Comparison of machine-learning methods for above-ground biomass estimation based on Landsat imagery

Article *in* Journal of Applied Remote Sensing · August 2016

DOI: 10.1117/1.JRS.10.035010

8 authors, including:

Chaofan Wu

13 PUBLICATIONS **204** CITATIONS

SEE PROFILE

Muye Gan

Zhejiang University

45 PUBLICATIONS **985** CITATIONS

SEE PROFILE

Jinxia Zhu

Zhejiang University of Finance and Economics

17 PUBLICATIONS **270** CITATIONS

SEE PROFILE

Journal of
Applied Remote Sensing

RemoteSensing.SPIEDigitalLibrary.org

Comparison of machine-learning methods for above-ground biomass estimation based on Landsat imagery

Chaofan Wu
Huanhuan Shen
Aihua Shen
Jinsong Deng
Muye Gan
Jinxia Zhu
Hongwei Xu
Ke Wang

SPIE.

Chaofan Wu, Huanhuan Shen, Aihua Shen, Jinsong Deng, Muye Gan, Jinxia Zhu, Hongwei Xu, Ke Wang, "Comparison of machine-learning methods for above-ground biomass estimation based on Landsat imagery," *J. Appl. Remote Sens.* **10**(3), 035010 (2016), doi: 10.1117/1.JRS.10.035010.

Comparison of machine-learning methods for above-ground biomass estimation based on Landsat imagery

Chaofan Wu,[a] Huanhuan Shen,[a] Aihua Shen,[b] Jinsong Deng,[a] Muye Gan,[a] Jinxia Zhu,[c] Hongwei Xu,[a] and Ke Wang[a,*]

[a]Zhejiang University, Institute of Applied Remote Sensing and Information Technology, Hangzhou 310058, Zhejiang, China

[b]Zhejiang Forestry Academy, Hangzhou, Zhejiang Province 310023, China

[c]Zhejiang University of Finance and Economics, Institute of Economic and Social Development, Hangzhou 310018, China

Abstract. Biomass is one significant biophysical parameter of a forest ecosystem, and accurate biomass estimation on the regional scale provides important information for carbon-cycle investigation and sustainable forest management. In this study, Landsat satellite imagery data combined with field-based measurements were integrated through comparisons of five regression approaches [stepwise linear regression, *K*-nearest neighbor, support vector regression, random forest (RF), and stochastic gradient boosting] with two different candidate variable strategies to implement the optimal spatial above-ground biomass (AGB) estimation. The results suggested that RF algorithm exhibited the best performance by 10-fold cross-validation with respect to R^2 (0.63) and root-mean-square error (26.44 ton/ha). Consequently, the map of estimated AGB was generated with a mean value of 89.34 ton/ha in northwestern Zhejiang Province, China, with a similar pattern to the distribution mode of local forest species. This research indicates that machine-learning approaches associated with Landsat imagery provide an economical way for biomass estimation. Moreover, ensemble methods using all candidate variables, especially for Landsat images, provide an alternative for regional biomass simulation. [DOI: 10.1117/1.JRS.10.035010]

Keywords: forest above-ground biomass; Landsat imagery; nonparametric approaches; spatial prediction.

Paper 16091 received Jan. 29, 2016; accepted for publication Jul. 21, 2016; published online Aug. 8, 2016.

1 Introduction

Since the first Earth Resources Technology Satellite, later named Landsat 1, was launched in 1972, remote-sensing imageries have been widely used in forest monitoring and management. Recently, applications have been changed from forest spatial extent extraction to more detailed forest category classification and biophysical property investigations.[1] With the rising awareness of valuable ecological service of forest ecosystem, human beings have realized that biomass, one of the most primary parameters, is not only closely related to timber productivity but also immediately interrelated to regional and even global carbon storage and cycle.[2] To emphasize the importance of forest biomass in global carbon balance and put forward a better understanding of carbon accounting, international mechanisms such as reducing emissions from deforestation and forest degradation (REDD) and REDD+[3] have been developed. Research on biomass estimation is growing, employing remote-sensing technology combined with field investigations to realize spatiotemporal biomass simulations. Traditional field-based measurements can provide the most accurate biomass values, but they are destructive and spatially limited. Compared to

*Address all correspondence to: Ke Wang, E-mail: kwang@zju.edu.cn

1931-3195/2016/$25.00

their difficult and time-consuming processes, remote-sensing techniques provide more efficient, macroscopical, and repeatable information for forest information retrieval, which expands the spatial coverage from a traditional quadrat scale to a regional scale. Therefore, it has become more comprehensive and practicable to unite these two approaches in the past few decades.[4]

The principle for remote-sensing-based biomass estimation is that the canopy reflectance of vegetation is "causally" related to the leaf area index of the canopy and covaries with above-ground biomass (AGB).[5] Yet there exists great uncertainty during the prediction process. Especially for optical remote-sensing imagery, no direct relationship has been developed that is universally applicable.[6] Consequently, empirical relationships between biomass and remote-sensing spectral bands have been explored by many researchers, and significant relationships have been found to promote remote-sensing-based biomass inversion.[7] Candidate variables that can be derived from satellite imageries includes primary multispectral bands, vegetation indices, texture information, multitemporal spectral information, and increasing syncretic features from different sensors, which are gradually integrated to establish unambiguous or hidden relationships for biomass prediction.[4] Du et al.[8] used moderate-resolution imaging spectroradiometer (MODIS) land-cover product (with a spatial resolution of about 5600 m) integrated with forest inventory data to produce the biomass map in 31 provinces in China. Compared with MODIS, Landsat imagery has a higher spatial resolution and exhibits more detailed information on spatial biomass distribution. Although multiple data including high spatial resolution images, hyperspectral data, light detection and ranging (LiDAR), and synthetic aperture radar data have been incorporated into biomass estimation, their preprocessing process is usually complicated and further popularization is always subject to the relatively limited availability and coverage. In contrast, the Landsat program has the advantage of appropriate spatial resolution suitable for forest biophysical parameter inversion, and its open access as well as long historic archives make great contributions to more practical biomass simulation applications.[9,10]

At present, improved mathematical models have been increasingly developed for spatially explicit biomass quantification through building theoretical relationships between remote-sensing variables and biomass.[11,12] A wide variety of algorithms, particularly machine-learning algorithms, have been introduced for establishing empirical relationships, especially with the prevalence of open-access software (R and so on). Contemporary approaches including artificial neural networks, classification and regression trees, *K*-nearest neighbor (KNN), support vector regression (SVR), and random forest (RF) are widely employed both for classification and regression because they make no assumptions about initial data and relationships between response variable and predictive variables which conforms to the complexity of forest ecosystem, especially when compared with traditional empirical models like linear regression that have a requirement of Gaussian distribution for the input variables.[13–15] In addition to this promising advantage, higher accuracy and delectable convenience have appealed to more researchers despite their indistinct implementation process that is difficult to interpret.[16,17]

Notwithstanding, successive progress has been made on biomass estimation and further attempts should be made to explore the most suitable approach for a specific study area and forest ecosystem. Foody et al.[15] investigated the transferability of predictive relations between estimated biomass and Landsat Thematic Mapper (TM) data for three sites in tropical forest and they found that the relations between sites were site-dependent, indicating that biomass estimation using statistical models is challenging because of environmental variation.[18] The objective of this study is to investigate and compare the effectiveness of four popular machine-learning approaches, KNN, SVR, RF, and stochastic gradient boosting (SGB), compared with traditional stepwise linear regression (Steplm) through two feature strategies using Landsat TM image and field-based measurements for spatial biomass estimation in northwestern Zhejiang Province, China. Among these nonparametric methods, KNN has a long history of application for biomass estimation,[19,20] while SVR is known for its especially distinctive capability when the number of samples is relatively small.[21,22] RF and SGB have been introduced to forest investigation in recent years for their comparable performance.[10,17,23] Previous studies almost inspected these methods for forest biomass estimation. As far it is known, there are seldom systematic comparisons for these approaches, especially in the subtropical forest ecosystems.

Under the circumstance that ground-based biomass samples are limited but have been confirmed and Landsat data are selected as the source imagery for its open access, the performance

of different representative modeling methods becomes worth exploration. The five approaches mentioned above have always been applied for both classification and regression, but the relationships between these two processes were seldom discussed. It is well known that biomass shows distinctions among different forest vegetation types, while previous studies provided the dissimilar impact of stratified class models on biomass estimation.[24,25] In this study, both the classification and biomass simulation of different machine-learning approaches were investigated to find if there is some relation between these two processes.

2 Material and Methods

2.1 *Study Area*

The study area (6915.76 km^2) is located between 29°36′ to 30°14′N and 119°9′ to 120°11′E in northwestern Zhejiang Province in China along the Qiantang River, with ~80% of the area covered with forest (Fig. 1). The region is characterized as a subtropical monsoon climate with an annual average precipitation of about 1502 mm and a mean temperature of 17.7°C. The elevation ranges from 1 to 1420 m. The main forest species are pine, Chinese fir, broadleaf, bamboo, and other secondary forest species where most of the primeval forest has been almost exhausted.

2.2 *Field Data*

In total, there are 299 sample plots in the study area provided by the Zhejiang Forestry Academy. The dataset is part of the provincial data collection using a stratified random sampling method during the growing season of 2010, by taking comprehensive consideration of qualitative forest species, geomorphology, ecological niche, and location within the whole province. The theory of errors is employed to assure the quantity of typical samples both in training and validation. The size of each inventory plot is 20 m × 20 m for trees, with three 2 m × 2 m subplots set in the diagonal line of each plot for shrubs and grasslands. Field measurements includes diameter at breast height (DBH), tree height, tree age, species, the center positions of plots, and so on, using measurement tools, e.g., DBH tape, height measuring instrument, Global Positioning System

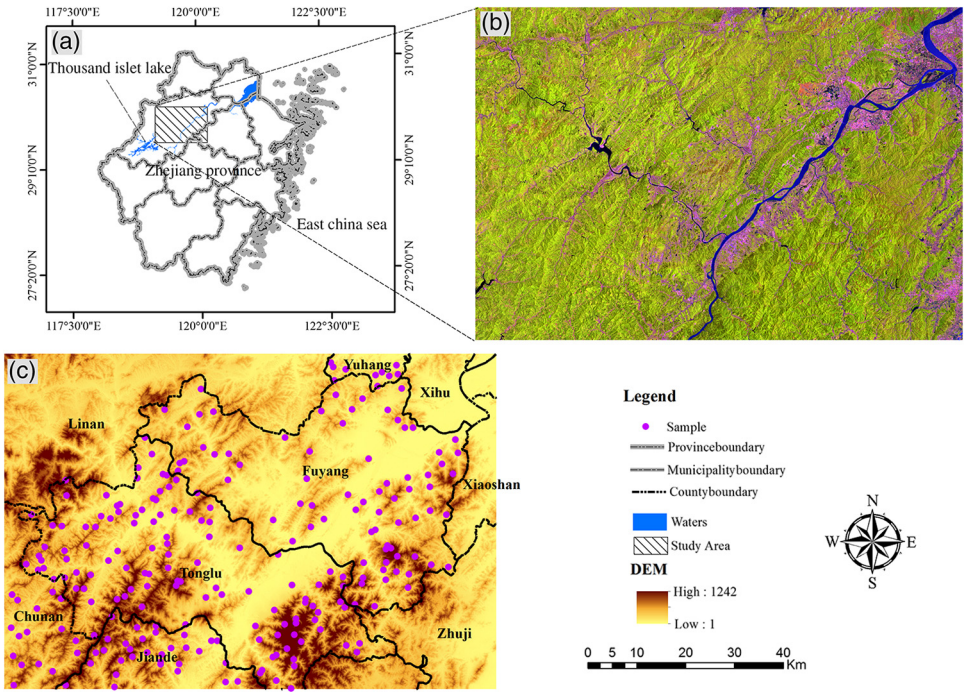

Fig. 1 (a) The location of study area in Zhejiang Province, China, (b) 2010 Landsat image (RGB:543), and (c) DEM and sample plots in the study area.

Table 1 Statistics of field-based AGB (ton/ha).

	Number of samples	Min	Max	Average	Standard deviation	CV (%)
In total	292	3.35	215.00	89.62	43.07	48.06
Broadleaf	99	8.05	195.94	92.00	43.95	47.77
Coniferous	44	21.10	190.76	97.99	39.45	40.26
Mix	103	25.45	215.00	102.07	38.91	38.12
Bamboo	30	16.14	108.04	60.50	21.25	35.12
Shrub	16	3.35	45.56	26.36	14.37	54.51

(GPS) and etc. Yuan et al.[26] established species-specific allometric equations to calculate single tree biomass for nine categories including pine (e.g., *Pinus massoniana*, *P. elliottii*), Chinese fir (e.g., *Cunninghamia lanceolata*), hardwood (e.g., *Cyclobalanopsis glauca*, *Cinnamomum camphora*), softwood (e.g., *Salix babylonica*), *Phyllostachys heterocycla* var. *pubescens*, miscellaneous bamboo, shrubs, and herbs. The total AGB was calculated as the sum of tree, shrub, and herb layers in each sample plot using the above corresponding models. All the plots were further checked from the high-resolution imagery through Google Earth for homogeneity. Moreover, the 3-σ (standard deviation) measurement, which is the assumption that values greater than the mean value plus three times the standard deviation or less than the mean value minus three times the standard deviation are considered as errors, was used to remove outliers. Ultimately, considering the 30-m resolution in the study area, there were 292 samples combined into five classes, viz. evergreen broadleaf forest (coniferous species proportion <10%), coniferous forest (pine/Chinese fir proportion greater than 90%), mixed forest (conifers proportion 10% to 90%), bamboo forest, and shrub forest (canopy density <0.3), as shown in Table 1.

2.3 *Remote-Sensing Data*

To avoid the influence of scan-line errors in Landsat Enhanced Thematic Mapper Plus (ETM+7), two cloud-free 1T Landsat TM 5 satellite images (path/row: 119/39) radiometrically and geometrically calibrated were acquired on May 24, 2010, and July 5, 2008, from the USGS.[27] Six spectral bands with 30-m spatial resolution including blue band (0.45 to 0.52 μm), green band (0.52 to 0.60 μm), red band (0.63 to 0.69 μm), near infrared band (0.76 to 0.90 μm), mid-infrared bands band (1.55 to 1.75 μm), and mid-infrared band (2.09 to 2.35 μm) were selected (except the thermal infrared band because of the coarser resolution). Atmospheric correction was conducted for 2010 imagery using fast line-of-sight atmospheric analysis of the spectral hypercube (FLAASH) model in the ENVI software.[28] The calibrated imagery was used as the reference and pseudoinvariant features were selected to complete the relative radiometric calibration for 2008 imagery.[29] Elevation data with the same resolution were acquired from the ASTER GDEM V2 product.[30]

To make full use of remote-sensing data, historic research was consulted to calculate related variables for AGB estimation. In total, there were 41 candidate predictor variables, including six multispectral bands, eight vegetation indices, 24 texture information, and three topographic features. Vegetation indices conducive to biomass simulation mentioned by previous studies were considered (Table 2). Normalized difference vegetation index (NDVI) is the most widely used index for its distinct explanation of vegetation coverage.[31] Two ratio indices, Ratio43 and Ratio54, were also calculated for their simplicity and effectiveness.[32,33] Tasseled cap (TC) transformation, which condenses six original bands into three components containing vegetation information named brightness (TCB), greenness (TCG), and wetness (TCW) were executed[17] as they were tested as the most correlated variables to biomass.[34] Another two indices, TC distance (TCD) reflecting vegetation composition and structure[35] and TC angle (TCA) representing the gradient of percent vegetation cover, were also generated.[36]

Table 2 Summary of vegetation indices used in the present study.

Formulation	Vegetation index	Reference
$NDVI = (NIR - Red)/(NIR + Red)$	NDVI	Rouse Jr et al.[31]
$Ratio43 = NIR/Red$	Ratio43	Tucker[33]
$Ratio54 = SIR/NIR$	Ratio54	Lu et al.[32]
$TCD = \sqrt{(TCG^2 + TCB^2)}$	TCD	Duane et al.[35]
$TCA = \arctan(TCG/TCB)$	TCA	Powell et al.[36]

Note: NIR is Landsat 5 TM band 4, Red is Landsat 5 TM band 3, and SIR is Landsat 5 TM band 5. TCG is greenness and TCB is brightness.

Lu[2] testified the importance of texture variables for biomass prediction, especially in complex structure forests. The positive effect of texture information over spectral bands alone on biomass simulation was also affirmed in southwest Colorado.[37] In this study, texture variables were extracted from three spectral bands, including the red band, the near-infrared band, and the short-wave infrared band, through the representative method gray-level co-occurrence (GLCM) calculation using a 3×3 pixels processing window including eight measurements (mean, variance, homogeneity, contrast, dissimilarity, entropy, second moment, and correlation).[37,38] Additionally, natural factors including geomorphology may influence the distribution of biomass,[39] so three topographic features were derived from the DEM: elevation, slope, and aspect.

2.4 *Modeling and Accuracy Assessment*

All remote-sensing derivatives responding to the sample plots were recalculated for the mean values using a 3×3 pixels moving window to reduce the alignment error from field collections. Moreover, to avoid the influence of other land covers, a forest mask was built based on the local land use map to assure that the regression was conducted in a forest area.

All the regression methods were implemented with two variable strategies, respectively. For the first strategy, all the variables were used for all the models. For the second modeling strategy, Pearson product moment correlation coefficients were calculated to choose variables that show significant importance between *in situ* measured biomass and Landsat-derived variables. Only these selected variables were employed for subsequent analysis. During the modeling process, the field measured AGB was treated as a dependent variable and remote-sensing derivates were used as independent variables. To compare with machine-learning methods, the Steplm was also conducted as it is the most representative method of traditional regression approaches.[4,40,41] A simple description of these four employed nonparametric approaches is as follows.

KNN has been extensively applied since it was successfully used for timber volume estimation in New Zealand.[19] The basic principle of KNN is to find the minimum spectral distance between an objective pixel and a reference pixel. The value of the objective pixel is always estimated by a weighted distance calculation based on the nearest neighbor pixels. In general, the spectral distance is inversely proportional to the spectral distance or square distance with different weights. Frequently used distance metrics include Euclidean distance, Mahalanobis distance, and Chebyshev distance. Consequently, the number of nearest neighbors and distance functions are two important parameters to determine the model performance.[34,42,43]

Mountrakis et al.[44] has presented an integrated description on SVR, which is famous for its excellent performance, especially in the case of limited reference data. It is originally designed as a binary classifier based on structural risk minimization to find the optimal separating hyperplane between two classes. Subsequently, it becomes progressively applied to multiclass classifications. The algorithm can not only be used for classification assignments, but can also be applied to solve regression problems. The wisdom of this method consists of developed kernel functions to project the feature space into high-dimensional space, creating nonlinear problems to be solved with linear solutions.[45] Commonly used kernel functions include linear, polynomial, sigmoid, and radial basis function (RBF) kernels. RBF is the most popular option because of its

steadiness. Once RBF is determined for the SVR, another two important parameters to control the algorithm accuracy are the penalization parameter C and the width of a Gaussian RBF kernel.[21,46]

RF was proposed as an ensemble of decision trees by Breiman[47] as an improvement of the classification and regression tree method. Through a random selection of variables at each node and bootstrap selection of samples without pruning (about $2/3$ of the samples were selected for training and the remaining $1/3$ of the samples were called out-of-bag samples for validation), the algorithm generates multiple decision trees growing to the maximum. There are two parameters named *mtry* and *ntree* to be adjusted, where *mtry* is the number of prediction variables used in each node and *ntree* is the number of regression trees. The randomness of sampling and variable selection proved to be prone to over-fitting compared to other methods. Another attractive characteristic of this algorithm is that RF can automatically calculate the features' relative importance using two feature evaluation mechanisms.[16,46]

SGB combines the advantages of boosting and bagging approaches. Starting from randomly selecting the samples, it utilizes the steepest gradient boosting algorithm to produce small-scale trees.[48] Since all the constructed trees emphasize misclassified training data that are closer to correct classifications, SGB is less sensitive to the distribution of the original dataset. Furthermore, critical variables can be voluntarily selected from the multiple variables. SGB has the same advantage as RF in that only part of the training data are randomly selected to improve estimation accuracy, which could effectively avoid over-fitting problems.[49,50]

The number of sample plots in this study was limited and the proportions of different forest species were quite different (e.g., the number of shrub samples is only 16, while the number of mixed forest is 103). Larger data volumes usually produce more accurate results for machine-learning approaches, therefore, to make use of all of the data and avoid the effect of different sample proportions on modeling performance,[51,52] the k-fold cross-validation ($k = 10$) with bootstrap was used to evaluate the prediction accuracy throughout the whole dataset.[53] In the 10-fold cross validation, the sample data were separated into 10 subsets, with each subset treated as the validation dataset and the other nine subsets used as training samples. This process was repeated until all the subsets had been traversed. Each model was implemented for 20 times via bootstrapping sampling. Modeling and accuracy assessment were conducted in the R open-source software through the "caret" package.[54,55] Root-mean-square error (RMSE), bias, and coefficient of determination (R^2) were calculated to compare the prediction performance of different approaches

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_e - y_f)^2}, \tag{1}$$

where y_e is the estimated biomass and y_f is the field measured biomass

$$\mathrm{Bias} = \bar{y}_e - \bar{y}_f, \tag{2}$$

where $\bar{y}_e$ is the mean value of the estimated biomass and $\bar{y}_f$ is the mean value of the field measured biomass. Generally, higher R^2 demonstrated more accurate estimates, while smaller RMSE and near zero bias present better results.[17]

Empirical models built on mono-temporal Landsat imagery and contemporaneous field inventory data can be effectively applied to time-series imageries for multitemporal biomass quantification.[17,56] To realize the bitemporal biomass estimation, the relationship between AGB and remote-sensing data established using the optimal method in 2010 was applied to the 2008 imagery. Furthermore, machine-learning methods were also used for forest classification to investigate whether there is a close relationship between regression and classification.

3 Results and Discussion

The values of field measured biomass ranged from 3.35 to 215 ton/ha (Table 1), and the majority of sample plots were within the forest types of mixed forest and broadleaf forest. The values of coefficients of variation (CV) indicated that there was a wide range of variability within each

Table 3 Correlations between field-based measured AGB and remote-sensing derivatives.

Variables	b2	b3	b4	Brightness	Greenness	Wetness	TCD	b3_mean	b4_mean
AGB	−0.280**	−0.233**	−0.258**	−0.353**	−0.231**	0.477**	−0.306**	−0.169**	−0.241**

Variables	b5	b7	Ratio54	b5_mean	b5_variance	b5_contrast	b5_dissimilarity	b4_correlation
AGB	−0.415**	−0.461**	−0.251**	−0.413**	0.129*	0.119*	0.118*	−0.116*

*Significant at 0.05 level.
**Significant at 0.01 level.

forest type. Among all forest types, bamboo had the lowest CV value of 35.12 ton/ha and shrub had the highest CV value of 54.51 ton/ha.

3.1 *Variables Analysis*

Table 3 showed that 17 of the 41 predictor variables were significantly related to the *in situ* AGB. Thus, these 17 selected variables were used in the second strategy for a comparison analysis with the first strategy using all the variables. In Table 3, all the multispectral bands (band2 to band5, band7) except for the blue band (band1) were selected as candidate variables, and the short-wave band (band7) had the highest negative correlation with biomass. Additionally, selected vegetation indices included wetness, brightness, greenness, TCD, and Ratio54. Among these related indices, wetness had the greatest positive correlation coefficients with biomass, while the others had negative relations. However, the most widely used NDVI was not selected in our study, indicating that NDVI may not be suitable to estimate AGB in this area. Moreover, seven texture variables were selected, but topographic variables were all eliminated. Except for the mean measurements, the other related texture indices had a positive relation with AGB. The reason that there was no significant relationship between AGB and topographic variables may be due to the limited coverage of our study area.

3.2 *Above-Ground Biomass Modeling by Different Methods*

3.2.1 *Comparison of models' above-ground biomass estimation performance with all variables*

AGB estimates using all the variables for different methods in terms of RMSE and R^2 were shown in Figs. 2(a) and 2(c) (the horizontal axis represented bootstrapped models for 20 times and the vertical axis represented three measurements of accuracy assessment), which indicated that RF performed better than the others, with the highest value of R^2 0.63 and the lowest RMSE 26.22 ton/ha. Meanwhile, the mean values of R^2 were 0.27, 0.21, 0.38, and 0.55, and the mean RMSE values were 41.12, 39.71, 34.61, and 28.64 ton/ha for Steplm, KNN, SVR, and SGB, respectively (Fig. 3). In Fig. 2(e), Steplm achieved the most stable bias values, while RF and SGB had much more accurate and stable performances than the other two approaches with the corresponding RMSE CV (RF: 0.98%, SGB: 0.98%) and R^2 CV (RF: 1.31%, SGB: 1.55%). The mean bias values for KNN, SVR RF, and SGB were −0.46, −1.69, 0.89, and 0.07, which affirmed the similar conclusion of a previous study that SGB has an incomparable stability in estimation performance.[10] Compared to SGB, KNN, SVR, and RF all achieved more fluctuant bias values.

3.2.2 *Comparison of models' above-ground biomass estimation performance with selected variables*

As shown in Figs. 2(b), 2(d), and 2(f), models that were completed with Pearson correlation selected variables had a comparable result to the first strategy. In general, the same conclusion was that RF outperformed the other algorithms with the highest R^2 values (the mean value is

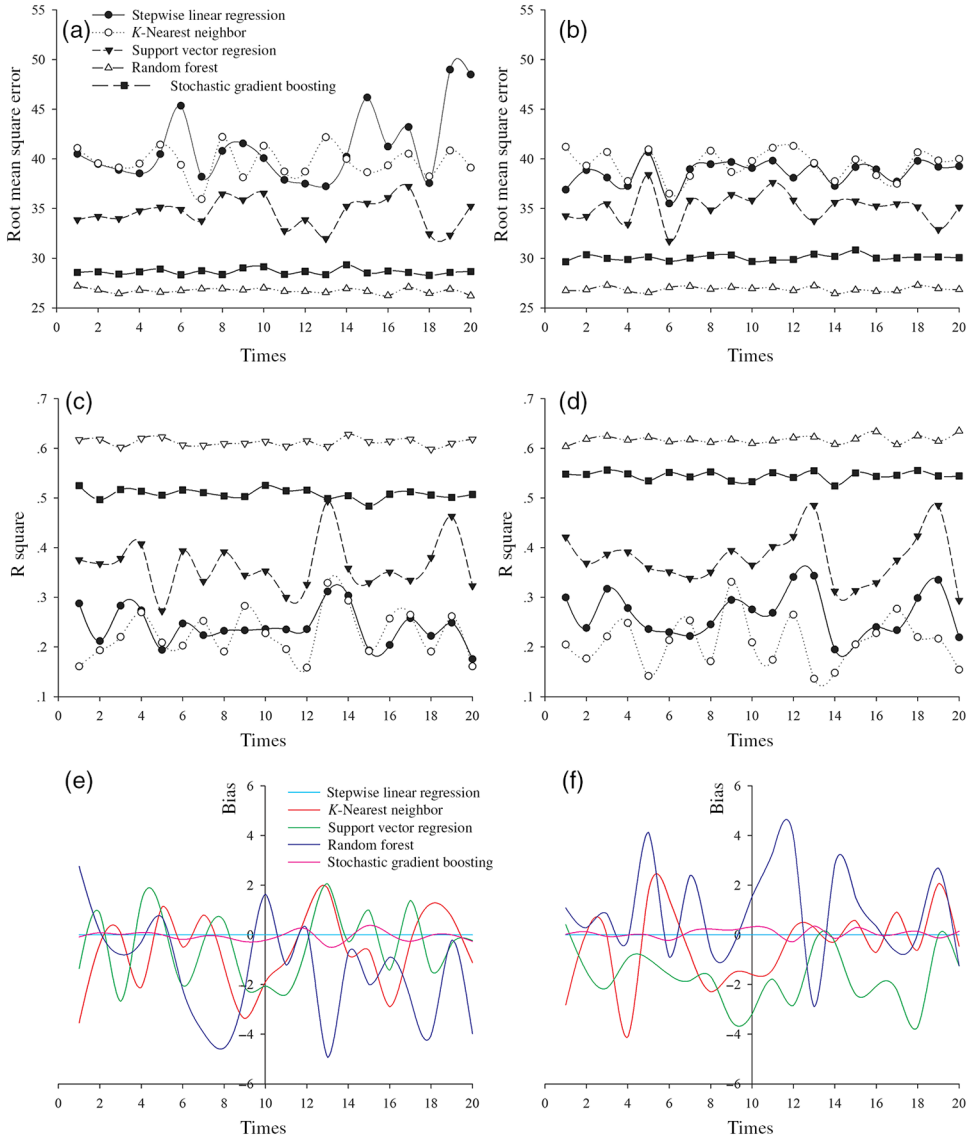

Fig. 2 Comparison of the (a) RMSE using all features, (b) RMSE using selected features, (c) R^2 using all features, (d) R^2 using selected features, (e) bias using all features, and (f) bias using selected features for Steplm, KNN, SVR, RF, and SGB.

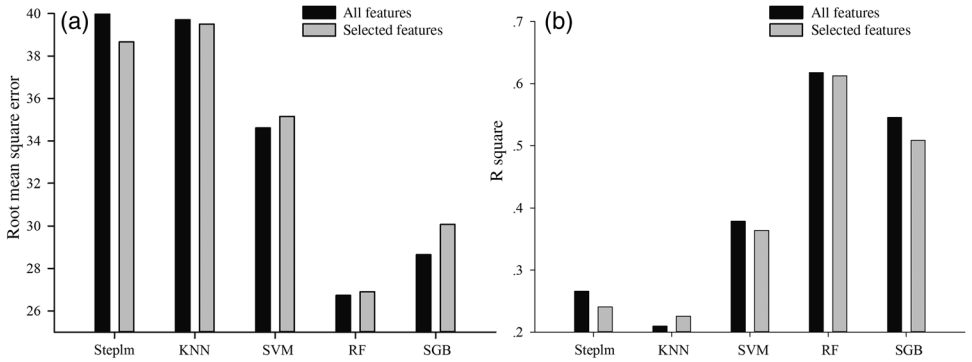

Fig. 3 Comparison of the (a) mean values of RMSE and (b) mean values of R^2 for five regression methods using two feature strategies.

0.63) and lowest RMSE values (the mean value is 26.44 ton/ha). Meanwhile, the mean values of RMSE in Fig. 3(a) were 38.67, 39.50, 35.15, and 30.07 ton/ha, and the mean values of R^2 in Fig. 3(b) were 0.24, 0.22, 0.36, and 0.51 for Steplm, KNN, SVR, and SGB, respectively. Moreover, the variance both in RMSE and R^2 revealed that the RF and SGB algorithms achieved more competitive representations, which was in accordance with the former results. In terms of bias shown in Figs. 2(e) and 2(f), SGB outperformed the other methods by obtaining a much smaller mean value (−0.06), followed by SVR (−0.45), KNN (−0.65), and RF (−1.45), respectively. It is unexpected that RF had the largest bias.

3.2.3 *Comparison of feature selection strategies' above-ground biomass estimation performance*

With regard to the R^2 and RMSE, Fig. 3 showed that reduced features produced higher RMSE and lower R^2 for SVR, RF and SGB, except for KNN and Steplm. One reason for this phenomenon may be the same as Latifi et al.[14] mentioned; that KNN was more sensitive to reduced variables. For Steplm, R^2 and RMSE showed different tendencies, where R^2 decreased but RMSE also decreased. However, for the other methods, the reduction of variables did not distinctly improve the accuracy. The increasing RMSE and decreasing R^2 even indicated that feature selection weakened the modeling performance.

To investigate whether the discarded variables in the second strategy have some effect on estimations, the most important features selected by RF and SGB were inspected. The top 20 variables were listed in Table 4 for RF and SGB, with an obvious difference observed for the two methods. Even compared to the result of Pearson correlation, the selected variables were evidently distinct, demonstrating that features individually irrelevant to biomass also made contributions to the machine-learning method biomass prediction. This result revealed that features derived from optical remote-sensing imagery had a potential relationship with biomass, but traditional linear or individual factor correlation analysis may hinder the relationship exploration between AGB and remote-sensing-derived variables, which may eliminate the potentially useful feature and affect the result of modeling, especially for the "self-learning" machine-learning approaches. Furthermore, the number of variables generated from the medium resolution optical remote-sensing image was countable when compared to high-dimensional data such as from hyperspectral imagery. Therefore, it is advisable to make use of all the Landsat imagery derivatives.

3.2.4 *Comparison of models' above-ground biomass estimation performance with additional derivatives from 2008 imagery*

To investigate the effect of multiple images from a similar period on biomass simulation, derivatives from 2008 were added to the 2010 images. Figure 4 showed the results of five regression methods using two datasets. In general, RMSE increased for all the methods after adding the variables derived from 2008 TM imagery though R^2 augmented at the same time, yet bias represented different changes for different methods. Wilcoxon signed-rank test was further calculated to evaluate whether there was significant improvement for adding the 2008 images to the

Table 4 Variable selection statistics in RF and SGB estimations.

Frequency	Selected features in RF	Selected features in SGB
20 times	Wetness, b7, b5, b5_mean, b4_mean, brightness, TCD, b2, b3_variance, greenness, b4_dissimilarity,b4, TCA, b4_contrast	Wetness, b7, b5_variance, Ratio54, b5_correlation, TCA, b4_variance, b4_correlation, b3_homo, b3_correlation, b5_mean, b5_dissimilarity, b4_mean, b5, b4_contrast, b2
1 to 19 times	Ratio54, b5_dissimilarity, b4_variance, b5_homo, b3, b1, b3_homo, b5_contrast, Ratio43, b4_homo, ndvic, ndvi, b5_variance	b3, b3_variance, ndvi, b5_contrast

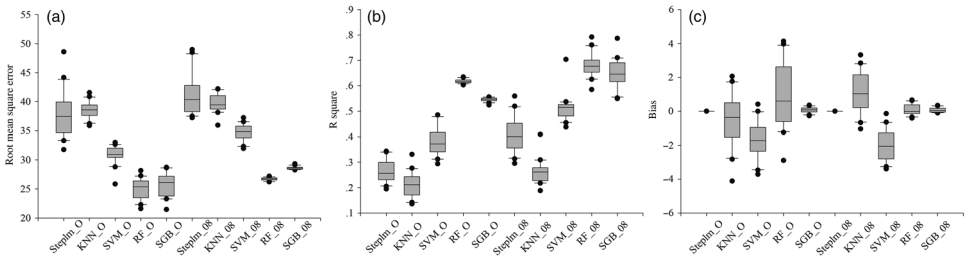

Fig. 4 Comparison of the (a) mean values of RMSE, (b) mean values of R^2, and (c) mean values of bias for five regression methods using original variables and additional 2008 imagery derivatives.

original 2010 data. In Table 5, only the values of bias for SVR, RF, and SGB showed significant differences, while all the others displayed no enhancement. The reason may be attributed to the limitation of supplemental images within the small time interval and similar seasons. Therefore, the original data in 2010 was used for the final AGB estimations.

3.2.5 *Comparison of above-ground biomass estimation performance and classification results*

Machine-learning approaches can be used for both regression and classification. Classification can be implemented when the predicted value is categorical, whereas in regression tasks the dependent variable is continuous. Both approaches are designed to find the most fitting modeling process. In this study, five methods were also applied to complete the classification processes using all the same variables adopted in a regression task. Compared to regression results, classification achieved a higher accuracy in general (Fig. 5), which may due to the fact that

Table 5 Summary of mean values comparison using Wilcoxon signed-rank test.

Measurement	Method	Original variable	With 2008 image derivatives	Sig.
Mean value of RMSE	lmStep	37.6922	41.1217	0.0050
	KNN	38.5813	39.7076	0.0370
	SVM	30.9250	34.6149	0.0000
	RF	25.0172	26.7289	0.0010
	SGB	25.7652	28.6446	0.0000
Mean value of R^2	lmStep	0.2657	0.4082	0.0000
	KNN	0.2097	0.2644	0.0040
	SVM	0.3785	0.5110	0.0000
	RF	0.6178	0.6823	0.0000
	SGB	0.5453	0.6472	0.0000
Mean value of bias	lmStep	0.0000	0.0000	0.0010
	KNN	−0.4640	1.0781	0.0000
	SVM	−1.6893	−2.0234	0.2960*
	RF	0.8930	0.0976	0.0930*
	SGB	0.0687	0.0642	0.6540*

*Significant at 0.05 level.

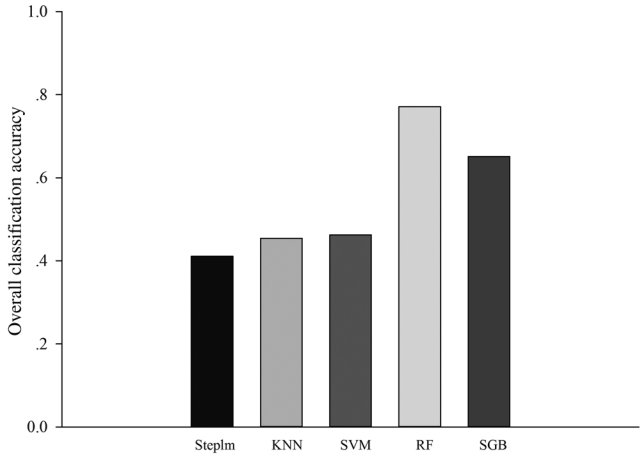

Fig. 5 Comparison of overall classification accuracy (percent correct) of different regression approaches.

classification was much less discrete and could be regarded as a regional unification of regression results. A Z-test was conducted to compare the differences between different methods (Table 6). The same conclusion was that RF achieved the best performance, and SGB, SVM, KNN, Steplm obtained descending accuracies, demonstrating that regression and classification had the same prediction tendency.

3.3 *Above-Ground Biomass Mapping*

As a result, in spite of more robust estimates for SGB, RF was selected as the final regression algorithm because of its highest R^2 and lowest RMSE. The predicted AGB map in the study area was produced using mean values of regression results for 20 times with all the features. Table 7 showed the statistics of the final estimated AGB in 2010, with the AGB values ranging from 21.11 to 191.73 ton/ha in the study area, with a mean value of 89.34 ton/ha and standard deviation of 18.88 ton/ha. The same models developed in 2010 were applied to images in 2008, and the statistics were appended to Table 6. There is no significant difference between the predicted AGBs in 2008 and 2010 ($Z = -0.0005$, p-value $= 0.9996$), which indicated that from 2008 to 2010, although the average AGB was increased, the growth was not significant.

Table 6 Z-test of four machine-learning methods.

	RF	SGB	SVM	KNN
RF	—	—	—	—
SGB	24.83 p-value < 2.2e – 16	—	—	—
SVM	54.77 p-value < 2.2e – 16	30.83 p-value < 2.2e – 16	—	—
KNN	55.82 p-value < 2.2e – 16	31.96 p-value < 2.2e – 16	1.19 p-value = 0.2333	—

Table 7 Statistical parameters of predicted AGB in 2010 and 2008.

	Maximum (ton/ha)	Minimum (ton/ha)	Average (ton/ha)	Standard deviation (ton/ha)	Skewness	Kurtosis
2010	191.73	21.11	91.72	26.63	0.33	3.25
2008	181.44	19.84	90.6	32.26	0.30	2.38

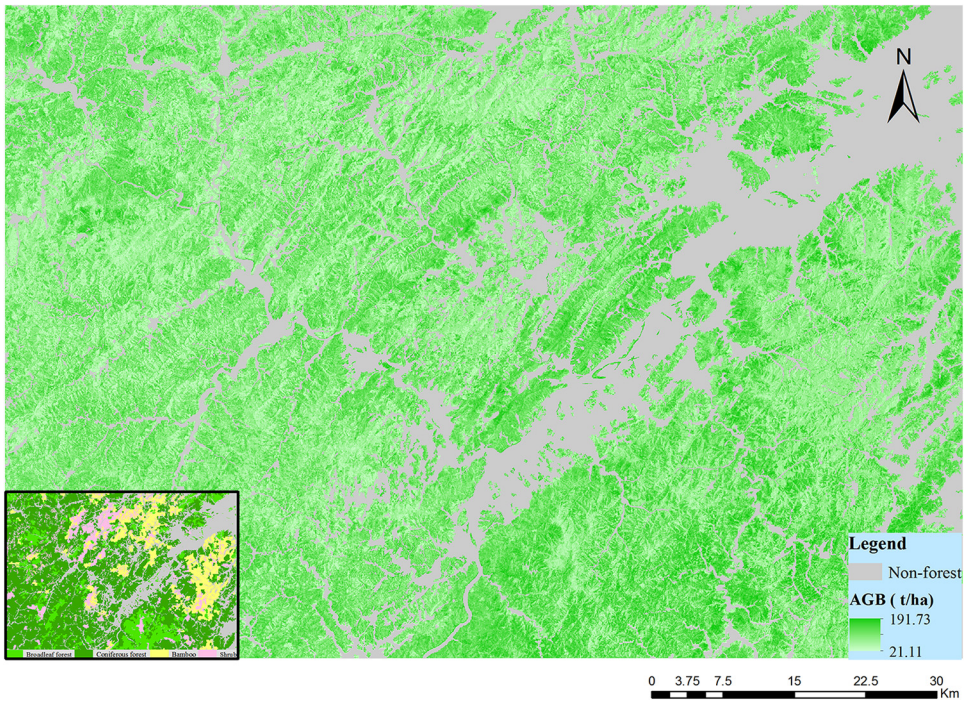

Fig. 6 Forest AGB map in northwestern Zhejiang produced by RF.

That means that the time interval of 2 years may be too short for biomass change inspection in spite of forest protection and management actions.

Figure 6 was the final map of the predicted AGB. Bamboo dominated in the eastern area and showed much lower biomass than those located in the south central area planted with broadleaf forest. Coniferous forest had higher AGB values than shrubs. The distribution of AGB was almost in accordance with the spatial pattern characteristic within different forest species, demonstrating that an AGB map estimated by combining Landsat images and field data through selected machine-learning methods is effective, which may provide helpful information to forest managers for sustainable development.

In the present study, an investigation of different regression approaches was completed to seek the most suitable combination of features and methods for the secondary forest. Compared to the traditional linear regression, machine-learning approaches, especially RF, have shown more accurate biomass predictions in terms of RMSE and R^2. Moreover, the reduction in a relatively small number of variables in the second feature strategy did not significantly improve the prediction accuracy. Therefore, we recommend using the whole features for machine-learning approach regressions for Landsat imagery.

3.4 *Future Works*

In this study, the "caret" package in the R software was employed because of its embedded function on comprehensive comparison between different modeling processes. Figure 7 showed the simple parameter optimization process for four machine-learning methods. For example, during one KNN modeling process, three default values of k (5, 7, 9) were compared, and 9 was selected as the final value based on the smallest RMSE result [Fig. 7(a)]. There were two parameters named *n.trees* and *interaction.depth* to be adjusted during the SGB regression process. Other parameters including *shrinkage* and *n.minobsinnode* were set as the default values (0.1 and 10) in the software [Fig. 7(d)]. In the future, more user-defined parameter values for different models should be required for more systematic investigations.

Simultaneously, there are some other limitations to be improved. First, the number of sample plots is limited as mentioned above, more samples can be collected for specific forest species because increasing the sampling numbers has a positive effect on estimate performance.[57] Zheng et al.[58] found that models for separate forest types had better performances compared to those for

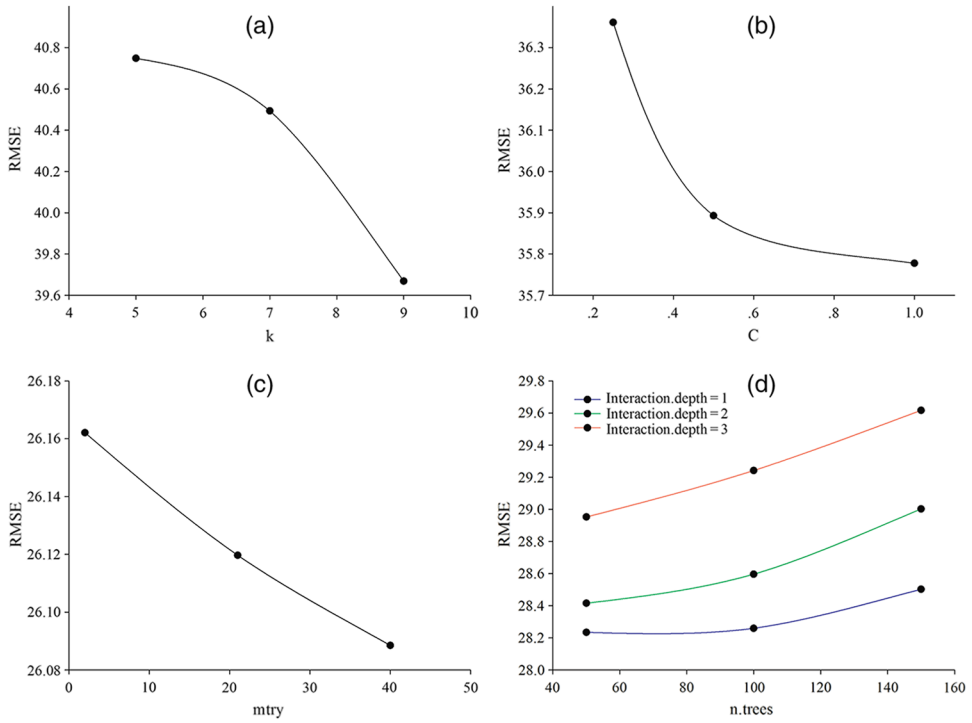

Fig. 7 Parameter adjustment of different approaches: (a) KNN, (b) SVM, (c) RF, and (d) SGB.

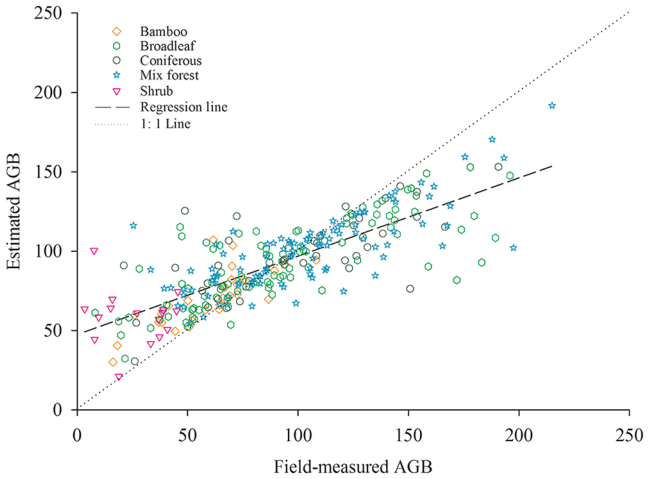

Fig. 8 Relationship between estimated and field measured AGB.

the entire forest. Second, more candidate features and feature selection methods could be integrated to determine whether there exists a threshold for the most appropriate number of Landsat imagery input variables. Third, although the accuracy of RF is relatively satisfactory, there is great potential for improvement. It should be pointed out that, as exhibited in Fig. 8, the estimated AGB result using RF still underestimated the highest biomass and overestimated the lowest biomass as the algorithm for regression averages the results of predictions.[47] To get a more precise biomass map combining the advantages of different algorithms, multimethod ensemble selection can be exploited in the future.[46]

4 Conclusion

Two sets of variables combined with traditional Steplm and four nonparametric algorithms including KNN, SVR, RF, and SGB, were investigated and compared to estimate forest

AGB in northwestern Zhejiang Province, China. As a result, ensemble methods RF and SGB showed more stable prediction performances, especially SGB. Moreover, RF produced the most accurate estimates in terms of RMSE and R^2. The results from two different feature strategies indicated that incorporating all variables derived from Landsat imagery, compared to the selected features using Pearson correction analysis, made more comprehensive use of relations between biomass and potential variables. The regional scale AGB estimation in this study developed with Landsat image and field-based AGB data provides a satisfactory result reflecting the spatial variability corresponding to the local tree species distribution patterns, revealing the great potential for large-area biomass estimation applications using machine-learning approaches with free and obtainable Landsat dataset, especially for developing countries.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (No. 41501190) and Zhejiang Provincial Natural Science Foundation of China (No. LQ14D010003). The authors are thankful to the USGS and NASA for the opening archives of the Landsat imagery, and we would like to acknowledge the R Development Team for the open-source package for the statistical analysis. The authors thank Editor Raymond Hunt and two anonymous reviewers for their constructive comments, suggestions and help in enhancing the manuscript.

References

1. D. S. Boyd and F. M. Danson, "Satellite remote sensing of forest resources: three decades of research development," *Prog. Phys. Geog.* **29**(1), 1–26 (2005).
2. D. Lu, "Aboveground biomass estimation using Landsat TM data in the Brazilian Amazon," *Int. J. Remote Sens.* **26**(12), 2509–2525 (2005).
3. L. P. Olander et al., "Reference scenarios for deforestation and forest degradation in support of REDD: a review of data and methods," *Environ. Res. Lett.* **3**(2), 0250112 (2008).
4. D. Lu et al., "A survey of remote sensing-based aboveground biomass estimation methods in forest ecosystems," *Int. J. Digital Earth* **9**, 1–43 (2014).
5. P. S. Roy and S. A. Ravan, "Biomass estimation using satellite remote sensing data—an investigation on possible approaches for natural forest," *J. Biosciences* **21**(4), 535–561 (1996).
6. G. Wang et al., "Uncertainties of mapping aboveground forest carbon due to plot locations using national forest inventory plot and remotely sensed data," *Scand. J. For. Res.* **26**, 360–373 (2011).
7. D. Lu, "The potential and challenge of remote sensing-based biomass estimation," *Int. J. Remote Sens.* **27**(7), 1297–1328 (2006).
8. L. Du et al., "Mapping forest biomass using remote sensing and national forest inventory in China," *Forests* **5**(6), 1267–1283 (2014).
9. C. Gómez et al., "Historical forest biomass dynamics modelled with Landsat spectral trajectories," *ISPRS J. Photogramm.* **93**, 14–28 (2014).
10. 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.* **101**, 36–46 (2015).
11. C. Yang, H. Huang, and S. Wang, "Estimation of tropical forest biomass using Landsat TM imagery and permanent plot data in Xishuangbanna, China," *Int. J. Remote Sens.* **32**(20), 5741–5756 (2011).
12. X. Tian et al., "Estimating montane forest above-ground biomass in the upper reaches of the Heihe River Basin using Landsat-TM data," *Int. J. Remote Sens.* **35**(21), 7339–7362 (2014).
13. C. J. Gleason and J. Im, "Forest biomass estimation from airborne LiDAR data using machine learning approaches," *Remote Sens. Environ.* **125**, 80–91 (2012).
14. H. Latifi, A. Nothdurft, and B. Koch, "Non-parametric prediction and mapping of standing timber volume and biomass in a temperate forest: application of multiple optical/LiDAR-derived predictors," *Forestry* **83**(4), 395–407 (2010).

15. G. M. Foody, D. S. Boyd, and M. E. J. Cutler, "Predictive relations of tropical forest biomass from Landsat TM data and their transferability between regions," *Remote Sens. Environ.* **85**(4), 463–474 (2003).
16. X. Zhu and D. Liu, "Improving forest aboveground biomass estimation using seasonal Landsat NDVI time-series," *ISPRS J. Photogramm.* **102**, 222–231 (2015).
17. 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).
18. G. Wang et al., "Mapping and spatial uncertainty analysis of forest vegetation carbon by combining national forest inventory data and satellite images," *Forest Ecol. Manage.* **258**(7), 1275–1283 (2009).
19. E. Tomppo, C. Goulding, and M. Katila, "Adapting Finnish multi-source forest inventory techniques to the New Zealand preharvest inventory," *Scand. J. For. Res.* **14**(2), 182–192 (1999).
20. H. Reese et al., "Applications using estimates of forest parameters derived from satellite and forest inventory data," *Comput. Electron. Agric.* **37**(1–3), 37–55 (2002).
21. M. Pal and P. M. Mather, "Support vector machines for classification in remote sensing," *Int. J. Remote Sens.* **26**(5), 1007–1011 (2005).
22. Y. Guo et al., "Optimal support vector machines for forest above-ground biomass estimation from multisource remote sensing data," in *IEEE Int. Geoscience and Remote Sensing Symp. (IGARSS 2012)*, 6388–6391 (2012).
23. J. Blackard et al., "Mapping U.S. forest biomass using nationwide forest inventory data and moderate resolution information," *Remote Sens. Environ.* **112**(4), 1658–1677 (2008).
24. H. Latifi et al., "Stratified aboveground forest biomass estimation by remote sensing data," *Int. J. Appl. Earth Obs.* **38**, 229–241 (2015).
25. P. Zhao et al., "Examining spectral reflectance saturation in Landsat imagery and corresponding solutions to improve forest aboveground biomass estimation," *Remote Sens.* **8**(6), 469 (2016).
26. W. Yuan et al., "Study on biomass model of key ecological forest in Zhejiang Province," *J. Zhejiang For. Sci. Technol.* **29**(2), 1–5 (2009).
27. U.S. Geological Survey, "Earth explorer," http://earthexplorer.usgs.gov/ (20 July 2015).
28. J. Yuan and Z. Niu, "Evaluation of atmospheric correction using FLAASH," in *Int. Workshop on Earth Observation and Remote Sensing Applications (EORSA 2008)*, pp. 302–307 (2008).
29. S. Myeong, D. J. Nowak, and M. J. Duggin, "A temporal analysis of urban forest carbon storage using remote sensing," *Remote Sens. Environ.* **101**(2), 277–282 (2006).
30. Japan METI and U.S. NASA, "The advanced spaceborne thermal emission and reflection radiometer (ASTER) global digital elevation model (GDEM)," http://gdem.ersdac.jspacesystems.or.jp/ (26 June 2009).
31. J. W. Rouse et al., "Monitoring vegetation systems in the great plains with ERTS," *NASA Spec. Publ.* **351**, 309 (1974).
32. D. Lu et al., "Relationships between forest stand parameters and Landsat TM spectral responses in the Brazilian Amazon Basin," *For. Ecol. Manage.* **198**(1–3), 149–167 (2004).
33. C. J. Tucker, "Red and Photographic infrared linear combinations for monitoring vegetation," *Remote Sens. Environ.* **8**(2), 127–150 (1979).
34. 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**(1–3), 129–144 (2006).
35. M. V. Duane et al., "Implications of alternative field-sampling designs on Landsat-based mapping of stand age and carbon stocks in Oregon Forests," *For. Sci.* **56**(4), 405–416 (2010).
36. S. Powell et al., "Quantification of impervious surface in the Snohomish water resources inventory area of Western Washington from 1972–2006," *Remote Sens. Environ.* **112**(4), 1895–1908 (2007).
37. K. Kelsey and J. Neff, "Estimates of aboveground biomass from texture analysis of Landsat imagery," *Remote Sens.* **6** (7), 6407–6422 (2014).
38. R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Trans. Syst. Man Cybern.* **3**(6), 610–621 (1973).

39. Y. Su et al., "Spatial distribution of forest aboveground biomass in China: estimation through combination of spaceborne lidar, optical imagery, and forest inventory data," *Remote Sens. Environ.* **173**, 187–199 (2016).
40. R. J. Hall et al., "Modeling forest stand structure attributes using Landsat ETM+ data: application to mapping of aboveground biomass and stand volume," *For. Ecol. Manage.* **225**(1–3), 378–390 (2006).
41. Y. Xie et al., "A comparison of two models with Landsat data for estimating above ground grassland biomass in Inner Mongolia, China," *Ecol. Model.* **220**(15), 1810–1818 (2009).
42. Z. Fazakas, M. Nilsson, and H. Olsson, "Regional forest biomass and wood volume estimation using satellite data and ancillary data," *Agr. For. Meteorol.* **98–99**, 417–425 (1999).
43. J. Jung et al., "Effects of national forest inventory plot location error on forest carbon stock estimation using k-nearest neighbor algorithm," *ISPRS J. Photogramm.* **81**, 82–92 (2013).
44. G. Mountrakis, J. Im, and C. Ogole, "Support vector machines in remote sensing: a review," *ISPRS J. Photogramm.* **66**(3), 247–259 (2011).
45. C. Chang and C. Lin, "LIBSVM: a library for support vector machines," *ACM Trans. Intel. Syst. Tech.* **2**(3), 1–27 (2011).
46. H. Feilhauer, G. P. Asner, and R. E. Martin, "Multi-method ensemble selection of spectral bands related to leaf biochemistry," *Remote Sens. Environ.* **164**, 57–65 (2015).
47. L. Breiman, "Random forests," *Mach. Learn.* **45** (1), 5–32 (2001).
48. G. G. Moisen et al., "Predicting tree species presence and basal area in Utah: a comparison of stochastic gradient boosting, generalized additive models, and tree-based methods," *Ecol. Model.* **199** (2), 176–187 (2006).
49. J. H. Friedman, "Stochastic gradient boosting," *Comput. Stat. Data Anal.* **38**(4), 367–378 (2002).
50. R. Lawrence, "Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis," *Remote Sens. Environ.* **90**(3), 331–336 (2004).
51. C. J. Gleason and J. Im, "Forest biomass estimation from airborne LiDAR data using machine learning approaches," *Remote Sens. Environ.* **125**, 80–91 (2012).
52. 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).
53. R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. of the 14th Int. Joint Conf. on Artificia*, Vol. 14, pp. 1137–1145 (1995).
54. M. Kuhn, "Building predictive models in R using the caret package," *J. Stat. Softw.* **28**(5), 1–26 (2008).
55. R Core Team, R: A language and environment for statistical computing, R foundation for statistical computing," http://www.R-project.org/ (20 September 2015).
56. L. Liu et al., "Improving artificial forest biomass estimates using afforestation age information from time series Landsat stacks," *Environ. Monit. Assess.* **186**(11), 7293–7306 (2014).
57. M. Schwieder et al., "Estimating fractional shrub cover using simulated EnMAP data: a comparison of three machine learning regression techniques," *Remote Sens.* **6**(4), 3427–3445 (2014).
58. D. Zheng et al., "Estimating aboveground biomass using Landsat 7 ETM+ data across a managed landscape in northern Wisconsin, USA," *Remote Sens. Environ.* **93**(3), 402–411 (2004).

Chaofan Wu received her bachelor's degree from Taiyuan University of Technology, China, in 2010. Currently, she is a PhD student at the Institution of Remote Sensing and Information System Application, Zhejiang University, China. Her research interests include remote-sensing image classification and forest AGB estimations based on remote-sensing imagery.

Huanhuan Shen obtained her bachelor's degree in 2013 from Chengdu Univerisity of Technology, China, and is a graduate student in Institution of Remote Sensing and Information System Application at Zhejiang University, China. Her research fields focus on land use planning and application of geographic information system (GIS).

Aihua Shen received his master's degree from College of Agriculture and Biotechnology of Zhejiang University, China. He is a deputy dean of Zhejiang Forestry Academy, China and a PhD student in Institution of Remote Sensing and Information System Application at Zhejiang University, China. His research fields focus on forest cultivation and forest ecology.

Jinsong Deng is an associate professor in Institution of Remote Sensing and Information System Application at Zhejiang University, China, and research staff in the School of Civil Engineering and Environmental Sciences and the School of Meteorology at the University of Oklahoma. He obtained his PhD from College of Environmental and Resource Sciences of Zhejiang University, China. His primary area of research is application of remote sensing and GIS.

Muye Gan earned his bachelor's degree from Nanjing Agricultural University and obtained his PhD from College of Environmental and Resource Sciences of Zhejiang University, China. He is a research associate in Institution of Remote Sensing and Information System Application at Zhejiang University, China. His research fields focus on urban vegetation landscapes based on remote sensing and GIS.

Jinxia Zhu obtained her PhD from College of Environmental and Resource Sciences of Zhejiang University, China, and is a teacher of Institute of Economic and Social Development at Zhejiang University of Finance and Economics. Her research fields focus on remote sensing classification and change detection.

Hongwei Xu is an associate professor in Institution of Remote Sensing and Information System Application at Zhejiang University, China. She obtained her PhD in 2004 from College of Environmental and Resource Sciences of Zhejiang University, China. Her primary area of research is land use planning and agricultural applications of remote sensing and GIS.

Ke Wang is a professor in College of Environmental and Resource Sciences of Zhejiang University, China, and a senior research scientist of the Institution of Remote Sensing and Information System Application of Zhejiang University, China. His research interests are application of remote sensing and GIS, rice nutrition diagnosis and land use planning.