Published June 24, 2014

The Use of Reflectance Data for In-Season Soybean Yield Prediction

Spyridon Mourtzinis,* Scott C. Rowntree, Justin J. Suhre, Nicholas H. Weidenbenner, Eric W. Wilson, Vince M. Davis, Seth L. Naeve, Shaun N. Casteel, Brian W. Diers, Paul D. Esker, James E. Specht, and Shawn P. Conley

ABSTRACT

Estimation of soybean [*Glycine max* (L.) Merr.] yield early in the growing season is an appealing idea for both, farmers and soybean-related industries. Prior attempts to predict soybean yield have had limited success, especially when using information early in the growing season. The objective of this study was to evaluate the release date and maturity group (MG) of the cultivar, digital imaging, reflectance, and weather data during successive stages of crop development as explanatory variables in a soybean yield prediction model. The data were collected in the North Central (NC) United States at Arlington, WI (2010–2011), and Lafayette, IN (2011), using 59 MG II cultivars (released 1928–2008) at Wisconsin, and 57 MG III cultivars (released 1923–2007) at Indiana that were planted in performance trials on two planting dates (May and June). A second order polynomial regression analysis followed by ridge regression was used to develop the soybean yield prediction equation. The model accounted for 80% of the yield variability in the NC U.S. data set. An additional dataset not used in the calibration was used to conduct a validation test of the predictive performance of the model. The average difference between the fitted and actual yields in the validation test was 67 kg ha^{-1}. Results from this study suggest that the use of cultivar release year, planting date, MG, near-infrared (NIR), visible red (RED), and Red-edge wavelength bands recorded at 77 d after planting, and weather data 30 d before and after the planting date can closely estimate soybean yields in the Midwest.

Soybean is the major crop in the United States after corn (*Zea mays* L.). It accounts for 90% of the country's oilseed production. Approximately 31.2 million hectares were planted in the United States in 2012, of which more than 25.1 million hectares were planted in the upper Midwest (USDA ERS, 2013). Harvesting these large crop areas every year results in millions of kilograms of soybean grain in a short period of time (late September and October). The types of industries that use these large quantities of soybean range from food production for human consumption, feed production for animals, as well as biodiesel plants. Therefore, a more precise estimation of the crop yield early in the growing season can benefit both farmers and the associated industries. Furthermore, a method that can accurately predict simultaneously the final yield of multiple soybean cultivars early in the growing season can be of great help for scientists with their yield trials and breeding programs. Farmers could also benefit from an early season estimation of the final yield since they could contract their product at more competitive prices.

The normalized difference vegetative index (NDVI) is an index calculated from crop reflectance measurements in the RED and NIR spectrum region (Rouse et al., 1974). Studies have shown that the cellular structure as well as the air-cell wall-protoplasm-chloroplast interfaces can be determined by the NIR (Kumar and Silva, 1973). Additionally, these cell characteristics are affected by the environment and the nutrient status of the plant (Gausman et al., 1969; Thomas et al., 1971; Blackmer et al., 1994). Several studies have shown that the photosynthetic capability of the plant can be estimated by red reflectance measurements (Benedict and Swidler, 1961; Thomas and Oerther, 1972; Filella et al., 1995).

There have been attempts to predict yields of various crops using the NDVI, but with variable results. The NDVI for growing degree days explained 73% of corn grain variability and 77% of biomass yield variability at V8 (eight-leaf) growth stage (Teal et al., 2006). However, in another study the combined use of NDVI with corn height accounted for only up to 43% of the yield variability (Freeman et al., 2007). In a cotton yield study, the relationship of yield with NDVI was linear and the coefficient of determination reached 0.70 (Mkhabela and Mkhabela, 2000). In a study where NDVI was used to predict soybean grain yield using measurements made at the

S. Mourtzinis, S.C. Rowntree, V.M. Davis, and S.P. Conley, Dep. of Agronomy, Univ. of Wisconsin, Madison, WI 53706; J.J. Suhre and B.W. Diers, Dep. of Crop Sciences, Univ. of Illinois, Urbana, IL 61801; N.H. Weidenbenner and S.L. Naeve, Dep. of Agronomy, Univ. of Minnesota, St. Paul, MN 55108; E.W. Willson and S.N. Casteel, Dep. of Agronomy, Purdue Univ., West Lafayette, IN 47907; P.D. Esker, Escuela de Agronomia, Universidad de Costa Rica, San Jose, Costa Rica; and J.E. Specht, Dep. of Agronomy, Univ. of Nebraska, Lincoln, NE 68583. Received 15 Dec. 2013. *Corresponding author (mourtzinis@wisc.edu).

Published in Agron. J. 106:1159–1168 (2014)
doi:10.2134/agronj13.0577

Abbreviations: MG, maturity group(s), MSE, mean squared error; NC, North Central; NDVI, normalized difference vegetative index; NIR, near-infrared; RED, visible red; VIF, variance inflation factor.

fifth reproductive stage of the plant (R5–begin seed-fill), the maximum explained variability during the first year of the experiment was a respective 65 and 80% for plants grown on North Gower clay loam or Granby sandy loam soil sites (Ma et al., 2001). In the second year of the same experiment, the maximum R^2 values were a respective 0.70 and 0.45 for plants grown on North Gower clay loam and Upland sandy loam soil sites. The authors observed that the association of soybean yield and NDVI can be affected by the growing conditions. Cold and dry weather before planting can delay emergence and potentially decrease yields, while optimal conditions can result in improved seed germination and quicker canopy coverage to capture solar radiation. Given the low to moderate R^2 values for the foregoing yield prediction models (regression equations), more research is necessary to improve the amount of explainable yield variability when using reflectance data.

The Crop Circle ACS 470 (Holland Scientific Inc., Lincoln, NE) is a sensing device that has the ability to generate a wide range of reflectance data due to its ability to measure six of the available wavelength bands, though only three can be used at one time (Cao et al., 2013). Apart from the NIR and RED band wavelengths, the Crop Circle instrument can calculate the Red-edge band values, and result in the calculation of multiple Red-edge indices. A recent study identified the superiority of a model using the normalized difference red-edge index (NDRE), which exhibited a 28% increase in R^2 value, when compared to a model using indices calculated based on the red bands (Eitel et al., 2010).

Another promising tool in nondestructive data acquisition process is the use of digital images. Computer software, such as Sigma Scan Pro 5.0 (Systat Software Inc., San Jose, CA) has the ability to process the qualitative information in digital images into quantitative numerical data. The idea behind the use of digital images in crop yield prediction is to correlate the measureable visible characteristics of crop canopy and the amount of biomass that covers the soil surface at some prior crop phenological time point with the crop yield measured at harvest. In a grapevine yield prediction study, Dunn and Martin (2008) found that the use of digital image data explained 85% of the yield variability. Nonetheless, similar studies in soybean are nonexistent.

An attempt to predict plot-size soybean yield using reflectance data have had limited success (Ma et al., 2001). Equations developed for individual years and/or individual soil types might perform well in a similar very specific environment; however, their use in larger regions with multiple soil types and wide range of environmental conditions would require extrapolation beyond the calibration data, and therefore would not be recommended. One must use data acquired from multi-location multi-year cultivar trials for the development and calibration of a model that is expected to predict crop yield with accuracy and precision. Other desirable features include being able to use the model well before the end of the crop growing season, and its applicability to soybean production systems that span both spatial and temporal conditions encountered by producers. This would result in a robust model with significant practical use from the farmers, scientists, and the industry. Therefore, the objective of this study was to evaluate the use of digital imaging, reflectance, and nominally available weather data collected at successive in-season crop development time points as explanatory variables in a model that can accurately estimate end-of-season soybean grain yield.

MATERIALS AND METHODS

Data were collected in the 2010 and 2011 cropping seasons at Arlington, WI, and in the 2011 season at Lafayette, IN. Location-specific information and soil characteristics for the two sites can be found in Table 1. In both years of the experiment, soybean followed corn harvested for grain at the Indiana location, while soybean followed corn harvested for silage at the Wisconsin location. All locations were fall-chiseled, and prepared in the spring with field cultivation. Fertility and pest management at each location was performed according to university management recommendations. At each of the three location-year sites, cultivars were seeded at two planting dates, with 1 May and 1 June chosen as the desired target dates. The 1 May planting date (early) was selected to represent planting

Table 1. Experimental details with respect to test sites, soils, and dates of planting and harvest.

Location	Arlington, WI		Lafayette, IN
Research site	Arlington Agricultural Research Station		Throckmorton Purdue Agricultural Center
	43°18′ N, 89°20′ W		40°17′ N, 86°54′ W
Soil Series	Plano silt loam		Throckmorton silt loam
Soil Family	fine-silty, mixed, superactive, mesic Typic Argiudoll		fine-silty, mixed, superactive, mesic mollic Oxyaquic Hapludalf
Soil fertility			
Phosphorus, mg kg^{-1}	44–56		39–66
Potassium, mg kg^{-1}	166–173		138–146
pH	6.9–7.1		6.0–6.1
Organic matter, g kg^{-1}	3.2		2.9–3.0
Field operations	2010	2011	2011
Planting date (May PD Treatment)	4 May	5 May	17 May
Planting date (June PD Treatment)	1 June	6 June	12 June
Harvest date (May PD Treatment)	8 Oct.	17 Oct.	11 Oct.
Harvest date (June PD Treatment)	13 Oct.	17 Oct.	11 Oct.
Planting date difference (days)	28	32	26

dates growers use currently, while the 1 June (late) planting was selected to represent planting dates used more commonly in the past (USDA-NASS, 2011). In both years, weather and soil moisture conditions resulted in planting occurring somewhat later than the target dates, though a 26- to 32-d differential in planting date was still achieved (Table 1).

At the Wisconsin location, 59 MG II soybean cultivars released over eight decades, from 1928 to 2008, were planted, whereas at the Indiana location, 57 MG III soybean cultivars released from 1923 to 2007 were planted. The cultivars used in the experiment, along with the plant introduction (PI) number and pedigree information, are listed in Table 2. Each cultivar used in the experiment was unique, novel, or widely-grown during the time period of introduction. Cultivars included plant introductions grown about 80 yr ago, along with public and proprietary cultivars derived from further cycles of selection and breeding since then. The soybean seed used for the experiment originated from public and private seed sources, with seed increases of all cultivars occurring in prior (i.e., 2009 and 2010) growing seasons. Seed of the MG II cultivars was increased at the University of Nebraska-Lincoln (Lincoln, NE), whereas seed of the MG III cultivars was increased at the University of Illinois at Urbana-Champaign (Urbana, IL). To provide an estimate of experimental error, 13 MG II cultivars and 15 MG III cultivars were replicated twice within each planting date of each MG at the three location-year sites. The limited number of replicated cultivars was due to limited seed supply and field space constraints, but was selected on the basis of uniformly spanning the 80-yr distribution of cultivar release years. The experiment was replicated by environment, defined as location within year, for each maturity group.

The 76-cm spaced four-row plots were mechanically seeded at a rate of 370,650 seeds ha^{-1}. Planted plot dimensions at all locations were 3.1 m wide by 4.6 m long. Post-emergent plant populations were recorded for all plots at the V1 (first trifoliolate) and then again at the R8 (95% pod maturity) stages, as defined by Fehr and Caviness (1977). The center two rows of each plot were mechanically harvested a few days after R8. Grain weight and seed moisture data were collected simultaneously at harvest so that seed yield could be expressed on a 130 g kg^{-1} seed moisture content basis.

The Crop Circle ACS 470 (Holland Scientific Inc., Lincoln, NE) was used to derive three fixed wavelength bands from every plot weekly after emergence for 17 successive weeks, until approximately 120 d after planting. To collect the data, a researcher would walk along the plot holding the sensor at 1.5 m above the crop canopy. The three wavelength bands were the RED band (670 nm), NIR (780nm), and Red-edge (730 nm). Digital images of a representative area of every plot were collected weekly at the same time with the Crop Circle data using a Nikon D3200 (Nikon Chiyoda Tokyo, Japan). The digital images were processed using the Sigma Scan software Pro 5.0 (Systat Software Inc., San Jose, CA) and the area that was covered by the soybean canopy was converted to continuous data (percent of soil coverage by the soybean canopy). No wavelength data was collected by the digital images. The exact method that the digital images were collected and converted to continuous data has been described by Purcell (2000). The three wavelength bands (RED, NIR, and Red-edge) and calculated vegetation indices, the area that was covered by the soybean canopy, and the precipitation sums and mean daily temperatures pre-plant for three 30-d periods: (i) just before planting, (ii) just after planting, and (iii) starting on the 31st day after planting were evaluated separately and together as potential independent variables during the regression model development. The vegetation index regression parameters are shown in Table 3, and were partially adopted by Gong et al. (2003). Since multiple soybean cultivars, released from 1923 to 2008, were used in this study, the release year was included as independent variable to account for the upward yield trend arising from genetic improvement (Rowntree et al. (2013) for this North Carolina data set, but see also Rincker et al. (2014) for genetic gain estimates for MG II and III at a greater number of locations). The soybean yield data was subjected to a second order polynomial regression analysis using the digital images and reflectance data at different growth stages of the plant. The stepwise selection technique in the REG procedure in SAS Version 9.3 (SAS Institute Inc., Cary, NC) was used due to the large number of variables that needed to be evaluated as potential independent variables. The statistical properties of the final model were evaluated and compared to other candidate models according to several statistical criteria such as the mean squared error, coefficient of variation, predicted residual sum of squares and Mallows' Cp (Mourtzinis et al., 2013). The inclusion of weather data as ancillary variables resulted in high multicolinearity. The highly collinear weather variables resulted in model of less than full rank. Therefore, ridge regression was used to mitigate multicolinearity (Variance Inflation Factor [VIF] < 5) and obtain more stable coefficients for the selected independent variables (Hoerl and Kennard, 1970). The predictive ability of the final regression equation was tested and evaluated using an additional dataset of plot yields (n = 40) which was not used during the development and calibration of the final model.

RESULTS AND DISCUSSION

In regression analysis and model development in agricultural research, the use of multi-source data sets (i.e., years, locations, planting dates, etc.) for model calibration is desirable because it ensures that the environmental variability that might be encountered when one uses the model will reflect the predictive boundaries of the model. In this study, there was significant environmental variation among the 3 site-years (Table 4). In 2010, at the Wisconsin location, the combination of increased precipitation and temperatures led to record soybean yields for the state (Rowntree et al., 2013). However, in 2011, the in-season precipitation was significantly lower than normal (30-yr average). At the Indiana location, the monthly average temperatures were higher than normal while the monthly cumulative precipitation was significantly lower than the 30-yr average. Because of this wide range of in-season weather conditions, the data were pooled for the subsequent regression analysis.

Based on results of Rowntree et al. (2013) and Rincker et al. (2014) that demonstrated that the release year of the soybean cultivar was a significant source of yield variability (i.e., upward yield trend), it was imperative then that the development of an in-season yield prediction model first examined the effect of the genetic factor in the model. When the release year of the cultivar was used to serve as a function of the genetic improvement

Table 2. List of cultivars, year of release, maturity group, plant introduction (PI) number, and pedigree (if available).

Cultivar	Year of release	Maturity group	PI no.†	Pedigree‡	Cultivar	Year of release	Maturity group	PI no.‡	Pedigree§
Korean§	1928	II	PI548360	From China	Dunfield§	1923	III	PI548318	PI 36846 (NE China)
Mukden§	1932	II	PI548391	PI 50523 (NE China)	Illini§	1927	III	PI548348	Sel. from A.K. in 1920
Richland§	1938	II	PI548406	PI 70502-2 (NE China)	AK (Harrow)§	1928	III	PI548298	Sel. from A.K. (by 1928)
Hawkeye§	1947	II	PI548577	Mukden × Richland	Mandell	1934	III	PI548381	Sel. from Manchu in 1926
Harosoy§	1951	II	PI548573	Mandarin (Ottawa)(2) × A.K. (Harrow)	Mingo	1940	III	PI548388	Sel. from Manchu in 1924
Lindarin	1958	II	PI548589	Mandarin (Ottawa) × Lincoln	Lincoln§	1943	III	PI548362	Mandarin × Manchu
Harosoy 63	1963	II	PI548575	Harosoy (8) × Blackhawk	Adams	1948	III	PI548502	Illini × Dunfield
Hawkeye 63	1963	II	PI548578	Hawkeye (7) × Blackhawk	Shelby	1958	III	PI548574	Lincoln (2) × Richland
Amsoy	1965	II	PI548506	Adams × Harosoy	Ford	1958	III	PI548562	Lincoln (2) × Richland
Corsoy§	1967	II	PI548540	Harosoy × Capital	Ross	1960	III	PI548612	Monroe × Lincoln
Beeson	1968	II	PI548510	C1253 (Blackhawk × Harosoy) × Kent	Wayne§	1964	III	PI548628	L49-4091 × Clark
Amsoy 71§	1970	II	PI548507	Amsoy (8) × C1253	Adelphia	1964	III	PI548503	C1070 × Adams
Wells	1972	II	PI548630	C1266R (Harosoy × C1079) × C1253	Calland§	1968	III	PI548527	C1253 × Kent
Harcor	1975	II	PI548570	Corsoy × × OX383 (Corsoy × Harosoy 63)	Williams§	1971	III	PI548631	Wayne × L57-0034 (Clark × Adams)
Private 2-7	1977	II	na	na	Woodworth§	1974	III	PI548632	Wayne × L57-0034
Private 2-8	1977	II	na	na	Private 3-1§	1978	III	na	na
Wells II	1978	II	PI548513	Wells (8) × Arksoy	Cumberland	1978	III	PI548542	Corsoy × Williams
Vickery	1978	II	PI548617	Corsoy (5) x (L65-1342 and Anoka × Mack)	Oakland	1978	III	PI548543	L66L-137 (Wayne × L57-0034) × Calland
Corsoy 79	1979	II	PI518669	Corsoy (6) × Lee 68	Pella	1979	III	PI548523	L66L-137 × Calland
Beeson 80	1979	II	PI548511	Beeson (8) × Arksoy	Williams 82§	1981	III	PI518671	Williams (7) × Kingwa
Century§	1979	II	PI548512	Calland × Bonus	Private 3-15	1983	III	na	na
Amcor	1979	II	PI548505	Amsoy 71 × Corsoy	Zane	1984	III	PI548634	Cumberland × Pella
Private 2-11	1982	II	na	na	Harper	1984	III	PI548558	F4 sel. from an unknown diallel-cross pop.
Century 84	1984	II	PI548529	Century (5) × Williams 82	Chamberlain§	1986	III	PI548635	A76-304020 × Land O Lakes Max
Elgin	1984	II	PI548557	F4 selection from AP6 population	Private 3-2	1986	III	na	na
Preston	1985	II	PI548520	Schechinger S48 × Land O' Lakes Max	Resnik	1987	III	PI534645	Asgrow A3127(4) × L24
Private 2-15	1985	II	na	na	Pella 86	1987	III	PI509044	From backcross of Pella(5) × Williams 82

Continued next page

Table 2 (continued).

Cultivar	Year of release	Maturity group	PI no.†	Pedigree‡	Cultivar	Year of release	Maturity group	PI no.‡	Pedigree§
Burlison	1988	II	PI533655	F4 selection from K74-113-76-486 × Century	Private 3-9	1989	III	na	na
Private 2-9	1988	II	na	na	Private 3-10	1990	III	na	na
Elgin 87	1988	II	PI518666	Elgin (5) × Williams 82	Private 3-16	1991	III	na	na
Conrad§	1988	II	PI525453	A3127 × Tri-Valley Charger	Dunbar	1992	III	PI552538	Platte × A3127
Jack§	1989	II	PI540556	Fayette × Hardin	Thorne	1992	III	PI564718	A80-344003 × A3127BC3F2-1
Kenwood	1989	II	PI537094	Elgin × A1937	Private 3-17	1992	III	na	na
Private 2-1	1989	II	na	na	Private 3-18	1993	III	na	na
Private 2-2	1990	II	na	na	Private 3-19	1994	III	na	na
RCAT Angora	1991	II	PI572242	B152 × T8112	Macon§	1995	III	PI593258	Sherman × Resnik
Private 2-6	1991	II	na	na	IA 3004	1995	III	na	Northrup King S23-03 × A86-301024
Private 2-5	1993	II	na	na	Maverick	1996	III	PI598124	LN86-4668 (Fayette × Hardin) × Resnik(3)
Private 2-10	1994	II	na	na	Private 3- 4	1996	III	na	na
Private 2-16	1994	II	na	na	Private 3-11	1996	III	na	na
IA 2021	1995	II	na	Elgin 87 × Marcus	Pana	1997	III	PI597387	Jack × Asgrow A3205
Savoy	1996	II	PI597381	Burlison × Asgrow A3733	Private 3- 5	1997	III	na	na
Private 2-12	1996	II	na	na	Private 3-12	1997	III	na	na
Dwight§	1997	II	PI597386	Jack × A86-303014	Private 3-6	1998	III	na	na
Private 2-18	1997	II	na	na	IA 3010	1998	III	na	Jaques J285 × Northrup King S29–39
IA 2038	1998	II	na	Pioneer 9301 × Kenwood	Private 3-7§	1999	III	na	na
IA 2050	2000	II	na	Northrup King S24-92 × A91-501002	Private 3-20	2000	III	na	na
IA 2052	2000	II	na	Northrup King S24-92 × Parker	U98-311442	2001	III	na	A94-773014 × Bell
Loda§	2001	II	PI614088	Jack × IA 3003	IA 3014	2001	III	na	LN90-4366 × IA3005
Private 2-4	2001	II	na	na	Private 3- 8§	2002	III	na	na
Private 2-17	2001	II	na	na	IA 3023	2003	III	na	Dairyland DSR-365 × Pioneer P9381
IA 2068	2003	II	na	AgriPro P1953 × LN94-10470	NE3001	2004	III	na	Colfax × A91-701035
Private 2-3	2004	II	na	na	Private 3-13§	2004	III	na	na
IA 2065	2005	II	na	na	IA 3024	2004	III	na	A97-553017 × Pioneer YB33A99
Private 2-19	2005	II	na	na	Private 3-22	2006	III	na	na
Private 2-20	2005	II	na	na	Private 3-23	2006	III	na	na
IA 2094	2006	II	na	AgriPro X0121B74 × A00-711036	Private 3-14	2007	III	na	na
Private 2-13	2008	II	na	na					
Private 2-14§	2008	II	na	na					

† na, not applicable.
‡ na, not available.
§ Cultivars replicated within location.

Table 3. Summary of the reflectance bands and vegetation indexes used for the model development.

Vegetation Index/ Wavelength band	Formula	Name
NIR	ρ_{NIR}	Near Infra-red band
RE	ρ_{RE}	Red-edge band
RED	ρ_{RED}	Red band
SRI	$\frac{\rho_{NIR}}{\rho_{RED}}$	Simple ratio index-Red
Chl I	$\frac{\rho_{NIR}}{\rho_{RE}}-1$	Chlorophyll index-Red-edge
NDVI	$\frac{\rho_{NIR}-\rho_{RED}}{\rho_{NIR}+\rho_{RED}}$	Normalized difference vegetation index
NDRE	$\frac{\rho_{NIR}-\rho_{RE}}{\rho_{NIR}+\rho_{RE}}$	Normalized difference Red-edge
NLI	$\frac{\rho_{NIR}^2-\rho_{RED}}{\rho_{NIR}^2+\rho_{RED}}$	Non-linear vegetation index
RDVI	$\frac{\rho_{NIR}-\rho_{RED}}{(\rho_{NIR}+\rho_{RED})^{0.5}}$	Re-normalized difference vegetation index
MSR	$\frac{\left(\frac{\rho_{NIR}}{\rho_{RED}}\right)-1}{\left(\frac{\rho_{NIR}}{\rho_{RED}}\right)^{0.5}+1}$	Modified simple ratio
NDVI × SRI	$\frac{\rho_{NIR}^2-\rho_{RED}}{\rho_{NIR}+\rho_{RED}^2}$	.

trend, the regression was significant and the R^2 for just this independent variable in a simple model was almost 0.25. Based on this result, release year was included as explanatory variable in all of the subsequent stages of regression analysis.

Reflectance Data Variable Evaluation

Models that included the release year of the cultivar and digital image-derived data were developed for each of the 7-d interval during the growing season (Fig. 1). Apart from the interval 70 d after planting (R3-beginning pod to R5-beginning seed), when maximum R^2 and minimum mean squared error (MSE) were observed ($R^2 = 0.356$, MSE = 618.76), these two parameters exhibited minimal fluctuation. However, when using stepwise regression selection with respect to the release date of the cultivar and the Crop Circle-derived data, including the calculated vegetative indices, the coefficient of determination and MSE exhibited inverse pairs of peaks and valleys (i.e., nadirs) at two timeframes during the growing season (Fig. 2). The R^2 peaks and corresponding MSE nadirs were observed at 35 and then at 77 d after planting. At 35 d (V3-fully expanded first trifoliolate leaflets at main stem node two to V5-third trifoliolate leaflets at main stem node four), R^2 attained peak of 0.63 whereas MSE dipped to a nadir 470.4. At 77 d (R4-full pod to R5.5-beans filling half the space in a pod at upper four main stem nodes), R^2 and MSE respectively reached a peak of 0.65 and a nadir of 455.9. These R^2 values of 0.63 and 0.65 for our model are comparable to most of the coefficients of determination reported by Ma et al. (2001). In their study, models were developed for individual site-years rather than using pooled data to develop a single model addressing the spatial (e.g., soil type) and temporal (e.g., in-season weather conditions) variability over all site-years.

The information in Fig. 1 and 2 indicate the superiority of Crop Circle-derived data in explaining soybean yield variability when compared to digital images during the growing season. Furthermore, the data suggest that between 28 and 91 d after planting (i.e., respectively V2 and R6) the use of Crop Circle reflectance data at 35 and 77 d has the highest potential in predicting final soybean yield.

Prediction Model Development

When the goal is to develop a yield prediction model that can be used for practical purposes by farmers, scientists, and the industry, the larger the amount of variability the greater likelihood of a more accurate yield prediction. In the previous section, the addition of Crop Circle data at 35 and 77 d after planting greatly improved the model's ability to predict in-season soybean yield. Still, users of a yield prediction model with a R^2 value of 0.65 might not consider this to be sufficient. To increase the R^2 value and reduce MSE, we evaluated the inclusion of more variables into the model.

A maturity group variable and a planting date variable (early or late) were considered to be significant sources of variation and were next included in the model development process for evaluation as ancillary variables. The MG variable was categorical (i.e., 2 or 3). In the North Carolina data sets, six different planting dates were used (Table 1). Irrespective of MG (i.e, 2 or 3) in Wisconsin and Indiana, the recommended and most common planting date is 1 May (Rowntree et al., 2013). The dates

Table 4. Mean monthly air temperature and total monthly precipitation at Arlington, WI, and Lafayette, IN, during the 2010 and 2011 growing seasons, and during the past 30 yr.

	Arlington, WI			Lafayette, IN	
Temperature and precipitation	2010	2011	30 yr	2011	30 yr
Air temperature, °C					
April	10.4	6.2	7.1	11.6	10.7
May	15.3	13.4	13.2	17.1	16.6
June	19.7	19.6	18.7	22.6	21.8
July	22.9	24	20.8	26	23.4
August	22.2	21	19.6	22.7	22.4
September	15.6	14.5	15.2	17.1	18.8
Precipitation, mm					
April	107.5	106.4	88.9	192.6	86.6
May	88.9	55.4	93.7	113.4	117.9
June	169.4	98.8	118.9	92.8	115.6
July	222.8	64.3	105.7	45.5	103.6
August	114	39.9	99.1	26.3	100.1
September	50.5	96.5	89.9	82.8	71.2

in early to mid-May were considered the early planting while the dates in early to mid-June were considered the late planting dates. Treating the planting date variable as categorical with just two levels (early = 1 or late = 2) would result in two separate prediction models. Therefore, the six planting dates were quantified (for data input purposes) on the basis of falling into one of a successive set of nine 5-d intervals beginning with 1 May and ending on 14 June. For example, the planting dates 4 and 5 May would fall in the same interval (i.e., the first 5-d interval), while the 17 May planting day would fall in the fourth interval, and the latest date in 12 June would fall in the last interval (ninth 5-d interval).

The in-season weather information was also included in the form of six independent variables in the model development process. These six variables consisted of precipitation sum and daily temperature average during (i) the 30 d before planting, (ii) the 30 d after planting, and (iii) the 30 d after the 31st day after planting. The new models which included inputs for MG, planting date, six weather variables, and the Crop Circle derived-data exhibited significantly improved statistical properties at both, 35 and 77 d after planting, and not much was gained in either case by adding the digital imaging data (Table 4). For the model at 35 d after planting, after performing ridge regression to mitigate the detected multicollinearity issues, the R^2 reached almost 0.79. For the model at 77 d after planting, the R^2 reached 0.80. Additionally, at 77 d after planting, the root-MSE and CV were greatly reduced when compared to the models at 35 d after planting further indicating improved model fit. In a final attempt to improve the fit of the regression equation, the digital image data were included in the model as additional independent variables. However, the improvement in the amount of explained variability, as well as the reduction of root-MSE and CV, was minimal (Table 5) so these data were left out of the final model. The

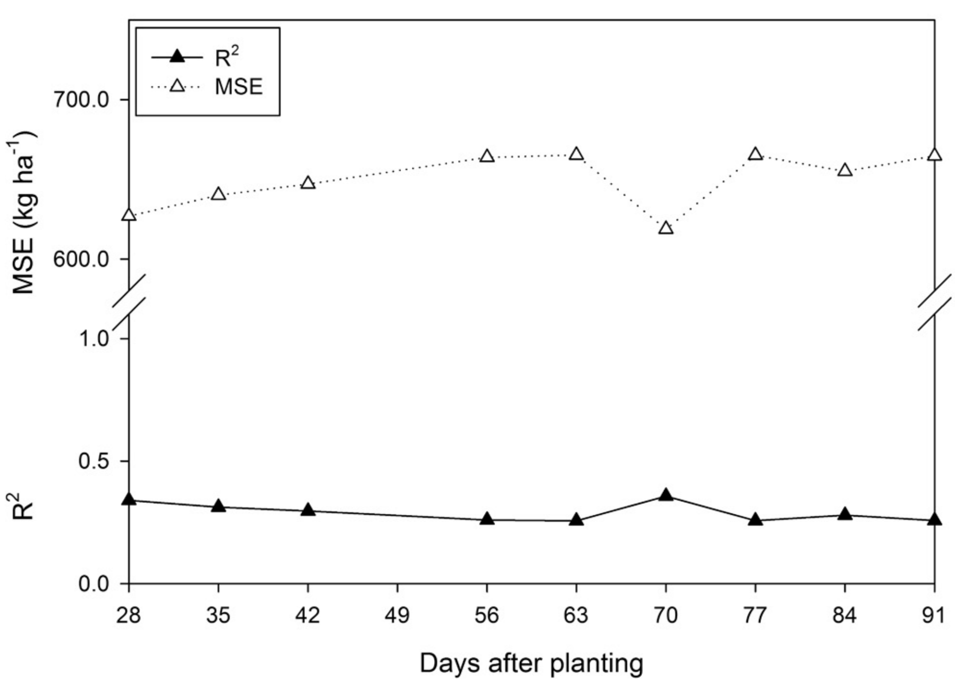

Fig. 1. Coefficient of determination (R^2) (black triangles) and mean squared error (MSE) (white triangles) for models using the release year of soybean cultivars (irrespective of maturity group) and digital image-derived data collected at 7-d intervals during a 28- to 91-d period after planting.

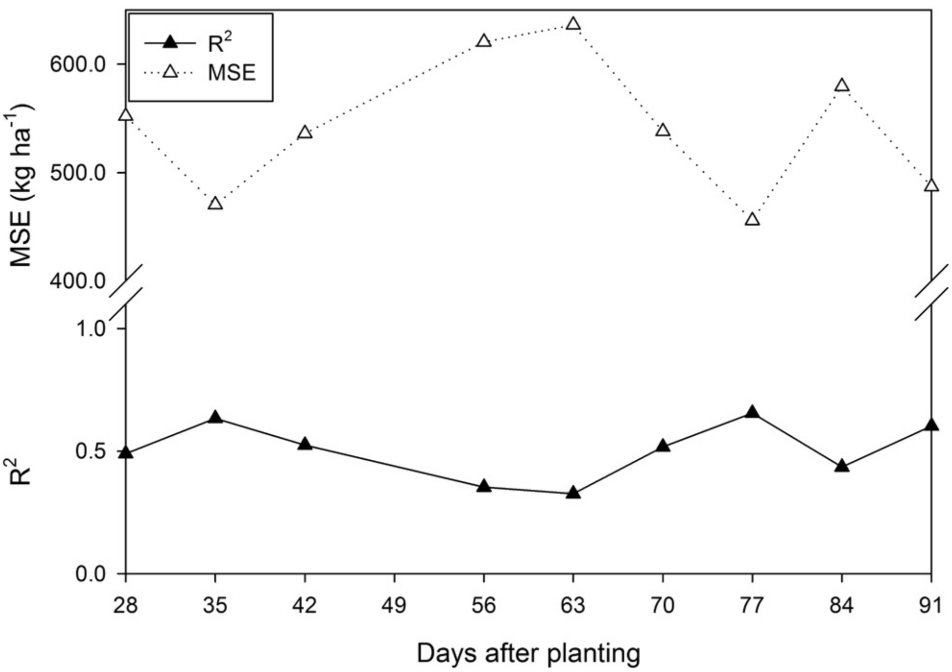

Fig. 2. Coefficient of determination (R^2) (black triangles) and mean squared error (MSE) (white triangles) for models using the release year of soybean cultivars (irrespective of maturity group) and Crop Circle-derived data collected at 7-d intervals during a 28- to 91-d period after planting.

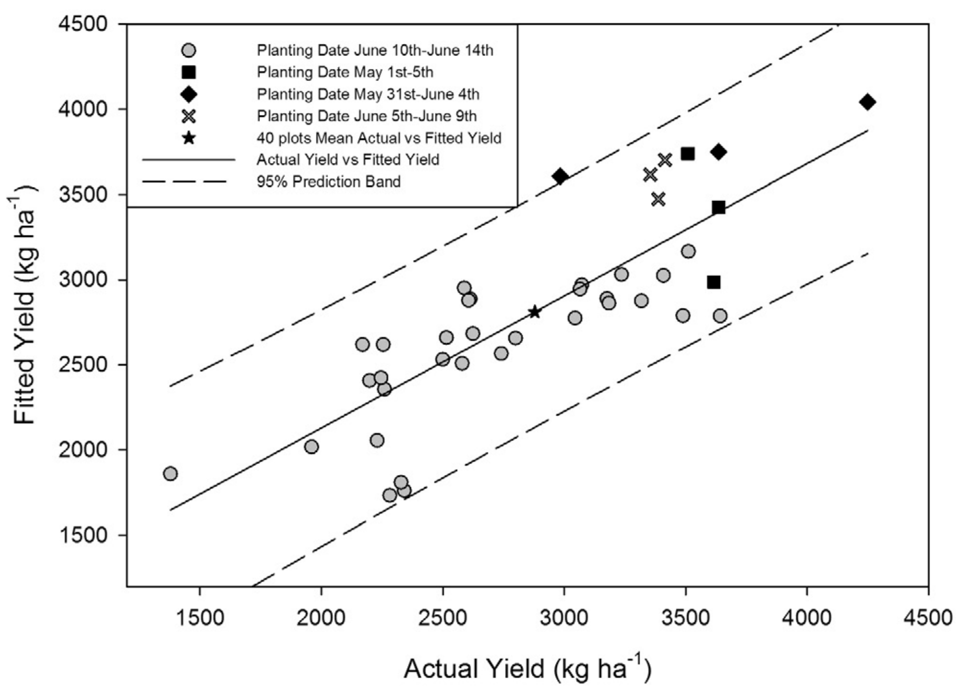

Fig. 3. Comparison of actual vs. fitted soybean yields in the 40 plot validation set.

Table 5. Statistical criteria for the evaluation of four soybean yield prediction models using maturity group (MG), planting date (PD), digital image data (DI), weather (W), and Crop Circle data (CC) collected at 35 and 77 d after planting.

Data used	Days after planting	R^2	Adjusted-R^2	Root mean squared error	Coefficient of variation
MG+PD+W+CC	35	0.7912	0.7876	371.74	11.85
MG+PD+W+CC	77	0.8004	0.7974	346.49	11.04
MG+PD+W+CC+DI	35	0.7918	0.7878	354.59	11.30
MG+PD+W+CC+DI	77	0.8040	0.8007	343.66	10.96

Table 6. Comparison of the fitted and actual soybean yields in the validation data set for constant and proportional differences.

Model	Variable	Coefficient	95% confidence limits
35 d	Intercept	593.03	35.608, 1150.455
	Slope	0.819	0.623, 1.015
77 d	Intercept	436.17	−132.653, 1005.008
	Slope	0.868	0.670, 1.067

Table 7. Comparison of the mean fitted and mean actual soybean yields in the validation set (n = 40).

Variable	Mean	SD
Actual yield, kg ha^{-1}	2878	605
Fitted yield, kg ha^{-1}	2811	572
Actual-Fitted, kg ha^{-1}	67	353
Difference, %	2.3	13

less complicated equation is preferred and therefore, the models with digital image data were not used.

Between the two models of using inputs associated with a 35-d (after planting) model, or inputs associated with a 77-d (after planting) model, the latter model exhibited improved R^2 and a significant reduction in MSE which indicate superiority in the fit of the data. The improvement in the statistical properties of the 77-d model are the result of including the two weather variables, cumulative precipitation and average temperature 30 d after the 31st day after planting. A model using information at 35 d after planting would be associated with a higher risk of inaccurate yield estimation since it omits much of the weather variability in the growing season which could greatly impact the final yield. Nevertheless, due to the similarity of the statistical criteria between the two regression equations, further evaluation was performed to identify the best model for soybean yield prediction.

Large R^2 values with similar adjusted R^2 along with small MSE and CV are desirable features of a regression equation and indicate a model with a good fit of the data. Nevertheless, the most appropriate way to test the predictive performance of a regression equation calibrated with an existing data set is to use that model on a validation data set. Therefore, a set of 40 plots randomly selected from the entire study, which were not used during the model development, were chosen to serve as the validation set. The actual yields for these 40 plots were regressed on the yields predicted by the 35- and the 77-d model (fitted values) to assess the degree (if any) and range of over- and underprediction by the model (Table 6). For the 35-d model, the 0.819 slope of the fitted and actual data and confidence limits that did include unity, indicate no statistically significant proportional over- or underestimation of the 40 actual soybean yields in the validation data set. However, the confidence limits of the intercept did not include zero (0), which indicated statistically significant constant overestimation of the actual soybean yields. For the 77-d model, the confidence limits of the intercept did include zero (0), which indicated no statistically significant constant over- or underestimation of the actual soybean yields. In addition, 1:1 plot of fitted and actual data exhibited a slope value of 0.868, and slope confidence limits that did include unity, thereby no statistically significant proportional over- or underestimation of the 40 actual soybean yields in the validation data set. Therefore, the model at 77 d after planting was chosen as the final soybean yield prediction equation. Figure 3 shows the scatter plot of fitted yields against actual values in the validation set. From the 40 observations, 39 for early and late planting dates fall inside the 95% prediction limits (dashed lines) with one falling on the upper 95% prediction limit. Furthermore, a direct comparison between the mean actual and mean fitted yields showed a 2.3% difference (Table 7). This translates in 67 kg ha^{-1} difference, which is of low practical importance.

The final model including the independent variables, associated coefficients and probability values is summarized in Table 8. From all the Crop Circle reflectance variables, only the interaction between the RED and NIR bands and the second power of the Red-edge band were significant in the model. Since in the developed model, both RED and Red-edge bands contributed significantly towards the soybean yield, this result is partially in agreement with Eitel et al. (2010) who suggested a possible superiority of the Red-edge band and Red-edge-based indices when compared to indices calculated based on the RED bands. Additionally, the 1 mo pre- and post-planting cumulative precipitation variables were significant in the model which indicates the degree to which rainfall and temperature impact the optimization of planting date and early crop development, and highlights the importance of including weather data in the prediction equation development. Finally, the small VIF values (<5) indicated that multicollinearity was not an issue (Montgomery et al., 2006).

Every prediction model has limitations and boundaries. The boundaries of the model also define its robustness. The developed regression equation should perform well (interpolate) within the range of observed weather conditions and soil types encountered at the 3 site-years. The addition of cultivar release year as a variable serves to adjust the yield prediction for the continual genetic yield gain that breeders have documented (Rincker et al., 2014). Moreover, the addition of a planting date variable serves to adjust the yield prediction for this

Table 8. Final model for soybean yield prediction (kg ha^{-1}) using information collected at 77 d after planting.

Model variable	Coefficient	P > |t|	Variance inflation factor
Intercept	−17821	<0.0001	.
Cultivar release year	13.42880	<0.0001	2.35
Cultivar maturity group	−770.98480	<0.0001	3.58
Planting date interval	−32.62526	<0.0001	2.72
NIR × RED†	−2.62890	0.0668	3.36
Red-edge^2	−9.33106	<0.0001	2.98
Precipitation sum 30 d before planting	−2.38992	0.0042	2.63
Precipitation sum 30 d after planting	3.95394	<0.0001	1.21

† NIR, near infra-red; RED, visible red.

management practice that agronomists have documented as a key factor influencing yield (Rowntree et al., 2013).

CONCLUSIONS

Reflectance-data acquisition methods have been widely used in agricultural research over the past years from the development of NDVI for crop yield prediction to the use of NIR spectroscopy for biomass compositional analysis. Results from this study suggest that the use of the Red-edge band can improve the fit of a soybean yield prediction model. However, the use of in-season reflectance data alone did not provide adequate information for accurate soybean yield prediction and the inclusion of ancillary variables was necessary for the development of a model with improved statistical criteria and acceptable predictive performance.

The results from this study revealed that a model whose inputs are confined to cultivar release year, cultivar maturity group, a planting date defined in terms of one of nine 5-d intervals of sowing delay after 1 May, Crop Circle-derived reflectance data obtained only at 77 d after planting, and a cumulative precipitation sum 30 d before and 30 d after planting can closely estimate the final soybean grain yield in the Wisconsin and Indiana locations of this study. Inclusion of additional sites and years of data, especially with soybean grown in conditions outside of the specified boundaries in this study, can further increase the range of the predictive ability of the model.

ACKNOWLEDGMENTS

The authors would like to thank the field personnel at the University of Wisconsin, University of Illinois, and Purdue University for their technical assistance and cooperation in this research. This research was partially funded by the Wisconsin Soybean Marketing Board, Illinois Soybean Program Operating Board, Indiana Soybean Alliance, and Pioneer Hi-Bred International, Inc.

REFERENCES

Benedict, H.M., and R. Swidler. 1961. Nondestructive method for estimating chlorophyll content of leaves. Science (Washington, DC) 133:2015–2016. doi:10.1126/science.133.3469.2015

Blackmer, T.M., J.S. Schepers, and G.E. Varvel. 1994. Light reflectance compared with other nitrogen stress measurements in corn leaves. Agron. J. 86:934–938. doi:10.2134/agronj1994.00021962008600060002x

Cao, Q., Y. Miao, S. Huang, H. Wang, R. Khosla, and R. Jiang. 2013. Estimating rice nitrogen status with the Crop Circle multispectral active canopy sensor. Precision agriculture. Wageningen Academic Publ.,Wageningen, the Netherlands.

Dunn, G.M., and S.R. Martin. 2008. Yield prediction from digital image analysis: A technique with potential for vineyard assessments prior to harvest. Aust. J. Grape Wine Res. 10:196–198. doi:10.1111/j.1755-0238.2004.tb00022.x

Eitel, J.U.H., R.F. Keefe, D.S. Long, A.S. Davis, and L.A. Virieling. 2010. Active ground optical remote sensing for improved monitoring of seedling stress in nurseries. Sensors (Basel Switzerland) 10:2843–2850. doi:10.3390/s100402843

Fehr, W.R., and C.E. Caviness. 1977. Stages of soybean development. Spec. Rep. 80. Coop. Ext. Serv., Agric. and Home Econ, Exp.Stn. Iowa State Univ., Ames.

Filella, I., L. Serrano, J. Serra, and J. Penluelas. 1995. Evaluating wheat nitrogen status with canopy reflectance indices and discriminant analysis. Crop Sci. 35:1400–1405. doi:10.2135/cropsci1995.0011183X003500050023x

Freeman, K.W., K. Girma, D.B. Arnall, R.W. Mullen, K.L. Martin, R.K. Teal, and W.R. Raun. 2007. By-plant prediction of corn forage biomass and nitrogen uptake at various growth stages using remote sensing and plant height. Agron. J. 99:530–536. doi:10.2134/agronj2006.0135

Gausman, H.W., W.A. Allen, V.I. Myers, and R. Cardenas. 1969. Reflectance and internal structure of cotton leaves, *Gossypium hirsutum* (L.). Agron. J. 61:374–376. doi:10.2134/agronj1969.00021962006100030011x

Gong, P., R. Pu, G.S. Biging, and M.R. Larrieu. 2003. Estimation of forest leaf area index using vegetation indeces derived from hyperion hyperspectral data. IEEETrans. Geosci. Remote Sens. 41:1355–1362. doi:10.1109/TGRS.2003.812910

Hoerl, A.E., and R.W. Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12:55–67. doi:10.1080/00401706.1970.10488634

Kumar, R., and L. Silva. 1973. Light ray tracing through a leaf cross section. Appl. Opt. 12:2950–2954. doi:10.1364/AO.12.002950

Ma, B.L., L.M. Dwyer, C. Costa, E.R. Cober, and M.J. Morrison. 2001. Early prediction of soybean yield from canopy reflectance measurements. Agron. J. 93:1227–1234. doi:10.2134/agronj2001.1227

Mkhabela, M.S., and M.S. Mkhabela. 2000. Exploring the possibilities of using NOAA, AVHRR data to forecast cotton yield in Swaziland. UNISWA J. Agric. 9:13–21.

Montgomery, D.C., E.A. Peck, and G.G. Vining. 2006. Introduction to Linear Regression Analysis. 4th ed. Wiley Series in Probability and Statistics, Hoboken, NJ.

Mourtzinis, S., F.J. Arriaga, K.S. Balkcom, and B.V. Ortiz. 2013. Corn grain and stover yield prediction at R1 growth stage. Agron. J. 105:1045–1050. doi:10.2134/agronj2012.0393

Purcell, L.C. 2000. Soybean canopy coverage and light interception measurements using digital imagery. Crop Sci. 40:834–837. doi:10.2135/cropsci2000.403834x

Rincker, K., R. Nelson, J. Specht, D. Sleper, T. Cary, S. Clanzio et al. 2014. Genetic improvement of US soybean in Maturity Groups II, III, and IV. Crop Sci. 54:in press. doi:10.2135/cropsci2013.10.0665

Rouse, J.W., R.H. Haas, J.A. Schell, and D.W. Deering. 1974. Monitoring vegetation systems in the Great Plains with ERTS. In: S.C. Freden and M.A. Becker, editors, Proceedings of the Third ERTS-1 Symposium. NASA SP-351, Greenbelt, MD. 10-14 Dec. 1973. p. 309–317.

Rowntree, S.C., J.L. Suhre, N.H. Weidenbenner, E.W. Wilson, V.M. Davis, S.L. Naeve et al. 2013. Genetic gain x management interactions in Soybean: I. Planting date. Crop Sci. 53:1128–1138. doi:10.2135/cropsci2012.03.0157

Teal, R.K., B. Tubana, K. Girma, K.W. Freeman, D.B. Arnall, O. Walsh, and W.R. Raun. 2006. In-season prediction of corn grain yield potential using normalized difference vegetation index. Agron. J. 98:1488–1494. doi:10.2134/agronj2006.0103

Thomas, J.R., L.N. Namken, G.F. Oerther, and R.G. Brown. 1971. Estimating leaf water content by reflectance measurements. Agron. J. 63:845–847. doi:10.2134/agronj1971.00021962006300060007x

Thomas, J.R., and G.F. Oerther. 1972. Estimating nitrogen content of sweet pepper leaves by reflectance measurements. Agron. J. 64:11–13. doi:10.2134/agronj1972.00021962006400010004x

USDA-ERS. 2013. Soybean and oil crops. USDA-ERS, Washington, DC. www.ers.usda.gov/topics/crops/soybeans-oil-crops/background (accessed 3 Dec. 2013).

USDA-NASS. 2011. Quick Stats 2.0. USDA-NASS, Washington, DC. www.nass.usda.gov/quickstats/(accessed 18 Oct. 2013).