Remote Sensing of Environment 112 (2008) 3971–3982

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

# Remote sensing of the distribution and abundance of host species for spruce budworm in Northern Minnesota and Ontario

Peter T. Wolter [a,*], Philip A. Townsend [a], Brian R. Sturtevant [b], Clayton C. Kingdon [a]

[a] Department of Forest and Wildlife Ecology, University of Wisconsin – Madison, Madison, Wisconsin, 53706, USA
[b] Forestry Sciences Laboratory, USDA Forest Service, 5985 Hwy K, Rhinelander, Wisconsin, 54501, USA

ARTICLE INFO

*Article history:*
Received 20 February 2008
Received in revised form 16 May 2008
Accepted 5 July 2008

*Keywords:*
Landsat
Multi-temporal
Partial least squares regression
Forest structure
Spruce budworm
Minnesota
Ontario

ABSTRACT

Insects and disease affect large areas of forest in the U.S. and Canada. Understanding ecosystem impacts of such disturbances requires knowledge of host species distribution patterns on the landscape. In this study, we mapped the distribution and abundance of host species for the spruce budworm (*Choristoneura fumiferana*) to facilitate landscape scale planning and modeling of outbreak dynamics. We used multi-temporal, multi-seasonal Landsat data and 128 ground truth plots (and 120 additional validation plots) to map basal area (BA), for 6.4 million hectares of forest in northern Minnesota and neighboring Ontario. Partial least-squares (PLS) regression was used to determine relationships between ground data and Landsat sensor data. Subsequently, BA was mapped for all forests, as well as for two specific host tree genera (*Picea* and *Abies*). These PLS regression analyses yielded estimates for overall forest BA with an $R^2$ of 0.62 and RMSE of 4.67 $m^2$ $ha^{-1}$ (20% of measured BA), white spruce relative BA with an $R^2$ of 0.88 (RMSE = 12.57 $m^2$ $ha^{-1}$ [20% of measured]), and balsam fir relative BA with an $R^2$ of 0.64 (RMSE = 6.08 $m^2$ $ha^{-1}$ [33% of measured]). We also used this method to estimate the relative BA of deciduous and coniferous species, each with $R^2$ values of 0.86 and RMSE values of 9.89 $m^2$ $ha^{-1}$ (23% of measured) and 9.78 $m^2$ $ha^{-1}$ (16% of measured), respectively. Of note, winter imagery (with snow cover) and shortwave infrared-based indices – especially the shortwave infrared/visible ratio – strengthened the models we developed. Because ground measurements were made largely in forest stands containing spruce and fir, modeled results are not applicable to stands dominated by non-target conifers such as pines and cedar. PLS regression has proven to be an effective modeling tool for regional characterization of forest structure within spatially heterogeneous forests using multi-temporal Landsat sensor data.

© 2008 Elsevier Inc. All rights reserved.

## 1. Introduction

Compared to other forest disturbances, insects and disease influence the largest area of forests in both the U.S. and Canada, affecting an estimated 20.2 million hectares in the U.S. with economic costs over $1.5 billion (Dale et al., 2001). The understanding and effective management of such disturbances requires knowledge of the distribution and patterns of host species for insects and diseases. This facilitates understanding of the potential for large-scale disturbances, such as severe insect outbreaks, but also provides the context to understand the likely consequences of outbreaks on a regional scale, such as changes in tree species composition, age structure, and fuel conditions (Hadley, 1994; White & Host, 2003; Williams & Birdsey, 2003). One of the most destructive insects to North American spruce–fir forests is the spruce budworm (*Choristoneura fumiferana*), whose widespread, recurrent outbreaks (see Blais, 1983; Erickson & Hastings, 1978; Williams & Birdsey, 2003) are a primary driving force shaping the structure, function, and fire history of these forests (Fleming et al., 2002). Because of their host-specific nature, spruce budworm outbreaks are responsive to the abundance and spatial distribution of their host and, consequently, are also suspected of sensitivity to feedback related to forest succession and change (Bergeron & Leduc, 1998; Hessburg et al., 1999).

Forest change, attributed primarily to increased effectiveness of fire suppression in the Border Lakes region of northern Minnesota and neighboring Ontario over the last century, has resulted in the conversion of pioneer species such as jack pine (*Pinus banksiana*) and quaking aspen (*Populus tremuloides*), to mixed-age, shade-tolerant species composed of white spruce (*Picea glauca*), black spruce (*Picea mariana*), balsam fir (*Abies balsamea*), and white cedar (*Thuja occidentalis*) (Baker, 1992; Frelich & Reich, 1995; Scheller et al., 2005). In recent decades, growth in demand for pulpwood has led to forest management strategies (e.g., clear-cutting) that promote growth of increasingly large, homogenous areas of aspen–fir forest associations (Blais, 1983; Wolter & White, 2002) which have greatly altered this region's landscape structure and dynamics (Pastor et al., 2005; White & Host, 2003; Wolter & White, 2002).

Because spruce budworm host-species such as balsam fir have become more dominant in the landscape, the probability of new

* Corresponding author. Tel.: +1 608 263 9107.
*E-mail address:* ptwolter@wisc.edu (P.T. Wolter).

0034-4257/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.rse.2008.07.005

outbreaks has also increased (Blais, 1983; Sturtevant et al., 2004). Early efforts to understand spruce budworm dynamics in this region identified three key stand characteristics that explained 56% of the variation in balsam fir mortality following an outbreak: percent basal area (BA) in spruce, percent BA in non-host species, and BA of balsam fir (Batzer, 1969). Spatially explicit forest landscape simulation models such as LANDIS are well suited to study the potential effects of multiple spatially interactive drivers of ecological change on future forest composition (Scheller & Mladenoff, 2005). Efforts to model and manage for insect impacts at landscape and regional scales therefore depend on the availability of spatially explicit data on host species distribution. Satellite remote sensing represents a valuable source for supplying input data for regional simulation modeling.

### 1.1. Study objective

The broad objective of the study is to map the distribution and abundance of spruce budworm host species (fir and spruce) in the northeastern Minnesota and adjacent Ontario to better understand the dynamics of this insect and identify landscape-scale management strategies that may minimize outbreak frequency and severity. The specific goal of this paper is to demonstrate a novel approach for modeling and mapping forest basal area (BA) and species abundance using readily available sources of remote sensing data. We employ partial least-squares (PLS) regression with multi-temporal Landsat sensor data to map spruce (*P. glauca*, *P. mariana*) and fir (*A. balsamea*) distribution and BA for a 6.4 million hectare area covering the Border Lakes region of northern Minnesota, U.S.A. and northwestern Ontario, Canada (Fig. 1). The strategy involves use of all the reflectance bands of Landsat-5 and -7 plus several spectral indices (SI) derived from these sensor data for multiple image dates per WRS-2 path and row. PLS is used with Landsat data and field data to produce models for mapping total forest BA (TBA), relative BA of fir (FIR), relative BA of spruce (SPRUCE), relative BA of deciduous forest (DEC), and relative BAs of coniferous forest (CON). While PLS has been used extensively with hyperspectral data (Coops et al., 2003; McDonald et al., 2003; Ourcival et al., 1999; Smith et al., 2002, 2003; Townsend et al., 2003), we demonstrate the capability of the algorithm to handle multi-temporal, broad band, satellite sensor data. PLS regression is convenient as it: 1) allows simultaneous modeling of multiple continuous predictor variables; 2) does not make unrealistic assumptions about spectral or ground measurement error, such as in ordinary least-squares regression (Cohen et al., 2003; Curran & Hay, 1986); and 3) addresses the problem of collinearity (dependence) among multiple independent and dependent variables (Helland, 1988).

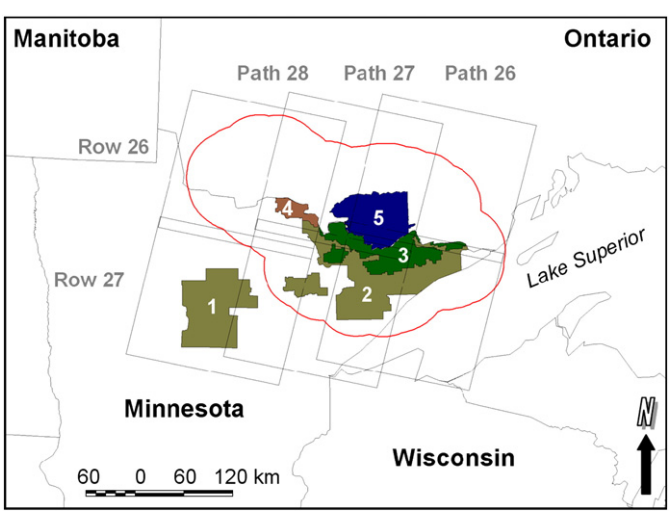


**Fig. 1.** Study area in northern Minnesota and northwestern Ontario showing the six Landsat footprints and the Border Lakes focus area in red. The region includes land ownerships with varying forest management strategies: Chippewa National Forest (1), Superior NF (2), BWCA Wilderness (3), Voyageurs National Park (4), and Quetico Provincial Park (5).

### 1.2. Background

#### 1.2.1. Partial least-squares regression

PLS is a predictive, 2-block regression strategy that uses estimated linear, latent variables or components, obtained through optimization of covariance measures (Nielsen, 2002), to simultaneously analyze two data sets (e.g., spectra and physical/chemical properties) collected from a single object of interest (Norgaard et al., 2000). PLS identifies a select number of eigenvectors from an independent data matrix capable of generating score values that capture predictor variance and are highly correlated with the response variables (Arenas-Garcia & Camps-Valls, 2007). Contrary to ordinary least-squares and multiple linear regression, PLS regression does not assume zero error in the predictor data (often falsely assumed for image data, Curran & Hay, 1986). PLS regression assumes that, if well sampled, vectors in the predictor space (irrespective of error) should provide superior predictive power for additional observations when there is a high degree of correlation among predictor variables (SAS, 2000). Ultimately, PLS regression seeks a balance between explaining variations in both response and predictor variables (SAS, 2000).

PLS regression was formulated out of a need to model information-scarce datasets in the social sciences (Wold, 1966, 1975). Kowalski et al. (1982) extended the use of PLS regression to chemometric applications using full-spectrum radiometer data. The PLS regression method is attractive because it provides a means to reduce a large number of collinear variables into relatively few relevant, non-correlated, latent structures or components (Norgaard et al., 2000). PLS regression differs from principal components regression (PCR) in that it uses the covariance between *X* and *Y* variables to form latent variables. As a result, variance among the *Y* variables is described better than the principal components of PCR that are based solely on the *X* variables (Zang et al., 2007). PLS regression is also superior to canonical correlation analysis (CCA) in situations where there are fewer observations with respect to variables, as PLS strives to maximize covariance, rather than correlation, between latent components (Zang et al., 2007).

In traditional multiple linear regression, when there are more samples than independent variables (e.g., full-spectrum remote sensing data) an exact solution for the regression coefficients (**B** vector) is not possible without minimizing the length of the residual vector. This may be accomplished using the least-squares method:

$$\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{1}$$

However, in the presence of collinearity among the **X** vector variables, an inverse for **X′X** may not be possible, causing instability among regression coefficients (Geladi & Kowalski, 1986). PLS regression reduces the rank of the **X′X** matrix by using a subset of the **X′X** eigenvectors, known as principal components or latent variables (hereafter referred to as components), to represent **X′X** in Eq. (1) (Wold et al., 1984) giving it the form of a generalized inverse (Marquardt, 1970). The resultant model is composed of two outer relations derived from the eigenstructure decomposition of **X** and **Y** as well as an inner relation coupling the **X** and **Y** score matrices (Geladi & Kowalski, 1986).

The key to PLS regression is deciding how many components to use for a given model complexity. While it is possible to compute as many components as there are predictor variables, a smaller initial number of components are typically computed (see SAS, 2000). This allows lower order components – often describing random measurement error as well as retaining collinearity problems (Geladi & Kowalski, 1986) – to be discarded and reduces the chance of over fitting the

model. The number of components to use for optimal model complexity is determined iteratively via cross-validation (Geisser, 1974; Stone, 1974; Wold et al., 1984).

Cross-validation can be initiated several ways, but in each strategy data points are withheld either individually or in groups followed by recursive model fitting with different numbers of components. Retained points are used to compute residuals and withheld points are used to calculate prediction error for each level of model complexity. Iterations cease after each data point has been withheld once. The cumulative sum of individual point prediction errors is known as the predicted residual sum of squares (PRESS) and provides a measure of model predictive capacity for a particular data set (Wold et al., 1984). The final level of model complexity (number of components) is chosen by statistically comparing the PRESS values of successively more complex models (Van der Voet, 1994). When addition of the next successive component (i.e., describing less of the covariance between each **X** score and corresponding **Y** score than the preceding component) fails to improve the PRESS statistic by some predetermined amount, the preceding model becomes final (Wold et al., 1984). Non-correlated components that are retained are then used in PLS regression to parameterize predictive models. The result is that the estimated regression coefficients are stabilized (Wold et al., 1984) making PLS models much more robust than classical multiple linear regression or principal components regression (Geladi & Kowalski, 1986).

#### 1.2.2. PLS applications in forest ecosystems

The use of PLS regression to calibrate models of biochemical and biophysical forest parameters using full-spectrum (i.e. hyperspectral) imagery (e.g., AVIRIS 224 bands and Hyperion 220 bands) has become routine (Coops et al., 2003; Martin et al., 2008; McDonald et al., 2003; Ourcival et al., 1999; Smith et al., 2002, 2003; Townsend et al., 2003). Smith et al. (2002) demonstrated that not only is it possible to measure whole-canopy nitrogen in small experimental forest ecosystems, but that it is also possible to calibrate multiple, contiguous models enabling landscape-scale forest productivity estimation for a diverse range of forest communities in New England. These same basic techniques have been employed successfully for a broad range of forest communities (Coops et al., 2003; Martin et al., 2008; McDonald et al., 2003; Townsend et al., 2003).

Recently, improvement in prediction potential has been achieved by first identifying a salient set of the hyperspectral bands (those showing elevated sensitivity to the response variable) prior to PLS regression analysis (Forina et al., 1999; Jarvis & Goodacre, 2005; Leardi & Gonzalez, 1998; Li et al., 2007; Lindgren et al., 1993; Swierenga et al., 1998). Pre-selection excludes bands showing low sensitivity to the response variable. Such variables adversely affect model calibration and result in both large relative bias toward zero and small additive bias away from the origin regardless of signal to noise distribution (Spiegelman et al., 1998). Thus, careful design of the pre-selection process is required. One popular method is the genetic algorithm (GA) of Holland (1975, also see Jarvis & Goodacre, 2005), which has been demonstrated to be especially useful for hyperspectral data (Li et al., 2007).

Many approaches have been developed to reduce the dimensionality of broad band, satellite sensor data (e.g., Landsat) for estimating forest biophysical parameters (Ardö, 1992; Clark et al., 2004; Cohen & Spies, 1992; Cohen et al., 2003; Hall et al., 1995; Healey et al., 2006; Li & Strahler, 1985; Peddle et al., 1999; Song, 2007; Song et al., 2007; Song & Woodcock, 2003; Woodcock et al., 1997; Zheng et al., 2004; and many more). Where some approaches strive to condense X-vector data into one meaningful independent variable prior to regression (Cohen et al., 2003), PLS regression provides a means to handle the X-vector space (multiple sensor bands and derivatives) to provide simultaneous estimates for the Y-vector dependent variables. However, examples of research using PLS regression for calibration of broad band, multi-temporal, remote sensing models with forest structural parameters (e.g., basal area, tree height, or canopy closure) have not yet been published. Nielsen (2002) provides an example of multiset canonical correlation analysis (CCA) using multi-temporal Landsat-5 sensor data to map forest change but only mentions PLS regression as being a similar approach. Thus, PLS regression applied to multi-temporal, Landsat sensor data represents a novel, robust, and appropriate approach for regional forest parameter prediction.

## 2. Methods

### 2.1. Study region

The 70,000 km$^2$ study region consists of the Border Lakes ecoregion, straddling northern Minnesota and Ontario, plus a 50 km surrounding buffer (Fig. 1). The total area, hereafter referred to as the Border Lakes (BL), is covered by portions of six Landsat scenes (each scene is 185 × 185 km) and includes all of the Superior National Forest (SNF), Boundary Waters Canoe Area (BWCA) wilderness, Voyageurs National Park (VNP), and Quetico Provincial Park (QPP) (Fig. 1). The BL region has a continental climate, with long, cold winters and relatively short summers (Heinselman, 1973). Average January and July temperatures are −16 °C and 19 °C, respectively. Sixty-seven percent of the average annual precipitation (61 cm) is received between May and September, with June being the wettest overall month (accounting for 17% of the total average). Numerous lakes, shallow soils over Precambrian bedrock, and gentle relief that is primarily of glacial origin characterize this region.

BL forest cover is considered transitional between the sub-boreal, Great Lakes–St. Lawrence forests and boreal forest (Baker, 1989; Heinselman, 1973). The BWCA, VNP, and QPP are protected from commercial logging and are largely composed of pine (*P. banksiana*, *P. resinosa*, and *P. strobus*), spruce (*P. glauca*, and *P. mariana*), and balsam fir (*A. balsamea*), with lesser amounts of white cedar (*T. occidentalis*), tamarack (*Larix laricina*), aspen (*P. tremuloides*, *P. grandidentata*, and *P. balsamifera*), paper birch (*Betula papyrifera*), and red maple (*Acer rubrum*). Unprotected forests are intensively managed for wood fiber and have a higher dominance of aspen, spruce, and fir forest type associations (Pastor et al., 2005; Wolter & White, 2002). The western portion of the BL region, on the U.S. side, is dominated by peatlands that support pure stands of black spruce (*P. mariana*) and tamarack, while the southeast region, within Lake Superior's North Shore Uplands ecological subsection, supports pure forests of sugar and red maple (*A. saccharium* and *A. rubrum*) (Wolter & White, 2002).

### 2.2. Field data

Field plot data ($n$ = 128) used to model forest structural parameters were collected during the summers of 2003 and 2004 and distributed evenly between Landsat WRS-2 paths 26 and 27 to facilitate comprehensive mapping (Fig. 1). Each plot consists of a cluster of five subplots located at the intersection and four end points of two crossing 50 × 50 m transect lines placed near center of large (≥7 × 7 pixels or 4.4 ha), homogenous stands. Sufficient stand size and homogeneity assured that any stand edge effects would be minimized during analysis, and that image misregistration errors, if greater than 15 m, would be inconsequential. Basal area (BA) by species was measured at each subplot using a metric factor 2 prism. In addition, percent cover of vegetation was visually estimated and placed in one of 10 cover classes according to Peet et al. (1998) for the canopy, subcanopy, shrub/sapling, and herb layers; the coniferous and deciduous portion was also estimated for each layer. Each successive percent cover classes approximately doubles the previous class (i.e. 0–1, 1–2, 2–5, 5–10, 10–25, 25–50, 50–75, 75–95, 95–100%), as the human mind is more attuned to these geometric increases in cover than to a linear scale (Peet et al., 1998). Additional measurements made at the center of each plot included tree heights, age, and visual estimates of cover by species. Total vegetative cover was also visually estimated at height intervals of 0–1 m,

1–2 m, 2–5 m, and at 5 m intervals to the top of the canopy to characterize forest vertical structure. Overall canopy openness was measured using a densiometer at four aspects at each of the five subplots and then averaged. Photos of each plot were taken and general site information was recorded (e.g., slope, aspect, etc.). Tree heights were measured using an Impulse 200 laser range finder (Laser Technology Inc., Edgewood, CO) mounted on a monopod, while tree age was determined from bole increment bore samples taken at 1.37 m above ground.

Field plot data were entered into spreadsheets, checked for errors, and then averaged across the five subplots per plot. The averaged data were analyzed to determine the total live and dead BA for each plot, as well as the relative BA of spruce, fir, deciduous, and coniferous species. Total cover in the understory (shrub layer plus subcanopy layer) of conifer species was also estimated as the product of total cover and estimated percent cover coniferous.

During the summers of 2006 and 2007, field data for an additional 120 plots were collected from random locations within the study and set aside for model validation. These field data were processed so as to match the forest structural attributes to be modeled using the 2003–2004 field data (Total BA of forest cover and the relative BA of *Abies*, *Picea spp*., deciduous, and coniferous).

In the two field campaigns, we recorded the location of each plot's center using a Thales Mobile Mapper (Thales Navigation, Inc., Santa Clara, CA) or Trimble Pro-XR (Trimble Navigation Ltd., Sunnyvale, CA) GPS receiver. We post-processed and averaged the coordinates to within one meter of true location using each company's respective software and the National Geodetic Survey's continuously operating reference station (NGS-CORS) data from Grand Marais, MN.

### 2.3. Image data

Field plots and satellite data cover six adjacent WRS-2 Landsat path/row combinations: paths 26–28 rows 26–27 (Fig. 1). Image data (Table 1) were paired by date for both rows in each of the three paths, and all images were acquired for a period that was proximal (2000–2003) to field data collection (2003–2004). All image data were coregistered to a master composite image using the Erdas Imagine AUTOSYNC routine (Leica GeoSystems, 2006). With this routine, 500 evenly distributed image-to-image tie points were automatically identified using sub-pixel correlation. A five meter RMSE threshold was set for each image prior to AUTOSYNC's calculation of a second-order block triangulation (also known as bundle adjustment, see Granshaw, 1980) to co-register input images to the master image. The master composite image was produced from six Global Land Cover Facility (GLCF) GeoCover Landsat images which are precision-orthorectified and geocorrected (source: www.landcover.org).

Following co-registration, each image was inspected for potential haze problems and the visible bands of problematic images (Table 1) were corrected using a haze removal algorithm (ATCOR, Richter, 2003). After this, top-of-atmosphere reflectance was calculated using calibration coefficients provided in the image header files.

For each field plot, differentially corrected GPS data were used to extract reflectance values from each Landsat image for nine pixel locations — the center pixel plus the eight neighboring pixels. From this, the image data and derivative variables used for model development (Table 1) included:

- Landsat bands (1–5, 7);
- TC1, TC2, and TC3 — Tasseled Cap derivatives of brightness, greenness, and wetness (Crist & Kauth, 1986);
- SAVI — soil-adjusted vegetation index (Huete, 1988) — similar to the normalized difference vegetation index (NDVI, Rouse et al., 1974; Tucker, 1979), but with a correction factor that limits soil background effects;
- SR — simple NIR/RED ratio (Jordan, 1969);
- MSI — moisture stress index (Rock et al., 1986) — originally used to discriminate and quantify forest decline (Vogelmann & Rock, 1988) and moisture stress (Hunt & Rock, 1989);
- GEMI — global environmental monitoring index (Pinty & Verstraete, 1992), which was first formulated to minimize atmospheric effects in AVHRR data; and
- SVR — shortwave infrared/visible ratio (derived by the authors), which uses the mean of the two shortwave infrared (SWIR) bands divided by the mean of the three visible bands.

Variance among reflectance values for each of the nine pixel locations was assessed prior to analyses. For each plot, pixel locations that had reflectance greater than one standard deviation from the focal mean were excluded from further analyses. The remaining pixel locations were then linked to the field data associated with the center pixel's location.

Most of the image derivatives were selected as they have been routinely used to study forest structural parameters (see Asner et al., 2003). In particular, indices using the SWIR bands were chosen as this region of the electromagnetic spectrum has been demonstrated to be sensitive to forest density and tree size (Cohen & Spies, 1992; Cohen et al., 1995; Hansen et al., 2001; Lu et al., 2004). The SVR index was developed by the authors specifically to provide a SWIR-based index that excluded NIR.

### 2.4. Initial forest stratification

In an effort to partition potential errors and maximize PLS model precision, both imagery and model development were stratified

**Table 1**
Landsat satellite sensor data and derivatives used in analyses

| Sensor | Path | Row | Date | BLU | GRN | RED | NIR | SWIR5 | SWIR7 | TC1 | TC2 | TC3 | GEMI | MSI | SAVI | SR | SVR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TM | 26 | 26:27 | 3/3/02* | x | x | x | | x | x | x | x | x | x | | x | | |
| ETM+ | 26 | 26:27 | 5/14/02* | x | x | | x | x | x | x | x | | | x | | | x |
| ETM+ | 26 | 26:27 | **7/1/02** | | | x | x | x | x | x | x | x | x | | x | x | x |
| TM | 26 | 26:27 | **9/27/02** | x | x | x | x | x | x | | x | | x | | x | | x |
| ETM+ | 27 | 26:27 | 2/27/01* | x | | | | x | | x | | x | | | x | x | x |
| ETM+ | 27 | 26:27 | **4/29/00*** | x | x | x | x | x | x | x | x | x | x | x | x | | x |
| ETM+ | 27 | 26:27 | 5/21/02 | | x | x | x | x | | x | | x | x | x | x | | x |
| ETM+ | 27 | 26:27 | 7/5/01 | x | | x | | | x | | x | | | x | | | |
| ETM+ | 27 | 26:27 | 8/25/02 | x | x | | | x | x | | x | | x | | | x | |
| TM | 27 | 26:27 | 10/7/03 | x | | x | x | x | | x | x | | x | x | | | x |
| TM | 28 | 26:27 | 2/10/02* | x | x | x | | x | x | x | | x | | | | x | x |
| ETM+ | 28 | 26:27 | 4/20/00* | x | x | | x | x | x | x | x | x | | x | | | x |
| ETM+ | 28 | 26:27 | 6/26/01 | x | x | x | | x | x | x | x | x | | | | x | x |
| ETM+ | 28 | 26:27 | **7/15/02** | | | x | x | x | x | x | x | | x | | x | x | |
| ETM+ | 28 | 26:27 | **8/26/00** | x | x | x | x | x | x | | x | x | | | | | x |

Images subject to haze correction in bold, and images used to stratify conifers from deciduous trees indicated with asterisks.

according to three broad structural classes: 1) pure conifer; 2) mixed forest where conifer and hardwood trees share the overstory; and 3) hardwood forest with an understory conifer component. To do this, Landsat imagery from spring (leaf-off and high sun angle) and winter (snow cover) (Table 1) were used to stratify coniferous forest area, and reference data, into the three classes listed above. Low altitude, digital, aerial photographs (~70% leaf-off), taken on 5 October 2006 (0.2 m pixel resolution), were used as ground truth for identifying the presence of coniferous forest in leaf-off spring Landsat imagery. However, because of the presence of copious amounts of sphagnum moss (*Sphagnum* spp.) and broad-leaved, evergreen, vegetation (e.g., *Ledum groenlandicum*, *Kalmia polifolia*, and *Chamaedaphne calyculata*) associated with treeless bogs and lowland black spruce stands, coniferous forest area was greatly overestimated. Thus, to mask out remaining non-forest vegetation, we used thresholds applied to Tasseled Cap brightness (TC1) calculated from winter Landsat data. Because snow depth archives for northeast Minnesota (http://climate.umn.edu) indicated between 0.5–1.5 m of snow cover for all three Landsat paths within the study area, we assumed that more prostrate vegetation, including *Ericaceae*-dominated brush and mosses, would be sufficiently covered by snow and, hence, hidden from view. Indeed, areas of high brightness (TC1), associated with snow-covered *Ericaceae*-dominated brush, were easily distinguished from coniferous forest using this methodology. Summer NDVI imagery (leaf-on) was then used to distinguish mixed forest where conifers shared the overstory from mixed stands where conifers occupied only the understory. Again, the 5 October 2006 aerial photography was used as reference data.

### 2.5. PLS regression model development

Separate PLS regression models were developed for the three forest structure classes within each of the three Landsat paths (26, 27, and 28) to predict total BA (TBA) of all standing live trees and the relative BAs of four specific forest components: fir (FIR), spruce (SPRUCE), coniferous (CON), and deciduous (DEC) trees. The stratification of PLS model development by Landsat path was necessary as each contained a different number and/or set of image dates (Table 1). Thus, a recursive backward elimination variable selection procedure, similar to that described by Forina et al. (1999), was used to pre-select the most relevant image variables from the full set available for each respective satellite path. To do this, an initial PLS regression model was fit using all image variables (**X**-vector), but holding the forest variables (**Y**-vector) to only the five listed above. In the first iteration the blue band was removed followed by fitting a new PLS model. If the PRESS statistic of this new model was lower than PRESS for the full model, the blue band was excluded from further analyses and the new, lower PRESS became the test criterion for the next iteration. This continued until all image data were assessed.

After the best set of image variables was determined for paths 26 and 27, a forward step procedure was used to iteratively assess additional ground variables for their ability to improve the PRESS statistic (Table 2). Because an insufficient number of field plots were available for path 28, path 27's modeled results from the path 27–28 overlap region (~70 km wide) were used instead to develop training data for path 28 PLS model development. Although such an approach to the development of training data is less preferred than the use of ground data, the decline in mapping accuracy using image-derived training data can be very small for immediately adjacent scenes (Knorn et al., 2008). In this case, the area modeled using these training data represented <17% of the total study area (see Fig. 1). Since these training data consisted of only the five structure variables listed above, iterations for the best PLS regression ended after all path 28 image data were assessed.

PLS regression model results are useful as diagnostic tools for determining salient spectral regions and, in this case, seasons for optimal mapping of forest structure using Landsat data. Thus, to facilitate these analyses, the absolute value of resulting PLS regression component loadings (hereafter referred to simply as loadings) for TBA, FIR, SPRUCE, DEC, and CON were aggregated within each Landsat path, summed by month and band, and then scaled by the associated number of image dates and bands.

**Table 2**
Forest variables used in PLS regression analyses by satellite path

| Path | TBA | FIR | SPRUCE | DEC | CON | *i*SPRUCE | *i*CON | iDEC |
|---|---|---|---|---|---|---|---|---|
| 26 | x | x | x | x | x | x | | |
| 27 | x | x | x | x | x | x | x | x |
| 28 | x | x | x | x | x | | | |

The prefix (*i*) indicates (% cover+relative BA)/2 for the respective variable.

**Table 3**
PLS regression model cross validation and statistics by Landsat path

| | Path | | | | | |
|---|---|---|---|---|---|---|
| | 26 | | 27 | | 28 | |
| Image variables used | 40 | | 51 | | 49 | |
| Forest variables used | 6 | | 8 | | 5 | |
| PRESS | 0.46 | | 0.47 | | 0.43 | |
| Factors used | 6 | | 10 | | 6 | |
| Model variation explained (%) | 89.9 | | 90.9 | | 82.5 | |
| Field data variation explained (%) | 79.7 | | 80.5 | | 82.6 | |
| Model Pr>F | 0.0001 | | 0.0001 | | 0.0001 | |
| Forest parameter statistics | $R^2$ | RMSE | $R^2$ | RMSE | $R^2$ | RMSE |
| Total BA (TBA) | 0.62 | 4.75 | 0.62 | 4.71 | 0.70 | 1.83 |
| Relative BA Abies (FIR) | 0.63 | 4.99 | 0.65 | 6.05 | 0.80 | 3.64 |
| Relative BA Picea (SPRUCE) | 0.88 | 12.32 | 0.88 | 12.52 | 0.87 | 9.00 |
| Relative BA deciduous (DEC) | 0.88 | 9.77 | 0.86 | 9.91 | 0.88 | 5.57 |
| Relative BA coniferous (CON) | 0.86 | 9.30 | 0.86 | 9.78 | 0.87 | 5.35 |

## 3. Results

### 3.1. PLS regression models

Initially, we stratified imagery and PLS model development according to three broad structural classes in an effort to isolate potential sources of error and improve model performance. However, this stratification approach did not yield model improvements over the pooled data. Thus, we report only results for the pooled models (summarized in Table 3). The map-derived "ground" data used for path 28, derived from neighboring path 27 prediction results, provided the best PLS models with 1) the lowest PRESS statistic, 2) highest percentage of variation in field data explained, 3) the best $R^2$ values for both TBA (0.70, Fig. 2A) and FIR (0.80, Fig. 2B), and 4) the lowest RMSE of prediction (Table 3). The overall PLS regression model results for paths 26 and 27 were similar with both explaining ~80% of the variation in the field data (Table 3). However, by definition, the lower PRESS statistic and fewer latent variables used for the path 26 PLS model indicates a better model fit than path 27 (Table 3). PLS prediction results for SPRUCE, DEC, and CON were all similar for the three different models (path 26, 27, 28) with $R^2$ values ranging between 0.86 and 0.89 (Table 3, Figs. 2C, 3A, B).

### 3.2. PLS component loadings

In each Landsat path the SWIR-visible ratio (SVR) had the highest loadings of all image variables, followed closely by Tasseled Cap wetness (TC3) (Fig. 4A). Conversely, the lowest loadings showed no obvious trends as most weak variables had been removed by the stepwise selection procedure discussed above. Seasonal analysis showed that Landsat data acquired in February (with snow cover) had the highest

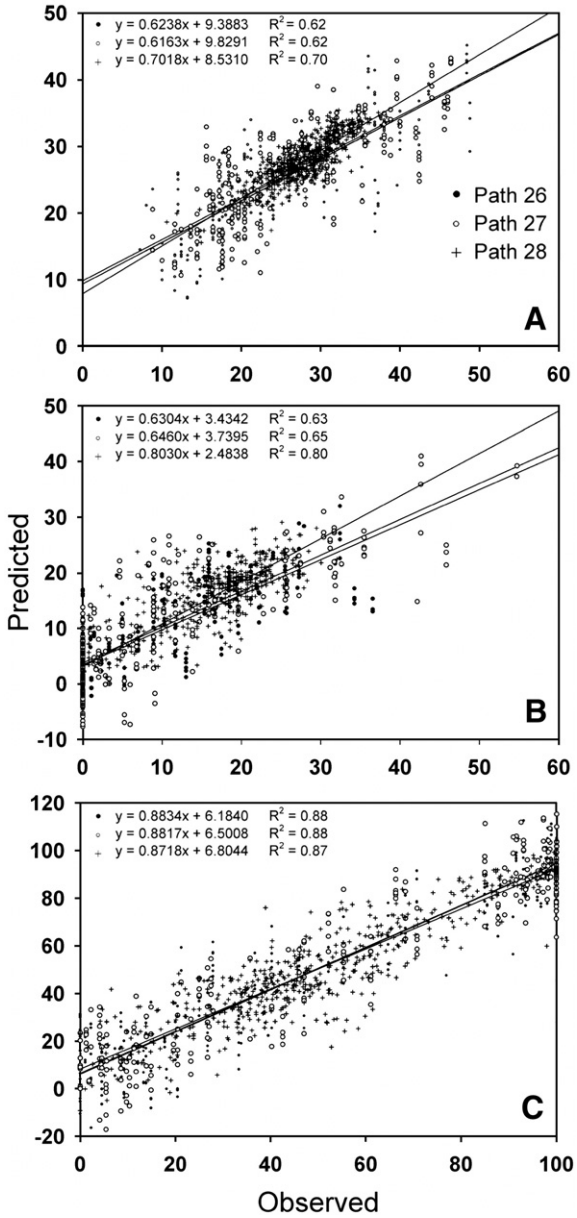


**Fig. 2.** PLS model basal area (A) and relative basal area (B and C) predictions versus observed data for TBA (A), FIR (B), and SPRUCE (C) in Landsat paths 26, 27, and 28.

loadings where they were used (Fig. 4B). March Landsat data (with snow cover) were used in place of February data for path 26. However, the highest component loadings recorded for path 26 were from September and May rather than the March winter data (Fig. 4B).

Path 26 ratio images produced from the 3 March 2002 Landsat-5 data revealed the presence of systematic noise patterns at approximately every eighth scan line throughout the entire image. These radiometric anomalies were dissimilar from, and in excess of, typical coherent noise levels that are characteristic of Landsat TM data. Visual inspection of the raw band data revealed only minor indications of these radiometric flaws, and image metadata provided no indication of any suboptimal sensor anomaly or band quality codes.

Assessment of aggregated PLS component loadings by band for the five separate forest structural attributes show SVR provided the highest absolute values for SPRUCE and FIR in each path (Fig. 5A, B, C), while TC3 and MSI rivaled SVR loadings for DEC and CON in paths 26 and 27, respectively. Visible blue ranked highest for TBA in paths 26 and 27, while MSI was highest for TBA in path 28. The high PLS loading on visible blue for TBA (Fig. 5A, B) came primarily from the February data in path 27 and from March and May data in path 26; all with $R^2$ values between 0.27 and 0.30 (Table 4). There was no consensus among the lowest absolute PLS loading for the five structural attributes; however, many were from visible green, visible red, and the greenness indices (i.e., GEMI, SAVI, SR, and TC2) (Fig. 5).

In terms of image date, February PLS loadings ranked the highest for almost all forest structural attributes within paths 27 and 28 (Fig. 6). The exceptions were FIR (April) and TBA (June) in paths 27 and 28, respectively (Fig. 6B, C). For Path 26, where winter imagery (March) was affected by excessive systematic noise, May ranked highest for SPRUCE, DEC, and CON; September for FIR; and July imagery for TBA (Fig. 6A). July Landsat data were among the least important in paths 26 and 28, while patterns were not as clear in path 27 (Figs. 6, 4B).

### 3.3. PLS model validation

In addition to the cross validation procedure performed as an intrinsic part of the PLS regression procedure (Table 3), an additional validation was performed using field data collected in the summers of 2006 and 2007. Forest structural data derived from these field data were used to assess the accuracy of model predictions for TBA, FIR, SPRUCE, DEC, and CON structure components. These field data are representative of all the major forest cover types found in the study region (120 plots); including forest cover types with spruce budworm host species (35 plots). The independent validation data are located within the overlap region of Landsat paths 26 and 27, for which the path 27 models were used in mapping structure. As such, the independent validation results presented below are relevant only for the path 27 PLS models; accounting for ~70% of the study area (Fig. 1).

Ordinary least-squares regression (OLR) was performed between PLS predictions of forest structural parameters and independent field validation data for 1) all forest cover types combined (Table 5 A) and 2) for plots composed primarily of host species (Table 5B). The relationship

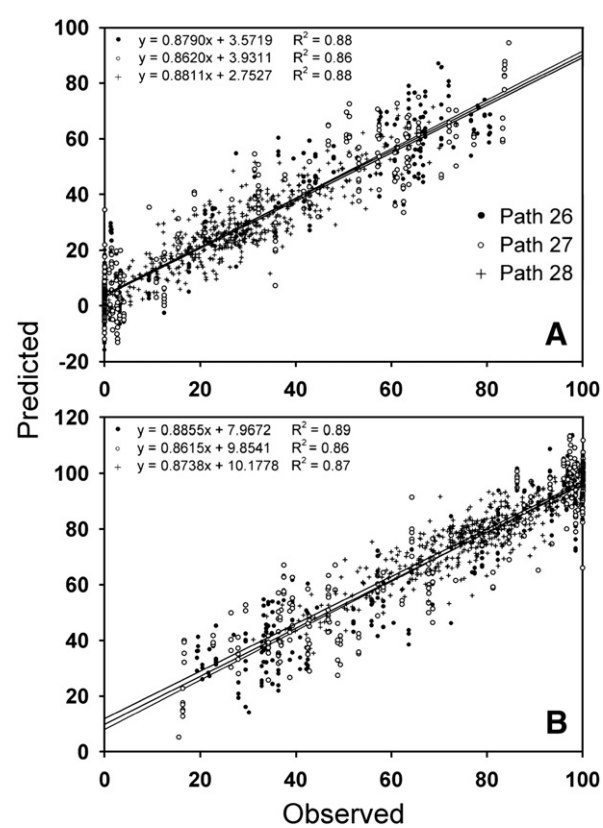


**Fig. 3.** PLS model relative basal area predictions versus observed data for DEC (A) and CON (B) in Landsat paths 26, 27, and 28.

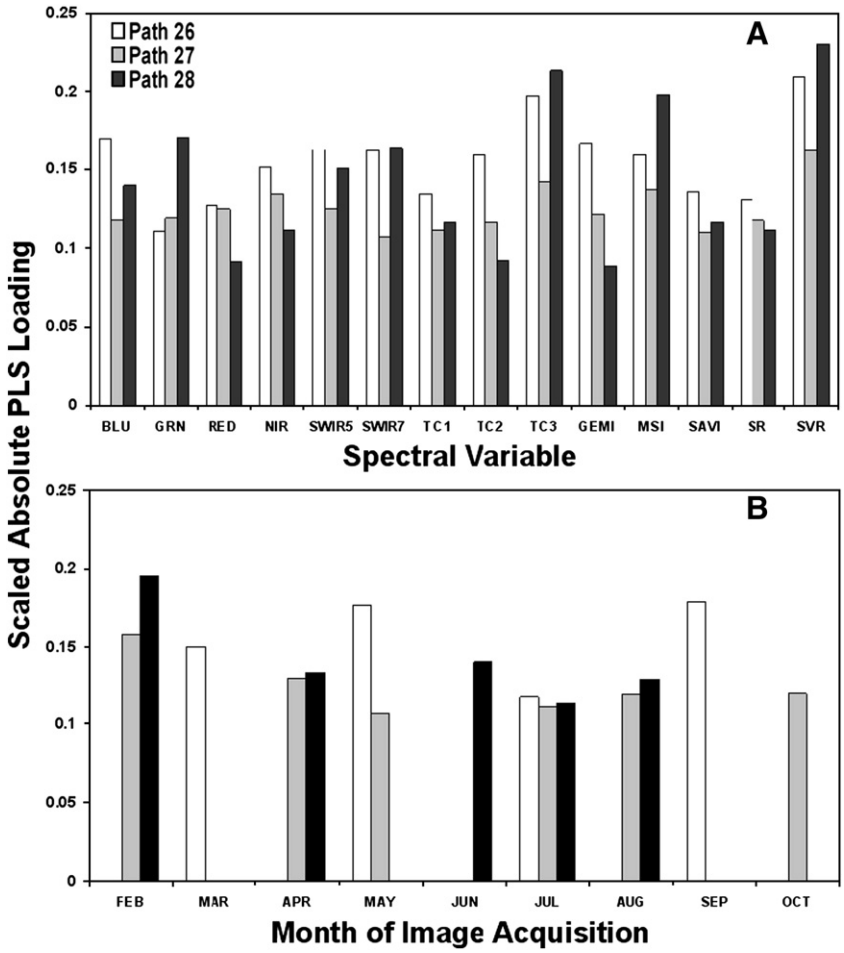


**Fig. 4.** Total absolute PLS loading weights by band (A) scaled by number of images per path and by month (B) scaled by the number of bands per image.

between predicted and measured TBA was weak when all cover types were combined ($R^2_{ALL}$=0.22, $RMSE_{ALL}$=7.82 $m^2$/ha), whereas the TBA relationship was clearly better when using only validation data taken from within host-species cover types ($R^2_{HOST}$=0.63, $RMSE_{HOST}$=4.81 $m^2$/ha). Relationships between predicted and observed values for the four relative BA variables were also stronger within host-specific cover types than when all forest cover types were combined (Table 5). Regression slope coefficients for host-specific validation results were very consistent (0.67–0.80, SE 0.03), while slopes among the non-host-specific plots were much more variable (0.41–1.62, SE 0.26) (Table 5).

In general, PLS regression models worked very well within stands composed primarily of host tree species (spruce and/or fir), but were less precise when applied to stands dominated by other conifer species. Within tamarack stands, independent validation revealed a strong relationship between observed and predicted TBA ($R^2$=0.80), with FIR predicted accurately at 0%. However, SPRUCE was consistently over estimated within tamarack stands at ~45% through the observed range from 5–28%. Conversely, for red/white pine and cedar, there was little to no relationship between observed and predicted TBA. The three field plots located in jack pine exhibited a much better relationship, but there were too few field plots to be conclusive. In non-host stands SPRUCE was over estimated at 40–75% for instances where observed SPRUCE was only 1–7%.

## 4. Discussion

### 4.1. Important spectral regions

The fact that indices containing SWIR (e.g., SVR, TC3, MSI) were important for predicting forest BA was not unexpected. Previous studies conducted in largely coniferous forests have shown that the SWIR bands of Landsat are especially responsive to forest vegetation density (Brockhaus & Khorram, 1992; Franklin et al., 2000; Horler & Ahern, 1986), volume (Ardö, 1992), and leaf area index (Brown et al., 2000). Landsat TM Wetness (TC3), a contrast of SWIR with the visible and near infrared bands, has been advanced as providing the most accurate structure information in closed canopy forests due, in part, to an apparent insensitivity to topographical effects (Cohen & Spies, 1992; Cohen et al., 1995). Hansen et al. (2001) also found good correlations between TC3 and a structural complexity index derived from the application of principal components analysis (PCA) on structure measurements of mature to old growth coniferous forests (>150 years).

The importance of SVR to prediction within each of the Landsat paths for all but one (TBA) of the forest structural parameters was unexpected (Fig. 5). The better performance of SVR over the other SWIR-related indices eludes specific explanation, though at least two factors may be relevant. First, using averages of both the three visible bands and the two SWIR bands in the calculation of SVR partially dampens the random effects of coherent noise (i.e., patterns of low-level, periodic noise present in all Landsat data) by averaging highly correlated bands with differing noise. Second, a spectral relationship between total foliar pigmentation, chlorophyll content, vegetation density, and canopy architecture may be enhanced using a ratio of these broad spectral regions. For example, both TM5 and TM7 are known to correlate inversely with conifer forest age ($r$=−0.62 and −0.59, respectively), while TM7 alone correlates inversely with conifer basal area ($r$=−0.48) (Brockhaus & Khorram, 1992), and both vary inversely with vegetation density (Ahern et al., 1991; Ardö, 1992; Franklin et al., 2000; Horler & Ahern, 1986).

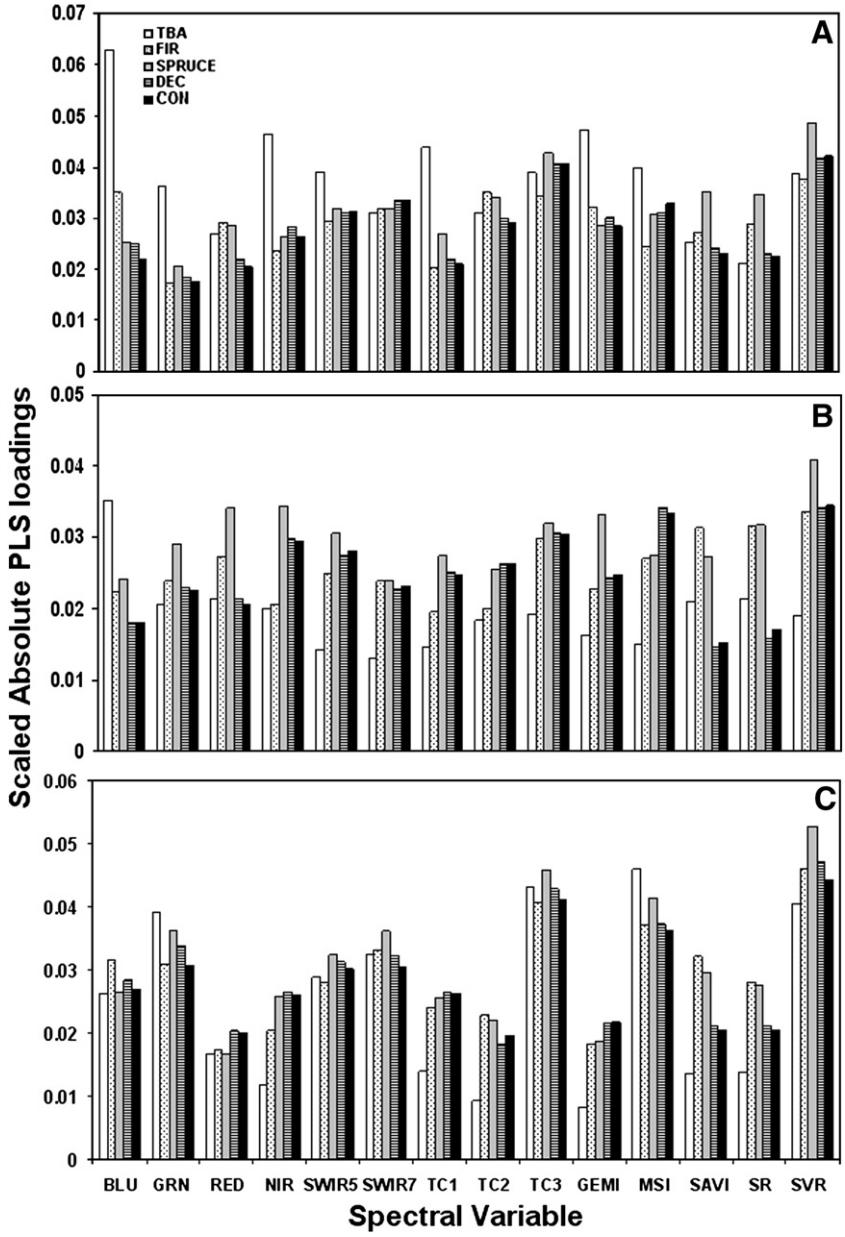


**Fig. 5.** Scaled absolute PLS band loadings for the five structure variables in paths 26 (A), 27 (B), and 28 (C).

The responsiveness of visible bands (especially red) to vegetation biomass (Roy & Ravan, 1996) and other structural properties is well documented (Brown et al., 2000; Goetz & Prince, 1996; Tucker, 1979; Turner et al., 1999). Specifically, coniferous forest reflectance in the visible varies inversely with biomass parameters such as basal area (Franklin, 1986). In addition to red, visible blue provides information relating to coniferous forest species and canopy condition (Nelson et al., 1984; Tucker, 1978). Nelson et al. (1984) found that when analyzed alone or in combination with NIR and/or SWIR, visible blue reflectance contained significant, unique spectral information for distinguishing northern forest cover types. Balsam fir foliage is similar to black spruce and white spruce in terms of visible blue reflectance, but has much lower visible green and red reflectance (see Pinard & Bannari, 2003). As such, balsam fir appears more bluish-green than either spruce or pine canopies in our low altitude color aerial photography. Here, the importance of visible blue with respect to FIR is evident in path 26 (Fig. 5A), which has the highest proportion of balsam fir of the three Landsat paths analyzed. In this case, visible blue and TC1 (Brightness) were essentially tied for second highest absolute PLS loading, behind SVR, among the 14 image variables analyzed (Fig. 5A). Image brightness (e.g., albedo; TC1, PCA1) is known to correlate with vegetation amount (Franklin, 1986) as well as being indicative of forest productivity (Ollinger et al., 2007).

We therefore conclude that the SVR ratio captures a unique spectral relationship linked to forest age/size class, density, and potentially species-specific absorption of visible light which is missed using individual SWIR bands or commonly used spectral derivatives (TC3, MSI or NDVI-related indices).

### 4.2. Seasonal factors

The unique advantage afforded by winter Landsat data (with snow cover) for forest structure modeling and mapping was anticipated based on earlier investigations using aerial photography (e.g., Sayn-

**Table 4**
Ordinary least-squares regression $R^2$ values between salient image variables and forest structure measurements

| Path | Band | Month | TBA | FIR | SPRUCE | DEC | CON |
|---|---|---|---|---|---|---|---|
| 26 | SVR | Mar | 0.01 | 0.21 | 0.42 | 0.35 | 0.36 |
| 26 | SVR | May | 0.32 | **0.45** | **0.75** | **0.73** | **0.74** |
| 26 | SVR | Jul | **0.44** | 0.33 | 0.57 | 0.52 | 0.55 |
| 26 | SVR | Sep | 0.26 | 0.35 | 0.57 | 0.53 | 0.54 |
| 26 | BLUE | Mar | **0.30** | 0.03 | 0.04 | 0.05 | 0.05 |
| 26 | BLUE | May | **0.30** | **0.28** | **0.48** | **0.48** | **0.49** |
| 26 | BLUE | Jul | 0.13 | 0.02 | 0.06 | 0.08 | 0.09 |
| 26 | BLUE | Sep | 0.15 | 0.05 | 0.10 | 0.10 | 0.10 |
| 27 | SVR | Feb | 0.02 | 0.26 | 0.51 | 0.45 | 0.45 |
| 27 | SVR | Apr | 0.20 | 0.21 | 0.49 | 0.57 | 0.57 |
| 27 | SVR | May | **0.32** | 0.25 | **0.58** | **0.67** | **0.67** |
| 27 | SVR | Jul | 0.27 | 0.26 | 0.50 | 0.51 | 0.53 |
| 27 | SVR | Aug | 0.31 | **0.28** | 0.52 | 0.53 | 0.55 |
| 27 | SVR | Oct | 0.18 | 0.18 | 0.24 | 0.24 | 0.25 |
| 27 | BLUE | Feb | **0.27** | 0.00 | 0.00 | 0.00 | 0.01 |
| 27 | BLUE | Apr | 0.08 | 0.09 | **0.17** | **0.23** | **0.23** |
| 27 | BLUE | May | 0.06 | 0.05 | 0.13 | 0.19 | 0.19 |
| 27 | BLUE | Jul | 0.00 | 0.01 | 0.00 | 0.01 | 0.00 |
| 27 | BLUE | Aug | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 |
| 27 | BLUE | Oct | 0.00 | **0.15** | 0.12 | 0.10 | 0.09 |
| 28 | SVR | Feb | 0.11 | 0.41 | 0.59 | 0.65 | 0.65 |
| 28 | SVR | Apr | 0.50 | 0.38 | 0.60 | **0.70** | **0.69** |
| 28 | SVR | Jun | **0.57** | **0.49** | **0.62** | 0.63 | 0.61 |
| 28 | SVR | Jul | 0.55 | 0.41 | 0.52 | 0.53 | 0.51 |
| 28 | SVR | Aug | 0.53 | 0.33 | 0.49 | 0.56 | 0.55 |
| 28 | BLUE | Feb | **0.27** | 0.02 | 0.00 | 0.00 | 0.00 |
| 28 | BLUE | Apr | 0.19 | **0.13** | **0.28** | **0.42** | **0.45** |
| 28 | BLUE | Jun | 0.05 | 0.03 | 0.05 | 0.07 | 0.08 |
| 28 | BLUE | Jul | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 28 | BLUE | Aug | 0.04 | 0.07 | 0.07 | 0.08 | 0.08 |
| 28 | MSI | Feb | 0.09 | 0.15 | 0.31 | 0.44 | 0.46 |
| 28 | MSI | Apr | **0.40** | 0.11 | 0.30 | **0.47** | **0.49** |
| 28 | MSI | Jun | 0.38 | **0.23** | **0.35** | 0.38 | 0.37 |
| 28 | MSI | Jul | 0.20 | 0.07 | 0.08 | 0.06 | 0.05 |
| 28 | MSI | Aug | 0.34 | 0.02 | 0.06 | 0.09 | 0.08 |

Highest values for each structure attribute in bold.

Wittgenstein, 1961) and Landsat TM imagery (Wolter et al., 1995; Wolter & White, 2002). In general, April and May leaf-off satellite imagery was critical for this study because understory conifers, particularly balsam fir, are not obscured by deciduous overstory foliage (Sayn-Wittgenstein, 1961). Images from these two months yielded the strongest individual relationships between salient spectral variables and forest structure (Table 4). However, the full model results for path 27 and 28 (Fig. 4B) show that imagery with sufficient snow cover may have provided an additional advantage by 1) covering spectrally variable forest undergrowth that can confound forest structure signatures (e.g., Brown et al., 2000; Chen & Cihlar, 1996; White et al., 1995) and 2) by providing a uniformly bright background that accentuates tree crowns and their shadows (Seely, 1949); as well as spectral factors linked to forest density, height, and age (Horler & Ahern, 1986).

The pitfalls of using winter Landsat data for forest structure mapping relate principally to the effect of low sun angle illumination on strong topography, but canopy snow retention may be equally problematic (Beaubien, 1979). The terrain within this study area is generally gentle and of minimal consequence. Nevertheless, our decision not to acquire imagery from December or January was intended to minimize sun angle effects. There is, however, a persistent risk of introducing spectral problems associated with snow retention in canopies (Sayn-Wittgenstein, 1961), particularly if imagery is acquired closely following a snowfall event. In this study, image acquisitions were all greater than two weeks after snowfall events, thus reducing this concern. It is possible that the lower March loadings (compared to February) observed for the full PLS model in path 26 (Fig. 6) are related to both poor radiometric quality of these winter Landsat data, as described above, and to retention of canopy snow, but specific determination of any canopy snow effects is not possible.

### 4.3. Accuracy of PLS regression models among non-host conifer stands

PLS models worked very well to predict BA within stands consisting of host tree species (spruce/fir), but were less precise when applied to stands dominated by other conifer species (except tamarack and jack pine). Concurrently, mapping accuracy for FIR within non-host cover types was more robust and encouraging. PLS regression models grossly overestimate the relative amount of spruce within non-host conifer types, such as red pine, white pine, and white cedar. It is important to note, however, that red and white pine dominated forests comprise only a fraction of total forest area compared to the vast spruce/fir-dominated forest type in this region, and the area of cedar is minuscule. The fact that modeled TBA results for jack pine may be better aligned with reality is significant, since jack pine is a close second in terms of total forest area after spruce/fir forests in this region. However, additional ground data are needed to assess the validity of any preliminary relationships regarding jack pine. The existence of a good stand type map (e.g., Wolter et al., 1995) can ensure that the results are interpreted and used correctly.

### 4.4. PLS model accuracy and balsam fir

The lower prediction accuracy of the PLS regression models for FIR in paths 26 and 27 (Fig. 2, Table 3) is likely related to smaller observed

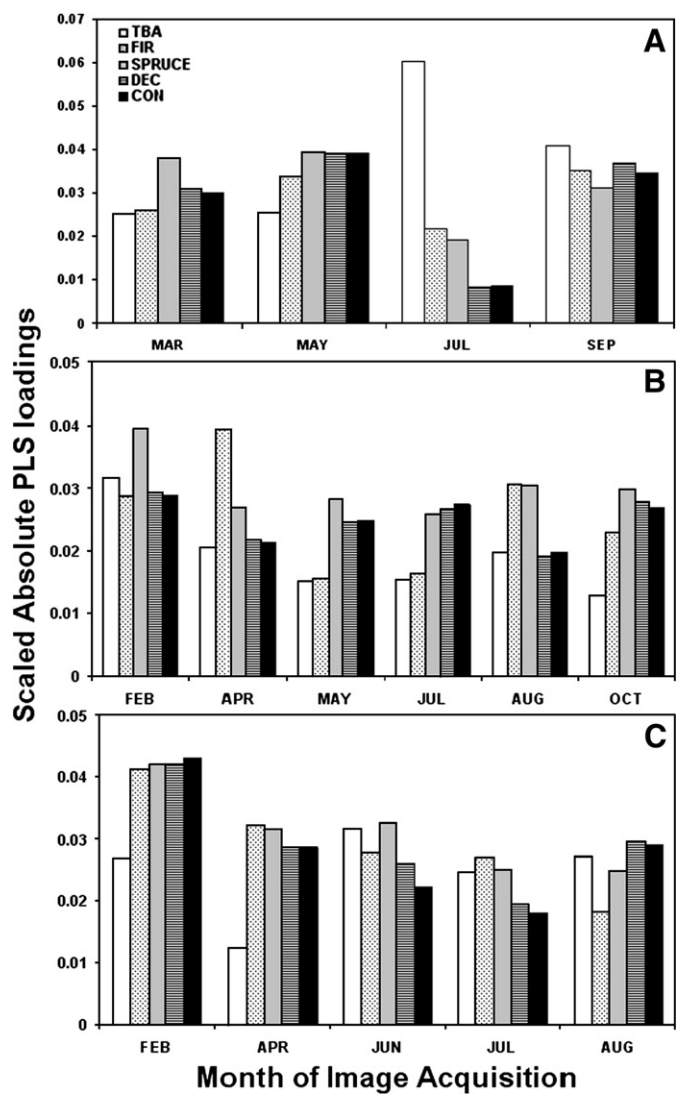


**Fig. 6.** Scaled absolute PLS band loadings by month for the five structure variables for paths 26 (A), 27 (B), and 28 (C).

**Table 5**
Results of independent model validation using 120 plots for all species (A) and for the 35 plots that contained primarily host tree species (B)

| A | Intercept | Slope | $R^2$ | RMSE |
|---|---|---|---|---|
| TBA | 13.84 | 0.27 | 0.22 | 7.82 |
| FIR | 7.67 | 0.41 | 0.69 | 5.55 |
| SPRUCE | 20.46 | 1.25 | 0.81 | 15.26 |
| DEC | 19.80 | 1.62 | 0.75 | 15.29 |
| CON | 10.47 | 1.19 | 0.62 | 18.32 |
| **B** | **Intercept** | **Slope** | $R^2$ | **RMSE** |
| TBA | 5.24 | 0.65 | 0.63 | 4.81 |
| FIR | 3.36 | 0.76 | 0.78 | 3.05 |
| SPRUCE | 12.58 | 0.80 | 0.91 | 10.54 |
| DEC | 5.73 | 0.67 | 0.96 | 5.24 |
| CON | 28.49 | 0.70 | 0.94 | 6.84 |

size classes for balsam fir compared to the spruces. Although, balsam fir is the most abundant conifer species in the region, it frequently occupies only an understory canopy position (<5 m tall) of relatively small stem diameter (<10 cm) usually below quaking and bigtooth aspen (*P. tremuloides* and *P. grandidentata*, respectively), paper birch (*B. papyrifera*), or older-growth white spruce. In other cases, balsam fir exists as either a sparse overstory component among the associates listed above, or rarely in pure stands. In contrast, white spruce in these forests generally consists of larger trees (e.g., 30–60 cm dbh). However, both prediction models slightly overestimate at low relative BA and underestimate at high relative BA (Fig. 2).

### 4.5. Practical ramifications and recommendations

Soon, researchers will have open access to literally hundreds of Landsat images as the United States Geological Survey (USGS) transitions to free electronic data retrieval from the national archive. Thus, PLS regression may become an invaluable tool for modeling a variety of ecological variables and processes given the sheer volume of data available in the Landsat archive.

In this study, we were able to acquire at least four different image dates (with good seasonal diversity) common to both rows in each path. However, in application it is likely that the PLS strategy may not perform as well in study areas where greater cloud frequency limits seasonal diversity of available imagery. For the application described in this study, the optimal number of image dates and seasons has yet to be determined.

As previously noted, it is apparent from both this work and from the literature that SWIR-based indices are especially important for characterizing forest structural attributes. Efforts to map structure in forest ecosystems should continue to include SWIR derivatives, but also investigate the usefulness of SVR. In addition, the importance of high quality winter satellite data (with snow cover) at these and higher latitudes is undeniable, and should be included in future studies of this nature.

## 5. Conclusions

Effective characterization of spruce budworm activity, especially in the context of anthropogenic disturbances, requires comprehensive information on the extent and abundance of its preferred host species, balsam fir, white spruce, and to a lesser extent black spruce. We demonstrate the utility of multi-temporal Landsat data and PLS regression to simultaneously measure relationships between multiple spectral and host-specific structural variables. In general, the PLS models performed very well within the limits of host-specific cover types, while extrapolation to red/white pine and cedar was poor. By and large, the ability to measure structural aspects of host species within pine- and cedar-dominated stands was poor, as ground data for these stands was not gathered, and hence, the spectral variability within these stands was not factored into the PLS models. Nevertheless, the TBA of pure tamarack stands and the FIR in non-host conifer stands were both predicted remarkably well using PLS models.

We identified important seasons and spectral variables for mapping host-species forest structure in this region using Landsat data. For example, imagery from February (with snow cover) and May were found to be very important for host-species structure mapping in Minnesota and neighboring Ontario apparently because 1) the bright snow cover accentuates conifer crowns, 2) visibility of understory conifers is improved for these leaf-off dates (particularly May due to the higher sun angle), while 3) uniform snow cover in February hides the otherwise spectrally diverse forest floor. In terms of spectral sensitivity, we found SVR was similar or more responsive to forest structure than Wetness (TC3), MSI, or either SWIR band alone, except TBA where visible blue was most important in two of the three Landsat paths analyzed.

As the availability of Landsat data expands, the PLS regression strategy provides an attractive alternative for handling multiple, collinear image variables. Future research should investigate whether PLS regression can sufficiently distinguish forest structure among a greater diversity of forest types, and/or whether data having higher spatial resolution (e.g., ASTER or SPOT) or longer wavelength (e.g., SAR) will improve results for all forest types in this region.

## Acknowledgments

This research was supported by funding from the U.S. Forest Service via the National Fire Plan (Project 01.NCS.C.2) and a grant from the USDA Cooperative State Research, Education and Extension Service (CSREES) Managed Ecosystems program (2005-35101-16342), as well as fellowship support from the Department of Forest and Wildlife Ecology at the University of Wisconsin. Thanks to Brian Miranda, Nathan Aspelin, and James Walter for assistance with fieldwork, to David Helmers and Peter Crump for programming support, to Rachel Spartz for data entry, and to the Natural Resources Research Institute of the University of Minnesota - Duluth for transportation assistance.

## References

Ahern, F. J., Erdle, T., Maclean, D. A., & Kerppeck, I. D. (1991). A quantitative relationship between forest growth rates and Thematic Mapper reflectance measurements. *International Journal of Remote Sensing*, *12*, 387–400.

Ardö, J. (1992). Volume quantification of coniferous forest compartments using spectral radiance recorded by Landsat Thematic Mapper. *International Journal of Remote Sensing*, *13*(9), 1779–1786.

Arenas-Garcia, J., & Camps-Valls, G. (2007). Feature extraction from remote sensing data using Kernel Orthonormalized PLS. *Proceedings of IEEE Geoscience and Remote Sensing Symposium (IGARSS), 23–28 July, Barcelona, Spain* (pp. 258–261).

Asner, G. P., Hicke, J. A., & Lobell, D. B. (2003). Per-pixel analysis of forest structure: Vegetation indices, spectral mixture analysis and canopy reflectance modeling. In M. Wulder & S.E. Franklin (Eds.), *Methods and Applications for Remote Sensing of Forests: Concepts and Case Studies* (pp. 209–254). New York: Kluwer Academic Publishers.

Baker, W. L. (1989). Landscape ecology and nature reserve design in the Boundary Waters Canoe Area, Minnesota. *Ecology*, *70*(1), 23–35.

Baker, W. L. (1992). Effects of settlement and fire suppression on landscape structure. *Ecology*, *73*, 1879–1887.

Batzer, H. O. (1969). Forest character and vulnerability of balsam fir to spruce budworm in Minnesota. *Forest Science*, *15*, 17–25.

Beaubien, J. (1979). Forest type mapping from Landsat digital data. *Photogrammetric Engineering and Remote Sensing*, *45*, 1135–1144.

Bergeron, Y., & Leduc, A. (1998). Relationships between change in fire frequency and mortality due to spruce budworm outbreak in the Southeastern Canadian boreal forest. *Journal of Vegetation Science*, *9*, 493–500.

Blais, J. R. (1983). Trends in the frequency extent and severity of spruce budworm outbreaks in eastern Canada. *Canadian Journal of Forest Research*, *13*, 539–547.

Brockhaus, J. A., & Khorram, S. (1992). A comparison of SPOT and Landsat-TM data for use in conducting inventories of forest resources. *International Journal of Remote Sensing*, *13*(6), 3035–3043.

Brown, L., Chen, J. M., Leblanc, S. G., & Cihlar, J. (2000). A shortwave infrared modification to the simple ratio for LAI retrieval in boreal forests: An image and model analysis. *Remote Sensing of Environment*, *71*, 16–25.

Chen, J. M., & Cihlar, J. (1996). Retrieving leaf area index of boreal conifer forests using Landsat TM images. *Remote Sensing of Environment*, *55*, 153–162.

Clark, D. B., Read, J. M., Clark, M. L., Cruz, A. M., Dotti, M. F., & Clark, D. A. (2004). Application of 1-M and 4-M resolution satellite data to ecological studies of tropical rain forests. *Ecological Applications*, *14*(1), 61–74.

Cohen, W. B., Maiersperger, T. K., Gower, S. T., & Turner, D. P. (2003). An improved strategy for regression of biophysical variables and Landsat ETM+ data. *Remote Sensing of Environment*, *84*, 561–571.

Cohen, W. B., & Spies, T. A. (1992). Estimating structural attributes of douglas–fir/western hemlock forest stands from Landsat and SPOT imagery. *Remote Sensing of Environment*, *41*, 1–17.

Cohen, W. B., Spies, T. A., & Fiorella, M. (1995). Estimating the age and structure of forests in a multi-ownership landscape of western Oregon, USA. *International Journal of Remote Sensing*, *16*, 721–746.

Coops, N. C., Smith, M. L., Martin, M. E., & Ollinger, S. V. (2003). Prediction of eucalypt foliage nitrogen content from satellite-derived hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, *41*(6), 1338–1346.

Crist, E. P., & Kauth, R. J. (1986). The tasseled cap de-mystified. *Photogrammetric Engineering and Remote Sensing*, *52*(1), 81–86.

Curran, P., & Hay, A. (1986). The importance of measurement error for certain procedures in remote sensing at optical wavelengths. *Photogrammetric Engineering and Remote Sensing*, *52*(2), 229–241.

Dale, V. H., Joyce, L. A., McNulty, S., Neilson, R. P., Ayres, M. P., Flannigan, M. D., et al. (2001). Climate change and forest disturbances. *Bioscience*, *51*, 723–734.

Erickson, G., & Hastings, A. (1978). Spruce budworm defoliation in Minnesota: 1954–1977. *U.S.D.A Forest Service, Research Note NC-243* 4 pp.

Fleming, R. A., Candau, J. N., & Mcalpine, R. S. (2002). Landscape-scale analysis of interactions between insect defoliation and forest fire in Central Canada. *Climatic Change*, *55*, 251–272.

Forina, M., Casolino, C., & Pizarromillan, C. (1999). Iterative predictor weighting (IPW) PLS: A technique for the elimination of useless predictors in regression problems. *Journal of Chemometrics*, *13*(2), 165–184.

Franklin, J. (1986). Thematic Mapper analysis of coniferous forest structure and composition. *International Journal of Remote Sensing*, *7*(10), 1287–1301.

Franklin, S. E., Moskal, L. M., Lavinge, M. B., & Pugh, K. (2000). Interpretation and classification of partially harvested forest stands in the Fundy Model Forest using multitemporal Landsat TM data. *Canadian Journal of Remote Sensing*, *26*(4), 318–333.

Frelich, L. E., & Reich, P. B. (1995). Spatial patterns and succession in a Minnesota southern–boreal forest. *Ecological Monographs*, *65*, 325–346.

Geisser, S. (1974). A predictive approach to the random effect model. *Biometrica*, *61*, 101–107.

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: A tutorial. *Analytica Chimica Acta*, *185*, 1–17.

Goetz, S. J., & Prince, S. D. (1996). Remote sensing of net primary production in boreal forest stands. *Agricultural and Forest Meteorology*, *78*(3), 149–179.

Granshaw, S. I. (1980). Bundle adjustment methods in engineering photogrammetry. *The Photogrammetric Record*, *10*(56), 181–207.

Hadley, K. S. (1994). The role of disturbance, topography, and forest structure in the development of a montane forest landscape. *Bulletin of the Torrey Botanical Club*, *121*(1), 47–61.

Hall, F. G., Townshend, J. R., & Engman, E. T. (1995). Status of remote sensing algorithms for estimation of land surface state parameters. *Remote Sensing of Environment*, *51*(1), 138–156.

Hansen, M. J., Franklin, S. E., Woudsma, C., & Peterson, M. (2001). Forest structure classification in the north Columbia Mountains using the Landsat TM Tasseled Cap Wetness component. *Canadian Journal of Remote Sensing*, *27*(1), 20–32.

Healey, S. P., Yang, Z., Cohen, W. B., & Pierce, D. J. (2006). Application of two regression-based methods to estimate the effects of partial harvest on forest structure using Landsat data. *Remote Sensing of Environment*, *101*, 115–126.

Heinselman, M. L. (1973). Fire in the virgin forests of the Boundary Waters Canoe Area, Minnesota. *Quaternary Research*, *3*, 329–382.

Helland, I. S. (1988). On the structure of partial least squares regression. *Communications in Statistics — Simulation and Computation*, *17*(2), 581–607.

Hessburg, P. F., Smith, B. G., Kreiter, S. D., Miller, C. A., Salter, R. B., McNicoll, C. H., et al. (1999). Historical and current forest and range landscapes in the Interior Columbia River Basin and Portions of the Klamath and Great Basins. *Part 1: Linking vegetation patterns and landscape vulnerability to potential insect and pathogen disturbances. USDA Forest Service, Pacific Northwest Research Station. Gen. Tech. Rep. PNW-GTR-458*.

Holland, J. H. (1975). Adaptation in natural and artificial systems. *University of Michigan, Ann Arbor, MI, Internal report (228 pp)*.

Horler, D. N. H., & Ahern, F. J. (1986). Forestry information content of thematic mapper data. *International Journal of Remote Sensing*, *7*, 405–428.

Huete, A. R. (1988). A soil adjusted vegetation index (SAVI). *Remote sensing of the environment*, *25*, 295–309.

Hunt, E. R., & Rock, B. N. (1989). Detection of changes in leaf water content using near- and middle-infrared reflectances. *Remote Sensing of Environment*, *30*, 43–54.

Jarvis, R. M., & Goodacre, R. (2005). Genetic algorithm optimization for pre-processing and variable selection of spectroscopic data. *Bioinformatics*, *21*(7), 860–868.

Jordan, C. F. (1969). Derivation of leaf area index from quality of light on the forest floor. *Ecology*, *50*, 663–666.

Knorn, J., Hostert, P., Radeloff, V., & Kuemmerle, T. (2008). Post-socialist forest cover change in the Ukrainian Carpathians. *Proceedings of 23rd Annual Landscape Ecology Symposium (US-IALE), 6–10 April, Madison, Wisconsin* (pp. 85).

Kowalski, B., Gerlach, R., & Wold, H. (1982). Chemical systems under indirect observation. In K. de Joreskog & S. Wold (Eds.), *Systems under indirect observation* (pp. 191–209). North-Holland, Amsterdam.

Leardi, R., & Gonzalez, A. L. (1998). Genetic algorithms applied to feature selection in PLS regression: how and when to use them. *Chemometrics and Intelligent Laboratory Systems*, *41*(2), 195–207.

Leica Geosystems (2006). *Imagine AutoSync* (pp. 1–20). White paper available at: http://gi.leica-geosystems.com/documents/pdf/IMAGINEAutoSyncWhitePapeFeb06.pdf

Li, X., & Strahler, A. H. (1985). Geometric-optical modeling of a conifer forest canopy. *IEEE Transactions on Geoscience and Remote Sensing*, *GE-23*, 705–721.

Li, L., Ustin, S. L., & Riaño, D. (2007). Retrieval of fresh leaf fuel moisture content using Genetic Algorithm Partial Least Squares (GA-PLS) modeling. *IEEE Transactions on Geoscience and Remote Sensing Letters*, *4*(2), 216–220.

Lindgren, F., Geladi, P., & Wold, S. (1993). The kernel algorithm for PLS. *Journal of Chemometrics*, *7*(1), 45–59.

Lu, D., Mausel, P., Brondízioc, E., & Moran, E. (2004). Relationships between forest stand parameters and Landsat TM spectral responses in the Brazilian Amazon Basin. *Forest Ecology and Management*, *198*(1–3), 149–167.

Marquardt, D. W. (1970). Generalized inverses, ridge regression, biased linear and non-linear estimation. *Technometrics*, *12*, 591–612.

Martin, M. E., Plourde, L. C., Ollinger, S. V., Smith, M. -L., & McNeil, B. E. (2008). A generalizeable method for remote sensing of canopy nitrogen across a wide range of forest ecosystems. *Remote Sensing of Environment*, *112*, 3511–3519.

McDonald, S., Niemann, K. O., Goodenough, D. G., Dyk, A., West, C., Han, T., et al. (2003). Hyperspectral remote sensing of conifer chemistry and moisture. *Proceedings of IEEE Geoscience and Remote Sensing Symposium (pp. 1, 552–554), 21–25 July* British Columbia, Canada: Victoria University.

Nelson, R. F., Latty, R. S., & Mott, G. (1984). Classifying northern forests using Thematic Mapper Simulator data. *Photogrammetric Engineering and Remote Sensing*, *50*(5), 607–617.

Nielsen, A. A. (2002). Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data. *IEEE transactions on image processing*, *11*, 293–305.

Norgaard, L., Saudland, A., Wagner, J., Nielsen, J. P., Munck, L., & Engelsen, S. B. (2000). Interval (iPLS): A comparative chemometric study with an example from near-IR spectroscopy. *Applied Spectroscopy*, *54*(3), 413–419.

Ollinger, S. V., Richardson, A. D., Martin, M. E., Frolking, S. E., Hollinger, D. Y., & Plourde, L. C. (2007). Canopy nitrogen, carbon assimilation and the Albedo of North American Forest Ecosystems. *Proceedings of the American Geophysical Union annual conference, 10–14 December, San Francisco, CA, 88(52), Supplement 1–2*.

Ourcival, J. M., Joffre, R., & Rambal, S. (1999). Exploring the relationships between reflectance and anatomical and biochemical properties in Quercus ilex leaves. *New Phytologist*, *143*(2), 351–364.

Pastor, J., Sharp, A., & Wolter, P. (2005). An application of Markov models to the dynamics of Minnesota's forests. *Canadian Journal of Forest Research*, *35*, 3011–3019.

Peddle, D. R., Hall, F. G., & LeDrew, E. F. (1999). Spectral mixture analysis and geometric-optical reflectance modeling of boreal forest biophysical structure — the influence of canopy closure, understorey vegetation and background reflectance. *Remote Sensing of Environment*, *67*(3), 288–297.

Peet, R. K., Wentworth, T. R., & White, P. S. (1998). A flexible, multipurpose method for recording vegetation composition and structure. *Castanea*, *63*(3), 262–274.

Pinard, V., & Bannari, A. (2003). Spectroradiometric analysis in a hyperspectral use perspective to discriminate between forest species. *Proceedings of IEEE Transactions on Geoscience and Remote Sensing Symposium (pp. 7, 4301–4303), 21–25 July, Toulouse, France*.

Pinty, B., & Verstraete, M. M. (1992). GEMI: A non-linear index to monitor global vegetation from satellites. *Vegetatio*, *101*, 15–20.

Richter, R. (2003). Atmospheric/Topographic Correction for Satellite Imagery: ATCOR-2/3 User Guide, Version 5.5. (200 pp.). Schläpfer, Wessling, Germany: ReSe Applications.

Rock, B. N., Vogelmann, J. E., Williams, D. L., Vogelmann, A. F., & Hoshizaki, T. (1986). Remote detection of forest damage. *BioScience*, *36*(7), 439–445.

Rouse, J. W., Haas, R. H., Schell, J. A., & Deering, D. W. (1974). Monitoring vegetation systems in the Great Plains with ERTS. *Proceedings of the 3rd ERTS Symposium (pp. 1, 309–317), NASA SP-351* Washington D.C.: U.S. Government Printing Office.

Roy, P. S., & Ravan, S. A. (1996). Biomass estimation using satellite remote sensing data — an investigation on possible approaches for natural forest. *Journal of Biosciences*, *21*(4), 535–561.

SAS Institute Inc. (2000). *SAS Users Guide, Version 8* (pp. 2693–2734). Cary, North Carolina: SAS Institute Incorporated.

Sayn-Wittgenstein, L. (1961). Phenological aids to tree species identification on air photographs. *Technical Note No. 104 (26 pp.)* Ottawa, Canada: Forest Research Branch, Canada Department of Forestry.

Scheller, R. M., & Mladenoff, D. J. (2005). A spatially interactive simulation of climate change, harvesting, wind, and tree species migration and projected changes to forest composition and biomass in northern Wisconsin, USA. *Global Change Biology*, *11*(2), 307–321.

Scheller, R. M., Mladenoff, D. J., Crow, T. R., & Sickley, T. A. (2005). Simulating the effects of fire reintroduction versus continued fire absence on forest composition and landscape structure in the Boundary Waters Canoe Area, northern Minnesota, USA. *Ecosystems*, *8*, 396–411.

Seely, H. E. (1949). Air photography and its application to forestry. *Photogrammetric Engineering*, *15*(4), 548–554.

Smith, M. L., Martin, M. E., Plourde, L., & Ollinger, S. V. (2003). Analysis of hyperspectral data for estimation of temperate forest canopy nitrogen concentration: Comparison between an airborne (AVIRIS) and a spaceborne (Hyperion) sensor. *IEEE Transactions on Geoscience and Remote Sensing*, *41*(6), 1332–1337.

Smith, M. L., Ollinger, S., Martin, M., Aber, J., Hallett, R., & Goodale, C. (2002). Direct estimation of aboveground forest productivity through hyperspectral remote sensing of canopy nitrogen. *Ecological Applications*, *12*, 1286–1302.

Song, C. (2007). Estimating tree crown size with spatial information of high resolution optical remotely sensed imagery. *International Journal of Remote Sensing*, *28*, 1–19.

Song, C., Schroeder, T. A., & Cohen, W. B. (2007). Predicting temperate conifer forest successional stage distributions with multitemporal Landsat Thematic Mapper imagery. *Remote Sensing of Environment*, *106*, 228–237.

Song, C., & Woodcock, C. E. (2003). Estimating tree crown size from multiresolution remotely sensed imagery. *Photogrammetric Engineering and Remote Sensing*, *69*(11), 1263–1270.

Spiegelman, C. H., McShane, M. J., Goetz, M. J., Motamedi, M., Yue, Q. L., & Coté, G. L. (1998). Theoretical justification of wavelength selection in PLS calibration: Development of a new algorithm. *Analytical Chemistry*, *70*(1), 35–44.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of Royal Statistical Society*, *36*, 111–133.

Sturtevant, B. R., Gustafson, E. J., Li, W., & He, H. S. (2004). Modeling biological disturbances in LANDIS: a module description and demonstration using spruce budworm. *Ecological Modeling*, *180*(1), 153–174.

Swierenga, H., de Groot, P. J., de Weijer, A. P., Derksen, M. W. J., & Buydens, L. M. C. (1998). Improvement of PLS model transferability by robust wavelength selection. *Chemometrics and Intelligent Laboratory Systems*, *41*(2), 237–248.

Townsend, P. A., Foster, J. R., Chastain, R. A., & Currie, W. S. (2003). Application of imaging spectroscopy to mapping canopy nitrogen in the forests of the central appalachian mountains using Hyperion and AVIRIS. *IEEE Transactions on Geoscience and Remote Sensing*, *41*(6), 1347–1354.

Tucker, C. J. (1978). A comparison of satellite sensor bands for vegetation monitoring. *Photogrammetric Engineering and Remote Sensing*, *44*, 1369–1380.

Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, *8*, 127–150.

Turner, D. P., Cohen, W. B., Kennedy, R. E., Fassnacht, K. S., & Briggs, J. M. (1999). Relationships between leaf area index and Landsat TM spectral vegetation indices across three temperate zone sites. *Remote Sensing of Environment*, *70*(1), 52–68.

Van der Voet, H. (1994). Comparing the predictive capacity of models using a simple randomization test. *Chemometrics and Intelligent Laboratory Systems*, *25*, 313–323.

Vogelmann, J. E., & Rock, B. N. (1988). Assessing forest damage in high-elevation coniferous forests in Vermont and New Hampshire using Thematic Mapper data. *Remote Sensing of Environment*, *24*, 227–246.

White, M. A., & Host, G. E. (2003). Changes in disturbance frequency, age and patch structure from pre-Euro–American to settlement to present in North-Central and North-Eastern Minnesota. *Minnesota Forest Resources Council Report LT-1203a, St. Paul, Minnesota, USA*.

White, H. P., Miller, J. P., Chen, J. M., & Peddle, D. R. (1995). Seasonal change in mean understory reflectance for boreal sites: preliminary results. *Proceedings of the 17th Canadian Symposium on Remote Sensing (pp. 189–194), 13–15 June, Saskatoon, Saskatchewan, Canada*.

Williams, D. W., & Birdsey, R. A. (2003). Historical patterns of spruce budworm defoliation and bark beetle outbreaks in North American coniferous forests: An atlas and description of digital maps. *U.S.D.A. Forest Service General Technical Report NE-308 (37 pp.)*.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In P. R. Krishnaiah (Ed.), *Multivariate Analysis* (pp. 391–420). New York: Academic Press.

Wold, H. (1975). Soft modeling by latent variables; the non-linear iterative partial least squares approach. In J. Gani (Ed.), *Perspectives in Probability and Statistics: Papers in Honour of M.S. Bartlett* (pp. 117–142). London: Academic Press.

Wold, S., Ruhe, A., Wold, H., & Dunn, W. J. (1984). The Collinearity problem in linear regression. The Partial Least Squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, *5*(3), 735–743.

Woodcock, C. E., Collins, J. B., Jakabhazy, V. D., Li, X., Macomber, S. A., & Wu, Y. (1997). Inversion of the Li–Strahler canopy reflectance model for mapping forest structure. *IEEE Transactions on Geoscience and Remote Sensing*, *35*, 405–414.

Wolter, P. T., Mladenoff, D. J., Host, G. E., & Crow, T. R. (1995). Improved forest classification in the northern Lake State using multi-temporal Landsat imagery. *Photogrammetric Engineering and Remote Sensing*, *61*(9), 1129–1143.

Wolter, P. T., & White, M. A. (2002). Recent forest cover type transitions and landscape structural changes in northeast Minnesota. *Landscape Ecology*, *17*, 133–155.

Zang, Y., Zhou, C., & Zhang, Y. (2007). A partial least-squares regression approach to land use studies in the Suzhou–Wuxi–Changzhou region. *Journal of Geographical Sciences*, *17*(2), 234–244.

Zheng, D., Rademacher, J., Chen, J., Crow, T., Bresee, M., Le Moine, J., et al. (2004). Estimating aboveground biomass using Landsat 7 ETM+ data across a managed landscape in northern Wisconsin, USA. *Remote Sensing Environment*, *93*, 402–411.