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\textsuperscript{a}, Renaud Mathieu\textsuperscript{b}, Russell Main\textsuperscript{a}, Waldo Kleynhans\textsuperscript{b}, Konrad Wessels\textsuperscript{b}, Gregory P. Asner\textsuperscript{c}, Brigitte Leblon\textsuperscript{d}

\textsuperscript{a}Ecosystem Earth Observation, Natural Resources and the Environment, CSIR, Pretoria, South Africa
\textsuperscript{b}Remote Sensing Unit, Meraka Institute, CSIR, Pretoria, South Africa
\textsuperscript{c}Department of Global Ecology, Carnegie Institution for Science, Stanford, CA, USA
\textsuperscript{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. This 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 need 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?
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 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.

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

<table>
<thead>
<tr>
<th>Band</th>
<th>X [N = 13761]</th>
<th>C [N = 1687]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm</td>
<td>R² (CI)</td>
<td>RMSE (CI)</td>
</tr>
<tr>
<td>LR</td>
<td>0.30 (0.002)</td>
<td>18.57 (0.023)</td>
</tr>
<tr>
<td>SVM</td>
<td>0.30 (0.002)</td>
<td>18.72 (0.036)</td>
</tr>
<tr>
<td>REPTree</td>
<td>0.36 (0.005)</td>
<td>17.74 (0.089)</td>
</tr>
<tr>
<td>ANN</td>
<td>0.39 (0.009)</td>
<td>17.29 (0.152)</td>
</tr>
<tr>
<td>RF</td>
<td>0.34 (0.003)</td>
<td>18.14 (0.040)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Band</th>
<th>L [N = 13954]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm</td>
<td>R² (CI)</td>
</tr>
<tr>
<td>LR</td>
<td>0.71 (0.002)</td>
</tr>
<tr>
<td>SVM</td>
<td>0.71 (0.003)</td>
</tr>
<tr>
<td>REPTree</td>
<td>0.78 (0.002)</td>
</tr>
<tr>
<td>ANN</td>
<td>0.79 (0.003)</td>
</tr>
<tr>
<td>RF</td>
<td>0.77 (0.001)</td>
</tr>
</tbody>
</table>
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


