Application of RF-KNN Optimal Technology for the Estimation of Forest Aboveground Biomass Using Multisource Remote Sensing Data

YING GUO, ZENGYUAN LI, ER-XUE CHEN, XINWEN YU and QISHENG HE

ABSTRACT

Quantifying the forest above ground biomass is critical for accurate carbon accounting. Qualities of the nonparametric K Nearest Neighbour technique make it an attractive tool for forest aboveground biomass (FAGB) estimation based on remote sense data. However, a thorough analysis of the usability of model parameter and feature optimal selection for the estimation of FAGB is missing. This study based on multisource data which included optical and LiDAR data and used random forest feature selection algorithm to build optimal KNN model for the estimation of FAGB estimation. The result showed the RMSE and $R^2$ of the optimal KNN model are 20.12 ton/hm$^2$ and 0.8 respectively fitted for FAGB estimation.

INTRODUCTION

It is well known that forest ecosystems play an important role in the earth. Continued deforestation and forest degradation will result in the loss of forest aboveground biomass (FAGB) magnifying the global negative effects of climate change [1]. The policy and management decisions governing these resources will play a critical role in mitigating these effects, thus requiring a robust method of

Ying Guo, Zengyuan Li, Er-xue Chen, Xinwen Yu, Institute of Forest Resources Information Techniques, Chinese Academy of Forestry, Beijing, 100091, China
Qisheng He, Hohai University, Nanjing, 211100, China
monitoring the spatial and temporal patterns of FAGB [2]. One solution is to develop robust approaches for estimating FAGB using remotely sensed data. The past three decades have produced significant advances in estimating FAGB including the application of different sensor data.

Multispectral remote sensing [3] have been used to map FAGB at moderate resolution and broad scales. However, passive optical sensors have difficulty penetrating beyond upper canopy layers [4] and are tend to be saturation in dense forest. LiDAR provides highly accurate measurements of forest structure [5]. However, due to the high cost of flight time, the need to limit scanning to near nadir in order to prevent ranging errors, they are not capable of imaging entire landscapes. The optimal strategy for mapping FAGB would include the finely detailed measurements of the vertical dimension as well as the broad spatial coverage of remote sensing. Thus, it is possible to map FAGB by statistically combining information from multiple sensors to take advantage of the highly detailed vertical measurements provided by LiDAR, the broad-scale mapping capabilities of passive optical sensors. Combining information from multiple sensors has yielded promising results for the estimation of forest structural characteristics [6]. Hudak [7] combined regression and co-kriging models from LiDAR and multispectral data; the results were more accurate than either data set alone. Wulder and Seeman [8] (2003) used texture metrics from Landsat TM images to improve LiDAR estimates of canopy height (from 61% to 67% variability explained).

Qualities of the k-nearest neighbor (kNN) method make it an attractive tool for multi-source remote sensing FAGB estimation [9]. The method assumes that similar forest exists within a large reference area covered by a satellite image and that the spectral radiometric responses of the pixels are only dependent on the state of the forest. Several examples can be found in the literature, including: Fazakas and Nilsson [10], Muinonen and Tokola [11], Nilsson [12], and Tomppo [13-16]. In Finland the kNN-based MSFI has been used for more than 10 years as a part of the NFI. Wall-to-wall maps are being used by the forest industry for timber procurement and ecologists use the maps for habitat analyses [17]. In Sweden, the method has been used to produce a complete map database for the whole country, named kNN Sweden. It has been used to improve forest statistics from the NFI by using post stratification based on stem volume strata derived from the database. The standard errors for estimates of total stem volume, stem volume for spruce, stem volume for pine, and woody biomass have been reduced by 10% to 30% at the county level.

Even though KNN method applications in the estimation of forest stand attributes using multisource remote sensing data have been intensively investigated during the last few years, a thorough analysis of the usability of model parameter and feature optimal selection for the estimation of FAGB is missing. The main objective of this paper is to describe methodology for using the kNN and related model parameter optimized techniques as well as random forest feature selection method, to improve FAGB estimation by using multi source remote sensing data, in the context of the FIA forest monitoring system.
METHODOLOGY

2.1 KNN Estimation Procedure

The KNN method is used here to generalize information from field plots to pixels for map production and local area estimation. A complete description of the procedure can be found in Franco-Lopez et al.[18] Thus, only the main features of the configurations used are described here. In the KNN estimation procedure, the variable $W_p$ for a specified pixel is predicted as the weighted average of the Values of the most spectrally similar reference points (the k nearest neighbours) according to the formula:

$$W_p = \sum_{i=1}^{k} \omega_{p_i,p} \times W_{p_i}$$  \hspace{1cm} (1)

Where $\omega_{p_i,p}$ are the weights of the k reference points, which may be computed as 1 minus the spectral distances or as inversely proportional to the spectral distances, as shown in formula:

$$\omega_{p_i,p} = \frac{1}{D_{p,p}} \left( \sum_{j=1}^{k} \frac{1}{D_{p,j}} \right)$$  \hspace{1cm} (2)

$$\omega_{p_i,p} = (1-D_{p,p}) \left/ \sum_{j=1}^{k} \frac{1}{(1-D_{p,j})} \right.$$  \hspace{1cm} (3)

Other factors which influence the performance of KNN procedures are the consideration of horizontal and/or vertical reference areas and the application of environmental stratifications for the selection of the k points. More importantly, several options exist regarding the form of spectral distance which is used to identify the k nearest neighbours[19]. In the current case, the effect of using three different spectral distances was evaluated, Abstract, Euclidean and Mahalanobis distance, according to formulae:

$$D_{p_i,p} = |x_{p_i} - x_p|$$  \hspace{1cm} (4)

$$D_{p_i,p} = \sqrt{\sum_{i=1}^{n}(x_{p_i} - x_p)^2}$$  \hspace{1cm} (5)

$$D_{p_i,p} = \sqrt{(x_{p_i} - x_p)^T H^{-1} (x_{p_i} - x_p)}$$  \hspace{1cm} (6)
2.2 Random Forest (RF)

RF models have been shown to reduce bias and over fitting, work well with random inputs, and in some cases tend to be more accurate than simple regression techniques for biomass estimation[20-21] and have been previously employed in forestry for modeling quantities[22]. RF models have roots in classification and regression trees (C4.5), which are a series of binary rule-based decisions that dictate how an input is related to its predictor variable. If the error associated with splitting a single rule into multiple rules is lower than using a single rule, the regression tree will ‘branch out’ and the tree will grow more rules. The decision tree stops growing when the minimum error versus the input data is obtained. The major advantage of such trees is their flexibility as regression trees can accurately describe complex relationships between variables at multiple scales and multiple predictor strengths. An aggregation of such trees usually leads to more accurate solutions, and the mechanics of the estimation can be found in Breiman. There is also a heuristic component to RF as the initial sample for each decision tree is chosen randomly, which may lead to different results each time the model is run.

The RF model fit for this study was implemented in java programming language based on the work of Breiman and Cutler. The importance of each node of in RF regression trees is determined by using input data to assess which variable at that node best characterizes the remaining observations.

3. TEST SITE AND DATA

3.1 Test Site

The study area is located in the XiShui Forest Farm, which belongs to the Qilian mountain national nature reserve in the Su'nan Yugu Autonomous County of GanSu province. The elevation ranges from 2700 to 3200 meters. The forest stands are mainly dominated by Picea crassifolia, which is natural mature coniferous forest.

3.2 Remotely Sensed Data

The remote sensing data used in this study consist of SPOT5 imagery and LiDAR point cloud statistical variables, which include high point digits, mean, variance, skewness, kurtosis, mean absolute deviation, standard deviation and canopy coverage, which is the point ratio of vegetation and all samples. For the computation of high point digits, the cloud points of vegetation around a sample area were ordered by height. Nineteen high point digits in every 5 percent ranged from 95 percent to 5 percent of height. To supplement the SPOT5 data, we also included the inverse and logarithm of spectra values, principle component, texture factor and vegetable indexes in the analysis. The adopted texture factors are mean, variance,
homogeneity, contrast, dissimilarity, entropy, second moment and correlation. The computing methods of vegetable index contained the ratio vegetation index (RVI), normalized difference vegetation index (NDVI), soil adjusted vegetation index (SAVI) and modified soil adjusted vegetation index (MSAVI), summed to 55 features involved in training the nonparametric models and estimating the FAGB.

3.3 Ground Survey Forest Plot Data

A set of 87 plots contained 5000 trees inside the coverage of the SOPT5 satellite image were available as ground reference data in August 2007 and June 2008. The plots were georeferenced with GPS, and the accuracy is expected to be within 10m for 99% of the plots. A series of parameters in each plot were measured, including height, DBH and crown of every tree. The FAGB was calculated on the basis of the full caliper measurements of trees within plots, the sample measurements of tree heights and the use of FAGB tables, which provided the estimated FAGB for each combination of tree DBH and height.

3.4 Accuracy Assessment

The estimation was evaluated using prediction error, which measures how well a model predicts the response value of a future observation. For every trial, the accuracy of our estimates of basal area and volume were examined using the root mean square error (RMSE) and the coefficient of determination ($R^2$).

4. RESULTS AND DISCUSSION

4.1 Distance Metric Evaluation

Figure 1 shows the RMSE values obtained in the estimation of forest biomass when abstract, Mahalanobis and Euclidean distance are used. This illustrates the typical behavior of the prediction errors RMSE associated with the KNN estimator. There is a rapid early increase in the precision of the technique with the addition of the first few neighbors, there after the marginal gains diminish and precision levels off. The Mahalanobis distance metric produced RMSE results at least 5% smaller than those for the abstract metric and Euclidean metric for any number of neighbors. We note that this is similar to the results reported by Nilsson (1997). Given this result, we set the Mahalanobis distance as the distance metric for use in subsequent trials.
4.2. Neighbor’s Weighting Functions

The results of applying two different weighting functions for forest AGB estimation are shown in Figure 2. The simple 1 minus the spectral distance estimator worked best. In order to take advantage of weighting functions such as inversely proportional to the distance or 1 minus the distance, it is necessary to have an identifiable close neighbor and some distance between this and the next one. It was apparent that instead of having one spectrally close neighbor, in most cases, there were several close neighbors. Thus, for subsequent trials, Eq. (3) was used to allocate equal weights to neighbors.
4.3 Number of Neighbors

The errors in KNN estimations for FAGB using Mahalanobis distance, 1 minus spectral distance weighting functions. It is clear that there was a rapid early gain in overall precision with the addition of neighbors. The minimum values for the RMSE is 20.19 when the number of neighbors was six; the maximum value for the RMSE is 29.13 when the number of neighbors was one.

For all values of k neighbors, divided the predicted FAGB value into 10 group from the minimum value to maximum value, counted the sample number of every interval and compared with the field-measure value distribution. It is clear with the increase of k, the distributions of FAGB estimation become centralized and the intervals become short. Although k=1, the bias was smaller, the value of RMSE is 29.13, larger than others. Thus, K equals to 6 is optimal value for FAGB estimation.

<table>
<thead>
<tr>
<th>RMSE</th>
<th>K VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAXIMUM RMSE VALUE</td>
<td>29.13</td>
</tr>
<tr>
<td>MINIMUM RMSE VALUE</td>
<td>20.19</td>
</tr>
</tbody>
</table>

TABLE I. THE K CORRESPOND TO THE MAXIMUM RMSE AND MINIMUM RMSE.
4.4 RF Feature Selection

Choosing predict or variables from all features derived forest measurements is important to assure the most accurate biomass models. In our study, variable reduction was performed by using RF feature selection method. The final covariates included measures of height and principle component. These variables are common in biomass estimation studies. Feature selection based on RF algorithm outperformed the use of nonfeature selection, which demonstrated that there is an advantage to an integrated analysis approach incorporating multiple remote sensing data.

<table>
<thead>
<tr>
<th>TABLE II. INVERSED BIOMASS ACCURACY BASED ON KNN AND MULTISOURCE REMOTE SENSING DATA BETWEEN FEATURE SELECTION AND NON-FEATURE SELECTION.</th>
</tr>
</thead>
<tbody>
<tr>
<td>NON-FEATURE SELECTION KNN</td>
</tr>
<tr>
<td>----------------------------</td>
</tr>
<tr>
<td>( R^2 )</td>
</tr>
<tr>
<td>RMSE</td>
</tr>
</tbody>
</table>

5. CONCLUSION

The KNN method is very promising for propagating FAGB through the landscape. The simplicity of this method and its role in post stratification provides a very feasible tool for wall-to-wall mapping of FAGB. The KNN method is a versatile technique with potential for combining different sources of remote sensing data, not only from passive optical data, but even from LiDAR. The combination of different remote sensors is straightforward since the method is based solely in the search for similar units.

Since the same neighbors are involved, the quality of the estimation relies on the correlation among the variables and the image features. The number of nearest neighbors to employ in an estimation problem is determined by the particular goals of a survey. When applying KNN for map production and using the six nearest neighbors, the estimator is unbiased and the range in variability of the sample is largely preserved.

The results of this study also show that RF is adept at identifying relevant features in high-dimensional data containing attributes on multiple scales of measurement. RF identifies features with small marginal effects. Relevant attributes may be selected from either data type, and there may be interactions across data.
types. RF is thus well-suited to the study of FAGB estimation based multisource remote sensing data.

ACKNOWLEDGEMENTS

This project was supported partially by Introduce international advanced agriculture science and technology plan (2012-4-79).

REFERENCES

15. Tomppo, E., 1997a, Application of remote sensing in Finnish national forest inventory. In: Application of remote sensing in European forest monitoring. International workshop


