Computers and Electronics in Agriculture 105 (2014) 103–110

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Employing broadband spectra and cluster analysis to assess thermal defoliation of cotton ☆

Reginald S. Fletcher [a,*], Allan T. Showler [b,1], Paul A. Funk [c,2]

[a] USDA-ARS-Crop Productions System Research Unit, P.O. Box 350, Stoneville, MS 38776, United States
[b] USDA-ARS-Knipling-Bushland U.S. Livestock Insects Research Laboratory, 2700 Fredericksburg Road, Kerrville, TX 78028, United States
[c] USDA-ARS-Cotton Ginning Research, 300 E. College Dr., Mesilla Park, NM 88047, United States

ARTICLE INFO

Article history:
Received 13 September 2013
Received in revised form 28 March 2014
Accepted 9 April 2014

Keywords:
Thermal defoliation
Unsupervised classification
Multispectral
Normalized difference vegetation index
Cotton

ABSTRACT

Growers and field scouts need assistance in surveying cotton (*Gossypium hirsutum* L.) fields subjected to thermal defoliation to reap the benefits provided by this nonchemical defoliation method. A study was conducted to evaluate broadband spectral data subjected to unsupervised classification for surveying cotton plots subjected to thermal defoliation. Ground-based reflectance measurements of thermally treated and non-treated cotton canopies were collected at two study Sites (Site 1 and Site 2) with a handheld hyperspectral spectroradiometer. The hyperspectral data were merged into eight broad spectral bands: coastal blue (400–450 nm), blue (450–510 nm), green (510–580 nm), yellow (585–625 nm), red (630–690 nm), red-edge (705–745 nm), near-infrared (770–895 nm), and panchromatic (450–800 nm). Also, a broadband normalized difference vegetation index (NDVI) was created with the red (630–690 nm) and near-infrared bands (770–895 nm). For each study Site, two datasets were analyzed: (1) two-class case (thermally treated cotton observations and non-treated cotton observations) and (2) five-class case (thermally treated cotton observations and non-treated cotton observations and three additional classes created with the weighted average of the thermally treated cotton observations and non-treated cotton observations). The clustering algorithm referred to as CLUES (CLUstEring based on local Shrinking) was employed to automatically group the data into clusters without the user selecting the number of clusters. Cluster validation was determined with the average silhouette width; also accuracy was assessed with contingency matrixes. Clustering analysis worked well in dividing the data into appropriate groups, with the best cluster structure occurring for the NDVI. User's and producer's accuracies for the NDVI were greater than 86%, indicating an excellent classification. Findings support future endeavors to assess air-borne and satellite-borne systems equipped with sensors sensitive to the wavelengths deemed useful in this study and unsupervised classification techniques that automatically determines the numbers of clusters to evaluate thermal defoliation of cotton fields.

Published by Elsevier B.V.

1. Introduction

Over the last ten years, thermal defoliation has shown promise as a nonchemical alternative for terminating cotton growth and defoliating cotton canopies (Funk et al., 2004, 2006; Showler et al., 2006), making it an ideal harvest aid for cotton grown under organic farming methods (Funk et al., 2006, 2012). The technique involves using propane to heat air that is applied directly to the plant canopy to quickly kill the leaves. Unlike traditional defoliation methods, it does not require optimal weather conditions to apply the treatment, and producers are able to harvest fields twenty-four hours after treatment if needed (Showler et al., 2006). Additionally, producers can use this form of defoliation in conventional systems to prepare fields threatened by severe weather, increasing their ability to harvest fields before the arrival of the inclement weather (Showler et al., 2006). Finally, thermal defoliation has shown potential for late-season pest control (Bundy et al., 2006; Funk et al., 2012).

There are several risks associated with incomplete thermal defoliation. Staining of the cotton fiber may occur in areas where leaf kill is incomplete, reducing the price grade of the cotton. Additionally, juices in green leaves may increase gum build-up on

☆ This project was funded by USDA-ARS; mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA.

* Corresponding author. Tel.: +1 662 686 5298; fax: +1 662 686 5422.
E-mail address: reginald.fletcher@ars.usda.gov (R.S. Fletcher).
[1] Tel.: +1 830 792 0341; fax: +1 830 792 0314.
[2] Tel.: +1 575 526 6381; fax: +1 575 525 1076.

http://dx.doi.org/10.1016/j.compag.2014.04.003
0168-1699/Published by Elsevier B.V.

picker spindles, requiring stoppage to clean the spindles and thus increasing harvesting time per field. Therefore, rapidly identifying areas, not responding to thermal treatment can support producer decision making regarding re-treating these areas. This would maintain the price grade of the cotton throughout treated fields and reduce the time required to harvest them.

Variation of cotton growth within the field is usually the same from year to year within a field. If re-treatment is not an option for the current year, growers do need a record showing areas not responding to treatment. In the following year, producers can adjust the defoliation process; for example, reducing the speed of the defoliating apparatus in areas not responding to treatment in the previous year would increase leaf kill. For these reasons, growers need assistance in surveying fields subjected to thermal defoliation to reap the benefits provided by this nonchemical defoliation method.

Ground-based and airborne remote sensing systems have shown potential as tools for monitoring defoliation of cotton. Yang et al. (2003) demonstrated applications of ground-based reflectance data and airborne color-infrared (CIR) and normalized difference vegetation index (NDVI) imagery to differentiate cotton plants exposed to different defoliants from a control and other chemical treatments. They noted that differences between control and treated plants were statistically significant for green, red, and NDVI recorded data. Ground-based reflectance indices based on red-edge measurements provided accurate and consistent defoliation estimates for cotton subjected to chemical treatment (Ritchie and Bednarz, 2005). Fletcher et al. (2007) showed that CIR photography had potential for differentiating thermally defoliated cotton plants from control plants. Color-infrared film is no longer sold, eliminating it as an option for consultants and producers to use for surveying thermally defoliated cotton fields. Most remote sensing research have focused on chemical defoliates and have concentrated on visual, statistical, and modeling efforts (Ritchie and Bednarz, 2005; Yang et al., 2003, 2011). No information is available on using classification algorithms as a means to group thermally defoliated plants based on their spectral properties.

Cotton plants exposed to thermal treatment may be perfect candidates for remote sensing instruments. Heat kills the leaves, causing them to turn brown within a few hours. Showler et al. (2006) reported that up to 60% of the desiccated leaves remain on the cotton plant until harvest.

Supervised and unsupervised classifications are general approaches used to derive maps from image data or group unknown reflectance data into classes. Supervised classification uses a set of user-defined spectral signatures to assign samples to groups (Campbell, 2002; Mather, 2005). Unsupervised classification requires minimum input from the operator; no training samples are used, and subdivision of the feature space is achieved by identifying natural groupings of the measurement vectors (Campbell, 2002; Mather, 2005). Clustering is the most widely used unsupervised procedure. It involves development of structures in unlabeled data by grouping pixel data into groups with similar properties. Pixel values within clusters are similar to each other, but are not similar to pixel values in other clusters.

Clustering algorithms such as K-means and the Iterative Self-Organizing Data Analysis Technique requires the user to input the number of clusters to create. For the user, deciding the number of clusters is often a difficult task. Researchers have developed clustering algorithms that automatically estimate the number of clusters without user input. These procedures are divided into methods that select the number of clusters by optimizing a measure of strength of the clusters (Tibshirani et al., 2000), that assemble the data into small clusters followed by merging these clusters until no further merging can occur (Frigui and Krishnapuram, 1999), that extract one cluster at time (Zhung et al., 1996), that uses a bump hunting technique to determine the number clusters (Wang et al., 2007), and that iteratively move data points toward cluster centers and then use the number of convergent points as the number of clusters (Wang et al., 2007).

Broadband (spectral bands greater than 10 nm) panchromatic and multispectral remote sensing products are readily available to the public through commercial companies via airborne or satellite platforms. It was hypothesized that cotton plants effectively treated by thermal defoliation could be separated from non-treated plants using broadband spectral data and unsupervised classification based on clustering analysis. The objective of this study was to evaluate ground-based broadband spectral data and cluster analysis as tools for surveying cotton plots subjected to thermal defoliation to prepare them for harvesting. Spectra simulating the broadbands of the WorldView 2 satellite were examined in this study: coastal blue (400–450 nm), blue (450–510 nm), green (510–580 nm), yellow (585–625 nm), red (630–690 nm), red-edge (705–745 nm), near-infrared (770–895 nm), and panchromatic (450–800 nm). These bands were chosen because the blue, green, red, near-infrared, and panchromatic bands are similar to those found on other high spatial resolution broadband satellite sensors. The coastal blue, yellow, and red-edge bands provide additional information relevant to identifying and mapping vegetation (Digital Globe, 2009). Also, the spectral bands are easy to fabricate for filters to use in multispectral cameras flown in aircraft. Furthermore, the study focused on using a clustering algorithm that determined the number of clusters without user input and application of multispectral, panchromatic, and vegetation index forms of the data as input into the clustering algorithm to separate adequately treated plants from inadequately treated or non-treated plants.

2. Materials and methods

2.1. Study sites

Data were collected from two study Sites, referred to as Site 1 and Site 2. The study Sites were located near the Kika de la Garza Subtropical Agricultural Research Center, Weslaco, Texas (26°09′N 97°57′W). These Sites were being used for on-going studies to compare thermal defoliation to chemical defoliation. The experiment at each study Site was a randomized complete block design consisting of three treatments (thermal defoliation, chemical defoliation, and control) replicated six times (blocks). Treatments within each replicate contained twelve rows planted to Deltapine 5415RR (Delta Pine and Land Co.; Scott, MS). Row spacing was 1 m, resulting in an area of 0.19 ha per treatment. The objective of the current study was to compare the thermally defoliated cotton and the control cotton; therefore, subsequent analyses focused on these two treatments. The thermal treatment occurred on 27 July 2005. It involved using the two row thermal defoliator prototype described by Funk et al. (2006) to apply heat at 193 °C to the cotton canopies.

2.2. Field spectra collection

Reflectance measurements (referred to as reflectance) of the non-treated and thermally treated cotton canopies were collected on 8 August and 1 August 2005 at Sites 1 and Sites 2, respectively, with a FieldSpec Handheld spectroradiometer (Analytical Spectral Devices, Inc., Boulder, CO) having a spectral range of 325–1075 nm. The device has a spectral resolution of 3 nm; however, the data is resampled to 1 nm intervals by the software used to operate the instrument. The data output range is between 0 (0% reflectance) and 1 (100% reflectance). The instrument has a 25°

viewing angle. During sampling, it was held approximately one meter above the plant canopies, resulting in a field of view of 0.696 m^2. At each study Site, the treatments were replicated six times; reflectance measurements were collected at five different locations in each non-treated and thermally treated replicate. Each location in a treatment is considered an observation (i.e., a sample); therefore, a total of thirty samples were collected per treatment at Sites 1 and 2. Each observation was an average of ten spectra. The instrument was calibrated at ten-minute intervals with a spectralon white reference panel (Labsphere, North Sutton, NH). Data acquisition occurred under sunny conditions ±2 h of solar noon.

2.3. Post processing of field spectral data

Post processing of the reflectance data involved using the ViewSpec Pro software (version 5.6.10, Analytical Spectral Devices, Inc., Boulder, CO) to qualitatively analyze the data and to check for erroneous data. During this review, it was discovered that data was only collected at twenty-nine locations for the thermally treated cotton canopies at Site 1; for this Site, subsequent analysis will focus on twenty-nine thermally treated cotton observations and thirty non-treated cotton observations. For Site 2, thirty spectra were recorded for the treated and for the non-treated cotton observations. Spectra were exported to text format for further processing. Wavelengths greater than 880 nm and less than or equal to 940 nm were moderately noisy and wavelengths greater than 940 nm were severely affected by noise. A Savitzky–Golay filter (Savitzky and Golay, 1964; Press et al., 2007) was applied to reduce the noise in the spectra. The parameters that worked best were a second order polynomial encompassing thirty-one points to create the new point. However, spectra greater than 900 nm were still noisy, leading to data greater than 900 nm not being used for further analysis.

Spectra simulating the broadbands of the WorldView 2 satellite were explored in this study: coastal blue (400–450 nm), blue (450–510 nm), green (510–580 nm), yellow (585–625 nm), red (630–690 nm), red-edge (705–745 nm), near-infrared (770–895 nm), and panchromatic (450–800 nm). For each observation, the band equivalent reflectance (BER, Trigg and Flasse, 2000) equation (see Eq. (1)) was used to convert the Savitzky–Golay filtered hyperspectral data to the band equivalent reflectance for the WorldView 2 spectral bands. Other researchers have employed BER equation to transform narrowband hyperspectral data to broadband spectral data (Trigg and Flasse, 2000; Eitel et al., 2007, 2011; Smith et al., 2010).

$$R_x = \sum_{i=\lambda\ min}^{\lambda\ max} r_i p_i \Bigg/ \sum_{i=\lambda\ min}^{\lambda\ max} r_i \tag{1}$$

Parameters of Eq. (1) are as follows: R_x represents BER for band x, λ min and λ max denotes minimum and maximum reflectance, respectively, of band *x's* filter function, r_i equals relative response for band x at wavelength i, and p_i indicate relative reflectance measured by the spectrometer at wavelength i (Trigg and Flasse, 2000). Relative spectral response values (i.e., values quantifying how the bands of various satellite sensors measure the intensity of the wavelengths of light), were obtained from Digital Globe, Inc., the commercial vendor and operator of the WorldView 2 satellite. The WorldView 2 satellite also has a sensor that collects data in a second near-infrared band (860–1040 nm). This band was not tabulated with the Savitzky–Golay filtered hyperspectral data because spectral data greater than 900 nm were too noisy to employ for further processing. The normalized difference vegetation index (NDVI, Rouse et al., 1973) was derived with the following equation.

$$NDVI = \frac{(near\text{-}infrared\ reflectance - red\ reflectance)}{(near\text{-}infrared\ reflectance + red\ reflectance)} \tag{2}$$

It was tabulated with the broadband red (630–690 nm) and broadband near-infrared (770–895 nm) bands created in the previous step.

2.4. Simulation study-five class case

As indicated earlier for the thermal defoliation, heat kills the leaves causing them to turn brown within a few hours. Many of them curl up and remain attached to the stalk, depending on treatment temperature. To enhance our knowledge of the effectiveness of using the broadband spectral data and the clustering algorithm to survey cotton plots subjected to thermal defoliation, the non-treated and treated canopy observations were used to develop three additional classes representing a mixture of green and brown leaves: (1) 75% non-treated – 25% thermally treated, 50% non-treated – 50% thermally treated, and 25% non-treated – 75% thermally treated. For interpretation purposes and an example, 75% non-treated – 25% thermally treated represents a thermally treated observation; yet, its appearance is more similar to the non-treated observation.

The following procedures were used to derive the three simulated classes for study Sites 1 and 2: (1) the thermally treated and non-treated observations Savitzky–Golay filtered hyperspectral data were employed to create a new spectral measurement representing a specific simulated class based on a weighted average and (2) then the spectrum created in step 1 was used as input into the BER equation (Eq. (1)) to produce the broadband spectra simulating the WorldView 2 bands. The replicates were equally represented for developing the new spectral classes. As indicated earlier, at Site 1, five samples each were collected from non-treated and thermally treated plots within replicate 1. The samples from those treatments were used to create five spectra for the three additional classes with the following technique. The samples were number one through five. Using a random number generator in Microsoft Excel, one sample was selected at random from the thermally treated and non-treated cotton spectral observations to create the new spectra for a particular class. The random selection was repeated until all spectra for that replication were created (Note: a non-treated and thermally treated cotton observation were used only once for spectra development for a particular class.). Then, the same procedure was repeated for replicates 2–6. The simulated spectra were calculated with Microsoft Excel (2007, Microsoft Corporation, Redmond, WA). Finally, the simulated hyperspectral data for each class were used as input into the BER equation to create the broad spectral bands for the simulated classes. The NDVI for the simulated classes were determined using Eq. (2) and the broadband simulated classes' near-infrared and red bands. The original data plus the three additional classes are referred to as the five-class data set.

2.5. Unsupervised classification

The R package (R version 2.15.2; R Core Team, 2012) CLUES (CLUstEring based on local Shrinking, Chang et al., 2010) was used to develop clusters. It uses nonparametric measures to employ shrinking (i.e., condensing clusters by moving observations toward the center of the cluster) and partitioning (i.e., grouping observations into clusters) procedures to approximate an optimal number of groups by using either the CH index or the Silhouette index, rather than relying on guessing a pre-specified number of groups (Chang et al., 2010). The following arguments were input into the algorithm to develop the clusters: (1) distance method – Euclidean and (2) strength method – silhouette width index. The distance method represents the dissimilarity measurement, differences between observations based on some criterion. For this study the Euclidean distance was used as the difference metric; its values range from 0 to infinity (Chang et al., 2010). The larger the value, the further apart are the two observations. Thus, observations with large values will not be grouped into the same cluster.

The average silhouette width index provided a measure of the compactness (i.e., a measurement of how well the observations are packed together in a cluster, cluster homogeneity) of the clusters and an indication of how well the clusters were separated. For each sample assigned to a cluster, the silhouette index is determined by calculating the dissimilarity of a data observation in one cluster to data observations assigned to the nearest neighboring cluster (Chang et al., 2010). This calculation is performed on all observations within clusters. The average silhouette index value is the mean of silhouette index values tabulated for all of the observations. Its values range from -1 to +1. The following scale was used to describe cluster strength based on the average silhouette index value: (1) 0.71–1.0, a strong structure has been found; (2) 0.51–0.70, a reasonable structure has been found; (3) 0.25–0.50, the structure is weak and could be artificial; and (4) less than 0.25, no substantial structure has been found (Anonymous, 2011).

For the unsupervised classification procedure, the data were partitioned into five band combinations: (1) a three-band combination consisting of blue, green, and red bands, (2) a three band-combination consisting of green, red, and near-infrared bands, (3) a seven-band combination consisting of coastal blue, blue, green, yellow, red, red-edge, and near-infrared bands, (4) a single band containing the panchromatic band, and (5) a single band consisting of the NDVI.

Additionally, contingency matrices were used to analyze the accuracy of the clustering analysis, including user's and producer's accuracies (Jones and Vaughan, 2010). The former refers to the probability that observation labeled as a certain land-cover by clustering is really that class; the latter refers to the probability that a certain land-cover of an area on the ground is classified as such. User's accuracy is calculated by dividing the total number of correct observations for a category by the total number of observations classified in that category. Producer's accuracy is determined by dividing the total number of correct observations within a specific category by the total number of observations as indicated by reference data for that specific category.

3. Results

3.1. Thermally treated versus non-treated (two-class case)

Summarized in Table 1 are the clustering results obtained at Site 1 for the thermally treated versus non-treated observations. Two clusters were established for the single and multiple band combinations used as input into the clustering algorithm (Table 1). The best cluster structure, as indicated by the silhouette index, was established with the NDVI data followed, in descending order, by the three-band combination consisting of the green, red, and near-infrared bands, the seven-band combination consisting of the coastal blue, blue, green, yellow, red, red-edge, and near-infrared bands, the panchromatic band, and the three-band combination consisting of the blue, green, and red bands. User's and producer's accuracies were equal to 100% for all of the band combinations, indicating an excellent classification; nevertheless, all of the band combinations except for the blue, green, and red bands achieved silhouette index values representing strong cluster structures (silhouette index range – 0.71–1.0).

With the exception of the panchromatic data, thermally treated and non-treated cotton observations were placed into two separate clusters for the spectral data analyzed with the clustering algorithm at Site 2; however, the results were variable (Table 2). For the panchromatic band, the clustering algorithm divided the data into four clusters. Cluster one consisted of fourteen non-treated cotton observations and one thermally treated cotton observation, cluster 2 contained thirteen non-treated cotton observations and eight thermally treated cotton observations, cluster three had three non-treated cotton observations and eleven thermally treated cotton observations, and cluster four consisted of ten thermally treated cotton observations. Clusters one and two represented the non-treated cotton class, and clusters three and four symbolized the thermally treated cotton class. Thus, clusters one and two were combined, and clusters three and four were merged.

The descending order of silhouette index values (Table 2) at Site 2 for the band combinations resulting in the development of two clusters were (1) NDVI, (2) green, red, and near-infrared band combination, (3) seven-band combination, and (4) blue, green, and red band combination. Note: four clusters were originally derived from the panchromatic dataset. Strong cluster structure (silhouette index range – 0.71–1.0) was attained by the NDVI and three-band combination consisting of the green, red, and near-infrared bands; reasonable cluster structure (silhouette index range – 0.51–0.70) was achieved for the other datasets.

User's and producer's accuracies of 100% were obtained for the three-band combination consisting of the green, red, and near-infrared bands, the seven-band combination consisting of the coastal blue, blue, green, yellow, red, red-edge, and near-infrared bands, and the NDVI. Compared with the other band combinations, higher errors occurred in discriminating the non-treated cotton observations from treated cotton observations for the three-band combination consisting of the blue, green, and red bands, and the panchromatic band. Including both datasets, the user's and producer's values ranged from 70% to 100%.

3.2. Five-class case

Summarized in Table 3 are the number of clusters and the average silhouette index values obtained for the five-class case for Site 1. Five clusters matching the original five classes were created with three of the band combinations: (1) the NDVI, (2) the green, red, and near-infrared spectral band combination, and (3) the seven-band combination. The remainder of the results for the five-class case of Site 1 concentrated on the NDVI, the green, red, and near-infrared spectral band combination, and the seven-band combination.

Using the silhouette index as a measure of cluster strength, NDVI was ranked first followed by the green, red, and near-infrared spectral band combination, and the seven-band combination consisting of the coastal blue, blue, green, yellow, red, red-edge, and near-infrared spectral bands. The NDVI had a strong cluster structure (silhouette index value between 0.71 and 1.0). The other two combinations were within the range identified as reasonable structure for clusters (silhouette index values of 0.51–0.70).

Summarized in Tables 4–6 are the contingency matrixes for the NDVI (Table 4), the three-band combination consisting of the green, red, and near-infrared bands (Table 5), and the seven-band combination containing the coastal blue, blue, green, yellow, red, red-edge, and near-infrared bands (Table 6), respectively, for Site 1. User's and producers accuracies ranged from 93.1% to 100%, from 96.7% to 100%, and from 93.3% to 100% for the NDVI, for the three-band combination consisting of the green, red, and near-infrared bands, and for the seven-band combination, respectively. Overall, the error matrix results were nearly equal for the aforementioned spectral data and indicated superb grouping of the data into the appropriate classes. However, the NDVI cluster structure was ten and twelve points greater (i.e., multiplying the silhouette

Table 1
Site 1 contingency matrixes and silhouette index values derived with the panchromatic, multispectral, and normalized difference vegetation index data as input into the clustering algorithm. Two-class case.

Band combination[a]		Reference data: Non-treated	Treated	Predicted total	User's accuracy (%)	Silhouette index
Pan	Cluster 1	30	0	30	100	0.88
	Cluster 2	0	29	29	100	
	Reference total	30	29			
	Producer's accuracy (%)	100	100			
B, G, R	Cluster 1	30	0	30	100	0.70
	Cluster 2	0	29	29	100	
	Reference total	30	29			
	Producer's accuracy (%)	100	100			
G, R, NIR	Cluster 1	30	0	30	100	0.91
	Cluster 2	0	29	29	100	
	Reference total	30	29			
	Producer's accuracy (%)	100	100			
Seven-band	Cluster 1	30	0	30	100	0.90
	Cluster 2	0	29	29	100	
	Reference total	30	29			
	Producer's accuracy (%)	100	100			
NDVI	Cluster 1	30	0	30	100	0.96
	Cluster 2	0	29	29	100	
	Reference Total	30	29			
	Producer's accuracy (%)	100	100			

[a] Pan – panchromatic, B – blue, G – green, R – red, NIR – near-infrared, and NDVI – normalized difference vegetation index. Seven-band – coastal blue, blue, green, yellow, red, red-edge, and near-infrared bands.

Table 2
Site 2 contingency matrixes and silhouette index values derived with the panchromatic, multispectral, and normalized difference vegetation index data as input into the clustering algorithm. Two-class case.

Band combination[a]		Reference data: Non-treated	Treated	Predicted total	User's accuracy (%)	Silhouette index
Pan	Cluster 1	27	9	36	75.0	0.68[b]
	Cluster 2	3	21	24	87.5	
	Reference total	30	30			
	Producer's accuracy (%)	90.0	70.0			
B, G, R	Cluster 1	25	0	25	100	0.60
	Cluster 2	5	30	35	85.7	
	Reference total	30	30			
	Producer's accuracy (%)	83.3	100			
G, R, NIR	Cluster 1	30	0	30	100	0.71
	Cluster 2	0	30	30	100	
	Reference total	30	30			
	Producer's accuracy (%)	100	100			
Seven-band	Cluster 1	30	0	30	100	0.69
	Cluster 2	0	30	30	100	
	Reference total	30	30			
	Producer's accuracy (%)	100	100			
NDVI	Cluster 1	30	0	30	100	0.88
	Cluster 2	0	30	30	100	
	Reference total	30	30			
	Producer's accuracy (%)	100	100			

[a] Pan – panchromatic, B – blue, G – green, R – red, NIR – near-infrared, and NDVI – normalized difference vegetation index. Seven-band – coastal blue, blue, green, yellow, red, red-edge, and near-infrared bands.
[b] The optimal number of clusters chosen by the algorithm for the panchromatic band was three; the silhouette index value shown in the table represents the three clusters. These clusters were merged into two clusters for accuracy comparisons.

index values by 100) than the respective three-band combination consisting of the green, red, and near-infrared bands (Table 5), and the seven-band combination, supporting it as the better tool for developing more compact clusters for the thermally defoliated data.

Tabulated in Table 3 are the numbers of clusters and the average silhouette index values obtained at Site 2 for each of the spectral datasets. Five clusters matching the five treatment classes were not obtained with none of the spectral datasets. Additionally, the silhouette index indicated that only the NDVI and the panchromatic data established clusters having a reasonable structure (silhouette index values of 0.51–0.70); whereas, the other spectral band combinations created clusters with weak and potentially artificial structures (silhouette index values of 0.26–0.50). Even though the criterion of the classes was not met, Site 2 results focused on the NDVI and panchromatic data because of their reasonable cluster structures.

The contingency matrixes obtained with the NDVI data for the five-class classification are presented in Table 7. Clusters one and two were combinations of the non-treated, N75-T25, and N50-T50

Table 3
The number of clusters and the average silhouette index derived from the spectral datasets evaluated for the five-class case at Sites 1 and 2.

Location	Band combination[a]	Number of clusters	Silhouette index
Site 1	Panchromatic	2	0.62
	B, G, R	6	0.38
	G, R, NIR	5	0.63
	Seven-band	5	0.61
	NDVI	5	0.73
Site 2	Panchromatic	3	0.61
	B, G, R	2	0.45
	G, R, NIR	3	0.46
	Seven-band	2	0.50
	NDVI	2	0.62

[a] B – blue, G – green, R – red, NIR – near-infrared, and NDVI – normalized difference vegetation index. Seven-band combination consisting of the coastal blue, blue, green, yellow, red, red-edge, and near-infrared bands.

cotton observations and the N25-T75 and thermally treated cotton observations, respectively. User's and producer's accuracies were within the range of 87–100%, indicating an excellent classification.

The three clusters derived for the panchromatic band were merged into two classes: (1) non-treated, N75-T25, and N50-T50 and (2) N25-T75 and treated (Table 8). User's and producer's accuracies ranged from 61.7% to 82%. Misclassifications were evident within the two clusters and were greater than the errors observed for clusters derived with the NDVI spectral data. These results support using the NDVI instead of the panchromatic data to assess thermal defoliation at Site 2.

4. Discussion

Based on the silhouette index, the NDVI produced the best final clustering in terms of compactness and separation of the clusters

Table 4
Site 1 contingency matrix derived with the normalized difference vegetation index five-class data as input into the clustering algorithm.

	Reference data						
Clusters	Non-treated	N75-T25[a]	N50-T50	N25-T75	Treated	Predicted total	User's accuracy (%)
Cluster 1	30	1	0	0	0	31	96.7
Cluster 2	0	28	2	0	0	30	93.3
Cluster 3	0	0	27	0	0	27	100
Cluster 4	0	0	0	29	0	29	100
Cluster 5	0	0	0	0	29	29	100
Reference total	30	29	29	29	29		
Producer's accuracy (%)	100	96.6	93.1	100	100		

[a] Non-treated = original plant canopies not subjected to thermal defoliation, N75-T25 = weighted average consisting of 75% non-treated spectra and 25% treated spectra, N50-T50 = weighted average consisting of 50% non-treated spectra and 50% treated spectra, N25-T75 = weighted average consisting of 25% non-treated spectra and 75% treated spectra, and Treated = original plant canopies subjected to thermal defoliation.

Table 5
Site 1 contingency matrix derived with the green, red, and near-infrared five-class data as input into the clustering algorithm.

	Reference data						
Clusters	Non-treated	N75-T25[a]	N50-T50	N25-T75	Predicted total	Total	User's accuracy (%)
Cluster 1	29	0	0	0	0	29	100
Cluster 2	1	28	0	0	0	29	96.6
Cluster 3	0	1	29	0	0	30	96.7
Cluster 4	0	0	0	29	0	29	100
Cluster 5	0	0	0	0	29	29	100
Reference total	30	29	29	29	29		
Producer's accuracy (%)	96.7	96.6	100	100	100		

[a] Non-treated = original plant canopies not subjected to thermal defoliation, N75-T25 = weighted average consisting of 75% non-treated spectra and 25% treated spectra, N50-T50 = weighted average consisting of 50% non-treated spectra and 50% treated spectra, N25-T75 = weighted average consisting of 25% non-treated spectra and 75% treated spectra, and Treated = original plant canopies subjected to thermal defoliation.

Table 6
Site 1 contingency matrix derived with the seven-band combination (coastal blue, blue, green, yellow, red, red-edge and near-infrared bands) five-class data as input into the clustering algorithm.

	Reference data						
Clusters	Non-treated	N75-T25[a]	N50-T50	N25-T75	Treated	Predicted total	User's accuracy (%)
Cluster 1	28	0	0	0	0	28	100
Cluster 2	2	28	0	0	0	30	93.3
Cluster 3	0	1	29	0	0	30	96.7
Cluster 4	0	0	0	29	0	29	100
Cluster 5	0	0	0	0	29	29	100
Reference total	30	29	29	29	29		
Producer's accuracy (%)	93.3	96.6	100	100	100		

[a] Non-treated = original plant canopies not subjected to thermal defoliation, N75-T25 = weighted average consisting of 75% non-treated spectra and 25% treated spectra, N50-T50 = weighted average consisting of 50% non-treated spectra and 50% treated spectra, N25-T75 = weighted average consisting of 25% non-treated spectra and 75% treated spectra, and Treated = original plant canopies subjected to thermal defoliation.

Table 7
Site 2 contingency matrix derived with the normalized difference vegetation index five-class data as input into the clustering algorithm.

	Reference data			
Clusters	Non-treated + N75-T25 + N50-T50[a]	N25-T75 + treated	Predicted total	User's accuracy (%)
Cluster 1	81	0	81	100
Cluster 2	9	60	69	87.0
Reference total	90	60		
Producer's accuracy (%)	90.0	100		

[a] Non-treated = original plant canopies not subjected to thermal defoliation, N75-T25 = weighted average consisting of 75% non-treated spectra and 25% treated spectra, N50-T50 = weighted average consisting of 50% non-treated spectra and 50% treated spectra, N25-T75 = weighted average consisting of 25% non-treated spectra and 75% treated spectra, and Treated = original plant canopies subjected to thermal defoliation.

Table 8
Site 2 contingency matrix derived with the panchromatic data as input into the clustering algorithm.

	Reference data			
Clusters	Non-treated + N75-T25 + N50-T50[a]	N25-T75 + treated	Predicted total	User's accuracy (%)
Cluster 1	74	23	97	76.3
Cluster 2	16	37	53	69.8
Reference total	90	60		
Producer's accuracy (%)	82.2	61.7		

[a] Non-treated = original plant canopies not subjected to thermal defoliation, N75-T25 = weighted average consisting of 75% non-treated spectra and 25% treated spectra, N50-T50 = weighted average consisting of 50% non-treated spectra and 50% treated spectra, N25-T75 = weighted average consisting of 25% non-treated spectra and 75% treated spectra, and treated = original plant canopies subjected to thermal defoliation.

(Tables 1–3) and resulted in the development of clusters matching the groups for the two-class case at both Sites and the five-class case at Site 1. User's and producer's accuracies for those NDVI datasets were greater than 93% (Tables 1, 2 and 4). At Site 2, the thermal treatment did not work as well in killing the leaves as at Site 1, leading to a mixture of dead and green leaves in the canopies. It is believed that this mixture affected the clustering algorithms ability to appropriately group the simulated spectra into individual classes. At this point, there has not been an accuracy standard established for dividing thermally defoliated fields into zones, but the results support using NDVI to complete the task. Other researchers have also indicated that vegetation indices are useful for monitoring defoliation of cotton (Ritchie and Bednarz, 2005; Yang et al., 2003, 2011). The study of Ritchie and Bednarz (2005) focused on models to assess defoliation and results indicated that red-edge vegetation indices worked well. That aspect was not evaluated in this study, but could be the focus of future research studies in which red-edge vegetation index data could be used as input into the clustering algorithm.

5. Perspective and conclusions

The automatic determination of clusters would be invaluable for the users of the remotely-sensed data since they would not need to guess the appropriate number of clusters. As with any unsupervised classification algorithm, the user has to visit the field to determine what land-cover features the clusters represents. The key to implementing remote sensing and the derivation of maps by clustering is to have a test strip that is not subjected to any treatment. Areas outside of the test strip that are grouped into clusters containing the test strip data could be evaluated further to determine the effectiveness of the thermal defoliation.

Within a control plot, variability in plant growth exists; therefore, a section of a control plot may have similar reflectance to a thermally treated area; especially if pests have damaged the plant. Nonetheless with the control strip and knowledge of how NDVI values relate to plant health, broadband spectral data and the clustering algorithm can be used as a tool to assess thermal defoliation of cotton fields. If time warrants, growers could opt to retreat areas in which the defoliation did not work well, or they could decide to use a precision harvesting technique. Barring major damage by natural disasters or major insect infestations, cotton plots usually have similar growth patterns from one year to the next. Data acquired from the previous year can, therefore, serve as a tool for treatment the following year. In areas where the treatment was not effective the previous year because of high biomass, the grower could opt to use a slower tractor speed so that the plants were exposed to the heat for a longer period of time, thus providing better leaf kill.

In this study, ground-based field data were evaluated to compare the spectra of thermally defoliated cotton canopies to non-treated cotton canopies. Digital brightness values and/or reflectance values recorded in a pixel by aerial and satellite sensors are affected more by leaf shape and orientation, background canopy reflectance, measurement geometry, and in-canopy shadowing (Asner, 1998). Additionally for thermally treated plants, the pixel value will be an integration of soil, background vegetation, and plant litter found underneath the plants. Furthermore, ground-based measurements do not account for atmospheric effects; nevertheless, the data do provide information related to the potential of spectral bands to separate the individual components and efficacy of using unsupervised classification in the form of cluster analysis to group the data. Findings support future endeavors to assess airborne and satellite-borne systems equipped with sensors sensitive to the wavelengths deemed useful in this study and unsupervised classification techniques that automatically determines the numbers of clusters to evaluate thermal defoliation of cotton fields.

Acknowledgments

The authors thank Veronica Abrigo, Fermin Alvarado, Raúl Cantú, Isabel Cavazos, Jaime Cavazos, Andy Cruz, Paul Delgado, Rene Davis, Armando Eldape, Jim Forward, Martín Galvan, Fred Gomez, Jaime Luna, Jaime Melendrez, and Kirk Zivkovich for their assistance in field data collection and Webushet Tadesse and Daniel Fisher for their review of this manuscript.

References

Anonymous, 2011. Cluster analysis. http://www.stat.berkeley.edu/classes/s133/Cluster2a.html.

Asner, G.P., 1998. Biophysical and biochemical sources of variability in canopy reflectance. Remote Sens. Environ. 64, 234–253.

Bundy, C., Funk, P., Steiner, R., 2006. Impact of Thermal Cotton Defoliation on Late-Season Insect Populations. In: Proceedings Beltwide Cotton Conference, Memphis, TN: National Cotton Council, pp. 1344–1351.

Campbell, J.B., 2002. Introduction to Remote Sensing, third ed. The Guilford Press, New York, NY.

Chang, F., Qiu, W., Zamar, R.H., Lazarus, R., Wang, X., 2010. Clues: an R package for nonparametric clustering based on local shrinking. J. Stat. Softw. 33 (4), 1–16.

Digital Globe, Inc. 2009. The benefits of the 8 spectral bands of WorldView-2. http://www.satimagingcorp.com/media/pdf/WorldView-2_8-Band_Applications_Whitepaper.pdf.

Eitel, J.U.H., Long, D.S., Gessler, P.E., Smith, A.M.S., 2007. Using in-situ measurements to evaluate the new RapidEye™ satellite series for prediction of wheat nitrogen status. Int. J. Remote Sens. 28, 4183–4190.

Eitel, J.U.H., Vierling, L.A., Litvak, M.E., Long, D.S., Schulthess, U., Ager, A.A., Krofcheck, D.J., Stoscheck, L., 2011. Broadband, red-edge information from satellites improves early stress detection in a New Mexico conifer woodland. Remote Sens. Environ. 115, 3640–3646.

Fletcher, R.S., Showler, A., Funk, P.A., 2007. Surveying thermally-defoliated cotton plots with CIR aerial photography. Crop Management. http://www.plantmanagementnetwork.org/pub/cm/research/2007/thermal/.

Frigui, H., Krishnapuram, R., 1999. A robust competitive clustering algorithm with applications in computer vision. IEEE Trans. Pattern Anal. Mach. Intell. 21, 450–465.

Funk, P.A., Armijo, C.B., McAlister, D.D., Lewis, B.E., 2004. Experimental thermal defoliator trials. J. Cotton Sci. 8, 230–242.

Funk, P.A., Armijo, C.B., Showler, A.T., Fletcher, R.S., Brashears, A.D., McAlister III, D.D., 2006. Cotton harvest preparation using thermal energy. Trans. ASABE. 49 (3), 617–622.

Funk, P.A., Armijo, C.B., Hawkes, G.M., Libbin, J.D., 2012. Cotton thermal defoliation economics. J. Am. Soc. Farm Managers Rural Appraisers 75 (1), 29–42.

Jones, H.G., Vaughan, R.A., 2010. Remote Sensing of Vegetation: Principles, Techniques, and Applications. Oxford University Press, New York, NY.

Mather, P.M., 2005. Computer Processing of Remotely-Sensed Images: An Introduction. John Wiley and Sons Ltd, England.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 2007. Numerical Recipes, third ed. Cambridge University Press.

R Core Team, 2012. R: a language and environment for statistical computing. R Foundation for Statistical Computing: Vienna, Austria, ISBN 3-900051-07-0. http://www.R-project.org/.

Ritchie, G.L., Bednarz, C.W., 2005. Estimating defoliation of two distinct cotton types using reflectance data. J. Cotton Sci. 9, 182–188.

Rouse, J.W., Haas, R.H., Schell, J.A., Deering, D.W., 1973. Monitoring vegetation systems in the great plains with ERTS, Third ERTS Symposium, NASA SP-351 I, pp. 309–317.

Savitzky, A., Golay, M.J.E., 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. 36, 1627–1639.

Showler, A., Funk, P., Armijo, C., 2006. Effect of thermal defoliation on cotton leaf desiccation, senescence, post-harvest regrowth, and lint quality. J. Cotton Sci. 10, 39–45.

Smith, A.M.S., Eitel, J.U.H., Hudak, A.T., 2010. Spectral analysis of charcoal on soils: implications for wildland fire severity mapping methods. Int. J. Wildland Fire 19, 976–983.

Tibshirani, R., Walther, G., Hastie, T., 2000. Estimating the number of clusters in a data set via the gap statistic. J. R. Statist. Soc. 63, 411–423.

Trigg, S., Flasse, S., 2000. Characterizing the spectral-temporal response of burned savannah using in situ spectroradiometry and infrared thermometry. Int. J. Remote Sens. 21, 3161–3168.

Wang, X., Qiu, W., Zamar, R.H., 2007. An iterative non-parametric clustering algorithm based on local shrinking. Comput. Stat. Data An. 52, 286–298.

Yang, C., Greenburg, S.M., Everitt, J.H., Sappington, T.W., Norman Jr., J.W., 2003. Evaluation of cotton defoliation strategies using airborne multispectral imagery. Trans. ASAE 46 (3), 869–876.

Yang, C., Greenburg, S.M., Everitt, J.H., Fernandez, C.L., 2011. Assessing cotton defoliation, regrowth control and root rot infection using remote sensing technology. IJABE 4 (4), 1–11.

Zhung, X., Huang, Y., Palaniappan, K., Lee, J.S., 1996. Gaussian mixture modeling, decomposition and applications. IEEE Trans. Signal Process. 5, 1293–1302.