The Society for engineering in agricultural, food, and biological systems

An ASAE Meeting Presentation

Paper Number: 053063

Genetic algorithms for Hyperspectral Range and Operator Selection

Brian L. Steward, Associate Professor

Agricultural and Biosystems Engineering Dept., Iowa State University, Ames, IA 50011 <bsteward@iastate.edu>

Amy L. Kaleita, Assistant Professor

Agricultural and Biosystems Engineering Dept., Iowa State University, Ames, IA 50011

Robert P. Ewing, Research Scientist

Agronomy Dept., Iowa State University, Ames, IA 50011

Daniel A. Ashlock, Professor

Dept. of Mathematics and Statistics, University of Guelph, Guelph, Ontario, Canada

Written for presentation at the
2005 ASAE Annual International Meeting
Sponsored by ASAE
Tampa Convention Center
Tampa, Florida
17 - 20 July 2005

Mention any other presentations of this paper here, or delete this line.

Abstract. *A novel genetic algorithm was developed using mathematical operations on spectral ranges to explore spectral operator space and to discover useful mathematical range operations for relating spectral data to reference parameters. For each range, the starting wavelength and length of the range, and a mathematical range operation were selected with a genetic algorithm. Partial least squares (PLS) regression was used to develop models predicting reference variables from the range operations. Reflectance spectra from corn plant canopies were investigated, with proportion of plants (1) with visible tassels and (2) starting to shed pollen as reference data. PLS models developed using the spectral range operator framework had similar fitness than PLS models developed using the full spectrum. This range/operator framework enabled identification of those*

spectral ranges with most predictive capability and which mathematical operators were most effective in using that predictive capability. Detection of operator locality may have utility in sensor and algorithm design and in developing breeding stock for other algorithms.

Keywords. hyperspectral analysis, remote sensing, spectroscopy, corn development, evolutionary computation

Introduction

The interest in hyperspectral or spectroscopic sensing for agricultural applications comes from the hope that the resulting high dimensional data will lead to new understanding that was not attainable using only broad spectral sensors. However, with the availability of high spectral resolution sensing technology comes the problem of how to extract meaningful relationships from such overwhelming quantities of data. One approach has been to use conventional linear methods such as principal component analysis (PCA) or partial least squares (PLS) to relate linear combinations of spectral bands to independent physical variables of interest. These "soft modeling" approaches are powerful, but they are limited because they (a) only apply linear relationships between spectral variables, (b) incorporate very little domain knowledge or physical first principles into the analysis (Martens and Martens, 2001), and (c) implicitly treat each channel as independent, rather than using information embedded in the structure of the spectrum.

A second approach, developed over years of remote sensing research and application, uses domain knowledge to calculate indices using simple mathematical operations consisting of ratios, differences, and sums such as the Normalized Difference Vegetation Index (NDVI; Rouse et al., 1973). Such an approach necessarily grew out of a period when only multispectral data – fewer than ten wavebands – were available, making tractable the process of experts finding waveband/operator combinations that were closely related to physical parameters of interest. However, with increasing numbers of spectral variables, the number of possible combinations grows rapidly, making impractical the use of trial-and-error by experts to find waveband/operator combinations that relate to a property or process of interest.

Evolutionary computation, and specifically the subset called genetic algorithms, are adaptive search methods that deal powerfully with difficult search problems without getting stuck in local extremes in the search space (Goldberg, 1989). Using a genetic model, search parameters are coded at random in a data structure called a chromosome. At each generation, each individual in a population is evaluated for fitness. Individuals with better fitness are selected for the next generation through crossover and mutation.

Evolutionary computation has rarely been used in spectral analysis for agricultural applications. Tang et al. (2000) used a genetic algorithm (GA) to identify a region in color space for vegetation segmentation in field images. Rauss et al. (2000) used genetic programming (GP) to develop a classification algorithm for an image with 28 spectral bands. Burman (1997, 1999) used evolutionary computation methods to perform automatic object detection and recognition in hyperspectral images. Yao and Tian (2003) used a GA to select hyperspectral bands to include in a principal components transformation for information extraction.

The few studies that have used evolutionary computation in spectral analysis have focused on linear combinations of the original data. In remote sensing, however, domain knowledge indicates that significant information can be gleaned by utilizing the data's spectral relationships. Numerous studies have addressed the utility of derivative approaches to remote sensing data analysis (Demetriades-Shah et al., 1990; Clevers et al., 2002). But to our knowledge, no studies to date have used evolutionary computation on spectral range operators such as slope or curvature.

The objective of this research was to use genetic algorithms to search the space of spectral ranges and operations to find the spectral regions with the most relatedness to the reference variable and determine the utility of range operators in hyperspectral data analysis.

Materials

The data set used for this work consisted of visible and near infrared (NIR) reflectance spectra from corn canopies. This data was collected in support of research to predict corn pollination several days in advance.

The corn observed in this study was grown at Iowa State University's Bruner Farm (42° 0.6' N., 93° 44.2' W.) during the 2003 growing season. Half of the field was planted with the cultivar Asgrow 740, the rest with Dekalb 611. Plots measured (6.1 x 12.2 m), and were replicated in three blocks. Each cultivar had three treatments: control, 50% detasseled, and 100% detasseled, with detasseling being done manually between the visible tassel and initial pollen shed growth stages. Five control areas were established within each plot, with each control area measuring approximately 2 m by 2 m and containing between one to seven plants, with three or four being the most common. During the tasseling and pollen shed period (late July and early August), plots were scouted repeatedly to determine fraction of tasseled plants within each control area that had reached the visible tassel (VT) or initial pollen shed (IPS) growth stage.

Visible and near infrared (NIR) reflectance spectra of these same control areas were recorded periodically during the same period with an ASD FieldSpec spectrophotometer (Advanced Spectral Devices, Boulder, CO). The downward looking sensor was mounted on a boom at a distance of 2 m above the canopy. An initial calibration using a Spectralon^{TM} panel was performed, and an upward looking sensor measured ambient light so that relative canopy reflectance could be calculated. The FieldSpec collected 751 spectral variables at 1 nm intervals across a spectral range from 325 nm to 1075 nm. In order to reduce sampling noise, 25 instantaneous spectra were averaged into one spectrum per sample. After initial inspection of the spectra, the range was truncated to 397 nm to 902 nm (505 variables). Reflectance data were transformed to absorbance using the $1/\log_{10}$ transform. In order to conserve computational time, the spectra were downsampled to include one out of every four wavelengths, for a total of 127 bands.

A total of 622 spectra and corresponding reference data (VT and IPS) were collected. Preliminary data analysis revealed little or no difference in reflectance spectra from the two cultivars. Thus, the analysis presented here used aggregated data from both cultivars. A potential limitation of this data set is that because of the relatively small number of plants in each control area, the reference variables have a limited number of discrete values; there are 15 unique values of VT from 0 to 1, and 17 unique values of IPS from 0 to 1.

Methods

When creating an evolutionary computation system for solving a given problem the choice of representation or problem encoding is critical. Direct encodings store explicit solutions in the populations of data structures being evolved. An indirect encoding suggests how to solve the problem without explicitly giving the solution. The representation used in this study is an indirect encoding. Rather than giving an explicit model for a given set of hyperspectral data, it selects a set of features which are the output of range operations to be passed to the PLS software. The PLS software uses that set of features to build the model.

In general, direct encodings are easier to design, implement, and control. Indirect encodings, in contrast, can leverage the power of other algorithms. A direct encoding is natural when nothing is known about the problem. The more specialized knowledge a researcher has available, the

more likely that an indirect encoding will be superior. Using indirect coding, specialized knowledge can be built into the algorithms used to construct the problem solution from the data.

The range operators used in this system exploit the mutual information of ranges of spectral frequencies as well as serving to compensate for noise in the data. If there is a particular spectral region that is relevant to detecting corn pollination then the particular data collection hardware, the weather and illumination on the day the data were taken, and even the particular strain of corn may perturb the data. By basing our model on abstractions of spectral data ranges we have potential for more robust modeling.

The notion of critical spectral ranges leads to another feature of this encoding that is useful. By plotting which range abstractions land in which parts of the target's spectrum, we can learn which frequency ranges are most information-rich with respect to which operators. This in turn permits the direct design of better models or more effective initialization of populations of evolvable models.

This range/operator enabled genetic algorithm (ROE-GA) was developed in Matlab script (The Mathworks, Natick, MA) to search the space of possible ranges and several basic mathematical operators on the range. Features extracted in the form of the multiple range operators were used as inputs into PLS which was used to calibrate a model for predicting associated reference variables (Figure 1). Cross validation was performed, using the standard error of cross validation (SECV) as the fitness for the GA. Multivariate analysis algorithms in PLS Toolbox (Eigenvector Research, Manson, WA) were used.

GA components and Operators

Chromosome: Chromosomes were composed of genes that described spectral range operators. Each gene consisted of three numbers representing (a) the spectral starting point *S*, (b) the length *L* of the range, with a maximum length of 100 bands, and (c) a code *O* for the operator; for convenience we refer to an individual gene as a *SLO*. The number of ranges used in the analysis was set at 25 which was arrived at through experimentation.

Five simple range operators were used: maximum, minimum, median, slope (determined from linear regression of fit to the spectral data over the range defined by *S* and *L*), and curvature (determined from a second-order polynomial fit to the spectral data over the range). A sixth operator was also used which returned a value of zero for that *SLO*, effectively turning it off. Individuals with more zero operators were rewarded in order to minimize ranges that added little contribution to the model.

We initially experimented with other operators including mean and variance, but discarded them because they were infrequently selected compared to other operators. Furthermore, in the case of the mean operator, there was likely some amount of content duplication when using mean in addition to median.

Population size: Goldberg's rule of thumb (1989) suggests a population equal to the length of the chromosome. For this corn canopy data, we used a population of 256 individuals, somewhat less than the chromosome length (~275 bits), but yielding a manageable 1 to 2 minute runtime per generation on the Matlab platform.

Selection and crossover: The population was ranked by fitness value. Absolute fitness replacement was implemented by sorting the population by fitness and discarding the less fit half of the population. The upper half was then replicated to form a new lower half, and children were formed using double point crossover.

Mutation: At each generation, the population of 256 individuals was sorted by fitness, and the less fit half of the population was discarded. The upper half was then replicated to form a new lower half, and children were generated using two-point crossover. Each child was selected for mutation. The mutation rate (the number of genes per individual that can mutate) was variable, starting with a very high mutation rate and then linearly approaching a low mutation rate near the end. When a *SLO* was mutated, either the starting point *S*, length *L*, or operator *O* was mutated. Mutations for *S* and *L* were limited to changes of 10-band intervals. Mutation of *O* was conversion to one of the operators selected at random.

Fitness: The fitness for an individual was the SECV associated with that individual calculated through a PLS cross-validation procedure (PLS Toolbox, Eigenvalue Research). The ROE-GA stopped when either 300 generations had passed or more than one half of the population contained duplicate individuals. Analysis was replicated five times using five different starting populations. Cross validation was accomplished using random segment selection in which the data was divided into five segments. Each segment was predicted using a calibration model calculated from the remaining four segments. This five segment cross-validation was repeated ten times. Model performance was measured using standard error of cross validation (SECV). The minimum SECV across full spectral models with 1 to 20 latent variables was used as a measure of model performance.

Analysis

The dataset was divided into calibration and prediction datasets consisting of three-quarters and one-quarter of the spectra, respectively. The ROE-GA method was applied to the calibration set resulting in ending populations of spectral range operators. The individual resulting from the ROE-GA method with the highest fitness was applied to the spectra in the calibration dataset and PLS was used to develop calibration models for both reference data using the spectral range features. The spectra of the prediction dataset were then applied to these models to predict the reference values. Predicted reference data was compared with measured data and model performance measures, R^2 and standard error of prediction (SEP) were calculated.

In addition, calibration models were developed using PLS directly operating on the down-sampled (by four times) spectral data with standard normal variate preprocessing. Cross validation was accomplished using random segment selection in which the data was divided into five segments. Each segment was predicted using calibration models calculated from the remaining four segments. This five segment cross-validation was repeated ten times. Model performance was measured using root-mean-squared error of cross validation (SECV). Typically, SECV decreased with increasing numbers of latent variable being added to the model until a minimum value was reached; then additional latent variables led to overfitting with SECV increases. For each of the full spectral data models, the minimum SECV was used as a measure of model performance. Similarly, the prediction data was applied to these models. Predicted reference data was compared with measured data and model performance measures were calculated.

Use of the ROE-GA approach allows us to analyze the portions of the spectral shape that may contain relevant information. The five fittest individuals from the each of the five replications were selected. Those range operations which were used by 15 or more of these top 25 individuals were found. These selected range operations were mapped onto the mean reflectance spectrum to determine which spectral features were being selected by the ROE-GA method. In addition, the VT and IPS data were divided into three classes based on their value. Data values less than 20% where placed in the first class; data values between 20% and 80% were placed in the second class; and data values greater than 80% were placed in the third

class. These class thresholds were chosen based on the salient stage of tassel development (Westgate et al., 2003). Normalized mean reflectance spectra for each of these classes were calculated to find differences in the spectra across plant development.

Results

The overall prediction performance of the two methods were similar. In each case, PLS analysis gave models with a good fit to the reference data. The coefficient of determination, R^2 was of 0.85 and 0.80 for VT and IPS, respectively. The SECV of the models were 0.19 and 0.15, and on the prediction set, SEP were 0.14 and 0.16 for VT and IPS, respectively. The coefficients of the PLS models did not show any consistent pattern across the spectrum (fig. 2). In comparison, the PLS models developed using the range operator features from the most fit individuals from five populations had an SEPs of 0.133 for VT and 0.175 for IPS and R^2 of 0.87 and 0.76 respectively.

The more salient aspect of the ROE-GA was observed after the most common spectral ranges were found from the five final populations. For the VT models, 12 dominant spectral range operations were found in 15 or more models out the top 25 (Table 1). For the IPS models, 10 dominant spectral range operations were found (Table 2). The proportion of operator types were different for the two reference data sets. For the VT models, 71% of the popular *SLO*s used slope or curvature operators, whereas for IPS, only 30% of the popular *SLO*s were slope or curvature operators, and the 60 percent of the SLOs used the median operator.

More particularly, the dominant SLOs tended to select regions of the reflectance spectrum where domain experts would expect changes to occur during this time of plant development. For example, several of the SLOs were near or including the green peak of the reflectance curve. Similarly, several SLOs included changes associated with the red edge or the near infrared plateau. Even more interestingly, the algorithm identified a small peak centered at approximately 610 nm. The reflectance changes at this location were not identified initially until the ROE-GA algorithm selected them. However, these changes across development were then also observed in the normalized class spectra (fig. 3 and 4). These results were similar to the changes in reflectance reported by Vina et al. (2004).

Conclusions

This work shows that a collection of a few simple range operators are sufficient to discover useful feature sets. When those features are used to develop PLS models, the models exhibit performance in cross-validation similar to that of full spectrum models. Another useful outcome is the association of particular operators with spectral ranges, which may prove useful in understanding how spectral ranges are associated with reference variables of interest.

While we have applied this representation to a hyperspectral modeling problem, the technique has far broader potential application. Any problem involving large noisy data sets that need to be modeled could conceivably benefit from the techniques used here. The steps are as follows:

Choose a collection of range selection and abstraction operators. There must be some encoding method indicating the where the boundaries associated with the range and what operation should be applied to the range. While in this work the data range along a spectral axis, the technique is not limited to data along one axis. Data with spatial attributes, such as images or georeferenced data, could also be treated by the technique but with ranges now being regions of interest.

Choose a modeling technique, such as PLS, principal components regression, or multiple-linear regression to generate a calibration or classification model for the reference variables in terms of the range operator-extracted features. Some measure of model performance must be calculated.

Evolve collections of range abstractions with model quality as a fitness function. The association of operators with spectral ranges may be useful in initializing subsequent evolutionary runs with those associated already embedded. Such expert initialization may enhance performance.

Acknowledgements

This research of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa, Project No. 3612, was supported by Hatch Act and State of Iowa funds. This research was also supported in part by the Iowa State University Special Research Initiation Grant Program and the Iowa Space Grant Consortium. Special thanks to M. E. Westgate, J. L. Hatfield, and the National Soil Tilth Laboratory for the corn canopy reflectance data and the ISU Grain Quality Laboratory for the corn kernel transmittance data.

References

Burman, J. A., 1997, Non-literal pattern recognition method for hyperspectral imagery exploitation using evolutionary computing methods, *In Imaging spectrometry III, SPIE Proceedings,* Vol. 3118,250-261, Bellingham, WA.: SPIE.

Burman, J. A., 1999, Hybrid pattern recognition method using evolutionary computing techniques applied to the exploitation of hyperspectral imagery and medical spectral data, *In Image and signal processing for remote sensing V*: *SPIE proceedings*, Vol. 3781, 348-357, Bellingham, WA: SPIE.

Clevers, J. G., S. M. De Jong, G. F. Epema, F. D. Van Der Meer, W. H. Bakker, A. K. Skidmore, and K. H. Scholte, 2002, Derivation of the red edge index using the MERIS standard band setting, *International Journal of Remote Sensing* 23: 3169-3184.

Demetriades-Shah, T. H., M. D. Steven, and J. A. Clark, 1990, High resolution derivative spectra in remote sensing, *Remote Sensing of Environment* 33: 55-64.

Goldberg, D. E., 1989, Sizing populations for serial and parallel genetic algorithms. In *Proc. 3rd Int. Conference on Genetic Algorithms*, 70-79, San Mateo, Calif: Morgan Kaufman.

Martens, H. and M. Martens, 2001, *Multivariate Analysis of Quality*, New York : John Wiley.

Rauss, P. J., J. M. Daida, and S. Chaudhary, 2000, Classification of spectral imagery using genetic programming, *Proc. of the Genetic and Evolutionary Computation Conference,* Las Vegas, NV, July 10-12, San Francisco: Morgan Kaufmann Publishers.

Rouse, J. W., R. H. Haas, J. A. Schell, and D. W. Deering, 1973, Monitoring vegetation systems in the Great Plains with ERTS, *Proceedings, 3rd ERTS Symposium*, 1: 48-62.

Tang, L., L. F. Tian, and B. L. Steward, 2000, Color image segmentation with genetic algorithm for in-field weed sensing, *Transactions of the ASAE* 43: 1019-1027.

Vina, A., A. A. Gitelson, D. C. Rundquist, G. Keydan, B. Leavitt, and J. Schepers. 2004. Monitoring maize (Zea mays L.) phenology with remote sensing. *Agronomy Journal* 96: 1139-1147.

Westgate, M.E., J. Lizaso, and W.D. Batchelor, 2003. Quantitative relationships between pollen shed density and grain yield in maize. *Crop Sci.*, 43: 934-942.

Yao, H. and Tian, L., 2003, "A genetic algorithm-based selective principal component analysis (ga-spca) method for high dimensional data feature extraction." *IEEE Transactions on Geoscience and Remote Sensing,* Vol. 41, pp. 1469 -1478.

Table 1. Regions identified as influential for VT prediction by the ROEGA. Only those SLOs which were used by 15 or more of the top 5 models from 5 populations are presented. Operator O, spectral range determined by S and L, and feature of the vegetation spectrum in that region are listed.

Operator	Range (nm)	No. uses	Spectral Feature
Minimum	633 – 673	20	Beginning of red edge
Median	593 – 633	15	Small yellow peak
	713 – 833	15	Red edge
	873 – 901	25	Far edge of spectra
Slope	433 – 513	15	Up to green peak
	753 – 901	15	Entire NIR plateau
	873 – 901	20	Far edge of spectra
Curvature	513 – 593	15	Green peak
	633 – 673	15	Between the yellow peak and the red edge
	593 – 713	20	Green peak to beginning half of red edge
	753 – 793	25	End of red edge and beginning of NIR plateau
	873 – 901	20	Far edge of spectra

Table 2. Regions identified as influential for IPS prediction by the ROEGA. Only those *SLO*s which were used by 15 or more of the top 5 models from 5 populations are presented. Operator *O*, spectral range determined by *S* and *L*, and feature of the vegetation spectrum in that region are listed.

Operator	Range (nm)	No. uses	Spectral Feature
Maximum	901	15	Last band in spectra
Median	473 – 713	15	Wide band from green peak through beginning of red edge
	513 – 593	15	Green peak
	593 – 713	21	Wide area around small yellow peak
	593 – 633	25	Small yellow peak
	673 – 873	17	Red edge through NIR plateau
	793 – 833	15	Within NIR plateau
Slope	553 – 593	25	Downslope of green peak before the small yellow peak
	833 – 901	20	Far end of NIR plateau
Curvature	753 – 793	20	End of red edge and beginning of NIR plateau

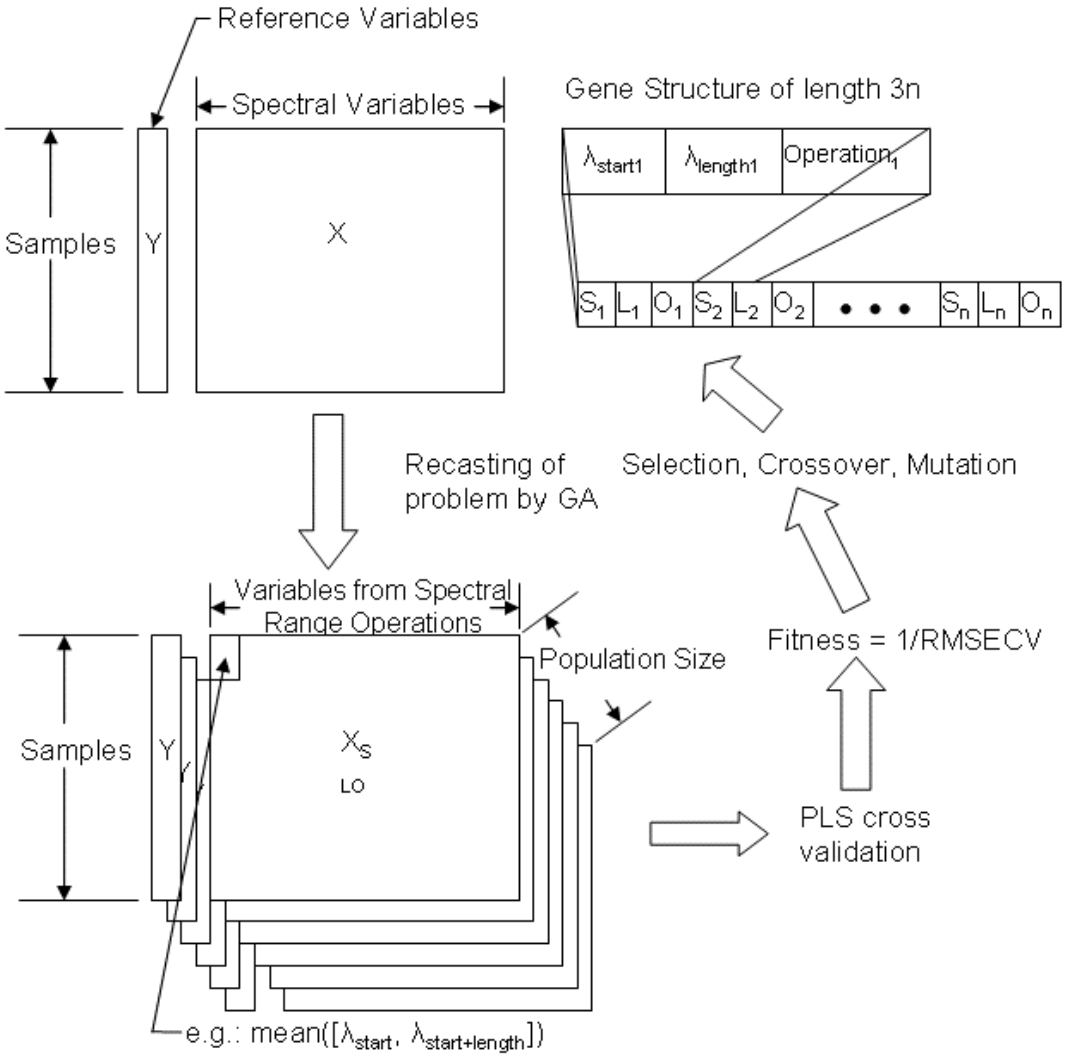

Figure 1. Diagram of range operator enabled genetic algorithm methodology.

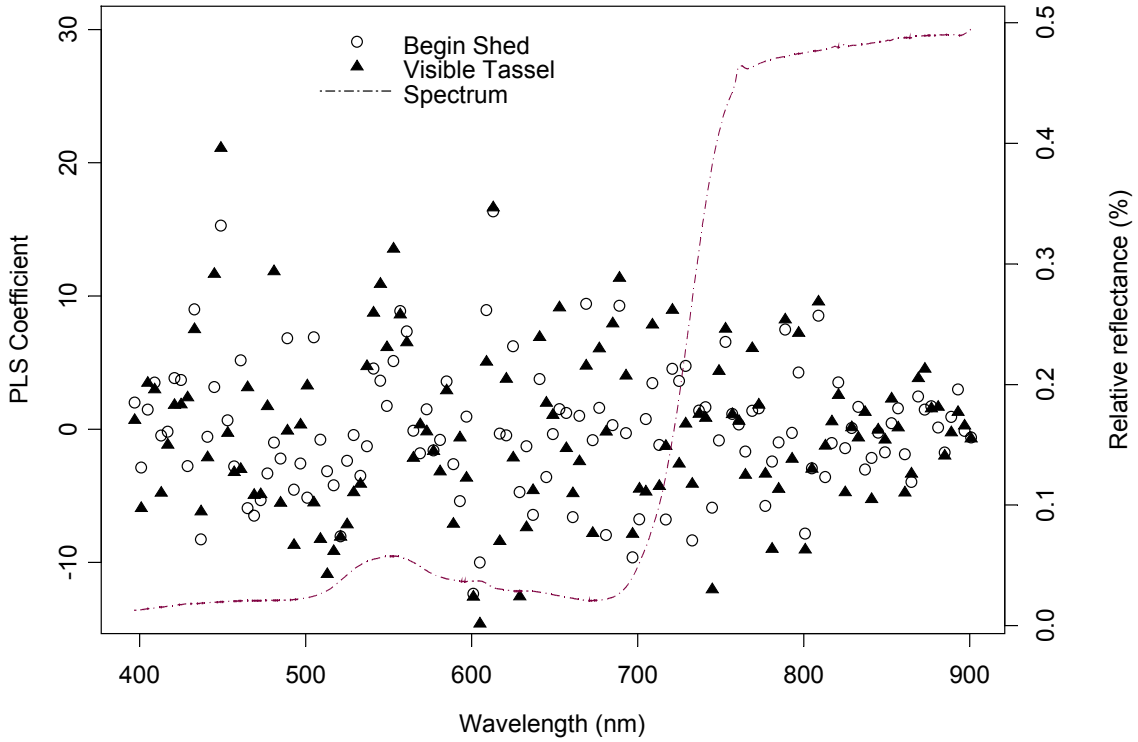

Figure 2. PLS coefficients for models of VT and IPS. The shape of an average reflectance spectrum is shown for reference.

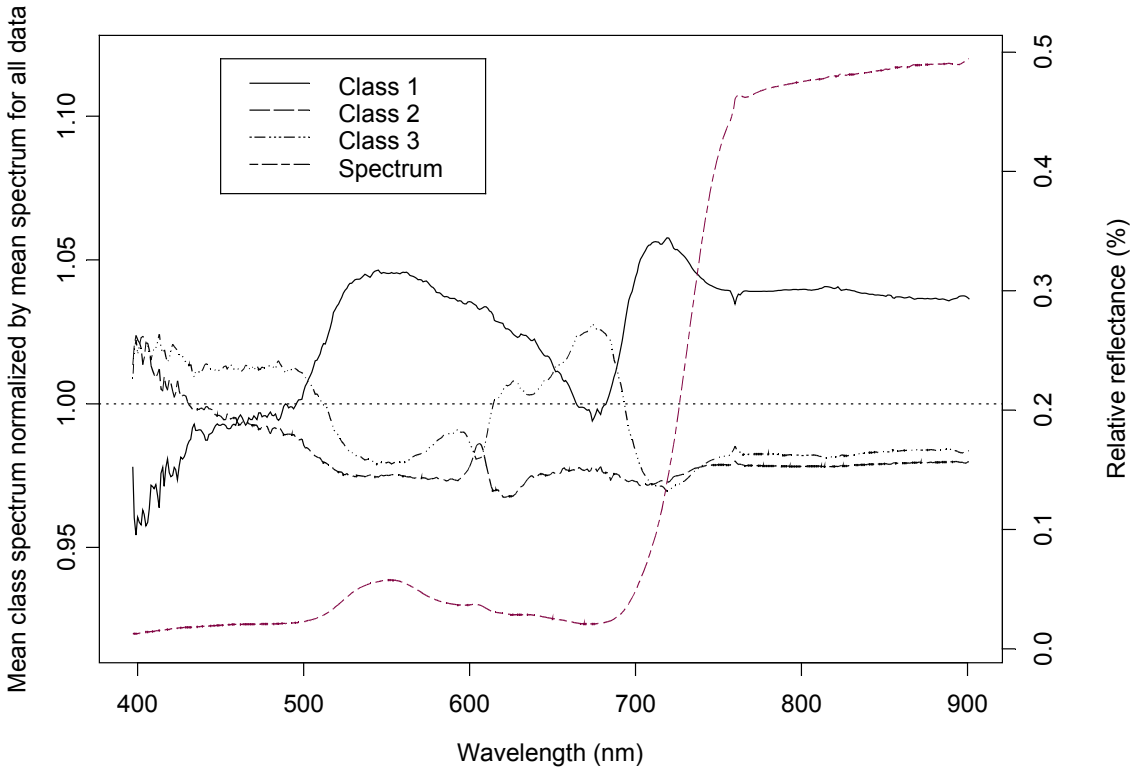

Figure 3. Mean normalized reflectance spectra for each of the three classes of VT: VT less than 20% (Class 1), VT between 20% and 80% (Class 2), and VT greater than 80% (Class 3). Spectra were normalized by the average of the three classes.

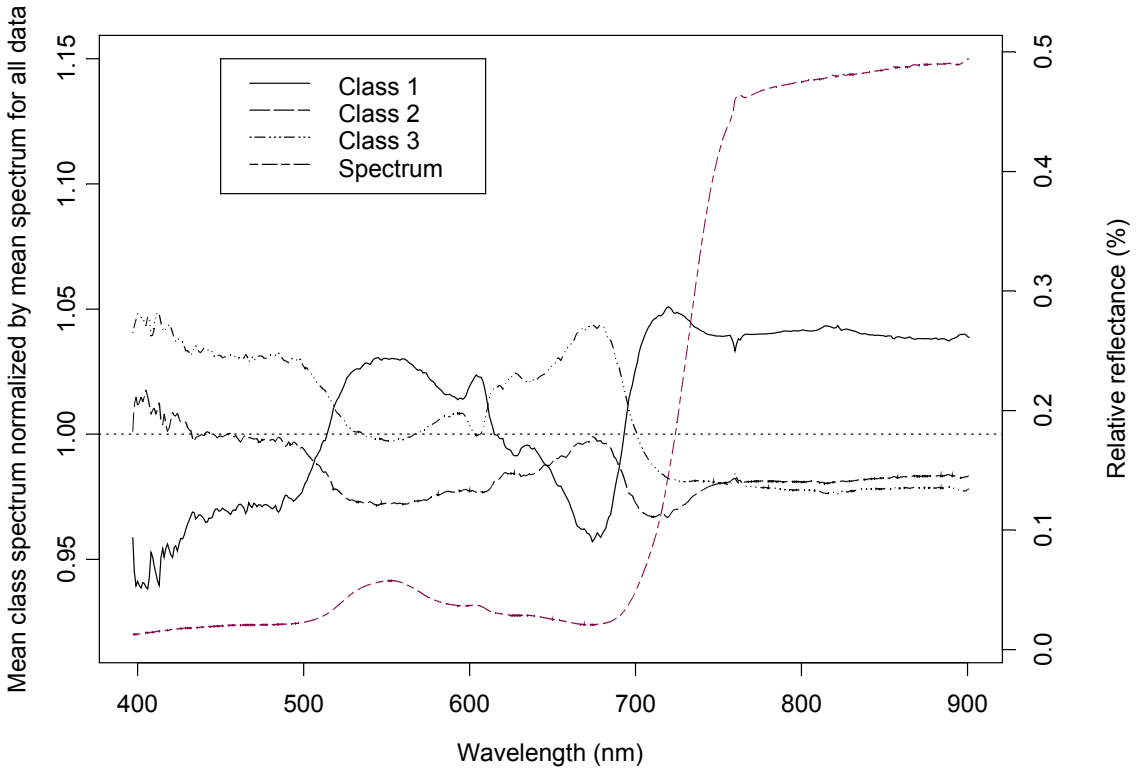

Figure 4. Mean reflectance spectra for each of the three classes of IPS: IPS less than 20\% (Class 1), IPS between 20% and 80% (Class 2), and IPS greater than 80% (Class 3). Spectra were normalized by the average of the three classes.