinants of woody cover in African savannas. *Nature* 438, 846–849. doi:10.1038/nature04070
- Sankaran, M., Ratnam, J., Hanan, N., 2008. Woody cover in African savannas: The role of resources, fire and herbivory. *Glob. Ecol. Biogeogr.* 17, 236–245. doi:10.1111/j.1466-8238.2007.00360.x
- Sankaran, M., Ratnam, J., Hanan, N.P., 2004. Tree-grass coexistence in savannas revisited - Insights from an examination of assumptions and mechanisms invoked in existing models. *Ecol. Lett.* 7, 480–490. doi:10.1111/j.1461-0248.2004.00596.x
- Santoro, M., Beer, C., Cartus, O., Schmullius, C., Shvidenko, A., McCallum, I., Wegmeller, U., Wiesmann, A., 2011. Retrieval of growing stock volume in boreal forest using hyper-temporal series of Envisat ASAR ScanSAR backscatter measurements. *Remote Sens. Environ.* 115, 490–507. doi:10.1016/j.rse.2010.09.018
- Santoro, M., Shvidenko, A., McCallum, I., Askne, J., Schmullius, C., 2007. Properties of ERS-1/2 coherence in the Siberian boreal forest and implications for stem volume retrieval. *Remote Sens. Environ.* 106, 154–172.
- Santos, J.R., Pardi-Lacruz, M.S., Araujo, L.S., Keil, M., 2002. Savanna and tropical rainforest biomass estimation and spatialization using JERS-1 data. *Int. J. Remote Sens.* 23, 1217–1229.
- Sawadogo, L., Savadogo, P., Tiveau, D., Dayamba, S.D., Zida, D., Nouvellet, Y., Oden, P.C., Guinko, S., 2010. Allometric prediction of above-ground biomass of eleven woody tree species in the Sudanian savanna-woodland of West Africa. *J. For. Res.* 21, 475–481. doi:10.1007/s11676-010-0101-4
- Scanlon, T.M., Caylor, K.K., Manfreda, S., Levin, S. a., Rodriguez-Iturbe, I., 2005. Dynamic response of grass cover to rainfall variability: Implications for the function and persistence of savanna

- ecosystems. *Adv. Water Resour.* 28, 291–302. doi:10.1016/j.advwatres.2004.10.014
- Schaaf, C.B., Gao, F., Strahler, A.H., Lucht, W., Li, X., Tsang, T., Strugnell, N.C., Zhang, X., Jin, Y., Muller, J.P., Lewis, P., 2002. First operational BRDF, albedo nadir reflectance products from MODIS. *Remote Sens. Environ.* 83, 135–148.
- Schmullius, C.C., Evans, D.L., 1997. Review article Synthetic aperture radar (SAR) frequency and polarization requirements for applications in ecology, geology, hydrology, and oceanography: A tabular status quo after SIR-C/X-SAR. *Int. J. Remote Sens.* 18, 2713–2722.
- Schoene, D., Killmann, W., von Lüpke, H., LoycheWilkie, M., 2007. Definitional issues related to reducing emissions from deforestation in developing countries. *Fao* 5, 1–26.
- Scholes, R., Walker, B., 1993. *An African Savanna: Synthesis of the Nylsvley Study*. Cambridge University Press, Cambridge.
- Scholes, R.J., Archer, S.R., 1997. Tree-Glass Interactions in Savannas<sup>1</sup>. *Annu. Rev. Ecol. Syst.* 28, 517–544. doi:10.1146/annurev.ecolsys.28.1.517
- Scholes, R.J., Biggs, R., 2004. *Ecosystem Services in Southern Africa: A Regional Assessment*, Southern African Millennium Ecosystem Assessment.
- Sexton, J.O., Noojipady, P., Song, X.-P., Feng, M., Song, D.-X., Kim, D.-H., Anand, A., Huang, C., Channan, S., Pimm, S.L., Townshend, J.R., 2015. Conservation policy and the measurement of forests. *Nat. Clim. Chang.* 6, 1–6. doi:10.1038/nclimate2816
- Sexton, J.O., Song, X.-P., Feng, M., Noojipady, P., Anand, A., Huang, C., Kim, D.-H., Collins, K.M., Channan, S., Dimiceli, C., Townshend, J.R., 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* 130321031236007. doi:10.1080/17538947.2013.786146
- Shackleton, C., Shackleton, S., 2004. The importance of non-timber forest products in rural livelihood security and as safety nets: A review of evidence from South Africa. *S. Afr. J. Sci.* 100, 658–664.
- Shackleton, C.M., 2000. Comparison of plant diversity in protected and communal lands in the Bushbuckridge lowveld savanna, South Africa. *Biol. Conserv.* 94, 273–285.
- Shackleton, C.M., Griffin, N.J., Banks, D.I., Mavrandonis, J.M., Shackleton, S.E., 1994. Community Structure and Species Composition Along a Disturbance Gradient in a Communally Managed South-African Savanna. *Vegetatio* 115, 157–167.

- Shackleton, C.M., Shackleton, S.E., Buiten, E., Bird, N., 2007. The importance of dry woodlands and forests in rural livelihoods and poverty alleviation in South Africa. *For. Policy Econ.* 9, 558–577.
- Shimabukuro, Y.E., Almeida-Filho, R., Kuplich, T.M., de Freitas, R.M., 2007. Quantifying optical and SAR image relationships for tropical landscape features in the Amazônia. *Int. J. Remote Sens.* 28, 3831–3840. doi:10.1080/01431160701236829
- Shimada, M., Itoh, T., Motooka, T., Watanabe, M., Shiraishi, T., Thapa, R., Lucas, R., 2014. New global forest/non-forest maps from ALOS PALSAR data (2007–2010). *Remote Sens. Environ.* 155, 13–31. doi:10.1016/j.rse.2014.04.014
- 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. doi:10.1109/JSTARS.2010.2077619
- Silva, J.F., Zambrano, A., Fariñas, M.R., 2001. Increase in the woody component of seasonal savannas under different fire regimes in Calabozo, Venezuela. *J. Biogeogr.* 28, 977–983.
- Sims, D.A., Gamon, J., 2002. Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages. *Remote Sens. Environ.* 81, 337–354.
- Skowno, A.L., Thompson, M.W., Hiestermann, J., Ripley, B., West, A.G., Bond, W.J., 2016. Woodland expansion in South African grassy biomes based on satellite observations (1990–2013): general patterns and potential drivers. *Glob. Chang. Biol.* 1–12. doi:10.1111/gcb.13529
- Song, X.-P., Huang, C., Feng, M., Sexton, J.O., Channan, S., Townshend, J.R., 2013. Integrating global land cover products for improved forest cover characterization: an application in North America. *Int. J. Digit. Earth* 00, 1–16. doi:10.1080/17538947.2013.856959
- Sousa, C.L., Ponzoni, F.J., 1998. Avaliação de índices de vegetação e de bandas TM/Landsat para estimativa de volume de madeira em floresta implantada de Pinus spp, in: IX Simposio Brasileiro de Sensoriamento Remoto. p. 11.
- Staver, A., Archibald, S., Levin, S., 2011. Tree cover in sub-Saharan Africa: rainfall and fire constrain forest and savanna as alternative stable states. *Ecology* 92, 1063–1072.
- Stevens, N., Bond, W., Hoffman, T., Midgley, G., 2015. Change is in the air: ecological trends and their drivers in South Africa. *New perspectives on global change for South Africa.*

- Stumpf, K., 1993. The estimation of forest vegetation cover descriptions using a vertical densitometer.
- Sun, G., Ranson, K.J., Guo, Z., Zhang, Z., Montesano, P., Kimes, D., 2011. Forest biomass mapping from lidar and radar synergies. *Remote Sens. Environ.* 115, 2906–2916. doi:10.1016/j.rse.2011.03.021
- Thompson, O.R.R., Paavola, J., Healy, J., Jones, J.P., Baker, T., Torres, J., 2013. Reducing Emissions from Deforestation and Forest Degradation ( REDD + ): Transaction Costs of Six Peruvian Projects 18.
- Touzi, R., Boerner, W., Lee, J., Lueneburg, E., 2004. A review of polarimetry in the context of synthetic aperture radar: concepts and information extraction. *Can. J. Remote Sens.* 30, 380–407.
- Townsend, P.A., 2002. Estimating forest structure in wetlands using multitemporal SAR. *Remote Sens. Environ.* 79, 288–304.
- Townshend, J.R., Hansen, M.C., Carroll, M., DiMiceli, C., Sohlberg, R., Huang, C., 2011. Vegetation Continuous Fields MOD44B, 2010 Percent Tree Cover [WWW Document]. Collect. 5, Univ. Maryland, Coll. Park. Maryl. URL [http://glcf.umd.edu/library/guide/VCF\\_C5\\_UserGuide\\_Dec2011.pdf](http://glcf.umd.edu/library/guide/VCF_C5_UserGuide_Dec2011.pdf) (accessed 4.8.15).
- Tsui, O.W., Coops, N.C., Wulder, M. a., Marshall, P.L., McCardle, A., 2012. Using multi-frequency radar and discrete-return LiDAR measurements to estimate above-ground biomass and biomass components in a coastal temperate forest. *ISPRS J. Photogramm. Remote Sens.* 69, 121–133. doi:10.1016/j.isprsjprs.2012.02.009
- Twine, W.C., 2005. Socio-economic transitions influence vegetation change in the communal rangelands of the South African lowveld. *African J. Range Forage Sci.* 22, 93–99. doi:10.2989/10220110509485866
- UNFCCC, 2001. Report of the Conference of the Parties on Its Seventh Session, Held At Marrakesh, Report of the conference of the parties on its seventh session, held at Marrakesh from 29 October to 10 November 2001. doi:10.1503/cmaj.109-3944
- Urbazaeu, M., Thiel, C., Mathieu, R., Naidoo, L., Levick, S.R., Smit, I.P.J., Asner, G.P., Schmullius, C., 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. doi:10.1016/j.rse.2015.06.013

- Valentini, R., Arneeth, a., Bombelli, a., Castaldi, S., Cazzolla Gatti, R., Chevallier, F., Ciais, P., Grieco, E., Hartmann, J., Henry, M., Houghton, R. a., Jung, M., Kutsch, W.L., Malhi, Y., Mayorga, E., Merbold, L., Murray-Tortarolo, G., Papale, D., Peylin, P., Poulter, B., Raymond, P. a., Santini, M., Sitch, S., Vaglio Laurin, G., Van Der Werf, G.R., Williams, C. a., Scholes, R.J., 2014. A full greenhouse gases budget of africa: Synthesis, uncertainties, and vulnerabilities. *Biogeosciences* 11, 381–407. doi:10.5194/bg-11-381-2014
- Van Wilgen, B.W., 2009. The evolution of fire management practices in savanna protected areas in South Africa. *S. Afr. J. Sci.* 105, 343–349.
- van Wilgen, B.W., Forsyth, G.G., Le Maitre, D.C., Wannenburg, A., Kotzé, J.D.F., van den Berg, E., Henderson, L., 2012. An assessment of the effectiveness of a large, national-scale invasive alien plant control strategy in South Africa. *Biol. Conserv.* 148, 28–38.  
doi:10.1016/j.biocon.2011.12.035
- Van Zyl, J.J., 1992. The effect of topography on radar scattering from vegetated areas. *Int. Geosci. Remote Sens. Symp.* 2, 1132–1134. doi:10.1109/IGARSS.1992.578363
- van Zyl, J.J., Chapman, B.D., Dubois, P., Shi, J., 1993. The Effect of Topography on SAR Calibration. *IEEE Trans. Geosci. Remote Sens.* 31, 1036–1043. doi:10.1109/36.263774
- Venter, F.J., Scholes, R.J., Eckhardt, H.C., 2003. The abiotic template and its associated vegetation pattern, in: Du Toit, J., Biggs, H., Rogers, K.H. (Eds.), *The Kruger Experience: Ecology and Management of Savanna Heterogeneity*. Island Press, London, pp. 83–129.
- Viergever, K.M., Woodhouse, I., Marino, A., Brolley, M., Stuart, N., 2008a. SAR Interferometry for estimating above ground biomass of savanna woodlands in Belize, in: *Geoscience and Remote Sensing Symposium (IGARSS)*. pp. 290–293.
- Viergever, K.M., Woodhouse, I.H., Stuart, N., 2008b. Monitoring the World's Savanna Biomass by Earth Observation. *Scottish Geogr. J.* 124, 218–225. doi:10.1080/14702540802425279
- Vollrath, A., 2010. Analysis of Woody Cover Estimations with regard to different Sensor Parameters using the SIR-C / X-SAR Dataset of Kruger National Park , RSA.
- Wang, C., Qi, J., 2008. Biophysical estimation in tropical forests using JERS-1 SAR and VNIR imagery. II. Aboveground woody biomass. *Int. J. Remote Sens.* 29, 6827–6849.  
doi:10.1080/01431160802270123

- Wang, X., Ge, L., Li, X., 2013. Pasture Monitoring Using SAR with COSMO-SkyMed, ENVISAT ASAR, and ALOS PALSAR in Otway, Australia. *Remote Sens.* 5, 3611–3636. doi:10.3390/rs5073611
- Ward, D., 2005. Do we understand the causes of bush encroachment in African savannas? *African J. Range Forage Sci.* 22, 101–105. doi:10.2989/10220110509485867
- Weepener, H., van den Berg, H., Metz, M., Hamandawana, H., 2011. The development of a hydrological improved Digital Elevation Model and derived products for South Africa based on the SRTM DEM. *Water Res. Comm. Rep. no. K5/1908* 1–52.
- Wessels, K.J., Colgan, M.S., Erasmus, B.F.N., Asner, G.P., Twine, W.C., Mathieu, R., van Aardt, J. a N., Fisher, J.T., Smit, I.P.J., 2013. Unsustainable fuelwood extraction from South African savannas. *Environ. Res. Lett.* 8, 014007. doi:10.1088/1748-9326/8/1/014007
- Wessels, K.J., Mathieu, R., Erasmus, B.F.N., Asner, G.P., Smit, I.P.J., van Aardt, J. a N., Main, R., Fisher, J., Marais, W., Kennedy-Bowdoin, T., Knapp, D.E., Emerson, R., Jacobson, J., 2011. Impact of communal land use and conservation on woody vegetation structure in the Lowveld savannas of South Africa. *For. Ecol. Manage.* 261, 19–29. doi:10.1016/j.foreco.2010.09.012
- Wigley, B., Bond, W., Hoffman, M., 2009. Bush encroachment under three contrasting land-use practices in a mesic South African savanna. *Afr. J. Ecol.* 47, 62–70.
- Williams, M.S., Patterson, P.L., Mowrer, H.T., 2003. Comparison of ground sampling methods for estimating canopy cover. *For. Sci.* 49, 235–246.
- Willis, C., 2002. Baseline Study on Woodlands in South Africa FINAL REPORT 1–51.
- Wood, E.M., Pidgeon, A.M., Radeloff, V.C., Keuler, N.S., 2012. Image texture as a remotely sensed measure of vegetation structure. *Remote Sens. Environ.* 121, 516–526. doi:10.1016/j.rse.2012.01.003
- Wu, H., Li, Z.L., 2009. Scale issues in remote sensing: A review on analysis, processing and modeling. *Sensors* 9, 1768–1793. doi:10.3390/s90301768
- Wulder, M.A., Bater, C.W., Coops, N.C., Hilker, T., White, J.C., 2008. The role of LiDAR in sustainable forest management. *For. Chron.* 84, 807–826. doi:10.5558/tfc84807-6
- Wulder, M.A., White, J.C., Bater, C.W., Coops, N.C., Hopkinson, C., Chen, G., 2012. Lidar plots — a new large-area data collection option: context, concepts, and case study. *Can. J. Remote Sens.* 38, 600–618. doi:10.5589/m12-049

Zeidler, J., Wegmann, M., Dech, S., 2012. Spatio-temporal robustness of fractional cover upscaling: a case study in semi-arid Savannah's of Namibia and Western Zambia, in: Proceedings of SPIE. pp. 1–10. doi:10.1117/12.970623

Zheng, D., Rademacher, J., Chen, J., Crow, T., Bresee, M., Le Moine, J., Ryu, S.R., 2004. Estimating aboveground biomass using Landsat 7 ETM+ data across a managed landscape in northern Wisconsin, USA. *Remote Sens. Environ.* 93, 402–411. doi:10.1016/j.rse.2004.08.008

## Appendices

### Appendix 2A: Example of poor VCF accuracy statistics according to stratified CC ranges and vegetation structural classes

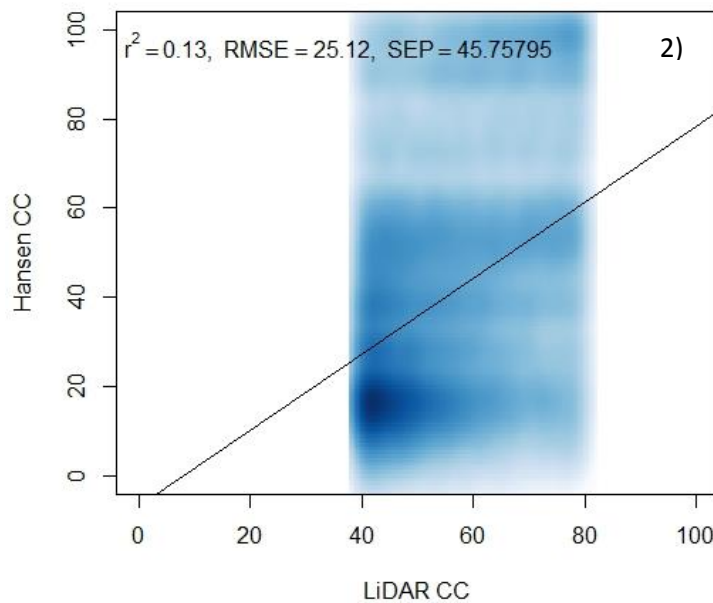
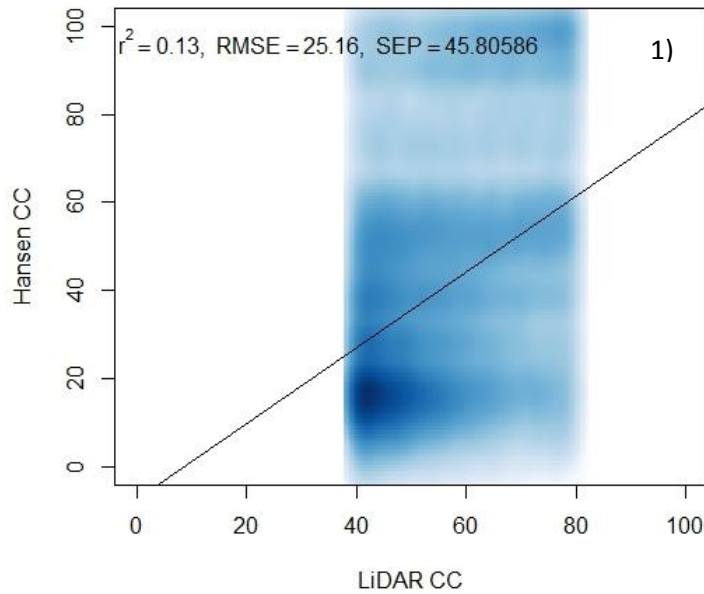


Figure 2A: XY density scatterplot of Landsat (Hansen) VCF CC values versus LiDAR CC for: 1) the 40 to 80 CC range and 2) the Woodland structural class

### Appendix 2B: FNF confusion matrix results based on new HV Forest class threshold of -19 dB

**Table 2B: Confusion matrix results of the new Forest/Non-Forest product using a HV threshold of -19dB**

New FNF	LiDAR		
	Forest (F)	Non-Forest (NF)	Grand Total
Forest (F)	21908	288	22196
Non-Forest (NF)	15060	10775	25835
Grand Total	36968	11063	48031
Producer's Acc.	59.26	97.40	
Overall Acc.			68.05

### Appendix 3A: Colgan et al. (2013) biomass allometric equation

$$M = 0.109D^{(1.39+0.14\ln(D))} H^{0.73} p^{0.80} \quad \text{Equation 3.1}$$

Where M = biomass in kg/Ha, D = Diameter above breast height (DBH) in cm, H = height of tree in metres and p = mean wood specific gravity (fixed at a mean value of 0.9) which is unitless.

### Appendix 3B: Plot level above ground biomass up-scaling factors

$$\text{Total 25m X 25m AGB plot} = X + Y + (Z*6.25) \quad \text{Equation 3.2}$$

Where X is the total AGB of stems  $\geq 10$ cm DBH, Y is total AGB of stems between 5 and 10cm DBH and Z is the total AGB of stems between 3 and 5cm DBH. The up-scaling factor of 6.25 was used as stems between 3 and 5cm were only sampled within the 10 by 10m (i.e. DBH zone 1) subplot and not sampled for the rest of the 25 by 25m grid (i.e. DBH zone 2). So  $625\text{m}^2$  (i.e. total area of the 25 X 25m sample plot) divided by  $100\text{m}^2$  (area of the 10 by 10m subplot) is 6.25. All remaining stems within the 25 by 25m sample plot, which subscribed to the remaining DBH conditions (i.e.  $\geq 5$ cm DBH), were measured and therefore did not require any up-scaling factors.

### Appendix 3C: The assessment of various data mining algorithms for modelling savannah woody cover

# THE ASSESSMENT OF DATA MINING ALGORITHMS FOR MODELLING SAVANNAH WOODY COVER USING MULTI-FREQUENCY (X-, C- AND L-BAND) SYNTHETIC APERTURE RADAR (SAR) DATASETS

*Laven Naidoo<sup>a</sup>, Renaud Mathieu<sup>a</sup>, Russell Main<sup>a</sup>, Waldo Kleynhans<sup>b</sup>, Konrad Wessels<sup>b</sup>, Gregory P. Asner<sup>c</sup>, Brigitte Leblon<sup>d</sup>*

<sup>a</sup>Ecosystem Earth Observation, Natural Resources and the Environment, CSIR, Pretoria, South Africa

Corresponding author contact details: L<sup>N</sup>aidoo@csir.co.za; (+27)12 841 2233

<sup>b</sup>Remote Sensing Unit, Meraka Institute, CSIR, Pretoria, South Africa

<sup>c</sup>Department of Global Ecology, Carnegie Institution for Science, Stanford, CA, USA

<sup>d</sup>Faculty of Forestry and Environmental Management, University of New Brunswick, Fredericton, Canada

## ABSTRACT

The woody component in African Savannas provides essential ecosystem services such as fuel wood and construction timber to large populations of rural communities. Woody canopy cover (i.e. the percentage area occupied by woody canopy or CC) is a key parameter of the woody component. Synthetic Aperture Radar (SAR) is effective at assessing the woody component, because of its capacity to image within-canopy properties of the vegetation while offering an all-weather capacity to map relatively large extents of the woody component. This study compared the modelling accuracies of woody canopy cover (CC), in South African Savannas, through the assessment of a set of modelling approaches (Linear Regression, Support Vector Machines, REPTree decision tree, Artificial Neural Network and Random Forest) with the use of X-band (TerraSAR-X), C-band (RADARSAT-2) and L-band (ALOS PALSAR) datasets. This study illustrated that the ANN, REPTree and RF non-parametric modelling algorithms were the most ideal with high CC prediction accuracies throughout the different scenarios. Results also illustrated that the acquisition of L-band data be prioritized due to the high accuracies achieved by the L-band dataset alone in comparison to the individual shorter wavelengths. The study provides promising results for developing regional savannah woody cover maps using limited LiDAR training data and SAR images.

*Index Terms*— *Woody canopy cover, Savannahs, Synthetic Aperture Radar, Multi-frequency, Non-parametric*

## 1. INTRODUCTION – BACKGROUND, AIMS AND OBJECTIVES

The woody component in African Savannahs provides essential ecosystem services such as fuelwood and construction timber to large populations of rural communities. The woody component is also an important physical attribute for many ecological processes and impact the fire regime, vegetation production, nutrient cycling, soil erosion and the water cycle of these environments [1]. In order to monitor and manage these fuelwood reserves and carbon stock, the structural parameters of the woody components needs to be estimated over large areas. Woody canopy cover (i.e. the percentage area occupied by woody canopy or CC) is a simple and key parameter of the woody component and is used for the estimation of above ground biomass by combining it with tree height [2].

Active remote sensing sensors such as Light Detection And Ranging (LiDAR) and Synthetic Aperture Radar (SAR) are effective at assessing the woody component, because of their capacity to image within-canopy properties of the vegetation [3], [4], [5]. SAR-based approach, furthermore, offers an all-weather capacity to map relatively large extents of the woody component, which cannot be easily achieved with LiDAR only [6]. In line with the protocols outlined in the GOFCC-GOLD Sourcebook [7], for extensive regional CC modelling, mapping potential and capacity to incorporate such diverse datasets, a robust but accurate modelling approach is needed. Both parametric and non-parametric modelling approaches can fulfill this criterion. Parametric approaches are based on particular assumptions about the input variable(s) distribution while in non-parametric approaches, the input variable(s) do not take a predetermined form but are built from information derived from the dataset(s) itself [8].

This study compared the modelling accuracies of woody canopy cover (CC), in South African Savannahs, through the assessment of a set of modelling approaches (from simple parametric Linear Regression to more complex non-parametric algorithms such as Support Vector Machines, REPTree decision tree, Artificial Neural Network and Random Forest) with the use of X-band (TerraSAR-X), C-band (RADARSAT-2) and L-band (ALOS PALSAR) datasets. Since this work feeds into a bigger programme for robust CC modelling development and automated mapping potential, minimal

algorithm parameter tuning and optimization was conducted. With this in mind, the default parameter values recommended by the various software proprietors were thus used in this study. Finally, CC was derived from airborne LiDAR data to train the models and evaluate the SAR modelling accuracies. The following research questions were posed in accordance to this study's main objectives:

- 1) Which modelling technique yielded the best CC modelled accuracies?
- 2) Which SAR frequency (e.g. X-, C- or L-band) yielded the highest accuracies for predicting CC?

## **2. MATERIALS AND METHODOLOGY**

Five 2012 TerraSAR-X X-band (Dual pol. StripMap), four 2009 Radarsat-2 C-band (Qual pol. Fine beam but only HH and HV data was used in this study) and two 2010 ALOS PALSAR L-band (Dual pol. FBD) images were acquired for the Southern Kruger National Park region (31°00' to 31°50' Long E; 24°33' to 25°00' Lat S). This area is made up of a mixture of communal rangelands (e.g. Bushbuckridge), private game reserves (e.g. Sabi Sands) and national parks (e.g. Kruger Park). The woody vegetation in the region is generally characterized as open forest with a canopy cover ranging from 20-60%, a predominant height range of 2 to 5m and biomass below 60T/ha [9]. The SAR imagery was acquired in winter when it is dry with the lowest moisture levels and leaf-off conditions. Dry conditions allow for minimal SAR signal noise from moisture variability [9]. The SAR intensity imagery underwent the following pre-processing steps: multi-looking (range and azimuth factor of 2:8 for L-band, 1:5 for C-band and 4:4 for X-band), radiometric calibration (conversion into  $\sigma^0$  backscatter values), geocoding and topographically normalization of the backscatter (90m SRTM DEM) and filtering (3X3m sigma Lee filter).

LiDAR data were acquired by the Carnegie Airborne Observatory AToMS sensor in summer 2012 and processed according to steps outlined in [10]. The LiDAR CC product was derived from a Canopy height model (CHM, pixel size of 1.12m) that was computed by subtracting a DEM from a Canopy Surface Model obtained from the raw point cloud. The percentage area of 25 x 25m area covered by woody canopy was calculated (using the CHM values above 0.5m to exclude the grass layer) to create the LiDAR CC product. For the modelling, the LiDAR CC and SAR datasets were combined using a fixed spatial grid of 105m cells, spaced 50m apart to avoid spatial autocorrelation [9].



Polygon shapefiles of the informal settlements, the main roads, rivers and dams were used to remove any grid cells occupying those features. Mean values within each 105m cell were extracted from the SAR and LiDAR CC datasets. This resulted in a dataset of approximately 21000 samples.

Five popular regression and data mining algorithms were applied to specific scenarios derived from the extracted data: linear regression (LR) [11], Support Vector Machines (SVM) [12], REPTree [13], Artificial Neural Network (ANN) [14] and Random Forest (RF) [15]. LR is the simplest to implement but are sensitive to outliers and are not suited to non-linearly distributed data. ANN (a feed-forward version used in this study with the hidden layer nodes set at a default value of 10), SVM (Polykernel algorithm with default RegSMOImproved optimizer) and RF are more suited to complex datasets but are 'black-box' in nature with specific software requirements. Additionally ANN and SVM are more computationally intensive and time consuming due to the level of complexity and customization that is required [16], [17]. REPTree decision tree (unconstrained with a default value of 3 number of folds for growing the rule set) have also been proven to be an effective technique [18] but, like most decision tree algorithms, are sensitive to small changes in the training datasets and are vulnerable to overfitting [19]. RF, however, is easier to implement as it only requires two main user-defined inputs – the number of trees in the forest (default = 500 trees) and the number of possible splitting variables for each node (default rule is the square root of number of predictor variables used i.e. 1 in this study) [20].

The various data input scenarios included X-band, C-band and L-band only. Models were computed in WEKA 3.6.9 and R rattle software. Data were split into a random 35% for model training and random 65% for model validation. The entire modelling process was repeated 10 times for robustness and cross-validation (allowing varying training/validation datasets) while calculating averaged coefficient of determination ( $R^2$ ), root mean square error (RMSE) and standard error of prediction (SEP) statistics (including their 95% confidence intervals or CI). Average predicted CC versus observed CC plots was also created.

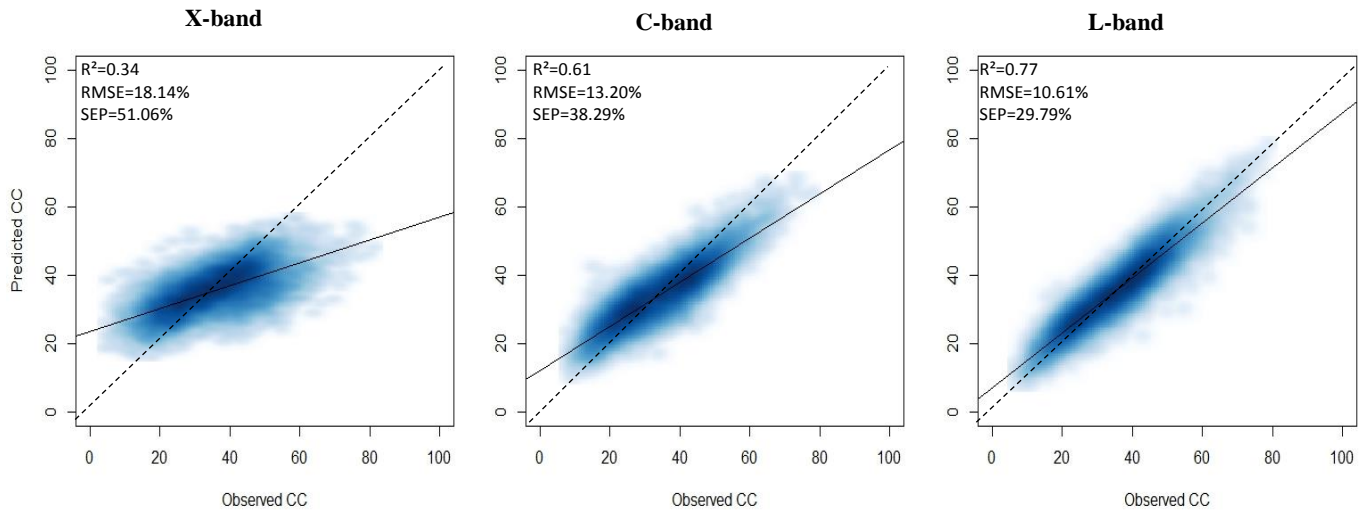


Figure 1: Mean RF predicted CC versus mean observed CC for each multi-frequency scenario (The dotted line refers to the 1:1 line)

Table 1: Validation accuracies for modelling CC across various SAR frequencies and algorithms (N= no. of observations)

Band	<b>X [N = 13761]</b>			<b>C [N = 11687]</b>		
Algorithm	<b>R<sup>2</sup> (CI)</b>	<b>RMSE (CI)</b>	<b>SEP (CI)</b>	<b>R<sup>2</sup> (CI)</b>	<b>RMSE (CI)</b>	<b>SEP (CI)</b>
<b>LR</b>	0.30 (0.002)	18.57 (0.023)	52.18 (0.084)	0.55 (0.002)	14.04 (0.034)	40.88 (0.123)
<b>SVM</b>	0.30 (0.002)	18.72 (0.036)	52.68 (0.112)	0.55 (0.002)	14.48 (0.099)	42.09 (0.280)
<b>REPTree</b>	0.36 (0.005)	17.74 (0.089)	49.86 (0.282)	0.63 (0.002)	12.91 (0.032)	37.53 (0.127)
<b>ANN</b>	0.39 (0.009)	17.29 (0.152)	48.52 (0.394)	0.65 (0.002)	12.56 (0.033)	36.50 (0.090)
<b>RF</b>	0.34 (0.003)	18.14 (0.040)	51.06 (0.153)	0.61 (0.002)	13.20 (0.031)	38.29 (0.117)
Band	<b>L [N = 13954]</b>					
Algorithm	<b>R<sup>2</sup> (CI)</b>	<b>RMSE (CI)</b>	<b>SEP (CI)</b>			
<b>LR</b>	0.71 (0.002)	11.88 (0.050)	33.36 (0.154)			
<b>SVM</b>	0.71 (0.003)	12.34 (0.083)	34.65 (0.246)			
<b>REPTree</b>	0.78 (0.002)	10.40 (0.045)	29.16 (0.145)			
<b>ANN</b>	0.79 (0.003)	10.15 (0.066)	28.49 (0.178)			
<b>RF</b>	0.77 (0.001)	10.61 (0.027)	29.79 (0.075)			

### 3. RESULTS AND DISCUSSION

In terms of the modelling algorithm results (table 1), LR and SVM both yielded poorer accuracies in comparison to REPTree, ANN and RF algorithms which obtained similarly high accuracies. This indicated that the implementation of mostly non-parametric algorithms (particularly ANN) were most suited for modelling CC in this heterogeneous savannah environment. LR performed poorly due to the fact that the relationships between the SAR predictor variables and CC were not linear (results not shown) while SVM's poor performance could be attributed to insufficient learning or training by the algorithm (requires the tuning of 'hyperparameters') [17]. Additional experimentation to find the optimal algorithm parameters (e.g. selecting a more effective kernel

algorithm and optimizer), instead of the implementation of the default parameters, could also have improved the SVM results. Preliminary results also showed that when datasets were combined, RF yielded higher accuracies than the other algorithms examined in this study, which indicate that RF is more suited for larger predictor datasets (to be explored in upcoming publications). Additionally, the overall low CI values indicated that the derived models were very robust and stable across the various iterations.

For the individual SAR frequencies, the L-band dataset yielded the highest modelled accuracies across all algorithms with the X-band dataset yielding the poorest results. This L-band result can be attributed to the ability of longer wavelengths to interact with the main tree structural constituents (particularly in tree canopies with patchy crown architectures of which the shorter wavelengths might not fully capture) thus resulting in a better correlation with the LiDAR CC metric. These modelling results were supported by the mean predicted versus mean observed CC scatterplots for each scenario (figure 1 – RF results). The levels of major CC over-prediction and under-prediction (in relation to the dotted 1:1 line where predicted CC equals observed CC) noticeably improved as one progressed from the X-band plot to the C-band and to finally the L-band band plot. These modelling results highlighted the important contribution of the L-band in CC modelling in this environment. The preference for L-band SAR datasets for tree structure modelling has been supported by numerous studies [21], [22] and this study's outcome corroborated those in [23]. The study provides promising results for developing regional savannah woody cover maps using limited LiDAR training data and SAR images.

#### **4. CONCLUDING REMARKS**

This study illustrated that the ANN, REPTree and RF non-parametric modelling algorithms were found to be robust while yielding consistently higher CC prediction accuracies throughout the different band scenarios. One of these algorithms could be implemented for continuous mapping potential of CC when future datasets become available. Results also illustrated that the acquisition of L-band data should be prioritized due to the high accuracies achieved by the L-band dataset alone in comparison to the individual shorter wavelengths (e.g. X-band and/or C-band). The recent launch of the ALOS PALSAR-2 (L-band) sensor will ensure further woody structure modelling potential for

future studies. The robust C-band results, however, still bode well for future work involving the Sentinel-1 sensor (recently launched) where free C-band data will be made available.

## 5. ACKNOWLEDGEMENTS

The authors will like to acknowledge the Council for Scientific and Industrial Research, Department of Science and Technology, South Africa (grant agreement DST/CON 0119/2010, Earth Observation Application Development in Support of SAEOS) and the European Union's Seventh Framework Programme (FP7/2007-2013, grant agreement no. 282621, AGRICAB) for funding this study. The X-band StripMap TerraSAR-X scenes were acquired under the general proposals submission programme (Proposal no. LAN 1504; August 2012) of the Deutsches Zentrum für Luft- und Raumfahrt (DLR) German Aerospace Center. The C-band Quad-Pol RADARSAT-2 scenes were provided by MacDonald Dettwiler and Associates Ltd. – Geospatial Services Inc. (MDA GSI) through the Canadian Space Agency (CSA) Science and Operational Applications Research (SOAR) program. The L-band ALOS PALSAR FBD scenes were acquired under the K&C Phase 3 Proposal of the Japanese Aerospace Exploration Agency (JAXA). The Carnegie Airborne Observatory is made possible by the Avatar Alliance Foundation, Margaret A. Cargill Foundation, John D. and Catherine T. MacArthur Foundation, Grantham Foundation for the Protection of the Environment, W.M. Keck Foundation, Gordon and Betty Moore Foundation, Mary Anne Nyburg Baker and G. Leonard Baker Jr., and William R. Hearst III. The LiDAR data was processed by T. Kennedy-Bowdoin, D. Knapp, J. Jacobson and R. Emerson at the Carnegie Institution for Science. The authors would also like to acknowledge SANParks (Dr Izak Smit), Sabi Sands Game Reserve (Michael Grover), WITS Rural facility (Rhian Twine and Simon Khosa), SAEON (Patrick Ndlovu and Mightyman Mashele), CSIR EO colleagues and Bushbuckridge local authorities and personnel. Personal thanks also go to Mr Mikhail Urbazhev for providing support in GAMMA scripting and processing of the SAR imagery.

## 6. REFERENCES

- [1] J.F. Silva, A. Zambrano, R. Mario, "Increase in the woody component of seasonal savannas under different fire regimes in Calabozo, Venezuela", *Journal of Biogeography*, 28, pp. 977-983, 2001
- [2] M.S. Colgan, G.P. Asner, S.R. Levick et al., "Topo-edaphic controls over woody plant biomass in South African savannas", *Biogeosciences*, 9, pp. 1809-1821, 2012

- [3] D. Lu, "The potential and challenge of remote sensing-based biomass estimation", *International Journal of Remote Sensing*, 27 (7), pp. 1297-1328, 2006
- [4] S.C. Popescu, K. Zhao, A. Neuenschwander, C. Lin, "Satellite LiDAR versus small footprint airborne LiDAR: comparing the accuracy of aboveground biomass estimates and forest structure metrics at footprint level", *Remote Sensing of Environment*, 115, pp. 2786-2797, 2011
- [5] O.W. Tsui, N.C. Coops, M.A. Wulder et al., "Using multi-frequency radar and discrete-return LiDAR measurements to estimate above-ground biomass and biomass components in a coastal temperate forest. *ISPRS Journal of Photogrammetry and Remote Sensing*, 69, pp. 121-133 2012
- [6] E.T.A Mitchard, S.S. Saatchi, S.L. Lewis et al., "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 Sensing of Environment*, 115, pp. 2861-2873, 2011
- [7] GOCF-GOLD, "Reducing greenhouse gas emissions from deforestation and degradation in developing countries: a sourcebook of methods and procedures for monitoring, measuring and reporting", *GOCF-GOLD Report Version COP14-2*, pp. 1-185, 2009
- [8] S. García, A. Fernández, J. Luengo, F. Herrera, "Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power", *Information Sciences*, 180 (10), pp. 2044-2064, 2010
- [9] R. Mathieu, L. Naidoo, M.A. Cho et al., "Toward structural assessment of semi-arid African savannahs and woodlands: the potential of multitemporal polarimetric RADARSAT-2 fine beam images". *Remote Sensing of Environment*, 138, pp. 215-231, 2013
- [10] G.P. Asner, D.E. Knapp, J. Boardman et al., "Carnegie Airborne Observatory-2: Increasing science data dimensionality via high-fidelity multi-sensor fusion", *Remote Sensing of Environment*, 124, pp. 454-465, 2012
- [11] N. Sugiura, "Further analysis of the data by Akaike's information criterion and the finite corrections". *Communications in Statistics – Theory and Methods*, 7 (1), pp. 13-26, 1978
- [12] S.K. Shevade, S.S. Keerthi, C. Bhattacharyya, K.R.K. Murthy, "Improvements to the SMO Algorithm for SVM Regression". *IEEE Transactions on Neural Networks*, 11 (5), pp. 1188-1193, 1999
- [13] F. Esposito, D. Malerba, G. Semeraro, V. Tamma, "The effects of pruning methods on the predictive accuracy of induced decision trees", *Applied Stochastic Models in Business and Industry*, 15, pp. 277-299., 1999
- [14] O. Intrator, N. Intrator, "Interpreting neural-network results: a simulation study", *Computational Statistics & Data Analysis*, 37, pp. 373-393, 1993
- [15] L. Breiman, "Manual on setting up, using and understanding Random Forests v4.0", [http://oz.berkeley.edu/users/breiman/Using\\_random\\_forests\\_v4.0.pdf](http://oz.berkeley.edu/users/breiman/Using_random_forests_v4.0.pdf) (accessed 08.02.11), 2003

- [16] V.J. Tu, "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes", *Journal of Clinical Epidemiology*, 49 (11), pp. 1225-1231, 1996
- [17] D. Anguita, A. Ghio, N. Greco, L. Oneto, S. Ridella, "Model selection for support vector machines: advantages and disadvantages of the machine learning theory", *International Joint Conference on Neural Networks*, IEEE, pp. 1-8, 2010
- [18] M.E. Keskin, O. Terzi, E.U. Kucuksille, "Data mining process for integrated evaporation model", *Journal of Irrigation and Drainage Engineering*, 135 (1), pp. 39-43, 2009
- [19] A.M. Prasad, L.R. Iverson, A. Liaw, "Newer classification and regression tree techniques: bagging and random forests for ecological prediction", *Ecosystems*, 9 (2), pp. 181-199, 2006
- [20] R. Ismail, O. Mutanga, L. Kumar, "Modelling the potential distribution of pine forests susceptible to *Sirex Noctilo* infestations in Mpumalanga, South Africa", *Transactions in GIS*, 14 (5), pp. 709-726, 2010
- [21] J.M.B. Carreira, J.B. Melo, M.J. Vasconcelos, "Estimating the above-ground biomass in Miombo savannah woodlands (Mozambique, East Africa) using L-band synthetic aperture radar data", *Remote Sensing Open Access*, 5, pp. 1524-1548, 2013
- [22] C.M. Ryan, T. Hill, E. Woollen, C. Ghee, E. Mitchard, G. Cassells, J. Grace, I.H. Woodhouse, M. Williams, "Quantifying small-scale deforestation and forest degradation in African woodlands using radar imagery", *Global Change Biology*, pp.1-15, 2011
- [23] R.M. Lucas, N. Cronin, A. Lee, M. Moghaddam, C. Witte, P. Tickle, "Empirical relationships between AIRSAR backscatter and LiDAR-derived forest biomass, Queensland Australia", *Remote Sensing of Environment*, 100, pp.407-425, 2006

## Appendices 4A and 4B: Random Forest optimisation attempts

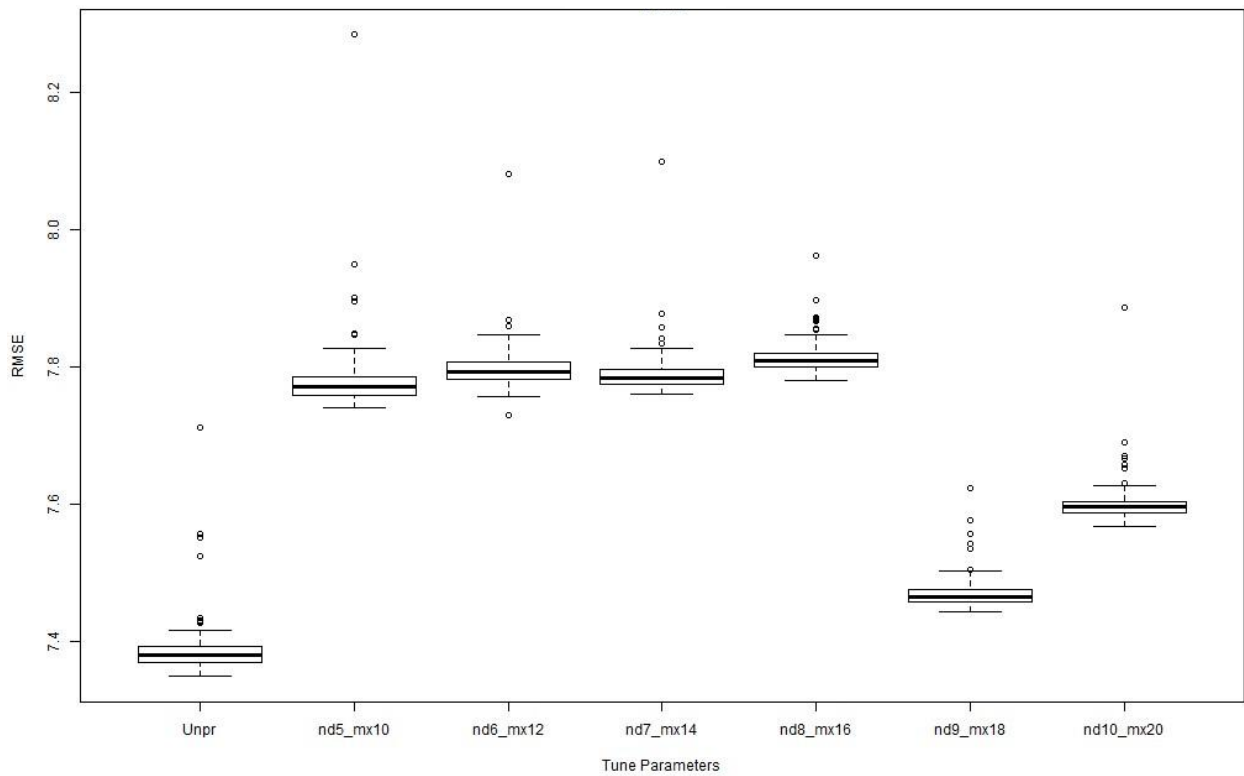
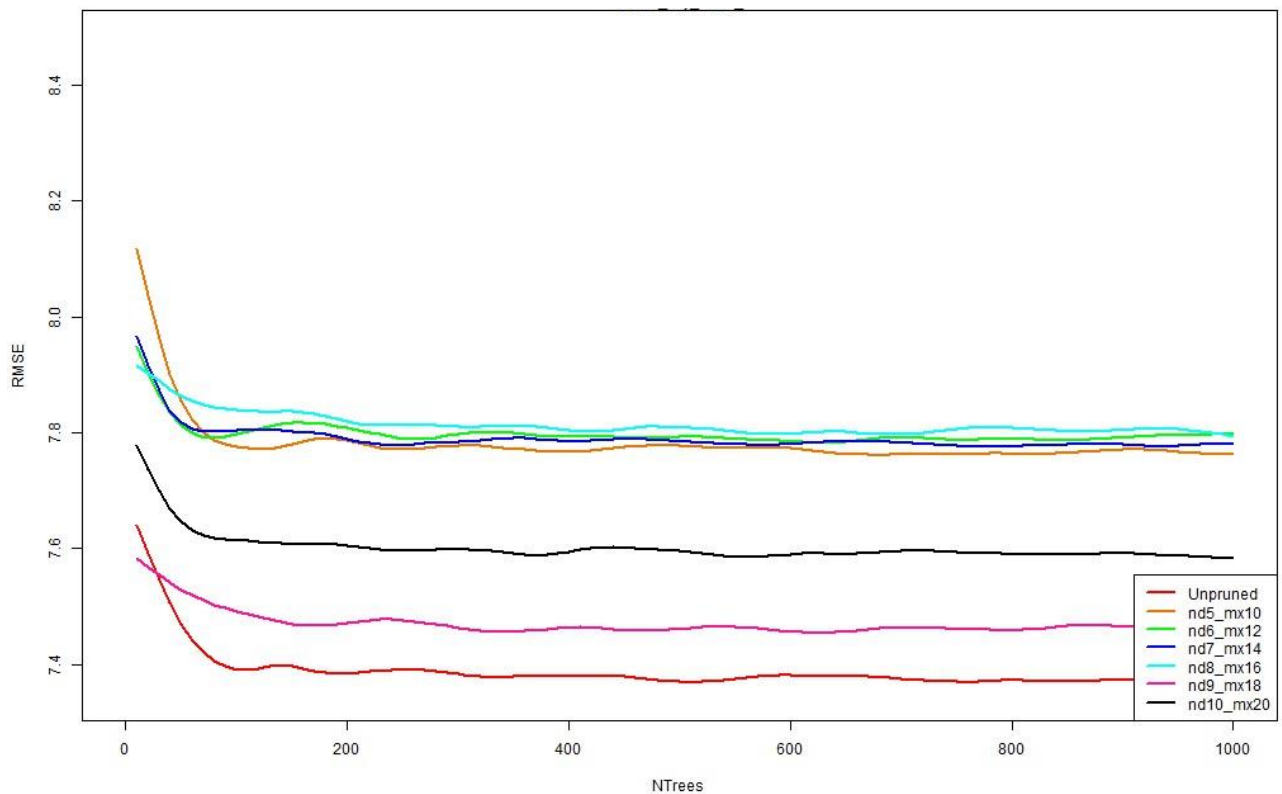


Figure 4A: Root Mean Square Error (RMSE) variability box plot, derived from modelled 2008 L-band SAR and LT5 summer reflectance data results, for analysing the different RF tuning parameters (Unpr = unpruned; nd = nodesize; mx = maxnodes; solid bar = mean RMSE; whiskers = max/min range; dots = outliers)



**Figure 4B: Root Mean Square Error (RMSE) line graph, derived from modelled 2008 L-band SAR and LT5 summer reflectance data results, for analysing the different RF tuning parameters across varying number of trees in the forest (NTrees = number of trees; nd = nodesize; mx = maxnodes)**

### Methodology for the creation of the EVI phenology graphs (Figures 4.4 and 4.5)

The EVI monthly values were extracted from all the 500m MODIS pixels which fell exactly within LiDAR coverages of specific grass ('L1' grass EVI values and phenology) and tree ('L8' tree EVI values and phenology) dominated landscapes between years 2005 and 2013. The multi-temporal EVI values were extracted from the WAMIS database (<http://wamis.meraka.org.za/>). Scene level EVI values were averaged for each date and plotted in figure 4.4 together with the monthly rainfall averages. The monthly average rainfall data was extracted from Graskop, Skukuza and Phalaborwa weather station, in Mpumalanga, and provided by the South African Weather Services. Figure 4.5 was created by subtracting the tree EVI values from the grass EVI values to get the difference EVI values which were plotted over the same time interval of figure 4.4. The approximate acquisition dates of the multi-temporal Landsat-5 imagery were also added to figure 4.5.