International Journal of Remote Sensing
Vol. 31, No. 1, 10 January 2010, 141–158

Taylor & Francis
Taylor & Francis Group

Use of AVHRR NDVI time series and ground-based surveys for estimating county-level crop biomass

E. LOKUPITIYA*†, M. LEFSKY‡ and K. PAUSTIAN†
†Department of Soil and Crop Sciences and Natural Resource Ecology Laboratory, Colorado State University, Fort Collins, CO 80523, USA
‡Department of Forest, Rangeland, and Watershed Stewardship, Colorado State University, CO 80523, USA

(*Received 1 January 2008; in final form 9 June 2008*)

Crop biomass and residue production are major components of cropland carbon dynamics that can be estimated using yield data from ground-based surveys. In the USA, surveyed yield data are available at county level and have been widely used for various research, economic and policy purposes, in addition to biomass estimation. However, survey data may be unavailable for certain times and/or locations and thus biomass estimates using remotely sensed data might be used to fill in any missing biomass data for estimating residue production and carbon dynamics in croplands. Compared to ground-based surveys, remotely sensed data are collected on a regular schedule and may also provide more spatially resolved data. We analysed composite biweekly Normalized Difference Vegetation Index (NDVI) data obtained using the Advanced Very High Resolution Radiometer (AVHRR) sensor and crop aboveground biomass (AGBM) estimated from available county-level yield data reported by the National Agricultural Statistics Service (NASS) for three crops (corn, soybean and oats) during 1992, 1997 and 2002. The aim of the study was to explore the relationships between NDVI and crop biomass to complete the missing biomass data in counties where no NASS-reported yields are available for biomass estimation.

AGBM was estimated from Pathfinder biweekly NDVI, using canonical correlation analysis (CCA) and best subset multiple regressions incorporating canonical variates from NDVI time series. Cross-validation of model estimates was performed by randomly splitting the dataset into training and application subsets, simulating a 10–40% range of missing values. NDVI and crop biomass in Iowa during a given year were well correlated, with coefficient of determination (R^2) values > 0.8 in most cases. Using the available (training) data from a single year or a combination of years to derive models for filling the missing (validation) data within the same time period yielded a mean estimated biomass with $< 1\%$ relative error and bias. However, models applied to out-of-sample years had lower (< 0.4) R^2 values for the relationships between biomass and NDVI, although the mean residuals were low.

1. Introduction

The use of remote sensing for crop forecasting dates back to the early 1970s (reviewed in MacDonald and Hall 1980). Since then, agricultural agencies in various countries (e.g. Canada, Hungary and the USA) have used sensors such

*Corresponding author. Email: erandi@atmos.colostate.edu

International Journal of Remote Sensing
ISSN 0143-1161 print/ISSN 1366-5901 online © 2010 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/01431160902882579

as the National Oceanic and Atmospheric Administration Advanced Very High Resolution Radiometer (NOAA AVHRR) and Landsat imagery to forecast crop yields and crop conditions (Csornai *et al.* 2002, Reichert and Cassey 2002, NASS 2009). Several studies have used multispectral and hyperspectral data for spatially explicit crop forecasting and yield estimation. Many have used the Normalized Difference Vegetation Index (NDVI) to estimate biomass (Lozano-Garcia *et al.* 1991, Hansen and Schjoerring 2003) and crop yields (Tucker *et al.* 1983, 1985, Quarmby 1993, Senay *et al.* 2000, Yang *et al.* 2000, Doraiswamy *et al.* 2003, 2004, 2005, Hill and Donald 2003, Knudby 2004), with most applications at the field scale. Some of these studies have used integrated NDVI over the crop growth period to estimate biomass (e.g. Tucker *et al.* 1983, 1985, Quarmby 1993).

The NDVI is a vegetation index that ranges between −1 and +1, and is the difference between near infrared (NIR) and red (R) channels normalized by their sum [i.e. NDVI = (NIR − R)/(NIR + R)]. Increasing positive values indicate increasing greenness and negative values indicate non-vegetated surface features such as water, ice, snow and clouds. The relationship between vegetation indices such as the NDVI and biomass depends on the relationship between the vegetation index and the Leaf Area Index (LAI) and the relationship between LAI and biomass. According to Curran (1981), NDVI is related directly to biomass when biomass is linearly correlated with LAI. In comparing several vegetation indices for plants in salt marshes, Modenese *et al.* (2005) found NDVI to give the highest correlation with aboveground biomass (AGBM). NDVI time series from the AVHRR have been used to study changes in vegetation properties over time (Myneni *et al.* 1998, Shabanov *et al.* 2002, Xiao and Moody 2005). Currently, the National Agricultural Statistics Service (NASS) of the United States Department of Agriculture (USDA) uses biweekly composite NDVI images from AVHRR to monitor crop condition and for crop forecasting, while Landsat imagery is mainly used to estimate crop areas at the county level, under its Cropland Data Layer Program (Allen *et al.* 2002).

Our study was carried out as part of an attempt to assess the carbon dynamics of agricultural soils in the conterminous US. Crop residues are the main component of carbon inputs to agricultural soils, and include about 50–60% of the total crop AGBM produced in a given year. Crop yields can be used to estimate above- and below-ground biomass (and hence residue carbon inputs) based on allometric relationships for different crops (Buvanovsky and Wagner 1986, Campbell and Jong 2001, Prince *et al.* 2001, Williams and Paustian, submitted). In our earlier study (Lokupitiya *et al.* 2007) we considered a 16-year period from 1982 to 1997, during which digital databases were available from both the NASS (annual data) and the Census of Agriculture (data reported every 5 years); these data sets were combined to derive a comprehensive county-level crop yield database. Such county-level data have been widely used for both research and policy applications, including analyses of carbon cycling and greenhouse gas inventories in agriculture. One limitation of using the available survey data for estimating biomass (and residue carbon production) is the missing data for certain years in certain counties. Therefore, in this study we explored the potential use of a combination of remote sensing and statistical approaches in estimating annual AGBM production in those counties that do not have yield information for estimating biomass. In doing this, we used the county-level crop production in Iowa as a test bed.

2. Materials and methods

2.1 *Data used*

The remotely sensed data consisted of NOAA AVHRR biweekly NDVI images (1 km, i.e. ~100 ha resolution) of the conterminous US and the 1992 National Land Cover Dataset (NLCD) produced by the US Geological Survey (USGS) using Landsat images. Biweekly AVHRR NDVI images, which have been corrected for cloud-contaminated pixels, were obtained from the Pathfinder dataset (James and Kalluri 1994) for the years 1992, 1997 and 2002. Currently, global-scale NDVI time series are also available from the MODerate Imaging Spectroradiometer (MODIS). MODIS has advantages over AVHRR with regard to its onboard calibration ability, and higher spatial (250 m), spectral (i.e. narrower bandwidth in the near-infrared (NIR) and red range compared to AVHRR) and radiometric (12-bit) resolution. However, we used AVHRR NDVI time series rather than MODIS because our objective was to derive county-level biomass estimates for several years starting from 1992, at a time when MODIS was not operational. Annual yield data reported by NASS for 1992, 1997 and 2002 were used as the ground data for estimating annual AGBM production. These years were chosen for both the remotely sensed and ground data because they were the most recent years that both NASS and the Census of Agriculture have reported crop yields, and because the use of multiple years allowed us to evaluate the impact of differences in crop phenology and/or other temporal variability in crop biomass in different years. The state of Iowa was chosen for the study because annual crops dominate the state's land cover and it is among the states with the most comprehensive reporting of yields and crop areas by NASS, with few missing data, thus making it a suitable place to cross-validate and evaluate relationships between crop AGBM and NDVI.

For three years (1992, 1997 and 2002), biweekly images collected during the growing season (beginning of April to end of October) were combined as bands within a single multitemporal image. A county map of the USA (source: www-atlas.usgs.gov/) was used to obtain the county boundaries and extract composite NDVI images for Iowa. The NLCD coverage for Iowa was recoded to exclude non-crop areas and mask the composite NDVI images from the three years. The cropland areas selected consisted of the NLCD categories for small grains (i.e. oats in Iowa) and row crops (corn and soybean). As the NLCD coverage with crop layers had 30 m resolution, pixels from NLCD crop layers were aggregated to 1000 m resolution for masking the NDVI images. If 75% of the area of an AVHRR pixel (1 km spatial resolution) was classified as cropland in the NLCD (30 m resolution), then the AVHRR pixel was classified as annual cropland, otherwise the pixel was classified as non-cropland. Average biweekly NDVI pixel values (from the separate biweekly layers of the composite image) were then calculated for each county. Image processing was performed using Erdas Imagine 8.6 (Leica Geosystems) and ArcGIS 8.1 (Environmental Systems Research Institute, Inc. (ESRI)).

NASS reports annual county-level yields by extrapolating yield data collected from a representative sample of farms in each county. Crop yields and areas for the years 1992, 1997 and 2002 reported by NASS (www.nass.usda.gov/Data_and_Statistics/index.asp) were used to estimate the percentage crop area and AGBM production in a given year. In Iowa, the major crops in all three years (1992, 1997 and 2002) were corn (*Zea mays* L.) and soybean (*Glycine max* L.); oats (*Avena sativa* L.) was also considered in this study as a third major crop. Iowa has 99

counties; for each county, the percentage of the county's total area occupied by any of the three crops and the percentage of each crop within the total annual crop area were estimated.

AGBM was calculated from the county-level yield data reported by NASS, using allometric equations relating grain yield to biomass for each crop, based on information from past studies (Lawes 1977, Wych and Stuthman 1983, Buvanovsky and Wagner 1986, Russel 1991, Bolinder *et al.* 1997, Bruce and Langdale 1997, Juma *et al.* 1997, Peters *et al.* 1997, Pierce and Fortin 1997, Vanotti *et al.* 1997, Campbell and Jong 2001, Prince *et al.* 2001, IPCC 2006). In doing so, the crop yields were corrected for moisture content and converted to biomass, dry matter yield (equation (1)); this yield was then used in crop-specific allometric equations incorporating harvest indices (S. A. Williams, personal communication) to estimate the above-ground non-harvested dry biomass (equations (2), (3) and (4)). The estimated total AGBM (i.e. both grain and residue dry biomass) was used for developing model relationships with NDVI values. The grain dry mass for a particular crop (GDM_{crop} in kg ha^{-1}) is given by:

$$GDM_{crop} = Y \times F_u \times F_{DM} \times 1.12 \quad (1)$$

where Y is the yield reported by NASS (bushels $acre^{-1}$), F_u is the unit conversion factor for converting bushels to pounds, and F_{DM} is the fraction of dry biomass (i.e. after removal of the fraction of moisture); lb $acre^{-1}$ was converted to kg ha^{-1} by multiplying by 1.12. The corresponding residue dry mass (RDM) for each crop (corn, soy and oats) is then given by:

$$RDM_{corn} = (GDM_{corn} \times 1.03) + 610 \quad (2)$$

$$RDM_{soybean} = (GDM_{soybean} \times 0.93) + 1350 \quad (3)$$

$$RDM_{oats} = (GDM_{oats} \times 1.06) + 849 \quad (4)$$

2.2 *Dependent and independent variables*

As the aim of the study was to derive general relationships between remotely sensed data and annual crop biomass production estimated from yield statistics, the following crop variables were considered as the dependent variables:

1. Mean annual AGBM per hectare of each crop (i.e. [C,S,O]AGBM kg ha^{-1}, where C,S,O denote corn, soybean and oats, respectively)
2. Area-weighted biomass (AWBM in kg ha^{-1}). This is the sum of the AGBM of the crops, weighted by their area fraction: AWBM = $\sum$ ([C,S,O]AGBM × [C,S,O]AF), where AF is the area of the particular crop as a fraction of the total crop area.

AWBM from the three crops was included, assuming it would match better with the NDVI signal at the pixel level, as it represents the mixed-crop biomass per hectare. Biomass data for the above variables were estimated for three different years (i.e. 1992, 1997 and 2002).

The NDVI pixel values for 2-week periods encompassing the growing season (April to October) of 1992, 1997 and 2002 were extracted at county level, and considered as

the independent variables. As there were slight differences in the beginning and end dates of the biweekly time intervals in the three years, biweekly periods of the growing season in 1997 and 2002 were matched with the corresponding 1992 periods that had the greatest date overlap for the analyses and interpretation of the results.

Initial analyses of the data showed significant correlations between temporally adjacent NDVI values, such that the values from different time periods cannot be treated as independent (i.e. they exhibit multicollinearity). To address the problem of multicollinearity, we used canonical correlation analysis (CCA) to model crop biomass as a function of NDVI.

2.3 *CCA*

CCA is a statistical approach that summarizes multiple variables from two datasets as pairs of canonical variates. Although CCA treats both sets of variables identically, it is convenient to label one dataset as independent and the other as dependent; in this case these are the remotely sensed NDVI values and crop biomass values, respectively. Pairs of canonical variates are created as linear combinations of the original variables in each dataset. CCA maximizes the correlation between linear combinations of variables from one set with linear combinations of variables from the second set. The advantage of CCA is that it quantifies the redundancy in each set of variables. This, makes it possible to analyse both X and Y variables in terms of their relationships to other variables within their own dataset and to variables in the other dataset.

In the CCA, biweekly NDVI values were considered as one set of variables, and county-level corn aboveground biomass (CAGBM), soybean aboveground biomass (SAGBM), oats aboveground biomass (OAGBM) and area-weighted biomass (AWBM) were considered as a second set of variables. As mentioned above, one advantage of CCA is it eliminates the multicollinearity associated with biweekly NDVI values. It also provides more interpretable results, as the patterns of correlation within and between datasets are reduced to a smaller number of variates that are ranked by their importance in explaining variance in each dataset. As an example, we could have used separate multiple regression analyses to predict AGBM for corn and soybean from NDVI values. However, doing so would obscure the fact that corn and soybean AGBM covary because of two effects: the limitation on total area of cropland in each county, and the fact that good years for corn are generally good years for soybean as well. In addition, the coefficients for the resulting regression equations would be difficult to interpret, as they would combine numerous effects into a single linear combination of the independent variables, whereas CCA separates out effects into separate canonical variates.

Best subset multiple regression analyses were performed to estimate each dependent variable using the canonical variates derived from the NDVI dataset. To obtain the model with the highest predictive power, the best subset of the entire set of independent variables was chosen based on Mallows' C_p value (a measure of model fit) and coefficient of determination (R^2) values for each subset of the independent variables; the subset that gave the lowest C_p (with $C_p \leq 1$, when $p =$ number of parameters) and the highest R^2 was chosen as the best.

These analyses were performed using datasets in which 10, 20 or 40% of the data were randomly removed to represent missing data (the validation dataset), and the remaining data were used as a training data set to derive regression equations. To obtain results that were insensitive to the particular selection of missing data,

canonical correlations and best subset multiple regression analyses were performed iteratively 100 times for cross-validation. Analyses were performed in the IDL software package (Research Systems 2005). The following scenarios were evaluated:

1. Analyses were carried out separately for each year: 10, 20 and 40% of the county-level biomass data from that year were removed from the dataset for each year and treated as missing data; as there were 99 counties, biomass data within 10, 20 and 40 counties were randomly removed for each iteration. Regression equations developed using the remaining (training) dataset in each iteration were then applied to the reserved counties. In addition, to determine the year-to-year consistency of the equations, these same equations for any single year were extrapolated to estimate AGBM values for the other two years.
2. Data from two years were combined and separated into training and validation sets with 10, 20 and 40% data missing (i.e. 20, 40 and 80 county-level data points missing from the combined data from the two years, for each iteration). The unused year's data were used to check the ability to extrapolate the regression equations beyond those years.
3. Data from all three years were combined; 10, 20 and 40% of the missing data from the same set (i.e. 30, 60 and 120 biomass data points were randomly removed from 297 data points created when the county-level data for all three years were combined, for each iteration) were considered as the validation data set for the models obtained using the remaining (training) data.

The above three levels of missing data (i.e. 10, 20 and 40%) were chosen for this study to test the usability and validity of the model relationships at different levels of missing data.

3. Results

The final composite NDVI images for the crop layers in 1992, 1997 and 2002 are shown in figure 1. The analyses included all 99 counties in Iowa and the average number of cropland pixels per county was 926 ± 367. When the temporal variation in NDVI was studied for each year, we found that NDVI increased from the beginning of April, and remained high from mid-June to early October in 1992 and 2002 and from mid-June to mid-September in 1997. The highest NDVI in all three years was observed from the end of July or early August until early September. The highest county-averaged NDVI observed for Iowa cropland was 0.63, 0.71 and 0.79 in 1992, 1997 and 2002, respectively. The increasing trend over the years was more conspicuous during the period of high NDVI; however, NDVI in 1997 was lower in certain biweekly periods towards the end of the growing season, compared to the corresponding periods in 1992 (figure 2).

According to the crop area information reported by NASS, about 80% of the counties (78 out of 99) in Iowa had more than 50% land cover with these crops, and the average crop area in the counties was 89%. The percentage areas occupied by corn within the cropland of a county were in the range 52–88% in 1992, 43–79% in 1997 and 38–70% in 2002. The percentage areas for soybean were 3–48% in 1992, 14–55% in 1997 and 24–59% in 2002, and those for oats were 0–12% in 1992, 4–9% in 1997 and 0–6% in 2002. In 1992, the average percentage crop areas for corn, soybean and oats were 61, 37 and 2%, respectively; in both 1997 and 2002, the average percentage annual crop areas for corn, soybean and oats were 53, 46 and 1%, respectively (figure 3).

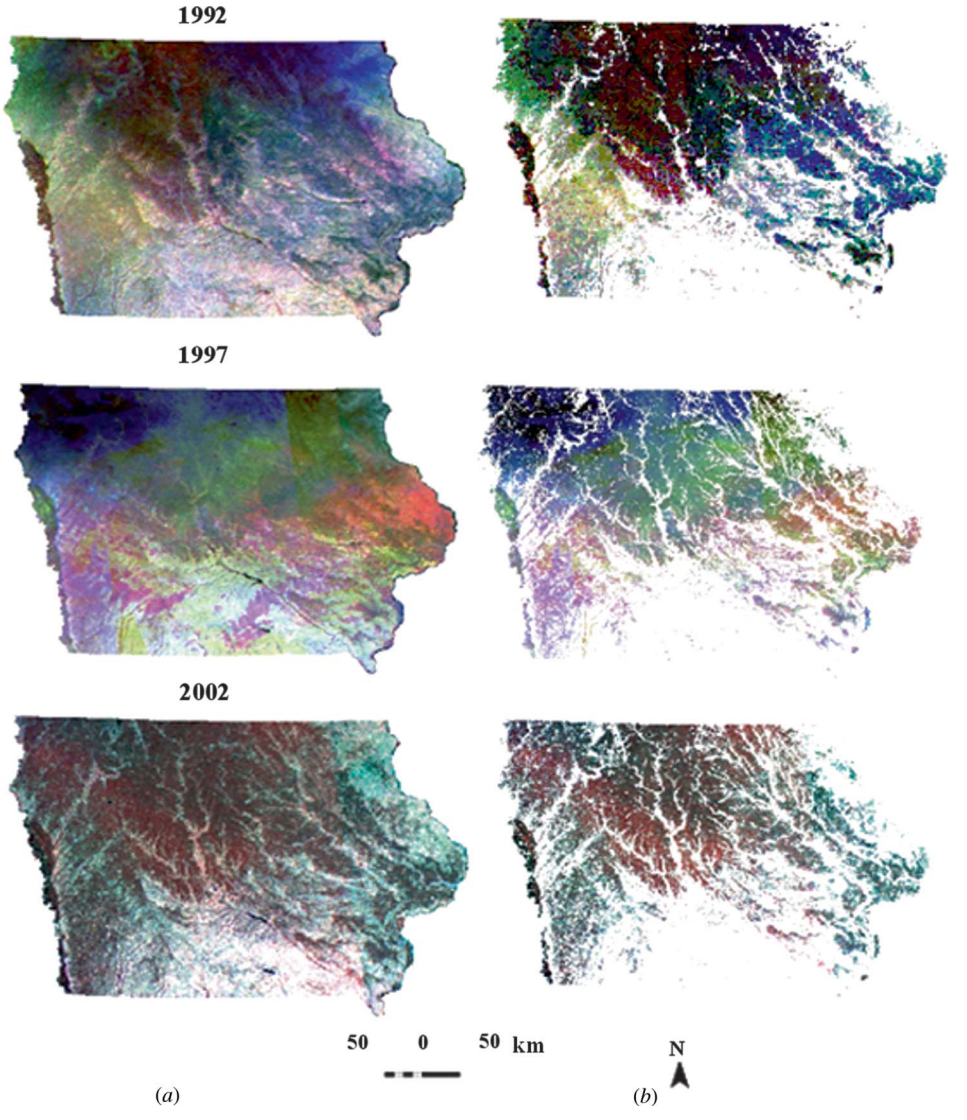

Figure 1. Images used in the data analyses. False-colour composites of (*a*) NDVI images during the crop season and (*b*) the same images after being masked for crop layers using the NLCD Landsat image for the years 1992, 1997 and 2002. Blue (B), green (G) and red (R) layers in the images correspond to the biweekly NDVI from the following periods within each year: 1992: 17–30 April (B), 1–14 May (G) and 15–28 May (R); 1997: 11–24 April (B), 25 April–8 May (G) and 9–22 May (R); 2002: 19 April–2 May (B), 3–16 May (G) and 17–30 May (R).

The county-level average yields reported by NASS for these three crops were slightly different between the three years, and the highest yields and estimated biomass values were found in 2002; average corn yields were 9, 8.5 and 10 Mg ha^{-1}, soybean yields were 2.9, 3 and 3.2 Mg ha^{-1}, and oats yields were 2.4, 2.6 and 2.7 Mg ha^{-1}in 1992, 1997 and 2002, respectively. Thus the yields of the crops increased over time, and this increased trend in yields was also reflected in the increased NDVI, with 2002 having higher overall NDVI than the other two years (figure 2).

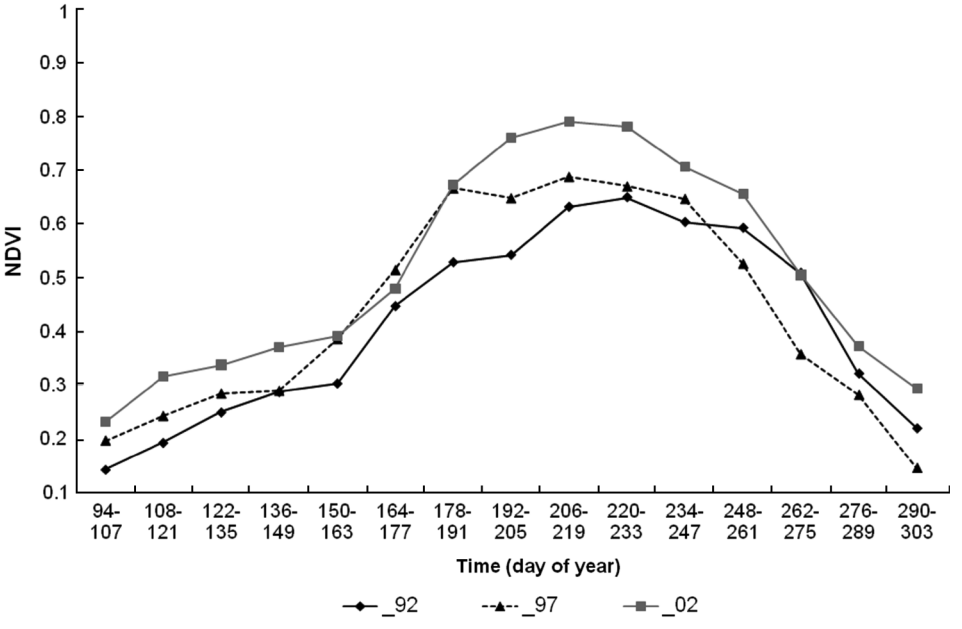

Figure 2. Variation in NDVI averaged for the whole state of Iowa during biweekly time periods in 1992 and corresponding time periods in 1997 and 2002. The same trend was observed at individual county level.

3.1 *CCA*

For any of the years or combination of the years, the highest correlation between the first NDVI and crop biomass canonical variates ranged between 0.9 and 0.95; the corresponding p-values of < 0.0001 rejected the null hypothesis that all the canonical correlations are zero. The results of the multivariate statistical tests also confirmed the significance of the canonical correlations obtained from the analyses.

Of the four NDVI canonical variates that contributed towards the observed model relationships with crop biomass (tables 1 and 2), the first canonical variate (CV1) had the highest loadings from the original NDVI dataset; CV1 had positive loadings from NDVI pixel values of the biweekly periods in early April to mid/end June, and early September to the end of October in all three years (figure 4). In all the models, CV1 was negatively correlated with crop biomass (tables 1 and 2). As CV1 has the highest (positive) loadings from NDVI at the early and end phases of the crop growth cycle, the negative coefficient of CV1 with biomass in all the models indicates that NDVI is negatively or less correlated with biomass during the early and final phases of the crop growth cycle. Thus CV1 seemed to represent a less green or non-green component of the crops, as it had the highest loadings when NDVI values were low. There was a positive correlation between crop biomass and original NDVI pixel values during the period from late June to late August; however, this correlation varied in value among the crops and different years, and ranged from 0.1 to 0.84. The loadings from the original NDVI pixel values on the second, third and fourth canonical variates were very low, except for the relatively high loadings on the second canonical variate in 1997 (figure 4). However, all four canonical variates seemed to follow the same pattern of variation. All four canonical variates contributed towards the model

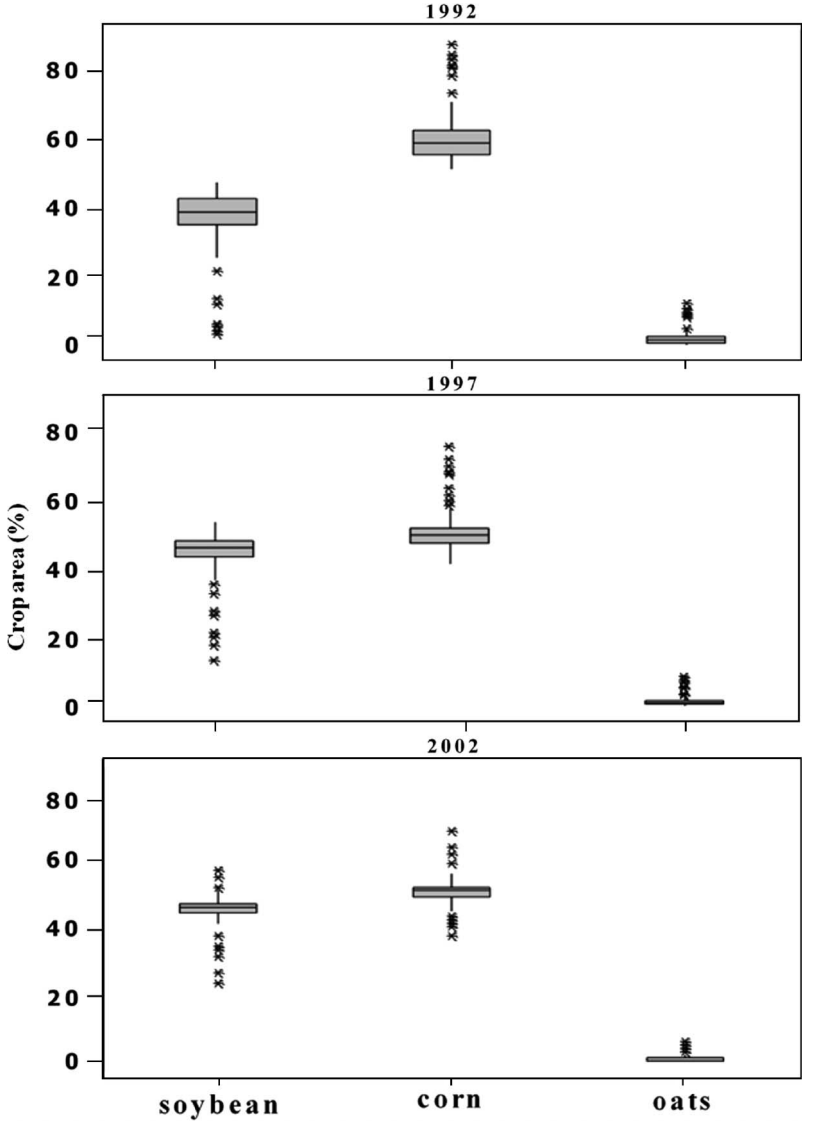

Figure 3. Box plots for crop area occupied by each crop as a percentage of total crop area in 1992, 1997 and 2002.

relationships with the biomass of each individual crop and area-weighted total biomass; different coefficients derived for each canonical variate in the model relationships (tables 1 and 2) predicted the biomass values separately for each specific crop.

3.2 *Model relationships between NDVI-derived canonical variates and crop biomass in relation to the extent of missing data*

Model relationships were obtained for 10, 20 and 40% missing data in biomass under the scenarios (1) data from a single year, (2) two years of data combined,

Table 1. Best subset multiple regression models between the canonical variates (CV1–CV4) from the relevant NDVI pixel values from different biweekly periods and AGBM in 1992 data when 10, 20 and 40% data were missing.

	R^2
40% data missing	
CAGBM = 25 941.4 – 354.7 × (CV1) – 144.2 × (CV2) – 174.4 × (CV3) – 132.7 × (CV4)	0.83
SAGBM = 15 324.9 – 72.8 × (CV1) – 27.4 × (CV2) – 8.2 × (CV3) – 27 × (CV4)	0.62
OAGBM = –12 382.4 – 244.1 × (CV1) – 22.7 × (CV2) – 22 × (CV3) + 4.4 × (CV4)	0.92
AWBM = 25 422.5 – 141.7 × (CV1) – 15.1 × (CV2) – 193 × (CV3) – 172.3 × (CV4)	0.61
20% data missing	
CAGBM = 29 598.2 – 324.8 × (CV1– 114.3 × (CV2) – 220.3 × (CV3) – 187 × (CV4)	0.81
SAGBM = 17 267.1 – 66.6 × (CV1) – 22.5 × (CV2) – 1.9 × (CV3) – 39.3 × (CV4)	0.59
OAGBM = –12 223.1 – 223 × (CV1) + 18.5 × (CV2) + 52.6 × (CV3) – 12.5 × (CV4)	0.92
AGBM = 28 438 – 128 × (CV1) – 21 × (CV2) – 237 × (CV3) – 191 × (CV4)	0.59
10% data missing	
CAGBM = 29 069.5 – 313.1 × (CV1) – 97.4 × (CV2) – 221.8 × (CV3) – 200.4 × (CV4)	0.81
SAGBM = 17 136.9 – 64.6 × (CV1) – 16.1 × (CV2) + 0.01 × (CV3) – 38.7 × (CV4)	0.58
OAGBM = –12 885.2 – 215.3 × (CV1) + 10.7 × (CV2) + 61 × (CV3) + 9 × (CV4)	0.92
AWBM = 28 509.7 – 124.4 × (CV1) – 26 × (CV2) – 251.1 × (CV3) – 228.8 × (CV4)	0.58

CAGBM, corn aboveground biomass; SAGBM, soybean aboveground biomass; OAGBM, oats aboveground biomass; AWBM, area-weighted biomass.

Table 2. Best subset multiple regression models between the canonical variates (CV1–CV4) from NDVI pixel values from different biweekly periods and AGBM in 1997 and 2002 for training data with 10% data missing.

	R^2
1997	
CAGBM = –16 067 – 1372.9 × (CV1) – 186.3 × (CV2) – 124.7 × (CV3) + 196.6 × (CV4)	0.83
SAGBM = –4869.7 – 375.7 × (CV1) + 171.8 × (CV2) + 194.7 × (CV3) + 39.2 × (CV4)	0.78
OAGBM = –9096.1 – 557.1 × (CV1) – 79.7 × (CV2) + 9.4 × (CV3) – 191.2 × (CV4)	0.55
AWBM = 25 422.5 – 141.7 × (CV1) – 15.1 × (CV2) – 193 × (CV3) – 172.3 × (CV4)	0.84
2002	
CAGBM = –100 639.1 – 824.4 × (CV1) – 29.4 × (CV2) + 45.3 × (CV3) + 4.3 × (CV4)	0.85
SAGBM = –43 935.6 – 221.4 × (CV1) – 35 × (CV2) – 22.1 × (CV3) – 1.3 × (CV4)	0.83
OAGBM = 29 324.2 – 216.8 × (CV1) – 23.1 × (CV2) – 23 × (CV3) + 17.8 × (CV4)	0.24
AWBM = –84 203.4 – 652.6 × (CV1) – 41.6 × (CV2) + 30.9 × (CV3) – 0.7 × (CV4)	0.86

CAGBM, corn aboveground biomass; SAGBM, soybean aboveground biomass; OAGBM, oats aboveground biomass; AWBM, area-weighted biomass.

and (3) all three years of data combined. In general, relatively high R^2 values were obtained for the models where both training and validation data sets were from the same year (e.g. tables 1 and 2, figure 5). However, when these model relationships were extrapolated to a different year or a combination of years other than those used in the model derivation, the same model relationships yielded relatively low R^2 values (figure 5). The extent of the difference between the mean estimated and observed values varied depending on the observed values in the year or the two years combined in the training dataset. For instance, under scenario 2, when the models for CAGBM from combined data in 1992/1997 were applied on the extrapolated data in 2002, the mean estimated values were 12% lower than the observed values [i.e. relative error 12%, where relative error = (observed – estimated)/observed]. When models from the 1992/

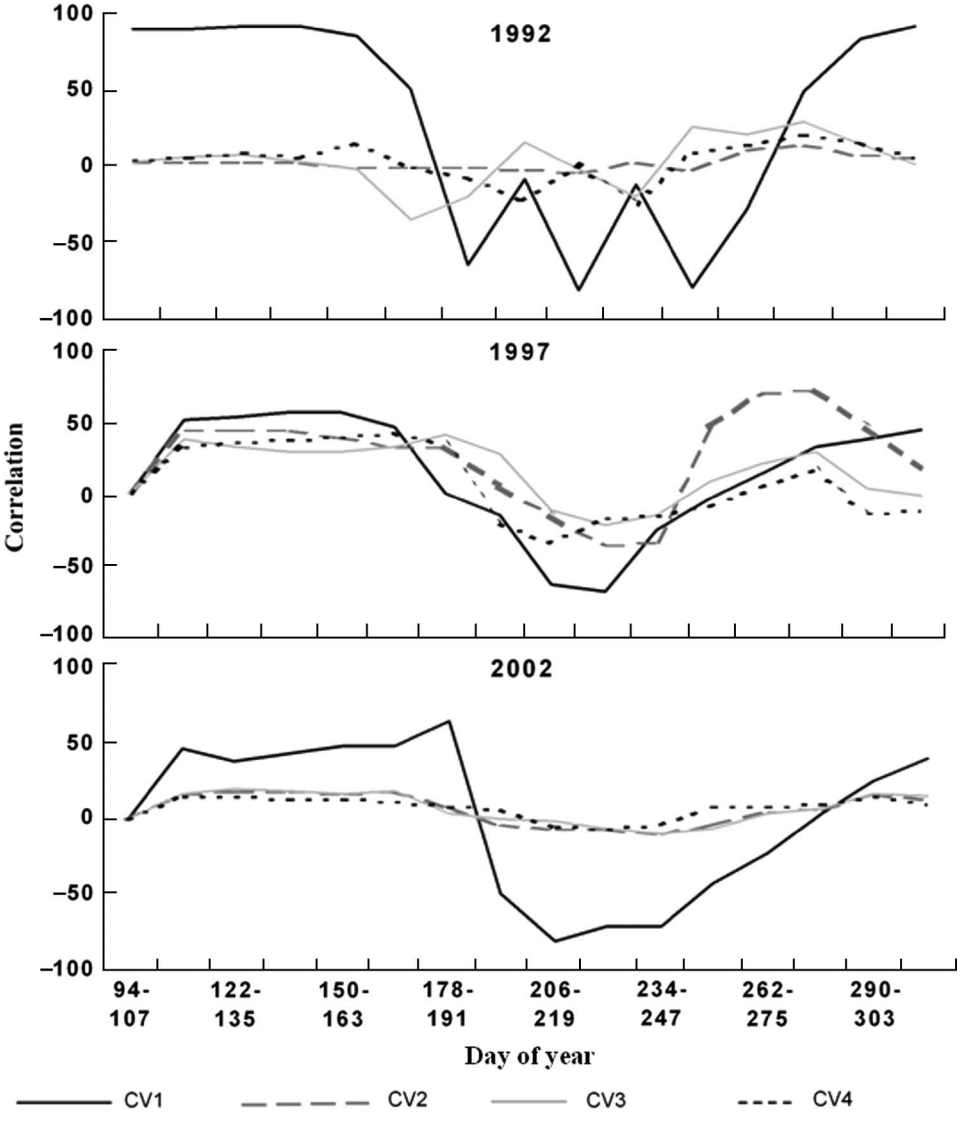

Figure 4. The correlation between the original NDVI pixel values from different biweekly periods and canonical variates with 90% data in the training dataset.

2002 combination were applied on the 1997 data, the mean estimated values were 3% higher than the observed values (i.e. relative error – 3%). The ratio of the root mean square error to the mean estimated value (RMSE/MPRED) was in the range 0.1–0.2 for corn, 0.05–0.2 for soybean, 0.1–0.3 for oats and 0.1–0.2 for AWBM, for the extrapolated datasets under the first and second scenarios. Under the third scenario, when the data from all three years were combined for the analyses, the model relationships estimated biomass values with $< 1\%$ relative error and values of < 0.05 for RMSE/MPRED in both training and validation data sets when 10, 20 and 40% data were missing. Thus the mean estimated values were very close (within 4% across all the biomass variables) to

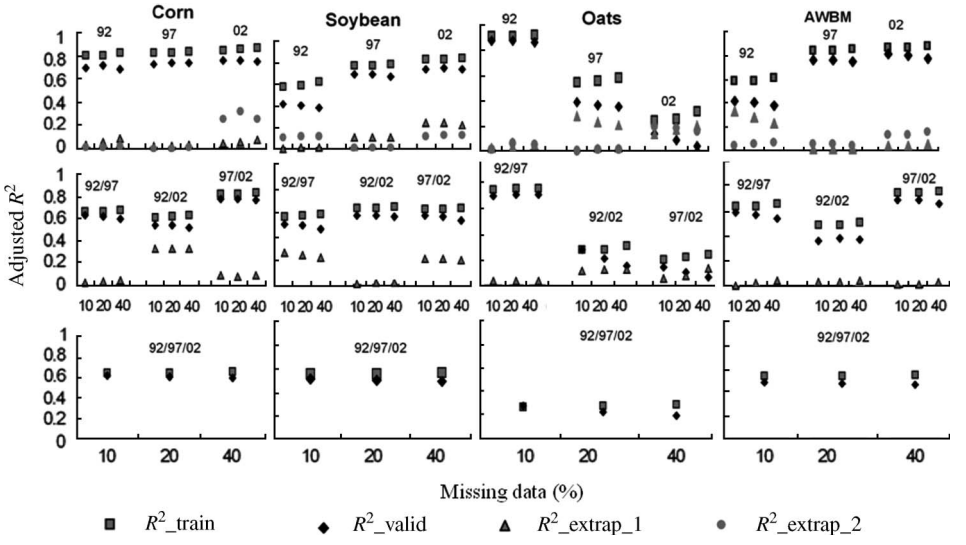

Figure 5. Adjusted R^2 values for the estimated model relationships between NDVI canonical variates and biomass of each crop (corn, soybean, oats) and area weighted biomass (AWBM) under the three data scenarios: single year data (top), two years combined (middle), and all three years combined (bottom). R^2_train is R^2 for the models derived using the training data when 10, 20 and 40% of the data were missing; R^2_valid is R^2 when the model from training data is applied on a data set with 10, 20 and 40% data missing randomly; R^2_extrap_1 and R^2_extrap_2 are R^2 when the model/s obtained using the training data from any single year or combination of years is/are applied on the data of the remaining year/s. R^2_extrap_1 corresponds to the earlier year of the remaining year/s, and R^2_extrap_2 corresponds to the later year of the remaining years.

the mean observed values for each biomass variable (table 3; figure 6). However, the R^2 values were slightly lower than those obtained for the training and validation data sets under the first two scenarios (i.e. the single year scenario and with combined data for two years; figure 5); the observed R^2 for the validation data under the third scenario was in the range 0.45–0.5 for AWBM, 0.6–0.7 for CAGBM, 0.5–0.6 for SAGBM and 0.2–0.3 for OAGBM.

We analysed the average bias (i.e. average residuals) for all three scenarios. Soybean always had the lowest bias (mostly within 5 kg ha^{-1}; figure 7). However, oats, being a minor crop with minimum crop area, showed the highest bias (still within 20 kg ha^{-1}) in relation to the mean observed biomass values. Corn had very low biases ($<$ 7 kg ha^{-1}) in 1992 and 1997, but the bias was slightly higher (close to 15 kg ha^{-1}) in 2002, in a year when the average corn biomass was much higher compared to the other two years; but this bias was negligible because the average observed corn biomass in 2002 was 18 353 kg ha^{-1}.

4. Discussion

We tested the feasibility of using remotely sensed AVHRR NDVI to estimate county-level crop biomass as a complement to ground-survey based data (e.g. to fill in missing data). Using raw NDVI pixel values as independent variables in models to estimate biomass was determined to be inappropriate because of the presence of multicollinearity among NDVI values for certain time periods, especially during early crop

Table 3. Mean values with standard errors from the observed and estimated values for training and validation biomass data when the data from all three years are combined for developing model relationships using canonical correlation analysis.

Variable	Mean observed biomass for all data (kg ha^{-1})	Missing data (%)	Mean estimated values for training data (kg ha^{-1})	Mean estimated biomass for validation data (kg ha^{-1})
CAGBM	17 281 ± 1880	10	16 910 ± 32	16 912 ± 262
		20	16 916 ± 51	16 894 ± 168
		40	16 916 ± 94	16 888 ± 132
SAGBM	6933 ± 523	10	6945 ± 8	6950 ± 65
		20	6948 ± 14	6943 ± 46
		40	6948 ± 25	6942 ± 33
OAGBM	5085 ± 1231	10	5090 ± 21	5105 ± 96
		20	5095 ± 36	5096 ± 85
		40	5091 ± 52	5087 ± 75
AWBM	12 646 ± 1292	10	12 407 ± 26	12 405 ± 169
		20	12 411 ± 43	12 400 ± 110
		40	12 412 ± 67	12 387 ± 94

CAGBM, corn aboveground biomass; SAGBM, soybean aboveground biomass; OAGBM, oats aboveground biomass; AWBM, area-weighted biomass.

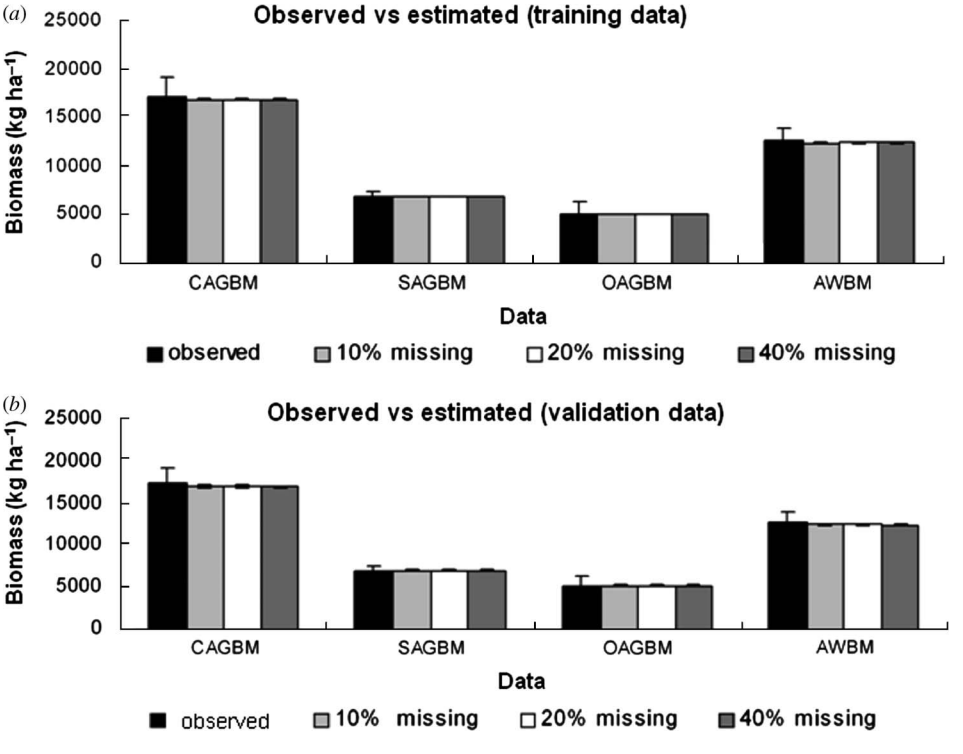

Figure 6. County-level mean observed biomass data compared against the biomass estimates from models based on different levels of missing data in (*a*) training data and (*b*) validation data, when all the data (i.e. from all three years) were combined.

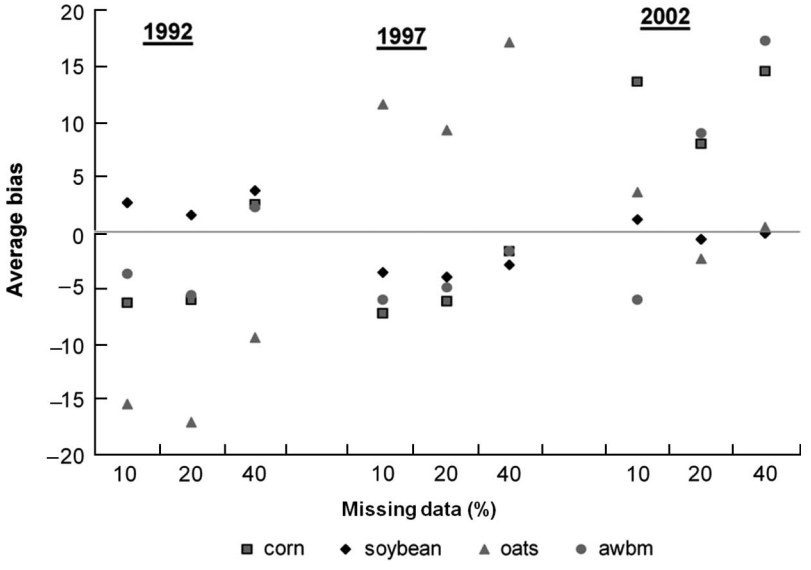

Figure 7. Average bias (i.e. average residuals) of the estimated values from 100 iterations when 10, 20 and 40% data were missing within a single year.

growth. Multicollinearity problems were avoided by using canonical correlation analyses that combined information from closely related, county-averaged biweekly NDVI pixel values during the crop growth period into separate canonical variates. AGBM variables derived from NASS yield data served as the dependent variables. As biomass was estimated using linear relationships incorporating observed yields, any spatial or temporal variation in crop production due to environment factors such as precipitation and temperature was directly reflected in the estimated biomass.

In analysing the correlations between original NDVI and individual crop biomass variables, the biomass of all three crops was positively correlated with NDVI from the end of June to the end of August when the NDVI was at a maximum. As corn and soybean had the highest crop areas (mostly $> 90\%$) and biomass, the largest contribution to NDVI must have come from these two crops. The usual harvest time for corn and soybean is October, and harvest dates for oats normally fall in July in Iowa. After the end of August, the NDVI was negatively correlated with biomass, during a period when highest biomass should be found in the crops, especially for soybean and corn, due to maturation and end phase of grain filling. The grain-filling period, which allocates 40–50% of the biomass to grain, usually falls within the past 50–60 days of the growth cycle in corn plants. Thus the maximum NDVI was observed near the beginning of the grain-filling period of the corn plants. The negative coefficient of the first canonical variate with biomass (and the observed correlation between biomass and original NDVI) denotes that biomass was negatively correlated with NDVI when the crop was close to harvest (in September–October). These results are in accordance with a study by Curran (1981), in which a negative correlation was found with high biomass in vegetation and NDVI. Canopy opening and similar reflectance from drier or senescing vegetation towards the end of the crop cycle and soil in the NIR and visible range (Todd *et al.* 1998, Campbell 2002) probably led to lower NDVI and the negative correlation with biomass.

The coefficients obtained for the first canonical variate for different crops and different years were distinct, indicating the differences among the crop growth cycles and the differences in the biweekly periods in terms of crop phenology in different years. Both NDVI and biomass depend on external environmental factors such as precipitation and temperature; crop biomass depends on the number of growing degree-days and the temperature during the grain filling. Therefore, when such environmental factors vary between different years, it makes it less accurate to use the model derived from one year or a combination of two years for predictions during a different year. Our study confirmed this by showing low R^2 values when the models from scenarios 1 and 2 were extended to a different year, although the mean residual values were low. Overall, the best results were obtained when the models from the training data were applied for the missing data within the same period (i.e. single year or the combination of years). According to the results of the analyses, the mean estimated values or R^2 values were not very dependent on the extent of the missing data; the results obtained for cross-validation using all levels (i.e. 10, 20 and 40%) of missing data were very close, indicating that our method could be used with even more than 40% of the values missing (i.e. as the basis for crop biomass estimates with a smaller quantity of training data).

Although the first canonical variate was the most useful canonical variate in predicting and interpreting the NDVI–biomass relationship, the purpose of the current study was to derive model relationships between NDVI and crop biomass. Thus the other canonical variates were also considered in the best subset multiple linear regression analyses, to select the subset of the canonical variates that would give the best model fit between the NDVI and biomass of each individual crop. The current study shows that CCA followed by best subset multiple regression analyses is a viable approach for predicting biomass from NDVI, when there are missing data in the reported crop statistics.

The mean estimated values for training data and validation data from 100 runs for the three selected levels of missing data were very close, within 5% relative error. When models were derived from the available data within the same time period (within the same year or the combination of the years that are relevant to the missing data), this approach was highly successful. R^2 values gradually decreased when we increased the time period for choosing training and validation data from one, two or three years. However, using the data from all three years was the best single approach to estimate missing data for all of the participating years, as there was no significant drop in the R^2 value when the models based on the training data were applied on the validation (i.e. missing) data (figure 5, bottom graphs), and the model relationships yielded the lowest relative errors. Although the single-year scenario had higher R^2 values for the model relationships, the fall in R^2 when the models were applied on the validation data was higher, making it less suitable in application for filling missing data. The deviation of the predicted biomass from the observed data was also higher for the single-year scenario, compared to the scenario that had the data from all three years combined. When using only one or two years' data to predict the missing data in a third year, R^2 values were low. As Iowa has fairly comprehensive reporting by NASS on crop production, the data set was well suited to validate the model relationships between NDVI and biomass under several 'forced' scenarios of missing data at the county level. As the method used gave promising results, further testing of this approach using states of more variable climate is warranted.

5. Conclusions

Overall, the method used in the current study yielded model relationships between NDVI canonical variates and biomass variables with high R^2 values, and estimated values with low relative errors (and RMSE/MPRED ratios). CCA between NDVI pixel values and biomass data and subsequent best subset regressions incorporating canonical variates were used as a methodology for avoiding the effect from multi-collinearity among adjacent biweekly NDVI pixel values. The results show that the model relationships derived from this approach are valid in predicting biomass values for up to 40% of the missing data. However, the missing data should be filled only with the models derived from the available data pertaining to the same time period, to better account for the specific phenological changes over the corresponding time period. Application of the models based on a single year or a combination of years on out-of-sample year/s proved to be less valid (with low R^2), suggesting an impact from the weather variability (e.g. drought conditions) and other external factors corresponding to the out-of-sample year/s. Of the scenarios considered in the study, the best results were found when the models based on the training data from all three years were used in filling the validation (missing) data for that period. CCA revealed that NDVI and crop biomass are well correlated during the middle of the crop growth from mid-June to the end of August, and the use of all the canonical variates from the original biweekly NDVI pixel values in subsequent best subset multiple regression analyses was needed in determining model relationships for biomass of individual crops. Overall, we found this approach suitable for filling missing biomass data at the county level, to be used in estimating residue carbon inputs or for similar purposes. As it incorporates low-resolution AVHRR NDVI data and available county-level yield data as the input data for model derivation, we find this is a better approach for regional- or national-scale studies than for field-scale studies. This approach could be further enhanced in the future, by using MODIS NDVI data that have higher spatial, spectral and radiometric resolution.

Acknowledgements

This research was supported by the Cooperative State Research, Education, and Extension Service, US Department of Agriculture, under Agreement No. 2001-38700-11092 to the Consortium for Agricultural Soils Mitigation of Greenhouse Gases (CASMGS). We thank Steve Williams, Sunil Kumar and Kay Dudek from Colorado State University.

References

Bolinder, M.A., Angers, D.A. and Dubuc, J.P., 1997, Estimating shoot to root ratios and annual carbon inputs in soils for cereal crops. *Agriculture, Ecosystems and Environment*, **63**, pp. 61–67.

Bruce, R.R. and Langdale, G.W., 1997, Soil carbon level dependence upon crop culture variables in a thermic-udic region. In *Soil Organic Matter in Temperate Agroecosystems*, E.A. Paul, K. Paustian, E.T. Elliott and C.V. Cole (Eds), pp. 247–261 (New York: CRC Press).

Buyanovsky, G.A. and Wagner, G.H., 1986, Post-harvest residue input to cropland. *Plant and Soil*, **93**, pp. 57–65.

Campbell, C.A. and de Jong, R., 2001, Root-to-straw ratios: influence of moisture and rate of N fertilizer. *Canadian Journal of Soil Science*, **81**, pp. 39–43.

Campbell, J.B., 2002, *Introduction to Remote Sensing* (New York: Guilford Press).

Csornai, G., Wirnhardt, Cs., Suba, Zs., Somogyi, P., Nádor, G., Martinovich, dr. L., Tikász, L., Kocsis, A., Tarcsai, B. and Zelei, Gy., 2002, Remote sensing based crop monitoring in Hungary. Available online at http://www.fomi.hu/internet/magyar/Projektek/remsensmonit.htm.

Curran, P.J., 1981, Multispectral remote sensing for estimating biomass and productivity. In *Plants and Daylight Spectrum*, H. Smith (Ed.), pp. 5–8, (New York: Academic Press).

Doraiswamy, P.C., Hartfield, J.L., Jackson, T.J., Akhmedov, B., Prueger, J. and Stern, A., 2004, Crop condition and yield simulations using Landsat and MODIS. *Remote Sensing of Environment*, **92**, pp. 548–559.

Doraiswamy, P.C., Moulin, S., Cook, P.W. and Stern, A., 2003, Crop yield assessment from remote sensing. *Photogrammetric Engineering and Remote Sensing*, **69**, pp. 665–674.

Doraiswamy, P.C., Sinclair, T.R., Hollinger, S., Akhmedov, B., Stern, A. and Prueger, J., 2005, Application of MODIS derived parameters for regional crop yield assessment. *Remote Sensing of Environment*, **92**, pp. 192–202.

Hansen, P.M. and Schjoerring, J.K., 2003, Reflectance measurement of canopy biomass and nitrogen status in wheat crop using normalized difference vegetation indices and partial least squares regression. *Remote Sensing of Environment*, **86**, pp. 542–553.

Hill, M.J. and Donald, G.E., 2003, Estimating spatio-temporal patterns of agricultural productivity in fragmented landscapes using AVHRR NDVI time series. *Remote Sensing of Environment*, **84**, pp. 367–384.

IPCC, 2006, 2006 IPCC guidelines for national greenhouse gas inventories, Prepared by the National Greenhouse Gas Inventories Programme, H. S. Eggleston, L. Buendia, K. Miwa, T. Ngara and K. Tanabe (Eds.), (Japan: IGES).

James, M.E. and Kalluri, S.N.V., 1994, The Pathfinder AVHRR land data set: an improved coarse-resolution data set for terrestrial monitoring. *International Journal of Remote Sensing*, **15**, pp. 3347–3364.

Juma, N.G., Izaurralde, R.C., Robertson, J.A. and McGill, W.B., 1997, Crop yield and soil organic matter trends over 60 years in a typic cryoboralf at Breton, Alberta. In *Soil Organic Matter in Temperate Agroecosystems*, E.A. Paul, K. Paustian, E.T. Elliott and C.V. Cole (Eds), pp. 273–281 (New York: CRC Press).

Knudby, A., 2004, An AVHRR-based model of groundnut yields in the Peanut Basin of Senegal. *International Journal of Remote Sensing*, **25**, pp. 3161–3175.

Lawes, D.A., 1977, Yield improvement in spring oats. *Journal of Agricultural Science*, **89**, pp. 751–757.

Lokupitiya, E., Breidt, F.J., Lokupitiya, R., Williams, S. and Paustian, K., 2007, Deriving comprehensive county-level crop yield and area data for U.S. cropland. *Agronomy Journal*, **99**, pp. 73–681.

Lozano-Garcia, D.F., Fernandez, R.N. and Johannsen, C.J., 1991, Assessment of regional biomass–soil relationships using vegetation indexes. *IEEE Transactions on Geoscience and Remote Sensing*, **29**, pp. 331–339.

MacDonald, R.B. and Hall, F.G., 1980, Global crop forecasting. *Science*, **208**, pp. 670–679.

Modenese, L., Camuffo, M., Belluco, E., Marani, A. and Marani, M., 2005, Relations between biomass and vegetation indices of halophytic plants in Venice salt marshes. *Geophysical Research Abstracts*, **7**, 09448. Available online at: http://www.cosis.net/abstracts/EGUOS/09448/EGUOS-J-09448.pdf.

Myneni, R.B., Tucker, C.J., Asar, G. and Keeling, C.D., 1998, Interannual variations in satellite-sensed vegetation index data from 1981 to 1991. *Journal of Geophysical Research*, **103**, pp. 6145–6160.

NASS, 2009, History of remote sensing for crop acreage. Available online at: http://www.nass.usda.gov/surveys/Remotely_Sensed_Data_Crop_Acreage.

Peters, S.E., Wander, M.M., Saporito, L.S., Harris, G.H. and Friedman, D.B., 1997, Management impacts on SOM and related soil properties in a long-term farming

systems trial in Pennsylvania: 1981–1991. In *Soil Organic Matter in Temperate Agroecosystems*, E.A. Paul, K. Paustian, E.T. Elliott and C.V. Cole (Eds), pp. 183–196 (New York: CRC Press).

Pierce, F.J. and Fortin, M.C., 1997, Long-term tillage and periodic plowing of a no-tilled soil in Michigan: impacts, yield, and soil organic matter. In *Soil Organic Matter in Temperate Agroecosystems*, E.A. Paul, K. Paustian, E.T. Elliott and C.V. Cole (Eds), pp. 141–149 (New York: CRC Press).

Prince, S.D., Haskett, J., Steninger, M., Strand, H. and Wright, R., 2001, Net primary production of US Midwest croplands from agricultural harvest yield data. *Ecological Applications*, **11**, pp. 1194–1205.

Reichert, C. and Cassey, D., 2002, A reliable Crop Condition Assessment Program (CCAP) incorporating NOAA AVHRR data, a geographic information system and the internet. Available online at: http://gis.esri.com/library/userconf/proc02/pap0111/p0111.htm.

Russell, W.A., 1991, Genetic improvement of maize yields. *Advances in Agronomy*, **46**, pp. 245–298.

Senay, G.B., Lyon, J.G., Ward, A.D. and Nokes, S.E., 2000, Using high spatial resolution multispectral data to classify corn and soybean crops. *Photogrammetric Engineering and Remote Sensing*, **66**, pp. 319–327.

Shabanov, N.V., Liming, Z., Knyazikhin, Y., Myneni, R.B. and Tucker, C.J., 2002, Analysis of interannual changes in northern vegetation activity observed in AVHRR data from 1981 to 1994. *IEEE Transactions on Geoscience and Remote Sensing*, **40**, pp. 115–130.

Todd, S.W., Hoffer, R.M. and Milchunas, D.G., 1998, Biomass estimation on grazed and ungrazed rangelands using spectral indices. *International Journal of Remote Sensing*, **19**, pp. 427–438.

Tucker, C.J., Vanpraet, C., Boerwinkel, E. and Gaston, A, 1983, Satellite remote sensing of total dry matter production in the Senegalese Sahel. *Remote Sensing of Environment*, **13**, pp. 461–474.

Tucker, C.J., Vanpraet, C.L., Sharman, M.J. and Van Ittersum, G., 1985, Satellite remote sensing of total herbaceous biomass production in the Senegalese Sahel: 1980–1984. *Remote Sensing of Environment*, **17**, pp. 233–249.

Vanotti, M.B., Bundy, L.G. and Peterson, A.E., 1997, Nitrogen fertilizer and legume-cereal rotation effects on soil production and organic matter dynamics in Wisconsin. In *Soil Organic Matter in Temperate Agroecosystems*, E.A. Paul, K. Paustian, E.T. Elliott and C.V. Cole (Eds), pp. 105–119 (New York: CRC Press).

Wych, R.D. and Stuthman, D.D., 1983, Genetic improvement in Minnesota-adapted oat cultivars released since 1923. *Crop Science*, **23**, pp. 879–881.

Xiao, J. and Moody, A., 2005, Geographical distribution of global greening trends and their climatic correlates: 1982–1998. *International Journal of Remote Sensing*, **26**, pp. 2371–2390.

Yang, C., Everitt, J.H., Bradford, J.M. and Escobar, D.E., 2000, Mapping grain sorghum growth and yield variations using airborne multispectral digital imagery. *Transactions of the ASAE (American Society of Agricultural Engineers)*, **43**, pp. 1927–1938.