International Journal of Applied Earth Observation and Geoinformation 15 (2012) 28–37

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag

A GEOBIA framework to estimate forest parameters from lidar transects, Quickbird imagery and machine learning: A case study in Quebec, Canada

Gang Chen [a,*], Geoffrey J. Hay [a], Benoît St-Onge [b]

[a] Foothills Facility for Remote Sensing and GIScience, Department of Geography, University of Calgary, 2500 University Dr., Calgary T2N 1N4, Canada
[b] Department of Geography, Université du Québec à Montréal, C.P. 8888, succ. Centre-Ville, Montreal H3C 3P8, Canada

ARTICLE INFO

Article history:
Received 4 October 2010
Accepted 18 May 2011

Keywords:
GEOBIA
Geo-intelligence
Forest parameters
Lidar transects
Quickbird
Machine learning

ABSTRACT

The *GEOgraphic Object-Based Image Analysis* (GEOBIA) paradigm continues to prove its efficacy in remote sensing image analysis by providing tools which emulate human perception and combine analyst's experience with meaningful image-objects. However, challenges remain in the evolution of this new paradigm as sophisticated methods attempt to deliver on the goal of automated geo-intelligence (i.e., geospatial content within context) from geospatial sources. In order to generate geo-intelligence from a forest scene, this article introduces a GEOBIA framework to estimate canopy height, above-ground biomass (AGB) and volume by combining lidar (light detection and ranging) transects, Quickbird imagery and machine learning algorithms. This framework is comprised three main components: (i) image-object extraction, (ii) lidar transect selection, and (iii) forest parameter generalization. The rational for integrating these methods is to provide a semi-automatic GEOBIA approach from which detailed forest information is obtained at the *individual tree crown or small tree cluster* level (i.e., mean object size of 0.04 ha); while also dramatically reducing airborne lidar data acquisition costs. Analysis is performed over a 16,330 ha forested study site in Quebec, Canada. Forest parameter estimation results derived from our GEOBIA framework demonstrate a strong relationship with those using the full lidar cover; where the highest estimates for canopy height ($R = 0.85$; RMSE = 3.37 m), AGB ($R = 0.85$; RMSE = 39.48 Mg/ha) and volume ($R = 0.85$; RMSE = 52.59 m^3/ha) were achieved using a lidar transect sample representing only 7.6% of the total study area.

1. Introduction

Remote sensing techniques allow for the collection of Earth surface information over a range of scales in a synoptic and timely manner (Wulder, 1998). Today, high spatial resolution (i.e., H-res pixels generally less than or equal to 5.0 m) remote sensing data are rapidly accessible from a variety of sources, such as satellite-based optical sensors and airborne lidar (light detection and ranging) systems. Over the last decade, the development of new image processing techniques increasingly referred to as *GEOgraphic Object-Based Image Analysis* (GEOBIA) (Hay and Castilla, 2008) have proven effective for analyzing high resolution data by incorporating analyst's experience, complimentary ancillary data, sophisticated geospatial analysis and methods that emulate the human perception of image-objects within a scene (i.e., based on size, shape, tone, color, texture, topology and context), rather than as isolated pixels of varying color (Hay and Castilla, 2008; Blaschke, 2010). However, the evolution of GEOBIA faces a growing challenge to develop semi/automated methods that bridge the gaps between straightforward segmentation – the extraction of image-objects – and the generation of *geo-intelligence* from geospatial sources. Here geo-intelligence refers to *geospatial content within context* (Hay and Blaschke, 2010).

As a dominant terrestrial sink for atmospheric CO_2, forests play an important role in the dynamics of the carbon cycle (Eamus and Jarvis, 1989). Similarly, precise forest management requires an accurate estimation of carbon content with an emphasis on above-ground biomass (AGB). To assess the commercial value of forests, volume is widely used to measure wood quantity; and an important parameter used to calculate AGB and volume, is canopy height. However, monitoring large-area forest parameters such as canopy height, AGB and volume, requires considerations of both accuracy and budget. Previous studies have proven promising to apply optical imagery and GEOBIA to retrieve forest parameters, such as forest height (Wulder et al., 2007; Mora et al., 2010), AGB (Addink et al., 2007; Kajisa et al., 2009), and volume (Mäkelä and Pekkarinen, 2001; Pekkarinen, 2002). Although it is cost effective estimating these parameters using only optical imagery, model accuracies are lower than those using airborne lidar data. To meet these challenges, recent research describes the combination of small-area lidar transects and wider extent optical imagery to provide cost-effective solutions. This is achieved by generalizing lidar-measured

* Corresponding author. Tel.: +1 403 210 8761; fax: +1 403 282 6561.
E-mail address: gangchen@ucalgary.ca (G. Chen).

1569-8432/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jag.2011.05.010

vertical canopy information from transects to the entire study site covered by an optical image (Hudak et al., 2002; Wulder and Seemann, 2003; Hilker et al., 2008; Stojanova et al., 2010; Chen and Hay, 2011). Recent studies (Chen and Hay, 2011, in press) have noted, that the ability to accurately extract this information depends on (i) the type of forest characteristics assessed, (ii) the ability to define appropriate lidar transects and (iii) the type of modeling and generalization methods used to relate transect samples back to the full scene.

Based on this brief background, the primary objective of this study is to present a GEOBIA framework to generate new forest geo-intelligence by estimating canopy height, AGB and volume from Quickbird imagery and airborne lidar transects. This framework builds upon prior research by incorporating three main components: (i) image-object extraction, (ii) lidar transect selection, and (iii) forest parameter generalization. Chen and Hay (2011, in press) first describe the use of a lidar transect selection algorithm and a *support vector regression* (SVR) generalization technique applied to a small (2601 ha) homogenous forest site (with two major tree species) in British Columbia, Canada. In this study, we build on this early work by presenting a more complete GEOBIA framework composed of one additional machine learning algorithm, and examine its performance over a larger (16,330 ha) more complex mixed forest site (with six major tree species), located in Quebec, Canada.

2. Data collection

2.1. Study area

Our 16,330 ha (14.2 km × 11.5 km) study site (48°30′N, 79°22′W) is located in the *Training and Research Forest of Lake Duparquet* (TRFLD), Quebec, Canada (Fig. 1), where it is characterized as a South-Eastern Boreal Forest composed of an abundance of mixed stands. The site is dominated by balsam fir (*Abiesbalsamea* L. [Mill.]), along with white spruce (*Piceaglauca* [Moench] Voss), black spruce (*Piceamariana* [Mill] B.S.P.), white birch (*Betulapaprifera* [Marsh.]), trembling aspen (*Populustremuloides* [Michx]), and jack pine (*Pinusbanksiana* Lamb.). The remainder of the site is composed of clearcuts, roads, rivers and lakes.

2.2. Field data

Field data were collected during the summer of 2003. A number of forest stands were visited and 37 plots were chosen to gather canopy height, DBH (diameter at breast height), species composition and stem density. A Panasonic SXBlue real-time differential GPS (Geneq, Montreal) was applied for plot positioning, with an average accuracy of 2–3 m under canopy. Most of the field plots were measured using a fixed size of 20 m × 20 m. However, the plot size of 10 m × 10 m was also used in several dense and uniform stands, where the two types of plot sizes would produce similar results. All measured canopy heights ranged from 1.4 m to 27.1 m, with an average height of 15.9 m and a standard deviation of 6.2 m. To further acquire AGB and volume at the plot level, biomass and volume equations were defined from the literature (Lambert et al., 2005; Perron, 2003) and used to calculate values at the individual tree level. AGB and volume per plot were then estimated by summarizing the individual values.

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. A forest canopy height model (CHM) was generated at a 1.0 m resolution. However, 1.0 m pixels do not represent the actual forest entities of interest – individual trees. To capture these basic forest objects, a *canopy height segmentation image* (CHS) was generated (from the 1.0 m resolution CHM) to describe forests at the *individual tree* crown *or small tree cluster* (hereafter, *crown/cluster*) level. This CHS was created by applying an algorithm provided by Chen et al. (in press), with the basic idea of adapting a watershed algorithm to delineate tree crowns and fill them with the average height values within the crown extents. Visual analysis revealed a strong spatial correlation between segmented tree crowns/clusters and individual trees and small tree groups. The latter being the case predominantly with deciduous trees, as their canopies appeared to overlap, and distinct crowns could not always be confidently isolated. Table 1 illustrates the proportion of each forest canopy height class in the study area; where the height classes were adopted from the local forest inventory height class codes.

2.4. Quickbird (QB) data

A QB scene of the study site was acquired on June 13, 2003 under clear skies. The off-nadir view angle was 14.0°. Four (2.44 m) multispectral bands [i.e., blue, green, red and near infrared (NIR)] and

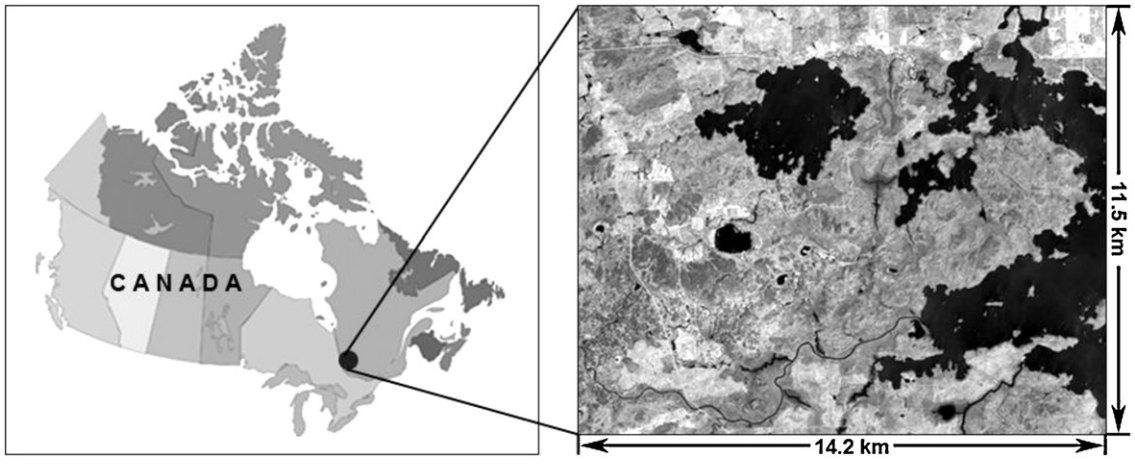

Fig. 1. Study area (left) located in the territory of the Training and Research Forest of Lake Duparquet (TRFLD), Quebec, Canada, with a Quickbird grayscale image (right) over the study site converted from a false color composite using near infrared (NIR), red and green bands.

Table 1
Proportion of each canopy height class in the study area, derived from the lidar CHS image and based on local forest inventory height classes.

Canopy height class (m)	Forest proportion (%)
1: 0.0–2.0	20.09
2: 2.1–4.0	15.18
3: 4.1–7.0	16.77
4: 7.1–12.0	19.49
5: 12.0–17.0	15.49
6: 17.1–22.0	9.13
7: >22.1	3.85

one (0.61 m) panchromatic band were used in this study. To exploit the combined multispectral and spatial content of this imagery, a principal components spectral sharpening technique (Welch and Ahlers, 1987) was used to fuse the multispectral and panchromatic bands. The pan-sharpened QB image was then resampled to a 1.0 m spatial resolution (the same as the CHS), and then geometrically co-registered to the CHS using 60 tie points, and yielding a RMSE of 0.9 m.

3. Data analysis

An important component of this project involves using optical imagery to generate 'pseudo-height' classes, from which to guide our selection and 'acquisition' of airborne lidar transects. This is based on research which describes useful relationships between optical imagery and canopy height (Franklin and McDermid, 1993; Hyde et al., 2006; Donoghue and Watt, 2006; Mora et al., 2010). Once transects are defined, forest height and species information (from the optical and lidar data covered by the transects) can be generalized to the entire study site covered by the optical imagery. To achieve this, the QB scene must first be segmented to derive image-objects at the small crown/cluster level. This is followed by an object-based classification to define conifer, deciduous and non-forest objects. Three types of variables are then extracted from the conifer and deciduous classes and used to simulate pseudo-height classes for each canopy type. By applying rules described in Chen and Hay (2011), optimal lidar transect locations and orientations can then be defined based on these pseudo-height classes.

To develop models within the transect-covered areas that link lidar-measured canopy height with the QB variables, two machine learning approaches mRMR (*minimal-redundancy–maximal-relevance*) and SVR (*support vector regression*) were evaluated. These provided appropriate model inputs from which transect canopy height was generalized to the entire site. Field measurements were then used to evaluate model performance by comparing the estimated forest parameters (i.e., canopy height, AGB and volume) with those using the full lidar scene. The flowchart in Fig. 2 summarizes these steps; while the following sub-sections provide greater details and explanations.

3.1. Image-object extraction

3.1.1. Image segmentation

Compared to medium- or low-resolution remote sensing imagery, high resolution data are able to better capture individual trees and detailed canopy variability. However, there are new challenges resulting from the enhanced spectral variability found in canopy and crown shadows. Fortunately, geo-object based image analysis provides suitable solutions to reduce these variations while still capturing meaningful forest characteristics. However, a more critical issue involves selecting an appropriate mean object size (MOS), or scales of analysis (Hay et al., 2001; Chen et al., in press). A large MOS (i.e., 2–20 ha) is able to simulate forest inventory polygons at the stand level (Wulder and Seemann, 2003); however,

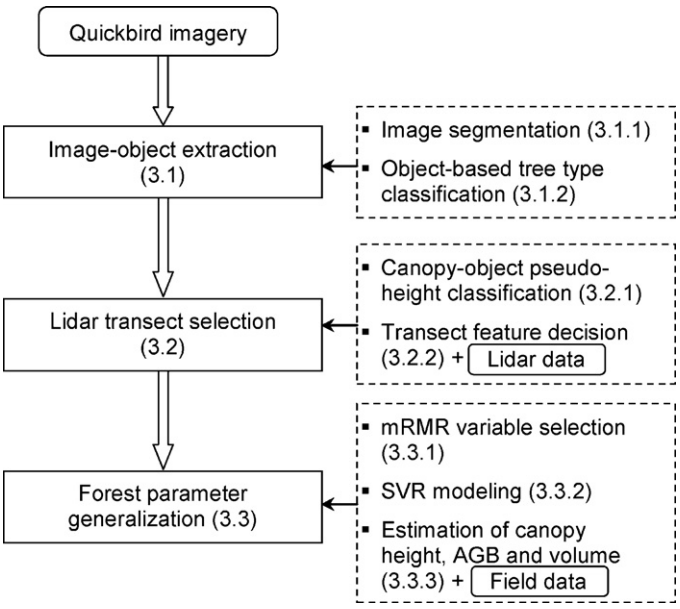

Fig. 2. Flowchart of the methodology process with reference to Section 3. mRMR is the abbreviation of *minimal-redundancy–maximal-relevance*, and SVR is *support vector regression*.

the merits of high resolution imagery may not be fully exploited, as detailed forest canopy variability tends to be ignored in large objects (Chen and Hay, in press). Conversely, a very small MOS may introduce large errors due to the difference of data acquisition geometries using lidar and optical remote sensing systems, especially in areas where large canopy height variation exists. As a result, Definiens Developer 7.0 (Definiens Imaging GmbH, Munich, Germany) was applied to segment the pan-sharpened multispectral Quickbird imagery at a relatively small MOS using its multiresolution segmentation algorithm. Two parameters of *shape* and *compactness* control the characteristics of similarity and heterogeneity for each image-object. In this study, the software default value of 0.1 was used for shape; while the compactness parameter was set at 0.8 to obtain smooth forest object boundaries. All four spectral bands were assigned the same weight during the segmentation. A scale parameter of 50 was used to derive image-objects at the crown/cluster level, with a MOS of 0.04 ha.

3.1.2. Object-based tree type classification

Popescu and Wynne (2004) recommend the development of separate models for conifer and deciduous trees, due to their distinct shapes and spectral reflectance. Therefore, all image-objects were classified into three categories: (i) conifer, (ii) deciduous and (iii) non-forest objects. This step was accomplished by applying the supervised nearest neighbor classification algorithm in Definiens Developer 7.0 using the four QB spectral bands. 50 object samples were manually selected for each class in this classification. To assess classification accuracy, 50 points were randomly extracted from each class and manually interpreted as the reference data. Due to a limitation of additional metadata (e.g., aerial images), these samples were also extracted from the Quickbird image; however, their interpretation was guided by the location of field plot data (see Section 2.2). Accuracy results are described in Section 4.1.

3.2. Lidar transect selection

3.2.1. Canopy-object pseudo-height classification

As its name suggests, the premise for using a lidar transect selection algorithm is to determine appropriate locations, from which lidar transects (i.e., height samples) should be collected, rather than

Table 2
Three types of variables derived from Quickbird imagery.

Variable type	Variable name	Description
Mean spectral bands	DN*i*	Average of DNs for the *i*th[a] band within image-objects
Image-texture	TXIT*i*	Internal standard deviation for the *i*th band within image-objects
	GEOTEX*i*	Neighboring standard deviation for the *i*th band within neighboring image-objects
Shadow fraction	SF	A quotient of the size of shadow areas and the size of corresponding entire image-objects

[a] *i* is the band number (i.e., 1 – blue band, 2 – green band, 3 – red band and 4 – NIR band).

collecting wall-to-wall coverage. To facilitate this, a base 'height' map needs to be created to define the canopy height variability of the entire study site. It should be noted that this map only needs to be a reasonable approximation of height variability; otherwise, it is meaningless to collect lidar data. Additionally, this approximation depends on the forest structure and the final accuracy requirements. In this step, QB imagery were used to provide a 'pseudo-height' base map, as previous studies have revealed that H-res optical imagery can provide detailed forest spectral, texture and shadow information, leading to promising results for estimating canopy height, AGB and volume (Franklin and McDermid, 1993; Pekkarinen, 2002; Hyde et al., 2006; Donoghue and Watt, 2006; Leboeuf et al., 2007; Chen et al., in press).

Specifically, Chen et al. (in press) described three types of QB-derived object-based variables (Table 2) which have proven beneficial for estimating canopy height. They are described as follows: (i) *mean spectral bands* (i.e., the average DN value within segments for each spectral band – blue, red, green and NIR); (ii) *image-texture* for spectral bands, which includes *internal-object texture* – a measure of the spatial variability of DNs within a segmented object, and *geographic object-based texture* (GEOTEX) – a measure of the spatial variability within neighboring objects; and (iii) *shadow fraction* – a quotient of the size of shaded areas and the size of the corresponding forest objects based on the DNs of the NIR band. In this study, these variables were extracted with custom IDL code (ITT Visual Information Solutions, Colorado, USA); and the pseudo-height classification was implemented in ENVI (ITT Visual Information Solutions, Colorado, USA) using ISODATA, an unsupervised classification algorithm. 14 height classes were generated, based on seven forest inventory height classes and the two tree types (i.e., conifer and deciduous) that occupy the site.

3.2.2. Transect feature decision

In this study, lidar transect selection requires the decision of four 'optimal' features: (i) sample size, (ii) orientation, (iii) location, and (iv) representative canopy height sampling. To best determine these features, the canopy-object pseudo-height classification image (derived from Section 3.2.1) was used as a proxy for forest height variability.

Based on the acquired airborne lidar data (Section 2.3), a swath width of 540 m was used to represent the minimum lidar extent for a single transect. Since fewer transects represent lower acquisition costs, four different sizes of transect samples were evaluated – based on differences in their total area resulting from one, two, three and four transects.

To evaluate if a directional bias existed in our rectangular study area (14.2 km × 11.5 km), two acquisition directions were also assessed. This resulted in four lidar transects evaluated in the N–S direction (representing 3.8%, 7.6%, 11.4% and 15.2% of the total study area), and four evaluated in the W–E direction (4.7%, 9.4%, 14.1% and 18.8%), thus, eight in total. The decision to test only four transects per direction was guided by results from Chen and Hay (2011) which revealed minimal model improvements with additional transects, though increasing lidar acquisition and processing costs.

With sample size and direction defined, the 'best' transect location(s), were determined from three main rules developed (in IDL code) from previous research (Chen and Hay, 2011). Specifically: (i) transects must contain no overlap with other transects; (ii) the transect-covered area must sample all canopy-object pseudo-height classes (with the non-vegetated objects masked out prior to analysis); and (iii) the canopy pseudo-height histogram derived from the transect-covered objects must have the highest correlation with the pseudo-height histogram derived from all objects for the entire site. Consequently, the best transect will use the smallest sample size to model the 'natural' height variability of the full scene. However, we note that this represents only one of many different rule scenarios. Conceptually, lidar collection could also be biased to maximize/minimize sample collection of one or more specific 'pseudo-height' classes, i.e., for the tallest/smallest height class, specific tree types, distance from roads, etc.

In practice, once the best transect sample size, location and direction have been defined from the pseudo-height image, this transect would be physically flown, and the acquired lidar data (referred hereafter as *lidar-measured canopy height*) would be used for modeling as described in the following sections. However, we note that in this research, a full lidar scene was already available from which the 'best' transects were extracted and evaluated.

3.3. Forest parameter generalization

3.3.1. Minimal-redundancy–maximal-relevance (mRMR) variable selection

Three types of QB-derived variables (Table 2) were used to link QB data with lidar-measured canopy height within the transect-covered area. As these variables have proven promising for canopy height estimation in eastern Canada (Chen and Hay, in press), our objective here is to evaluate their efficacy for predicting canopy height in complex western Canadian forests. To ensure that, high correlation (i.e., redundancy) is minimized between these variables to avoid model over-fitting, potentially decreasing model performance (Pal and Foody, 2010). *Variable selection* was used before the modeling step. Typically, variable selection (also know as *feature selection*) focuses on constructing and selecting an appropriate variable subset from all input variables to improve the model prediction performance (Guyon and Elisseeff, 2003). In this study, a machine learning technique named the mRMR (minimal-redundancy–maximal-relevance) approach (Peng et al., 2005) was applied to select the *best* variable subset, with which to separately model the canopy height for conifer and deciduous trees. Originally developed for gene selection, this algorithm has consistently shown promising results when evaluated with several commonly used classification methods (e.g., support vector machines and naive Bayes) and different types of data sets (e.g., handwritten digits and cancer cell lines) (Peng et al., 2005). mRMR has also shown a comparable performance with other feature selection methods (e.g., Random Forest and Correlation-Based algorithm) for remote sensing hyperspectral data (Pal and Foody, 2010). Additionally, mRMR is available as free open source software, is easy to use, and is computationally efficient.

The basic idea of using mRMR was to select variables with the minimal similarity between each other, and the maximal relevance with the target classes. In our evaluation, the result is a list of variables in order of predicting power, where the first variable or the combination of variables in the top of the list is more statistically

relevant to canopy height than those in the bottom. To define the best variable subset, we further conducted an extensive evaluation using different numbers and combination of variables (i.e., 1, 2, 3, ... 13) from the mRMR resulting list for conifer and deciduous trees, separately. The resulting model (see details in Section 3.3.2) estimation errors (i.e., RMSE) were compared using training data with in the eight lidar transect extents (see Section 3.2.2). The best number of variables for this study was selected based on two considerations: (i) if the errors continue decreasing with the addition of variables, the best number should represent a threshold where the use of more variables is less able to considerably increase model performance. This criterion facilitates computational efficiency in model development. (ii) If the errors decrease first and then increase significantly after reaching a peak value (Pal and Foody, 2010), the number of variables corresponding to this peak value is selected as the best.

3.3.2. Support vector regression (SVR) modeling

Support vector regression (SVR), also known as *support vector machines* (SVM) *for regression*, is an extended application of SVM used to solve complex nonlinear regression problems (Vapnik, 1998). SVR essentially transforms the nonlinear regression problem into a linear one using kernel functions to map the original input space into a new feature space with higher dimensions, while having two major advantages: (i) robustness in generalization, even when the training data are noisy; and (ii) it guarantees a unique global solution, which is not trapped in multiple local minima (Cristianini and Shawe-Taylor, 2000). Compared to the typically used multiple regression, SVR has demonstrated its robustness in several remote sensing applications including the estimation of evapotranspiration, ocean chlorophyll concentration, moisture transport over marine atmospheres and forest canopy height (Camps-Valls et al., 2006; Yang et al., 2006; Xie et al., 2008; Chen and Hay, in press).

In this study, SVR was applied to model forest height by developing nonlinear models linking segmented QB data with lidar-measured canopy height for both conifers and deciduous trees within the transect-covered area. The SVM open source software of LIBSVM was used for modeling (Chang and Lin, 2001). Specifically, (i) we used the model type ε-SVR, based on its wide use and good results; and (ii) we selected the *radial basis function* (RBF) as the kernel function, as it typically has better performance and requires fewer input parameters than other types of kernels (Chang and Lin, 2001). To determine the optimal SVR model parameters (i.e., C: penalty parameter, ε: precision parameter, and γ: kernel parameter), we have followed a two-step grid-search technique (Hsu et al., 2009) using training samples within the lidar transect extents. This technique is comprised (i) a coarse search using relatively large grid intervals, followed by (ii) a fine search using relatively small grid intervals within the selected large interval (that was defined from the coarse search). In this study, the best parameter combination was found at $C = 8.0$, $\varepsilon = 0.5$ and $\gamma = 1.0$. For more details about SVR basics, please refer to Gunn (1998), and Smola and Schölkopf (2004). After applying the SVM model, lidar-measured canopy height was generalized from the 'best' transect area to estimate canopy height for the entire study site.

3.3.3. Estimation of canopy height, AGB and volume

Nonlinear models with a natural logarithm form have been widely used with lidar data to estimate forest biophysical parameters (Næsset, 1997; Means et al., 1999; Lim et al., 2003). As a result, Eq. (1) was used to build the relationship between the estimated canopy height and our field measurements:

$$M_f = \beta_0 h_E^{\beta_1} \quad (1)$$

where M_f are the field measurements (canopy height, AGB or volume), h_E is the estimated lidar-measured canopy height (from Section 3.3.2); and β_0 and β_1 are coefficients. As our study area is covered by abundant mixed forest stands, and many field plots contain both conifers and deciduous trees, three types of models were developed for estimating forest canopy height, AGB and volume, without distinguishing forest type. To compare the field estimation results derived from our GEOBIA framework (i.e., the combination of segmentation, QB imagery, lidar transects and machine learning models) with those from the full lidar scene, Eq. (1) was further used to directly build relationships between the field measurements and the lidar canopy height segmentation image (CHS) for the entire site. The relationship between these two types of estimation results was then evaluated using a correlation coefficient (R) and RMSE.

4. Results and discussion

4.1. Image-objects

Fig. 3(a) represents a sample area in our study site, covered by deciduous trees, conifers, roads and forest gaps. Fig. 3(b) shows the corresponding area overlaid by image-object boundaries, derived from the segmentation procedure. Fig. 3(c) represents an object-based image, where the spectral values within each image-object are averaged. We note that most image-objects in this figure have jagged boundaries, which are distinctly different from the boundary delineation results from many other larger geographic features, such as roads and clear-cuts. This strongly suggests that the image-objects were particularly influenced by the high spectral variability within the canopy. Over small areas, manual tree crown delineation still produces better crown-objects than current automated segmentation tools (Castilla et al., 2008). However, image-objects increasingly provide improved results for vegetation parameter estimation, classification and change detection than traditional pixel-based approaches (Addink et al., 2007; Yu et al., 2008; Johansen et al., 2010; Chen et al., in press). Fig. 3(d) shows the classification result in the sample area, where white represents deciduous canopies, grey represents conifer canopies, and black represents non-forest areas. As our study area is characterized by mixed forests, deciduous and conifer stands are often mingled with each other in small groups. Consequently, a small MOS well captures the complex crown structure of this site. The overall classification accuracy for conifer's versus deciduous is 80.5%, with an overall kappa statistic of 0.78.

4.2. Spatial distribution of selected transects

By applying the transect selection algorithm (Section 3.2), Fig. 4 illustrates the four 'best' (N–S oriented) lidar transect combinations and their canopy height histograms; which represent (1) 3.8%, (2) 7.6%, (3) 11.4%, and (4) 15.2% of the total study area. To facilitate a comparison between the histograms derived from the best transects [i.e., Fig. 4(1b)–(4b)] and from the full-cover lidar data Fig. 4(5a), correlation coefficients were calculated based on canopy height frequency. High correlations were found with R values ranging from 0.95 to 0.98. Similarly, Fig. 5 illustrates the locations of the four (W–E oriented) lidar transect combinations using lidar samples of (1) 4.7%, (2) 9.4%, (3) 14.1%, and (4) 18.8%, and compares their histograms [i.e., Fig. 5(1b)–(4b)] with the full-area lidar data [i.e., Fig. 5(5b)]. A similar trend (to the N–S) exists in all histograms, although correlations are slightly lower, with R values ranging from 0.90 to 0.95. This may be caused by more masked areas in this orientation, resulting in smaller height class sample sizes, even though the areas of W–E transects were 23% larger than the N–S (due to the rectangular shape of the study area).

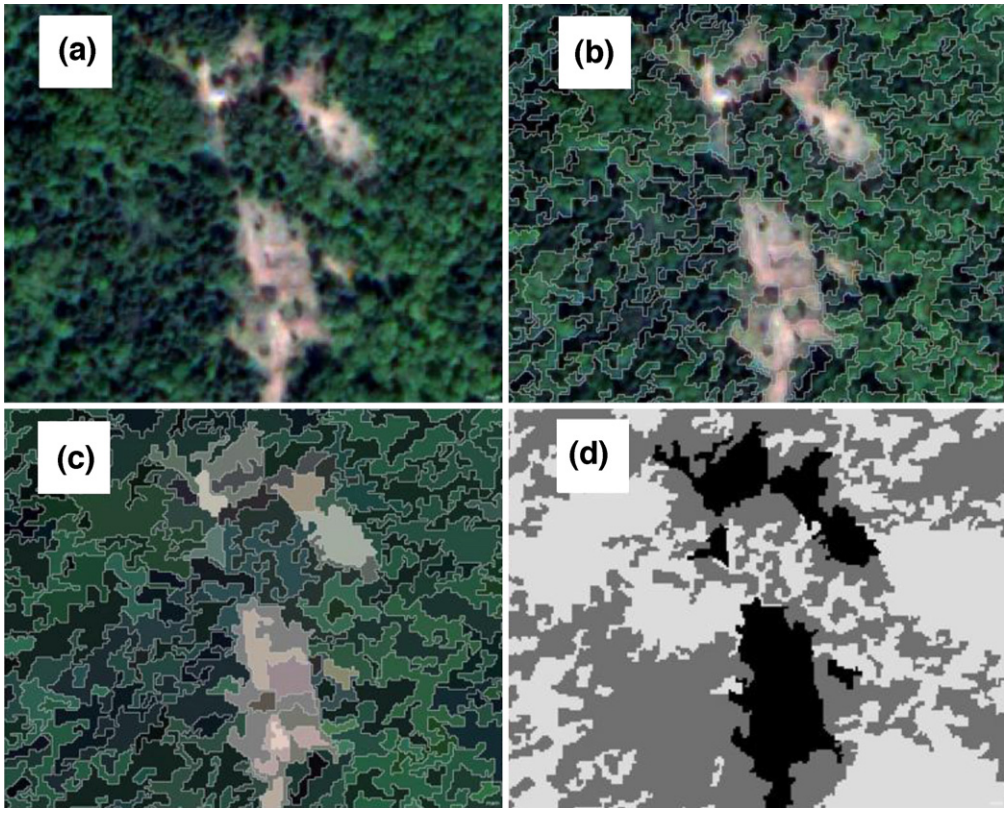

Fig. 3. (a) A sample area in the study site, (b) overlaid by image-object boundaries (grey lines) using a scale parameter of 50 (MOS: 0.04 ha), and (c) each object is filled with the average spectral values within the object extent. (d) The object-based classification result, where light grey represents deciduous canopies, the dark grey represents conifer canopies, and the black represents non-forest areas.

4.3. Selected modeling variables

Fig. 6 shows canopy height estimation errors derived from evaluating different numbers of variables for conifers and deciduous trees based on training data within the eight lidar transects. Results reveal that errors continue decreasing with the addition of variables for both tree types using all types of transect combinations. However, there exists a variable number threshold of six, with which the canopy height estimation performance increases by 12.2% and 12.0% (averaged from the results from all transect combinations) for conifers and deciduous trees, respectively. The addition of the remaining seven variables can only increase the estimation accuracy by 4.9% and 7.4% (averaged from the results from all transect combinations). Based on our predefined selection criterion [see (i) in Section 3.3.1] that considers both model accuracy and computational efficiency, the top six variables within the mRMR-derived resulting lists were chosen as the *best* variable subsets.

The subset variable names for eight different transect combinations as well as the full-cover lidar data are illustrated in Table 3. Specifically, for deciduous trees, all three types of variables (i.e., spectral, texture and shadow fraction) proved to be effective for linking Quickbird imagery with lidar data. However, we did not expect that the variable of shadow fraction to have such a low

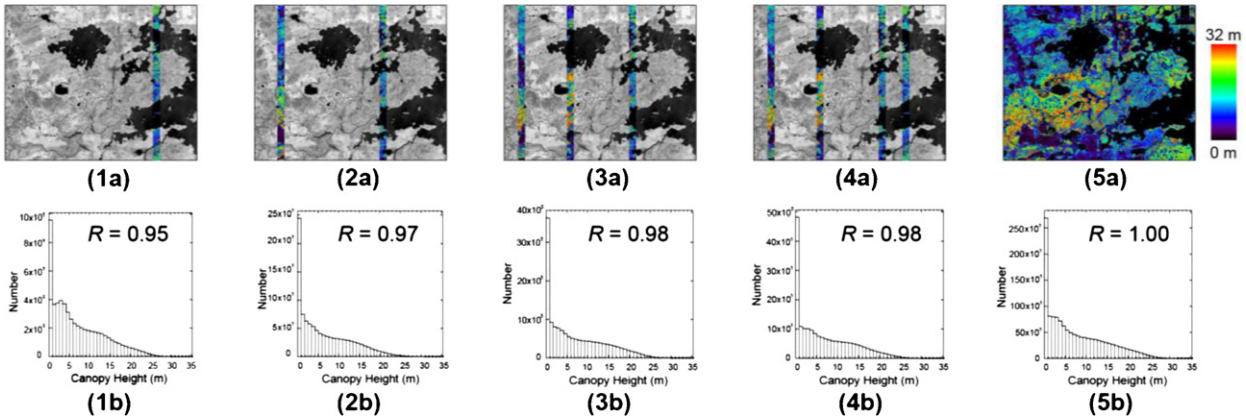

Fig. 4. (1a)–(4a) Illustrate the four lidar transect combinations (i.e., location and sample area in the N–S direction) derived from the lidar transect selection algorithm (Section 3.2), and the full lidar cover (5a), which represent various transect numbers and their (%) extents: (1) 3.8%, (2) 7.6%, (3) 11.4%, (4) 15.2% and (5) 100.0%. For illustrative purposes, the QB image was used as the base layer with lidar transects overlaid. (1b)–(5b) Illustrate the canopy height histograms derived from the corresponding lidar transect data in (1a)–(5a) and their correlation (R) with the height histogram of the full scene (5b).

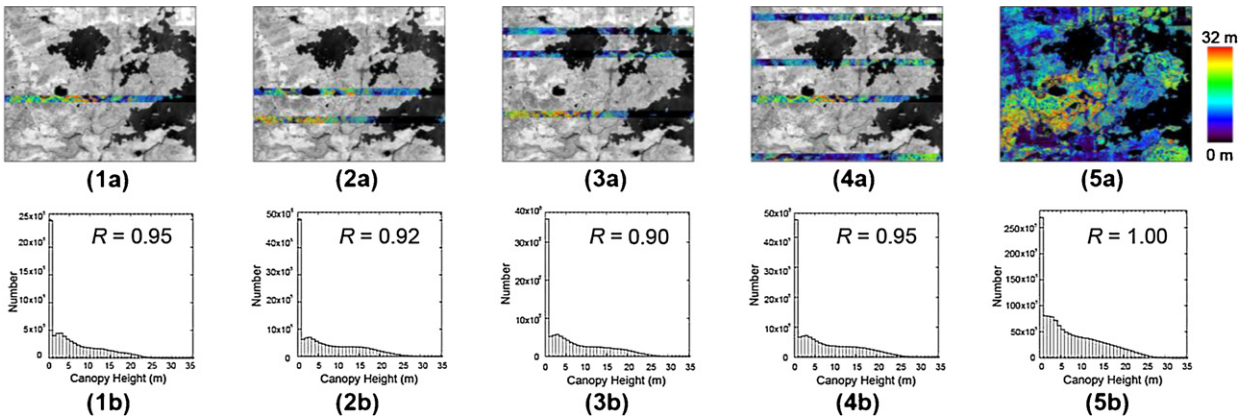

Fig. 5. (1a)–(5a) Illustrate the four lidar transect combinations (i.e., location and sample area in the W–E direction) derived from the lidar transect selection algorithm (Section 3.2) and the full-cover lidar data, which represent various transect numbers and their (%) extents: (1) 4.7%, (2) 9.4%, (3) 14.1%, (4) 18.8% and (5) 100.0%. For illustrative purposes, the QB image was used as the base layer with lidar transects overlaid. (1b)–(5b) Illustrate the canopy height histograms derived from the corresponding lidar transect data in (1a)–(5a) and their correlation (R) with the height histogram of the full scene (5b).

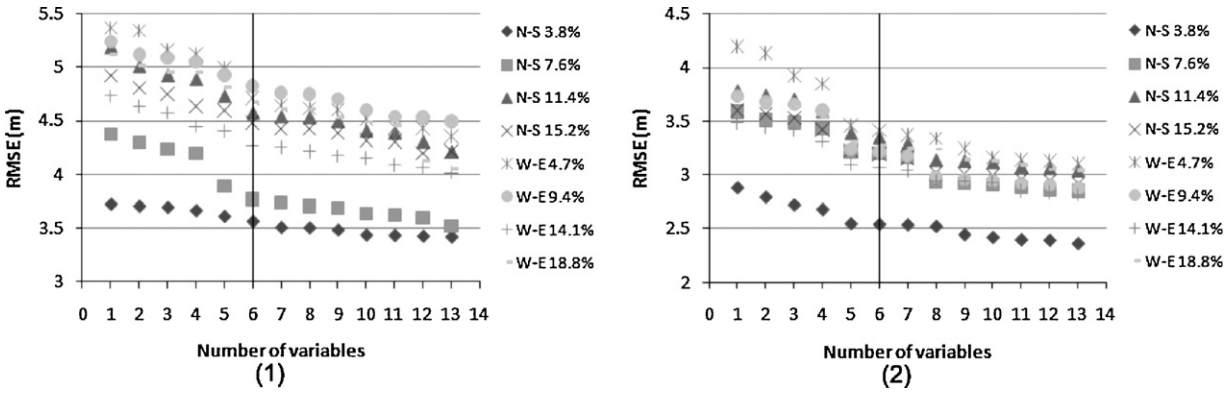

Fig. 6. Canopy height estimation errors derived from evaluating different numbers of variables (i.e., 1, 2, 3, . . . 13) for (1) conifers and (2) deciduous trees based on training data within eight lidar transect extents.

contribution for estimating conifer canopy height. In contrast with Pacific Rim forests, where conifers tend to be tall and dominate the landscape (Chen et al., in press), our Eastern study area is comprised mixed stands, where numerous tall broad crown deciduous trees dominate. As a result, the shadow information for relatively small conifers was affected by the neighboring larger deciduous trees; which may explain why shadow fraction was not as effective for conifers as the other variables. Another phenomenon, which can be found in Table 3, is that most variables maintain the same importance when a different lidar sample size was used. This suggests

Table 3
Variables selected for different lidar transect combinations, and ranked in the order of their predicting power from highest (left) to lowest (right).

Transect direction	Transect extent	mRMR selected variables for conifers (C) and deciduous trees (D)
N–S	3.8%	C: GEOTEX4, TXIT3, DN4, TXIT4, DN3, TXIT1 D: DN3, TXIT2, GEOTEX1, TXIT4, SF, GEOTEX4
	7.6%	C: DN3, TXIT3, GEOTEX3, TXIT4, DN4, TXIT1 D: DN3, TXIT3, GEOTEX4, TXIT4, GEOTEX3, SF
	11.4%	C: DN3, TXIT2, GEOTEX3, TXIT3, DN4, TXIT4 D: DN3, TXIT3, GEOTEX3, TXIT4, GEOTEX4, SF
	15.2%	C: DN3, TXIT2, GEOTEX3, TXIT3, DN4, TXIT4 D: DN3, TXIT3, GEOTEX3, GEOTEX4, TXIT4, SF
W–E	4.7%	C: TXIT3, GEOTEX3, DN3, TXIT4, DN4, TXIT1 D: TXIT3, DN4, GEOTEX3, TXIT4, GEOTEX4, DN3
	9.4%	C: DN3, GEOTEX3, TXIT3, TXIT4, DN4, TXIT1 D: DN3, TXIT3, GEOTEX3, TXIT4, GEOTEX4, SF
	14.1%	C: DN3, GEOTEX3, TXIT2, TXIT3, DN4, TXIT4 D: DN3, TXIT3, GEOTEX3, GEOTEX4, TXIT4, SF
	18.8%	C: DN3, TXIT2, GEOTEX3, TXIT3, DN4, TXIT4 D: DN3, TXIT3, GEOTEX3, TXIT4, GEOTEX4, SF
Full-cover lidar data	100.0%	C: DN3, TXIT2, GEOTEX3, TXIT3, DN4, TXIT4 D: DN3, TXIT3, GEOTEX4, TXIT4, GEOTEX3, SF

DNi = mean spectral value for the ith band; TXITi = internal-object texture value for the ith band; GEOTEXi = geographic object-based texture (GEOTEX) value for the ith band; SF = shadow fraction; and i is the band number (i.e., 1 – blue band, 2 – green band, 3 – red band and 4 – NIR band).

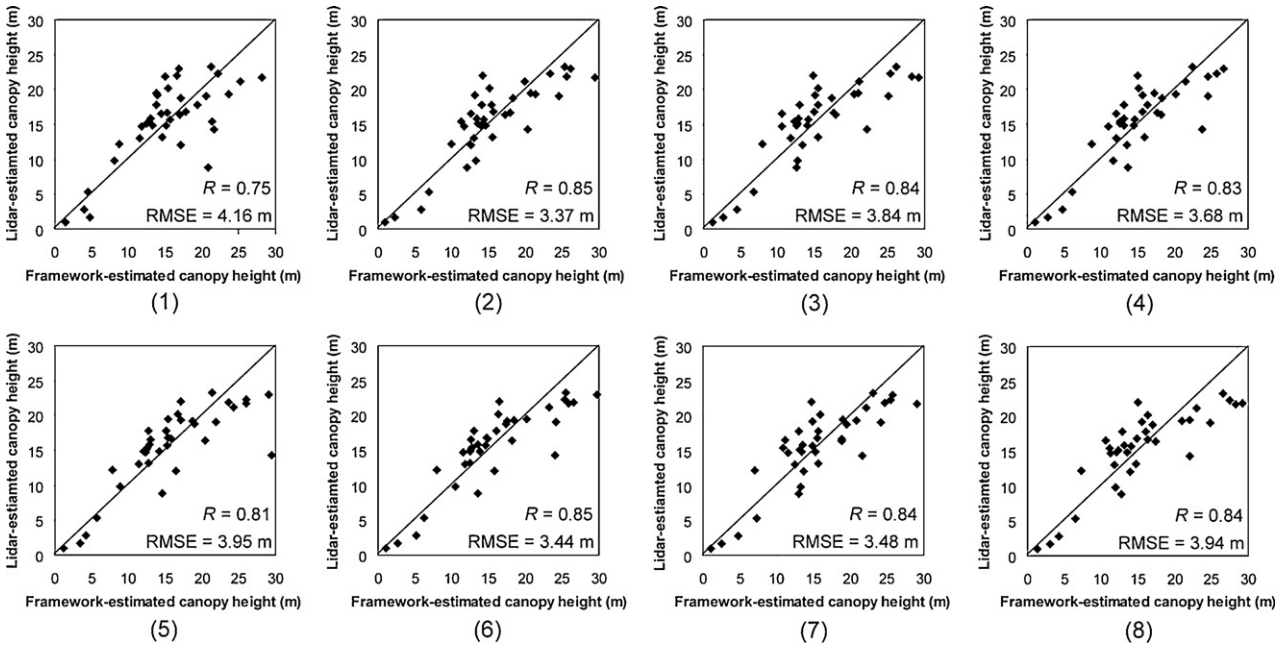

Fig. 7. Scatterplots of estimated canopy height (using lidar transects and Quickbird data) versus lidar-estimated canopy height (using full-cover lidar data). (1)–(4) Represent lidar transect extents in the N–S direction: (1) 3.8%, (2) 7.6%, (3) 11.4% and (4) 15.2%; while (5)–(8) represent lidar transect extents in the W–E direction: (5) 4.7%, (6) 9.4%, (7) 14.1% and (8) 18.8%.

that the mRMR variable selection approach can provide consistent inputs for our models which is robust to different transects sizes.

4.4. Comparison of model performance

In order to compare the model performance of our GEOBIA framework with that using all lidar data, Fig. 7 shows the scatterplots of our estimated canopy height (from lidar transects and Quickbird data) versus lidar-estimated canopy height (using the full lidar coverage). As the estimation of AGB and volume were both based on forest canopy height [see Eq. (1)], the correlation between the GEOBIA framework- and lidar-estimated AGB/volume was almost the same as the relationship of canopy height in Fig. 7. Therefore, only the canopy height scatterplots are presented in this paper. However, the estimation performance of AGB and volume are presented in Table 4.

Fig. 7 reveals that the smallest lidar transect sample has the lowest model performance, and largest height error. Specifically, both Fig. 7(1) and (5) illustrate canopy height estimation results from a single transect representing lidar samples of 3.8% and 4.7% (of the total site area) in N–S and W–E direction, where the relatively low correlation (R = 0.75, 0.81) and high errors (RMSE = 4.16 m, 3.95 m) were located. With an increase in transect sampled area the model performance for estimating forest canopy height also increases. However, the correlation only changes in a small range (i.e., between 0.83 and 0.85). This could be explained in two ways: (i) the selected transects and their related training samples are sufficient to develop robust models. For example, two transects (in N–S direction) include over 20,000 image-objects (i.e., training samples). If these samples can represent the height and species variability of the entire area, it is possible for the model to work well. This also indicates the importance of selecting appropriate lidar transects, as arbitrarily defined transects may not represent this variability. However, we note that it is difficult to determine how many training samples are 'sufficient', only from this study, as this number depends on several conditions, such as landscape complexity and accuracy requirement. (ii) Another reason is that the machine learning SVR approach has a strong generalization ability, which facilitates the collection of lidar data over a relatively small-area in order to estimate canopy height over a relatively large area.

The best performance (i.e., highest correlation and lowest error) of our model to estimate canopy height (R = 0.85; RMSE = 3.37 m) was found in Fig. 7(2), where two lidar transects (in N–S orientation) represent lidar cover of 7.6%. Correspondingly, the best AGB (R = 0.85; RMSE = 39.48 Mg/ha) and volume (R = 0.85; RMSE = 52.59 m^3/ha) estimation results were also found using the same transect features. Fig. 7 also reveals that our models tend to overestimate the (lidar-estimated) forest parameters for tree canopies that are lower than 5.0 m, or taller than 20 m. However,

Table 4
Estimation performance [R (correlation coefficient) and RMSE (root mean squared error)] of above ground biomass (AGB) and volume for eight different lidar transect combinations in two different directions.

Transect direction	Transect extent	R	RMSE of AGB (Mg/ha)	RMSE of volume (m^3/ha)
N–S	3.8%	0.75	46.86	71.93
	7.6%	0.85	39.48	52.59
	11.4%	0.84	44.71	66.91
	15.2%	0.83	41.92	53.28
W–E	4.7%	0.81	42.24	76.24
	9.4%	0.85	39.04	78.76
	14.1%	0.84	40.25	54.29
	18.8%	0.84	45.62	65.45

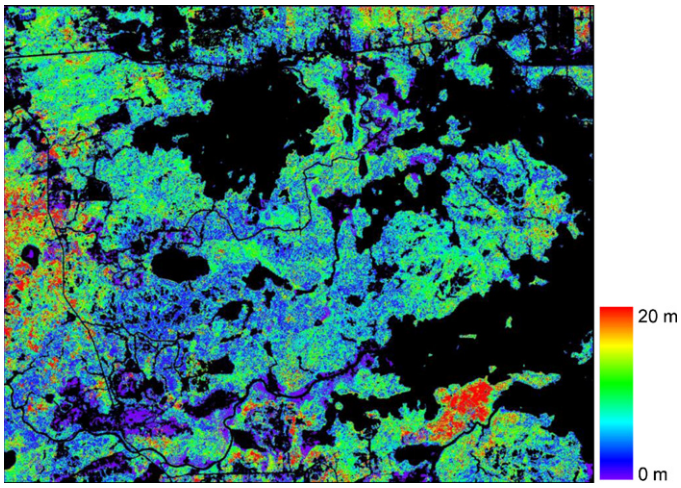

Fig. 8. The error image derived from the 'best' lidar transect sample, that represents 7.6% of the entire study area. Black represents non-forest areas, while the other colors represent different errors in height estimation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

we note that most field plots contain canopies within 5–20 m. Additionally, the average error (RMSE = 3.37 m) is lower than the ≈5.0 m forest inventory height class interval for this area.

To better understand the spatial distribution of the canopy height error derived from our framework, an error image (Fig. 8) was created (using the 7.6% lidar sample) by illustrating the difference between our canopy height estimates and the lidar CHS. Fig. 8 shows that our framework produced errors that are lower than 5.0 m over 50.1% of the forest-covered area. However, it has difficulties with some forest stands covered by low density deciduous trees lower than 8.0 m, such as those illustrated in the red color region close to the lower right corner of our study area. This may be caused by misclassification, as the vegetated ground patches appear as trees. However, errors that are larger than 15.0 m only account for 2.7% of the forests, and typically appear near roads, water boundaries and cut blocks, where these types of landscape edge features are complex. We also note that the lidar data contain errors (i.e., 1.30 m) when compared to the field measurements.

5. Conclusions

In this study, we have generated geo-intelligence from a forest scene by reducing airborne lidar data acquisition costs, providing meaningful geospatial information related to the size, orientation and location to best acquire lidar transects, and have applied novel machine learning algorithms to model important forest parameters over a large area. A semi-automatic GEOBIA framework is presented to extract forest information (i.e., canopy height, AGB and volume) at the small crown/cluster level (i.e., MOS: 0.04 ha). This framework is comprised three main components: (i) image-object extraction, (ii) lidar transect selection, and (iii) forest parameter generalization. Although our study area is dominated by complex mixed forest stands composed of six major species, the integration of the object-based paradigm with a lidar transect selection algorithm and machine learning mRMR and SVR techniques has produced promising results. In particular, the GEOBIA framework derived estimates of forest parameters show a strong relationship with those generated from lidar data covering the full (16,330 ha) study site. The highest correlation and lowest error for canopy height (R = 0.85; RMSE = 3.37 m), AGB (R = 0.85; RMSE = 39.48 Mg/ha) and volume (R = 0.85; RMSE = 52.59 m^3/ha) were obtained using a N–S direction lidar sample representing 7.6% (1240 ha) of the entire study area. Due to the difficulties in field logistics and budget limitations, we note that the field plots were derived with a bias to forest stands accessible from roads, as well as to tree heights between 10 and 25 m (Fig. 7). However, these field data sample a wide range of canopy heights from 1.4 m to 27.1 m, which represent all study site height classes. In this research, the selection of a small mean object size (0.04 ha) ensured that forest variability was well represented, as large image-objects would have resulted in information loss, and an inability to exploit the information potential inherent to high resolution remotely sensed imagery. Additionally, the mRMR variable selection approach has proven to provide consistent inputs for our models which is robust to different transects sizes. Future work will consider improving the performance of this framework by modeling dominant tree species individually and incorporating additional variables/features.

Acknowledgments

This research has been funded by an Alberta Informatics Circle of Research Excellence (iCore) Ph.D. scholarship awarded to Gang Chen. Dr. Hay acknowledges support from a Natural Sciences and Engineering Research Council (NSERC) Discovery Grant and an AIF New Faculty Award. Dr. Benoît St-Onge acknowledges support from the BIOCAP Foundation of Canada. We also thank the anonymous reviewers for their valuable suggestions.

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 (8), 905–912.

Blaschke, T., 2010. Object based image analysis for remote sensing. ISPRS Journal of Photogrammetry and Remote Sensing 65 (1), 2–16.

Camps-Valls, G., Bruzzone, L., Rojo-Alvarez, J.L., Melgani, F., 2006. Robust support vector regression for biophysical variable estimation from remotely sensed images. IEEE Geoscience and Remote Sensing Letters 3 (3), 339–343.

Castilla, G., Hay, G.J., Ruiz, J.R., 2008. Size-constrained region merging (SCRM): an automated delineation tool for assisted photointerpretation. Photogrammetric Engineering and Remote Sensing 74 (4), 409–419.

Chang, C.-C., Lin, C.-J., 2001. LIBSVM: A Library for Support Vector Machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chen, G., Hay, G.J., Castilla, G., St-Onge, B., Powers, R. 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 press.

Chen, G., Hay, G.J., 2011. An airborne lidar sampling strategy to model forest canopy height from Quickbird imagery, lidar transects and GEOBIA. Remote Sensing of Environment 15 (6), 1532–1542.

Chen, G., Hay, G.J. A support vector regression approach to estimate forest biophysical parameters at the object level using airborne lidar transects and Quickbird data. Photogrammetric Engineering and Remote Sensing, in press.

Cristianini, N., Shawe-Taylor, J., 2000. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press, Cambridge, UK, 189p.

Donoghue, D.N.M., Watt, P.J., 2006. Using LiDAR to compare forest height estimates from IKONOS and Landsat ETM+ data in Sitka spruce plantation forests. International Journal of Remote Sensing 27 (11), 2161–2175.

Eamus, D., Jarvis, P.G., 1989. The direct effects of increase in the global atmospheric CO_2 concentration on natural and commercial temperate trees and forests. Advances in Ecological Research 19, 1–55.

Franklin, S.E., McDermid, G.J., 1993. Empirical relations between digital SPOT HRV and CASI spectral response and lodgepole pine (Pinuscontorta) forest stand parameters. International Journal of Remote Sensing 14 (12), 2331–2348.

Gunn, S.R., 1998. Support Vector Machines for Classification and Regression. Technical Report, 66p.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. Journal of Machine Learning Research 3, 1157–1182.

Hay, G.J., Blaschke, T., 2010. Forward: special issue on geographic object-based image analysis (GEOBIA). Photogrammetric Engineering and Remote Sensing 76 (2), 121–122.

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. Springer-Verlag, Berlin, pp. 77–92.

Hay, G.J., Marceau, D.J., Dubé, P., Bouchard, A., 2001. A multiscale framework for landscape analysis: object-specific analysis and upscaling. Landscape Ecology 16 (6), 471–490.

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 (1), 5–12.
Hsu, C.-W., Chang, C.-C., Lin, C.-J., 2009. A Practical Guide to Support Vector Classification. Technical Report, 15p.
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 (2–3), 397–416.
Hyde, P., Dubayah, P., Walker, W., Blair, J.B., Hofton, M., Hunsaker, C., 2006. Mapping forest structure for wildlife habitat analysis using multi-sensor (LiDAR, SAR/InSAR, ETM+, Quickbird) synergy. Remote Sensing of Environment 102 (1–2), 26–36.
Johansen, K., Arroyo, L.A., Phinn, S., Witte, C., 2010. Comparison of geo-object based and pixel-based change detection of riparian environments using high spatial resolution multi-spectral imagery. Photogrammetric Engineering and Remote Sensing 76 (2), 123–136.
Kajisa, T., Murakami, T., Mizoue, N., Top, N., Yoshida, S., 2009. Object-based forest biomass estimation using Landsat ETM+ in Kampong Thom Province, Cambodia. Journal of Forest Research 14 (4), 203–211.
Lambert, M.-C., Ung, C.-H., Raulier, F., 2005. Canadian national tree aboveground equations. Canadian Journal of Forest Research 35 (8), 1996–2018.
Leboeuf, A., Beaudoin, A., Fournier, R.A., Guindon, L., Luther, J.E., Lambert, M.-C., 2007. A shadow fraction method for mapping biomass of northern boreal black spruce forests using QuickBird imagery. Remote Sensing of Environment 110 (4), 488–500.
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 (5), 658–678.
Mäkelä, H., Pekkarinen, A., 2001. Estimation of timber volume at the sample plot level by means of image segmentation and Landsat TM imagery. Remote Sensing of Environment 77 (1), 65–75.
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 (3), 298–308.
Mora, B., Wulder, M.A., White, J.C., 2010. Segment-constrained regression tree estimation of forest stand height from very high spatial resolution panchromatic imagery over a boreal environment. Remote Sensing of Environment 114 (11), 2474–2484.
Næsset, E., 1997. Estimating timber volume of forest stands using airborne laser scanner data. Remote Sensing of Environment 61 (2), 246–253.
Pal, M., Foody, G.M., 2010. Feature selection for classification of hyperspectral data by SVM. IEEE Transactions on Geoscience and Remote Sensing 48 (5), 2297–2307.
Pekkarinen, A., 2002. Image segment-based spectral features in the estimation of timber volume. Remote Sensing of Environment 82 (2–3), 349–359.
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 (8), 1226–1238.
Perron, J.-Y., 2003. Tarif De Cubage Général – Volume marchand brut (in French). Available at: http://www.mrn.gouv.qc.ca/publications/forets/connaissances/tarif-de-cubage-64.pdf.
Popescu, S.C., Wynne, R.H., 2004. Seeing the trees in the forest: using lidar and multispectral data fusion with local filtering and variable window size for estimating tree height. Photogrammetric Engineering and Remote Sensing 70 (5), 589–604.
Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Statistics and Computing 14 (3), 199–222.
Stojanova, D., Panov, P., Gjorgjioski, V., Kobler, A., Dzeroski, S., 2010. Estimating vegetation height and canopy cover from remotely sensed data with machine learning. Ecological Informatics 5 (4), 256–266.
Vapnik, V., 1998. Statistical Learning Theory. Wiley, New York, 736p.
Welch, R., Ahlers, W., 1987. Merging multiresolution SPOT HRV and Landsat TM Data. Photogrammetric Engineering and Remote Sensing 53 (3), 301–303.
Wulder, M.A., 1998. Optical remote sensing techniques for the assessment of forest inventory and biophysical parameters. Progress in Physical Geography 22 (4), 449–476.
Wulder, M.A., Han, T., White, J.C., Sweda, T., Tsuzuki, H., 2007. Integrating profiling LIDAR with Landsat data for regional boreal forest canopy attribute estimation and change characterization. Remote Sensing of Environment 110 (1), 123–137.
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 (5), 536–543.
Xie, X., Liu, W.T., Tang, B., 2008. Spacebased estimation of moisture transport in marine atmosphere using support vector regression. Remote Sensing of Environment 112 (4), 1846–1855.
Yang, F., White, M.A., Michaelis, A.R., Ichii, K., Hashimoto, H., Votava, P., Zhu, A.X., Nemani, R.R., 2006. Prediction of continental-scale evapotranspiration by combining MODIS and ameriflux data through support vector machine. IEEE Transactions on Geoscience and Remote Sensing 44 (11), 3452–3461.
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 (8), 1007–1018.