PARAMETRIC AND NON-PARAMETRIC FOREST BIOMASS ESTIMATION FROM POLINSAR DATA

Maxim Neumann, Sassan S. Saatchi, Lars M. H. Ulander[†] *and Johan E. S. Fransson*[*]

Jet Propulsion Laboratory (JPL), California Institute of Technology (CALTECH),
Radar Science and Engineering Section, 4800 Oak Grove Dr, Pasadena, CA 91109
E-mail: maxim.neumann@jpl.nasa.gov

[†] Chalmers University of Technology, SE-412 96 Göteborg, Sweden

[*] Swedish University of Agricultural Sciences (SLU), SE-901 83 Umeå, Sweden

ABSTRACT

Biomass estimation performance from model-based polarimetric SAR interferometry (PolInSAR) using generic parametric and non-parametric regression methods is evaluated at L- and P-band frequencies over boreal forest. PolInSAR data is decomposed into ground and volume contributions, estimating vertical forest structure, and using a set of obtained parameters for biomass regression. The considered estimation methods include multiple linear regression, support vector machine and random forest. The biomass estimation performance is evaluated on DLR's airborne SAR data at L- and P-bands over Krycklan Catchment, a boreal forest test site in Northern Sweden. The combination of polarimetric indicators and estimated structure information has improved the root mean square error (RMSE) of biomass estimation up to 28% at L-band and up to 46% at P-band. The cross-validated biomass RMSE was reduced to 20 tons/ha.

***Index Terms*—** Forest biomass, polarimetric SAR interferometry, regression, support vector machine, random forest

1. INTRODUCTION

The present estimate by IPCC (Intergovernmental Panel on Climate Change) is that deforestation amounts to between 10% and 30% of the total anthropogenic carbon dioxide (CO_2) flux. The range of uncertainty is large due to the lack of accurate observational techniques. Therefore, several space-borne remote sensing missions are proposed to address the need for global monitoring of forest carbon stocks and changes, such as NASA's L-band mission DESDynI and ESA's P-band mission BIOMASS.

In radar remote sensing, above ground biomass (AGB) has been estimated from SAR backscatter, polarimetry, and InSAR coherence, by means of empirical and model-based approaches. Polarimetric SAR interferometry (PolInSAR) has demonstrated the possibility to estimate forest height, which can be related to AGB. Furthermore, it is possible to separate the ground and volume contributions and to estimate vertical structure components, compensating for temporal and thermal decorrelation.

In this paper, we systematically investigate the improvement of AGB estimation, using parameter sets from model-based PolInSAR inversion of vertical structure and ground/volume polarimetric signatures. The analyzed estimation methods include multiple linear regression (LR), support vector machine (SVM) regression and random forest (RF) regression. The last two approaches are non-parametric in the sense that no parametric model or distribution is assumed. Both SVM and RF are two promising machine learning techniques, which have been applied successfully to several real-world classification and regression problems. The AGB estimation performance is evaluated using the leave-one-out cross-validation and the distributions of predicted and mapped biomass.

2. METHODOLOGY

Given n interferometric tracks with p used polarimetric channels, second-order statistics of multi-baseline polarimetric interferometric SAR (MB-PolInSAR) data can be represented in the covariance matrix $\mathbf{C}_{MB} \in \mathbb{C}^{pq \times pq}$. The MB-PolInSAR Random Volume over Ground (RVoG) model provides the means to decompose the covariance matrix $\mathbf{C}_{MB}$ into the ground ($_g$) and volume ($_v$) contributions [1] :

$$\mathbf{C}_{MB} = \mathbf{C}_{MB/g} + \mathbf{C}_{MB/v} \iff \begin{cases} \mathbf{C} = \mathbf{C}_g + \mathbf{C}_v \\ \mathbf{\Omega}_{ij} = \gamma_{ij/g}\mathbf{C}_g + \gamma_{ij/v}\mathbf{C}_v \end{cases} \tag{1}$$

where $\mathbf{C}, \mathbf{C}_g, \mathbf{C}_v \in \mathbb{C}^{p \times p}$ are polarimetric covariance matrices for total backscattering and the ground and volume contributions, respectively, $\gamma_{ij/\{g,v\}} \in \mathbb{C}$ are the ground/volume coherences of the interferometric pairs i and j.

Polarimetric covariance matrices provide the means to compute some parameters which characterize the layers, in-

This research was supported by an appointment to the NASA Postdoctoral Program at the Jet Propulsion Laboratory, administered by Oak Ridge Associated Universities through a contract with the National Aeronautics and Space Administration.

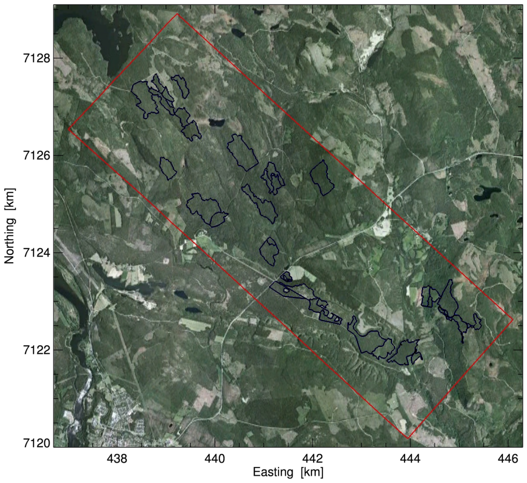


Fig. 1. Optical image of the Krycklan Catchment area, including the swath of radar data (red) and the location of the forest stands (black).

cluding the Pauli-basis intensities (*HH+VV, HH-VV, HV*) and the scattering-mechanism-type indicator angle α. From the PolInSAR coherences, it is possible to retrieve information about the forest height h_v and the ground-volume-ratio μ.

After the ground-volume separation and parameter retrieval, AGB is regressed through LR, SVM and RF. LR is solved by least squares. SVM and RF are based on machine learning classification approaches, adapted to regression problems. SVM is a framework for non-parametric and non-linear classification and regression. The basic idea consists in transforming the input data into a higher-dimensional feature space, where the problem can be addressed in linearized manner. In the end, training a regression SVM involves solving a quadratic optimization problem. In this work, the Gaussian radial basis function (RBF) kernel is used for the transformation. RF is an ensemble learning technique, where many decision trees are constructed based on random sub-sampling of the given data set. In addition, each node of every tree is split based on another random subset of parameters. This two-layer randomization provides a certain level of robustness to outliers and overfitting. The regression result is usually aggregated by taking the average of the predictions from all trees.

3. EXPERIMENTAL RESULTS AND DISCUSSION

As part of the studies for the proposed ESA P-band mission BIOMASS, the BioSAR 2008 campaign in Northern Sweden was carried out in October of 2008 to evaluate the possibilities for biomass estimation in boreal forests. The test site is located in the Krycklan Catchment and consists of mainly managed forest. Within the 30 km^2 test site 27 forest stands (Fig. 1), which are fully covered by radar data, are used for biomass estimation.

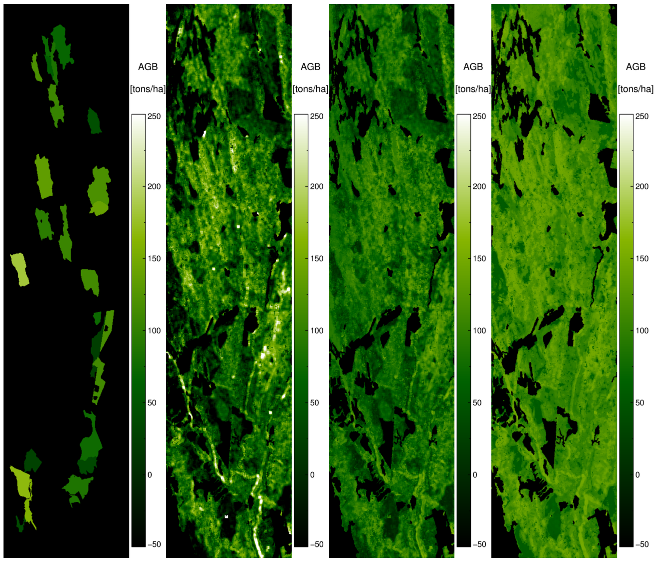
The radar data are acquired by DLR's E-SAR sensor at L- and P-bands. For the experiments presented in this paper, data from two data sets with opposite headings are combined, in order to double the number of forest stand samples to 54. The resolution is about 1.5 m × 0.92 m in slant range and azimuth directions for L-band, and 1.5 m × 1.47 m for P-band. The radar images were multi-looked to achieve pixels with about 50×50 m^2 resolution, on which parameter inversion and AGB regression is performed.

In order to evaluate regression performance, several sets of retrieved parameter sets ζ are used. The data are grouped based on increasing number of polarimetric parameters (from ζ={*HV*} to ζ={*HH+VV, HH-VV, HV*, α}) on one hand and the increasing complexity of PolInSAR derived parameters and covariance matrices on the other hand (Tables. 1-3). As shown in the first row of Table 1, the polarimetric parameters are first estimated only for the total covariance matrix. In the second series of tests the regression parameter set is extended by inverted structure parameters, h_v and μ. In the final two test series, the polarimetric parameters from the ground and volume covariance matrices are added as well. Given this configuration, the number of parameters used for AGB regression is between 1 and 14.

Table 1 presents the root mean square error (RMSE) of biomass estimation for multiple linear regression at L- and P-bands. In every cell there are two values: the first represents the RMSE using the entire data set, the second value represents the cross-validated RMSE using leave-one-out approach. In the leave-one-out case, for every single sample, the regression function is estimated using all other samples except for the one sample, which is then used for testing. This provides reasonable accuracies assessments of the overall performance of the forest biomass estimation.

Using only a single backscatter value (HV), the cross-validated AGB RMSE of L and P frequencies is 34.9 and 39.7 tons/ha, respectively. If using more polarimetric information from $\mathbf{C}$, the RMSE value stays higher than 27 tons/ha at both frequencies. With the addition of height and ground-volume ratio information one can observe a significant improvement in AGB estimation. At L-band the RMSE improves up to 20.9 tons/ha, and at P-band up to 25.0 tons/ha. With increasing number of parameters, one can observe the tendency to overfit the data, which is responsible for the difference between the RMSEs obtained with and without cross-validation. Including in addition the polarimetric parameters from ground and volume covariance matrices $\mathbf{C}_g$ and $\mathbf{C}_v$ improves the RMSE to 19.9 tons/ha and 20.8 tons/ha, respectively.

Overall, using multiple linear regression on the given data set, the inclusion of structure information in form of height and ground-volume characteristics improved the RMSE of AGB estimation by 20% to 25% at L-band, and 9% to 32% at P-band. Using parameters from the estimated ground and volume covariance matrices improved the RMSE up to 28% and 46% at L- and P-bands, respectively. The improvement of RMSE by using more polarimetric parameters (instead of only HV) was up to 22% at L-band, and 31% at P-band. But these results may be slightly different if other combination of



(a) AGB(*in situ*) (b) AGB(LR) (c) AGB(SVM) (d) AGB(RF)

Fig. 2. Maps of (a) *in situ* AGB at forest stand level, and predicted AGB using (b) Linear Regression (LR), (c) SVM, (d) RF.

parameters are used in the analysis.

Evaluating non-parametric approaches, one can observe that they perform better when applied to the training data than LR because of the potential to fit the data better to the model and the errors or variations, that may exist in dependent and independent variables. However, when tested over an independent data set, the performance may be worse than the regression model.

RMSE of training data is improved significantly, up to 9.5 tons/ha. However, note the difference between training RMSE and testing RMSE which indicates overfitting of the data in some instances, reaching up to 24 tons/ha difference.

Using many more samples and reducing the number of parameters might compensate for the overfitting. However, usually the number of *in situ* AGB stands in a local environment is limited, requiring robust and noise-tolerant methods.

Based on the regression of AGB using the different methods, biomass values for the entire study area can be predicted. As an example, Fig. 2 shows the predicted AGB maps for the ascending L-band data set using the regression parameter combination $\zeta(\mathbf{C}), \zeta(\mathbf{C}_g), h_v, \mu$ with $\zeta = \{$*HH+VV, HH-VV, HV,* $\alpha\}$ (10 parameters). Visually, the LR predicted AGB map provides more contrast, while the RF AGB map is more homogeneous over the forested areas.

As can be seen in the histograms of the predicted AGB values (Fig. 3), the distributions of AGB for the different regression methods are quite different, despite using same input data sets, and having comparably similar cross-validation RMSE with biomass. In particular, linear regression often predicts negative biomass values, due to linear extrapolation of very low backscatter and height values. This shows that a single LR relation is insufficient to characterize biomass over the entire range, and suggests to use several LR parameterizations for different biomass ranges.

SVM and RF have more limited ranges of predicted AGB. Though some of SVM biomass values are still negative, it is only a very small portion. However, only linear regression distribution follows approximately the Gaussian distribution. The other two approaches provide more artificial distributions. In particular the RF histogram is characterized by concentration of predicted AGB at about a few values. These values probably correspond to the output of a few most likely RF decision trees and nodes, which introduces bias in the result.

4. CONCLUSION

In this paper, we have evaluated the performance of AGB estimation from model-based PolInSAR data at L- and P-bands. The estimated vertical structure, encompassing forest height and ground-volume ratio, and ground and volume polarimetric scattering characteristics, have been shown to enhance biomass estimation. Different regression approaches have been evaluated, including multiple linear regression, support vector machine and random forest regression.

Overall, the combination of polarimetry with interferometry provides the potential for significant improvement of biomass estimation. While in the end the performances at L- and P-bands were similar, the internal relationships are different. Using forest height and ground and volume characteristics in a multiple linear regression improved AGB estimation as expected, decreasing biomass RMSE by 26% to 28% at L-band and 19% to 46% at P-band. LR provided best cross-validated results for AGB estimation, but in mapping, it often predicted negative AGB. SVM and RF provided more reasonable AGB predictions, but their cross-validation performance was slightly worse. The analysis of mapped AGB distributions suggests that using a more adaptive parametric estimation technique, e.g. physical model based, could still improve biomass prediction.

5. REFERENCES

[1] M. Neumann, L. Ferro-Famil, and A. Reigber, “Estimation of Forest Structure, Ground and Canopy Layer Characteristics from Multi–Baseline Polarimetric Interferometric SAR Data,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 3, pp. 1086–1104, Mar. 2010.

[2] M. Neumann, S. S. Saatchi, L. M. H. Ulander, and J. E. S. Fransson, “Boreal Forest Biomass Regression from Model-Based Polarimetric Interferometric SAR Data at L- and P-Bands,” *IEEE Transactions on Geoscience and Remote Sensing*, 2011, submitted.

[3] G. Sandberg, L. M. H. Ulander, J. E. S. Fransson, J. Holmgren, and T. Le Toan, “L- and P-Band Backscatter Intensity for Biomass Retrieval in Hemiboreal Forest,” *Remote Sens. Env.*, 2011, in press.

Table 1. Root mean square error (RMSE) in tons/ha for estimation of AGB using LR regression without/with cross-validation.

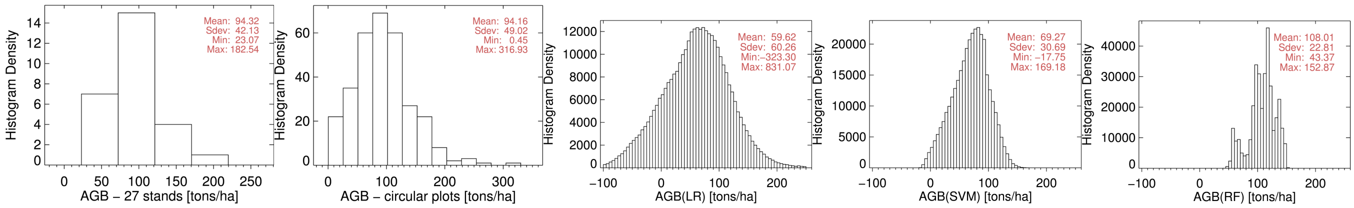
L-Band — ζ	$\zeta(\mathbf{C})$	$\zeta(\mathbf{C})$, h_v, μ	$\zeta(\mathbf{C})$,$\zeta(\mathbf{C}_g)$,h_v,μ	$\zeta(\mathbf{C})$,$\zeta(\mathbf{C}_g)$,$\zeta(\mathbf{C}_v)$,h_v,μ
HV	34.1/34.9	24.5/26.1	23.7/26.0	23.1/25.6
HH-VV, HV	31.2/32.2	23.8/25.8	21.1/24.5	19.4/24.0
HH+VV, HH-VV, HV	25.7/27.2	18.7/20.9	17.5/21.2	15.7/19.9
HH+VV, HH-VV, HV, α	25.7/27.5	18.4/21.1	17.4/21.9	15.1/19.9
P-Band — ζ	$\zeta(\mathbf{C})$	$\zeta(\mathbf{C})$, h_v, μ	$\zeta(\mathbf{C})$,$\zeta(\mathbf{C}_g)$,h_v,μ	$\zeta(\mathbf{C})$,$\zeta(\mathbf{C}_g)$,$\zeta(\mathbf{C}_v)$,h_v,μ
HV	38.53/39.7	25.40/27.2	22.55/24.0	21.88/24.2
HH-VV, HV	36.78/38.6	25.39/27.5	22.06/24.7	18.14/20.8
HH+VV, HH-VV, HV	25.52/27.3	22.66/25.0	20.72/26.4	17.96/22.2
HH+VV, HH-VV, HV, α	25.17/28.1	22.27/25.7	19.02/23.9	17.24/22.9

Table 2. Root mean square error (RMSE) in tons/ha for estimation of AGB using SVM regression without/with cross-validation.

L-Band — ζ	$\zeta(\mathbf{C})$	$\zeta(\mathbf{C})$, h_v, μ	$\zeta(\mathbf{C})$,$\zeta(\mathbf{C}_g)$,h_v,μ	$\zeta(\mathbf{C})$,$\zeta(\mathbf{C}_g)$,$\zeta(\mathbf{C}_v)$,h_v,μ
HV	33.9/35.5	22.8/27.6	23.1/27.4	22.7/27.1
HH-VV, HV	31.6/33.7	22.0/26.0	21.5/25.9	20.6/26.1
HH+VV, HH-VV, HV	30.2/30.1	21.1/25.2	19.2/23.6	18.5/23.8
HH+VV, HH-VV, HV, α	26.0/30.6	19.3/23.0	18.5/24.5	17.3/24.3
P-Band — ζ	$\zeta(\mathbf{C})$	$\zeta(\mathbf{C})$, h_v, μ	$\zeta(\mathbf{C})$,$\zeta(\mathbf{C}_g)$,h_v,μ	$\zeta(\mathbf{C})$,$\zeta(\mathbf{C}_g)$,$\zeta(\mathbf{C}_v)$,h_v,μ
HV	37.9/40.4	24.0/26.3	22.4/27.2	21.6/27.1
HH-VV, HV	36.9/41.1	23.8/28.3	22.2/27.7	20.8/27.7
HH+VV, HH-VV, HV	30.6/32.2	22.7/29.3	20.4/28.4	19.0/26.9
HH+VV, HH-VV, HV, α	25.7/31.3	21.0/29.2	18.9/28.1	16.6/26.8

Table 3. Root mean square error (RMSE) in tons/ha for estimation of AGB using RF regression without/with cross-validation.

L-Band — ζ	$\zeta(\mathbf{C})$	$\zeta(\mathbf{C})$, h_v, μ	$\zeta(\mathbf{C})$,$\zeta(\mathbf{C}_g)$,h_v,μ	$\zeta(\mathbf{C})$,$\zeta(\mathbf{C}_g)$,$\zeta(\mathbf{C}_v)$,h_v,μ
HV	18.1/34.7	10.9/25.7	10.9/25.6	11.1/25.9
HH-VV, HV	17.1/36.3	10.7/24.9	10.3/24.8	10.2/24.7
HH+VV, HH-VV, HV	15.7/35.2	10.7/25.3	10.2/24.6	10.7/25.6
HH+VV, HH-VV, HV, α	13.9/31.9	10.2/24.3	10.5/24.6	10.8/25.6
P-Band — ζ	$\zeta(\mathbf{C})$	$\zeta(\mathbf{C})$, h_v, μ	$\zeta(\mathbf{C})$,$\zeta(\mathbf{C}_g)$,h_v,μ	$\zeta(\mathbf{C})$,$\zeta(\mathbf{C}_g)$,$\zeta(\mathbf{C}_v)$,h_v,μ
HV	22.2/43.6	12.4/28.1	11.1/25.6	11.2/26.1
HH-VV, HV	20.0/44.0	12.0/28.2	11.2/26.5	11.6/27.0
HH+VV, HH-VV, HV	14.1/33.3	11.5/27.7	11.4/26.9	11.3/26.1
HH+VV, HH-VV, HV, α	10.7/24.3	10.1/24.4	10.2/23.4	9.5/22.7



(a) forest stands (b) circular plots (c) LR (d) SVM (e) RF

Fig. 3. Histograms of *in situ* AGB on (a) forest stand level (resolution between 2-26 ha), (b) circular plots (resolution 0.03 ha), and predicted AGB for (c) LR, (d) SVM and (e) RF (resolution about 0.25 ha).