ELSEVIER

Available online at www.sciencedirect.com
SCIENCE DIRECT®

Remote Sensing of Environment 84 (2003) 561–571

Remote Sensing of Environment

www.elsevier.com/locate/rse

An improved strategy for regression of biophysical variables and Landsat ETM+ data

Warren B. Cohen[a,*], Thomas K. Maiersperger[b], Stith T. Gower[c], David P. Turner[b]

[a] *Forestry Sciences Laboratory, Pacific Northwest Research Station, USDA Forest Service, 3200 SW Jefferson Way, Corvallis, OR 97331, USA*
[b] *Department of Forest Science, Forestry Sciences Laboratory 020, Oregon State University, Corvallis, OR 97331, USA*
[c] *Department of Forest Ecology and Management, 1630 Linden Drive, University of Wisconsin, Madison, WI 53706, USA*

Received 5 December 2001; accepted 15 July 2002

Abstract

Empirical models are important tools for relating field-measured biophysical variables to remote sensing data. Regression analysis has been a popular empirical method of linking these two types of data to provide continuous estimates for variables such as biomass, percent woody canopy cover, and leaf area index (LAI). Traditional methods of regression are not sufficient when resulting biophysical surfaces derived from remote sensing are subsequently used to drive ecosystem process models. Most regression analyses in remote sensing rely on a single spectral vegetation index (SVI) based on red and near-infrared reflectance from a single date of imagery. There are compelling reasons for utilizing greater spectral dimensionality, and for including SVIs from multiple dates in a regression analysis. Moreover, when including multiple SVIs and/or dates, it is useful to integrate these into a single index for regression modeling. Selection of an appropriate regression model, use of multiple SVIs from multiple dates of imagery as predictor variables, and employment of canonical correlation analysis (CCA) to integrate these multiple indices into a single index represent a significant strategic improvement over existing uses of regression analysis in remote sensing.

To demonstrate this improved strategy, we compared three different types of regression models to predict LAI for an agro-ecosystem and live tree canopy cover for a needleleaf evergreen boreal forest: traditional (Y on X) ordinary least squares (OLS) regression, inverse (X on Y) OLS regression, and an orthogonal regression method called reduced major axis (RMA). Each model incorporated multiple SVIs from multiple dates and CCA was used to integrate these. For a given dataset, the three regression-modeling approaches produced identical coefficients of determination and intercepts, but different slopes, giving rise to divergent predictive characteristics. The traditional approach yielded the lowest root mean square error (RMSE), but the variance in the predictions was lower than the variance in the observed dataset. The inverse method had the highest RMSE and the variance was inflated relative to the variance of the observed dataset. RMA provided an intermediate set of predictions in terms of the RMSE, and the variance in the observations was preserved in the predictions. These results are predictable from regression theory, but that theory has been essentially ignored within the discipline of remote sensing.

Keywords: Regression analysis; Biophysical variables; Landsat ETM+

1. Introduction

Biogeochemical cycling models are increasingly run in a spatially explicit mode, requiring as model drivers moderate to high spatial resolution surfaces of land cover and leaf area index (LAI) derived from satellite imagery (Bonan, 1993; Reich, Turner, & Bolstad, 1999: Running, Baldocchi, Turner, Gower, Bakwin, & Hibbard, 1999). Mapping of continuous variables like LAI from high-resolution imagery such as Landsat TM or ETM+ has largely depended on modeling empirical relationships derived from single-date spectral vegetation indices (SVIs). The most important of these are the normalized difference vegetation index (NDVI) and its counterpart, the simple ratio (SR) (Chen & Cihlar, 1996; Fassnacht, Gower, MacKenzie, Nordheim, & Lillesand, 1997; White, Running, Nemani, Keane, & Ryan, 1997). These and other ratio-based indices, although important, utilize only a fraction of the spectral information available in many image datasets (Cohen, Spies, & Fiorella, 1995). Moreover, with the cost of ETM+ data substantially reduced from that of its predecessor, TM, there are increasing opportunities to utilize multiple dates of imagery in these analyses.

* Corresponding author. Tel.: +1-541-750-7322; fax: +1-541-758-7760.
E-mail address: warren.cohen@orst.edu (W.B. Cohen).

0034-4257/02/$ - see front matter © 2002 Elsevier Science Inc. All rights reserved.
PII: S0034-4257(02)00173-6

Traditional methods for empirical modeling of continuous variables, such as LAI, from SVIs rely on ordinary least squares (OLS) regression (Steel & Torrie, 1980), a technique that has important limitations for such applications (Curran & Hay, 1986). In particular, a violation of assumptions about measurement error can have undesirable effects on OLS estimates of the biophysical variable. In spite of the cogent arguments against the use of OLS regression in remote sensing offered by Curran and Hay (1986), we could find only one subsequent remote sensing paper that heeded their advice (Larsson, 1993). Alternative regression models may provide improved estimates of biophysical variables in remote sensing. Several such models are discussed in the literature, but almost all of that literature is outside of our discipline.

When conducting regression analyses that utilize multiple SVIs and multi-date data, it would be useful to construct a single, integrated index to represent the multiple predictor variables. This would facilitate visual assessment of model strength and whether the integrated relationship is linear. An integrated index could also help in subsequent analyses, or for screen viewing and interpretation (similar to the NDVI or SR). Additionally, an integrated index would be useful for comparisons among possible model formulations. Most important, however, is that certain regression procedures are best conducted in a simple linear context, and thus rely on a single predictor variable. These needs can be met using a statistical tool known as canonical correlation analysis (CCA).

1.1. Objective

The goal of this paper is to demonstrate an improved strategy for regression modeling of biophysical variables in remote sensing. That strategy includes use of multiple SVIs from multiple dates of ETM+ imagery, development of a CCA-based index that integrates these, and choice of an appropriate type of regression model. We test three regression-modeling approaches: traditional OLS (Reg_T), inverse OLS (Reg_I), and reduced major axis (RMA). The test is done for two biophysical variables, one in each of two different biomes, to highlight the general applicability of the analyses and results. At an agricultural site, we model LAI for two separate dates; at a boreal forest site, live tree cover is modeled. The objective is to compare and contrast the three regression approaches in terms of basic statistical characteristics of the predicted variables relative to the statistical characteristics of the observed variables. Numerous examples of such comparisons exist in the general literature, and the lessons learned are applied in various disciplines. However, the two papers in remote sensing literature that address this issue (Curran & Hay, 1986; Larsson, 1993) have been essentially ignored. With continued use of regression in remote sensing, and an increased reliance on Landsat imagery to drive ecosystem process models with regression-derived surfaces, it is imperative that we consider the weaknesses of our common methods and the potential strengths of alternative methods.

1.2. Background

1.2.1. Spectral vegetation indices (SVIs) and related linear combinations

The value of SVIs for modeling the relationship between vegetation variables and reflectance data is well established. In particular, since their inception, the SR (Birth & McVey, 1968) and the NDVI (Rouse, Haas, Schell, & Deering, 1974) have dominated the remote sensing and related literature (e.g., Chen & Cihlar, 1996; Huete, Jackson, & Post, 1985; Sellers, 1987; Tucker, 1979; Turner, Cohen, Kennedy, Fassnacht, Briggs, 1999). Modifications to these indices have been proposed to account for background effects associated with incomplete canopy cover (Huete, 1988), some of which take advantage of shortwave-infrared reflectance (Brown, Chen, Leblanc, & Cihlar, 2000; Nemani, Pierce, Running, & Band, 1993).

Although these "ratio-based" indices have the advantage of being simple to understand and apply, an alternative set of indices, called "*n*-space indices" (Jackson, 1983), is designed to more fully exploit the spectral domain of reflectance data. Numerous such indices exist, including the Perpendicular Vegetation Index (Richardson & Wiegand, 1977) and the widely used Tasseled Cap, which consists of the brightness, greenness (Kauth & Thomas, 1976), and wetness (Crist & Cicone, 1984) indices. The Tasseled Cap indices, in particular, provide standardized coefficients for all spectral bands of Landsat MSS and TM data. One study, across a forested scene containing bare ground, brush, and broadleaf and needleleaf forests of varying ages, demonstrated that brightness, greenness, and wetness accounted for 85% of the total spectral variability contained in a single date of TM reflectance data (Cohen et al., 1995); this, in comparison to only 52% in the red and near-infrared bands.

The temporal domain of spectral data can greatly enhance our ability to map vegetation (Helmer, Brown, & Cohen, 2000; Lefsky, Cohen, & Spies, 2001; Loveland et al., 2000; Oetter, Cohen, Berterretche, Maiersperger, & Kennedy, 2001). Incorporating a temporal series of data into SVIs directly, however, has been given minimal attention. Malila (1980) described the change magnitude and angle calculations, which are indices of a sort, required for change vector analysis (CVA). Whereas CVA was designed for two spectral dimensions and two dates of imagery, it has been crudely extended to three spectral dimensions (Virag & Colwell, 1987), and the magnitude calculation has been generalized to *n*-dimensions by Lambin and Strahler (1994). The concept for generalizing CVA angles and magnitudes for *n* spectral and/or temporal dimensions was described by Cohen and Fiorella (1998), but they stopped short of implementing the procedure. Collins and Woodcock (1996) developed a two-date Tasseled Cap transformation

for use in change detection, but an *n*-date transformation was not attempted.

Principal components analysis (PCA) is an attractive means of incorporating spectral data from numerous dates into a small set of axes that contain most of the spectral information contained in the full multispectral, multitemporal dataset (Eastman & Fulk, 1993; Richards, 1984). Although not vegetation indices per se, the value of PCA for reducing the size of the spectral–temporal dataset is great. However, there are two important problems using PCA for spectral–temporal analyses. First, the resulting axes are dataset-dependent. Although this means that it is difficult to generalize the interpretation of PCA axes to other datasets, this is not unique to PCA, as the same problem is common to all correlation-based, empirical analyses. The second and more meaningful problem is that the coefficients for PCA axes are normally obtained without regard for the axes' relationships with the variable we are interested in predicting (e.g., LAI).

1.2.2. Regression and related analysis

In the simple linear case, OLS regression analysis is an empirical approach for modeling the relationship between two observed variables, X and Y. The form of the OLS regression model is

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad (1)$$

where Y is the variable to be predicted, X is the variable Y is predicted from, β_0 is the intercept, β_1 is the slope of the relationship between X and Y, and ε is error. Data for the analysis are supplied by paired observations of the two variables. Commonly, one variable is difficult or costly to measure (e.g., vegetation attributes from field sampling), and the other is relatively easy or inexpensive to observe (e.g., SVIs from remote sensing). Although often the intent of OLS regression is to determine the feasibility of estimating or predicting the expensive variable from observations of the inexpensive one, sometimes the analysis is used simply to determine the form and strength of the relationship between the two variables. If the objective is the latter, it is not particularly important which variable is X and which is Y, and common in the remote sensing literature are both X and Y representing the vegetation variable and the SVI (Butera, 1986; Chen & Cihlar, 1996; Cohen, 1991; White et al., 1997). If we are interested in actually using the regression model to predict one variable from the other, however, the distinction between X and Y becomes very important. This is because of specifications and assumptions associated with OLS regression.

One specification is that Y is the dependent variable and X is the independent variable (Steel & Torrie, 1980). Although it can be argued that spectral response is dependent on vegetation state and not the other way around, much of the remote sensing literature reports the vegetation attribute being modeled as the dependent variable. Curran and Hay (1986) discuss this as the "specification problem". Specification in this manner is important because of an assumption associated with OLS regression: that the independent variable, X (e.g., an SVI), is measured without error (Steel & Torrie, 1980). As the coefficients for the regression equation are calculated by minimizing the sums of squares of error in Y (e.g., LAI), illustrated graphically by Curran and Hay (1986), the result is an attenuation (or compression) of the variance of LAI predictions. In other words, values above the mean of Y tend to be underpredicted and values below the mean tend to be overpredicted (e.g., Cohen, Maiersperger, Spies, & Oetter, 2001; Hudak, Lefsky, Cohen, & Berterretche, 2002).

An alternative form of OLS regression is inverse estimation (Brown, 1979), also known as inverse regression (Cohen, 1991) and calibration (Scheffé, 1973). Curran and Hay (1986) refer to this as X on Y regression and illustrate the concept graphically. With inverse estimation, the specification problem is addressed in that the dependent and independent variables are properly assigned, or "specified" (e.g., X is the vegetation variable and Y is the SVI). In practice then, to predict the vegetation variable, the coefficients for the OLS regression model are derived using Eq. (1) and then the equation must be inverted to solve for X, such that $X=(Y-\beta_0)/\beta_1$. The error term in Eq. (1), ε, is expressed as prediction residuals for each observation.

Although for remote sensing, inverse estimation eliminates the specification problem, it does not address the more important problem that X is assumed to be measured without error. Curran and Hay (1986) provide an in-depth discussion of sources of error for both X and Y in remote sensing. The impact of measurement error in X when using inverse estimation is known to be the opposite to that of its effect using the Y on X form of OLS regression, i.e., amplification of the variance of predicted biophysical values such that values above the mean of X are overestimated and those below the mean are underestimated.

Recognizing that there are errors in both X and Y, Curran and Hay (1986) tested three alternative methods to predict grassland LAI from the SR: Wald's grouping method, RMA, and an alternative least squares procedure that incorporates a priori knowledge of relative errors in X and Y. They recommended using Wald's method or RMA if no estimates of measurement error are available, and the alternative least squares procedure if such estimates are available. From a practical perspective, it will be rare for analysts to have precise estimates of error from all the various sources associated with measurements of vegetation and spectral variables. As such, it is perhaps more prudent to make no assumptions regarding the relative amounts of measurement error and use RMA or Wald's method. In spite of the convincing arguments made by Curran and Hay (1986), we could find only one subsequent remote sensing article that used one of these methods (Larsson, 1993), where RMA was used to predict woodland canopy cover from single-date NDVI measurements.

RMA is one of a class of similar models known as orthogonal regression, total least squares regression, or errors-in-variables modeling, depending on the discipline in which the specific technique was developed (Van Huffel, 1997). Orthogonal regression minimizes the sum of squared orthogonal distances from measurement points to the model function. The RMA version of orthogonal regression is graphically depicted in Curran and Hay (1986). Van Huffel (1997) contains examples of orthogonal regression's usage in astronomy, meteorology, 3-D motion estimation, biomedical signal processing, and multivariate calibration. RMA, specifically, is quite commonly applied in allometry (Conrad & Gutmann, 1996; Gower, Kucharik, & Norman, 1999; Nicol & Mackauer, 1999; Niklas & Buckman, 1994). Besides making no assumptions about errors in X and Y, RMA likewise makes no assumptions about dependency. Conrad and Gutmann (1996) refer to RMA as geometric mean regression, in that the slope (β_1) is defined as the ratio of sample standard deviation for Y over the sample standard deviation for X, thus preserving in the model the relative variance structure of the sample dataset. The effect of this is to minimize or eliminate any attenuation or amplification of predictions. For RMA, β_0 is defined as the sample mean of Y minus the quantity β_1 times the sample mean of X. One important component of the slope term (β_1) is that it must be given the sign (+/−) of the correlation between X and Y (Conrad & Gutmann, 1996), which is not given by Curran and Hay (1986) or Larsson (1993). The form of the regression model is identical to Eq. (1), but the calculations of β_0 and β_1 are different. Mathematical similarities in the formulations of the two OLS and the RMA regression models mean that the model intercepts are all equivalent, as are the coefficients of determination. What differ among these models are the root mean square errors (RMSEs) and the slopes of the relationships.

1.2.3. Canonical correlation analysis (CCA)

OLS regression has both simple (single X) and multiple (several X) forms (Steel & Torrie, 1980). The use of OLS regression in its multiple form, Y on multiple X, is familiar to most remote sensing analysts conducting regression modeling. Although much less familiar, there is also a formulation for multiple X inverse calibration (Brown, 1979). A simple application of RMA requires one X and one Y. Thus, to incorporate n-space indices and/or temporal datasets into an RMA, the multiple X dataset must be linearly combined into a single X variable. In essence, we must develop a new, integrated index that is a linear combination of the multiple X indices (or bands) from a single date or multiple dates. As discussed earlier, this need is directly facilitated by CCA.

CCA is a generalized form of multiple regression that permits the examination of interrelationships between two sets of variables (multiple X's and multiple Y's) (Tabachnick & Fidell, 1989). CCA maximizes the correlation between a composite of variables from one set with a composite of variables from another set. When there is only one X (i.e., vegetation variable, such as LAI), CCA provides a set of coefficients for the Y's that aligns them with the variation in the X variable. When those coefficients are applied to the Y variables, the result is a CCA score for each observation. CCA scores are indexed values in the same way that brightness, greenness, or wetness (or NDVI) values are indexed values. However, with CCA, the alignment is dataset-specific, whereas with the Tasseled Cap or NDVI, the formulations are generalized and fixed.

2. Methods

This work was conducted in a temperate broadleaf agro-ecosystem, consisting of corn and soybeans, and a boreal needleleaf evergreen forest. The biophysical variable of interest within the agro-ecosystem was LAI, which was modeled for two separate measurement dates. The dominant tree species in the boreal forest is black spruce, and the variable we modeled was percent tree cover. This work was done in the context of the BigFoot project, which was designed to provide local validation of global estimates of biophysical variables and processes using MODIS data (Cohen & Justice, 1999).

2.1. Study sites, sampling design, and field measurements

The study sites and sampling design were described in Campbell, Burrows, Gower, and Cohen (1999). The agricultural site (AGRO) was a 5 × 5 km area located just south of Champaign, IL. The boreal forest site was a similarly sized area surrounding the northern old black spruce (NOBS) site of the Boreal Ecosystem Atmosphere Study (Sellers et al., 1997), approximately 40 km west of Thompson, Manitoba, Canada. The sample design was a nested spatial series (Burrows et al., in press) that permits explicit examination of spatial covariation among field-measured ecosystem properties using variograms and cross-variograms (Cressie, 1991). At each site, there were approximately 100 25 × 25 m plots where land cover, LAI, absorbed radiation, and net primary production were measured/observed at five to nine subplots per plot. Subplot measurements were averaged to provide a single value for each measured variable at each plot. Plot locations were determined using a real-time differential GPS. The accuracy of the system was <0.5 m in both the x and y dimensions.

At the AGRO site, LAI was measured at five subplots per plot using standard, direct harvest methods described by Gower et al. (1999). Measurements were made at several time periods during the growing season in 2000. We used data from July and August. At NOBS, percent tree cover was measured at nine systematically spaced subplots using an upward-looking digital camera. The imaged canopy projection area was dependent on tree height and the field of view of the camera, which was 30°. At approximately 10-m height, this means that among the nine subplots, nearly

100% of the canopy area in each plot was imaged. In the lab, each of the nine photos per plot was sampled using a grid of 99 points to derive the percent live tree canopy cover at each plot (Berterretche, 2002).

2.2. Image data and processing

For AGRO, ETM+ data from four dates were used to capture the growing season from April through September (Table 1). At NOBS, two images were used, one from March and one from June. The images were georeferenced, radiometrically calibrated, and translated into Tasseled Cap brightness, greenness, and wetness. All images were acquired at level 1G processing, with a cell size of 30 m, and UTM (WGS84) projection. At AGRO, positional accuracy of the native map projection of the June image was judged by direct comparison with USGS digital orthophoto quadrangles (DOQs) at a 9×9 km area centered on the study site. A systematic local shift of -37.5 m in the x-direction and -127.5 m in the y-direction was applied to the ETM+ image to register it to the DOQs. Subsequently, all other image dates were positionally shifted to match the June date. At NOBS, a panchromatic IKONOS image was registered to the earth's surface with the same projection parameters as at AGRO using several GPS points collected in the field. The June image was then positionally shifted to match the IKONOS image, and the March image was shifted to match the June image.

The COST absolute radiometric correction model of Chavez (1996) was applied to each image to convert digital counts to reflectance. Radiometrically "dark" objects were assumed to have 2% reflectance across all bands. For AGRO, the June image was selected as a reference image and all other dates of imagery were relatively normalized to it, as a fine-tuning for multidate, inter-image calibration. The method used was similar to that of Oetter et al. (2001) and of the Ridge Method of Kennedy and Cohen referred to by Song, Woodcock, Seto, Pax Lenney, and Macomber (2001), which are an adaptation of standard band-by-band relative normalization procedures based on co-located bright and dark targets. As the COST model is not appropriate for low sun angle situations, the March image from NOBS was converted to reflectance using a more basic dark-object-subtraction model. Further, no relative normalization was performed for the NOBS dataset due to major spectral property differences between the two dates, given the backdrop of ice and snow for the March image and of vegetation and water for the June image.

Table 1
ETM+ images used in this study

Site	Path/row	Date
AGRO	22/32	April 26, 2000
	22/32	June 29, 2000
	22/32	July 15, 2000
	22/32	September 1, 2000
NOBS	34/21	March 13, 2000
	33/21	June 6, 2000

No published transformation exists to convert atmospherically-corrected ETM+ spectral data to Tasseled Cap indices. However, Crist (1985) derived coefficients for brightness, greenness, and wetness from ground-based spectral data that can be applied to atmospherically corrected Landsat data. Slight differences in spectral band width and position, as well as calibration, exist between Landsat TM and ETM+ (Teillet et al., 2001; Vogelmann et al., 2001), but they are similar enough to assume that the differences in Tasseled Cap indices derived for data from the two different sensors are small. We tested this assumption using TM and ETM+ images acquired within a few days of each other (Path 46/Row 29) over western Oregon in 1999. First, we converted atmospherically corrected TM DN data to the Tasseled Cap indices using the coefficients in Crist and Cicone (1984). We then converted the atmospherically corrected TM DN data to reflectance using published coefficients and formulae, before using the Crist (1985) coefficients to convert the reflectance data to Tasseled Cap indices. Finally, we atmospherically corrected the ETM+ data and then converted the reflectance data to the Tasseled Cap indices using the Crist (1985) coefficients. A comparison of the brightness, greenness, and wetness images from the three methods showed that they were highly intercorrelated at a level of roughly 95%.

2.3. Variable selection and model development, execution, and comparison

With both OLS and RMA regression done in a multiple-Y (i.e., multiple SVIs) context, there is the issue of variable selection. Not all Y variables are needed or are significant in the presence of other Y variables, and some may need to be culled from the dataset. For this we used forward stepwise regression. In each case, brightness, greenness, and wetness from all dates of available imagery were used as potential variables for a model. To avoid overfitting a given model, we imposed the rule that the number of variables to enter the model be less than one-third the number of observations. Prior to conducting stepwise regression, bivariate plots of all potential Y variables against LAI or canopy cover were evaluated to determine if transformations were required to linearize relationships. Where necessary, standard log and square root transformations were used.

Once the variables for a given dataset were selected, Reg_T, Reg_I, and RMA regression models were developed. For all three modeling approaches, the CCA axis derived from the same Y-variable set was used. To compare the three modeling approaches, predicted versus observed plots were developed and overall bias and variance ratios were calculated. Bias was calculated as the mean of the predicted values minus the mean of the observed values, such that a positive bias equated to a mean overprediction and vice versa. Variance ratio was calculated as the standard devia-

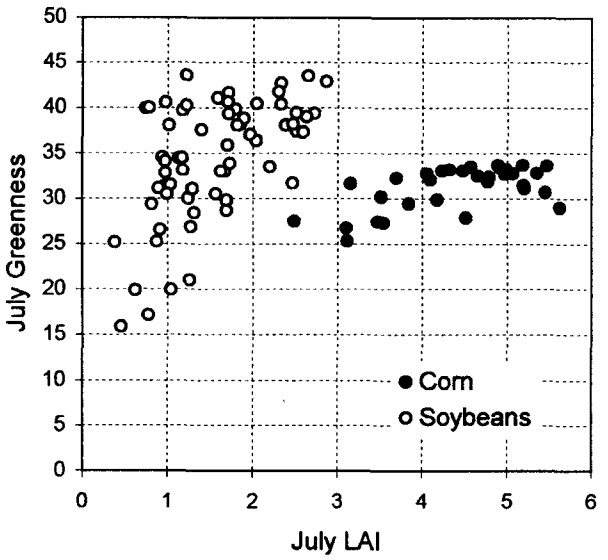

Fig. 1. Tasseled Cap greenness as a function of July LAI at AGRO.

tion of the predicted values divided by the standard deviation of the observed values. As such, a ratio of greater than one meant that the prediction variance was greater than the observed variance.

Field data are expensive to collect and process, so using them prudently is essential. There is a trade-off between using all available observations to develop a regression model and having no independent observations to test the model, versus excluding a predetermined number of observations to test the model, but having a less robust model because it was developed on fewer points. The statistical literature provides several alternative, but related ways to address this problem: cross-validation, bootstrapping, and jackknifing (Efron & Gong, 1983). We used the cross-validation procedure, which provides a nearly unbiased estimator of prediction error (Efron & Gong, 1983). This required, for each dataset and regression variant, that (where $n = 100$) 100 separate models be developed, each time with data from 99 observations. Then, each model was used to predict the observation that was left out, thus providing the predictions for all 100 plot observations that were needed to compare against the observed values. This provided an error characterization equivalent to the PRESS statistic (SAS, 1990).

Table 2
Regression model statistics for each model type and dataset

Model	Type	Slope	Intercept	R^2
Soy, July	Reg_T	0.49	1.54	0.58
	Reg_I	1.19/0.84	−1.83/1.54	0.58
	RMA	0.64	1.54	0.58
Corn, July	Reg_T	0.63	4.41	0.61
	Reg_I	0.96/1.04	−4.24/4.41	0.61
	RMA	0.81	4.41	0.61
Soy, August	Reg_T	0.49	3.44	0.27
	Reg_I	0.56/1.79	−1.91/3.44	0.27
	RMA	0.93	3.44	0.27
Corn, August	Reg_T	0.45	4.00	0.64
	Reg_I	1.44/0.70	−5.76/4.00	0.64
	RMA	0.56	4.00	0.64
Canopy cover	Reg_T	15.08	38.89	0.68
	Reg_I	0.045/22.1	−1.76/38.89	0.68
	RMA	18.26	38.89	0.68

For the Reg_I models, the original slope and intercept are given along with the back-inverted slope and intercept for comparison with other model types.

3. Agricultural example—LAI at AGRO

A scatterplot of corn and soybean greenness from the July measurement date (Fig. 1) revealed that these two crops represented different populations and were best modeled separately. This was done for both July and August dates, yielding four separate modeling sets (Table 2). For all four model sets, the three regression approaches had equivalent

Table 3
Cross-validation results for each model type and dataset

	Model type			n
	Reg_T	Reg_I	RMA	
Soy, July				64
R	0.74	0.74	0.74	
RMSE	0.42	0.59	0.47	
Bias	−0.01	−0.03	−0.01	
Variance ratio	0.80	1.40	1.06	
Corn, July				31
R	0.76	0.76	0.76	
RMSE	0.53	0.68	0.56	
Bias	0.00	0.03	0.01	
Variance ratio	0.76	1.31	1.01	
Combined, July				95
R	0.95	0.92	0.94	
RMSE	0.46	0.62	0.50	
Bias	−0.01	−0.01	0.00	
Variance ratio	0.96	1.07	1.00	
Soy, August				64
R	0.47	0.49	0.49	
RMSE	0.82	1.57	0.95	
Bias	−0.01	−0.02	−0.02	
Variance ratio	0.53	1.95	1.02	
Corn, August				31
R	0.79	0.79	0.79	
RMSE	0.35	0.46	0.37	
Bias	0.00	0.03	0.01	
Variance ratio	0.81	1.33	1.02	
Combined, August				95
R	0.59	0.54	0.57	
RMSE	0.70	1.32	0.81	
Bias	−0.01	−0.01	−0.01	
Variance ratio	0.62	1.81	1.01	
Canopy cover				103
R	0.82	0.82	0.82	
RMSE	10.41	12.68	10.93	
Bias	0.02	0.03	0.03	
Variance ratio	0.83	1.22	1.01	

For the agricultural site, five plots were not corn or soybeans. For the forest site, there were three extra plots.

coefficients of determination and model intercepts. The differences among approaches were expressed in the slope term, with Reg_T, having the least, Reg_I having the greatest, and RMA being intermediate. As mentioned earlier, these are anticipated results that are provided here for demonstration purposes.

A summary of cross-validation predictions revealed the effect of the different modeling approaches (Table 3), again, for demonstration purposes. Presented are results from the crop-specific models and from the combined set of predictions across the two crop-specific models. For the July date, corn, soybeans, and combined, the correlation coefficients (R) between predicted LAI and observed LAI were essentially the same for all three approaches. The only difference, which was minimal, was for the combined model. This difference is attributable to the cross-validation procedure. Bias was near zero in all cases, indicating that the observed mean of the samples was preserved in the predictions. The differences among the modeling approaches were related to the different slope terms (from Table 2), and were expressed in both the RMSE and the variance ratio. Reg_T, by design, had the lowest RMSE in predictions of Y (LAI). Similarly, because Reg_I also minimized the sums of squares of error in Y (this time, SVI), it yielded the greatest RMSE in LAI. As

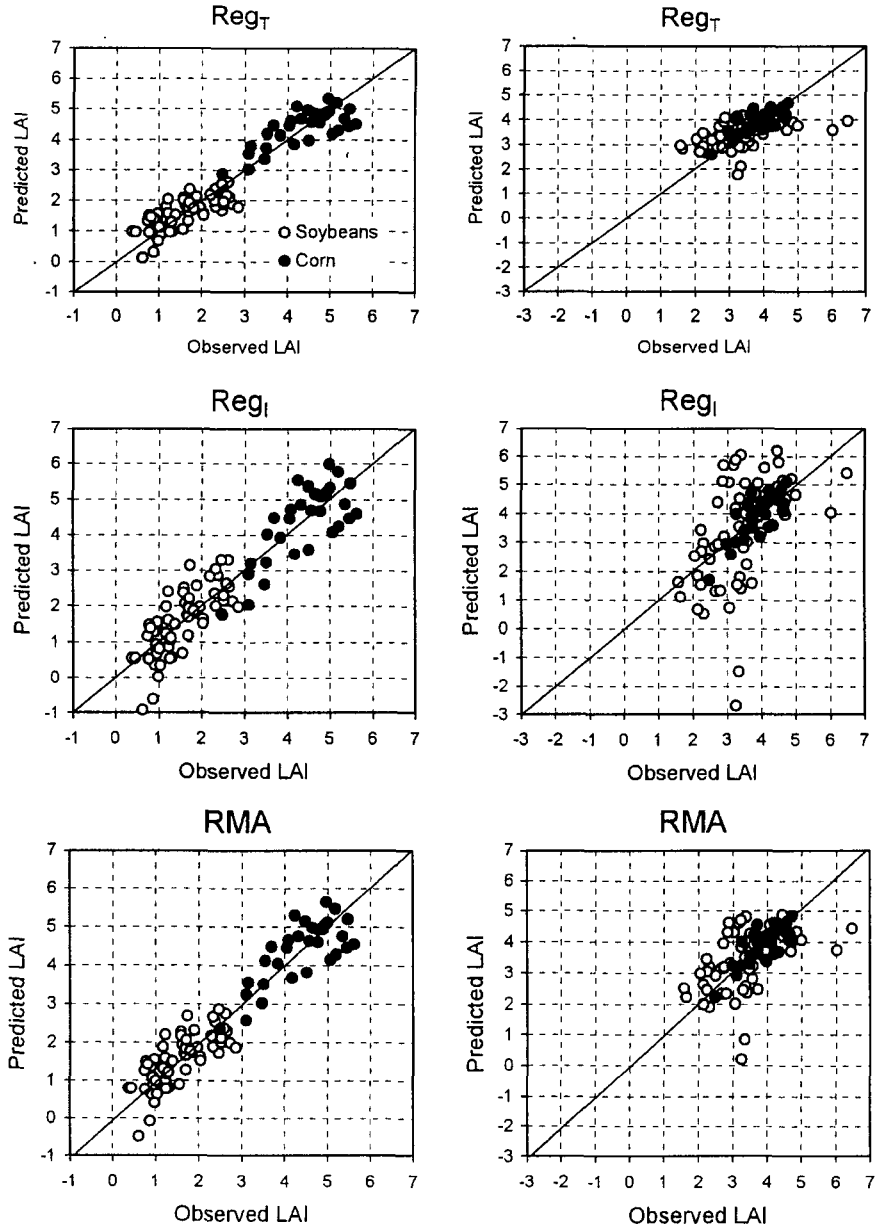

Fig. 2. Predicted (from cross-validation) versus observed July and August LAI at AGRO. Reg_T is traditional OLS regression, Reg_I is inverse OLS regression, and RMA is reduced major axis regression. Left is July; right is August.

expected, RMA was a compromise solution, having intermediate values of RMSE. With respect to the variance ratio, RMA always exhibited a value close to 1.0, indicating that the variance structure of the observed values was preserved in the predicted values. Deviations from unity for the OLS methods were greatest when the correlation coefficients were lowest. This latter point was particularly evident in the results from August, where correlations were lower than in July. For all cases, Reg_I had variance ratios greater than 1.0 and Reg_T had values less than 1.0, indicating greater and lesser variance, respectively, relative to observed values.

With regression models, it is always possible to have individual predictions outside the range of observed values. This mostly occurs when observed "independent" variables have values outside the range on which models were constructed (an increased possibility with cross-validation), or when significant outliers exist in the model dataset. Here, this is evident from the predictions of negative LAI (Fig. 2). Of course, this "problem" is amplified with Reg_I and is suppressed with Reg_T.

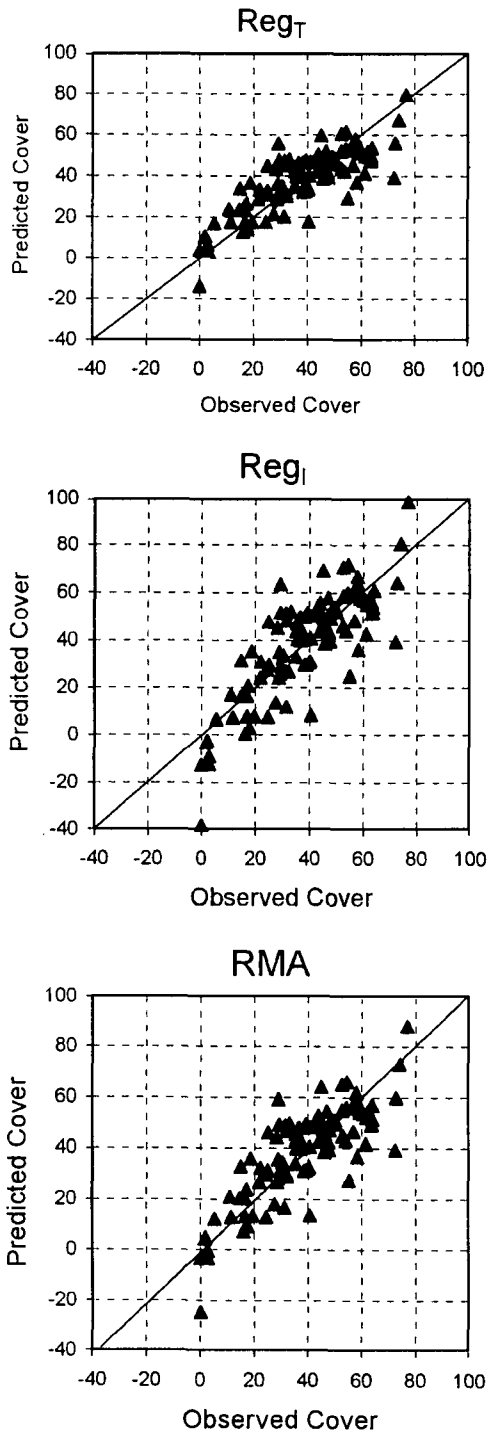

Fig. 3. Predicted (from cross-validation) versus observed percent live tree canopy cover at NOBS. Reg_T is traditional OLS regression, Reg_I is inverse OLS regression, and RMA is reduced major axis regression.

4. Boreal forest example—tree cover at NOBS

The models for tree cover at NOBS exhibited similar relative characteristics as those at AGRO (Table 2). The coefficients of determination and the intercepts were all identical. The only difference among modeling approaches was the slope of the relationships, with Reg_T having the lesser value, Reg_I having the greater value, and RMA having an intermediate value. Likewise, the cross-validation results indicated identical correlations between tree cover and the CCA axis and essentially no overall bias for any of the models. Again, RMSE was lowest for Reg_T, highest for Reg_I, and intermediate for RMA.

Some predictions outside of the observed range occurred for tree cover, as it had for LAI at AGRO (Fig. 3). Again, this was most evident using the Reg_I approach, which amplified the variance of the predictions relative to the observed values. For tree cover, there was a slight tendency toward an asymptote in the predictions, especially for the Reg_T model.

5. Discussion and summary

This paper presents to a remote sensing audience a verification of existing regression theory. The remote sensing literature contains little of regression theory, and even less of the numerous options for its application. With rare exception, the remote sensing literature contains rote application of OLS (ordinary least squares) regression, without ever questioning its disadvantages relative to other forms of regression. Questioning these disadvantages and demonstrating alternative regression approaches were the main purposes of this paper. Numerous examples of alternative regression models exist in many other scientific and technical fields, but remote sensing community has largely ignored the important work of Curran and Hay (1986).

Most of the regression-based remote sensing literature is focused on minimizing error (e.g., RMSE) in the predictions

of biophysical variables. Traditional OLS regression is an excellent means for accomplishing that, but there are two big problems with OLS. First, it assumes no error in measurements of vegetation reflectance and/or the biophysical variable of interest. Second, traditional OLS provides attenuated variance in predictions of that variable. The statistical literature strongly suggests that if there are errors in the measurements of both variables (e.g., reflectance and biophysical), then OLS regression is the wrong model to use. Because it is nearly impossible to defend any claim that either reflectance or biophysical variables are measured without error, application of OLS regression is inappropriate in remote sensing.

Compression of variance by OLS becomes critical if the regression model is used to build a map of a biophysical variable, that in turn drives a functional/mechanistic model. If the mechanistic model involves nonlinear functions of the biophysical variable, attenuation of variance in the biophysical variable introduces error in the mechanistically modeled outputs. Because OLS attenuates the variance, it will introduce such error. The degree of attenuation is essentially a linear function of the correlation between the spectral data and the biophysical variable, low correlation, much attenuation, and vice versa. Many of the relationships between reflectance and biophysical variable are poorly correlated, so this is not a non-issue.

The remote sensing literature contains a great number of examples of empirical, regression modeling relating SVIs to measures of a myriad of vegetation variables modeled across an assortment of sites and biomes. Most studies used SVIs based on red and near-infrared reflectance. More often than not, a single SVI from a single date was used. The spectral depth of ETM+ is essentially three-dimensional (Cohen et al., 1995; Crist & Cicone, 1984) and we have increasing numbers of temporal image series available for greater predictive power (Lefsky et al., 2001). As such, we should be expanding our use of multiple regression over simple regression techniques. Multiple regression in an RMA context requires a single-integrated index of multiple bands or indices. This need is directly facilitated by CCA. Canonical correlation analysis has rarely been used in remote sensing. However, for those contemplating the use of CCA for deriving a dataset-dependent index, there should be a clear understanding of what the procedure does to the dataset.

A simple test on any appropriate single-Y (in this case, e.g., LAI), multiple-X (in this case, SVIs) dataset illustrates that CCA scores are perfectly correlated to predicted Y values from traditional OLS multiple regression on that dataset. The difference is that one provides predictions of the Y variable, whereas the other is simply a set of index scores that are maximally correlated with the observed Y variable. If one then conducts traditional simple OLS regression with the CCA scores as X and the LAIs as Y, they will derive exactly the same predicted values for LAI as those predicted from the original multiple OLS regression. In both cases, RMA would be required to balance the variance ratio at a value of 1.0. Traditional OLS regression provides biophysical predictions, but the variance of those predictions is unbalanced vis-à-vis the observed variance in the biophysical variable. CCA provides an index that is maximally correlated with the biophysical variable of interest, but it does not provide predictions. RMA can either provide predictions from a CCA index that have a balanced variance, or it can balance (or calibrate) a set of unbalanced predictions derived previously from traditional OLS regression conducted on a CCA index or from multiple OLS regression. For the latter case, the CCA index would be superfluous. Thus, the only important reason for conducting the CCA is if the index itself is desired, for which there may be numerous reasons. For this study, it was desirable to have a single index for the convenience of comparative analysis among methods to derive a single regression slope term for each method. It is important to keep in mind that whereas CCA is dataset-specific, NDVI or SR, or other SVIs such as brightness, greenness, and wetness, are more generalizable in terms of their biophysical meaning.

In this study, we illustrated an improved regression modeling strategy that incorporates all the recommended steps for deriving mapped estimates that have their errors characterized. This strategy includes the collection of georeferenced field data, image georeferencing, image radiometric calibration, translation of reflectance into SVIs, testing for significance of each SVI in regression models, and using cross-validation to provide nearly unbiased testing of robust models. Applying the RMA models to the CCA indices provided high-quality maps of LAI for the agricultural site, with means and variances well preserved in each important land cover class (corn and soybeans). Additionally, an important forest variable was mapped, tree cover, which will subsequently be used at the forest site to help derive a land cover map using classes that are largely based on percent tree cover. Although any given study may weigh these various processing components differently, two considerations are critical. (1) Is it acceptable for a predicted variable to have a different variance structure from that of empirical observations? (2) Is there a compelling reason to limit the analysis to a single SVI from a single date? If the answer to Question 2 is "no", then CCA may be an important aid in your analysis.

Acknowledgements

This research was funded by NASA's Terrestrial Ecology Program. We greatly thank Karin Fassnacht for thoughtful and lively discussion and for her review of early drafts, and Robert Kennedy for his contributions to the discussion.

References

Berterretche, M. (2002). Comparison of regression and geostatistical methods to develop LAI surfaces for NPP modeling. Master of Science thesis in Forest Science, Oregon State University, Corvallis, OR, 218 pp.

Birth, G. S., & McVey, G. R. (1968). Measuring the color of growing turf with a reflectance spectrophotometer. *Agronomy Journal, 60*, 640–643.

Bonan, G. (1993). Importance of leaf area index and forest type when estimating photosynthesis in boreal forests. *Remote Sensing of Environment, 43*, 303–314.

Brown, G. (1979). An optimization criterion for linear inverse estimation. *Technometrics, 2*, 575–579.

Brown, L., Chen, J., Leblanc, S., & 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.

Burrows, S., Gower, S., Clayton, M., Mackay, D., Ahl, D., Norman, J., Diak, G. Application of geostatistics to characterize LAI from flux tower to landscape scales using a cyclic sampling design. *Ecosystems* (in press).

Butera, M. K. (1986). A correlation and regression analysis of percent canopy closure versus TMS spectral response for selected forest site in the San Juan National Forest, Colorado. *IEEE Transactions on Geoscience and Remote Sensing, 24*, 122–129.

Campbell, J. L., Burrows, S., Gower, S. T., & Cohen, W. B. (1999). *BigFoot: characterizing land cover, LAI, and NPP at the landscape scale for EOS/MODIS validation. Field Manual Version 2.1.* Oak Ridge, TN: Environmental Sciences Division, Oak Ridge National Laboratory (104 pp.).

Chavez Jr., P. (1996). Image-based atmospheric corrections—revised and improved. *Photogrammetric Engineering and Remote Sensing, 62*, 1025–1036.

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.

Cohen, W. (1991). Response of vegetation indices to changes in three measures of leaf water stress. *Photogrammetric Engineering and Remote Sensing, 57*, 195–202.

Cohen, W., & Fiorella, M. (1998). Comparison of methods for detecting conifer forest change with Thematic Mapper imagery. In R. Lunetta, & C. Elvidge (Eds.), *Remote Sensing Change Detection, Environmental Monitoring Methods and Applications* (pp. 89–102). Chelsea, MI: Sleeping Bear Press.

Cohen, W., & Justice, C. (1999). Validating MODIS terrestrial ecology products: linking in situ and satellite measurements. *Remote Sensing of Environment, 70*, 1–3.

Cohen, W., Maiersperger, T., Spies, T., & Oetter, D. (2001). Modelling forest cover attributes as continuous variables in a regional context with Thematic Mapper data. *International Journal of Remote Sensing, 22*, 2279–2310.

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

Collins, J. B., & Woodcock, C. E. (1996). An assessment of several linear change detection techniques for mapping forest mortality using multitemporal Landsat TM data. *Remote Sensing of Environment, 56*, 66–77.

Conrad, R., & Gutmann, J. (1996). *Conversion Equations between Fork Length and Total Length for Chinook Salmon (Oncorhynchus tshawytscha).* Northwest Indian Fisheries Commission. Project Report Series No. 5. Olympia, WA. 32 pp.

Cressie, N. (1991). *Statistics for Spatial Data.* New York: Wiley.

Crist, E. P. (1985). A TM tasseled cap equivalent transformation for reflectance factor data. *Remote Sensing of Environment, 17*, 301–306.

Crist, E. P., & Cicone, R. C. (1984). A physically-based transformation of thematic mapper data—the TM tasseled cap. *IEEE Transactions on Geoscience and Remote Sensing, GE-22*, 256–263.

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

Eastman, J. R., & Fulk, M. (1993). Long sequence time series evaluation using standardized principal components. *Photogrammetric Engineering and Remote Sensing, 59*, 991–996.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician, 37*, 36–48.

Fassnacht, K., Gower, S., MacKenzie, M., Nordheim, E., & Lillesand, T. (1997). Estimating the leaf area index of north central Wisconsin forests using the Landsat Thematic Mapper. *Remote Sensing of Environment, 61*, 229–245.

Gower, S., Kucharik, C., & Norman, J. (1999). Direct and indirect estimation of leaf area index, fAPAR, and net primary production of terrestrial ecosystems. *Remote Sensing of Environment, 70*, 29–51.

Helmer, E., Brown, S., & Cohen, W. (2000). Mapping montane tropical forest successional stage and land use with multi-date Landsat imagery. *International Journal of Remote Sensing, 21*, 2163–2183.

Hudak, A., Lefsky, M., Cohen, W., Berterretche, M., 2002. Integration of lidar and Landsat ETM+ data for estimating and mapping forest canopy height. *Remote Sensing of Environment* 82, 397–416.

Huete, A. R. (1988). A soil-adjusted vegetation index (SAVI). *Remote Sensing of Environment, 25*, 295–309.

Huete, A. R., Jackson, R. D., & Post, D. F. (1985). Spectral response of a plant canopy with different soil backgrounds. *Remote Sensing of Environment, 17*, 37–53.

Jackson, R. D. (1983). Spectral indices in n-space. *Remote Sensing of Environment, 13*, 409–421.

Kauth, R.J., & Thomas, G.S. (1976, 6 June–2 July). The tasseled cap—a graphic description of the spectral–temporal development of agricultural crops as seen by Landsat. *Proc. Second Ann. Symp. Machine Processing of Remotely Sensed Data* (pp. 41–51). West Lafayette, IN: Purdue U. Lab. App. Remote Sens.

Lambin, E. F., & Strahler, A. H. (1994). Change-vector analysis in multitemporal space: a tool to detect and categorize land-cover change processes using high temporal-resolution satellite data. *Remote Sensing of Environment, 48*, 231–244.

Larsson, H. (1993). Linear regressions for canopy cover estimation in Acacia woodlands using Landsat-TM, -MSS, and SPOT HRV XS data. *International Journal of Remote Sensing, 14*, 2129–2136.

Lefsky, M., Cohen, W., & Spies, T. (2001). An evaluation of alternative remote sensing products for forest inventory, monitoring, and mapping of Douglas-fir forests in western Oregon. *Canadian Journal of Forest Research, 31*, 78–87.

Loveland, T., Reed, B., Brown, J., Ohlen, D., Zhu, Z., Yang, L., & Merchant, J. (2000). Development of a global land cover characteristics database and IGBP DISCover from AVHRR data. *International Journal of Remote Sensing, 21*, 1303–1330.

Malila, W. A. (1980, 3–6 June). Change vector analysis: an approach for detecting forest changes with Landsat. In P. G. Burroff, & D. B. Morrison (Eds.), *Proc. Sixth Ann. Symp. Machine Processing of Remotely Sensed Data. Soil Information Systems and Remote Sensing and Soil Survey* (pp. 326–335). West Lafayette, IN: Purdue U. Lab. App. Remote Sens.

Nemani, R. R., Pierce, L., Running, S., & Band, L. (1993). Forest ecosystem processes at the watershed scale: sensitivity to remotely-sensed leaf area index estimates. *International Journal of Remote Sensing, 14*, 2519–2534.

Nicol, C., & Mackauer, M. (1999). The scaling of body size and mass in a host-parasitoid association: influence of host species and stage. *Entomologia Experimentalis et Applicata, 90*, 83–92.

Niklas, K., & Buchman, S. (1994). The allometry of saguaro height. *American Journal of Botany, 81*, 1161–1168.

Oetter, D., Cohen, W., Berterretche, M., Maiersperger, T., & Kennedy, R. (2001). Land cover mapping in an agricultural setting using multiseasonal Thematic Mapper data. *Remote Sensing of Environment, 76*, 139–155.

Reich, P., Turner, D., & Bolstad, P. (1999). An approach to spatially distributed modeling of net primary production (NPP) at the landscape scale and its application in validation of EOS NPP products. *Remote Sensing of Environment, 70*, 69–81.

Richards, J. A. (1984). Thematic mapping from multitemporal image data

using principal components transformation. *Remote Sensing of Environment, 16*, 35–46.

Richardson, A. J., & Wiegand, C. L. (1977). Distinguishing vegetation from soil background information. *Photogrammetric Engineering and Remote Sensing, 43*, 1541–1552.

Rouse, J. W., Haas, R. H., Schell, J. A., & Deering, D. W. (1974, 10–14 Dec. 1973). Monitoring vegetation systems in the Great Plains with ERTS. In S. C. Freden, E. P. Mercanti, & M. A. Becker (Eds.), *NASA SP-351: Proc. Third Earth Resources Tech. Satellite-1 Symp. Vol. 1: Technical Presentations Sec. A* (pp. 309–317). Washington, DC: NASA Science and Technology Information Office.

Running, S., Baldocchi, D., Turner, D., Gower, S., Bakwin, P., & Hibbard, K. (1999). A global terrestrial monitoring network integrating tower fluxes, flask sampling, ecosystem modeling and EOS data. *Remote Sensing of Environment, 70*, 108–127.

[SAS] SAS Institute (1990). SAS/STAT® User's Guide, Version 6, 4th ed., vols. 1–2, Cary, North Carolina, USA, 943 pp. and 846 pp.

Scheffé, H. (1973). A statistical theory of calibration. *The Annals of Statistics, 1*, 1–37.

Sellers, P. J. (1987). Canopy reflectance, photosynthesis, and transpiration II. The role of biophysics in the linearity of their interdependence. *Remote Sensing of Environment, 21*, 143–183.

Sellers, P. J., Hall, F. G., Kelley, R. D., Black, A., Baldocchi, D., Berry, J., & Ryan, M. (1997). BOREAS in 1997: experiment overview, scientific results, and future directions. *Journal of Geophysical Research, 102* (D24), 28731–28769.

Song, C., Woodcock, C., Seto, K., Pax Lenney, M., & Macomber, S. (2001). Classification and change detection using Landsat TM data: when and how to correct atmospheric effects? *Remote Sensing of Environment, 75*, 230–244.

Steel, R., & Torrie, J. (1980). Principles and Procedures of Statistics—A Biometrical Approach. (2nd ed.). New York: McGraw-Hill.

Tabachnick, B., & Fidell, L. (1989). Using Multivariate Statistics. (2nd ed.). United Kingdom: Harper Collins Publishers.

Teillet, P. M., Barker, J. L., Markham, B. L., Irish, R. R., Fedosejevs, G., & Storey, J. C. (2001). Radiometric cross-calibration of the Landsat-7 ETM+ and Landsat-5 TM sensors based on tandem data sets. *Remote Sensing of Environment, 78*, 39–54.

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

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

Van Huffel, S. (Ed.) (1997). *Recent Advances in Total Least Squares Techniques and Errors-In-Variables Modeling*. Philadelphia: Society for Industrial and Applied Mathematics.

Virag, L. A., & Colwell, J. E. (1987, 26–30 October). An improved procedure for analysis of change in Thematic Mapper image-pairs. *Proceedings, Twenty-First International Symposium on Remote Sensing of Environment* (pp. 1101–1110). Ann Arbor, MI: ERIC.

Vogelmann, J. E., Helder, D., Morfitt, R., Choate, M. J., Merchant, J. W., & Bulley, H. (2001). Effects of Landsat 5 Thematic Mapper and Landsat 7 Enhanced Thematic Mapper Plus radiometric and geometric calibrations and corrections on landscape characterization. *Remote Sensing of Environment, 78*, 55–70.

White, J., Running, S., Nemani, R., Keane, R., & Ryan, K. (1997). Measurement and remote sensing of LAI in Rocky Mountain montane ecosystems. *Canadian Journal of Forest Research, 27*, 1714–1727.