Journal of Applied Remote Sensing

RemoteSensing.SPIEDigitalLibrary.org

Comparison of satellite-based estimates of aboveground biomass in coppice oak forests using parametric, semiparametric, and nonparametric modeling methods

Amir Safari
Hormoz Sohrabi
Scott Powell

SPIE.

Amir Safari, Hormoz Sohrabi, Scott Powell, "Comparison of satellite-based estimates of aboveground biomass in coppice oak forests using parametric, semiparametric, and nonparametric modeling methods," *J. Appl. Remote Sens.* **12**(4), 046026 (2018), doi: 10.1117/1.JRS.12.046026.

Comparison of satellite-based estimates of aboveground biomass in coppice oak forests using parametric, semiparametric, and nonparametric modeling methods

Amir Safari,[a] Hormoz Sohrabi,[a,*] and Scott Powell[b]
[a]University of Tarbiat Modares, Department of Forestry, Faculty of Natural Resources, Tehran, Iran
[b]Montana State University, Department of Land Resources and Environmental Sciences, Bozeman, Montana, United States

Abstract. Accurate estimates of forest biomass are essential for several purposes, ranging from carbon accounting and ecological applications to sustainable forest management. There are, however, critical steps for mapping aboveground forest biomass (AGB) based on optical satellite data with an acceptable degree of accuracy, such as selecting the proper statistical modeling method and deriving spectral information from imagery, at known field locations. We compare nine modeling techniques including parametric, semiparametric, and nonparametric methods for remotely estimating AGB based on various spectral variables derived from Landsat 8 Operational Land Imager (OLI). We conduct this research in Zagros oak forests on two sites under different human disturbance levels: an undegraded (UD) forest site and a highly degraded (HD) forest site. Based on cross-validation statistics, the UD site exhibited better results than the HD site. Support vector machine (SVM) and Cubist regression (CR) were more precise in terms of coefficient of determination (R^2), root-mean-square error (RMSE), and mean absolute error (MAE), though these approaches also result in more biased estimates compared to the other methods. Our findings reveal that if the degree of under- or over-estimation is not problematic, then SVM and CR are good modeling options ($R^2 = 0.73$; RMSE = 31.5% of the mean, and MAE = 3.93 ton/ha), otherwise, the other modeling methods such as linear model, *k*-nearest neighbor, boosted regression trees, generalized additive model, and random forest may be better choices. Overall, our work indicates that the use of freely available Landsat 8 OLI and proper statistical modeling methods is a time- and cost-effective approach for accurate AGB estimates in Zagros oak forests. © *2018 Society of Photo-Optical Instrumentation Engineers (SPIE)* [DOI: 10.1117/1.JRS.12.046026]

Keywords: aboveground biomass; Landsat 8 Operational Land Imager; parametric; semiparametric; nonparametric; Zagros.

Paper 180497 received Jun. 12, 2018; accepted for publication Nov. 13, 2018; published online Dec. 11, 2018.

1 Introduction

Accurate estimation of aboveground biomass (AGB) at broad scales is critical for understanding forests' contribution to the global carbon cycle.[1–3] Therefore, developing robust, simple, reliable, and cost-effective approaches to accurately estimate AGB is a critical task.[1,4–6] AGB estimation in forest ecosystems is commonly done by field survey, however, due to high data acquisition costs and intensive field labour, this method is inefficient.[4,5,7,8] Fortunately, the combination of derived variables from remotely sensed data and statistical model-based techniques provides an accurate means for quantifying AGB and AGB dynamics, particularly at broad scales.[7–11] The spectral information, recorded by remote sensing and vegetation indices, has been shown to correlate with forest AGB estimation.

*Address all correspondence to Hormoz Sohrabi, E-mail: hsohrabi@modares.ac.ir

1931-3195/2018/$25.00 © 2018 SPIE

There are the numerous studies of AGB estimation from local to global scales, based on remotely sensed data, from passive optical sensors to active sensors such as synthetic aperture radar and light detection and ranging.[12,13] However, most of these efforts have focused on estimation of biomass in dense forests (e.g., temperate, tropical, and boreal),[14] and relatively few studies have focused on remotely sensed AGB estimates in other regions, such as forests in arid and semiarid climates.[7,14,15] In these areas, sparse canopy vegetation often results in mixed spectral reflectance that is strongly affected by bare soil or shadowing.[1,7,16,17] One such ecosystem is the six million hectare semiarid oak forests of the Zagros Mountains of Iran.[18,19] Here traditional human activities such as the conversion of forest to agricultural land[20] and the cutting of branches and sprouts to provide fuel and animal feed, continue to change forest structure and function.[20,21] Estimates of the distribution and total AGB for these forests across different disturbance levels have yet to be established. Therefore, cost-efficient methods are needed for wall-to-wall AGB mapping. The freely available Landsat 8 Operational Land Imager (OLI) data are a practical option for AGB estimation, because compared to earlier Landsat sensors, the OLI sensor has several key improvements, including (i) narrower near-infrared wavebands, (ii) higher signal-to-noise ratio, and (iii) improved radiometric sensitivity.[15,16,22]

For AGB estimation using remotely sensed data, there are two important issues. The first issue relates to remotely sensed derived spectral information,[2,23,24] which are used as independent variables in modeling methods. The second issue concerns appropriate statistical modeling methods.[8,25,26] Many previous studies have applied different Landsat-derived spectral information (e.g., raw bands, vegetation indices, and linear transformation)[11,27–31] and statistical modeling methods ranging from parametric to semiparametric and nonparametric for AGB estimates in dense forests.[1,8,9,11,13,16,28–32] However, the application of these approaches in sparse forests with low-biomass density remains a challenge.[1,7,22,33]

As observed in many previous studies,[8,11,13,15,29,31,32,34] the comparison of methods has been faced with two primary challenges: (1) no single modeling method has been determined to be the best and (2) the performance of these methods is affected by forest type, forest structure, and sample size.[8,11,35] Nonetheless, Fassnacht et al.[8] concluded that the modeling method is as important as the data type in deriving accurate AGB estimates.

The main objective of this study, therefore, was to compare different statistical modeling methods for generating estimates of AGB in Zagros coppice oak forests with low-biomass density. Within this primary goal, there were two main objectives: (1) to determine the most suitable set of predictor variables for the different modeling methods and (2) to assess the effect of anthropogenic activities on Landsat-based estimates of AGB.

2 Materials and Datasets

2.1 *Study Area*

This study was conducted in dry oak forests located in the Zagros Mountains of western Iran. These forests are mostly sparse and open park-like forests comprised of various oak species (*Quercus* spp.).[19] To evaluate the accuracy of AGB estimates in forest stands affected by human activities, we selected two sites with different degrees of disturbance. The first site is located in the Gahvareh region with undegraded (UD) forests, and the second site is located in the Sarfiruz-Abad region with highly degraded (HD) forests (Fig. 1). Both of these sites are dominated by coppice oak (*Quercus brantii* Lindle) trees. The UD and HD sites have semihumid and semiarid climates, respectively. At the HD site, forest lands have been converted to farm lands and oak trees are mostly used to provide charcoal and fuel wood. Overgrazing is also common at this site; however, at the UD site, only minor cattle grazing occurs.

2.2 *Field Data*

We stratified both test sites (UD and HD) by a Landsat-derived leaf area index (LAI) map based on a developed global model,[36] into three strata (low, moderate, and high LAI). Field sampling was carried out using a systematically gridded design with a 200 m × 200 m grid in each stratum.

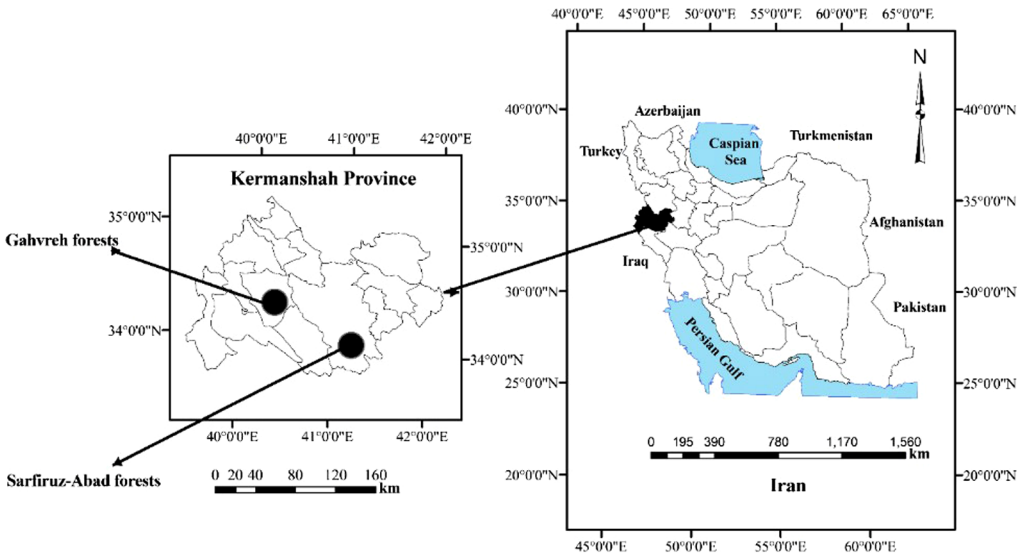

Fig. 1 Location map of the study areas subject to two different disturbance levels in Kermanshah province, western Iran.

In June and July of 2015, we measured 124 georeferenced plots (61 plots at the HD site and 63 plots at the UD site), each 30 m × 30 m (corresponding to a Landsat pixel). To reduce GPS horizontal locational error (~5 to 10 m), final plot positions were determined based on the criterion that the forest structure and composition in a 10-m buffer around the plot be the same as within the plot.[7,37] The AGB per tree was calculated by applying species-specific allometric equations developed for Zagros dry oak trees.[38] Then we calculated AGB in tons per plot by summing each tree's AGB in each plot and converting the per plot AGB values to tons/hectare (t/ha).

2.3 *Satellite Image Acquisition and Preprocessing*

Our test sites were covered by one Landsat 8 tile (path/row: 167/36), which was downloaded from the USGS Earth Resources Observation and Science (EROS) Centre archive (http://earthexplorer.usgs.gov/). This cloud-free image was acquired by Landsat 8 on August 10, 2015, which is in the peak of Zagros forests growing season and near the date of the forest inventory. Image preprocessing involved georeferencing, radiometric calibration, and topographic correction. First, the image was geometrically corrected to a digital map. Then the digital numbers were calibrated to at-sensor radiance via the orbital and sensor parameters. Atmospheric corrections were not applied because: (1) the two test sites were covered by one single-image tile and the atmospheric conditions were, therefore, assumed to be the same; (2) the absolute surface reflectance is not required in empirical methods; and (3) if accurate atmospheric characteristics are not known, the atmospheric correction may increase the uncertainties.[11] The recorded radiance by optical sensors in mountainous areas is strongly affected by topography and leads to large uncertainties in remote sensing-based forest attribute retrieval.[11,39,40] In this study, the C-correction method was used to conduct topographic correction, which is recommended for Landsat images.[40]

For each sample plot location, we extracted radiance values in four categories: raw bands, simple band ratios, vegetation indices, and linear transformations [principle component analysis (PCA) and tasseled cap (TC)]. In total, 50 explanatory variables were derived from the Landsat 8 image (see Table 1).

2.4 *Predictor Variables*

We followed four steps in selecting the best predictor variables for each modeling method, in an approach similar to Görgens, Montaghi, and Rodriguez.[44] First, we calculated a correlation

Table 1 Landsat 8 OLI derived spectral information used in aboveground biomass estimation

Categories	Details	References
Raw bands	B, G, R, NIR, SWIR 1 and 2	30
Simple band ratios	B/G, B/R, B/NIR, B/SWIR1, B/SWIR2, G/R, G/NIR, G/SWIR1, G/SWIR2, R/NIR, R/SWIR1, R/SWIR2, NIR/SWIR1, NIRR/SWIR2, SWIR1/SWIR2	13, 15, and 41
Vegetation indices	ALBEDO, ARVI, EVI, MSAVI2, NGDI, NDVI, OSAVI, PVI, RDVI, SARVI, SAVI, TNDVI, TSAVI, DVI	13, 15, 30, and 42
Linear transformations: principal component analysis and tasseled cap	PC1, PC2, PC3, PC4, PC5, PC6, TCB,TCG, TCW, TCA, TCD	6, 15, 24, 29, 41, and 43

Note: B, blue; G, green; R, red; NIR, near-infrared; SWIR 1 and 2, shortwave-infrared 1 and 2; Albedo: total shortwave, infrared, and visible; ARVI, atmospherically resistant vegetation index; EVI, enhanced vegetation index; MSAVI2, modified soil adjusted vegetation index; NGDI, normalized green difference vegetation index; NDVI, normalized difference vegetation index; OSAVI, optimized soil-adjusted vegetation index; PVI, perpendicular vegetation index; RDVI, renormalized difference vegetation index; SARVI, soil and atmospherically resistant vegetation index; SAVI, soil adjusted vegetation index; TNDVI, transformed normalized difference vegetation index; DVI, difference vegetation index; TCB, tasseled cap brightness; TCG, tasseled cap greenness; TCW, tasseled cap wetness; TCA, tasseled cap angle [= arc tan(TCG/TCB)]; and TCD, tasseled cap distance [$= (TCG^2 + TCB^2)^{0.5}$].

matrix between the AGB and each of the derived remotely sensed variables in each category, and variables with a correlation lower than the median were removed. Second, we constructed all subset models with one, two, and three independent variables in each category and calculated the adjusted coefficient of determination. Third, among all of the models from step 2, we selected the best model based on the adjusted coefficient of determination, and the selected variables in the best models were considered to be the most important variables in each category. Finally, for the selected subset of independent variables, we repeated steps 1 to 3 to determine the most important variables among all categories for the final modeling.

2.5 *Modeling Methods*

We tested nine different modeling methods for estimating AGB, including parametric, semiparametric, and nonparametric approaches. The methods we tested follow.

2.5.1 *Parametric models*

Linear modeling. Among parametric modeling methods, linear regression is the most straightforward. We used ordinary least squares regression to predict AGB as it is commonly used to estimate forest attributes using remotely sensed data.[11]

2.5.2 *Semiparametric*

Generalized additive models. A generalized additive model (GAM) is well-suited to reveal complex curvilinear relationships without a predefined assumption of model shape.[45] The framework is based on a simple approach: (i) the relationships between the explanatory variables and the target variable follow a smooth pattern that can be linear or curvilinear and (ii) the GAM can estimate these smooth relationships simultaneously and then construct predictions by simply adding them up.

2.5.3 *Nonparametric*

Random forest. A random forest (RF) model is a tree-based model that applies a collection of rule-based decisions to evaluate the relationships between a target variable and its explanatory variables.[32,46] It generates a large number of small trees constructed by a different randomly permuted sample from the input dataset.[46,47] The target data are categorized through two offspring at each node split to maximize homogeneity, and the best split is selected. Finally, the target data for each tree are achieved using bootstrap resampling.[47,48] Applying unique tree bagging and selection of a random subset of covariates results in minimization of within group variance and overcoming the over-fitting problem.[47] A comprehensive review of applications of RF in the field of remote sensing can be found in Belgiu and Dragut.[49]

Support vector machine. Support vector machine (SVM) is a statistical learning algorithm, which assumes that each set of explanatory variables has a unique relationship to the target variable, and groupings of predictors can be applied to recognize the rules for predicting a target from a set of explanatory variables.[50] It transforms the input data into a multidimensional hyperplane space using a kernel function to separate groups of input data with similar targets to predict a target variable.[48] Hyperplanes are a multidimensional space constructed from axes that represent each predictor variables. Each response variable can be located in such a space by plotting it according to its explanatory variable values. The SVM applies "support vectors" to assign each target to a well-fragmented space.[32,44,50] The main idea behind SVM is to minimize structural risk and moderate the overfitting problem.[23,51] A comprehensive review of applications of SVM in field of remote sensing can be found in Mountrakis, Im, and Ogole.[50]

k-nearest neighbor. k-nearest neighbor (kNN) is a multivariate approach, which classifies or predicts the response values using the information about the weighted mean of response values of the most similar neighbor(s), where the similarity is identified in a feature space encompassed by selected explanatory variables.[31,51,52] The performance of kNN depends on "k," which determines the number of neighbors used by the method.[31] When $k = 1$, a predicted target value is estimated by the closest training value and when $k > 1$, prediction is estimated with a majority vote. A comprehensive review of applications of kNN in the field of remote sensing can be found in Chirici, Mura, McInerney, Py, Tomppo, Waser, Travaglini, and McRoberts.[52]

Boosted regression trees. Boosted regression trees (BRT) originated from decision theory.[46] It constructs additive regression models by successively fitting the selected base learner to current "pseudo"-residuals by least squares regression at each iteration. BRT utilizes the integration of bagging (for improving the model stability and final predictive accuracy) and boosting (to model the nonlinear relationships, to improve predictive performance of multiple single models, and to reduce the risk of over-fitting).[30,53–56]

Multivariate additive regression splines. Multivariate additive regression splines (MARS) is an automated, multivariate, and adaptive nonparametric classification/regression method introduced by Friedman.[53] It is a flexible technique, which models conventional functions through nonlinear regression models by fitting a weighted sum of multivariate spline basis functions. With this method, it is easier to demonstrate variable interactions and relationships between predictor variables and targets in high-dimensional space[13,57,58] because the original variables, as well as interactions between them, can be found directly in the results.

Cubist regression. Cubist regression (CR) is a hybrid tree-based regression approach, which models the relationships between response and predictor variables based on linear least square regression.[13] It retrieves a set of rules associated with sets of multivariate linear models, then a specific set of explanatory variables will select an actual prediction model based on the rule that best fits the predictors.[59,60] Cubist is a commercial product, and its

algorithm documentation is unknown in comparison to other methods tested in this study.[32,59] Nonetheless, it has been successfully utilized for estimating AGB when using remotely sensed data.[59,61] More details about CR model can be found by Walton.[60]

Gaussian processes regression. Gaussian processes regression (GPR) provides a probabilistic method for learning generic regression problems with kernels.[51] In GPR, there are three components, including explanatory variables applied in the training stage, weight assigned to each variable, and the function assessing the similarity between the predictor dataset from the test data and the overall training samples.[51] It has three major advantages: (i) it provides simultaneous pixel-wise predictions and confidence intervals; (ii) it overcomes the black-box problem that is common to nonparametric regression methods; and (iii) it efficiently and automatically generates model hyper-parameters and model weights by maximizing the marginal likelihood in the training set.[51,62] More details about GPR can be found in Verrelst, Muñoz, Alonso, Delegido, Rivera, Camps-Valls, and Moreno.[62]

2.6 *Statistical Measures for Evaluation of Models Performance*

Given the number of field plots, we used leave-one-out cross validation (LOOCV) to avoid overfitting the model. The model performance was evaluated using four quality statistics:

1. The mean error (Bias) [Eqs. (1)and (2)].
2. The mean absolute error (MAE) [Eq. (3)].
3. The root-mean-square error (RMSE) [Eqs. (4)and (5)].
4. The coefficient of determination (R^2) [Eq. (3)].

$$\text{Bias} = \sum_{i=1}^{n} (\hat{y}_i - y_i)/n, \tag{1}$$

$$\text{Bias}(\%) = \frac{\text{Bias}}{\overline{y}} \times 100, \tag{2}$$

$$\text{MAE} = \sum_{i=1}^{n} |\hat{y}_i - y_i|/n, \tag{3}$$

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)}{n}}, \tag{4}$$

$$\text{RMSE}(\%) = \frac{RMSE}{\overline{y}} \times 100, \tag{5}$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \overline{y})^2}, \tag{6}$$

where y_i is the observed AGB on validation plots, $\hat{y}_i$ is the predicted AGB, n is the number of validation plots, and $\overline{y}$ is the mean observed AGB.

All statistical computing was implemented using open-source R software.[63]

3 Results

3.1 *Field-Measurement of Aboveground Biomass*

The descriptive statistics for the field-measured AGB values, calculated from 124 sample plots (TS) across the two study sites, is summarized in Table 2. The UD site had higher mean and

Table 2 Summary statistics of aboveground biomass (tons/ha) based on field data from the two different study sites.

	Statistical parameters					
Study area	No. sample plots	Mean	Minimum	Maximum	Range	Standard deviation
HD	61	12.6	4.7	20.5	15.8	5.0
UD	63	20.5	0.0	53.0	53.0	11.9
TS	124	16.6	0.0	53.0	53.00	9.9

Note: HD, highly degraded; UD, undegraded; and TS, total surveyed sample plots across two test sites.

Table 3 Modeling assessment using LOOCV for aboveground biomass estimation

Modeling methods	Site	R^2	RMSE (RMSE%)	Bias (Bias%)	MAE
LM	HD	0.47	3.47 (27.99)	0.01 (0.05)	2.77
	UD	0.73	6.07 (29.65)	0.03 (0.15)	5.08
	TS	0.70	5.44 (32.87)	0.01 (0.04)	4.16
GAM	HD	0.49	3.42 (27.63)	0.09 (0.77)	2.71
	UD	0.73	6.10 (29.59)	0.00 (−0.01)	4.89
	TS	0.70	5.44 (32.85)	−0.01 (−0.05)	4.16
RF	HD	0.45	3.56 (28.70)	0.06 (0.46)	2.87
	UD	0.74	5.98 (29.18)	−0.11 (−0.54)	4.57
	TS	0.71	5.32 (32.16)	−0.14 (−0.87)	4.15
SVM	HD	0.48	3.54 (28.53)	−0.67 (−5.40)	2.87
	UD	0.77	5.66 (27.62)	−0.72 (−3.52)	4.47
	TS	**0.73**	**5.21 (31.45)**	**−0.62 (−3.74)**	**3.93**
kNN	HD	0.47	3.50 (28.18)	−0.02 (−0.17)	2.82
	UD	**0.77**	**5.64 (27.49)**	**−0.40 (−1.95)**	**4.46**
	TS	0.70	5.42 (32.71)	−0.08 (−0.50)	4.22
BRT	HD	0.42	3.66 (29.50)	0.04 (0.34)	3.04
	UD	0.74	5.99 (29.26)	0.13 (0.64)	4.58
	TS	0.70	5.41 (32.65)	0.02 (0.12)	4.09
MARS	HD	0.45	3.53 (28.46)	−0.10 (−0.83)	2.91
	UD	0.72	6.24 (30.45)	−0.05 (−0.24)	4.81
	TS	0.70	5.51 (33.27)	0.09 (0.55)	4.14
CR	HD	**0.56**	**3.17 (25.60)**	**0.03 (0.25)**	**2.52**
	UD	0.74	6.07 (29.44)	−0.88 (−4.26)	4.95
	TS	**0.73**	**5.19 (31.32)**	**−0.42 (−2.56)**	**3.93**
GPR	HD	0.44	3.59 (28.88)	0.02 (0.19)	2.93
	UD	0.69	6.60 (32.17)	−0.22 (−1.11)	5.06
	TS	0.71	5.38 (32.46)	−0.02 (−015)	4.13

Note: HD, highly degraded; UD, un-degraded; and TS, total surveyed sample plots across two test sites; best results are shown in bold format.

standard deviation (20.5 and 11.8 ton/ha, respectively) than the HD site (12.6 and 5.0 ton/ha, respectively). The mean AGB at the HD site was 39% lower than at the UD site.

3.2 *Comparative Analysis of the Models*

3.2.1 *Comparative analysis of modeling methods by LOOCV*

A comparative assessment of the nine tested modeling methods revealed differences among these methods in terms of the accuracy of AGB estimates.

Evaluation of the level of agreement between observed and predicted values (R^2) showed that for each of the modeling methods tested, the UD (0.69 to 0.77) and TS sites (0.70 to 0.73) had higher R^2 values compared to the HD site (0.42 to 0.56). (Table 3, Figs. 2–4). For the HD site, the CR method ($R^2 = 0.56$; RMSE% = 25.60% of the mean; Bias% = 0.25% of the mean; and MAE% = 20% of the mean) slightly outperformed all of the other methods (Table 3, Fig. 2). For the UD site, the R^2, RMSE%, and MAE% of the kNN method were slightly better than the other methods ($R^2 = 0.77$; RMSE% = 27.49% of the mean, and MAE = 21.75% of the mean, respectively). However, the CR yielded Bias and Bias% higher than other methods at the

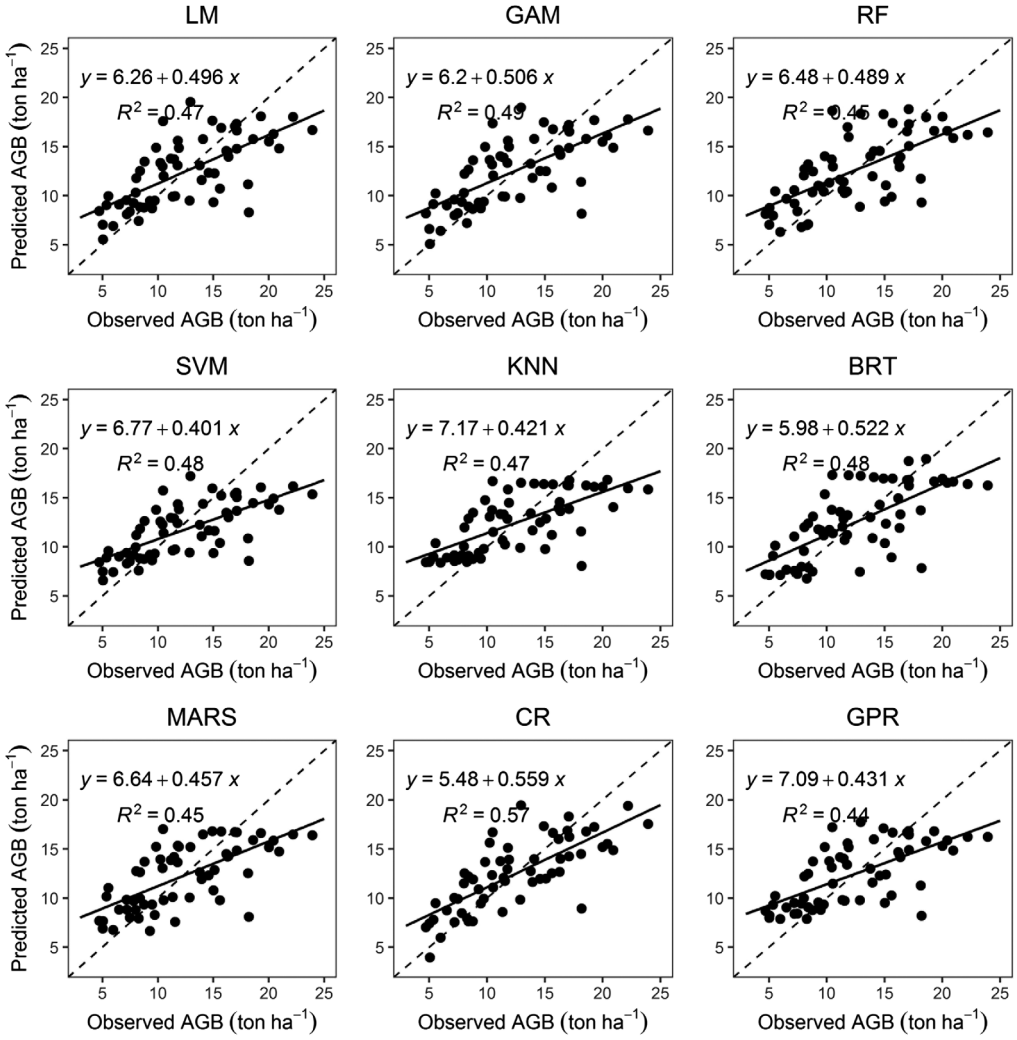

Fig. 2 Observed AGB versus predicted AGB by LM, GAM, RF, SVM, KNN, BRT, MARS, CR, and GPR via LOOCV for the HD site. The solid line indicates the optimal regression of observed versus predicted AGB (the corresponding equation and coefficient of correlation are also represented) and the dashed line indicates the 1:1 line of perfect agreement.

HD site (Table 3, Fig. 3). Across all plots combined (TS), both the SVM and CR methods resulted in R^2 values of 0.73; RMSE of ~5.2 ton/ha (~31% of the mean); and MAE of 3.93 ton/ha (23.67% of the mean) (Table 3, Fig. 4).

Bias was only slightly different among methods, typically below 2% for each test site (which equates to ~0.25 and 0.41 ton/ha for HD and UD, respectively). SVM and CR, however, did yield more biased results (Table 3). The SVM CR models tended to under-estimate AGB by more than 3.5% for some sites (Table 3). In most cases, however, the modeling methods yielded unbiased predictions of AGB. (Table 3).

3.3 *Predictor Variables*

The results of the study indicated that the extracted Landsat spectral variables had different utility in each of the modeling methods for AGB estimation across sample sites. Overall, the selected variables included simple band ratios, raw bands, linear transformation (i.e., TC and PCA), and vegetation indices (Table 4).

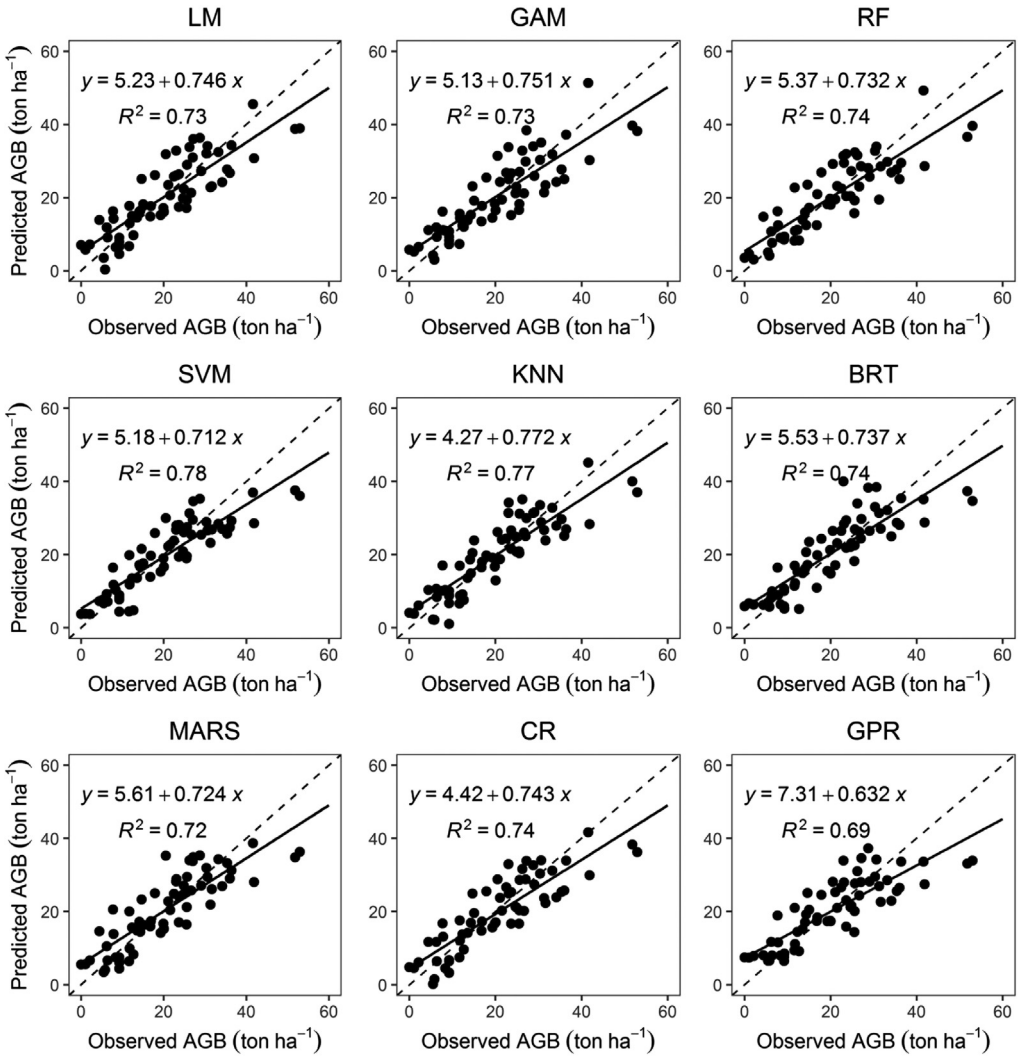

Fig. 3 Observed AGB versus predicted AGB by LM, GAM, RF, SVM, KNN, BRT, MARS, CR, and GPR via LOOCV for the UD site. The solid line indicates the optimal regression of observed versus predicted AGB (the corresponding equation and coefficient of correlation are also represented) and the dashed line indicates the 1:1 line of perfect agreement.

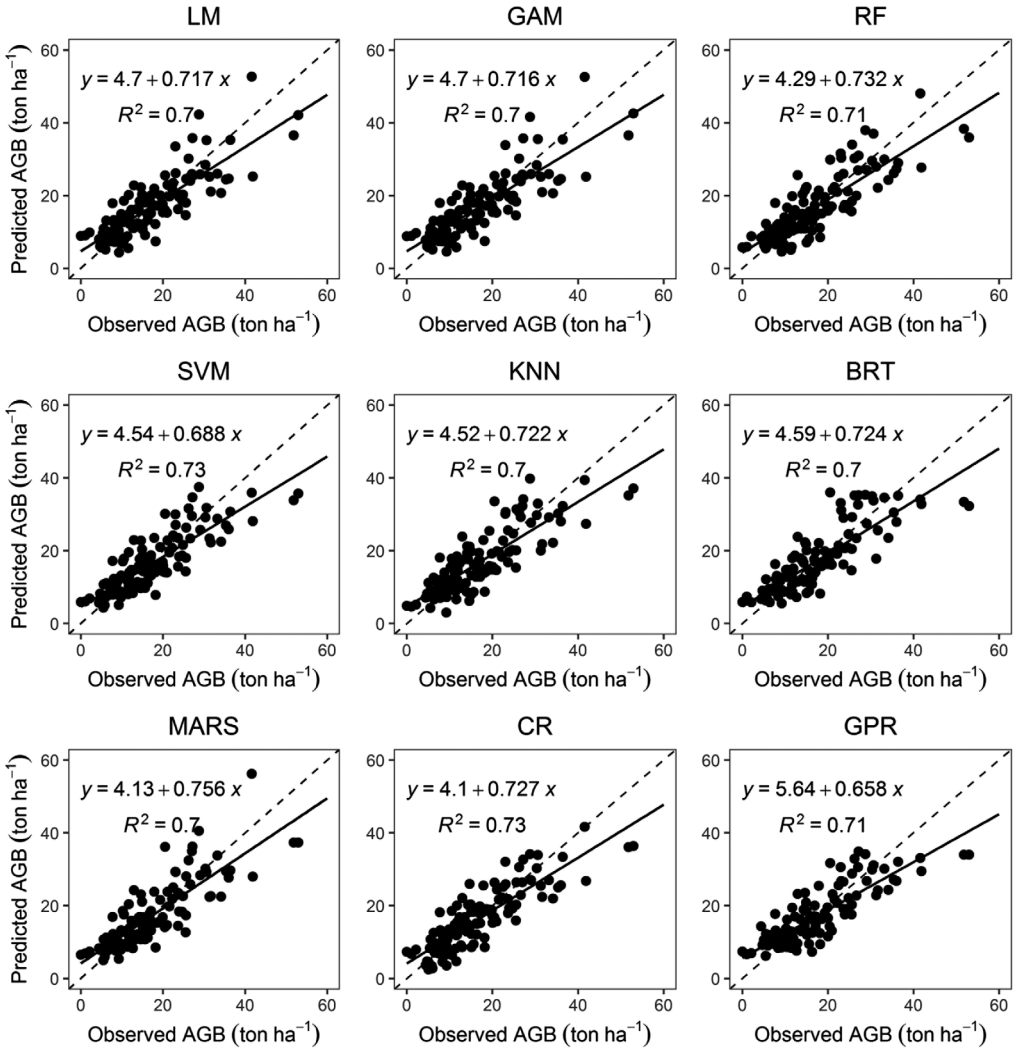

Fig. 4 Observed AGB versus predicted AGB by LM, GAM, RF, SVM, KNN, BRT, MARS, CR, and GPR via LOOCV for the TS site. The solid line indicates the optimal regression of observed versus predicted AGB (the corresponding equation and coefficient of correlation are also represented) and the dashed line indicates the 1:1 line of perfect agreement.

4 Discussion

Accurate and timely estimation of forest biomass has increasingly been highlighted as critical for improved forest ecosystem management, considering the role of forests in the global carbon cycle. AGB estimation using remotely sensed predictor variables faces several uncertainties, and its accuracy is affected by the choice of modeling method, the selection of predictor variables, and the inherent forest stand characteristics.[2,8] In this paper, we assessed how these factors influence AGB estimates in Zagros coppice oak forests. For each combination of factors, we calculated R^2, RMSE, Bias, and MAE diagnostics using LOOCV.

Although it is difficult to compare our results to other studies with different forest conditions, sampling designs, and modeling methods, the highest R^2 and lowest RMSE% results from our study (UD $\sim$ 0.77 and 27.5%, respectively; TS $\sim$ 0.73% and 31.1%, respectively) are generally consistent with results from other Landsat-based forest AGB estimates.[10,11,14,24,28,29,33,41,64] The accuracy of the results in our study is likely attributable, in part, to the low-biomass density at our study sites. In most studies of AGB estimation based on passive optical remote sensing, data saturation has been reported as a major challenge or source of uncertainty in forests with closed canopies and complex stand structure.[11,12,22,23,28,35] The relatively low biomass in Zagros forests

Table 4 Predictor variables selected for each modeling method.

Modeling method	Site		
	HD	UD	TS
LM	N/R, R	S1/S2, MSAVI2	N/R, GREENESS
GAM	R, N	G/S1, N/S2	N/R, S1/S2
RF	N/R, R	S1/S2, MSAVI2	N/R, GREENESS
SVM	N/R, R	S1/S2, MSAVI2	N/R, GREENESS
kNN	R, NDVIC	TCA, GREENESS	N/R, G/N, S1/S2
BRT	N/R, NDVI	TCA, S1/S2	N/R, G/N, S1/S2
MARS	N/R, B/R	TCA, S1/S2	N/R, S1/S2
CR	G, DVI, GREENESS	S1, S2, GREENESS	N, R, B
GPR	N/R, R, NDVIC	TCA, N/R	PC5, GREENESS, S1/S2

Note: HD, highly degraded; UD, undegraded; and TS, total surveyed sample plots across two test sites

resulted in fairly low-reflectance values recorded for the NIR band, therefore, data saturation was not an issue.[28]

Another potential reason for the high accuracy of AGB estimates in our study might be related to the allometric equations that we used to calculate tree-level AGB. Previous studies have frequently used allometric functions for tree biomass based on characteristics such as the diameter of the trunk or the height of the tree, but here we used allometric equations that were based on crown canopy diameter. Optical-sensors such as Landsat 8 OLI record the spectral reflectance of properties related to canopy cover, therefore, these datasets are efficient for horizontal vegetation structure such as crown canopy rather than vertical vegetation structure (e.g., diameter or height of tree).[1,2,13,23]

4.1 *Predictive Performance of the Modeling Methods*

The importance of modeling methods in remote sensing-based forest biomass estimates is a complicated issue. Some studies have revealed that the method of prediction is not an issue in AGB estimation,[2,11,32,44] whereas other researchers have reported that the modeling method plays a key role in remote sensing-based forest ecosystem biomass estimates.[8,13,25]

In this study, various modeling methods were used for remote sensing-based forest biomass estimation, with no clear conclusion on what algorithms perform best. The accuracy estimates of these methods depend upon the forest stand characteristics, the sample size, and the method of accuracy evaluation.[11] Here we investigated the effect of field sites (two cases) and validation metric (four cases) on the accuracy of Landsat-based AGB prediction by nine different modeling methods.

The results demonstrate that it is difficult to reach a solid conclusion as to which modeling method is the most accurate. For instance, for the HD site, in terms of R^2, RMSE, and MAE, the results of LM, GAM, SVM, CR, and GPR methods were only slightly different. However, SVM yielded the most biased prediction (−5.40% of the mean). We also documented only slightly different results for the UD site for LM, RF, SVM, kNN, and CR. In terms of R^2, RMSD, and MAE, all of these methods except GPR yielded approximately the same values. All in all, considering the bias at the two test sites (and their combination) by LOOCV, the SVM and CR methods were markedly poorer than the other tested models. This phenomenon can be explained by the fact that SVM is strongly affected by parameterization, and it can yield the best results only if its parameters are optimal.[31]

LM yielded acceptable results compared to semi- and nonparametric modeling methods, which can be explained by the fact that the relationships between Landsat-derived predictors

and observed AGB are likely linear, and therefore, well modeled by LM. It might be a result of canopy cover-based allometric equations that were used and also the fact that AGB did not reach the spectral saturation point. Spectral reflectance is sensitive to forest crown canopies, therefore, it can better predict canopy characteristics.[13,23] The highest AGB values measured in sample plots (HD and UD sites: 20.49 and 52.95 ton/ha, respectively) are lower than saturation values, such as 150 ton/ha, reported elsewhere.[22,35]

Our findings illustrate that the model performance depends on both field data and modeling methods. Therefore, because of the low-biomass density in the HD site and the absence of spectral saturation, AGB can be well-modeled linearly by spectral variables. However, for the HD site, CR was the best prediction method ($R^2 = 0.56$, RMSE% = 25.60%, and Bias% = 0.25%). Gleason and Im[32] and Güneralpet al.[13] reported weak performance of CR methods, whereas Blackard et al.[65] and Chenet al.[66] selected CR as the best method for biomass modeling. However, the modeling method is best selected based on the research goal. If the precision of the prediction is the most important factor, then SVM and CR methods should be considered; in contrast, if the degree of under- or over-prediction of the modeling method is the most important factor, then other methods such as BRT, RF, kNN, and GAM should be considered.

4.2 *Predictor Variables*

It is essential to understand how remote sensing data are related to AGB. Different spectral data sets are available for AGB estimation from remote sensing data such as original raw bands, simple band ratios, vegetation indices, and linear transformations. In this study, we used a large diversity of spectral information. Several researchers have shown that the original raw bands and vegetation indices are efficient for remote sensing-based biomass estimates.[11,22,28,30,42,67] Our findings demonstrate that it is not possible to select specific variables or variable groups as the best explanatory variables.

Among explanatory variables that we tested here, original raw bands and their simple ratio combinations (e.g., red/near-infrared and shortwave-infrared 1/shortwave-infrared2) were commonly selected by different models. This is in agreement with other studies in dryland and low-biomass forests.[7,14,22,28,68] We rarely used NDVI and soil adjusted indices as the final variables in the models. Sensitivity of NDVI to AGB is reduced by background soil reflectance and shadow in low-biomass forests.[7,64] Zandler et al.[15] and Gasparriet al.[28] found that soil adjusted indices were not useful in AGB estimates because of the complexity of building a generalized soil line. Many studies have identified shortwave-infrared bands as being critical variables for AGB estimates for different forest types.[7,14,69,70]

4.3 *Effect of Human-Activities on Aboveground Biomass Estimate Accuracies*

Fragile ecosystems, such as Zagros forests, which are affected by traditional human activities, require continuous monitoring. Since human activities create different forest structures and composition, different spectral data sets and modeling methods are needed. As this study demonstrates, the accuracy of AGB estimates for the HD site was lower than for the UD site as well as for all sites combined. The main reasons for this are the low canopy cover of degraded stands and, therefore, a high likelihood of error for biomass calculations based on these field data. The allometric equations used in this work were derived based on the assumption that trees have symmetric crowns, but because of traditional rural harvesting practices such as pollarding of trees at the HD site, this assumption was not always met, which resulted in higher error for biomass calculation, as well as weaker relationships between tree crown parameters and biomass. The low canopy cover at the HD site leads to mixed pixels containing background reflectance from vegetation, soil and understory vegetation.[14,15,35]

5 Conclusion

Accurate estimation of AGB at broad scales is critical for understanding forests' contribution to the carbon cycle and the terrestrial carbon budget and for improving decision making processes.

Constructing an accurate and cost-efficient model for AGB estimates in low-biomass forests is an important task to meet these needs. This study explored a comparative analysis of nine modeling methods and evaluation methods for AGB estimation in Zagros forests under different levels of human activity. An extensive range of spectral information was derived from Landsat 8 OLI imagery and used as input to parametric, semiparametric, and nonparametric models. The following conclusions were drawn:

- It is possible to estimate AGB using Landsat 8 OLI data with relatively high accuracy.
- The accuracy of AGB estimates related to site conditions based on human disturbance levels.
- If the research goal is to minimize prediction error, then SVM and CR should be considered; if the goal is to minimize bias, then LM, kNN, BRT, GAM, and RF should be considered.
- The preferred models were derived based on different sets of predictor variables; therefore, it was not possible to select the best variable or group of variables for all cases.

Acknowledgments

The authors wish to thank Dr. James Halperin for valuable comments on the manuscript. Also, they appreciate help from Dr. Shaban Shataee and Dr. Jalil Alvai. Sincerely, they are grateful to the two anonymous reviewers for the constructive criticism, which improved an earlier version of this paper. No potential conflict of interest was reported by the authors. This work is a part of A. Safari's PhD thesis supported by Tarbiat Modares University, Iran.

References

1. O. Fernández-Manso, A. Fernández-Manso, and C. Quintano, "Estimation of aboveground biomass in Mediterranean forests by statistical modelling of ASTER fraction images," *Int. J. Appl. Earth Obs. Geoinf.* **31**, 45–56 (2014).
2. X. Wang et al., "An application of remote sensing data in mapping landscape-level forest biomass for monitoring the effectiveness of forest policies in northeastern China," *Environ. Manage.* **52**(3), 612–620 (2013).
3. J. Carreiras, J. B. Melo, and M. J. Vasconcelos, "Estimating the above-ground biomass in miombo savanna woodlands (Mozambique, East Africa) using L-band synthetic aperture radar data," *Remote Sens.* **5**(4), 1524–1548 (2013).
4. D.-A. Kwak et al., "Estimating stem volume and biomass of Pinus koraiensis using LiDAR data," *J. Plant Res.* **123**(4), 421–432 (2010).
5. M. A. Stelmaszczuk-Górska et al., "Non-parametric retrieval of aboveground biomass in Siberian Boreal Forests with ALOS PALSAR interferometric coherence and backscatter intensity," *J. Imaging* **2**(1), 1 (2016).
6. M. González-Roglich and J. J. Swenson, "Tree cover and carbon mapping of Argentine savannas: scaling from field to region," *Remote Sens. Environ.* **172**, 139–147 (2016).
7. C. Eisfelder, C. Kuenzer, and S. Dech, "Derivation of biomass information for semi-arid areas using remote-sensing data," *Int. J. Remote Sens.* **33**(9), 2937–2984 (2012).
8. F. E. Fassnacht et al., "Importance of sample size, data type and prediction method for remote sensing-based estimations of aboveground forest biomass," *Remote Sens. Environ.* **154**, 102–114 (2014).
9. N. R. Jachowski et al., "Mangrove biomass estimation in Southwest Thailand using machine learning," *Appl. Geogr.* **45**, 311–321 (2013).
10. V. Avitabile et al., "Capabilities and limitations of Landsat and land cover data for aboveground woody biomass estimation of Uganda," *Remote Sens. Environ.* **117**, 366–380 (2012).
11. X. Zhu and D. Liu, "Improving forest aboveground biomass estimation using seasonal Landsat NDVI time-series," *ISPRS J. Photogramm. Remote Sens.* **102**, 222–231 (2015).

12. R. M. Lucas, A. L. Mitchell, and J. Armston, "Measurement of forest above-ground biomass using active and passive remote sensing at large (subnational to global) scales," *Curr. For. Rep.* **1**(3), 162–177 (2015).
13. I. Güneralp, A. M. Filippi, and J. Randall, "Estimation of floodplain aboveground biomass using multispectral remote sensing and nonparametric modeling," *Int. J. Appl. Earth Obs. Geoinf.* **33**, 119–126 (2014).
14. M. Karlson et al., "Mapping tree canopy cover and aboveground biomass in Sudano-Sahelian woodlands using Landsat 8 and random forest," *Remote Sens.* **7**(8), 10017–10041 (2015).
15. H. Zandler, A. Brenning, and C. Samimi, "Quantifying dwarf shrub biomass in an arid environment: comparing empirical methods in a high dimensional setting," *Remote Sens. Environ.* **158**, 140–155 (2015).
16. N. F. Glenn et al., "Landsat 8 and ICESat-2: performance and potential synergies for quantifying dryland ecosystem vegetation cover and biomass," *Remote Sens. Environ.* **185**, 233–242 (2016).
17. H. Abd-El Monsef and S. E. Smith, "A new approach for estimating mangrove canopy cover using Landsat 8 imagery," *Comput. Electron. Agric.* **135**, 183–194 (2017).
18. A. Henareh Khalyani et al., "Deforestation and landscape structure changes related to socioeconomic dynamics and climate change in Zagros forests," *J. Land Use Sci.* **8**(3), 321–340 (2013).
19. K. Sagheb-Talebi, M. Pourhashemi, and T. Sajedi, *Forests of Iran*, Springer Science, London (2014).
20. K. Majumda and B. K. Datta, "Effects of anthropogenic disturbances on vegetation diversity and structure: a case study in the remnant forests surrounding the village ecosystems of Tripura, Northeast India," *Chin. J. Popul. Res. Environ.* **4**, 332–340 (2015).
21. A. Valipour et al., "Traditional silvopastoral management and its effects on forest stand structure in northern Zagros, Iran," *For. Ecol. Manage.* **327**, 221–230 (2014).
22. B. Gizachew et al., "Mapping and estimating the total living biomass and carbon in low-biomass woodlands using Landsat 8 CDR data," *Carbon Balance Manage.* **11**(1), 13 (2016).
23. D. Lu et al., "A survey of remote sensing-based aboveground biomass estimation methods in forest ecosystems," *Int. J. Digital Earth* **9**(1), 63–105 (2016).
24. R. J. Frazier et al., "Characterization of aboveground biomass in an unmanaged boreal forest using Landsat temporal segmentation metrics," *ISPRS J. Photogramm. Remote Sens.* **92**, 137–146 (2014).
25. Z. Shao and L. Zhang, "Estimating forest aboveground biomass by combining optical and SAR data: a case study in Genhe, Inner Mongolia, China," *Sensors* **16**(6), 834 (2016).
26. M. Alrababah et al., "Estimating east Mediterranean forest parameters using Landsat ETM," *Int. J. Remote Sens.* **32**(6), 1561–1574 (2011).
27. D. Pflugmacher et al., "Using landsat-derived disturbance and recovery history and lidar to map forest biomass dynamics," *Remote Sens. Environ.* **151**, 124–137 (2014).
28. N. I. Gasparri et al., "Assessing multi-temporal Landsat 7 ETM+ images for estimating above-ground biomass in subtropical dry forests of Argentina," *J. Arid. Environ.* **74**(10), 1262–1270 (2010).
29. S. L. Powell et al., "Quantification of live aboveground forest biomass dynamics with Landsat time-series and field inventory data: a comparison of empirical modeling approaches," *Remote Sens. Environ.* **114**(5), 1053–1068 (2010).
30. T. Dube and O. Mutanga, "Evaluating the utility of the medium-spatial resolution Landsat 8 multispectral sensor in quantifying aboveground biomass in uMgeni catchment, South Africa," *ISPRS J. Photogramm. Remote Sens.* **101**, 36–46 (2015).
31. P. M. López-Serrano et al., "A comparison of machine learning techniques applied to landsat-5 TM spectral data for biomass estimation," *Can. J. Remote Sens.* **42**(6), 690–705 (2016).
32. C. J. Gleason and J. Im, "Forest biomass estimation from airborne LiDAR data using machine learning approaches," *Remote Sens. Environ.* **125**, 80–91 (2012).

33. J. Zhu et al., "Mapping forest ecosystem biomass density for Xiangjiang river basin by combining plot and remote sensing data and comparing spatial extrapolation methods," *Remote Sens.* **9**(3), 241 (2017).
34. D. Gagliasso, S. Hummel, and H. Temesgen, "A comparison of selected parametric and non-parametric imputation methods for estimating forest biomass and basal area," *Open J. For.* **4**(1), 42–48 (2014).
35. Y. Feng et al., "Examining effective use of data sources and modeling algorithms for improving biomass estimation in a moist tropical forest of the Brazilian Amazon," *Int. J. Digital Earth* **10**, 996–1016 (2017).
36. A. Nolè et al., "Application of the 3-PGS model to assess carbon accumulation in forest ecosystems at a regional level," *Can. J. For. Res.* **39**(9), 1647–1661 (2009).
37. S. Eckert, "Improved forest biomass and carbon estimations using texture measures from WorldView-2 satellite data," *Remote Sens.* **4**(4), 810–829 (2012).
38. Y. Iranmanesh, "Assessment on biomass estimation methods and carbon sequestration of Quercus brantii Lindl. in Chaharmahal & Bakhtiari Forests, Tarbiat Modares University," PhD Thesis, Tarbiat Modares University, p. 113 (2013).
39. R. Richter, T. Kellenberger, and H. Kaufmann, "Comparison of topographic correction methods," *Remote Sens.* **1**(3), 184–196 (2009).
40. D. Riaño et al., "Assessment of different topographic corrections in Landsat-TM data for mapping vegetation types (2003)," *IEEE Trans. Geosci. Remote Sens.* **41**(5), 1056–1061 (2003).
41. S. Labrecque et al., "A comparison of four methods to map biomass from Landsat-TM and inventory data in western Newfoundland," *For. Ecol. Manage.* **226**(1), 129–144 (2006).
42. A. A. Wani, P. K. Joshi, and O. Singh, "Estimating biomass and carbon mitigation of temperate coniferous forests using spectral modeling and field inventory data," *Ecol. Inf.* **25**, 63–70 (2015).
43. C. Gómez et al., "Historical forest biomass dynamics modelled with Landsat spectral trajectories," *ISPRS J. Photogramm. Remote Sens.* **93**, 14–28 (2014).
44. E. B. Görgens, A. Montaghi, and L. C. E. Rodriguez, "A performance comparison of machine learning methods to estimate the fast-growing forest plantation yield based on laser scanning metrics," *Comput. Electron. Agric.* **116**, 221–227 (2015).
45. S. Wood, *Generalized Additive Models: An Introduction with R. 2006*, Chapman and Hall/CRC Press, Chicago (2006).
46. L. Breiman, "Random forests," *Mach. Learn.* **45**(1), 5–32 (2001).
47. S. Park et al., "Drought assessment and monitoring through blending of multi-sensor indices using machine learning approaches for different climate regions," *Agric. For. Meteorol.* **216**, 157–169 (2016).
48. K. Were et al., "A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape," *Ecol. Indic.* **52**, 394–403 (2015).
49. M. Belgiu and L. Dragut, "Random forest in remote sensing: a review of applications and future directions," *ISPRS J. Photogramm. Remote Sens.* **114**, 24–31 (2016).
50. G. Mountrakis, J. Im, and C. Ogole, "Support vector machines in remote sensing: a review," *ISPRS J. Photogramm. Remote Sens.* **66**(3), 247–259 (2011).
51. H. Latifi et al., "Stratified aboveground forest biomass estimation by remote sensing data," *Int. J. Appl. Earth Obs. Geoinf.* **38**, 229–241 (2015).
52. G. Chirici et al., "A meta-analysis and review of the literature on the k-nearest neighbors technique for forestry applications that use remotely sensed data," *Remote Sens. Environ.* **176**, 282–294 (2016).
53. J. H. Friedman "Multivariate adaptive regression splines," *Ann. Stat.* **19**, 1–67 (1991).
54. J. Elith, J. R. Leathwick, and T. Hastie, "A working guide to boosted regression trees," *J. Animal Ecol.* **77**(4), 802–813 (2008).
55. J. Leathwick et al., "Dispersal, disturbance and the contrasting biogeographies of New Zealand's diadromous and non-diadromous fish species," *J. Biogeogr.* **35**(8), 1481–1497 (2008).

56. W. Zhang et al., "Boosted regression tree model-based assessment of the impacts of meteorological drivers of hand, foot and mouth disease in Guangdong, China," *Sci. Total Environ.* **553**, 366–371 (2016).
57. R. Put et al., "Multivariate adaptive regression splines (MARS) in chromatographic quantitative structure-retention relationship studies," *J. Chromatogr. A* **1055**(1), 11–19 (2004).
58. A. M. Filippi, İ. Güneralp, and J. Randall, "Hyperspectral remote sensing of aboveground biomass on a river meander bend using multivariate adaptive regression splines and stochastic gradient boosting," *Remote Sens. Lett.* **5**(5), 432–441 (2014).
59. P. T. Noi, J. Degener, and M. Kappas, "Comparison of multiple linear regression, cubist regression, and random forest algorithms to estimate daily air surface temperature from dynamic combinations of MODIS LST data," *Remote Sens.* **9**(5), 398 (2017).
60. J. T. Walton, "Subpixel urban land cover estimation," *Photogramm. Eng. Remote Sens.* **74**(10), 1213–1222 (2008).
61. M. Li et al., "Forest biomass and carbon stock quantification using airborne LiDAR data: a case study over Huntington Wildlife Forest in the Adirondack Park," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **7**(7), 3143–3156 (2014).
62. J. Verrelst et al., "Machine learning regression algorithms for biophysical parameter retrieval: opportunities for Sentinel-2 and-3," *Remote Sens. Environ.* **118**, 127–139 (2012).
63. R Core Team, "R: a language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, 2015, www.R-project.org (2016).
64. H. Du et al., "The responses of Moso bamboo (Phyllostachys heterocycla var. pubescens) forest aboveground biomass to Landsat TM spectral reflectance and NDVI," *Acta Ecol. Sin.* **30**(5), 257–263 (2010).
65. J. Blackard et al., "Mapping US forest biomass using nationwide forest inventory data and moderate resolution information," *Remote Sens. Environ.* **112**(4), 1658–1677 (2008).
66. X. Chen et al., "Estimating aboveground forest biomass carbon and fire consumption in the US Utah High Plateaus using data from the Forest Inventory and Analysis Program, Landsat, and LANDFIRE," *Ecol. Indic.* **11**(1), 140–148 (2011).
67. A. Günlü et al., "Estimating aboveground biomass using Landsat TM imagery: a case study of Anatolian Crimean pine forests in Turkey," *Ann. For. Res.* **57**(2), 289–298 (2014).
68. G. P. Asner, "Biophysical and biochemical sources of variability in canopy reflectance," *Remote Sens. Environ.* **64**(3), 234–253 (1998).
69. A. Baccini et al., "A first map of tropical Africa's above-ground biomass derived from satellite imagery," *Environ. Res. Lett.* **3**(4), 045011 (2008).
70. M. Steininger, "Satellite estimation of tropical secondary forest above-ground biomass: data from Brazil and Bolivia," *Int. J. Remote Sens.* **21**(6–7), 1139–1157 (2000).

Amir Safari received his BS and MS degrees in forest science from the University of Guilan and the University of Kurdistan in 2005 and 2009, respectively, and his PhD in forestry from Tarbiat Modares University, Iran, in 2018. His current research interests include the use of remote sensing and statistical modeling methods for quantifying and monitoring the structural parameters of forest stands, in particular carbon stocks.

Hormoz Sohrabi received his BA degree in forest science form Kurdistan University, his MSc degree in forest science (statistical ecology), and his PhD in forest biometrics and remote sensing from Tarbiat Modares University. He is currently an associate professor in the Department of Forest Science at Tarbiat Modares University. He is the author of more than 40 journal papers and has written a book. His research focus is stand and tree level biomass estimation using satellite and UAV images.

Scott Powell received his BA degree in biology from Macalester College, his MEM degree in resource ecology from Duke University, and his PhD in ecology from Montana State University. He is currently an assistant professor in the Department of Land Resources and Environmental Sciences at Montana State University. His research focuses on the use of geospatial data to characterize ecosystem and landscape processes, including quantification of land use and carbon sequestration at broad spatial and temporal scales.