2003 BELTWIDE COTTON CONFERENCES, NASHVILLE, TN – JANUARY 6-10

MULTITEMPORAL REMOTE SENSING FOR PREDICTING COTTON YIELD

J. Alex Thomasson, James R. Wooten, Swapna Gogineni, and Ruixiu Sui
Agricultural and Biological Engineering Department
Mississippi State University
Mississippi State, MS

Abstract

Remotely sensed images of a Mississippi cotton field were collected with high spatial and temporal resolution during the growing season. The image data were overlaid in a GIS (geographic information system) with historical spatial data from the field: topography, soil texture, and historical cotton yield. The combination of all these data was studied to determine relationships with yield data collected at the end of the season. The motivation was to develop the ability to predict yield, well in advance of harvest, on a spatially variable basis. Such an ability would be an excellent new farm-management tool, allowing producers to better understand, in a spatially variable context, the monetary risks and returns involved in applying costly inputs such as pesticides, fertilizers, *etc*. Statistical analyses were conducted at grid-cell sizes from 10 m square (100 m^2) to 100 m square (10,000 m^2) in 10-m increments. The relationships at each cell size were calculated with data available at the beginning of the season, at the first image date, at the second image date, and so on until the last image date. Stepwise linear regression was used as the procedure for selecting variables at each date that would make up the most appropriate model to predict yield. Results indicated that the accuracy of the models at most dates was highest at the 100-m cell size. Remotely sensed data apparently added a great amount of information to the models, with most of that information being provided in the first 2 months after harvest. The highest R^2 value produced by any model was 0.92, and the prediction error associated with that model was on the order of 150 lbs/ac, that in a field with yields varying from about 1500 to 2600 lbs/ac.

Introduction

Background
Precision agriculture seeks to optimize profits on a site-specific basis within farm fields by improving revenue and/or reducing costs. As crop growth progresses in a field, variations in the vigor of the crop appear from site to site. These variations are related to various plant stresses, which ultimately influence yield through their effects on plant growth. If variations that affect yield can be ascertained during the growing season via remote sensing, then it may be possible to predict yield with some reasonable level of accuracy. A map of predicted yield would be very advantageous to a farmer, allowing him to make site-specific management decisions with some knowledge of the monetary risks and returns associated with them.

Literature Review
Several researchers have attempted yield estimation with remote sensing data. According to Wiegand and Richardson (1990) and Wiegand et al. (1991, 1992), vegetation indices calculated from remote spectral observations are good yield predictors. Anderson and Yang (1996) reported strong correlation ($R^2 = 0.90$) between red band reflectance and yield of a grain sorghum crop. Wiegand et al. (1991) reported fairly strong correlation between satellite-based spectral vegetation indices and cotton boll counts ($R^2 = 0.76$). It is on this basis that cotton yield estimates have been attempted with remote sensing data. Thomasson et al. (2000) found a high correlation ($R^2 = 0.88$) between Landsat band 4 (infrared) digital numbers (DNs) and cotton yield monitor data averages at a given DN. Other significant relationships were discovered between yield and soil fertility and texture, and between Landsat band images and soil fertility and texture. Therefore, it appears reasonable that a combination of remote-sensing data and soil-property data could enable reasonable predictions of cotton yield, particularly in irrigated fields wherein soil-moisture content is relatively controlled. It is also expected that multi-temporal spectral reflectance data taken throughout the growing season will enhance the accuracy of crop yield estimates. In fact, Yang and Everett (2000) reported that the correlations between multi-temporal remote-sensing data and grain sorghum yield steadily increased from the beginning of the growing season to the end.

Objectives
It is apparent that remote-sensing data of a cotton crop contain information on spatial variability that pertains to yield. Sequentially collected images have added more and more such information for the prediction of sorghum yield, so it is conceivable that cotton yield could be predicted with a series of remotely sensed images collected during the growing season, with the aid of historical site-specific data. Thus, the objectives of this study were as follows:

- To develop techniques to "predict" yield with multiple remotely sensed images and other available site-specific data.
- To evaluate the techniques developed.

Materials and Methods

Study Site
A field of approximately 275 acres in a cooperating producer's operation near Vance, in the northern part of Mississippi's Delta region, was selected for study. This field has been in a rotation of cotton in even years with double-cropped soybeans and wheat in odd years. Most of the field is irrigated with a center-pivot system, while the northwestern corner is irrigated with furrow irrigation, and the northeastern corner is not irrigated.

Available Data
All harvesting was done with a John Deere four-row cotton picker, on which cotton yield monitors were mounted. Yield-map data were available for production years 1998 and 2000 (Figures 1 and 2). In 1998, yield data were collected with a Micro-Trak cotton yield monitor, while 2000 yield data were collected with a Zycom cotton yield monitor. Elevation data were also available (Figure 3), having been collected on a 50-ft by 50-ft grid with a laser-plane survey system in 1999. Furthermore, soil texture data (% clay and % sand) had been collected on a 1.0-acre grid in 1998 (Figures 4 and 5), and so these data were available as well.

Remotely Sensed Data Collected in 2002
Four-band multispectral images of the field were acquired from an aircraft equipped with the GeoVantage multispectral sensor package, at a ground-distance resolution of approximately 0.5 m, roughly every two weeks beginning on May 22, 2002. The four spectral bands had center wavelengths and bandwidths at half peak respectively as follows: band 1, 450 nm and 10 nm; band 2, 550 nm and 10 nm; band 3, 650 nm and 10 nm; and band 4, 850 nm and 20 nm. The GeoVantage system of hardware and software has the capability to automatically geo-register and mosaic images. Figures 6 through 14 are original color-infrared mosaic images based on the GeoVantange multispectral imagery at individual dates. No attempt was made to atmospherically correct the image data.

A few observations about the images are worth noting. In the May 22 image, it is apparent that the image was acquired in the middle of a tillage operation, in which half the field had been tilled and the other half had not. The tilled portion of the field had moist soil exposed to the sky, which appeared as the darker portion of the field in the image. In the August 9 image, it is apparent that the image was acquired during an irrigation event. The portion of the field just west of the center pivot system appears slightly darker, again because of higher moisture, this time in and/or on the plants. Finally, the mosaicing process resulted in dark regions where original scenes had been joined together. This was apparently caused indirectly by the low altitude from which images were acquired. Early in 2002, the data acquisition team of the Remote Sensing Technologies Center at Mississippi State University had decided to collect images at 0.5-m rather than 2-m resolution, which was the requested maximum ground-distance resolution. This decision was made largely because of the windier conditions at the higher altitude required for 2-m resolution. Flying at the lower altitude caused two conditions that became problems during image analysis. First, flying lower causes an image to be collected over a wider camera angle, which can cause vignetting (darkening) at image edges. Second, flying lower required the collection and mosaicing of many more images in order to construct one scene for the entire field. The dark regions in the mosaic scenes are most notable in the August 9 image, but they can be seen in most of the other images also.

Image Processing
Because of the aforementioned problems with image mosaics, new mosaics were created from the original smaller scenes. Each was modified so as to reduce vignetting effects, and then several attempts were made to mosaic the small scenes together in such a way as to minimize visible seams. This process was not completely successful in removing all seam-lines from the mosaics, but it provided apparent improvements over the mosaics that the system had provided automatically. Each mosaic image was then re-geo-corrected, and mosaic images from all nine dates were overlaid in Arcview GIS.

Data Fusion
Also in Arcview, the image mosaics were overlaid with historical data to produce three data sets: data set A included remote-sensing, historical-yield, soil-texture (% clay and % sand), and topographic (elevation and slope) data averaged over 10-m square areas (100 m^2) centered at each soil sample location (Figure 15); data set B included remote-sensing and historical-yield data averaged over areas ranging in size from 10 m square (100 m^2) to 100 m square (10,000 m^2) in 10-m increments; and data set C was developed in the same way as data set B, but the remote-sensing data were converted to NDVI (normalized difference vegetation index) in accordance with Equation 1, and difference in NDVI from one date to the next.

$$NDVI = (NIR-Red) / (NIR+Red) \quad (1)$$

Data set A comprised the 272 soil sample locations, so the number of data points was 272. Data sets B and C comprised a varying number of data points depending on the resolution used (Table 1). The three principal differences between data set A and data sets B and C were as follows: (1) data set A was a sampling of the field while data sets B and C covered the entire field; (2) data set A included only one set of 272 data points, while data sets B and C included ten subsets of varying numbers of data points; and (3) data set A could be expected to be more accurate with respect to soil texture because its data points coincided geographically with actual soil-texture measurements, whereas data sets B and C included only interpolated values between actual soil-sample locations, and the high level of soil-reflectance variability evident in images of the field indicated that interpolating provided considerably less accuracy. In creating these data sets, the number of original data points, mean, and standard deviation of yield and remote-sensing data were calculated for each cell in the field. For example, the number of image pixels in a 10-m square area was approximately 400, and for each 10-m cell the exact number was calculated along with mean and standard deviation of pixel values. Table 2 lists available data in each data set.

Data Analysis

Data analyses were conducted in such a way as to see how accurately cotton yield could be "predicted" at various points during the growing season with data that might be reasonably available to the producer (such as those available in this study: historical yield; soil texture, which typically does not change appreciably over time; topography, which can also be expected to remain constant; and periodic in-season remote-sensing data). Three analyses were performed to predict yield with the available data. In analysis 1, data set A was used so that as accurate data as possible would be available for a limited number of samples. In analysis 2, data set B was used so that similar data would be available for the entire field. In analysis 3, data set C was used so that an analysis could be done over the entire field with remote-sensing data (NDVI and related variables) that should be more repeatable than raw multispectral data. In each case, several predictions were made, the first with data available at the beginning of the season, the second with data available after the first remote-sensing flight, the third with data available after the second remote-sensing flight, and so on. The SAS procedure STEPWISE was used to select to select an appropriate multiple-linear-regression model for cotton yield based on available data.

Results

Analysis 1

When yield prediction was attempted with data set A, including only data available at the beginning of the growing season, a significant relationship (at the 5% level) was found, but the R^2 value was only 0.11. When the first remotely sensed image, which was of predominantly bare soil, was added to the list of possible regressors, the R^2 increased to 0.26. When all nine images were included in the list of possible regressors, the R^2 value increased to 0.48, with 19 variables in the model. Figures 17 and 18 depict the relatively steady increase in R^2 and number of regressors throughout the season. It is apparent in Figure 17 that the R^2 value tends toward a plateau at approximately 60 days after planting.

Analysis 2

When yield prediction was attempted with data set B, including only data available at the beginning of the growing season, a significant relationship was found, and the R^2 was 0.34 in a model that included four variables on an 80-m by 80-m grid. When the first remotely sensed image was added to the list of possible regressors, the R^2 value increased to 0.47, in a model that included seven variables on a 100-m by 100-m grid. When all nine images were included in the list of possible regressors, the R^2 value increased to 0.92, in a model that included 26 variables on a 100-m by 100-m grid. Figures 19 and 20 depict the relatively steady increase in R^2 and number of regressors throughout the season. It is apparent in Figure 19 that, as with data set A, the R^2 value tends toward a plateau at approximately 60 days after planting. Furthermore, Figure 21 depicts the predicted values and regression line for the best model (100-m resolution, all available data, $R^2 = 0.92$). It is apparent that this "focus model" explains most of the yield variability evident after yield was aggregated into 100-m by 100-m cells.

Analysis 3

When yield prediction was attempted with data set C, including only data available at the beginning of the growing season, a significant relationship was found, and the R^2 was 0.34 in a model that included eight variables on an 80-m by 80-m grid. When data from the first remotely sensed image were added to the list of possible regressors, the R^2 value increased to 0.43, in a model that included nine variables on an 80-m by 80-m grid. When data from all nine images were included in the list of possible regressors, the R^2 value increased to 0.72, in a model that included 15 variables on a 100-m by 100-m grid. Figures 22 and 23 depict the relatively steady increase in R^2 and number of regressors throughout the season. It is apparent in Figure 19 that, as with data sets A and B, the R^2 value tends toward a plateau, but with data set C it is at closer to 75 days after planting. While the "best" model had and R^2 value of 0.72, another model was found to be nearly as good ($R^2 = 0.68$), and it was better in two ways: (1) it required no data beyond 79 days after planting, which means its predictive power was maximized relatively early in the growing season; and (2) it required only 12 variables overall (elevation, clay content, mean 2000 yield, mean 1998 yield, NDVI from date 1, standard deviation of NDVI from date 2, difference in standard deviation of NDVI between dates 2 and 1, standard deviation of NDVI from date 3, difference in NDVI between dates 3 and 2, NDVI on date 4, the number of pixels at each cell on date 5, and difference in NDVI between dates 5 and 6). This group of variables is seen as being much more likely to be repeatable than raw remote-sensing values. Furthermore, Figure 24 depicts the predicted values and regression line for this data-set-C focus model, which explains most of the yield variability evident after yield was aggregated into 100-m by 100-m cells.

Conclusions

It is apparent from this study that remote sensing provides much information about cotton yield. It is also apparent that images taken throughout the growing season add more and more information about yield, but that a plateau tends to appear near the middle of the growing season. These results confirm previous findings in cotton and other crops. The addition of other reasonably accessible ground-based data, such as historical yield, soil texture, and topography, appear to have increased the predictability of cotton yield on a spatially variable basis. While it is important to develop techniques to make models broadly applicable (i.e., so that they work on other fields and during other growing seasons), the success of the NDVI-related models in this project give hope toward that end. That is because indices like NDVI are much more repeatable than raw remote-sensing measurements.

Acknowledgements

This project was funded by the Remote Sensing Technologies Center (RSTC) at Mississippi State University, which is sponsored by NASA.

References

Anderson, G. L., and C. Yang. 1996. Multispectral videography and geographic information systems for site-specific farm management. In: *Proc. 3rd Int. Conf. on Precision Agric.*, 681-692. Madison, WI: Agronomy Society of America.

Thomasson, J. A., J. Chen, J. R. Wooten, S. A. Shearer, and D. A. Pennington. 2000. Cotton yield prediction improvement with remote sensing. In: *Proc. Beltwide Cotton Conf.*, P. Dugger and D. Richter, eds., 419-421. Memphis, Tenn.: National Cotton Council of Am.

Wiegand, C. L., and A. J. Richardson. 1990. Use of spectral vegetation indices to infer leaf area, evapotranspiration and yield. II. Results. *Agron. J.* 82(3):630-636.

Wiegand, C. L., A. J. Richardson, D. E. Escobar, and A. H. Gerbermann. 1991. Vegetation indices in crop assessments. *Remote Sensing Environ.* 35(2/3):105-119.

Wiegand, C. L., J. H. Everitt, and A. J. Richardson. 1992. Comparison of multispectral video and SPOT-1 HRV observations for cotton affected by soil salinity. *Int. J. Remote Sens.* 13:1511-1525.

Yang, C., and J. H. Everitt. 2000. Relationships between yield monitor data and airborne multispectral digital imagery. In: *Proc. 5th Int. Conf. on Precision Agric.* Bloomington, Minn.: Univ. of Minnesota Precision Agric. Center.

Table 1. Number of data points for each ground-distance resolution in data sets B and C.

Ground-distance resolution (m)	**Number of data points**
10	9435
20	2354
30	1053
40	589
50	376
60	261
70	194
80	149
90	122
100	98

Table 2. Data included in each data set, along with cell sizes considered.

Data set	Independent variables	Dependent variable	Cell size
A	4-band multispectral images, 0.5-m resolution, 9 dates. Number of image pixels per cell, 9 dates. Standard deviation of pixel values, 4 bands, 9 dates. Mean cotton yield, 1998 and 2000. Number of yield data points per cell, 1998 and 2000. Standard deviation of cotton yield within cell, 1998 and 2000. % clay content. % sand content. Elevation. Slope.	Mean 2002 cotton yield.	10 m square (100 m^2)
B	4-band multispectral images, 0.5-m resolution, 9 dates. Number of image pixels per cell, 9 dates. Standard deviation of pixel values, 4 bands, 9 dates. Mean cotton yield, 1998 and 2000. Number of yield data points per cell, 1998 and 2000. Standard deviation of cotton yield within cell, 1998 and 2000. % clay content. % sand content. Elevation. Slope.	Mean 2002 cotton yield.	10 m square (100 m^2) to 100 m square (10,000 m^2)
C	NDVI, 0.5-m resolution, 9 dates. Number of image pixels per cell, 9 dates. Standard deviation of NDVI, 9 dates. Difference in NDVI from one date to next, 8 differences. Standard deviation of difference in NDVI, 8. Mean cotton yield, 1998 and 2000. Number of yield data points per cell, 1998 and 2000. Standard deviation of cotton yield within cell, 1998 and 2000. % clay content. % sand content. Elevation. Slope.	Mean 2002 cotton yield.	10 m square (100 m^2) to 100 m square (10,000 m^2)

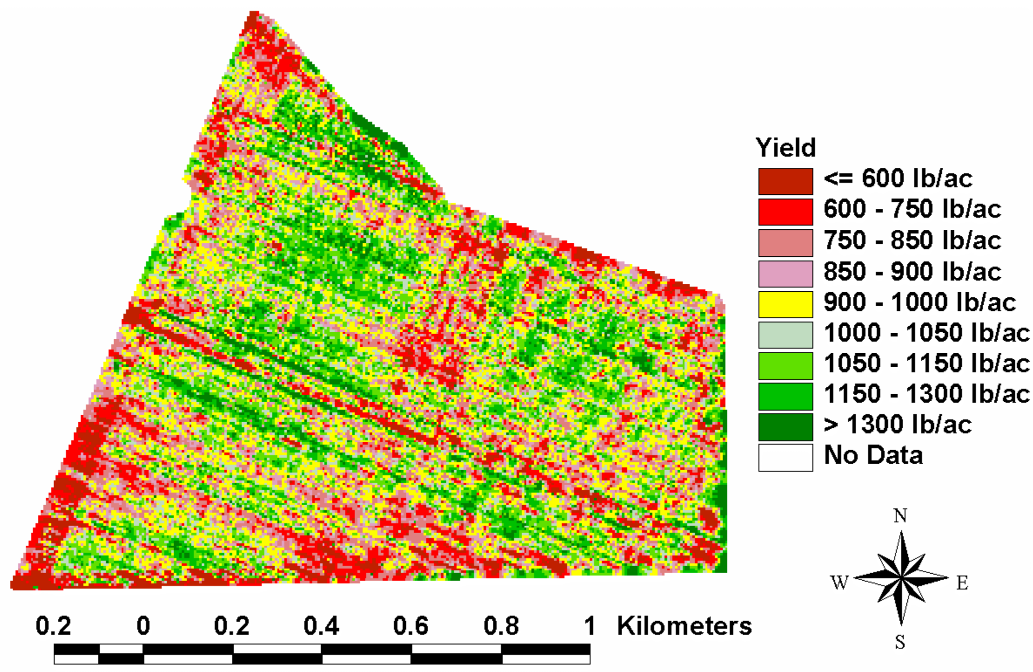

Figure 1. Yield map of 1998 cotton production at study site.

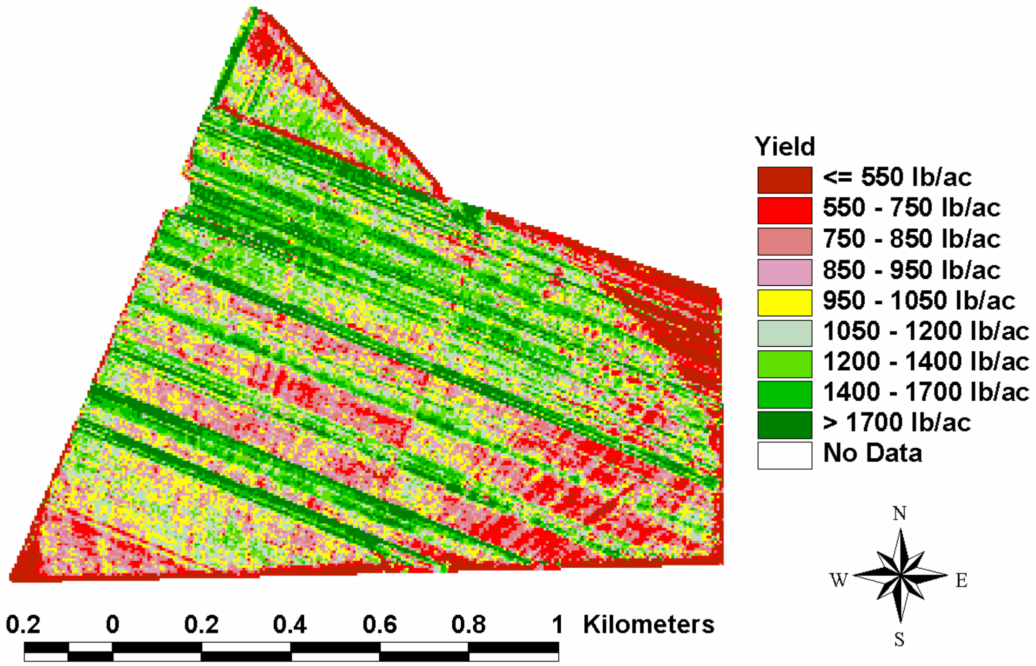

Figure 2. Yield map of 2000 cotton production at study site.

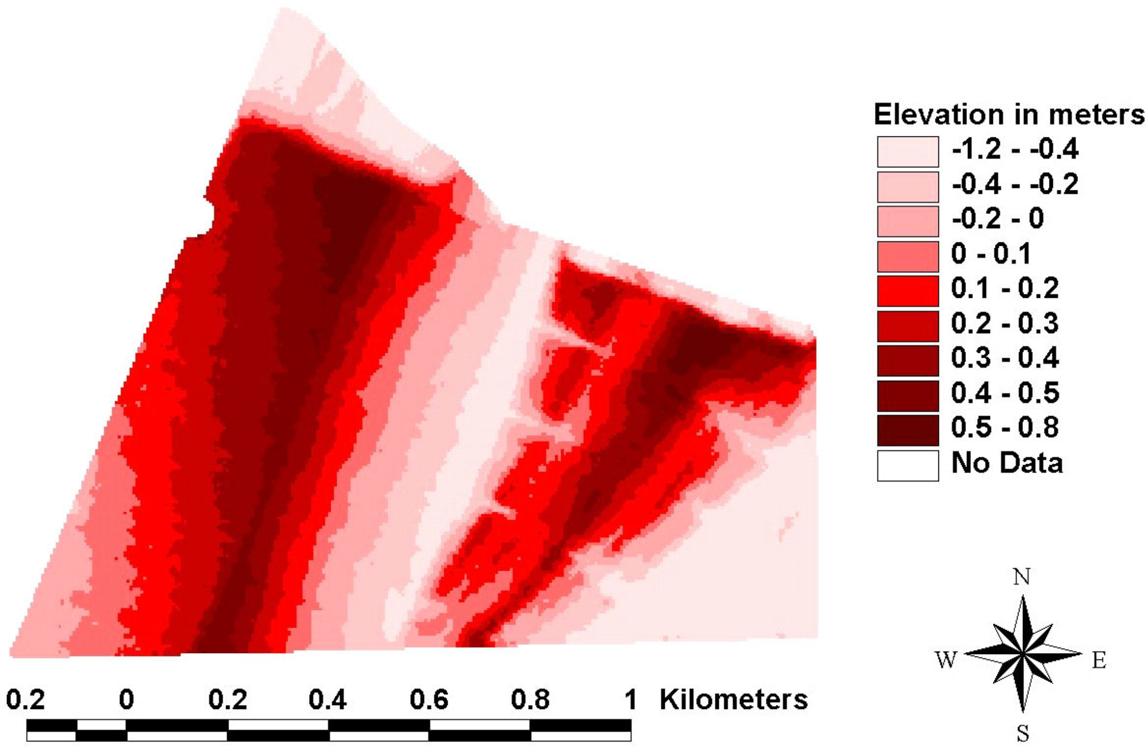

Figure 3. Elevation map of study site.

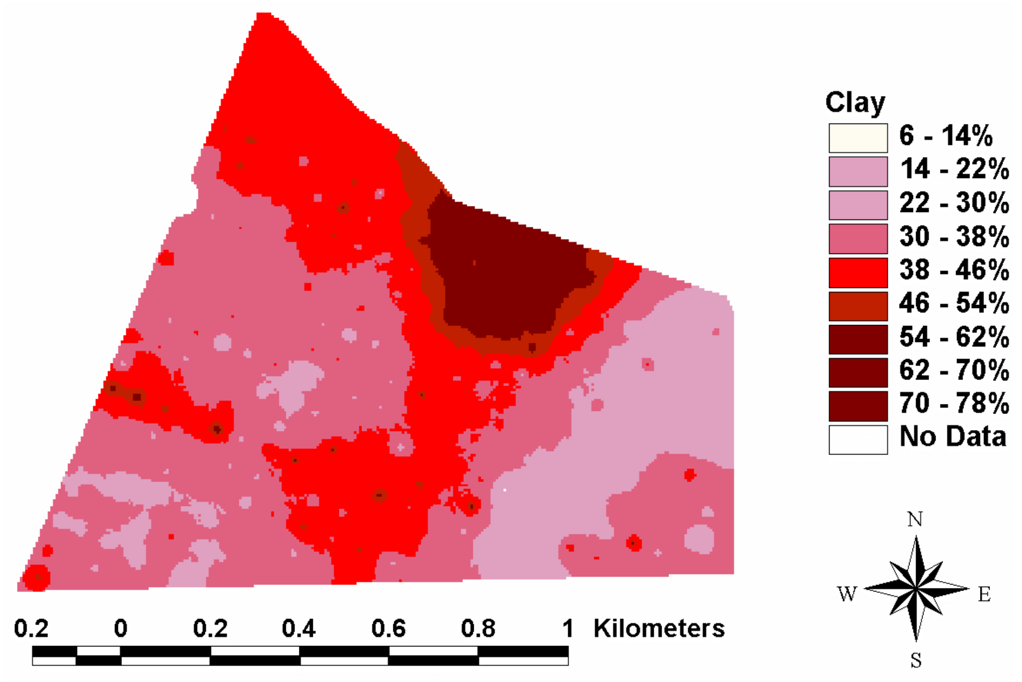

Figure 4. Soil clay content map of study site.

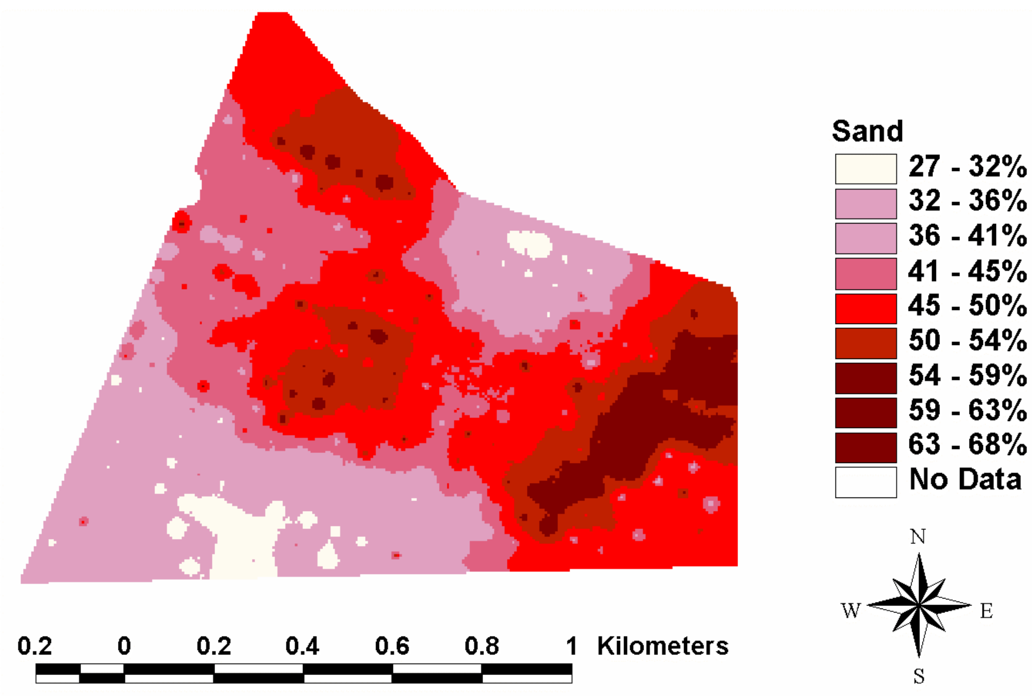

Figure 5. Soil sand content map of study site.

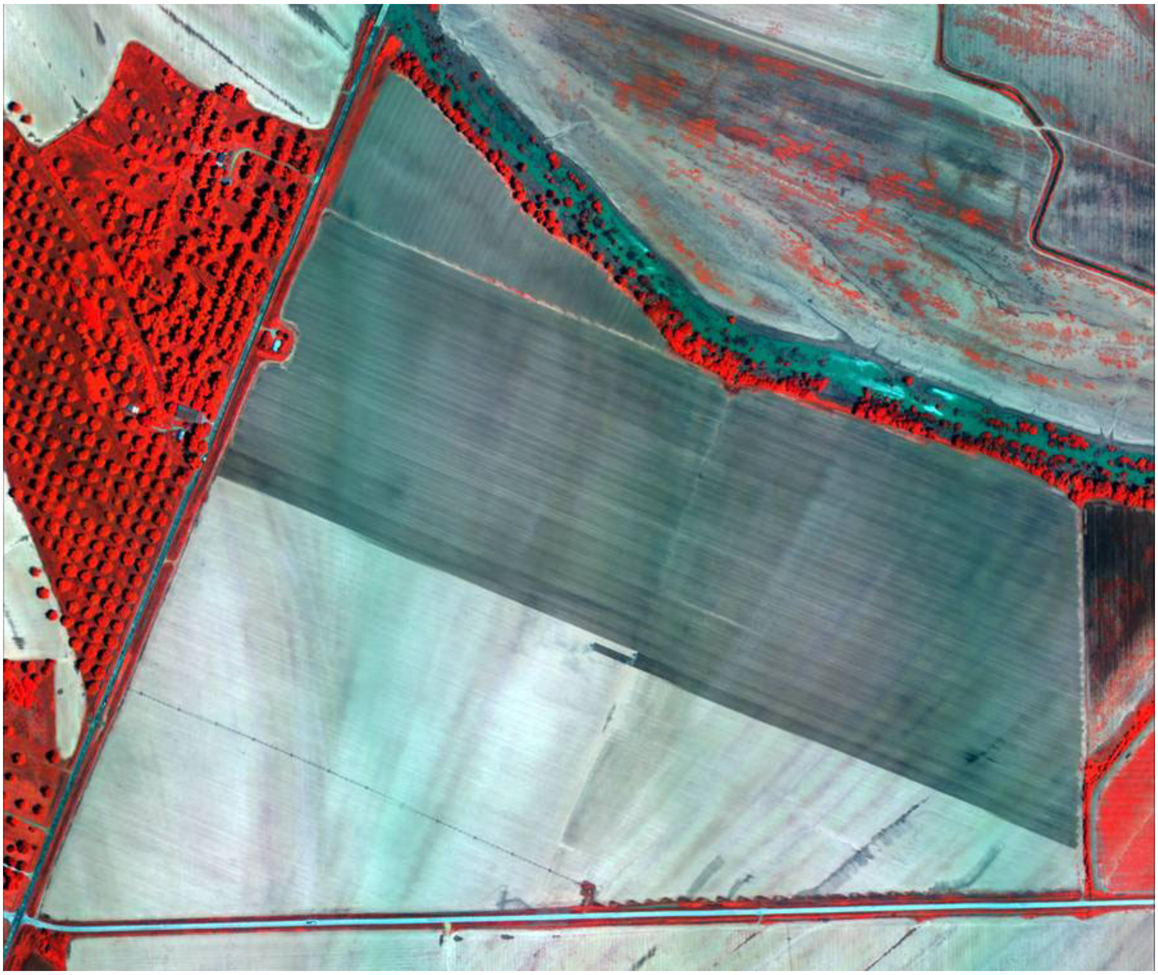

Figure 6. Color-infrared image of study site May 22, 2002, during tillage just prior to planting.

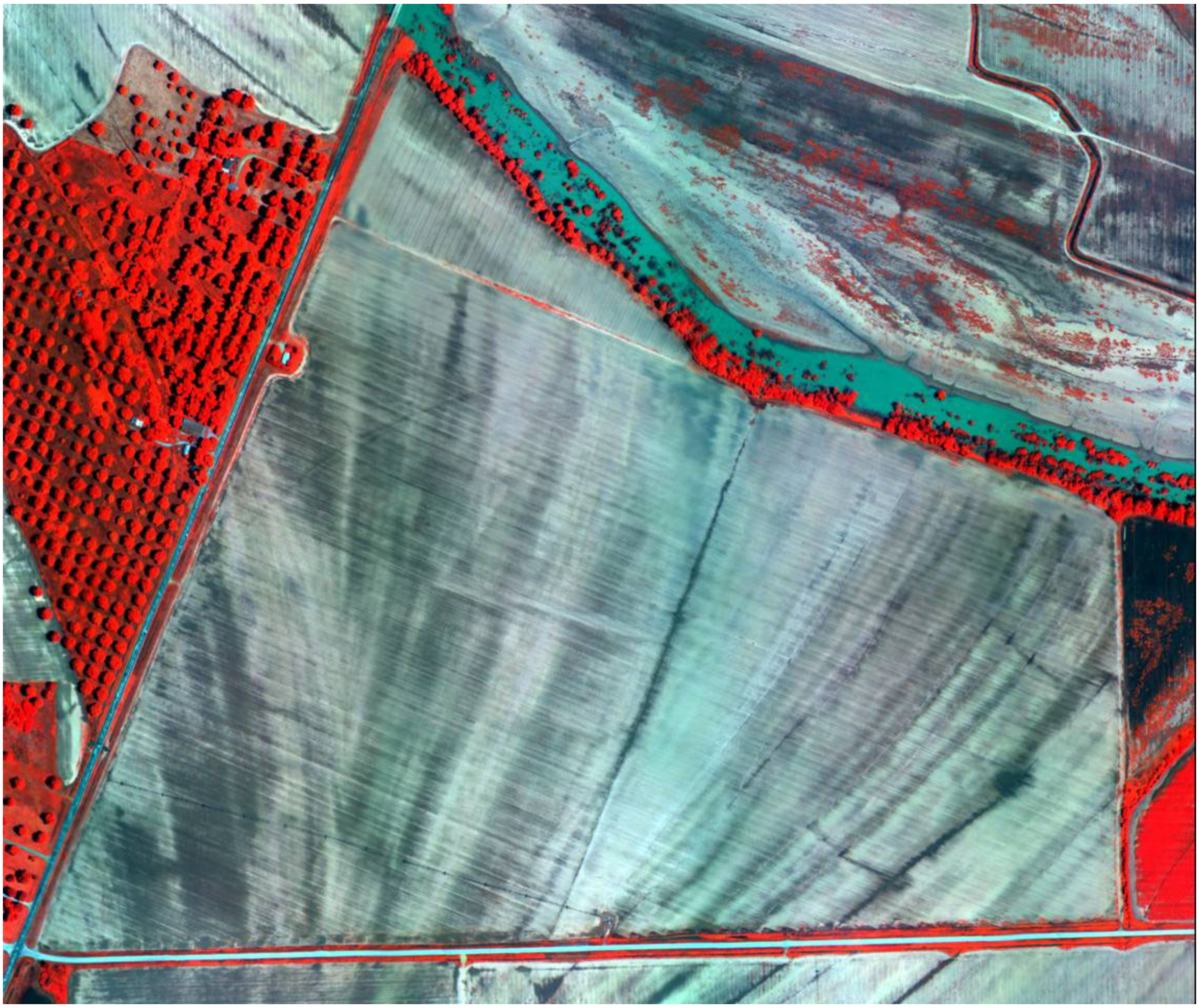

Figure 7. Color-infrared image of study site June 1, 2002, soon after planting.

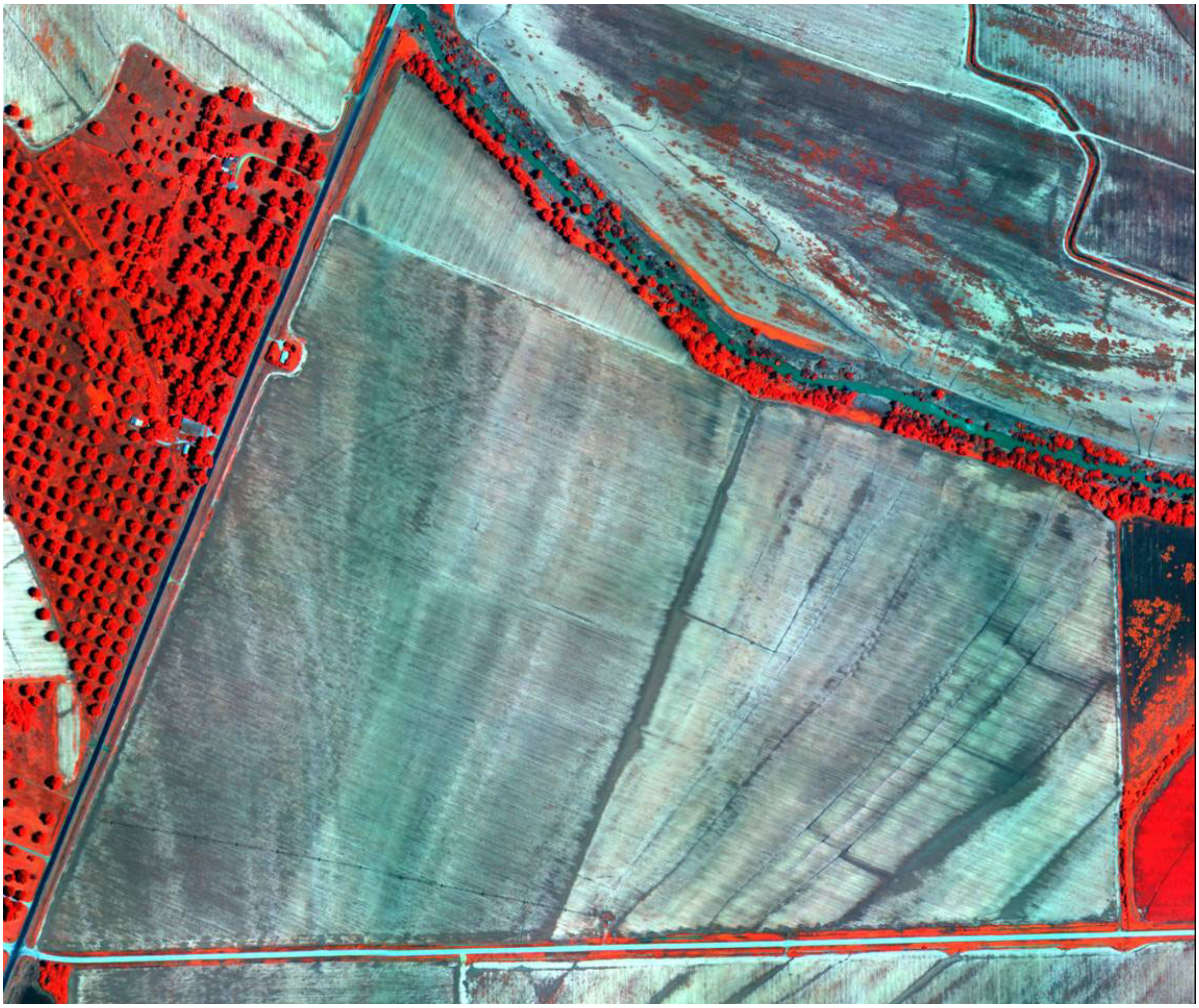

Figure 8. Color-infrared image of study site June 15, 2002.

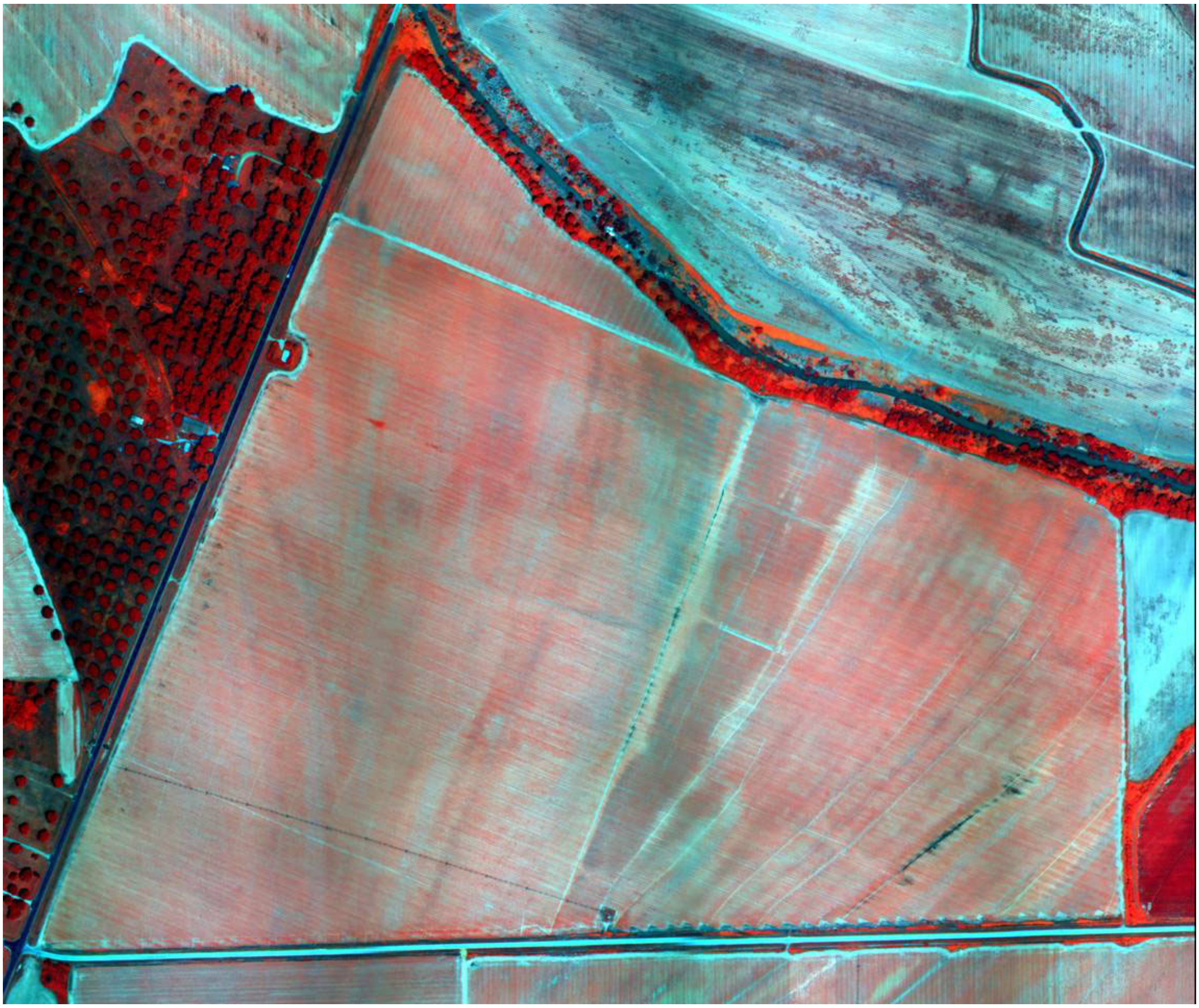

Figure 9. Color-infrared image of study site July 2, 2002.

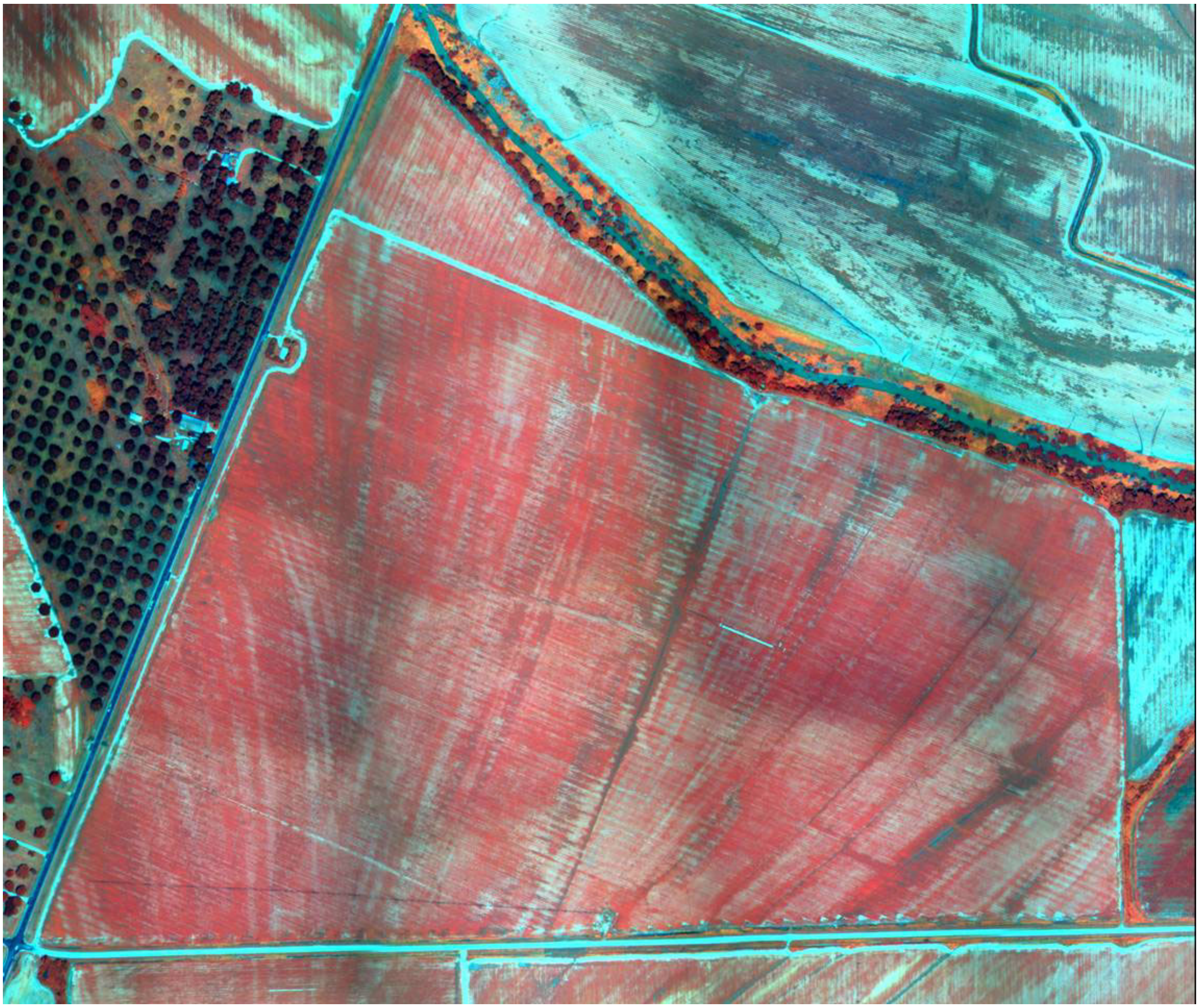

Figure 10. Color-infrared image of study site July 17, 2002.

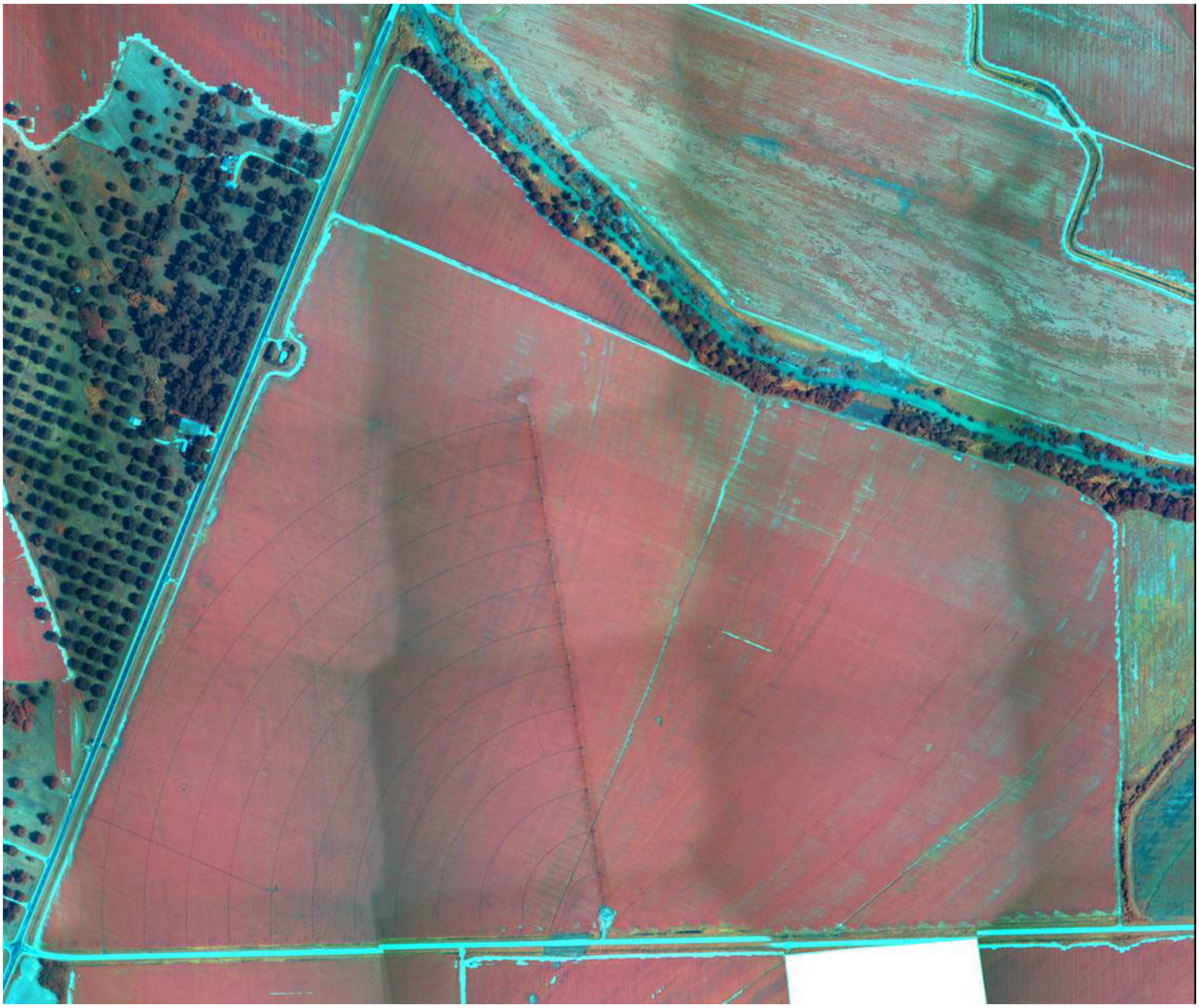

Figure 11. Color-infrared image of study site August 9, 2002.

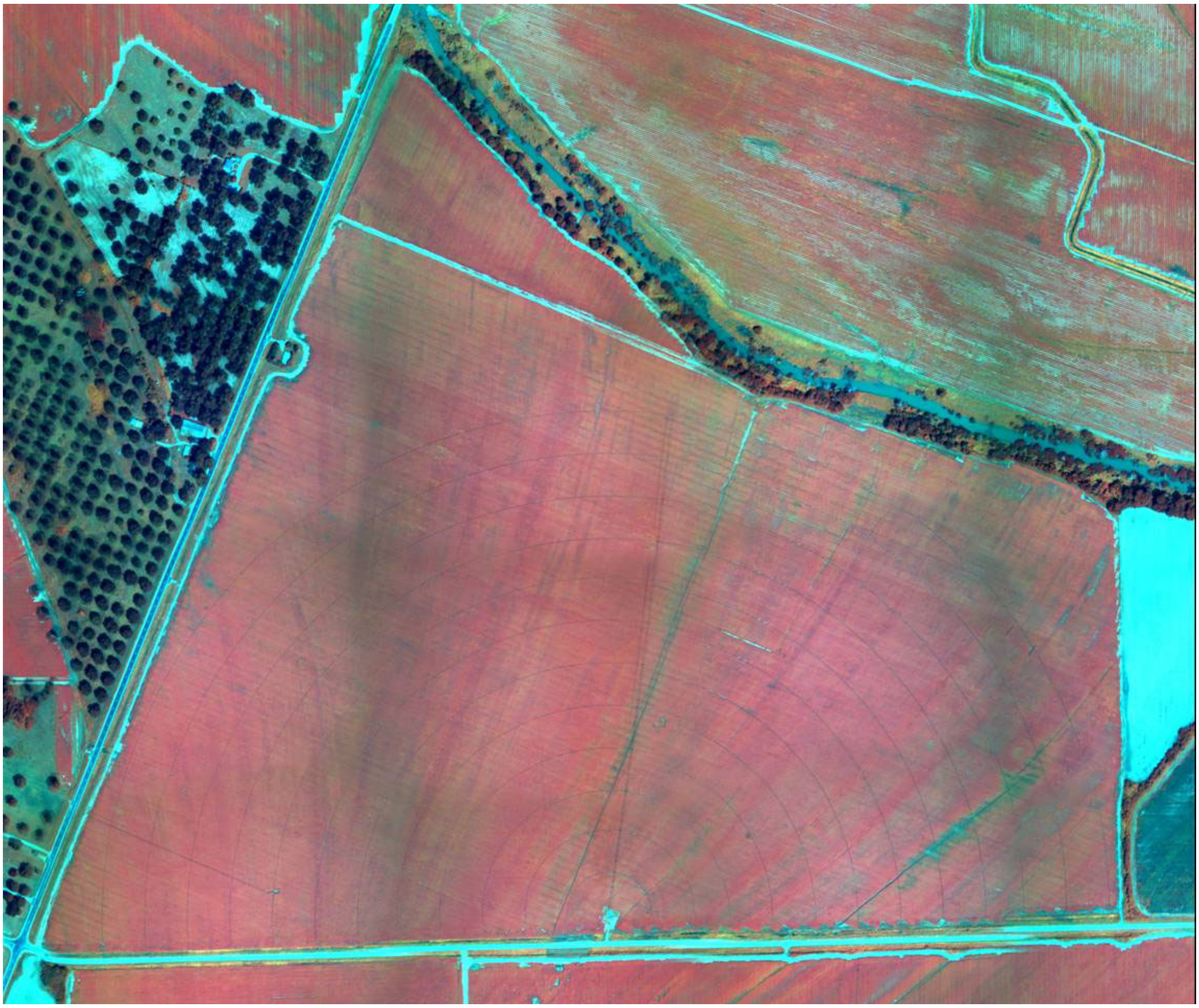

Figure 12. Color-infrared image of study site August 21, 2002.

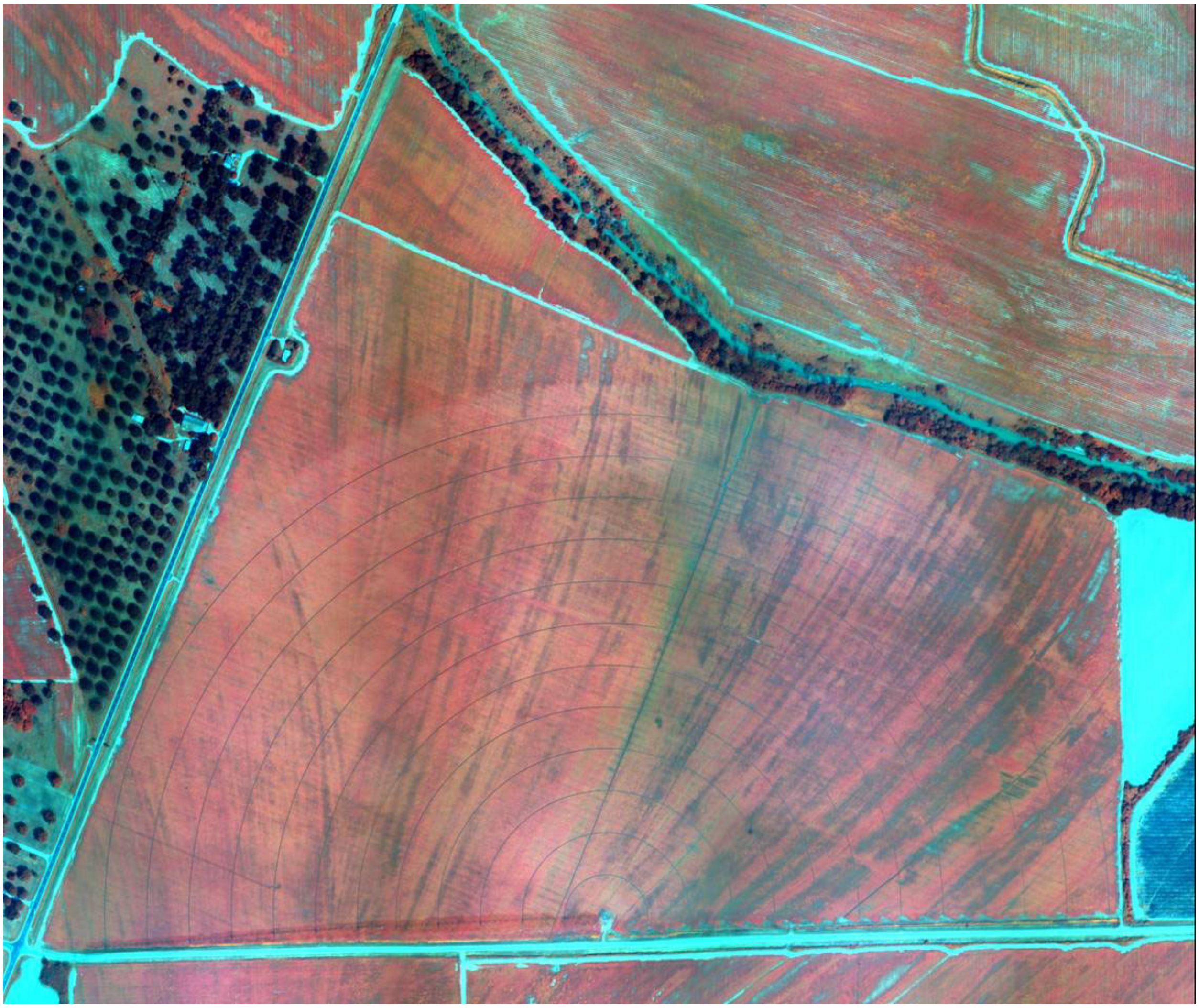

Figure 13. Color-infrared image of study site September 5, 2002.

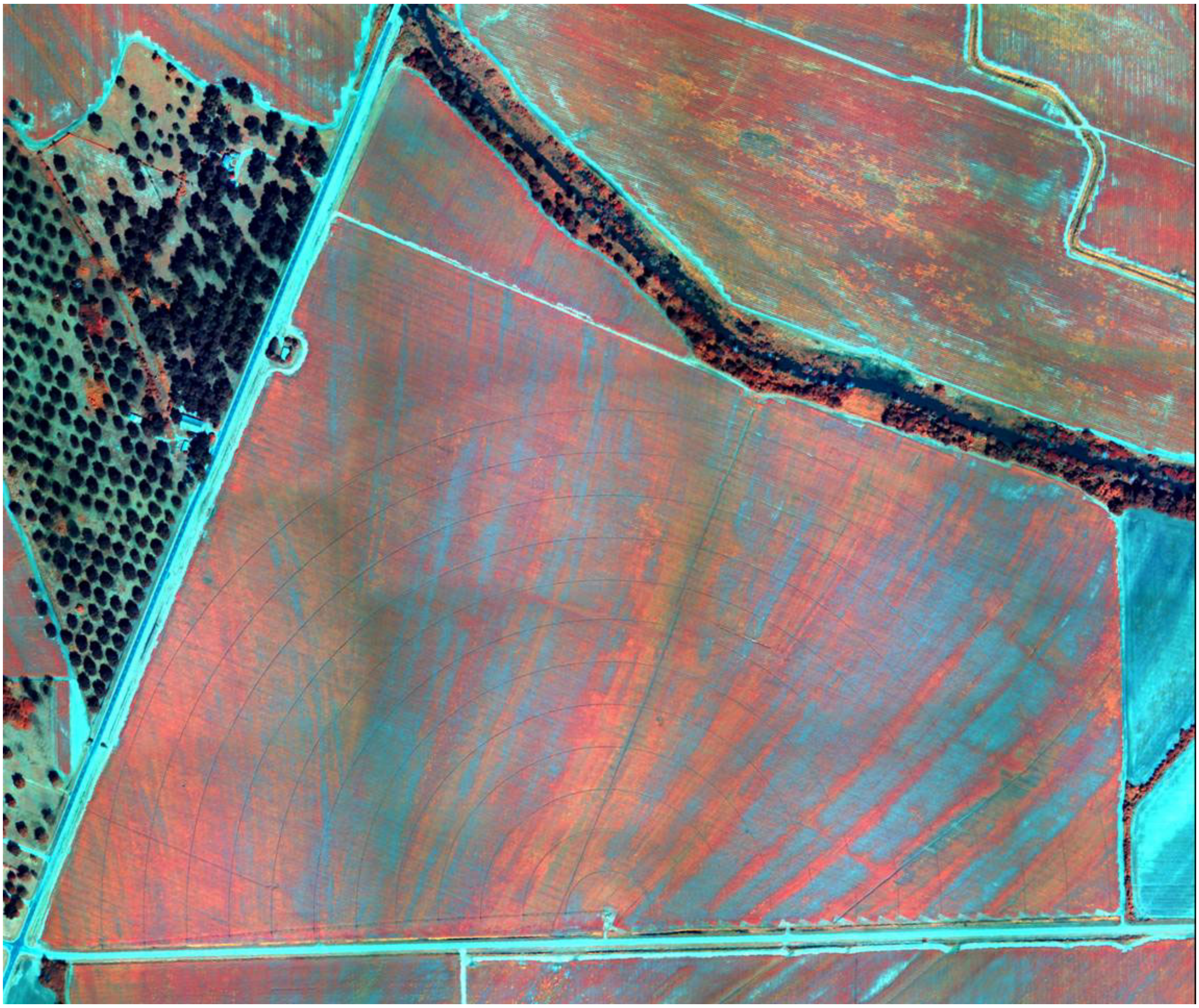

Figure 14. Color-infrared image of study site September 23, 2002.

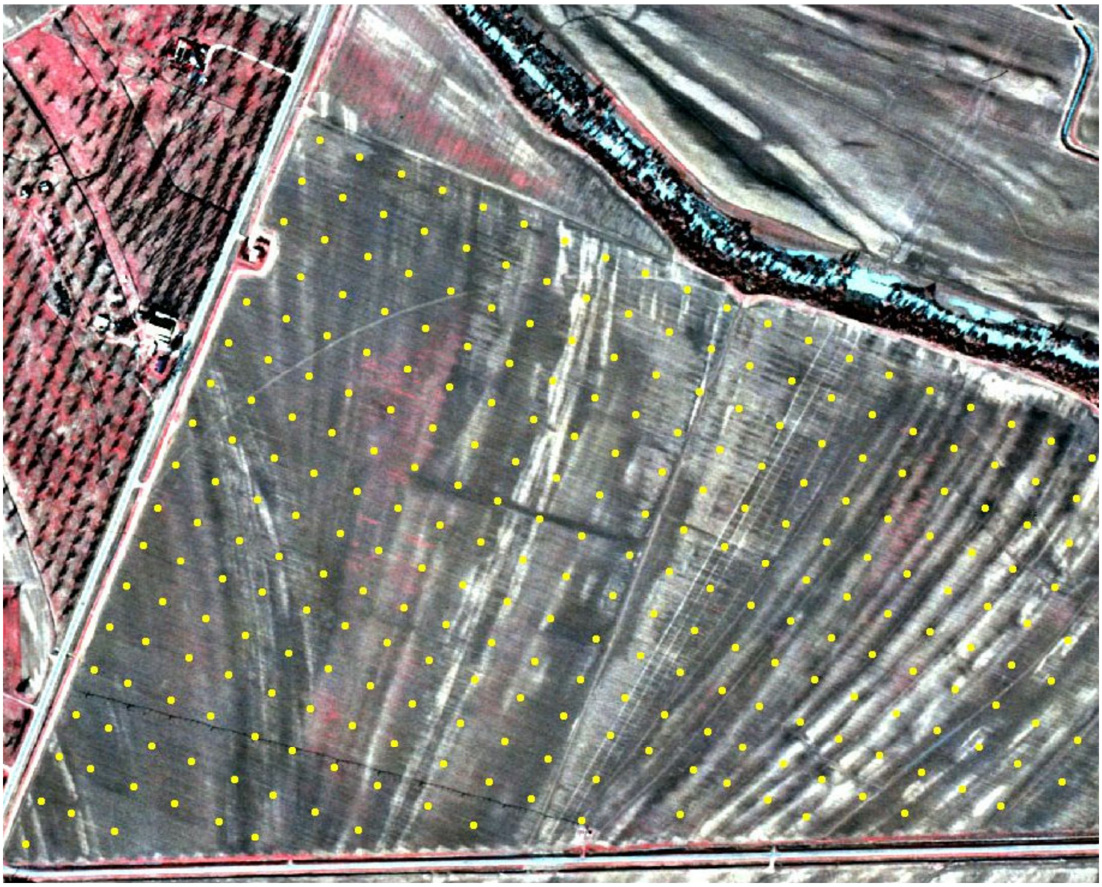

Figure 15. Digital Ortho Quad image of study site in 1992, showing predominantly bare soil, with soil sample locations overlaid as yellow dots.

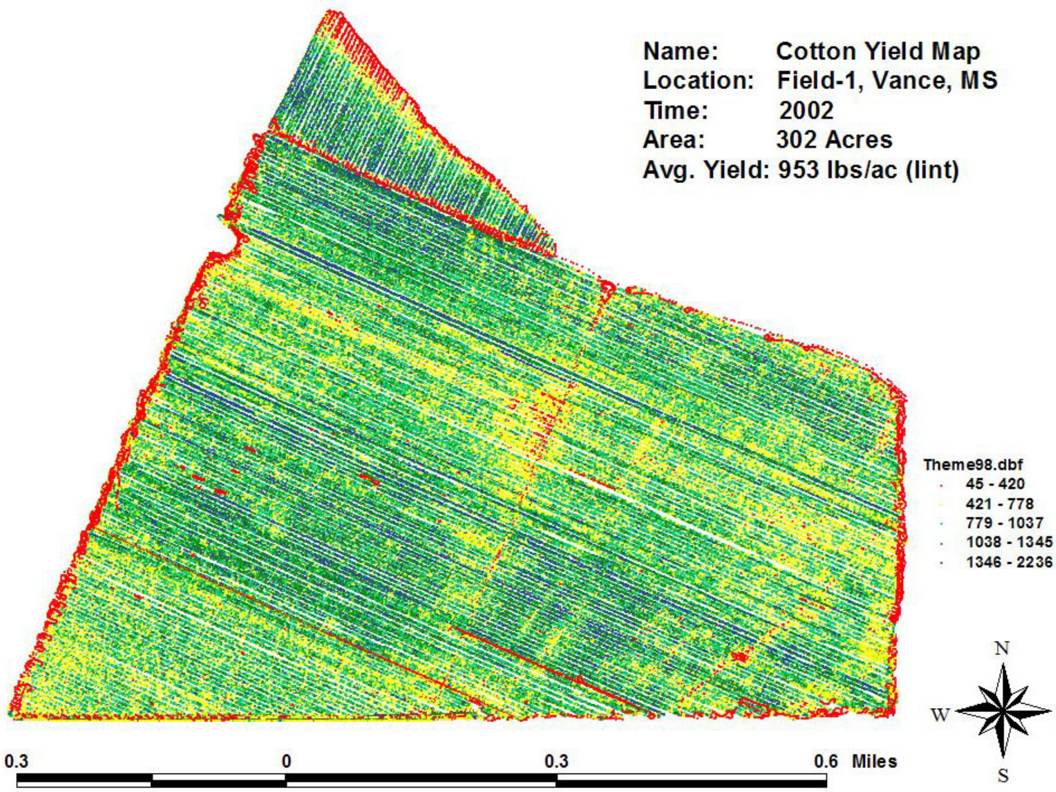

Figure 16. Cotton yield map of commercial field in Mississippi.

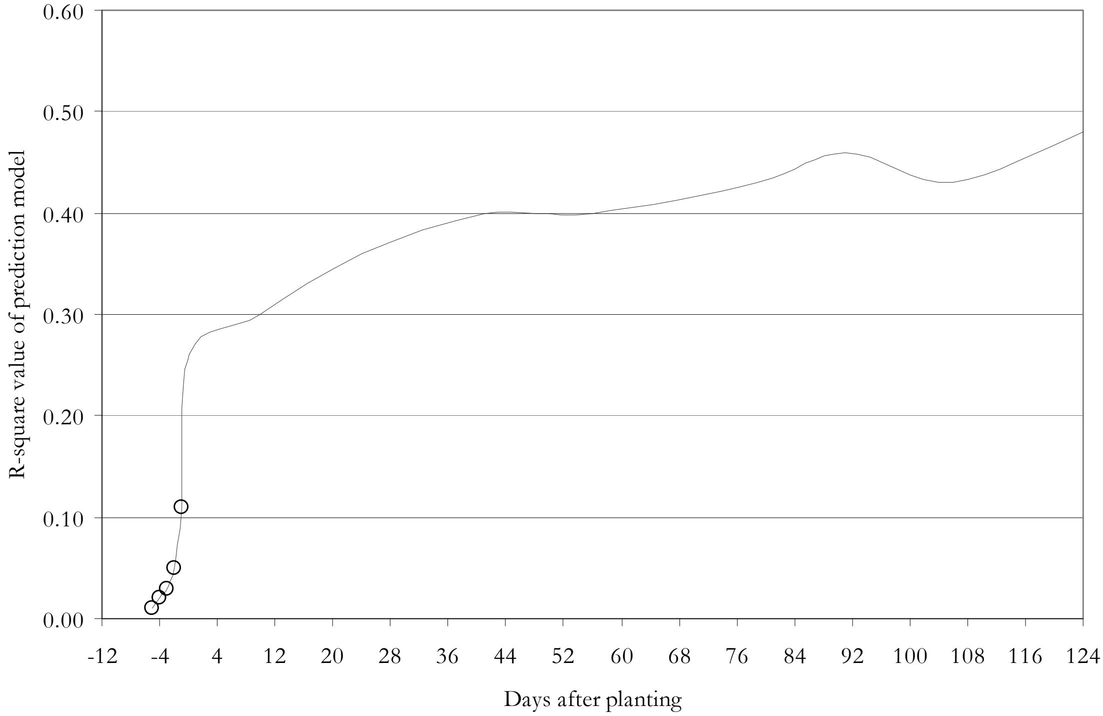

Figure 17. R^2 values of models, at numerous dates throughout the growing season, to predict cotton yield based on data set A. Negative days included for data available prior to planting.

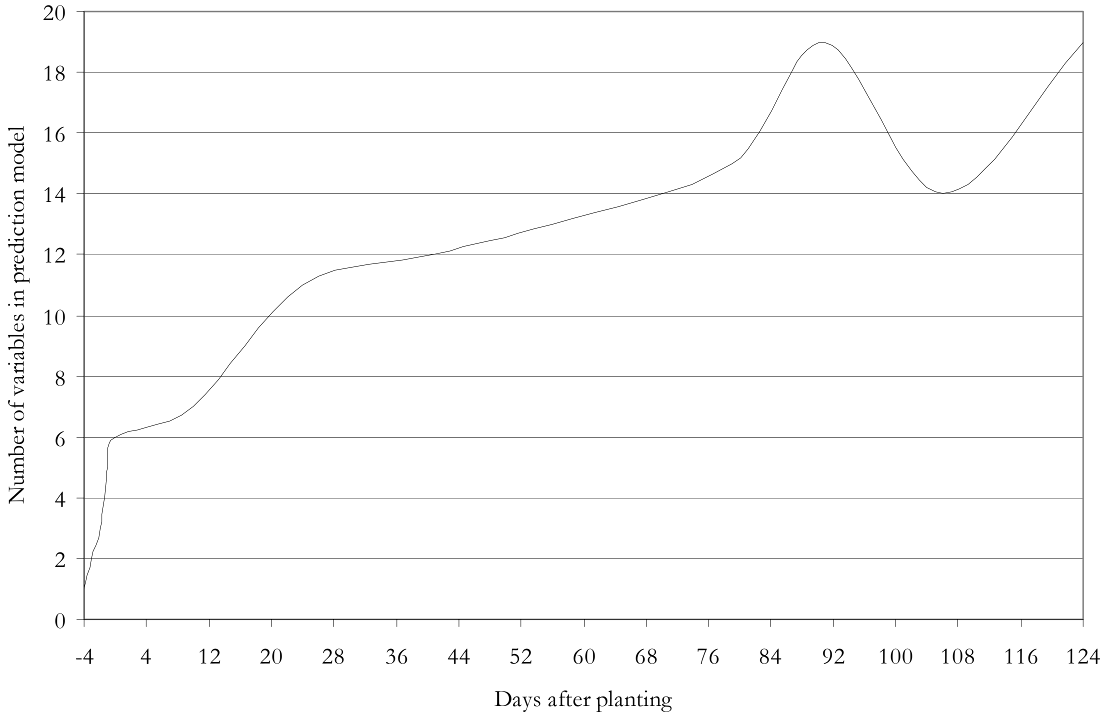

Figure 18. Number of variables in models, at numerous dates throughout the growing season, to predict yield based on data set A. Negative days included for data available prior to planting.

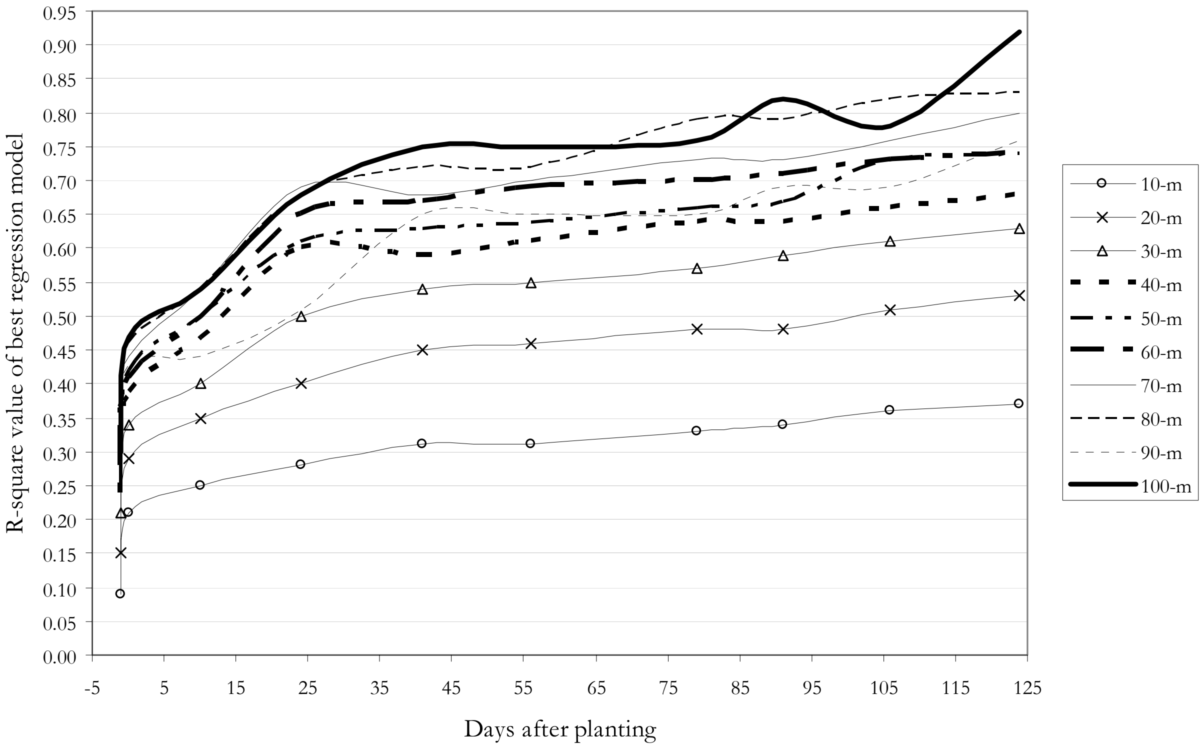

Figure 19. R^2 values of models, at numerous spatial resolutions and dates throughout the growing season, to predict cotton yield based on data set B. Negative days included for data available prior to planting.

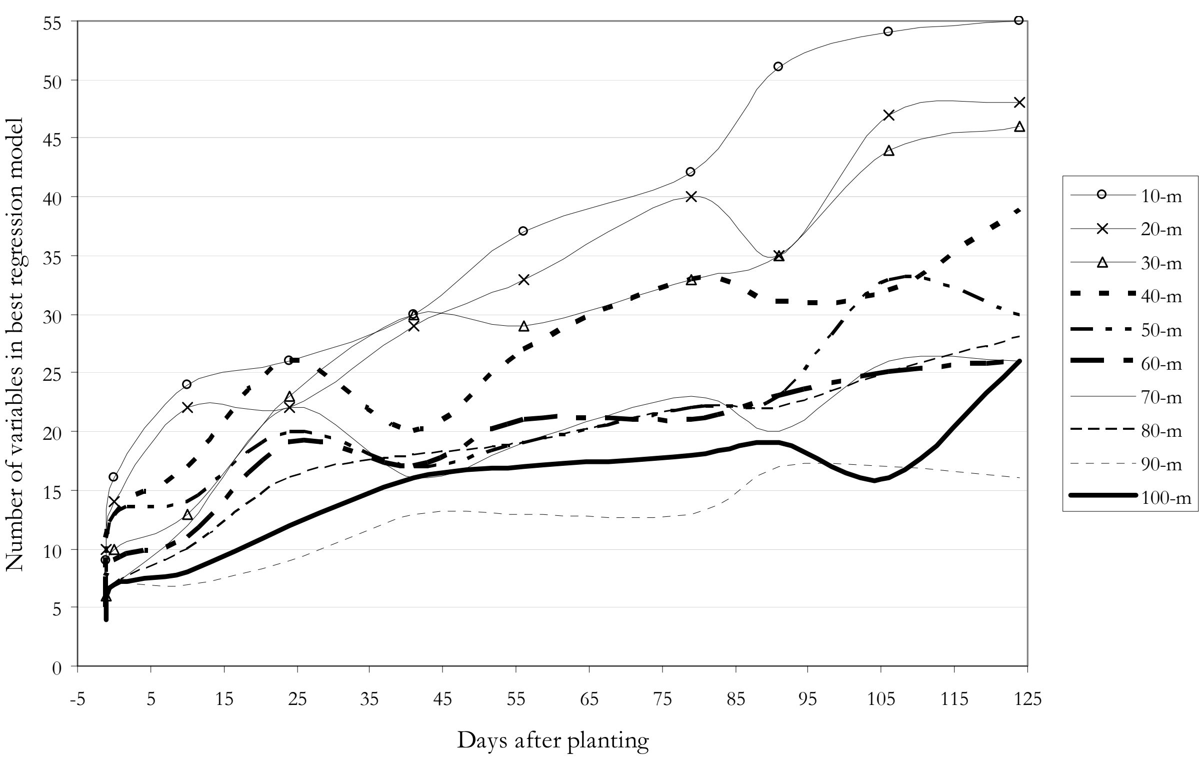

Figure 20. Number of variables in models, at numerous spatial resolutions and dates throughout the growing season, to predict yield based on data set B. Negative days included for data available prior to planting.

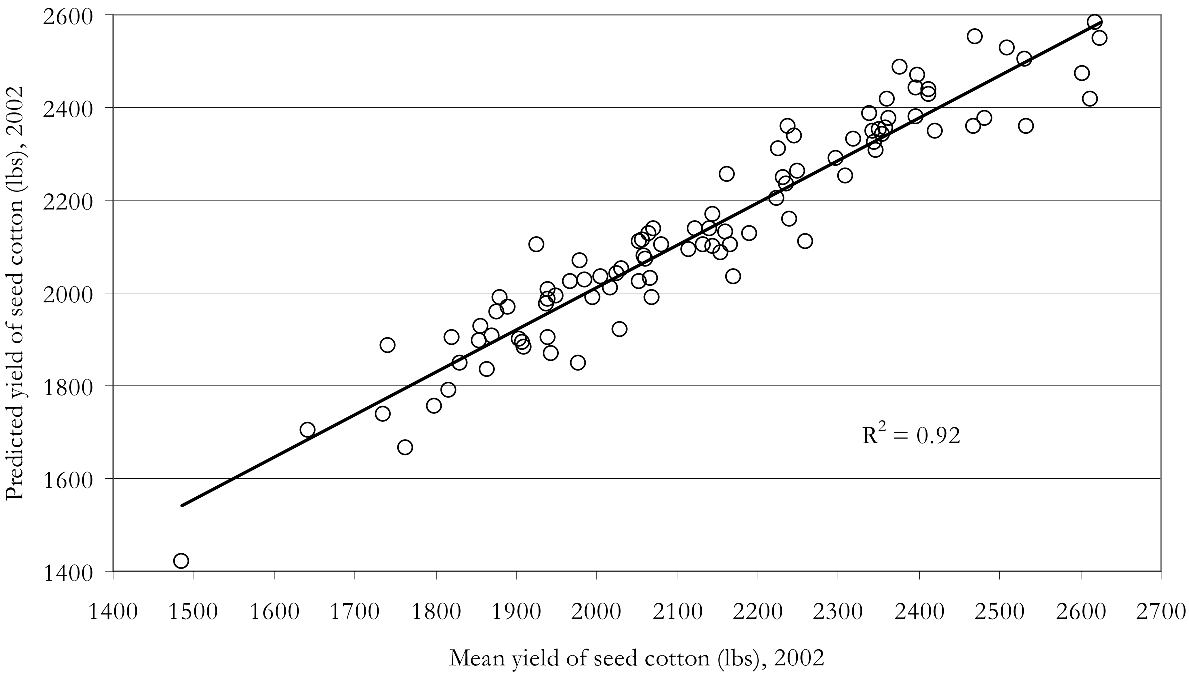

Figure 21. Predicted values and regression line of focus model from data set B: 100-m ground-distance resolution, including historical-yield, soil-texture, topographic, and remote-sensing data from all possible dates during growing season.

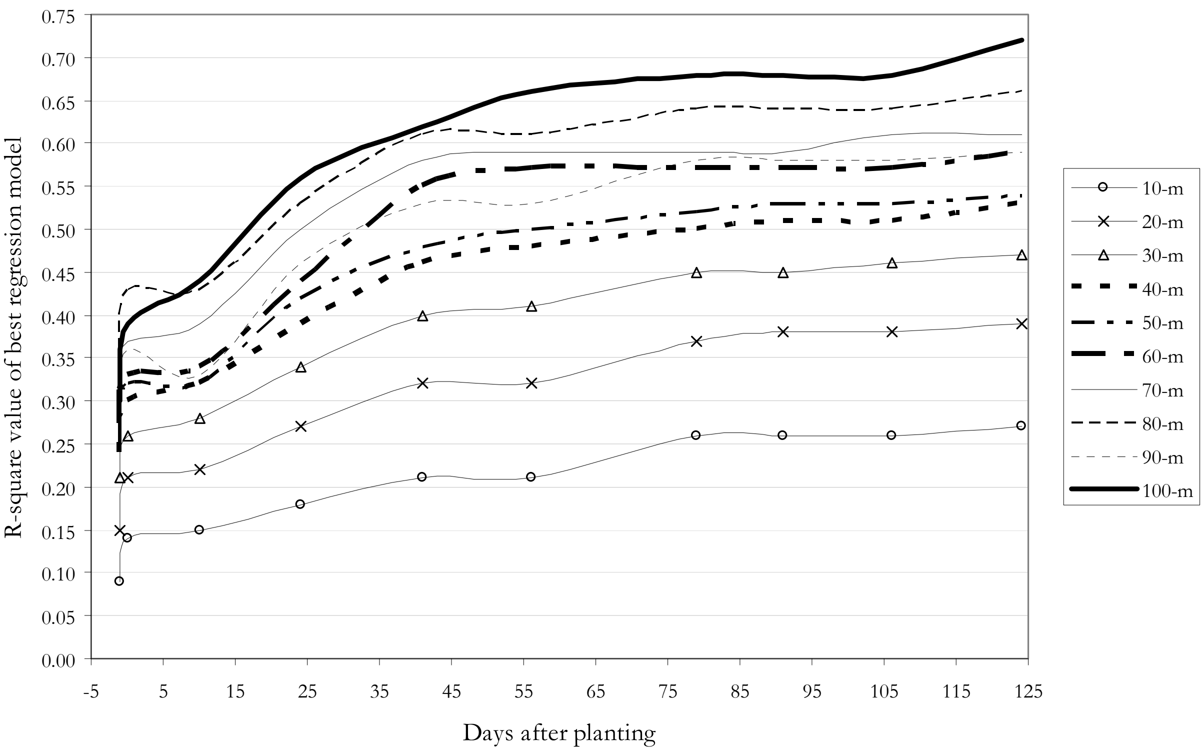

Figure 22. R^2 values of models, at numerous spatial resolutions and dates throughout the growing season, to predict cotton yield based on data set C. Negative days included for data available prior to planting.

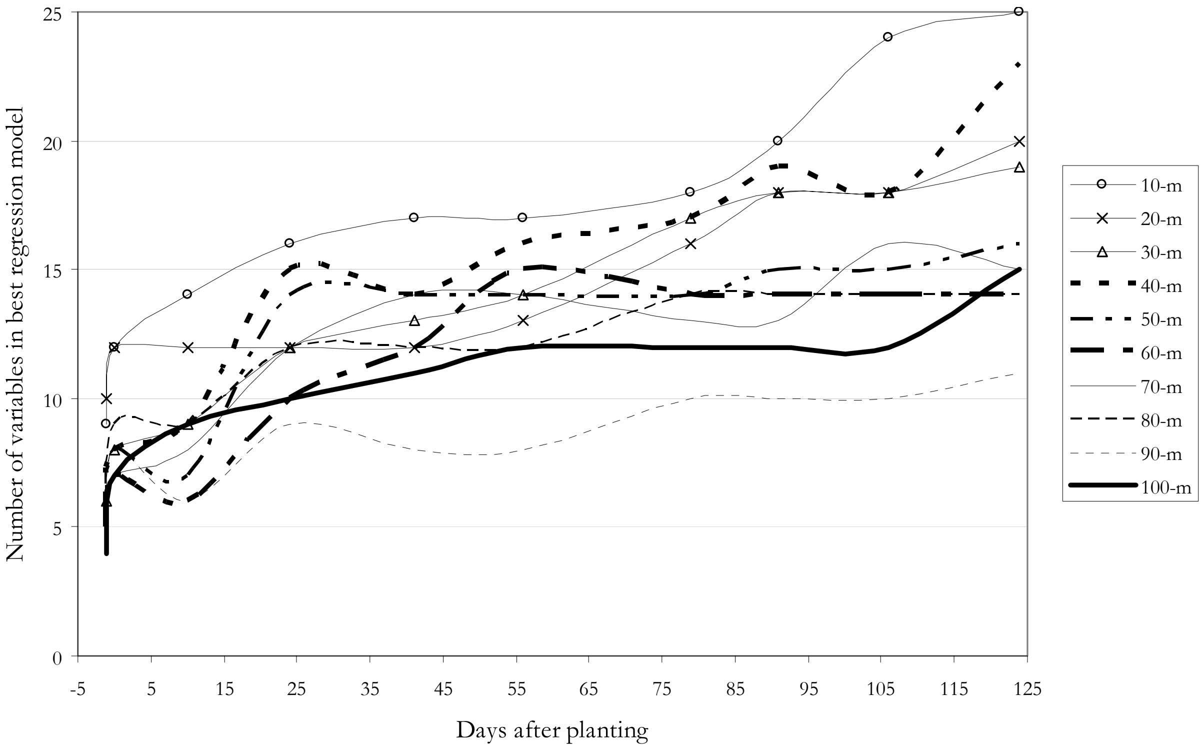

Figure 23. Number of variables in models, at numerous spatial resolutions and dates throughout the growing season, to predict yield based on data set C. Negative days included for data available prior to planting.

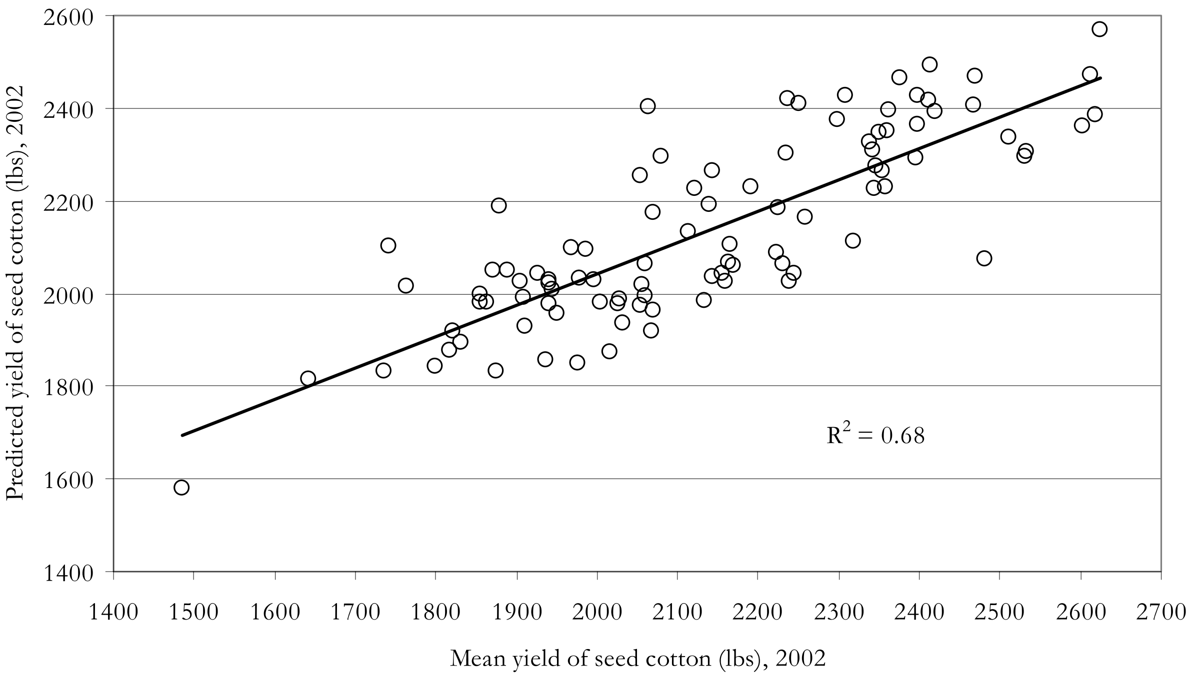

Figure 24. Predicted values and regression line of focus model from data set C: 100-m ground-distance resolution, including historical-yield, soil-texture, topographic, and remote-sensing indices through 79 days after planting.