American Journal of Plant Sciences, 2015, 6, 3193-3204
Published Online December 2015 in SciRes. http://www.scirp.org/journal/ajps
http://dx.doi.org/10.4236/ajps.2015.619311

Testing Leaf Multispectral Reflectance Data as Input into Random Forest to Differentiate Velvetleaf from Soybean

Reginald S. Fletcher

United States Department of Agriculture, Agricultural Research Service, Crop Production Systems Research Unit, Stoneville, USA
Email: reginald.fletcher@ars.usda.gov

Received 22 September 2015; accepted 11 December 2015; published 14 December 2015

Abstract

Velvetleaf (*Abutilon theophrasti* Medic.) infestations negatively impact row crop production throughout the United States and Canada's eastern provinces. To implement management strategies to control velvetleaf, managers need tools for differentiating it from crop plants. 5 Band, 7 Band, 8 Band, and 16 Band multispectral datasets simulating LANDSAT 3 plus a blue band, LANDSAT 8, WorldView 2, and WorldView 3 spectral bands, respectively were tested as input into the random forest algorithm for velvetleaf soybean [*Glycine max* L. (Merr.)] discrimination. During two separate greenhouse experiments in 2014, leaf reflectance measurements were obtained at the vegetative growth stage of velvetleaf plants and two soybean varieties. The reflectance measurements were collected with a plant contact probe attached to a hyperspectral spectroradiometer. Leaf hyperspectral reflectance measurements were convolved to the four multispectral datasets with computer software. Overall, user's, and producer's accuracies and kappa coefficient were employed to determine classification accuracies. Using the multispectral datasets as input, the random forest algorithm differentiated velvetleaf from the soybean varieties with accuracies ranging from 86.7% to 100%. 7 Band, 16 Band, 8 Band, and 5 Band datasets ranked or tied for the highest accuracies seventeen, sixteen, twelve, and one time, respectively. Kappa coefficients indicated an almost perfect agreement (*i.e.*, kappa value, 0.81 - 1.0) to substantial agreement (*i.e.*, kappa value, 0.61 - 0.80) between reference data and model predicted classes. This study was the first to demonstrate the application of the random forest machine learner and leaf multispectral reflectance data as tools to distinguish velvetleaf from soybean and to identify multispectral band combinations providing the best accuracies. Findings support further application of the random forest machine learner along with remotely-sensed multispectral data as tools for velvetleaf soybean discrimination with future implications for site-specific management of velvetleaf.

How to cite this paper: Fletcher, R.S. (2015) Testing Leaf Multispectral Reflectance Data as Input into Random Forest to Differentiate Velvetleaf from Soybean. *American Journal of Plant Sciences*, **6**, 3193-3204.
http://dx.doi.org/10.4236/ajps.2015.619311

Keywords

***Glycine max*, *Abutilon theophrasti*, Machine Learning, Supervised Classification, Ensemble Technique**

1. Introduction

Velvetleaf (*Abutilon theophrasti* Medic.), a broadleaf plant native to China, was introduced into the United States from India as a fiber crop. It escaped cultivation and now has become a problem weed in row crops, especially in corn (*Zea mays* L.) and soybean (*Glycine max* (L.) Merr.) fields throughout the United States and Canada's eastern provinces. The summer annual weed grows to heights ranging from 0.3 to 2.0 m. The plant reproduces from seed and can develop up to 17,000 seeds that may remain viable for up to sixty years. Velvetleaf grows best in warm regions and invades vacant lots, gardens, and cultivated fields. Once established, it is a problem weed for many years to come.

Velvetleaf infestations negatively impact a crop and field in several ways. Velvetleaf plants emerging before or at the same time as crop plants are highly competitive for water and plant nutrients and thus can outgrow the crop. A 25% decrease in crop yield can occur when the velvetleaf plant population is equivalent to 1 plant per 30 cm [1]. Seeds, adult plants, and decaying plant parts contain or produce allelopathic (toxic) chemicals that inhibit water uptake and chlorophyll production of some crop plants, particularly soybean, thus preventing growth. The chemicals enter the soil during rain events.

Producers commonly use preemergence and postemergence measures to manage or control velvetleaf infestations. Detecting and eliminating the plant before seeding is vital because of the long-term dormancy of the seeds and the future problems they may cause. Therefore, field managers need additional techniques besides the common field survey for detecting velvetleaf infestation in crop fields.

Remote sensing technology has gained popularity as a tool for weed detection in agricultural systems [2]-[8]. The technology involves using ground, airborne, or satellite-borne sensors to obtain light reflectance measurements of plant leaves and canopies to differentiate between weed and crop plants. Detecting weeds with remote sensing technologies requires that differences in spectral reflectance exist between weeds and their environment and that the spatial and spectral resolution of remote sensing equipment is sufficient to detect these differences [9].

Soybean weed discrimination has been the focus of several remote sensing studies including velvetleaf as one of the weeds of interest. Reference [10] determined from statistical analysis of multispectral data spanning the visible to near infrared region of the light spectrum that weed-free soybean plots could be distinguished from soybean plus velvetleaf plots, soybean plus mixed weed plots, and soybean plus grass plots. The separation only occurred with red/infrared ratios. None of the soybean plus weed plots, however, could be distinguished from each other with single bands or red/infrared ratios.

Reference [11] obtained mixed results differentiating soybean from velvetleaf and foxtail (*Setaria faberi* Herrm.) in a controlled experiment. At one study site, they reported a classification error less than 17% for the weeds; at the other study site they achieved classification errors of 17% and 39% for foxtail and velvetleaf, respectively. Their study focused on using airborne multispectral imagery collected within the visible green (520 to 600 nm), visible red (630 to 690 nm), and near infrared (760 to 900 nm) wavebands. They concluded that if weed differentiation was not an issue for the weed management program then remote sensing techniques have good potential to differentiate weeds from crops.

Reference [4] demonstrated that a single decision tree approach based on the classification and regression technique could use vegetation indices as input to discriminate between corn, corn and velvetleaf, corn and a mixture of various grass species, corn and mixture of random predominant weed species, soybean, soybean and velvetleaf, soybean and mixture of various grass species, velvetleaf, mixed grass, and mixtures of random predominant weed species. The classification success rate was 85% ± 6%. The study focused on using twenty-four narrowband multispectral bands within the visible and near infrared regions of the light spectrum. From those bands, sixty-five normalized difference vegetation index bands were created and were used as input for classification. Accurate results were achieved; however, the sample size was small at only three plots per treatment.

Based on the above studies, more information is needed on the potential of using remote sensing technology for soybean weed discrimination, especially in the case of velvetleaf. Currently, information is lacking on the comparison of multispectral systems wavebands for soybean velvetleaf discrimination. Additionally, no information exists on including shortwave infrared spectral data to discriminate soybeans from velvetleaf. The shortwave infrared region of the light spectrum (1300 - 2500 nm) is sensitive to the water content in plants [12]. Finally, no information is available on the role that soybean variety may play in differentiating it from velvetleaf.

Another key aspect of using remote sensing technology is the computer algorithm employed to process the data. The success or failure of using the technology is affected by the algorithm selected to analyze the data. In this study, it is proposed to use the random forest machine learner for soybean velvetleaf discrimination. Random forest has gained popularity as a tool to use for classification problems because it is fully automated, and users have the ability to design powerful models with little experience in using the machine learner. Random forest has been ranked as one of the best learners to employ for classification and regression problems [13]. Researchers have successfully used it in genetics, clinical medicine, bioinformatics, agriculture, and remote sensing applications [14]-[17].

Random forest is an ensemble learning method based on the principle that a group of "weak learners" can come together to develop a "strong learner" [18]. Thus, it uses multiple decision trees to make a consensus prediction, hence the name random forest. Each decision tree in the "so-called forest" is derived from a bootstrap sample (*i.e.*, a percentage of the original data is selected for training, and the non-selected data are used for testing) of the original data (sampling with replacement). The splitting of each tree node is determined by the Gini criterion (*i.e.*, a measurement of node purity). For the splitting process, the algorithm selects a subset of the predictor variables at each node and then the best-splitting variable is chosen from that subset. Samples not selected in the bootstrap process for a tree (*i.e.*, approximately 36.8% of the original samples), known as "out-of-bag" (OOB) samples, are used to test the accuracy of the classifier. Random forest assigns an OOB sample to a class by using the decision trees in which the sample was OOB. The votes of each tree are tallied, and the OOB sample is assigned to the class receiving the largest votes. Compared with other machine learners, the random forest algorithm does not need an independent test set because the OOB samples serve as the test set [18]. Random forest also provides a variable importance reading representing the importance of each predictor variable to the model.

Currently, no information is available on using leaf multispectral reflectance data as input into random forest for soybean velvetleaf discrimination. The objective of this investigation was to evaluate leaf multispectral reflectance data as input into the random forest machine learner to differentiate velvetleaf from soybean. Specifically, the study focused on evaluating multispectral data mimicking the spectral bands of satellite sensors to discriminate the velvetleaf from two soybean varieties. Spectral bands of satellite sensors were chosen because the bands are strategically placed in different regions of the light spectrum for land cover mapping, thus providing different spectral band combinations for the model to test for separating velvetleaf from soybean.

2. Materials and Methods

2.1. Plant Descriptions

Two Progeny (P) brand LibertyLink (LL) soybean varieties (P4928LL and P5460LL, Progeny Ag Products, Wynne, Arkansas) and non-glyphosate resistant velvetleaf (United States Department of Agriculture, Agricultural Research Service, Stoneville, MS) were grown for the study. All three plants are characterized as pubescent plants, consisting of gray, light tawny, and white hairs for soybean P4928LL, soybean P5460LL, and velvetleaf, respectively. Soybean P4928LL is characterized as having an indeterminate growth habit (*i.e.*, a continuation of vegetative growth after flowering) and soybean P5460LL as having a determinate growth habit (*i.e.*, vegetative growth completed prior to flowering). The maturity group assigned to soybean P4928LL and soybean P5460LL are 4.9 and 5.4, respectively.

2.2. Greenhouse Experiment

The study was conducted at the United States Department of Agriculture, Agricultural Research Service, Stoneville, MS facility. Data were collected from two separate greenhouse experiments initiated on June 13, 2014, and

August 28, 2014. Soybean and velvetleaf seeds were sown in plugs containing commercial potting mix (Pro-Mix, Ultimate Potting Mix, Quakertown, Pennsylvania). Ten days after germination, thirty plants of each soybean variety and weed species were transplanted to individual 1 L pots filled with the commercial potting mix. Plants were watered at three- to four-day intervals. The potting mix consisted of a slow release nitrogen, phosphorus, and potassium fertilizer. The plants were grown at a temperature and photoperiod of 26.6˚C and 14-h, respectively.

2.3. Data Collection

Leaf reflectance measurements were obtained with a full range hyperspectral spectroradiometer (FieldSpec 3, PANalytical Boulder, Boulder, CO). The instrument's fiber optic was attached to a plant probe (PANalytical Boulder, Boulder, CO) equipped with a light source. The plant probe has a 1 cm field of view. A leaf clip (PANalytical Boulder, Boulder, CO) was fastened to the contact probe. This device has a trigger lock/release gripping system designed to hold the leaf in place without removing it from the plant or causing damage to the plant. The leaf clip has a two-sided rotating head. One side of the head contains a black panel face, and the other side has a white panel face. The black and white panels are ideal for reflectance and transmittance measurements, respectively. The former was employed in this study.

The spectroradiometer obtained continuous spectra in the range of 350 - 2500 nm. Its sampling interval and spectral resolution were 1.4 nm and 3 nm, respectively, within the 350 nm to 1000 nm spectral range. The sampling interval and spectral resolution were 2 nm and 10 nm, respectively, within the 1000 nm to 2500 nm spectral range. The proprietary software operating the instrument resampled the reflectance data to 1 nm wavelengths.

Reflectance measurements were collected from the most recently matured leaf of each plant. Soybean has a trifoliate leaf, therefore, the center leaflet of the most recently matured leaf was chosen for data collection. At the selected sample spot of each plant leaf, reflectance measurements were an average of fifteen readings. Leaf reflectance measurements were obtained on June 30, 2014, and September 17, 2014, for the first and second experiments, respectively. For velvetleaf, it is important to identify and treat the plant prior to seeding. Measurements were obtained for all plants during the vegetative growth stage. The instrument was calibrated with a white spectralon panel (PANalytical Boulder, Boulder, CO) at 15-minute intervals.

2.4. Development of Multispectral Bands

The hyperspectral reflectance measurements of the soybean and velvetleaf leaves were resampled to four multispectral datasets (**Table 1**), referred to as 5 Band, 7 Band, 8 Band, and 16 Band. The green, red, and near infrared bands of the 5 Band dataset were replicates of the green, red, and near infrared bands obtained by LANDSAT 1 - 4 multispectral scanners [19]. The blue band was added to the 5 Band dataset to represent a broad region of the blue spectrum. Also, it represented blue light reflectance data obtained by many commercial handheld cameras. The 7 Band, 8 Band, and 16 Band datasets simulated the spectral bands of the LANDSAT 8 Operational Land Imager, WorldView 2 sensors, and WorldView 3 sensors, respectively [19]-[21]. The datasets were unique because they represented different regions of the light spectrum and provided spectral resolutions ranging from 20 to 300 nm. The multispectral bands were created by resampling the original hyperspectral bands using a Gaussian distribution function and the lower and the upper bounds of each satellite sensor spectral bands. The resampled spectral data were created with the hsdar package [22] of the R software [R version 3.2.0 (April 16, 2015) Full of Ingredients].

2.5. Classification Model Development

The conditional inference version of random forest (cforest) was used to create the models evaluated in this study. Reference [15] recommended using cforest instead of the original version of random forest if the prediction variables were highly correlated. Some of the variables were highly correlated in each dataset, thus justifying the use of cforest for model creation. Strong correlation between variables biases the variable importance rankings provided by random forest for classification or regression problems. Cforest implementation of random forest was designed to better handle correlation among variables, thus providing more accurate and unbiased rankings of the variable of importance [15]. The cforest technique utilizes conditional inference trees as base learners, in contrast to random forest, which employs classification and regression trees as base learners [15] [23]

Table 1. Spectral band descriptions and wavelengths of the multispectral datasets used in this study.

Spectral Band	Wavelengths of Each Dataset			
	5 Band[a]	7 Band	8 Band	16 Band
Coastal		430 - 450 nm	400 - 450 nm	400 - 450 nm
Blue	400 - 500 nm	450 - 510 nm	450 - 510 nm	450 - 510 nm
Green	500 - 600 nm	530 - 590 nm	510 - 580 nm	510 - 580 nm
Yellow			585 - 625 nm	585 - 625 nm
Red	600 - 700 nm	640 - 670 nm	630 - 690 nm	630 - 690 nm
Red-edge			705 - 745 nm	705 - 745 nm
Near infrared 1	700 - 800 nm	850 - 880 nm	770 - 895 nm	770 - 895 nm
Near infrared 2	800 - 1100 nm		860 - 1040 nm	860 - 1040 nm
Shortwave infrared 1		1570 - 1650 nm		1195 - 1225 nm
Shortwave infrared 2		2110 - 2290 nm		1550 - 1590 nm
Shortwave infrared 3				1640 - 1680 nm
Shortwave infrared 4				1710 - 1750 nm
Shortwave infrared 5				2145 - 2185 nm
Shortwave infrared 6				2185 - 2225 nm
Shortwave infrared 7				2235 - 2285 nm
Shortwave infrared 8				2295 - 2365 nm

[a]5 Band-simulates LANDSAT 3 spectral bands plus an additional blue band, 7 Band-simulates LANDSAT 8 spectral bands, 8 Band-simulates WorldView 2 spectral bands, and 16 Band-simulates WorldView 3 spectral bands.

[24]. Furthermore, instead of using bootstrap samples to construct its decision trees, cforest utilizes subsampling without replacement for constructing unbiased decision trees for the forest. Finally, the cforest algorithm uses the conditional permutation scheme described by [15] to determine the variable of importance ranking.

The number of samples to evaluate at each split of the tree (*mtry*) and the number of trees to use for creating the model (*ntree*) were the two parameters needed to be set before completing the classification. For this study, the default *mtry* value of 5 was used for each dataset. The default *ntree* value of 500 was employed as the starting point and was adjusted accordingly to obtain consistent variable importance rankings.

The following procedure was used to test the robustness of the models relative to variable importance [15]. A model was created using the default *mtry* and *ntree* values, the variable importance rankings were tabulated, and then the model was rerun using the same *mtry* and *ntree* values and a different starting seed (*i.e.*, the random generator used as a starting point for sampling). The model parameters were accepted if the variable importance ranking was similar between the first and the second runs. If the variable importance rankings were not consistent between runs, then the *ntree* value was increased by 1000, and the model was retested using the same *mtry* and seed values. This process was continued until a stable variable importance ranking was obtained.

2.6. Accuracy Assessment

Classification accuracies of the selected models were determined by evaluating the user's, producer's, and overall accuracies and kappa coefficient [24]. User's accuracy represents the percentage of predicted samples classified correctly. Producer's accuracy characterizes the percentage of reference samples correctly identified. The overall accuracy is a measure of the total number of correctly classified samples divided by the total number of samples. The kappa coefficient quantifies the variation between the observed agreement of the reference data and predicted data and the chance agreement between the two. The accuracy values were tabulated from the "out of bag" samples, those samples not used to train the model. Model development and evaluation were determined with the party package of the R software [25]-[27].

3. Results

3.1. Accuracy Assessment

The accuracy assessment results of the random forest classification for the June 30, 2014, dataset are summarized in **Table 2** for the velvetleaf soybean P4928LL classification. Overall, user's, and producer's accuracies greater than 90% were achieved for all of the multispectral datasets. The highest overall classification accuracy of 96.7% was obtained with the 7 Band, 8 Band, and 16 Band datasets; the lowest overall classification accuracy of 95% occurred for the 5 Band dataset. The same ranking order of the datasets was observed for the kappa coefficients (**Table 2**). The user's and producer's accuracy ranged from 93.3% to 100%. For the velvetleaf class, a tie occurred between the 7 Band and the 16 Band multispectral datasets for the highest user's accuracy; whereas, the 8 Band dataset ranked best in the producer's accuracy (**Table 2**). The 8 Band, and the 7 Band and 16 Band multispectral datasets achieved the greatest user's and producer's accuracies, respectively, for the soybean P4928LL class (**Table 2**).

The random forest classification results of the velvetleaf soybean P4928LL classes are tabulated in **Table 2** for the September 17, 2014, multispectral datasets. The 7 Band dataset obtained the highest measurement accuracies with 93.3%, 0.867, 90.6%, 96.4%, 96.7%, and 90.0% for overall accuracy, kappa coefficient, velvetleaf user's accuracy, soybean P4928LL user's accuracy, velvetleaf producer's accuracy, and soybean P4928LL producer's accuracy, respectively. The other multispectral datasets were tied for second in the measurement accuracies.

Overall, user's, and producer's accuracies and the kappa coefficients are presented in **Table 3** for the June 30, 2014, velvetleaf soybean P5460LL classification. The 7 Band, 8 Band, and 16 Band datasets ranked best in all accuracy categories. Their user's, producer's, and overall accuracies ranged from 96.7% to 100%, and the kappa coefficients were 0.967. The 5 Band dataset obtained the lowest accuracies, with user's, producer's, and overall accuracies ranging from 93.5% to 96.7%. The kappa value was 0.9.

The September 17, 2014, dataset for the velvetleaf soybean P5460LL classification indicated that the 16 Band dataset model was ranked or tied for first in all of the accuracy categories (**Table 3**). The 7 Band and 8 Band dataset models were tied for first for the soybean P5460LL producer's accuracy. They obtained the second highest accuracies for the other categories. The 5 Band dataset ranked last in all of the accuracy assessment categories.

Table 2. Accuracy assessment of the velvetleaf versus soybean P4928LL classification based on leaf multispectral data input into the random forest classifier.

Classification	Date	Accuracy Measurement	Multispectral Dataset[a]			
			5 Band	7 Band	8 Band	16 Band
Velvetleaf-soybean P4928LL	June 30, 2014	User's accuracy velvetleaf	93.5%	96.7%	93.8%	96.7%
		User's accuracy soybean P4928LL	96.6%	96.7%	100%	96.7%
		Producer's accuracy velvetleaf	96.7%	96.7%	100%	96.7%
		Producer's accuracy soybean P4928LL	93.3%	96.7%	93.3%	96.7%
		Overall accuracy	95.0%	96.7%	96.7%	96.7%
		Kappa coefficient	0.900	0.933	0.933	0.933
Velvetleaf-soybean P4928LL	September 17, 2014	User's accuracy velvetleaf	90.3%	90.6%	90.3%	90.3%
		User's accuracy soybean P4928LL	93.1%	96.4%	93.1%	93.1%
		Producer's accuracy velvetleaf	93.3%	96.7%	93.3%	93.3%
		Producer's accuracy soybean P4928LL	90.0%	90.0%	90.0%	90.0%
		Overall accuracy	91.7%	93.3%	91.7%	91.7%
		Kappa coefficient	0.833	0.867	0.833	0.833

[a]Refer to **Table 1** for the spectral band designations of the multispectral datasets.

Table 3. Accuracy assessment of the velvetleaf versus soybean P5460LL classification based on leaf multispectral data input into the random forest classifier.

Classification	Date	Accuracy Measurement	Multispectral Dataset[a]			
			5 Band	7 Band	8 Band	16 Band
Velvetleaf-soybean P5460LL	June 30, 2014	User's accuracy velvetleaf	93.5%	96.7%	96.7%	96.7%
		User's accuracy soybean P5460LL	96.6%	100%	100%	100%
		Producer's accuracy velvetleaf	96.7%	100%	100%	100%
		Producer's accuracy soybean P5460LL	93.3%	96.7%	96.7%	96.7%
		Overall accuracy	95.0%	98.3%	98.3%	98.3%
		Kappa coefficient	0.900	0.967	0.967	0.967
Velvetleaf-soybean P5460LL	September 17, 2014	User's accuracy velvetleaf	87.1%	93.1%	93.1%	93.3%
		User's accuracy soybean P5460LL	89.7%	90.3%	90.3%	93.3%
		Producer's accuracy velvetleaf	90.0%	90.0%	90.0%	93.3%
		Producer's accuracy soybean P5460LL	86.7%	93.3%	93.3%	93.3%
		Overall accuracy	88.3%	91.7%	91.7%	93.3%
		Kappa coefficient	0.767	0.833	0.833	0.867

[a]Refer to **Table 1** for the spectral band designations of the multispectral datasets.

3.2. Model Parameters

For fourteen out of the sixteen classification models, the default *mtry* and *ntree* values were adequate for obtaining stable variable importance readings (**Table 4**). The two exceptions were the random forest models used to complete the velvetleaf soybean P4928LL and the velvetleaf soybean 5460LL classifications based on the 8 Band and 16 Band datasets, respectively, for September 17, 2014. Three thousand five-hundred and 4500 trees were used to complete the classifications of the former and latter, respectively.

3.3. Variable Importance

The variable importance rankings of the random forest models used for the June 30, 2014, velvetleaf versus soybean P4928LL were as follows (**Figure 1**). The green (G) and near infrared two (NIR2) spectral bands were relevant to the model and had similar variable importance scores for the 5 Band dataset. The NIR1, G, and shortwave infrared one (SWIR1) spectral bands were important to the 7 Band dataset model while noticeable differences occurred in their variable importance scores. NIR1 and 2, G, and yellow (Y) spectral bands were needed by the model for the 8 Band dataset; the NIR2 and G spectral bands had similar variable importance scores and appeared in the top tier of variable importance scores. The NIR1 and Y spectral bands had variable importance scores similar to each other and appeared in the second tier of variable importance scores. SWIR1 to 4, NIR1 and 2, G, and Y spectral bands were the most important variables in the 16 Band dataset model. The spectral bands were grouped into six tiers: tier one-SWIR1, tier two-NIR2, tier three-G, tier four-NIR1 and Y, tier five-SWIR3, and tier six-SWIR2 and 4.

Variable importance rankings of the random forest models are shown in **Figure 2** for the September 17, 2014, velvetleaf versus soybean P4928LL classification. The G spectral band was the most important variable in the 5 Band dataset model. The G and blue (B) spectral bands were needed by 7 Band dataset model, with the G band ranked best. The G, Y, and B spectral bands were ranked most useful by the random forest model for the classification with the 8 Band dataset. Noticeable differences occurred in their importance scores. Eight spectral bands including G, Y, B, NIR1 and 2, red (R), SWIR1, and coastal (CA) were ranked important to the 16 Band dataset model. The G spectral band ranked first followed by the Y, B, NIR2, R, NIR1, SWIR1, and CA spectral bands.

Figure 3 illustrates the variable importance rankings of random forest models used in the classification of the June 30, 2014, velvetleaf and soybean P5460LL classes. The G and NIR2 spectral bands were ranked most im-

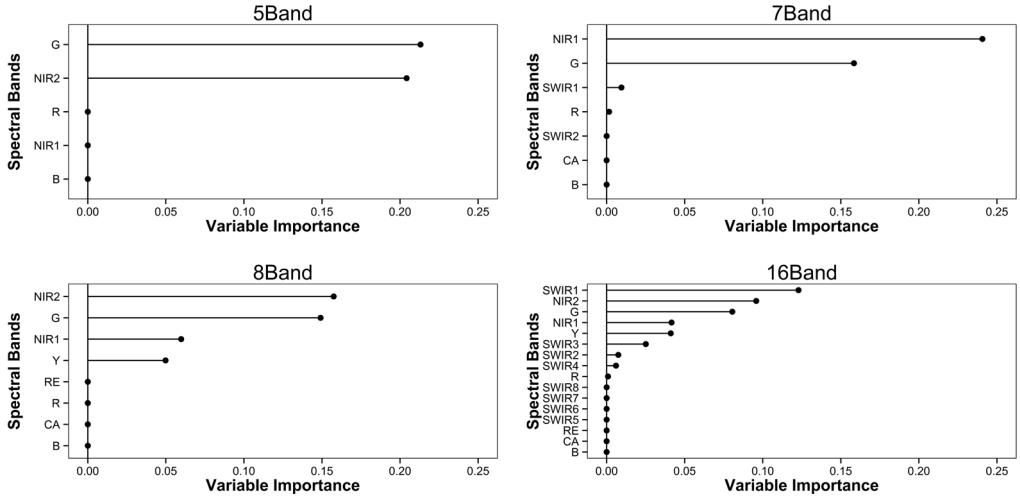

Figure 1. Variable importance rankings per multispectral dataset derived by the random forest model used for the velvetleaf and soybean P4928LL classification, June 30, 2014. CA = coastal, B = blue, G = green, Y = yellow, R = red, RE = red-edge, NIR = near infrared, and SWIR = shortwave infrared.

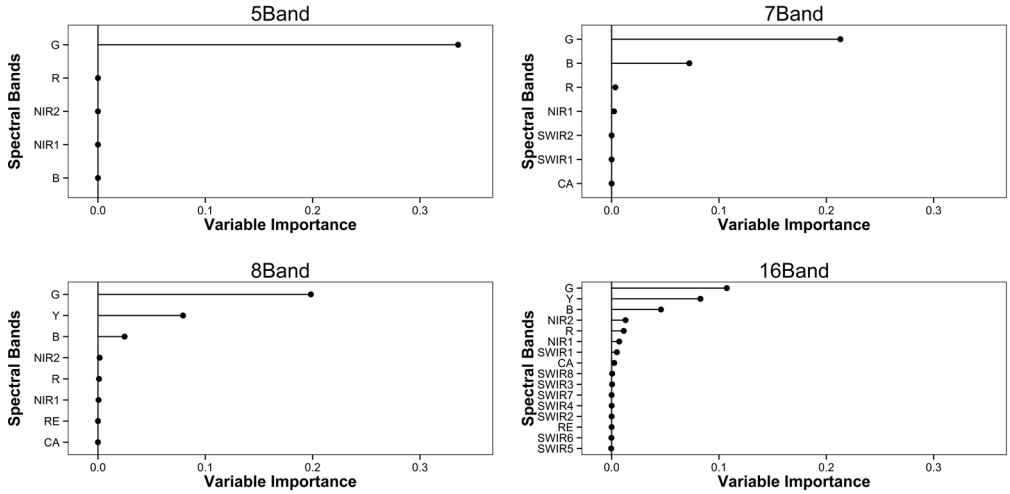

Figure 2. Variable importance rankings per multispectral dataset derived by the random forest model used for the velvetleaf and soybean P4928LL classification, September 17, 2014. CA = coastal, B = blue, G = green, Y = yellow, R = red, RE = red-edge, NIR = near infrared, and SWIR = shortwave infrared.

Table 4. Random forest model parameters used with the multispectral datasets to distinguish velvetleaf from two soybean varieties.

Classification	Dataset[a]	*mtry*[b]	*Ntrees* (June 30, 2014)	*Ntrees* (September 17, 2014)
Velvetleaf-soybean P4928LL	5 Band	5	500	500
Velvetleaf-soybean P4928LL	7 Band	5	500	500
Velvetleaf-soybean P4928LL	8 Band	5	500	3500
Velvetleaf-soybean P4928LL	16 Band	5	500	500
Velvetleaf-soybean P5460LL	5 Band	5	500	500
Velvetleaf-soybean P5460LL	7 Band	5	500	500
Velvetleaf-soybean P5460LL	8 Band	5	500	500
Velvetleaf-soybean P5460LL	16 Band	5	500	4500

[a]Refer to **Table 1** for the spectral band designations of the multispectral datasets. [b]*mtry* = number of randomly preselected variables; *ntrees* = number of trees used in the classification.

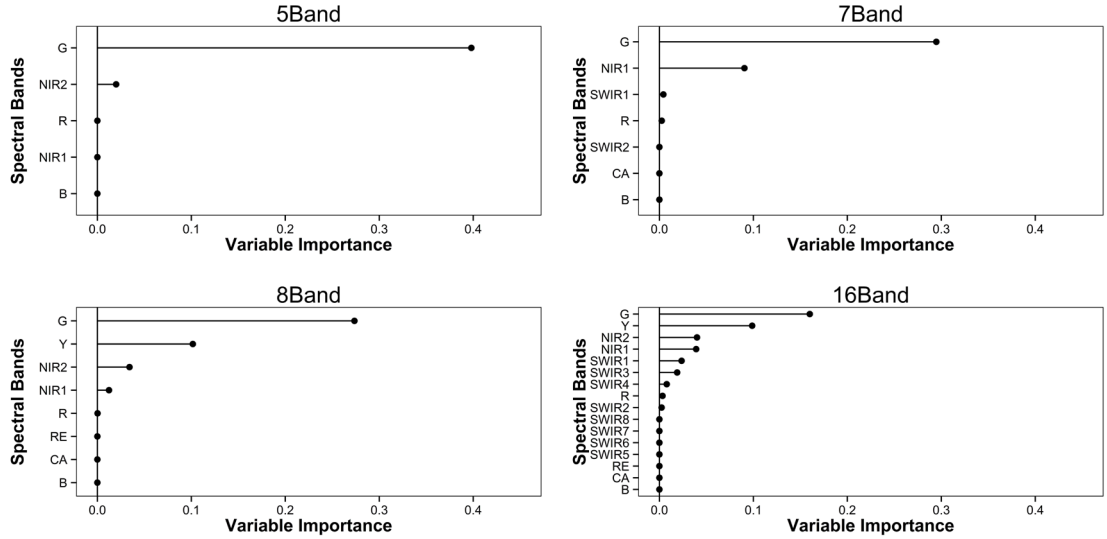

Figure 3. Variable importance rankings per multispectral dataset derived by the random forest model used for the velvetleaf and soybean P5460LL classification, June 30, 2014. CA = coastal, B = blue, G = green, Y = yellow, R = red, RE = red-edge, NIR = near infrared, and SWIR = shortwave infrared.

portant to the model for the 5 Band multispectral dataset. Distinct differences were observed in the scores, with the G band ranked the most important. Essential spectral bands for the 7 Band dataset model in descending order were G, NIR1, SWIR1, and R. The 8 Band dataset random forest model selected the G, Y, and NIR1 and 2 spectral bands as valuable variables for the classification; the rankings appeared in four distinct tiers: tier one-G, tier two-Y, tier three-NIR2, and tier four-NIR1. Five class tiers was observed for the most important rankings for the 16 Band dataset including the G spectral band in tier one, the Y spectral band in tier two, NIR1 and 2 spectral bands in tier three, SWIR spectral bands one and three in tier four, and the SWIR4 band in tier five.

Variable importance scores of the random forest models are shown in **Figure 4** for the September 17, 2014, velvetleaf P5460LL classification. The NIR2 and G spectral bands were the most useful to the model when using the 5 Band dataset, and their scores were nearly identical. The spectral bands critical to the classification model using the 7 Band dataset were as follows in descending order: NIR1, G, and B. There was an obvious difference in the variable importance scores. Four spectral bands were relevant to the model using the 8 Band dataset: NIR1, NIR2, G, and Y. NIR1, NIR2 and G, and Y spectral bands appeared in the first, second, and third tiers of the rankings, respectively. Eight spectral bands were relevant to the model using the 16 Band dataset, and their rankings in descending order were NIR2, NIR1, G, Y, SWIR1, B, SWIR8, and SWIR7.

4. Discussion

The objective of this study was to evaluate leaf multispectral reflectance data as input into the random forest classification algorithm to differentiate soybean from velvetleaf, an invasive weed affecting soybean production throughout the United States and eastern provinces of Canada. The study emphasized using different multispectral band combinations as input into the algorithm to differentiate velvetleaf from two different soybean varieties. The algorithm achieved overall, user's, and producer's accuracies that were greater than 85% for velvetleaf soybean discrimination (**Table 2** and **Table 3**), which was comparable to soybean weed discrimination studies using statistical methods [11] and single decision trees [4] to classify airborne imagery. Kappa values indicated that an almost perfect agreement (*i.e.*, kappa value range 0.81 - 1.0) to substantial agreement (*i.e.*, kappa value range 0.61 - 0.80) occurred between the reference data and predicted data (**Table 2** and **Table 3**). The latter was observed only for the 5 Band dataset for the velvetleaf soybean P5460LL classification occurring on September 17, 2014.

Generally, for all the datasets, the G and NIR spectral bands were ranked as important variables to the models for discriminating velvetleaf from soybean. Plant leaf reflectance and absorption of green light are influenced by leaf chlorophyll content [12], and may have been responsible for soybean velvetleaf differentiation. The inter-

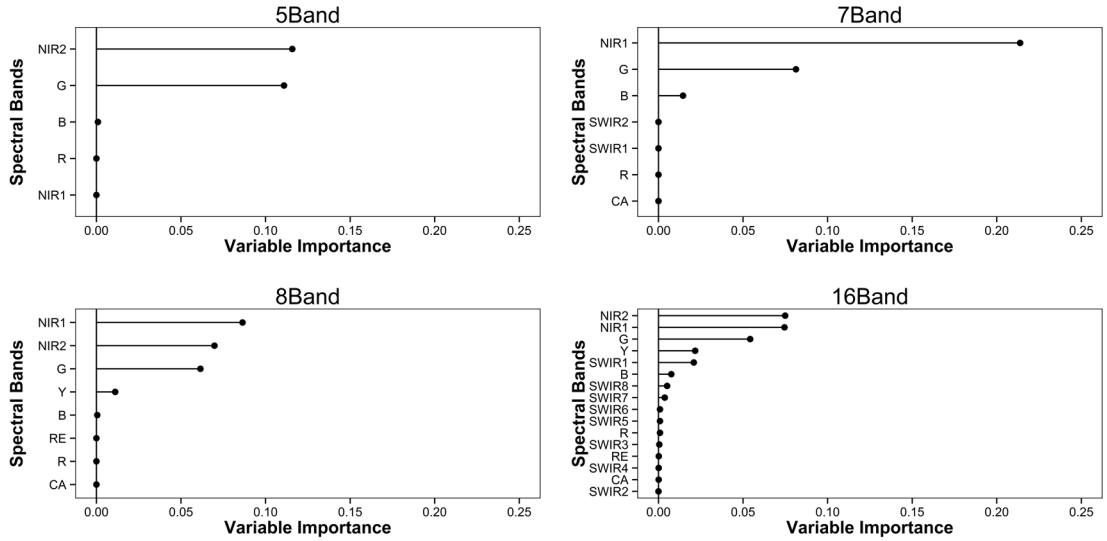

Figure 4. Variable importance rankings per multispectral dataset derived by the random forest model used for the velvetleaf and soybean P5460LL classification, September 17, 2014. CA = coastal, B = blue, G = green, Y = yellow, R = red, RE = red-edge, NIR = near infrared, and SWIR = shortwave infrared.

micellular spaces of plant leaves affect their ability to reflect and absorb near infrared light [12]. Therefore, leaf pigment and internal structure appear to be important components for distinguishing soybean from velvetleaf. Additionally, for the 7 Band and 16 Band datasets, the SWIR bands were important to the models for velvetleaf soybean discrimination; however, the SWIR bands' importance to a model was date specific. The shortwave infrared reflectance of plant leaves is affected by the water content of the leaf tissues [12]. Furthermore, for the 8 Band and 16 Band datasets, the Y spectral band was consistently ranked as an important variable to the models. Plant leaves reflectance of yellow light is also affected by chlorophyll content of the leaves.

With the increase in the number of spectral bands, more variables were ranked important to the random forest models (**Figures 1-4**); however, the increase in the number of bands per se did not always result in an increase in classification accuracy. For example, the number of accuracy test results completed for both dates and soybean varieties equal twenty-four. The 7 Band, 16 Band, 8 Band, and 5 Band datasets ranked or tied for the highest accuracies seventeen, sixteen, twelve, and one time, respectively. The differences in overall, user's, and producer's accuracies ranged from 0% to 6.6%, with the lowest accuracies occurring 95% of the time for the 5 Band dataset. For the kappa coefficients, the 5 Band model ranked last 100% of the time. The lower classification accuracies observed for the 5 Band dataset were most likely a result of the broader bandwidths (*i.e.*, 100 nm or greater). Also, the findings indicated that reliable accuracies generally can be achieved using the default *mtry* and *ntree* values (**Table 4**).

To put this study into perspective, leaf multispectral reflectance data were used as input into the random forest model for differentiating the velvetleaf from the soybean varieties. Leaf reflectance measurements represent pure reflectance measurements. Plant canopy response is affected by leaf angle, leaf positioning in the plant canopy, inter-canopy shadowing, soil background, and intermixing of plant canopies. Those aspects could lead to a different variable importance ranking of the spectral bands for plant canopy studies. Additionally, the study focused on binary classifications of soybean versus velvetleaf. Future studies need to focus on determining the potential of discriminating more than one weed at a time from soybean. Overall, this study provided valuable information on using the machine learning technique and on the influence of using different multispectral band combinations as input into the model for velvetleaf soybean discrimination.

5. Conclusion

This study provided new information on using the random forest algorithm with leaf multispectral reflectance data for differentiating velvetleaf from soybean. It demonstrated that the random forest algorithm could be used with a complement of multispectral datasets to separate velvetleaf from soybean. The best accuracies were achieved with multispectral datasets sensitive to visible (green and yellow spectral bands), near infrared, and

shortwave infrared light. Findings support further application of the random forest machine learner along with remotely-sensed multispectral data as tools for velvetleaf soybean discrimination with future implications for site-specific management of velvetleaf.

Acknowledgements and Disclaimer

The author is grateful to Dr. Vijay Nandula for supplying the velvetleaf seed, Mr. Milton Gaston Jr., Mr. Arrington Smith, Ms. Keysha Hamilton, Mr. David Fisher, Ms. Raven Thompson, and Ms. Keyanna Nealon for their assistance in data collection, and Dr. Ken Fisher and Dr. Chenghai Yang for their critical review of the manuscript. Mention of trade names or commercial products in this report is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture.

References

[1] Lanini, W.T. and Wertz, B.A. (2015) Velvetleaf. Penn State Extension. http://extension.psu.edu/pests/weeds/weed-id/velvetleaf

[2] Koger, C.H., Bruce, L.M., Shaw, D.R. and Reddy, K.N. (2003) Wavelet Analysis of Hyperspectral Reflectance Data for Detecting Pitted Morning Glory (*Ipomoea lacunosa*) in Soybean (*Glycine max*). *Remote Sensing Environment*, **86**, 108-119. http://dx.doi.org/10.1016/S0034-4257(03)00071-3

[3] Smith, A.M. and Blackshaw, R.E. (2003) Weed-Crop Discrimination Using Remote Sensing: A Detached Leaf Experiment. *Weed Technology*, **17**, 811-820. http://dx.doi.org/10.1614/WT02-179

[4] Yang, C.C., Prasher, S.O. and Goel, P.K. (2004) Differentiation of Crop and Weeds by Decision-Tree Analysis of Multi-Spectral Data. *Transactions of the ASAE*, **47**, 873-879. http://dx.doi.org/10.13031/2013.16084

[5] Iqbal, J., Owens, P.R. and Ali., I. (2006) Application of Remote Sensing Data to Assess Weed Infestation in Cotton. *Agricultural Journal*, **1**, 186-191.

[6] Gómez-Casero, M.T., Castillejo-González, I.L. and García-Ferrer, A. (2010) Spectral Discrimination of Wild Oat and Canary Grass in Wheat Fields for Less Herbicide Application. *Agronomy for Sustainable Development*, **30**, 689-699. http://dx.doi.org/10.1051/agro/2009052

[7] Nieuwenhuizen, A.T., Hofstee, J.W., van de Zande, J.C., Meuleman, J. and van Henten, E.J. (2010) Classification of Sugar Beet and Volunteer Potato Reflection Spectra with a Neural Network and Statistical Discriminant Analysis to Select Discriminative Wavelengths. *Computers Electronics in Agriculture*, **73**, 146-153. http://dx.doi.org/10.1016/j.compag.2010.05.008

[8] de Castro, A.I., Jurado-Expósito, M., Gómez-Casero, M.T. and López-Granados, F. **(**2012) Applying Neural Networks to Hyperspectral and Multispectral Field Data for Discrimination of Cruciferous Weeds in Winter Crops. *Science World Journal*, Article ID: 630390. http://dx.doi.org/10.1100/2012/630390

[9] Lamb, D.W. and Brown, R.B. (2001) Remote-Sensing and Mapping of Weeds in Crops. *Journal of Agricultural Engneering Research*, **78**, 117-125. http://dx.doi.org/10.1006/jaer.2000.0630

[10] Goel, P.K., Prasher, S.O., Patel, R.M., Smith, D.L. and Di Tommaso, A. (2002) Use of Airborne Multi-Spectral Imagery for Weed Detection in Field Crops. *Transactions of American Society of Agricultural Engineers*, **45**, 443-449.

[11] Gibson, K.D., Dirks, R., Medlin, C.R. and Johnston, L. (2004) Detection of Weed Species in Soybean Using Multispectral Digital Images. *Weed Technology*, **18**, 742-749. http://dx.doi.org/10.1614/WT-03-170R1

[12] Gausman, H. (1985) Plant Leaf Optical Properties. Texas Tech Press, Lubbock.

[13] Fernández-Delgado, M., Cernadas, E., Barro, S. and Amorim, D. (2014) Do We Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research*, **15**, 3133-3181.

[14] Gislason, P.O., Benediktsson, J.A. and Sveinsson, J.R. (2006) Random Forests for Land Cover Classification. *Pattern Recognition Letters*, **27**, 294-300. http://dx.doi.org/10.1016/j.patrec.2005.08.011

[15] Strobl, C., Malley, J. and Tutz, G. (2009) An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests. *Psychological Methods*, **14**, 323-348. http://dx.doi.org/10.1037/a0016973

[16] Goldstein, B.A., Polley, E.C. and Briggs, F.B.S. (2011) Random Forest for Genetic Association Studies. *Applications in Genetics and Molecular Biology*, **10**, 1-34. http://dx.doi.org/10.2202/1544-6115.1691

[17] Ok, A.O., Akar, O. and Gungor, O. (2012) Evaluation of Random Forest Method for Agricultural Crop Classification. *European Journal of Remote Sensing*, **45**, 421-432.

[18] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. http://dx.doi.org/10.1023/A:1010933404324

[19] US Geological Survey (2015) Frequently Asked Questions about the Landsat Missions. http://landsat.usgs.gov/best_spectral_bands_to_use.php

[20] Digital Globe (2010) The Benefits of the Eight Spectral Bands of WorldView 2. http://global.digitalglobe.com/sites/default/files/DG-8SPECTRAL-WP_0.pdf

[21] Digital Globe (2014) WorldView 3 Data Sheet. https://dg-cms-uploads-production.s3.amazonaws.com/uploads/document/file/95/DG_WorldView3_DS_forWeb_0.pdf

[22] Lehnert, L.W., Meyer, H. and Bendix, J. (2015) Hsdar: Manage, Analyse and Simulate Hyperspectral Data in R. R Package Version 0.3.0. https://cran.r-project.org/web/packages/hsdar/index.html

[23] Hothorn, T., Buehlmann, P., Dudoit, S., Molinaro, A. and van der Laan, M. (2006) Survival Ensembles. *Biostatistics*, **7**, 355-373. http://dx.doi.org/10.1093/biostatistics/kxj011

[24] Congalton, R. and Green, K. (2009) Assessing the Accuracy of Remotely Sensed Data: Principles and Practices. 2nd Edition, CRC/Taylor & Francis, Boca Raton, 183 p.

[25] Hothorn, T., Hornik, K. and Zeileis, A. (2006) Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, **15**, 651-674. http://dx.doi.org/10.1198/106186006X133933

[26] Strobl, C., Boulesteix, A.L., Zeileis, A. and Hothorn, T. (2007) Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinformatics*, **8**, 25. http://dx.doi.org/10.1186/1471-2105-8-25

[27] Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T. and Zeileis, A. (2008) Conditional Variable Importance for Random Forests. *BMC Bioinformatics*, **9**, 307. http://www.biomedcentral.com/1471-2105/9/307 http://dx.doi.org/10.1186/1471-2105-9-307