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[a], Renaud Mathieu[a], Russell Main[a], Waldo Kleynhans[b], Konrad Wessels[b], Gregory P. Asner[c], Brigitte Leblon[d]

[a]Ecosystem Earth Observation, Natural Resources and the Environment, CSIR, Pretoria, South Africa
Corresponding author contact details: LNaidoo@csir.co.za; (+27)12 841 2233
[b]Remote Sensing Unit, Meraka Institute, CSIR, Pretoria, South Africa
[c]Department of Global Ecology, Carnegie Institution for Science, Stanford, CA, USA
[d]Faculty of Forestry and Environmental Management, University of New Brunswick, Fredericton, Canada

ABSTRACT

The woody component in African Savannahs 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 Savannahs, 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 GOFC-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 σ0 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.

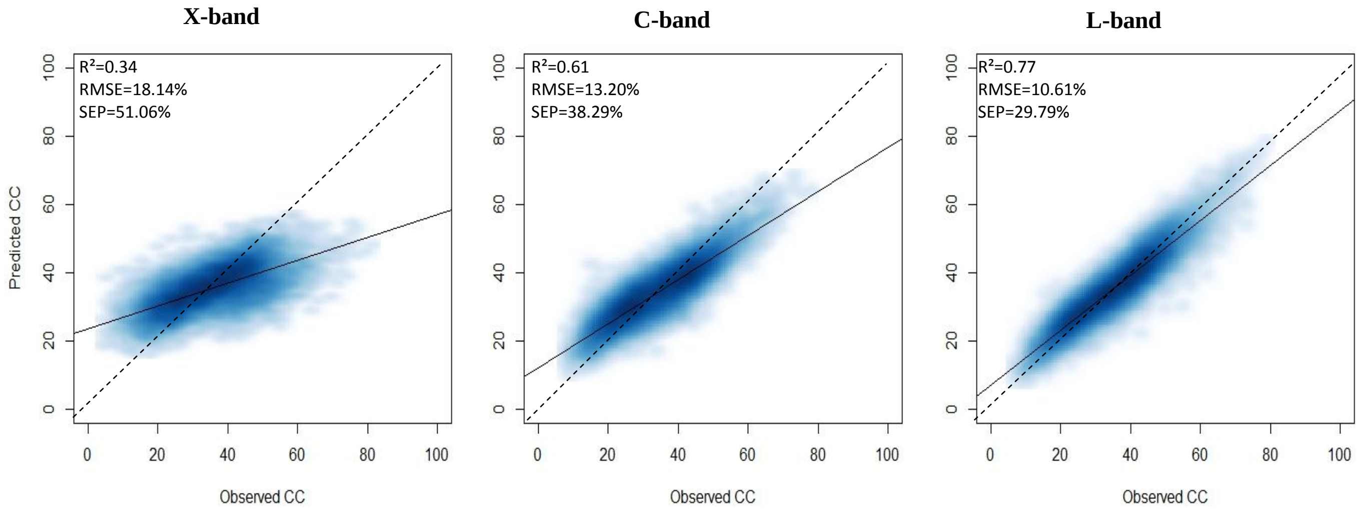

Figure 1: Mean RF predicted CC versus mean observed CC for each multi-frequency scenario (The dotted line refers to the 1:1 line)

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

Band	**X** *[N = 13761]*			**C** *[N = 11687]*		
Algorithm	**R² (CI)**	**RMSE (CI)**	**SEP (CI)**	**R² (CI)**	**RMSE (CI)**	**SEP (CI)**
LR	0.30 (0.002)	18.57 (0.023)	52.18 (0.084)	0.55 (0.002)	14.04 (0.034)	40.88 (0.123)
SVM	0.30 (0.002)	18.72 (0.036)	52.68 (0.112)	0.55 (0.002)	14.48 (0.099)	42.09 (0.280)
REPTree	0.36 (0.005)	17.74 (0.089)	49.86 (0.282)	0.63 (0.002)	12.91 (0.032)	37.53 (0.127)
ANN	0.39 (0.009)	17.29 (0.152)	48.52 (0.394)	0.65 (0.002)	12.56 (0.033)	36.50 (0.090)
RF	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	**L** *[N = 13954]*					
Algorithm	**R² (CI)**	**RMSE (CI)**	**SEP (CI)**			
LR	0.71 (0.002)	11.88 (0.050)	33.36 (0.154)			
SVM	0.71 (0.003)	12.34 (0.083)	34.65 (0.246)			
REPTree	0.78 (0.002)	10.40 (0.045)	29.16 (0.145)			
ANN	0.79 (0.003)	10.15 (0.066)	28.49 (0.178)			
RF	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 fur 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 Urbazaev 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] GOFC-GOLD, “Reducing greenhouse gas emissions from deforestation and degradation in developing countries: a sourcebook of methods and procedures for monitoring, measuring and reporting”, *GOFC-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 (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