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.

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).

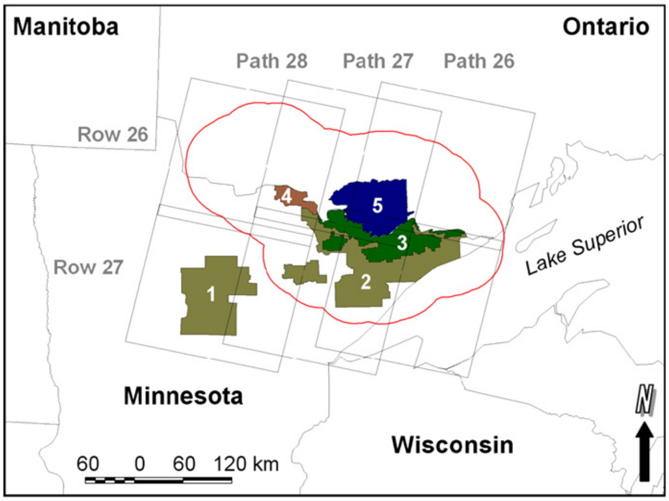

Fig. 1. Study area in northern Minnesota and northwestern Ontario showing the six Landsat footprints and the Border Lakes focus area in red. The region includes land ownerships with varying forest management strategies: Chippewa National Forest (1), Superior NF (2), BWCA Wilderness (3), Voyageurs National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

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

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