REGIONAL FOREST ABOVE-GROUND BIOMASS RETRIEVAL BY OPTIMIZED K-NN ALGORITHM IN NORTHEAST CHINA

*Xin Tian[(1)(2)] *, Erxue Chen[(1)] *, Zengyuan Li[(1)], Z. Bob Su[(2)], Lina Bai[(1)], Christiaan van der Tol[(2)]*

(1) Institute of Forest Resource Information Techniques, Chinese Academy of Forestry, Wanshoushanhou, 100091, Beijing, P.R. China

(2) Faculty of Geo-Information Science and Earth Observation, University of Twente, Hengelosestraat 99, 7500 AA, Enschede, The Netherlands

ABSTRACT

This study explores retrieval of wall-to-wall forest above-ground biomass (AGB) over Jilin province in Northeast China, using the optimized non-parametric *k*-NN method, the 7th National Forest Inventory (NFI) data, and the remote sensing data: Landsat-TM/ETM+ images. For pixel-based validation, the estimated result was compared to the NFI data by leave-one-out process and R^2 = 0.40 and RMSE = 54.29 tons/hm^2. For county-scale validation, the result was verified by the intensive forest sub-compartment data of eight county and R^2 = 0.80 and RMSE = 34.26 tons/hm^2.

Index Terms*—k*-NN method, forest above-ground biomass, optimized configuration.

1. INTRODUCTION

Describing and quantifying forest AGB has become of importance to many scientific and societal tasks such as sustainable forest management, timber management, forest ecosystem productivity estimation, carbon sink evaluation, and studies of the role of forest in the global carbon cycle, and links between hydrology and ecology, etc. In conventional techniques on basis of statistical assessment (i.e., tree species, vertical structure, stand height, and stand density), the forest AGB information comes from expensive and time consuming field surveys by the high sampling intensity [1]. As an alternative, remote sensing is a valuable tool for estimating forest AGB in remote areas [2]. Two methods exist for establishing (calibrating) the relation between remote sensing data and forest AGB with ancillary data: parametric and non-parametric. The parametric method is conceptually simple, but the success largely depends on the statistical robustness of the relationships. In reality a change in AGB rarely directly results in a change of remote sensing signatures. Consequently, the parametric method applied to remote sensing data usually fails to map forest AGB satisfactorily. The non-parametric method (i.e., *k*-Nearest Neighbors, *k*-NN) is based on more flexible assumptions than the parametric method, and does not suffer from the same limitations [3].

This study investigated the performances of *k*-NN method for estimating forest AGB by use of Landsat TM/ETM+ data over a complex forest area, Jilin province. The objectives of this study are (1) to evaluate different configurations of the *k*-NN algorithm for forest AGB retrieval at pixel level and county scale, (2) to identify the best *k*-NN algorithm for routine application of forest AGB mapping in a heterogeneous forest over Jilin.

2. STUDY AREA AND DATASET

As one of the key forestry provinces in China, Jilin province was chosen to be study area (40 ° 52 '~ 46 ° 18' N, 121 ° 38' ~ 131 ° 19' E). The geographic and geomorphic conditions in Jilin province are significantly different. Its topography is sloping from southeast to northwest and shows clear signs that the south east is high and the northwest is low. Jilin province is located east of mid-latitude Eurasia, has a temperate continental monsoon climate and four distinctive seasons, with hot rainy season. It is dry and windy in spring, hot and rainy in summer, high sky and fine weather in autumn and long cold in winter. The average temperature is below -11 ℃ in winter, The annual average precipitation is 400 to 600 mm.

As a pilot province for ecological construction, the province's forest coverage is more than 40% and it has abundant forest resources, mainly composed of natural forest (i.e., *Pinus koraiensis, Picea asperata, Abies fabri, Larix olgensis*), secondary forest (i.e., *Picea asperata, Abies fabri, Quercus mongolica, Betula platyphylla, Populus ussuriensis, Tilia amurensis, Juglans mandshurica*) and plantation (i.e., *Larix gmelinii, Populus ussuriensis*).

The Landsat-TM/ETM+ images were downloaded from USGS website: http://glovis.usgs.gov/. Totally, 19 frames level.1 products were obtained and the acquired time was from 2004 to 2005. These products were further processed by radiometric correction, atmospheric correction (FLAASH) (See Fig.1). Preparing for running *k*-NN, the land cover map was generated from these images by supervised method. Validated by the 7th NFI data (8870 samples) investigated from 2003 to 2004, the overall accuracy of above forest/non-forest map was 92.35% and Kappa coeffient was 0.87. The NFI data was also used to calculate the forest AGB based on the tree biomass equations.

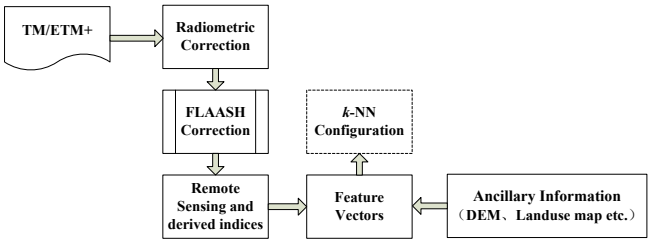

Fig.1 Preparation of spectral feature vectors for *k*-NN configuration

3. METHODOLOGY

The *k*-NN method is one of the most widely known and routinely utilized non-parametric estimation methods for updating different scale forest inventories, even in complex landscapes [4, 5]. The attractiveness of the method is that it does not depend on the assumptions on the nature of the relationships linking ground reference forest plot measurement to the spectral data. Using the forest inventory data, parameters such as total wood volume, biomass, and age can be calculated for every pixel within the forest area. The estimated forest parameter value for a specified pixel (V_p) is calculated as a weighted mean value of the parameter's measurements of the ith reference pixel (pi) ($V_{pi,p}$) at the k nearest samples in spectral space $\{i_1(p)...i_k(p)\}$:

$$V_p = \sum_{i=1}^{k} W_{pi,p} V_{pi,p} \quad (1)$$

where $W_{pi,p}$ are the weights assigned to each of the k samples proportional to the inverse squared distance ($d_{pi,p}$) between the pixel to be estimated and the reference plot:

$$W_{pi,p} = \frac{1/d_{pi,p}^2}{\sum_{j=1}^{k}\left(1/d_{pj,p}^2\right)} \quad (2)$$

and $d_{pi,p}$ is multidimensional distance between target and reference pixel.

For each frame Landsat-TM/ETM+ image, the *k*-NN method was applied in 1800 different configurations, varying both the mathematical formulation and the remote sensing feature input (Table 1). Then, the quality of the performance of each configuration was evaluated by Leave One Out (LOO) cross-validation against the NFI data.

Table 1 *k*-NN configurations used in this study

k value	Distance Measures	Feature Type	Feature Extraction Method
1-25	ED, MD, FD	(1) 6 TM/ETM+ bands (1~5 and 7 band) (2) 6 TM/ETM+ bands, IRI, NDVI (3) 3 PC bands, (4) 3 PC bands, IRI, NDVI	Pixel-wise, 3 x 3, 5 x 5,7 x 7, 9 x 9, 11 x 11

The feature types that were varied included different original spectral information, principle components (PC), derived indices such as Infrared Index (IRI), Normalized Difference Infrared Index (NDVI). The ED, MD, FD distance measures are Euclidean Distance, Mahalanobis Distance and Fuzzy Distance respectively [6, 7]. The first one is ED, the most used and simplest distance is defined as:

$$d_{pi,p}(ED) = \sum_{t=1}^{T}(x_{pit} - x_{pt})^2 \quad (3)$$

where x_{pit} and x_{pt} are the value of the t^{th} feature space variable for pixel p and for reference plot respectively.

The second distance is MD. Contrary to ED, MD takes correlations between variables and variances of the variables into account, and is scale-invariant. The variance-covariance matrix of the feature space variables, C, is used to correct the factor of the multicollinearity of the feature space variables:

$$d_{pi,p}(MD) = (x_{pi} - x_p)'C^{-1}(x_{pi} - x_p) \quad (4)$$

where x_{pi} and x_p are the feature space vector for target and reference pixels respectively.

The third distance, FD, enhances the importance of the most informative bands for the specific parameters to be estimated, which is the modification of MD where the variance-covariance matrix is computed in a fuzzy way.

$$d_{pi,p}(FD) = (x_{pi} - x_p)'C^{*-1}(x_{pi} - x_p) \quad (5)$$

where C^* is the fuzzy variance-covariance matrix,

$$C^* = \frac{\sum_{j=1}^{N} Fz_j (X_j - M^*)(X_j - M^*)'}{\sum_{j=1}^{N} Fz_j} \quad (6)$$

where N is number of training pixels, X_j is spectral vector of training pixel j and M^* is fuzzy mean spectral vector of all training pixels,

$$M^* = \frac{\sum_{j=1}^{N} Fz_j X_j}{\sum_{j=1}^{N} Fz_j} \quad (7)$$

and Fz_j is membership grade of each reference pixel j,

$$Fz_j = (2\pi)^{-1/2} D_z^{-1} e^{-1/2(Z_i - M_z)^2/D_z} \quad (8)$$

where Z_i is value of the parameter at training pixel j, M_z is mean value of the parameter and D_z is the standard deviation of the parameter.

Five different compositions were used: (1) the 6 bands TM/ETM+ spectrum, (2) feature type (1) plus the most

sensitive ancillary information: the IRI and NDVI, which was determined by stepwise multiple linear regression process from all the ancillary information (including IRI, NDVI, DEM and etc.), (3) the most significant PC and ancillary information, (4) feature type (3) plus the IRI and NDVI. The first three PC of the TM/ETM+ images were selected because they can describe more than 95.00% of the original spectral variance.

All factors (feature types, multidimensional distance measures, number of *k* nearest neighbors, and the extraction methods) affected the performance of *k*-NN estimates [8]. The optimal *k*-NN configuration was decided by the Pearson correlation index (R) and root mean square error (RMSE) between *k*-NN estimates and NFI data on the basis of LOO procedure.

4. RESULTS AND DISCUSSION

The optimized *k*-NN was applied to estimate the forest AGB for each frame TM/ETM+ image. The mosaic of overall estimation results was shown in Fig.2 and the validation was plotted in Fig.3.

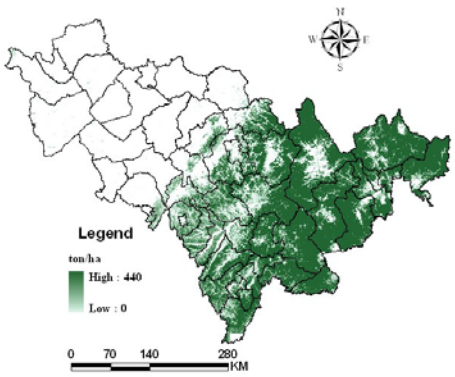

Fig.2 Mosaic retrievals of forest above-ground biomass over Jilin province

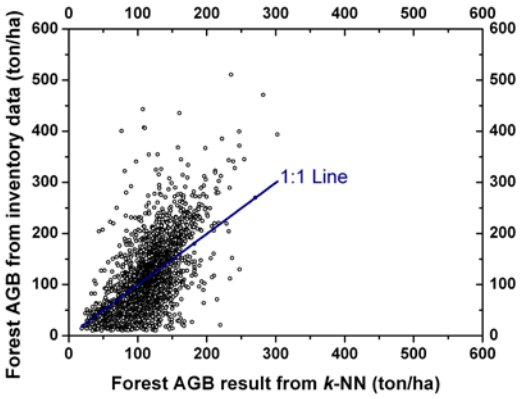

Fig.3 Validation of forest above-ground biomass estimates against forest inventory data
(R^2 = 0.40， RMSE = 54.29 ton/ha)

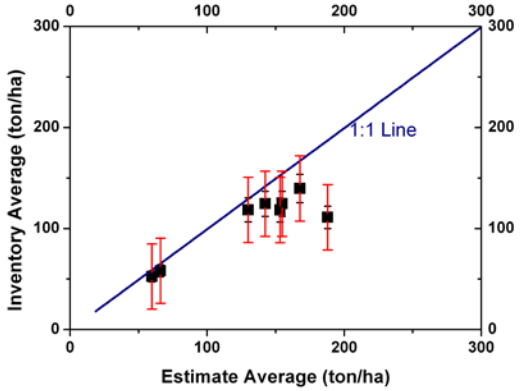

Fig.4 Validation of forest above-ground biomass estimates at county scale
(R^2 = 0.80， RMSE = 34.26 ton/ha)

Tab.2 Statistics of *k*-NN estimates at county scale

County	Plot Average	Estimates Average	Relative Accuracy
Baihe	154.74	124.51	80.47%
Dashitou	142.41	124.51	87.43%
Daxinggou	130.01	118.46	91.12%
Dongfeng	66.00	58.19	88.16%
Dongliao	59.76	52.50	87.86%
Dunhua	167.68	139.63	83.27%
Hongshi	187.93	111.09	59.11%
Huichun	153.21	118.3	77.21%

The quality of the performance of each configuration was evaluated by LOO cross-validation against ground measurements. For each frame TM/ETM+, the comparsion showed that the inclusion of more informative variables does not necessarily make the estimation better. The comparison of configurations of mathematical setup showed that the pixel-wise extraction method performed consistently worse than the bigger window (such as 3x3, 5x5) extractions. When $k > 5$, there seems no clear relationship between the value for *k* and estimation performance. Higher values of *k* imply averaging over a larger number of ground measurements. This results in values for AGB that move away from the extreme values in the ground dataset. This averaging effect becomes smaller with increasing *k*. The comparison of distance measures also showed that MD and FD outperformed the simplest ED measure (for k>1). Normally, small modification of the variance-covariance matrix normally could enhance the importance of the most informative input variables without hindering the stability of the response variables. However, when the information content of the feature space variables is relatively uniform, the improvement becomes marginal. In this study large differences between FD and MD occurred for values of *k* between 2 and 5, suggesting that the information content of the feature space was inhomogeneous.

At pixel scale, about 3400 NFI plots data was used to validate the retrievals, the overall R^2 was 0.40 and RMSE was 54.29 ton/ha, significant at 0.01. At county scale, the addtional intensive forest sub-compartment plots data (about 12000 plots) which was investigated during 2007 to 2010 were averaged at 8 counties, and then the mean values were compared to the corresponding forest AGB average retrievals at each county. It showed that the accuracy was largely improved, with the overall R^2 of 0.80 and RMSE of 34.26 ton/ha, significant at 0.01 (see Fig.4).

The pixel-wise validation process discovered that the *k*-NN underestimated the forest AGB for a large number of forest plots. It can be explained by that the signals of optical remote sensing data are easy to be saturated at low forest AGB level, or the terrain conditions affected the remote sensing signals to some extent.

To improve the remote sensing signal saturated level on forest AGB retrivals, further study can focus on integrating the satelitte LiDAR (i.e. ICE-GLAS) data and polarimetric SAR (i.e. ALOS PALSAR) data by use of above optimized *k*-NN method. For terrain effects, with the DEM information support, SCS+C model was found to be able to alleviate it [9]. Currently, the ASTER GDEM products can provide similar resoultion DEM information to these satellite data. With the same resolution to Landsat TM/ETM+ data, it is our future interest to reduce the terrain impact by use of ASTER GDEM.

With comparsion, the parametric estimation is easier to perform than *k*-NN, but it requires that a statistical relationship between the measurements and feature space variables exists. In Northeast China, the statistical relationship of the forest AGB measurements and remote sensing observation is rather variable, generally would bring out poor results for the regression. The *k*-NN method does not suffer from the same limitations, which makes it more robust in complex environments, and reasonable results can be achieved.

5. CONCLUSION

This study explores the predictive power of Landsat TM/ETM+ data for the retrieval of forest AGB over Jilin province by opimizing the non-parametric *k*-NN algorithm (varying both feature inputs and mathematical factors).

As a whole, the overall R and RMSE were satisfying and then it is concluded that the non-parametric method with Landsat-TM/ETM+ data is able to map forest AGB operatively over the Northeast China, where has the heterogeneous forest (i.e., the various terrain and environmental condition, tree species, density). The optimizing strategy used in this study might be more applicable to areas where forests are more homogenous. There is a real prospect that the non-parametric forest AGB results obtained in this study encourage further interests to investigate other forest attributes (i.e., basal area and LAI) with this optimized strategy.

6. ACKNOWLEDGEMENT

This study is financially supported by National Natural Science Foundation of China (41101379), National 973 Program (2013CB733404) and National 863 Program of China (2011AA120405).

7. REFERENCES

[1] P. Schroeder, S. Brown, J.M. Mo, R. Birdsey and C. Cieszewski, “Biomass estimation for temperate broadleaf forests of the US using inventory data”, *Forest Science*, Vol.43, pp.424-434, 1997.

[2] J. E.Luther, R. A. Fournier, D. E. Piercey, L. Guindon, and R.J. Hall, “Biomass mapping using forest type and structure derived from Landsat TM imagery”, *International Journal of Applied Earth Observation Geoinformation*, Vol.8, pp.173-187, 2005.

[3] X. Tian, Z. Su, E.X. Chen, Z.Y. Li, C. van der Tol, J.P. Guo, Q.S. He, “Estimation of forest above-ground biomass using multi-parameter remote sensing data over a cold and arid area”, *International Journal of Applied Earth Observation and Geoinformation*, Vol.14, pp.160-168, 2012.

[4] F. Maselli, G. Chirici, L. Bottai, P. Corona, M. Marchetti, “Estimation of Mediterranean forest attributes by the application of *k*-NN procedures to multitemporal Landsat ETM+ images”. I*nternational Journal of Remote Sensing,* Vol.26, pp.3781-3796, 2005.

[5] F. Maselli, M. Chiesi, “Evaluation of statistical methods to estimate forest volume in a mediterranean region. *IEEE Transactions on Geoscience and Remote Sensing*, Vol.44, pp.2239-2250, 2006.

[6] F. Maselli, “Extension of environmental parameters over the land surface by improved fuzzy classification of remotely sensed data”, *International Journal of Remote Sensing*, Vol.22, pp.3597-3610, 2001.

[7] F. Maselli, M. Chiesi, “Evaluation of statistical methods to estimate forest volume in a mediterranean region”, *IEEE Transactions on Geoscience and Remote Sensing*, Vol.44, pp.2239-2250, 2006.

[8] G. Chirici, A. Barbati, P. Corona, M. Marchetti, D. Travaglini, F. Maselli and R. Bertini, “Non-parametric and parametric methods using satellite imagery for estimating growing stock volume in alpine and Mediterranean forest ecosystems”, *Remote Sensing of Environment*, Vol.112, pp.2686-2700, 2008.

[9] S.A. Soenen, D.R. Peddle and C.A. Coburn, “SCS+C: A modified Sun-Canopy-Sensor topographic correction in forested terrain”, *IEEE Transaction on Geoscience and Remote Sensing*, Vol.43, pp.2148-2159, 2005.