MODELLING STAND VARIABLES OF PINE FOREST USING SENTINEL-2A DATA AND THE RANDOM FOREST APPROACH

S. Arellano-Pérez[1], M.A. González-Rodríguez[1], F. Castedo-Dorado[2], C.A. López-Sánchez[3], C. Pérez-Cruzado[1], J.G. Álvarez-González[1] and A.D. Ruiz-González[1]

[1] Escuela Politécnica Superior de Ingeniería. Universidad de Santiago de Compostela. Campus Universitario s/n. 27002, Lugo (Spain). stefano.arellano@gmail.com, miguelangel.gonzalez.rodriguez@rai.usc.es, cesar.perez@usc.es, juangabriel.alvarez@usc.es, anadaria.ruiz@usc.es.

[2]Escuela Superior y Técnica de Ingeniería Agraria. Universidad de León. Avda. Astorga s/n. 24400, Ponferrada (Spain). fcasd@unileon.es

[3]Escuela Politécnica de Mieres. Universidad de Oviedo. C/ Gonzalo Gutiérrez de Quirós s/n. 33600, Mieres (Spain). lopezscarlos@uniovi.es

Introduction

Quantification of stand variables such as volume and biomass is an important issue in forest management. Reliable information is required for accurate stand variables estimation and, traditionally, forest inventory has been carried out by systematically designed field measurements. This approach is expensive and time-consuming; however, the use of auxiliary remotely sensed data linked to the development of statistical frameworks to ensure a rigorous uncertainty assessment (Gregoire et al., 2016) has allowed an accuracy estimation of stand variables at lower cost and over larger areas (Kauranne et al., 2017; Puliti et al., 2018). In this paper, models to estimate the main stand variables of a thinning trial of pine species using data from the moderate image resolution Sentinel-2 satellite have been developed using the random forest approach.

Materials and methods

The study area was located in the North-west of Spain. The dataset corresponds to 41 thinning trial locations installed in pure, even-aged stands of *Pinus pinaster* (22 locations) and *Pinus radiata* (19 locations). At each location, three rectangular plots (1000 m^2 in size) were established and georeferenced by using a differential GPS. A different treatment was applied to each plot: unthinned control, moderate thinning (20% of the basal area removed) and heavy thinning (40% of the basal area removed). The plots were thinned from below in 2010 and were re-measured at different age intervals, although in this study, only the measurement carried on the summer of 2016 was used.

Diameter at breast height (d) of all the trees was measured to the nearest 0.1 cm. Total tree height (h) was measured to the nearest 0.1 m in a randomized sample of 30 trees and in an additional sample of dominant trees (the proportion of the 100 largest diameter trees per hectare, depending on plot size). Total tree height for the remaining trees was estimated using the h-d model developed for these species (Diéguez-Aranda et al., 2009). The number of stems per hectare (N), stand basal area (G), mean height ($\bar{h}$) and stand dominant height (H, defined as the mean height of the 100 thickest trees per hectare) were calculated from tree variables and the tree volume and tree biomass equations developed for these species in Galicia (Diéguez-Aranda et al. 2009) were used to estimate the stand volume (V) and stand aboveground biomass (W). The mean, maximum, minimum values and standard deviation for the main tree and stand variables are shown in Table 1.

Table 1. Statistics of the main tree and stand variables. Std. dev is the standard deviation; d = tree diameter, h = tree height, t = stand age; N = stem density; G = stand basal area; H = dominant height (defined as the mean height of the 100 thickest trees per ha); $\bar{h}$ = mean height; V = stand volume and W = stand aboveground biomass.

Species	Statistic	Tree variables (n = 10831)			Stand variables (n = 123)					
		d (cm)	h (m)	t (years)	N (stems ha^{-1})	G (m^2 ha^{-1})	H (m)	$\bar{h}$ *(m)*	V (m^3 ha^{-1})	W (Mg ha^{-1})
P. pinaster	Mean	21.12	14.96	23.67	1028.22	39.08	16.08	14.82	252.87	137.19
	Std. dev.	6.50	2.96	4.18	480.64	9.34	2.69	2.65	86.96	44.01
	Min.	3.90	5.60	18	407	20.04	10.98	4.53	98.13	54.18
	Max.	48.50	23.7	38	3095	66.08	21.35	13.03	517.08	267.85
P. radiata	Mean	21.92	20.13	22.43	838.47	34.57	22.67	19.91	308.40	152.07
	Std. dev.	6.78	3.88	2.06	427.53	9.88	2.78	2.75	96.37	47.65
	Min.	2.75	5.80	18	316	15.59	16.15	14.40	120.20	58.91
	Max.	47.55	30.10	28	2899	66.63	27.23	25.54	509.60	259.14

Cloud-free S-2A multispectral instrument (MSI) level 1-C (L1C) imagery acquired on July 19 and August 1, 2016 (Table 2), matching the field inventory of the sample plots, were downloaded from the U.S. Geological Survey (USGS) Global Visualization Viewer at http://glovis.usgs.gov/.

Table 2. Acquisition dates, solar elevation and azimuth angles of S-2 scenes.

Scene	Acquisition Date	Solar Elevation (◦)	Solar Azimuth (◦)
S2A_tile_20160801_29TMH	8/01/2016	61.92	148.41
S2A_tile_20160719_29TNG	7/19/2016	63.25	144.23
S2A_tile_20160801_29TNH	8/01/2016	61.92	148.41
S2A_tile_20160719_29TPG	7/19/2016	63.25	144.23
S2A_tile_20160719_29TPH	7/19/2016	63.25	144.23
S2A_tile_20160719_29TPJ	7/19/2016	63.25	144.23
S2A_tile_20160719_29TQH	7/19/2016	63.25	144.23

The 12-bit S-2A MSI image has 13 spectral bands in the visible (VIS), near infrared (NIR), and shortwave infrared (SWIR) wavelength region, with spatial resolutions from 10 to 60 m. For this study, the three "atmospheric" bands (B1, B9 and B10) were discarded. The data preparation involved the resampling of the S2 bands acquired at 20 m to obtain a layer stack of 10 spectral bands at 10 m using the ESA's S-2 toolbox from ESA Sentinel Application Platform (SNAP) and then converted to ENVI format (Table 3).

Table 3. Spatial and spectral resolutions of S-2/MSI.

Sentinel-2/MSI (um)	Band	Resolution (m)
Band 2 (0.46–0.52)	Blue	10
Band 3 (0.54–0.58)	Green	10
Band 4 (0.65–0.68)	Red	10
Band 5 (0.7–0.71)	Red-edge-1	20
Band 6 (0.73–0.75)	Red-edge-2	20
Band 7 (0.76–0.78)	Red-edge-3	20
Band 8 (0.78–0.90)	NIR	10
Band 8A (0.85–0.87)	NIR plateau	20
Band 11 (1.56–1.65)	SWIR-1	20
Band 12 (2.10–2.28)	SWIR-2	20

Since atmospherically corrected images are essential to assess spectral indices with spatial reliability and products comparison, L1C data were processed to level-2A (L2A, bottom-of-atmosphere reflectance) taking into account the effects of aerosols and water vapor on reflectances. These corrections were performed using the Sen2Cor tool [46] for the S-2 images. Then, the rectified images were used to calculate five vegetation indices (VIs): NDVI, SAVI, MSAVI, EVI and RENDVI. The normalised difference vegetation index (NDVI) [47] is currently the most widely used VI as a feature for modelling many target variables. The soil adjusted vegetation index (SAVI) [48] is a refinement of the NDVI aimed to account for uncertainty due to variation in background condition. SAVI incorporated the correction factor L into the NDVI formula. L accounts for soil variation by varying the factor between 1, for low vegetation, and 0, for dense vegetation. In this study, a value of 0.5 was assumed. Qi et al. [49] presented a modified version of the SAVI (MSAVI) which utilised a self-adjusting L factor. The enhanced vegetation index (EVI) was developed by Huete et al. [50] to account for aerosol variation and also as a vegetation index less prone to saturation on dense green vegetation canopy. Finally, the red-edge NDVI (RENDVI) is based on NDVI but it uses the red-edge spectral band B6 instead of the red spectral band B4 [51]. The formulation and the bands used to estimate the VIs are shown in Table 4.

Five metrics (mean, standard deviation, minimum, median and maximum) were computed from the area-weighted values of the pixels completely contained or intersected by plot boundaries, using the 10 bands (B2 to B8, B12 and B13) and 5 VIs previously described. This lead to 75 different auxiliary features for modelling canopy and surface fuel variables.

Table 4. Formulation of the vegetation indices (VIs) calculated from S-2 image data. NDVI: Normalised difference vegetation index, SAVI: Soil adjusted vegetation index, MSAVI: modified version of the SAVI, EVI: Enhanced vegetation index, B4, B6, B8: 12-bit S-2A bands.

Vegetation index	Formulation	S-2 bands used
NDVI [47]	$\frac{(NIR - Red)}{(NIR + Red)}$	$\frac{(B8 - B4)}{(B8 + B4)}$
SAVI [48]	$\frac{(1 + L)(NIR - Red)}{(NIR + Red + L)}$	$\frac{(1.5)(B8 - B4)}{(B8 + B4 + 0.5)}$
MSAVI [49]	$\frac{(1 + L)(NIR - Red)}{(NIR + Red + L)}$ $L = 1 - \frac{(NIR - Red)(NIR - 0.5Red)}{(NIR + Red)}$	$\frac{\left(2 - \frac{(B8 - B4).(B8 - 0.5B4)}{(B8 + B4)}\right)(B8 - B4)}{\left(B8 + B4 + 1 - \frac{(B8 - B4).(B8 - 0.5B4)}{(B8 + B4)}\right)}$
EVI [50]	$\frac{G \cdot (NIR - Red)}{(NIR + C1 \cdot Red - C2 \cdot Blue + L)}$	$\frac{2.5(B8 - B4)}{(B8 + 6 \cdot B4 - 7.5 \cdot B2 + 1)}$
RENDVI [51]	$\frac{(NIR - Red\_edge)}{(NIR + Red\_edge)}$	$\frac{(B8 - B6)}{(B8 + B6)}$

Random Forest was used to relate the stand variables with S-2 bands and VIs by using the randomForest package (Liaw and Wiener, 2002) of the R software (R Core Team, 2017) and setting the number of trees to 1000. Tree species and thinning intensity were initially considered as potential covariates, however, their inclusion in the models implies the need to obtain a classification system of the S-2 images to differentiate between species and thinning intensities, therefore, a classification tree was fitted using the rpart package (Therneau et al., 2017) of the R software (R Core Team, 2017). Two goodness-of fit statistics were used to evaluate the performance of the models: the percentage of the root mean squared error (rRMSE) over the mean value and the square of the correlation coefficient between observed and estimated stand variables (ρ^2).

Results

The confusion matrix of the RF classification of species and thinning treatments is show in Table 5. The overall accuracy of this model was 83.74% with a Kappa value of 0.8044, while the relative accuracies for species and treatments classification were 96.75% and 83.74%, respectively (Kappa values of 0.9343 and 0.7561). The most important auxiliary features were the red-edge normalized difference index (RENDVI) and metrics related to two bands in the visible (B2, B4) and two bands in the shortwave infrared spectral range (B11, B12). The predicted species and treatment classes were then considered as auxiliary features to fit the MARS and RF models of the four fuel variables analysed.

Table 5. Confusion matrix related to the RF model fitted to differentiate between species and thinning treatments (C - Control, MT – Moderate thinning and HT - Heavy thinning).

			Observed						
			P. pinaster			*P. radiata*			
			C	LT	HT	C	LT	HT	User's accuracy
Predicted	*P. pinaster*	C	19	4	1	0	1	0	79.17%
		MT	2	18	3	1	0	0	75.00%
		HT	1	0	18	0	2	0	84.21%
	P. radiata	C	0	0	0	16	0	0	100%
		MT	0	0	0	1	16	3	80.00%
		HT	0	0	0	1	0	16	94.12%
Producer's accuracy			86.36%	81.82%	81.82%	84.21%	84.21%	84.21%	83.74%

The predicted classification was considered as a new feature to fit the models. The goodness-of-fit statistics of the random forest models are shown in table 6.

Variable	rRMSE(%)	ρ^2
N (stems/ha)	43.4888	0.2330
G (m^2/ha)	24.9248	0.1273
$\bar{h}$ (m)	14.2474	0.5714
H (m)	13.6585	0.6383
V (m^3/ha)	29.5370	0.2546
W (t/ha)	29.1775	0.1642

Table 6. Goodness-of-fit statistics of the random forest models.

The observed variability explained by the models ranged from 13% to 64% with the best results for dominant height (H) and mean height ($\bar{h}$). The relative importance of each feature to estimate the four fuel variables using both approaches is shown in Figure 2.

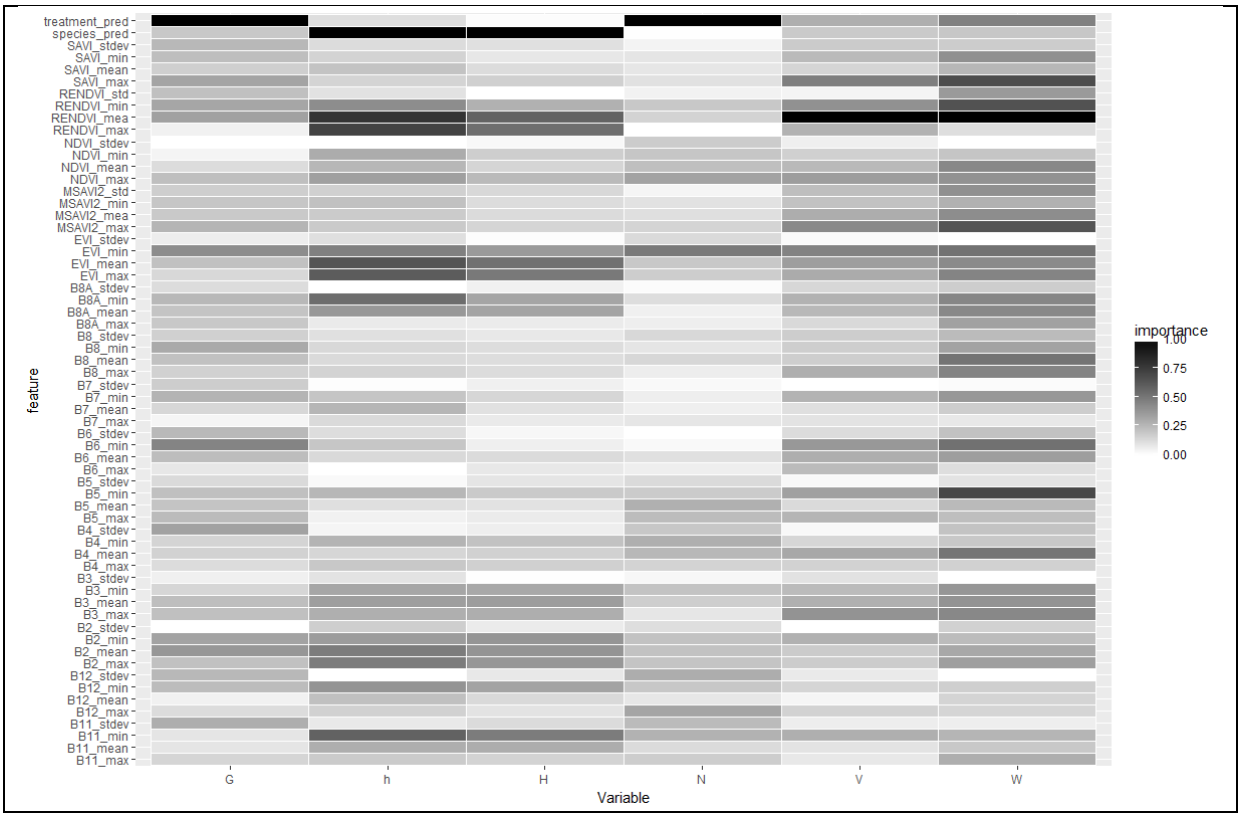

Figure 2. Relative importance of each auxiliary feature to model the main stand variables using the RF modelling approach.

The most important variables to estimate the stand variables were the predicted species and treatment values obtained with the classification tree and features related to the RENDVI and EVI vegetation indices.

Conclusions

The results obtained in this study indicated the models obtained using only features related to Sentinel-2 data are not accuracy enough for stand variables estimation except for stand heights. These poor results contrast with those obtained for other authors using moderate resolution Landsat images (e.g. Zheng et al., 2004; Hall et al., 2006); however, sample plots from a thinning trial of two different pine species have been used in this study, implying a great variability of the stand variables analysed. Therefore, due to the enhanced spatial, spectral and temporal characteristics of Sentinel-2 compared with Landsat, this sensor provides a great opportunity for stand variables estimation and updating and there is a need for further investigation to evaluate the potential of Sentinel-2 data to estimate these variables

References

Diéguez-Aranda, U., Rojo Alboreca, A., Castedo-Dorado, F., Álvarez González, J.G., Barrio-Anta, M., Crecente-Campo, F., et al., 2009. Herramientas selvícolas para la gestión forestal sostenible en Galicia. Consellería do Medio Rural, Xunta de Galicia. Santiago de Compostela, España.

Gregoire, T.G., Næsset, E., McRoberts, R.E., Ståhl, G., Andersen, H.-E., Gobakken, T., Ene, L., Nelson, R., 2016. Statistical rigor in Lidar-assisted estimation of aboveground forest biomass. Remote Sensing of Environment, 173, pp.98–108.

Hall, R.J., Skakun, R.S., Arsenault, E.J., Case, B.S., 2006. Modeling forest stand structure attributes using Landsat ETM+ data: Application to mapping of aboveground biomass and stand volume. Forest Ecology and Management, 225, pp.378-390.

Kauranne, T., Joshi, A., Gautam, B., Manandhar, U., Nepal, S., Peuhkurinen, J., Hämäläinen, J., Junttila, V., Gunia, K., Latva-Käyrä, P., Kolesnikov, A., Tegel, K., Leppänen, V., 2017. LiDAR-Assisted Multi-Source Program (LAMP) for Measuring Above Ground Biomass and Forest Carbon. Remote Sensing, 9, pp.154.

Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R News 2(3), pp.18-22.

Puliti, S., Saarela, S., Gobakken, T., Ståhl, G., Næsset, E., 2018. Combining UAV and Sentinel-2 auxiliary data for forest growing stock volume estimation through hierarchical model-based inference. Remote Sensing of Environment, 204, pp.485-497.

R Core Team, 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (Available on: https://www.R-project.org/.)

Therneau, T., Atkinson, B., Ripley, B., 2017. rpart: Recursive Partitioning and Regression Trees. R package version 4.1-11. https://CRAN.R-project.org/package=rpart

Zheng, D., Rademacher, J., Chen, J., Crow, T., Bresee, M., Le Moine, J., Ryu, S.-R., 2004. Estimating aboveground biomass using Landsat 7 ETM+ data across a managed landscape in northern Wisconsin, USA. Remote Sensing of Environment, 93, pp.402–411