Implementation of a Sentinel-2 Based Exploratory Workflow for the Estimation of Above Ground Biomass

Giulia Molisse

Institute of Earth Observation, Eurac Research, Bolzano, Italy

Dzhaner Emin

IABG, Dresden, Germany

Hugo Costa

Direção-Geral do Território, Lisbon, Portugal

NOVA Information Management School, Universidade Nova de Lisboa, Lisbon, Portugal

<u>This is the final, accepted version of the conference paper published by *IEEE*</u>

Molisse, G., Emin, D., & Costa, H. (2022). Implementation of a Sentinel-2 Based Exploratory Workflow for the Estimation of Above Ground Biomass. In 2022 IEEE Mediterranean and Middle-East Geoscience and Remote Sensing Symposium, M2GARSS 2022 - Proceedings (pp. 74-77). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/M2GARSS52314.2022.9839897

IMPLEMENTATION OF A SENTINEL-2 BASED EXPLORATORY WORKFLOW FOR THE ESTIMATION OF ABOVE GROUND BIOMASS

1st Giulia Molisse
Institute of Earth Observation, Eurac Research
Bolzano, Italy
gmolisse@eurac.edu

2nd Dzhaner Emin
IABG
Dresden, Germany
emin@iabg.de

3rd Hugo Costa
1. Direção-Geral do Território
2. NOVA Information Management School
Universidade Nova de Lisboa, Lisbon, Portugal
hcosta@dgterritorio.pt

***Abstract*—This work presents a Sentinel-2 based exploratory workflow for the estimation of Above Ground Biomass (AGB) in a Mediterranean forest. Up-to-date and reliable mapping of AGB has been increasingly required by international commitments under the climate convention, and in the last decades, remote sensing-based studies on the topic have been widely investigated. After the generation of several vegetation and topographic features, the proposed approach consists of 4 major steps: 1) Feature selection 2) AGB prediction with k-Nearest Neighbour (kNN), Random Forest (RF), Extreme Gradient Boosting (XGB), and Artificial Neural Networks (ANN); 3) hyper-parameters fine-tuning with Bayesian Optimization; and finally, 4) model explanation with the SHapley Additive exPlanations (SHAP) package. The following results were obtained: 1) before hyper-parameters optimization, the Deep Neural Network (DNN) yielded the best performance with a Root Mean Squared Error (RMSE) of 42.30 t/ha; 2) after hyper-parameters fine-tuning with Bayesian Optimization, the Extreme Gradient Boosting (XGB) model yielded the best performance with a RMSE of 37.79 t/ha; 3) model explanation with SHAP allowed for a deeper understanding of the features impact on the model predictions. Finally, the predicted AGB throughout the study area showed an average value of 83 t/ha, ranging from 0 t/ha to 346.56 t/ha.**

***Index Terms*—Above Ground Biomass, Sentinel-2, Extreme Gradient Boosting, SHAP**

I. INTRODUCTION

International commitments under the climate convention [1] as well as sustainable forest management practices require accurate and up-to-date mapping of forested areas. Forests are the largest terrestrial carbon pool, counting about 85% of the total land vegetation biomass, and in them it is found 73% of the global soil carbon [2]. Additionally, global warming reports require detailed information on the forests' carbon content [3]. Hence, reliable and up-to-date Above Ground Biomass (AGB) mapping has become worldwide essential.

This work aims to asses the capabilities of Sentinel-2 derived measures for the estimation AGB in a Mediterranean forest. More specifically, the authors want to identify an optimal Season for satellite data acquisition, a ranking of the most influential satellite-generated features, as well as building an optimal regression model.

II. DATA AND METHODS

A. Study area

The description of the study area was provided by a study from Torralba et al. (2018) [4]. The area covers a total of 3741.5 ha and is located in the Natural Park of Serra de Espadan, in the eastern Spain province of Castellon. This natural park is a Mediterranean forest with soft and rounded hills, abandoned farming with artificial terraces, and mountain peaks up to 1100 meters of altitude. The area displays a heterogeneous landscape dominated by pure and mixed native coniferous and deciduous forests, with species of the genera *Pinus* and *Quercus*.

B. Field data and field-based Above Ground Biomass

A field inventory with measured AGB at the plot level was provided by the Geo-Environmental Cartography and Remote Sensing Group (CGAT) of the Universitat Politècnica de València (UPV) [1]; the collection of this forestry inventory was funded by the Spanish Ministerio de Economía y Competitividad, in the framework of the project CGL2016-80705-R. The field data were collected in September 2015 for a total of 73 circular plots with a radius of 15 m distributed throughout the study area. For each species or forest type within a plot, AGB was estimated in $t/plots$ using species-specific and forest type-specific allometric equations from Monteiro et al. (2005) [5]. The field-based AGB was then converted from $t/plots$ into t/ha. The field-based AGB ranges from a minimum of 0.35 *t/ha* to a maximum of 274.50 *t/ha*, with a mean value of 92.49 *t/ha*;

C. Sentinel-2 and Digital Elevation Model collection

The study area is covered by one single Sentinel-2A Level-1C tile. One Summer (August 2015) and one Autumn (November 2016) image were downloaded from the Copernicus Open Access Hub [2]. The images were converted from Top of the Atmosphere (ToA, L1C) to Bottom of the Atmosphere (BoA, L2A) reflectance, using Sen2Cor. Further preprocessing was

[1] http://cgat.webs.upv.es/

[2] https://scihub.copernicus.eu/

carried out using the Sentinel Application Platform (SNAP); red-edge and SWIR bands were re-sampled from 20 to 10 m using the nearest neighbor method; the 3 bands with a spatial resolution of 60 m (band 1, 9 and 10) were excluded from the analysis.

The European Digital Elevation Model (DEM) and derived slope were downloaded in the section *Imagery and Reference Data* of the Copernicus website[3]. These products have a spatial resolution of 25m. DEM and slope were re-sampled to the same spatial resolution as the Sentinel-2 images (10m) by making sure that cell size and cell positioning matched.

D. Methodology

1) Features generation: Vegetation Indices (VI) and biophysical parameters have been proved to increase the performance of regression algorithms for AGB estimation in different ecological zones and forest types [6, 7, 8]. Particular care was taken in including VIs which required SWIR and red-edge bands. Hence, VIs were calculated for both dates, that is August 2015 and November 2016. A total of 10 VIs were generated: Normalized Difference Vegetation Index (NDVI), Green Normalized Difference Vegetation Index (GNDVI), Soil Adjusted Vegetation Index (SAVI), Modified Soil Adjusted Vegetation Index (MSAVI), Global Environmental Monitoring Index (GEMI), Normalized Difference Vegetation Index red-edge 1, 2 and 3 (NDVIre1, NDVIre2 and NDVIre3) , Chlorophyll red-edge index (Clre), and Normalized Difference Water Index (NDWI).

Additionally, 5 biophysical parameters - Leaf Area Index (LAI), Canopy Water Content (LAI cwc), Canopy Chlorophyll Content (LAI cab), Fraction of absorbed photo-synthetically active radiation (FAPAR) and Fraction of vegetation cover (FCOVER) - were calculated for each image by using the *biophysical processor* in SNAP. Such variables have been found to enhance the estimation of biomass by describing spatial distribution and dynamics of vegetation [7].

Further spatial predictors were included. This allows to consider not only the spectral response of different surfaces, but also the spatial relationships among these surfaces. Therefore, texture measures derived from the Gray Level Co-occurrence Matrix (GLCM) were included, as they have been widely used for enhancing remote sensing-based classification and regression forestry-related problems [9, 10]. Therefore, contrast, entropy, and GLCM-mean were derived from both Sentinel-2 band-2 (blue) and NDVI generated for August 2015 and November 2016.

2) Features selection: In order to build a reliable model, careful evaluation is required when deciding which and how many features to include in the said model. To the best of our knowledge, step-wise regression and Random Forest (RF) measures are amongst the most frequently used approaches for feature selection in the field of AGB [11, 12, 13, 7]. Despite step-wise regression is still widely used for this purpose, it was decided not to include it. Such a decision is legitimized by the work of several authors proving its unsuitability for feature selection [14, 15].

Feature ranking is performed using the RF algorithm, which provides 3 measures of impurity: Gini index, entropy, and variance. Variance, or residual sum of squares, is used to measure node impurity in regression problems, and it represents the total reduction of the variance of the target variable due to the split of a certain feature at the node [16]. Impurity measures can be used for feature selection by evaluating the extent to which each feature contributes to decreasing the averaged impurity in each tree composing the forest [16], so as to calculate the MDI for each feature. Those features able to account for more variance decrease are going to be at the top of the ranking [16]. Therefore, the MDI can be seen as the total decrease in node impurity from splitting on the variable, averaged over all trees [17].

3) Above Ground Biomass prediction: The ranking yielded by the MDI method was tested and evaluated using several non-parametric Machine Learning (ML) algorithms: K-Nearest Neighbour (kNN), Random Forest (RF), Extreme Gradient Boosting (XGB) and, lastly, 3 Artificial Neural Networks (ANN). Hence, each model was cross-validated by using cross-validation and a scaling function.

kNN is a ML algorithm and it does not assume normal distribution or linear relationships; furthermore, it can be used for regression and classification problems. This non-parametric algorithm has been widely tested for the prediction of AGB [18, 19]. The default value of 5 was used for ”k”.

RF is an ensemble ML algorithm constituted of several decision trees [20], and it is widely used for classification, regression, and other tasks. Default hyper-parameters values were used for the testing of the RF, with 100 estimators.

The Extreme Gradient Boosting (XGB) is a popular variants of the Gradient Boosting algorithm, and it has been the winner of many ML competitions [21]. In order to test the feature selection ranking, an Extreme Gradient Boosting algorithms was trained using default hyper-parameters, with 300 estimators, and a learning rate of 0.3.

Finally, 3 ANN were generated. These had the following architecture: a Linear Neural Network (LNN) - with no hidden layers -, a Shallow Neural Network (SNN) - with 3 hidden layers of 8,4 and 1 neurons -, and Deep Neural Network (DNN) - with 5 hidden layers of 128, 64, 32,16, 8 and 1 neurons.

4) Hyper-parameters fine-tuning: Bayesian optimization was implemented by using a python package by Nogueira (2004) [22]; who defines it as a constrained global optimization tool which needs a performance metric to be maximized in as few iterations as possible, through Bayesian inference and Gaussian process. For this work, the negative RMSE (-RMSE) was set as the value to be maximised. Subsequently, the hyper-parameters to be tested for each algorithm and their respective ranges were identified. Hence, Bayesian optimization builds a probabilistic model for the selected performance metric and search for the hyper-parameters combination which maximises the value of such a chosen metric. This approach is here used

[3]https://land.copernicus.eu/

as an alternative to the long computational time characterizing a Grid Search and to the pure randomness which we deal with when using Random Search.

5) Model explanation with Shapley values: Ideally, a ML model should be highly accurate and simple to interpret. By ”simple to interpret” is intended the ability to expose its performance in an understandable and intuitive way. Unfortunately, the more the model complexity increases, the harder it is to understand how certain values were predicted and which features had contributed more to those predictions. When working with a simple model, the impact that a feature has on the model output is easily interpretable by looking at its weights; whereas, complex models such as ensemble methods or deep networks are not as easy to understand; in this scenario, a model explainer can help interpreting the model results [23].

The SHapley Additive exPlanations (SHAP) package was used as model explainer for the best performing model. SHAP was created by Lundberg (2020) [23][4] and, according to its author, its implementation allows for appropriate user trust, provides insights for model improvement, and supports the understanding of the problem being modelled.

The proposed method was implemented using scikit-learn and executed in Anaconda 2020.10 on a computer with NVIDIA GeForce GTX 1650 4 GB RAM.

III. Results

1) Features selection: Figure 1. shows the MDI that each feature brings to the model. We can observe that the first 9 selected features, apart from the DEM, were all extracted from the Summer image. Furthermore, biophysical parameters were often selected, specifically Canopy Chlorophyll Content (LAIcb), Canopy Water Content (LAIcw), and a chlorophyll index calculated using red-edge bands (Clre). A SWIR band (b12), the DEM, and a texture measure, the entropy measure derived from the summer NDVI, were also included in the top selected features.

2) Above Ground Biomass prediction: The worst performing predictive model was the LNN, which yielded 77.2 t/ha as its lowest RMSE, using the first 40 features selected by the MDI method. For the KNN, the lowest RMSE (44.5t/ha) was reached when the first 25 features selected by MDI were included; The RF algorithm yielded the best predictive performance when only the first 5 features selected by the MDI where included, with a RMSE of 44.29. Whereas, for the XGB algorithm, after an initial arise of the model performance, the error started erratically increasing once more than 23 features were included in the model. The lowest RMSE corresponds to 45.44 t/ha and it was achieved by using the first 23 features of the ranking generated by MDI. The SNN ended up in a RMSE of 43.23 t/ha when using 15 variables. Finally, for the DNN, the RMSE had a very erratic pace, and after an initial decrease, it started increasing once more than 5 features were added. The best performing DNN was achieved when using the

[4]https://github.com/slundberg/shap

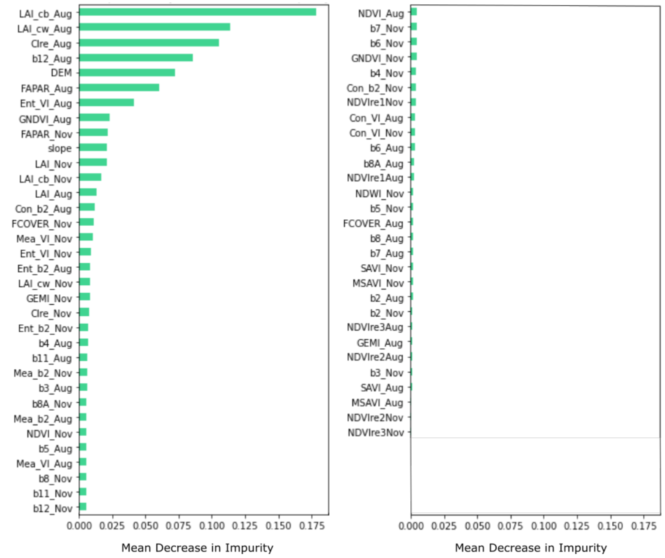

Fig. 1. Feature selection based on Mean Decrease in Impurity

5 features identified by MDI, which yielded an error of 42.30 t/ha. Furthermore, the LNN, SNN and DNN were tested for over-fitting, without showing important signs of it.

3) Hyper-parameters fine-tuning: When working with neural networks, hyper-parameters fine tuning is a complex and time-consuming task, due to the high number of hyper-parameters to be taken into account. Therefore, the Bayesian Search algorithm was applied only to the remaining 3 regression models. After optimization, RF showed an error of 44.16 t/ha, KNN of 42.68 t/ha, and the XGB achieved the lowest error, 37.79 t/ha.

Hyper-parameters optimization for XGB was performed by using as model input the 23 previously selected variables from the MDI feature selection method, and by fine-tuning 6 hyper-parameters within their respective range of values: Alpha (from 0 to 0.9), Gamma (from 0 to 0.9), learning rate (from 0 to 0.9), Number of estimators (from 30 to 1000), maximum depth (from 1 to 8), and the sub-sample (from 0 to 0.9).

Finally, the AGB throughout the study area was predicted using the XGB with Bayesian Optimization and feature selection using the MDI method. The predictions show an average value of 83 t/ha, ranging from 0 t/ha to 346.56 t/ha.

4) Model explanation with Shapley values: By observing the SHAP summary plot in Figure 2, which contain 20 of the 23 selected features, these can be divided into 2 groups: 1) features which increasing in values pushed the model to output AGB higher than the base value; 2) features which increasing in values would push the model to output AGB lower than the base value.

Belonging to the first group are: The Chlorophyll index based on red-edge bands from the Summer month (ClreAug), the Fraction of absorbed radiation from both months (FAPARAug and FAPARNov), the Canopy Chlorophyll Content from the Summer month (LAIcbAug), and the slope; while, band

12 from the Summer image (b12Aug) and the DEM belong to the second group. The remaining features do not show any clear visual pattern in the way they impacted the model output.

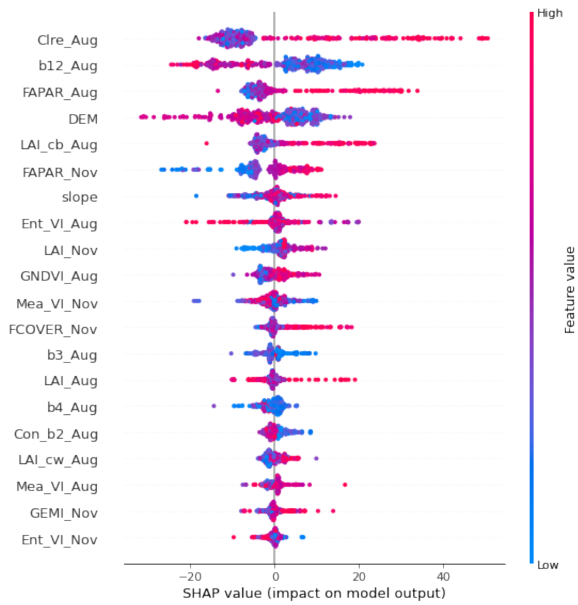

Fig. 2. Summary plot with SHAP

IV. Discussion and Conclusion

Overall, Summer features had a higher position in the ranking than Winter features when using the MDI method. Out of the 63 variables, only 23 were used for the final model. These can be observed in figure 1, as figure 2 only shows the first 20 features, due to a limitation of the SHAP algorithm.

Bayesian Search improved the XGB model predictions; this search method had advantageous computation time compared to the Grid Search, yet its implementation is complex. The XGB algorithm was found to be the best performing algorithm when used in combination with MDI and Bayesian Search.

For all the tested models, AGB values lower than 40-50 t/ha were slightly overpredicted, whereas values higher than 150-160 t/ha were underpredicted. This is a typical issue when estimating AGB with the use of ML and satellite images [8, 19, 7], and it is exacerbated by a limited number of representative samples for low and high values of the forestry inventory.

The insights gathered with SHAP showed the utility of a model explainer for the scientific community. Further work should be focusing on the exploration of model explainers so as to allow for appropriate user trust, and to support the understanding of the problem being modelled [23].

Acknowledgment

This study was carried out in the framework of the MAIL project, which was funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 823805.

References

[1] UNFCCC (1997), 'Kyoto protocol to the united nations framework convention on climate change'.

[2] Schimel, D., House, J., Hibbard, K., Bousquet, P., Ciais, P., Peylin, P., Braswell, B.,Apps, M., Baker, D., Bondeau, A., Canadell, J., Churkina, G., Cramer, W., Denning,A., Field, C., Friedlingstein, P., Goodale, C., Heimann, M., Houghton, R., Melillo, J.,Moore, B., Murdiyarso, D., Noble, M., Pacala, S., Prentice, I., Raupach, M., Rayner,P., Scholes, R., Steffen, W. and Wirth, C. (2001), 'Recent patterns and mechanismsof carbon exchange by terrestrial ecosystems',Nature, 414, 6860, 169–172..

[3] IPCC (2006), '(intergovernmental panel climate change), guidelines for national green-house gas inventories volume 4 agriculture,forestry and other land use', 4, 48.

[4] Torralba and Crespo-Peremarch (2018), Assessing the use of discrete , full-waveform lidar and tls to classify mediterranean forest species composition, Revista de Teledeteccion, Asociación Española de Teledetección, 52, 27-40.

[5] Montero Gregorio and RuizPeinado Ricardo and Muñoz Marta (2005), Producción de biomasa y fijación de CO2 por los bosques españoles. MONOGRAFÍAS INIA: SERIE FORESTAL nº 13.

[6] Pandit, S., Tsuyuki, S. and Dube, T. (2018), 'Estimating above-ground biomass in sub-tropical buffer zone community forests, Nepal, using sentinel-2 data',Remote. Sens., 10, 601.

[7] Forkuor, G., Zoungrana, B., Dimobe, K., Ouattara, B., Vadrevu, K. and Tondoh, E.(2020), 'Above-ground biomass mapping in west african dryland forest using sentinel-1and 2 datasets - a case study',Remote Sensing of Environment, 236, 111496.

[8] Galidaki, G., Zianis, D., Gitas, I., Radoglou, K., Karathanassi, V., Tsakiri–Strati, M.,Woodhouse, I. and Mallinis, G. (2017), 'Vegetation biomass estimation with remote sensing: focus on forest and other wooded land over the mediterranean ecosystem',International Journal of Remote Sensing 38(7), 1940–1966.

[9] Pandit, S., Tsuyuki, S. and Dube, T. (2019), 'Exploring the inclusion of sentinel-2 msi texture metrics in above-ground biomass estimation in the community forest of Nepal', Geocarto International, 30, 1832-1849.

[10] Kelsey, K. and Neff, J. (2014), 'Estimates of above ground biomass from texture analysis of Landsat imagery',Remote Sensing, 6, 6407–6422.

[11] Zheng, D., Rademacher, J., Chen, J., Crow, T., Bresee, M. K., Moine, J. L. and Ryu, S. R.(2004), 'Estimating above ground biomass using landsat 7 etm+ data across a managed landscape in northern wisconsin, usa',Remote Sensing of Environment, 93, 402–411.

[12] Vicharnakorn, P., Shrestha, R., Nagai, M., Salam, A. P. and Kiratiprayoon, S. (2014),'Carbon stock assessment using remote sensing and forest inventory data in savan-nakhet, lao pdr',Remote Sens.6, 5452–5479.

[13] Puliti, S., Hauglin, M., Breidenbach, J., Montesano, P., Neigh, C., Rahlf, J., Solberg,S., Klingenberg, T. F. and Astrup, R. (2020), 'Modelling above-ground biomass stock over norway using national forest inventory data with arctic dem and sentinel-2 data',Remote Sensing of Environment, 236.

[14] Babyak, M. (2004), 'What you see may not be what you get: A brief, non technical introduction to over-fitting in regression-type models', Psychosomatic medicine, 66, 411–21.

[15] Smith, G. (2018), 'Step away from stepwise',Journal of Big Data5, 32.

[16] Lewinson, E. (2019), 'Explaining feature importance by example of a random forest', Towards Data Science.

[17] Hong Han, Xiaoling Guo and Hua Yu (2016), Variable selection using mean decrease ac-curacy and mean decrease gini based on random forest,in'2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)', pp. 219–224.

[18] Gao, Y., Lu, D., Li, G., Wang, G., Chen, Q., Liu, L. and Li, D. (2018), 'Comparative anal-ysis of modeling algorithms for forest above ground biomass estimation in a subt-ropical region', Remote Sens. 10, 627.

[19] Salem, I., Basam, D., Taoufik, K. and Nazmi, S. (2020), 'A review of terrestrial carbon assessment methods using geo-spatial technologies with emphasis on arid. lands',Remote Sens. 12, 2008.

[20] Breiman, I. (2001), 'Random forests.', Machine Learning (1), 5–32.

[21] Seif, G. (2019), 'A beginner's guide to xgboost',Towards Data Science.

[22] Nogueira, F. (2014), 'Bayesian Optimization: Open source constrained global optimiza-tion tool for Python'.

[23] Lundberg, S., Erion, G., Chen, H., DeGrave, A., Prutkin, J., Nair, B., Katz, R., Him-melfarb, J., Bansal, N. and Lee, S.I. (2020), 'From local explanations to global under-standing with explainable AI for trees', Nature Machine Intelligence.