A GEOBIA APPROACH TO ESTIMATE LARGE AREA FOREST CANOPY HEIGHT USING LIDAR TRANSECTS AND QUICKBIRD IMAGERY

G. Chen [a], *, G. J. Hay [a]

[a] Foothills Facility for Remote Sensing and GIScience, Department of Geography, University of Calgary
Calgary, T2N 1N4 Canada - (gangchen, gjhay)@ucalgary.ca

Commission VI, WG IV/4

KEY WORDS: GEOBIA, Canopy height, Lidar transect, Machine learning, Forest

ABSTRACT:

Numerous applications of small footprint airborne lidar (light detection and ranging) have provided highly accurate results for estimating forest height. However, the associated acquisition cost remains high, which limits its use for wall-to-wall large area mapping. In this study, we developed a novel framework by integrating GEOBIA (*GEOgraphic Object-Based Image Analysis*), lidar transects and Quickbird imagery to estimate large area canopy height. Model results (from eight different lidar transect combinations in two different directions, N-S and W-S) were compared with the corresponding canopy height from the full lidar scene. Results show that the highest correlation ($R = 0.85$) was achieved using a lidar transect cover of 7.6% of the full scene (i.e., two transects in N-S direction), while the lowest correlation ($R = 0.75$) was obtained from a lidar transect cover of 3.8%.

1. INTRODUCTION

Remote sensing techniques have been widely used to estimate large-area forest vertical structure. Among these techniques, lidar (light detection and ranging) has proven the ability to accurately characterize forest canopy height (Means et al., 1999). However, an important issue associated with lidar is that small-footprint airborne lidar acquisition remains expensive, which limits wall-to-wall lidar mapping over large forest areas. Compared to lidar data of similar extent and resolution, optical imagery (even high-resolution data) are less costly to acquire. Therefore, an integration of lidar transects and optical imagery provides a potential solution to estimate forest canopy height for large areas. To date, there have been only a few studies on this subject (Hudak et al., 2002; Wulder and Seemann, 2003; Hilker et al., 2008). Though promising results have been reported, two critical issues have rarely been discussed: (i) the selection of appropriate lidar transect features (e.g., location, direction and cover); and (ii) the investigation of the relationship between optical and lidar variables. Specifically, (i) lidar transects were arbitrarily selected in previous studies, without considering forest height variability of the entire area, which could possibly reduce model robustness; and (ii) the commonly used multiple regression approach may fail to capture the complex relationship between the two data types for forest studies, which requires the use of techniques with strong generalization capacity, such as machine learning methods.

GEOgraphic Object-Based Image Analysis (GEOBIA), a sub-discipline of Geographic Information Science (GIScience), has been through a dramatic development in recent years driven by the advent of high-spatial resolution imagery and the availability of powerful data processing software (Hay and Castilla, 2008; Blaschke, 2010). In remote sensing forest studies, GEOBIA has proven successful for decreasing internal spectral variance within each geographic object and therefore dramatically increase model performance in several aspects, such as characterizing tree biophysical parameters and classifying forest cover types (Addink et al., 2007; Yu et al., 2008).

Based on these ideas, the goal of this study was to develop a framework to accurately estimate large-area forest canopy height by integrating small-area airborne lidar transects and Quickbird imagery. To achieve this goal, (i) a GEOBIA approach was used to characterize canopy height at the small tree cluster level, as the (upper scale) stand-level estimates will be available directly by aggregating the tree-cluster-level results; (ii) the best lidar transect locations were determined by using a lidar transect selection algorithm, which was built upon our prior research (Chen and Hay, 2010). However, emphasis in this study was placed on evaluating the algorithm performance in mixed forests with a larger extent; and (iii) we further investigated the potential of using two machine learning techniques – a *minimal-redundancy-maximal-relevance* (mRMR) method and *support vector regression* (SVR), to select suitable model variables and further generalize forest vertical information from lidar transects to the large study area.

2. DATA COLLECTION

2.1 Study area

Our study site is located at the *Training and Research Forest of Lake Duparquet* (TRFLD), Quebec, Canada. The study area is 16,330 ha (14.2 × 11.5 km) and is characterized by south-east boreal forests with an abundance of mixed stands. This study site is dominated by balsam fir (*Abies balsamea* L. [Mill.]), along with small proportions of white spruce (*Picea glauca* [Moench] Voss), black spruce (*Picea mariana* [Mill] B.S.P.), white birch (*Betula paprifera* [Marsh.]), trembling aspen (*Populus tremuloides* [Michx]), and jack pine (*Pinus banksiana* Lamb.). The remainder of the site is composed of clearcuts, roads, rivers and lakes.

---

* Corresponding author.

2.2 Field data

Field data were collected during the summer of 2003. A number of forest stands were visited and canopy height was defined from 37 field plots. Most of these plots were measured using a fixed size of 20 × 20 m; however, a plot size of 10 × 10 m was also used in several dense and uniform stands, where the two types of plot sizes would produce similar results.

2.3 Lidar data

Lidar data were acquired from August 14 to 16, 2003, by a discrete-return Optech ALTM2050 system. This mission was carried out at a flying attitude of 1000 m, with a pulse repetition frequency of 50 kHz, a beam divergence of 0.2 mrad, and a maximum scale angle of 15° (i.e., swath width of 540 m). First and last returns were recorded, with average densities of 3.0 and 0.2 hit(s)/m^2, respectively. These were used to generate digital surface and elevation models. A forest canopy height model (CHM) was derived from them at a 1.0 m resolution.

2.4 Quickbird (QB) data

QB imagery of the study site were acquired on June 13, 2003. Four multispectral bands [i.e., blue, green, red and near infrared (NIR)] and one panchromatic band were used in this study. To increase the spatial resolution while maintaining the multispectral information, a principal components spectral sharpening technique (Welch and Ahlers, 1987) was used to fuse the QB multispectral bands with the panchromatic band. The pan-sharpened QB image was then resampled to 1.0 m, consistent with the lidar CHM. The QB image was then geometrically co-registered to the lidar CHM using 60 tie points, yielding a RMSE of 0.9 m.

3. METHODOLOGY

3.1 Image segmentation and tree type classification

To obtain meaningful forests image-objects, Definiens Developer 7.0 (Definiens Imaging GmbH, Munich, Germany) was applied to segment the pan-sharpened multispectral Quickbird imagery using the multiresolution segmentation algorithm. Two parameters of shape and compactness control the characteristics of similarity and heterogeneity for each image-object. In this study, we used the software default value of 0.1 for shape; while the compactness parameter was set as 0.8 to obtain smooth boundaries for forest objects. All four spectral bands were assigned the same weight during the segmentation. The scale parameter was adjusted to derive image-objects at the small tree cluster level, with the mean object size (MOS) of 0.04 ha. All image-objects were further classified into three categories: conifer, deciduous and non-forest objects. This step was accomplished by applying the hierarchical classification algorithm in Definiens Developer 7.0 using the QB four spectral bands.

3.2 Canopy-object pseudo-height classification

The purpose of this step is to simulate the canopy height variability of the entire study area by using QB imagery only, from which the lidar transect features can be appropriately determined in the next step (see Section 3.3). Three types of QB-derived object-based variables have proven the potential of estimating lidar-measured canopy height in our previous study (Chen et al., 2009): (i) mean spectral bands; (ii) image-texture

for spectral bands, which includes internal-object texture and geographic object-based texture (GEOTEX); and (iii) shadow fraction. Then, 14 height classes were generated using an unsupervised classification algorithm of ISODATA in ENVI (ITT Visual Information Solutions, Colorado, USA).

3.3 Lidar transect selection

In this study, lidar transect selection requires the decision of three important transect features: (i) cover, (ii) direction and (iii) location. To determine appropriate features, the canopy pseudo-height classification result (derived from Section 3.2) was used as a proxy for forest height class variability. Since a small area lidar acquisition represents a relatively smaller cost, only four types of small-area lidar cover were evaluated in this study using 1-4 transect samples. Two directions of N-S (north-south) and W-E (west-east) were further assessed. This resulted in lidar transect cover (as a percentage of the entire scene area) of 3.8%, 7.6%, 11.4% and 15.2% in N-S direction, and 4.7%, 9.4%, 14.1% and 18.8% in W-E direction. The best transect location(s) for each combination of cover and direction was determined by using the rules developed by Chen and Hay (2010).

3.4 mRMR selecting variables

Three types of QB-derived variables (the same variables as those used in Section 3.2) were used to link QB data with lidar-measured canopy height within the transect-covered areas. To determine an appropriate variable subset from all input variables, a machine learning technique named the *minimal-redundancy-maximal-relevance* (mRMR) approach (Peng et al., 2005) was applied to select the best subset of six variables from a total of 13 variables. These were used separately to model the canopy height for conifer and deciduous tree classes.

3.5 Support vector regression (SVR) modelling

In this study, SVR was applied to develop models estimating lidar-measured canopy height for both conifers and deciduous trees. An SVM open source software of LIBSVM was used to perform the modelling (Chang and Lin, 2001), with the best parameter combination found at C (penalty parameter) = 8.0, ε (precision parameter) = 0.5 and γ (kernel parameter) = 1.0.

3.6 Estimation of forest height

Nonlinear models, which have been widely used to estimate forest biophysical parameters using lidar data in previous studies (Næsset, 1997; Means et al., 1999; Lim et al., 2003), were employed to build a relationship between the previously estimated canopy height and the field measurements. To compare the estimation results derived from our framework with those using the full lidar scene, (i) a nonlinear model was developed to build the relationship between field measurements and actual data; and (ii) the difference between these two types of estimation results (using our framework versus actual lidar data) was evaluated using correlation (R) and RMSE.

4. RESULTS AND DISCUSSION

4.1 Spatial distribution of the selected transects

By applying the lidar transect selection algorithm in this area, Figure 1 shows the results illustrating four transect

combinations and canopy height histograms using lidar cover of (1) 3.8%, (2) 7.6%, (3) 11.4%, and (4) 15.2% in N-S direction. It should be noted that these transect locations were considered as the best for each type of lidar cover in this study, based on our lidar transect selection algorithm (Section 3.3). To further ascertain whether these algorithm selected transects can well model the full scene canopy height variability, their canopy height histograms [Figures 1(1b) – (4b)] were compared with the canopy height histogram derived from the entire lidar dataset [Figure 1(5b)]. Clearly, the comparison shows that all histograms are highly correlated and have a similar trend. This confirms that the transect locations for different types of lidar cover were well selected. Since lidar transects selected in the W-E direction had a similar condition, they are not presented in this paper.
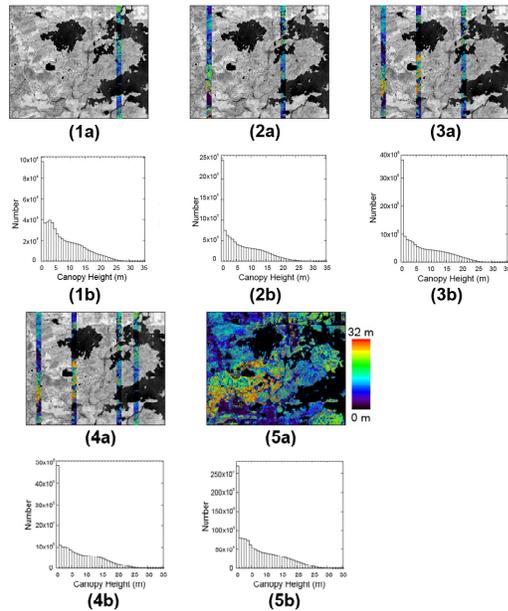
Figure 1. (1a) – (5a) shows four (N-S direction) lidar transect combinations - and associated coverage - (1a) 3.8%, (2a) 7.6%, (3a) 11.4%, (4a) 15.2% and (5a) 100.0% - derived from the transect selection algorithm (section 3.3). For illustrative purposes, the QB image was used as the base layer (B/W) with lidar transects overlaid (colour). (1b) – (5b) illustrate the canopy height histograms derived from the corresponding lidar transect data in (1a) – (5a).

4.2 Model performance

To evaluate the transect model performance with that using all lidar data, Figure 2 shows the scatterplots of estimated canopy height (using lidar transects and Quickbird data) versus the canopy height (using full-cover lidar data). Figure 2 shows that small lidar transect cover cannot produce high model performance, as the training samples may not be large enough. For example, Figures 2(1) used only one transect representing a lidar cover of 3.8% in N-S direction, where relatively low correlations ($R = 0.75$) and high errors (RMSE = 4.16 m) were located. With the increase of lidar transect cover, the model performance for estimating forest canopy height increased as well. However, the correlation changed in a relatively small range (i.e., between 0.81 and 0.85). This could be explained in two ways: (i) the amount of lidar transect data (i.e., training samples) were large enough to develop relatively robust models; and (ii) the machine learning SVR approach had a strong
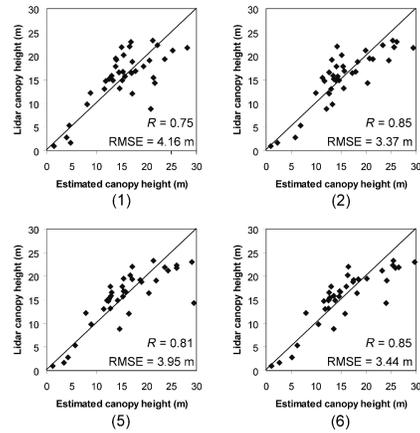
Figure 2. Scatterplots of estimated canopy height (using lidar transects and Quickbird data) versus lidar canopy height (using full-cover lidar data). (1) - (4) represent lidar transect cover in N-S direction: (1) 3.8%, (2) 7.6%, (3) 11.4% and (4) 15.2%.

generalization ability, which facilitated the use of relatively small lidar transect cover to estimate large-area canopy height. The best performance (i.e., highest correlation and lowest error) of our framework to estimate canopy height ($R = 0.85$; RMSE = 3.37 m) is shown in Figure 2(2), where two lidar transects were selected that represent a lidar cover of 7.6% in N-S direction. Additionally, Figure 2 also reveals that our framework tended to overestimate the (lidar-modelled) forest canopy height, for tree canopies that were lower than 5 m or taller than 20 m, although most field-measured canopies were located between this height range and the average error (RMSE = 3.37 m) was lower than the ~ 5.0 m forest inventory height class interval used in this area.

5. CONCLUSION

In this study, we developed a novel framework by integrating GEOBIA, small-area lidar transects and Quickbird imagery to estimate large-area forest canopy height. By using optical Quickbird imagery and the lidar transect selection algorithm, we first determined the best lidar transect locations for eight different lidar samples in two transect directions. Although our study area was dominated by mixed forest stands, the algorithm selected transects well modeled similar canopy height variability as that using all lidar data. To further generalize the accurate canopy height information from lidar transects to the large study area, two machine learning approaches – minimal-redundancy-maximal-relevance (mRMR) and support vector regression (SVR) were used. Based on a comparison between the canopy height estimation performance using our framework with that using the full-scene lidar data, the highest correlation and lowest error ($R = 0.85$; RMSE = 3.37 m) were derived using a lidar transect cover of 7.6% (i.e., two transects) in N-S direction. With the GEOBIA approach, all estimates in this study were made at the small tree cluster level of 0.04 ha, which is similar to a typical field plot size. However, the geo-objects derived from GEOBIA can better delineate tree clusters with similar internal characteristics, rather than using arbitrarily designed square/circular plots. Larger object sizes were not evaluated in this study, as the within-object variability decreases

with the increase of object sizes, which would definitely result in the loss of forest variability within large geo-objects.

REFERENCES

Addink, E.A., de Jong, S.M., & Pebesma, E. J., 2007. The Importance of Scale in Object-based Mapping of Vegetation Parameters with Hyperspectral Imagery. *Photogrammetric Engineering and Remote Sensing*, 73, pp. 905-912.

Blaschke, T., 2010. Object based image analysis for remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65, pp. 2-16.

Chen, G., Hay, G.J., Castilla, G., St-Onge, B., & Powers, R., 2009. A multiscale geographic object-based image analysis (GEOBIA) to estimate lidar-measured forest canopy height using Quickbird imagery. *International Journal of Geographical Information Science*. In Review.

Chen, G., & Hay, G.J., 2010. Modeling large-area forest canopy height using lidar transects, Quickbird data and geographic object-based image analysis (GEOBIA). *Remote Sensing of Environment*. In Review.

Hay, G.J., & Castilla, G., 2008. Geographic Object-Based Image Analysis (GEOBIA). In Blaschke, T., Lang, S., & Hay, G. J. (Eds.), *Object-Based Image Analysis - Spatial concepts for knowledge-driven remote sensing applications* (pp. 77-92). Berlin: Springer-Verlag.

Hilker, T, Wulder, M.A., & Coops, N.C., 2008. Update of forest inventory data with lidar and high spatial resolution satellite imagery. *Canadian Journal of Remote Sensing*, 34, pp. 5-12.

Hudak, A.T., Lefsky, M.A., Cohen, W.B., & Berterreche, M., 2002. Integration of LIDAR and Landsat ETM+ data for estimating and mapping forest canopy height. *Remote Sensing of Environment*, 82, pp. 397-416.

Lim, K., Treitz, P., Baldwin, K., Morrison, I., & Green, J., 2003. Lidar remote sensing of biophysical properties of tolerant northern hardwood forests. *Canadian Journal of Remote Sensing*, 29, 658-678.

Means, J.E., Acker, S.A., Harding, D.J., Blair, J.B., Lefsky, M.A., Cohen, W.B., Harmon, M.E., & McKee, W.A., 1999. Use of large-footprint scanning airborne lidar to estimate forest stand characteristics in the Western Cascades of Oregon. *Remote Sensing of Environment*, 67, pp. 298-308.

Næsset, E., 1997. Estimating timber volume of forest stands using airborne laser scanner data. *Remote Sensing of Environment*, 14, pp. 175–181.

Peng, H., Long, F., & Ding, C., 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27, pp. 1226-1238.

Welch, R., & Ahlers, W., 1987. Merging Multiresolution SPOT HRV and Landsat TM Data. *Photogrammetric Engineering and Remote Sensing*, 53, pp. 301-303.

Wulder, M. A., & Seemann, D., 2003. Forest inventory height update through the integration of LIDAR data with segmented Landsat imagery. *Canadian Journal of Remote Sensing*, 29, pp. 536-543.

Yu, Q., Gong, P., Tian, Y.Q., Pu, R.L., & Yang, J., 2008. Factors Affecting Spatial Variation of Classification Uncertainty in an Image Object-based Vegetation Mapping. *Photogrammetric Engineering and Remote Sensing*, 74, pp. 1007-1018.