European Space Agency's 8th Advanced Training Course on Land Remote Sensing

Sentinel 2 and Machine Learning for biomass modelling

Davor Korman[1], Alen Berta[1], Zrinka Mesić[1], Ana Ostojić[1], Nela Jantol[1], Ivona Žiža[1],Vladimir Kušan[1]

1 Oikon ltd Institute of Applied Ecology

INTRODUCTION

Remote sensing techniques have been used in forestry for last several decades, but due to the variety of available imagery data and data processing algorithms, the single best method for aboveground biomass (AGB) assessment is not existent. Furthermore, there is no quorum whether is it more appropriate to create AGB assessment models with data averaged on seasonal or monthly temporal scale in order to create the most accurate model. Besides, classical statistical methods often result in models with moderate or poor performance, while on the other hand Machine Learning techniques, coupled with the advancements in computer hardware and an increase in processing power, present new opportunities in remote sensing and AGB assessment modelling.

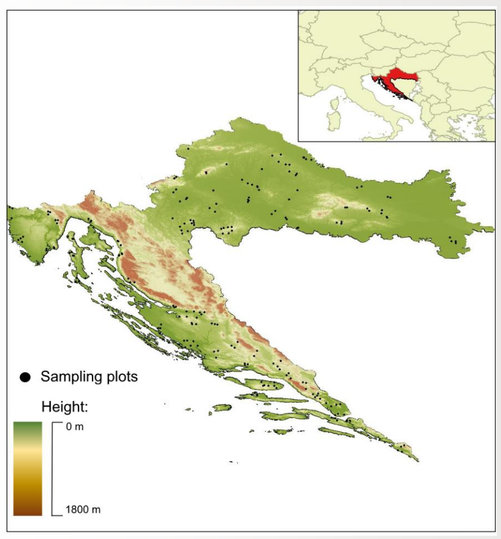
RESEARCH OBJECTIVE

In this research we have tested the effectiveness of using Sentinel 2 imagery for AGB assessment and designed a practical approach (algorithm) for extracting information from large amounts of satellite data.

Our goals were to:

1) create models for aboveground biomass assessment for thicket, maquis and continental forests of first age class using Sentinel 2 data
2) investigate and compare the suitability of monthly and seasonally averaged data for AGB assessment
3) explore the possibility of improving model accuracy by implementing various machine learning techniques.

STUDY AREA



METHODS – GROUND TRUTH DATA

Field data was gathered on 204 sample plots covering six different forest stands across Croatia:

Management class	Number of sample plots for field measurement AGB assessment
Common beech	24
Sessile oak	24
Common oak	20
Narrow-leafed ash	16
Continent total	84
Thicket	76
Maquis	44
TOTAL	204

Gathered data was analysed in several steps for the purpose of obtaining total AGB per sample plot:

I. Measurement of tree height and diameter of each tree inside the sample plot
II. Calculation of individual tree volumes using Schumaher-Hall volume equation

$$V = a*db*hc * f$$

III. Calculation of total volume per sample plot

Response variable for image based modelling

METHODS – REMOTE SENSING DATA

For the purpose of the research and for the comparison of seasonal and monthly based AGB models, all available Sentinel 2 imagery in 2016 growing season was downloaded

(9 Sentinel 2 bands + 110 vegetation indices) * month | season

1100 independent variables

METHODS – DATA DOWNLOAD AND PROCESSING

DATA PROCESSING STEPS

1 DATA DOWNLOAD
- Python API „sentinelsat"
- Available at Github

TERABYTES OF SATELLITE IMAGERY

2 ATMOSPHERIC CORRECTION
- sen2Cor stand-alone software
- Operable as a Windows batch script

3 CONVERSION TO TIFF
- GDAL library
- Runs in Windows Command Line environment

BIG DATA

4 DATA EXTRACTION ALGORITHM
- R(Studio) software package
- Batch pixel values extraction algorithm

7 000 000 PIXEL VALUES EXTRACTED

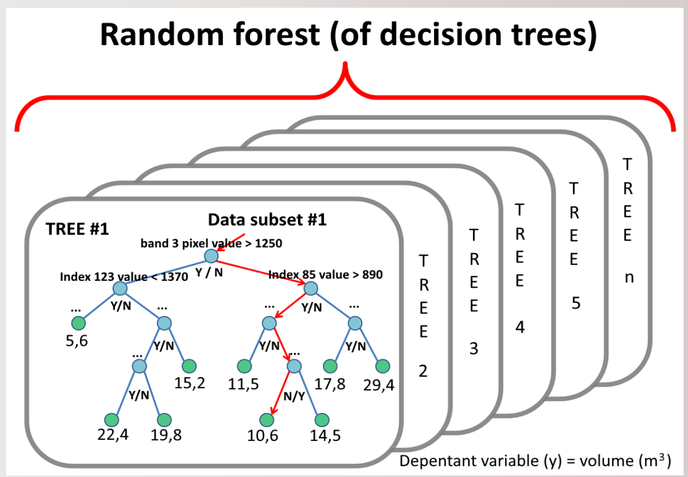
OPEN SOURCE SOFTWARE USED ONLY

Code available at https://github.com/DKorman

METHODS – PRIMARY MODEL BUILDING PIPELINE

I. Random forest for selecting best among ~1100 variables
II. Stepwise regression for additional variable reduction
III. Creation of multiple regression model:

$$Volume = a + b1* VegetationIndex1 + b2*VI2 \ldots$$



High interpretability, but... (see results)

PRIMARY MODEL RESULTS & CONCLUSION

Seasonal vs. monthly models

Monthly averaged data has been selected as the most appropriate independent variable in all cases

Model accuracy

Forest stands	R^2 adj.	Percent relative standard error
Common beech and sessile oak	0.65	41.75
Common oak and narrow-leafed ash	0.80	30.43
Thicket in Istria and Kvarner	0.64	33.12
Thicket in Dalmatia	0.55	36.87
Maquis	0.53	61.99

Sentinel 2 based AGB asessment resulted in models with moderate or even poor performance

IMPROVING ACCURACY WITH MACHINE LEARNING

Machine learning technique	Hyperparameters tuned	Role in the model creation pipelines
Booststrap aggregating	Number of trees, tree depth, number of observation in nodes	Model creation and variable selection
Random forest	-||-, number of features selected	Model creation and variable selection
Gradient boosting	Shrinkage,depth, number of observations in node, number of trees	Model creation and variable selection
Extreme gradient boosting	Depth, eta, number of trees, lambda, alpha	Model creation and variable selection
Genetic algorithm	Percentage of mutation , elitism, crossover	Model creation and variable selection
Support vector machines	Epsilon, cost, gama	Model creation
Princ. component analysis		Dimensionality reduction

RESULTS OF MACHINE LEARNING IMPLEMENTATION

Some of the pipelines tested (example on 2 forest stands

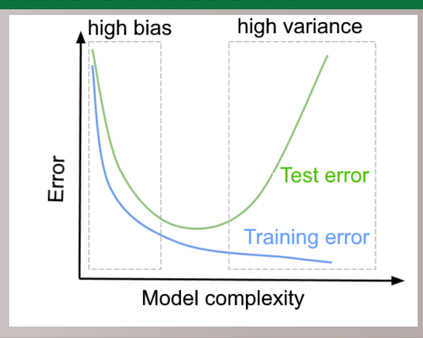
ML PIPELINE USED	RMSE	
	sessile oak	thicket
Random forest - Linear model (PRIMARY MODEL)	22,04	5,25
Support vector machines	21,32	6,013
Random forest - Random forest	11,61	3,78
PCA -Random forest	18,32	5,73
Genetic algorhitm - LR	19,08	5,64
PCA -> LR	17,83	5,85

ML PIPELINE USED	RMSE	
	sessile oak	thicket
Gradient boosting – LR	17,80	5,87
Extreme gradient bosting (tree) – LR	18.73	5,87
Extreme g. bosting (linear) - LR	16,33	6,29
Random forest - Extreme g.boosting	10,28	3,80
Random forest - Support vector machines	13,19	6,24
Random forest - gradient boosting	10,59	3,78

MACHINE LEARNING CONCLUSION

MACHINE LEARNING CAN BE AN EXCELLENT TOOL FOR BIOMASS MODELLING

BUT



Contact: Davor Korman,
Professional associate
M +385 91 6054 652
E korman.davor@gmail.com

https://www.linkedin.com/in/davorkorman
https://www.researchgate.net/profile/Davor_Korman
https://github.com/DKorman

OikonLtd. Institute forApplied Ecology

Trg senjskih uskoka 1-2
HR –10020 Zagreb
T +385 1 5507 100

F +385 1 5507 101
E oikon@oikon.hr
W www.oikon.hr