Kansas State University Libraries

New Prairie Press

Conference on Applied Statistics in Agriculture 2014 - 26th Annual Conference Proceedings

MULTIVARIATE STATISTICAL ANALYSIS OF COLEOPTERA SPECTRAL REFLECTANCE

Sarah E.M. Herberger
University of Idaho

Bahaman Shafii
University of Idaho

Stephen P. Cook
University of Idaho

Christopher J. Williams
University of Idaho

William J. Price
University of Idaho

See next page for additional authors

Recommended Citation

Herberger, Sarah E.M.; Shafii, Bahaman; Cook, Stephen P.; Williams, Christopher J.; and Price, William J. (2014). "MULTIVARIATE STATISTICAL ANALYSIS OF COLEOPTERA SPECTRAL REFLECTANCE," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1002

Author Information

Sarah E.M. Herberger, Bahaman Shafii, Stephen P. Cook, Christopher J. Williams, and William J. Price

Multivariate Statistical Analysis of Coleoptera Spectral Reflectance

Sarah E.M. Herberger[1], Bahman Shafii[1,2,3], Stephen P. Cook[2], Christopher J. Williams[1], William J. Price[3]

[1]Department of Statistics

[2]Department of Plant, Soil, and Entomological Sciences

[3]Statistical Programs

University of Idaho, Moscow, ID 83844

Abstract

The insect order Coleoptera, commonly known as beetles, comprises 40% of all insects which in turn account for half of all identified animal species alive today. Coleopterans frequently have large elytra (the hardened front wings) that can have a wide range of colors. Spectral reflectance readings from these elytra may be used to uniquely identify coleopteran taxonomic groups. Multiple samples of eleven species of wood boring beetles were selected from the University of Idaho William Barr Entomology Museum. Spectrometer readings for each specimen were then fit to normal distribution mixture models to identify multiple peak reflectance wavelengths. Eighteen prominent peaks were identified across all taxonomic groups and genders creating a multivariate response structure. Multivariate statistical procedures including principal component and discriminant analyses were employed to assess the differentiation of taxonomic groups and genders based on spectral reflectance. The first three axes of the principal component analysis accounted for 96% of the variation and provided a clear clustering of genus and gender for a subset of taxonomic groups. The linear discriminant analysis under an assumption of multivariate normality provided a distinct classification of taxonomic groups resulting in an overall 4% misclassification rate; while the nearest neighbor discriminant analysis with a proportional prior gave an overall error rate of 5.2%. Internal bootstrap validation of the latter discriminant model yielded an average error rate of 3.5%. An external cross validation of the same model, conducted on independent samples of the same species with new individuals resulted in an average misclassification error rate of only 6.5%. Given the low error rates of misclassification, such multivariate statistical approaches are recommended for analysis of spectral reflectance in Coleoptera and other similar insect groups.

Introduction

Insects are one of the most abundant, diverse, and necessary life forms on earth. They play an integral role in pollination, degradation of waste, maintenance of pests, and medicine. The order Coleoptera makes up over 50% of known insect species, with 350,000 species of Coleoptera having been formally described. Coleoptera can be found in every terrestrial climate in the world with species diversity often increasing in tropical locations (Vigneron et al. 2006). Estimates on the number of Coleoptera species range from 600,000 to 3 million (Seago et al. 2009). The majority of Coleoptera species are undescribed, even when using conservative estimates.

The methods typically used for identification and classification of Coleoptera are often derived from antennal, tarsi, mouthparts (labial and maxillary palpi), ventral characters (sterna, pleura, coxae), and other morphological characteristics (Choate 1999). There is a high potential for misclassification that can occur in the process of identification. For example, long-horned beetles which do not have long antennae, snout beetles which do not have snouts, ground beetles that live in trees, or aquatic beetles that are never in the vicinity of water (Choate 1999). Morphology of an insect has to be painstakingly analyzed in order to identify them accurately, i.e. antenna measured, veins on wings analyzed, carapace shape diagramed, etc. Memorizing or locating references for morphology and then applying that knowledge for the process of identification can be very time consuming. Coupled with human error, and the ever expanding number of described species, this may lead to misclassifications. For example, one of the taxonomic groups chosen for this study, the genus *Callidium* within the family Cerambycidae, has been viewed by three different entomological experts with each one identifying it differently.

One of the most accurate ways to differentiate Coleoptera species is through their color. In fact, entomologists have created more than 30 different terms that are used to describe the color brown (Seago et al. 2009). With such a strong emphasis on color, the most distinguishable coloration is often seen in the hard front wing, or elytra, of Coleoptera. The elytra typically have a relatively uniform coloration with the most frequent colors being blue or green (Piszter 2010). The elytra are composed of chitin, with elements such as carbon, hydrogen, nitrogen, oxygen, calcium, and magnesium present to achieve a particular color (Piszter 2010). Elytral color is exposed to some of the strongest evolutionary pressures (Piszter 2010) which include, but are not limited to crypsis, aposematic, sexual signals, polarized signaling (for conspecific communication), thermoregulation, and confusion of depth perception of predators (Seago et al. 2009). Coloration in Coleoptera has been observed to change during development or as a result of environmental conditions (Seago et al. 2009). Elytral color has also been shown to vary along geographical gradients (Kawakami et al. 2013).

All of the specimens selected for this project were wood borers or predators of wood boring insects. Wood-boring beetles, or Woodborers, are often considered pests in trees and some wooden structures. The mandibles of these species are specifically designed for chewing wood.

The taxonomic groups selected from the University of Idaho William Barr Entomology Museum for this study included species in the families Cerambycidae (*Callidium* sp., *Desmocerus piperi* Webb, *Prionus californicus* Motschulsky, and *Spondylis upiformis* Mannerheim), Buprestidae (*Dicerca tenebrica* Kirby, *Melanophila atropurpurea* Say, *Buprestis lyrata* Casey, and *Trachykele blondeli blondeli* Marseul), Lucanidae (*Lucanus capreolus* Linnaeus, *Lucanus mazama* LeConte), and Trogossitidae (*Temnochila chlorodia* mannerheim).. Under most museum conditions, beetles have been shown to retain their color (Seago et al. 2009). Previous research has indicated that near infrared reflectance can be used for rapid identification of wheat pests (Dowell et al. 1999; Vigneron et al. 2006). Slight variations of color can allow one to distinguish between closely related species, as well as genders within the same species (Vigneron et al. 2006).

The objective of this study was to differentiate Coleoptera taxonomic groups, as listed above, based upon spectral reflectance of the elytra. It was also intended to potentially differentiate the gender of the aforementioned taxonomic groups using the same methodology.

Material and Methods

Source and description of data

The insect order Coleoptera was selected because of their overwhelming commonality and unique identifying body parts (e.g. the large elytra). Specifically, primarily wood boring species were selected from the William F. Barr Entomological Museum (College of Agricultural and Life Sciences, University of Idaho, Moscow, Idaho), controlling for the location and year collected within a taxa. The collections at the William F. Barr Entomological Museum date back to 1893. The holdings are a substantial regional and national resource for specimens from the intermountain west, in addition to containing a worldwide representation of select taxa. Given its breadth of specimens, the museum provided a unique opportunity to examine several families beetles. Table 1 provides the taxa, year, collection location, number of individuals and abbreviations for the species used in this study.

Table 1. List of Coleoptera taxa used with year collected, location collected, number of individual specimens measured and respective abbreviations

Family: Genus Species	Year	Location	Number of individuals	Abbreviation
Buprestidae: *Buprestis lyrata*	1982	5 min west of paradise pt. Palouse range ID	24	PC

Buprestidae: *Dicerca tenebrica*	1954	bear creek camp 10 min north of Leslie	20	DI
Buprestidae: *Melanophila atropurpurea*	2012	I 84 rest stop nearest to Utah border	18	ME
Buprestidae: *Trachykele blondeli blondeli*	1966	Marion county Oregon	10	TR
Cerambycidae: *Callidium* sp.	1990	Clark mountain	18	CA
Cerambycidae: *Desmocerus piperi*	1963	lost trail pass Idaho	18	DE
Cerambycidae: *Prionus californicus*	2008	Parma research center	27	PR
Cerambycidae: *Spondylis upiformis*	1976	3.4 miles west of clarkia Idaho	19	SP
Lucanidae *Lucanus mazama*	2006	Kanal Utah	22	LM
Lucanidae: *Lucanus capreolus*	2006	Camden AR	7	LC
Trogossitidae: *Temnochila chlorodia*	1977		26	TE

The data collection was carried out in dark room laboratories in order to control the lighting. Specimens were enclosed in an area painted with Krylon Ultra-Flat black paint. This paint was chosen because it does not register on the spectral instrument used and, therefore, provides a null background for the desired readings.

Each insect was attached through the right, front wing with a standard insect mounting pin. A description of the insect's scientific name, collection date and location was attached below the insect. Spectrometer readings of insects were collected with a FieldSpec® Pro Full Range model, with the spectral acquisition range of 350 to 2500 nanometers (nm). This instrument has a resolution of 3 nm at 700 nm, and 10 nm at 1400 nm and 2100 nm (ASD Inc. 2012).

Fiber optics, connected to the spectrometer, were maintained at a 3 centimeter distance from the specimens and manipulated through a pistol grip control affixed at a 90^{o} angle to the specimen’s target area. Each specimen was sequentially illuminated across a spectrum of 400 to 700 nm. The light source used was a Smith Vector Corp Photographic Light Model 750-SG outfitted with a full spectrum light bulb and placed at a 45^{o} angle, one meter away from the specimens. The experimental setup is presented in figure 1.

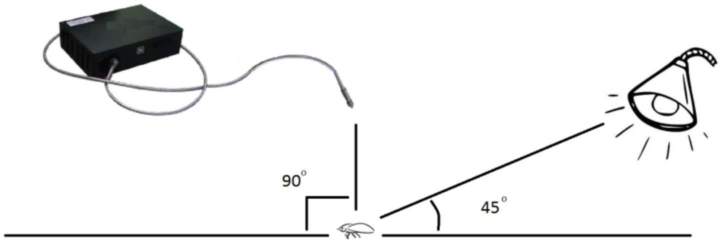

Figure 1. The diagram shows the experimental set-up of light, specimen & spectrometer. This set up was used to reduce the direct light from the source while still fully illuminating the specimen.

Each specimen’s elytral spectral relative reflectance (%) was recorded at each wavelength (nm). The relative reflectance was the percentage of a white 8^{o} hemispherical spectral reflectance factor for SRT-99-050. After every third spectrometer reading, the hemispherical spectral reflectance factor was recorded. This ensured that the machine’s calibration remained constant.

Each specimen was measured three times with the spectrometer. The instrument recording software (R^3), itself, averaged three additional shots for each of these observations. Following data collection, the three manual observations per specimen were averaged, effectively giving one spectral data point based on nine spectrometer readings. This was intended to reduce any potential measurement errors. Eleven taxa were measured, and each included approximately the same number of male and female specimens. Replication (individuals per taxa) ranged from a minimum of three to a maximum of 12, for a total of 210 insects. An overall multispectral database was subsequently created from these specimens that encompassed reflectance measurements of 2150 wavelengths.

Statistical Analysis

Finite Mixture Models (FMM)

In order to approximate the multi-modal spectral data series, finite mixture models were used, assuming normal distribution components. Finite mixture models have the general form of $\sum_{i=1}^{q} p_i \, f_i(x_j)$. Assuming a normal distribution model basis, the finite mixture model becomes:

$$\mathrm{pr}(x_j) = \sum_{i=1}^{q} p_i \left(\frac{1}{\sqrt{2\pi\sigma_i^2}}\right) e^{\left(\frac{-\left(x_{ij}-\mu_i\right)^2}{2\sigma_i^2}\right)} = p_1 \left(\frac{1}{\sqrt{2\pi\sigma_1^2}}\right) e^{\left(\frac{-\left(x_{1j}-\mu_1\right)^2}{2\sigma_1^2}\right)} + \cdots +$$

$$p_q \left(\frac{1}{\sqrt{2\pi\sigma_q^2}}\right) e^{\left(\frac{-\left(x_{qj}-\mu_q\right)^2}{2\sigma_q^2}\right)} \qquad (1)$$

where μ_i is the i^{th} component peak (mean and mode), σ_i is the associated standard deviation, and the subscript q denotes the maximum number of components, x_{ij} represents the observed response of the i^{th} component at the j^{th} wavelength, and p_i is the proportion accounted for by the i^{th} mixture component, where $0 \leq p_i \leq 1$, and $\sum_{i=1}^{q} p_i = 1$, satisfying the necessary conditions for a complete probability distribution. The above represents a univariate method for identifying the multiple peaks in the original wavelength data. Finite mixture models have been previously used to describe and compare other biological responses, such as the length distributions of mountain white fish (Shafii et al. 2010). Also, Royle and Link (2005) created a Gaussian mixture model of Anuran call surveys to predict species abundance.

Procedure FMM in SAS 9.3 was used to fit a varying number of normal curves mixture model components separately to the data for 22 separate taxa and gender groups within the data, i.e. 11 species, both male and female. Following adequate model estimation, the wavelengths at the corresponding model component peaks, $\mu_{i,}$ were chosen as the basis for further analysis. This provided a way to reduce the number of wavelengths from 2150 down to a more manageable database where false positives were less likely to occur.

Principal Component analysis (PCA)

Principal component analysis is designed to take multidimensional data sets and reduce their dimensions by determining one or more linear combinations of the variables that account for the largest variation in the data. In our case, wavelengths $x_1, x_2, x_3, \ldots, x_p$, selected by the FMM procedure may potentially be correlated and can lack the ability to show any underlying data structure individually. PCA can help define potential unobserved latent variables in the data by reducing the inherent dimensions of the problem through a centering of the data origin to $\bar{\mathrm{x}}$ and subsequently rotating the data using:

$$\mathrm{z_l} = \mathrm{A}(\mathrm{x_p} - \bar{\mathrm{x}}) \qquad (2)$$

where A is an orthogonal matrix of coefficients, and $\mathrm{z_l}$ is $\mathrm{x_l}$ rotated. The rotation is done such that $(\mathrm{z_1}, \mathrm{z_2}, \mathrm{z_3}, \dots, \mathrm{z_p})$ are uncorrelated to one another and the covariance matrix of $(\mathrm{z_1}, \mathrm{z_2}, \mathrm{z_3}, \dots, \mathrm{z_p})$ will be defined as:

$$\mathrm{S_z} = \mathrm{ASA'} = \begin{bmatrix} \mathrm{s_{z1}^2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathrm{s_{zp}^2} \end{bmatrix} \qquad (3)$$

$\mathrm{s_{zp}^2}$ is the Eigenvalue, λ_p, where λ_1 has the largest variance and λ_p has the smallest variance (Rao 1964). The SAS procedure PRINCOMP was used for PCA estimation, based on the underlying variance-covariance matrix.

Other ordination techniques, including multidimensional scaling based on a dissimilarity matrix were also attempted. However, the results were not satisfactory, and hence are not reported here.

Multivariate Discriminant Analysis

Linear discriminant analysis

Normal discriminant analysis is typically of the form

$$\mathrm{L_k} \propto (2\pi)^{-\frac{\mathrm{k}}{2}}|\mathrm{V_k}|^{-\frac{1}{2}}\exp(-0.5\mathrm{M_k}) \qquad (4)$$

With $\mathrm{L_k}$ representing the likelihood that an individual belongs to species k and $\mathrm{V_k}$ the variance-covariance matrix obtained from the k^th species. $\mathrm{M_k}$ is interpreted as the Mahalanobis distance given by:

$$\mathrm{M_k} = (\mathrm{d} - \mathrm{x_k})'\mathrm{V_k^{-1}}(\mathrm{d} - \mathrm{x_k}) \qquad (5)$$

The Mahalanobis distance measures the distance between the data response vector, d, and a known vector of responses from the k$^{\text{th}}$ species, $\mathrm{x_k}$ (Lachenbruch 1979).

Multidimensional normal discriminant analysis has aided in the identification of insects prior to this study. For example, the identification of the Africanized honey bees in the U.S. based upon the insect’s characteristics (Daly and Balling 1978).

Bayesian Nearest Neighbor or K-Nearest Neighbor

The nearest neighbor rule was first introduced by Fix and Hodges in 1951. Subsequently, a nearest neighbor discriminant analysis was proposed for selecting the Kth nearest points using the distance function:

$$\mathrm{M}_k = (\mathrm{x_m} - \mathrm{x_n})' \mathrm{S}_{\mathrm{pl}}^{-1} (\mathrm{x_m} - \mathrm{x_n}) \qquad (6)$$

where $\mathrm{S}_{\mathrm{pl}}^{-1}$ is the inverse of the pooled sample variance-covariance matrix from the defined sample, $\mathrm{x_m}$ is a data point of interest, and $\mathrm{x_n}$ is all other data points. The purpose of this technique is to classify each $\mathrm{x_m}$, using the k points nearest to $\mathrm{x_{n.}}$. That is, if the majority of the k points belong to group 1, assign $\mathrm{x_m}$ to group 1; otherwise, assign $\mathrm{x_m}$ to another group, etc. K-Nearest Neighbor Discriminant Analysis, or Non-parametric Discriminant Analysis, dispenses with the need to make probabilistic assumptions for likelihood determinations.

Bayesian discriminant analysis modifies (4) through the addition of a prior assumption on group assignments. A base model for this would use a uniform or uninformed prior for discriminant analysis resulting in the following posterior distribution:

$$\mathrm{p}\big(\text{insect is in } \mathrm{k}^{\mathrm{th}} \text{ species}\big|\mathrm{d}\big) = \frac{\mathrm{q_k L_k}}{\sum \mathrm{q_k L_k}} \qquad (7)$$

This will produce a probability between 0 and 1 with $\mathrm{q_k}$ as the prior probability, where k represents the number of species as follows:

$$\mathrm{q_k} = \frac{1}{\mathrm{k}} \qquad (8)$$

Alternatively, a proportional prior for discriminant analysis can be defined as the proportion of observations from each group in the input data set (Hinich 1979). When data are balanced, the proportions for each group are the same, so this method will be equivalent to using a uniform prior. This can be seen in (7), where q_k is the prior probability for species k, defined as follows:

$$q_k = \text{number of individuals in species } k \text{ / total number of individuals} \qquad (9)$$

The nearest neighbor methods aid in the prediction of species based upon multivariate spectrometer readings. Procedure DISCRIM in SAS 9.3 was used for all discriminant analyses estimations and validations.

Validation

Internal Validation

Bootstrap is a resampling technique, with replacement, that is done when one is unsure about the behavior of the target population (Efron 1979). By randomly selecting a subsample (X_i*) from the sample (X_i), a new sample is produced which is selected from a known population. By analyzing the relationship between the sample and subsample, conclusions can be drawn about the actual population. Gathering data on the population would require a census, which would be impractical for a subject such as Coleoptera due to the number of possible individuals. Bootstrap simulation procedures, therefore, provide a practical means of assessing the differences between Coleopteran species and the analyses carried out in this study.

An internal bootstrap of proportional discriminant analysis was performed through bootstrap sampling with replacement and data splitting. Data splitting was completed using 60% of the data to construct the model, while the remainder of the data were utilized for validation. The bootstrap sample, X_i*, was selected from the data X_i at predefined proportions of sex and species in the database. For each bootstrap sample, two types of misclassification were possible: omission (Type I), or commission (Type II). An error of omission occurred when an observation, X_i*, was classified outside of its true type, while an error of commission occurred when an observation was placed in the wrong type. Confidence intervals, means, and standard deviations were created from B = 5000 bootstrap simulations.

External Validation

A new independent database was created from 180 insects of the same species that were not previously sampled. External validation was carried out using these data and the same methodology as the internal validation, that is, a bootstrap simulation of discriminant analysis assuming a proportional prior. Unlike the first database, however, the insects chosen for inclusion were not controlled for location or year. This validation provided a robust confirmation of the adequacy of the estimated discriminant model.

Results and Discussion

Finite Mixture Model (FMM)

As an objective means of identifying the peaks (modes) of the spectral distributions, equation (1) was fitted separately to each species-gender combination assuming spectral reflectance values were proportional to their probability of being observed.

The number of normal curve components was allowed to vary and were ultimately estimated from the distribution of the data. The final number of components ranged from 3 to 8 distributions per species - gender group. Thus, each of the 22 groups had a different set of fitted normal curves. The peaks (e.g. the means) were selected from the normal curves as a technique for quantifying the strongest wavelengths in the spectrum. The set of peak bands from each spectrum could then be used as a basis for comparing species-gender combinations. An example for the female *Lucanus capreolus* data set is given in Figure 2. In that case, six peaks were identified and ranged from 977 nm to 2133 nm.

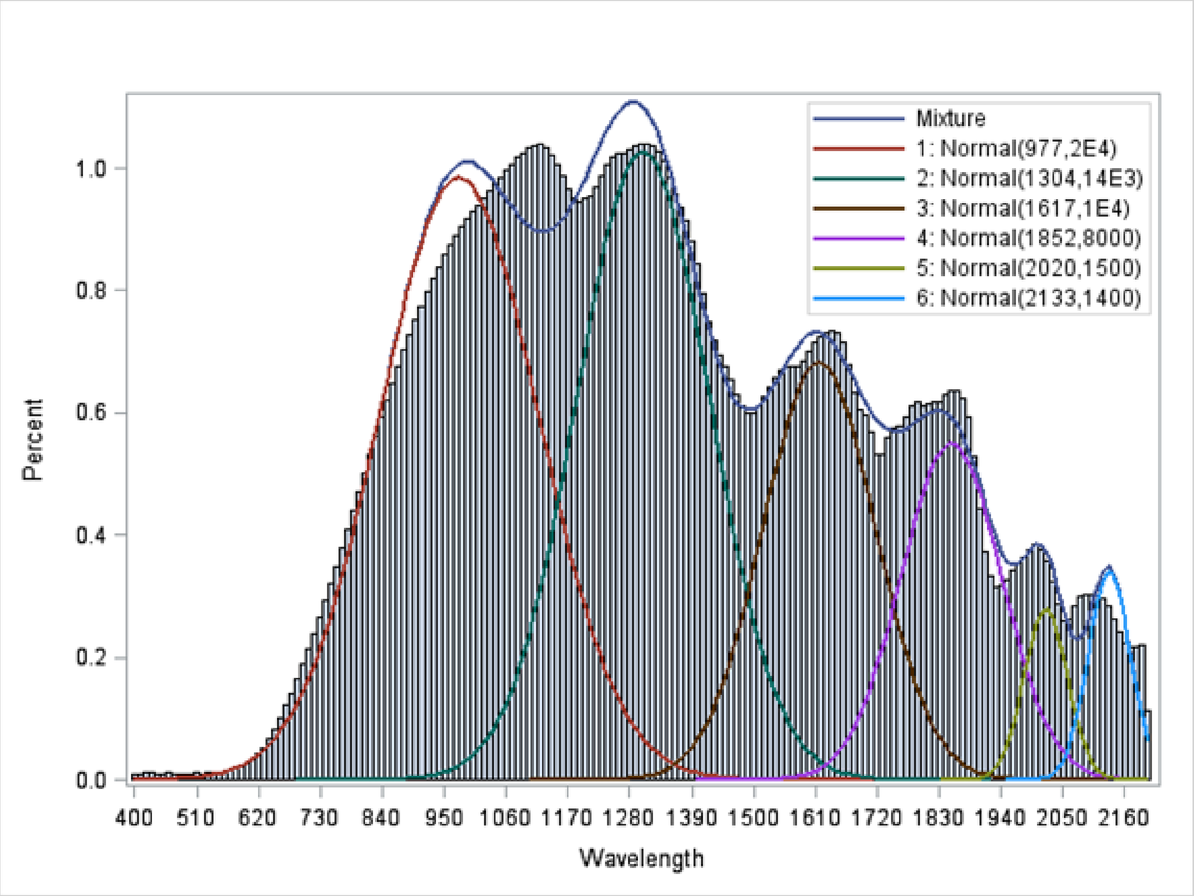

Figure 2. Example fit of normal curves fitted to the female *Lucanus capreolus* distribution.

Overall, a large number of peaks were identified. To assess any commonalities among the 22 species-gender combinations, peak placement in relation to the wavelength was graphed (Figure 3).

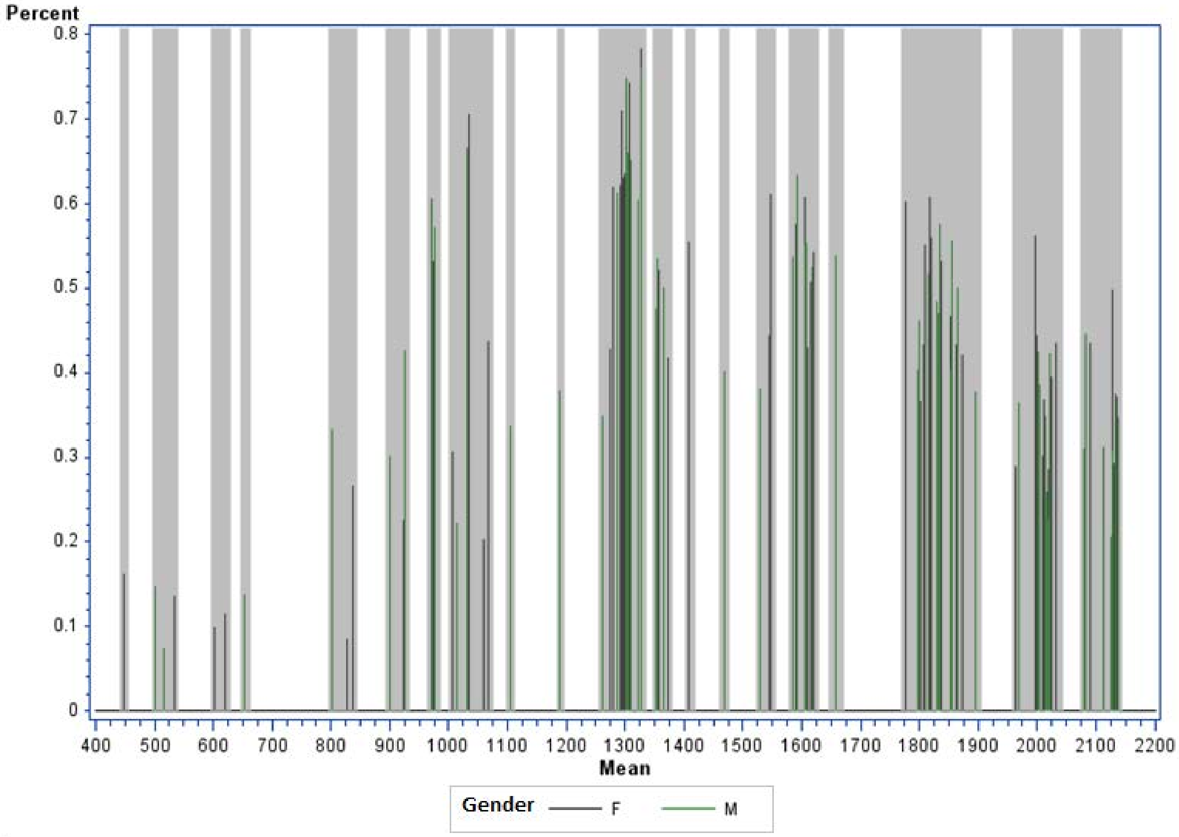

Figure 3. Peak locations based on the Finite Mixture Model in relative reflectance (percent) by wavelength. The green lines are male and the black lines are female. The lines are representing the relative reflectance at peak locations as identified by equation (1). The grey shaded area is emphasizing the aggregation of the 18 peak observations.

From figure 3, it was determined that the peaks showed some aggregation and hence, it led to the creation of 18 common peaks$(R_1, R_2, R_3, \ldots, R_{18})$, i.e. 18 different bandwidths selected as a common dataset across species. A detailed outline of the 18 variables generated from FMM procedure and their corresponding bandwidths are given in Table 2.

Table 2. Detailed outline of the 18 variables generated from FMM procedure
and their corresponding bandwidths

Bandwidth	Lower Limit	Upper Limit	New Variable	Wavelength	Mean
10	440	450	R1	445	445
50	500	550	R2	525	525
60	600	660	R3	630	630
50	800	850	R4	825	825
30	900	930	R5	915	915
20	960	980	R6	970	970
125	1000	1125	R7	1063	1062.5
50	1175	1225	R8	1200	1200
80	1250	1330	R9	1290	1290
30	1350	1380	R10	1365	1365
25	1400	1425	R11	1413	1412.5
20	1460	1480	R12	1470	1470
25	1525	1550	R13	1538	1537.5
45	1580	1625	R14	1603	1602.5
25	1650	1675	R15	1663	1662.5
125	1775	1900	R16	1838	1837.5
90	1950	2040	R17	1995	1995
65	2075	2140	R18	2108	2107.5

The new 18-variable database provided a manageable number of variables for subsequent analyses.

Principal Component Analysis (PCA)

The relationship between the 18 variables created by FMM was investigated using principle components analysis (PCA). The analysis implemented equations (2) and (3) in order to obtain the Eigen vectors or PCA axis. The first PCA axis explained 66.84% and the second PCA axis explained 19.88% of the total variability in the data. The third axis explained 10.3% of the variability, while the amount of variability explained by PCA axes 4 through 18 was less than 5%. The retention of three PCA axis, or a three dimensional space, explained 96.3% of the variability. The third axis would normally not have been considered, however, the 10.3% of the variability explained by that axis provided an increased separation between species and genders. The scree plot further detailing the first six PCA axes is given in Figure 4.

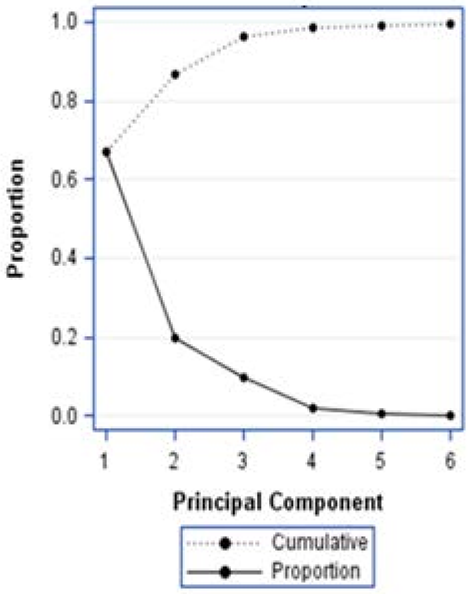

Figure 4. The PCA scree plot showing the variance explained by the first six PCA axes.

Plots of the resulting first three PCA axes, coded by species, are given in Figures 5 and 6. The ellipses represent an approximate 95% confidence region for each species, assuming bivariate normality. The separation of species seen in PCA axis two vs. PCA axis three (Figure 6) and PCA axis one vs. PCA axis two (Figure 5) is more closely related to PCA axis two. In both figures (5 and 6), LC (*Lucanus capreolus*), LM (*Lucanus mazama*), and PR (*Prionus californicus*) separate from the rest of the species. The angle of their respective ellipses also varies from other species in these plots.

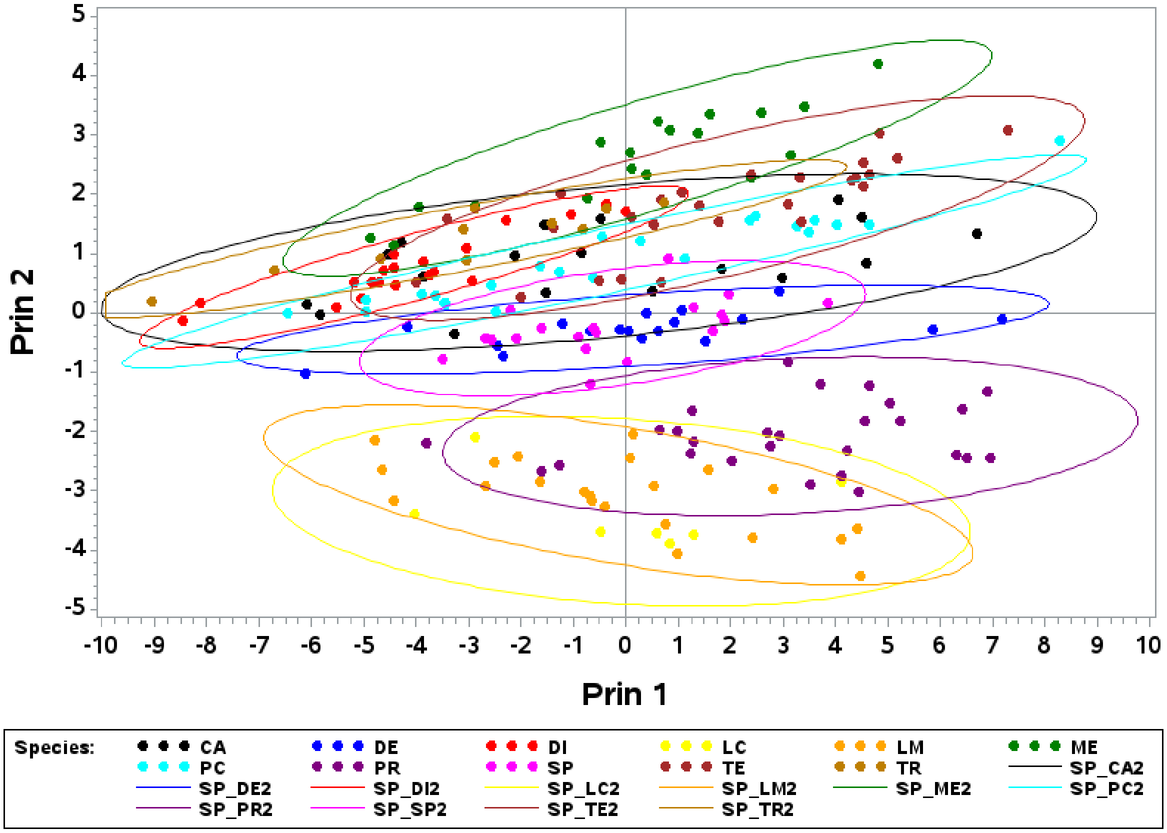

Figure 5. The 95% prediction ellipse displays PCA axis one vs PCA axis two. The points are the original data points projected into the PCA space. The abbreviations represent the following species: *Callidium* sp. (CA, SP_CA2), *Desmocerus piperi* (DE, SP_DE2), *Dicerca tenebrica* (DI, SP_DI2), *Lucanus capreolus* (LC, SP_LC2), *Lucanus mazama* (LM, SP_LM2), *Melanophila atropurpurea* (ME, SP_ME2), *Buprestis lyrata* Casey (PC, SP_PC2), *Prionus californicus* (PR, SP_PR2), *Spondylis upiformis* (SP, SP_SP2), *Temnocheila chlorodia* (TE, SP_TE2), *Trachykele blondeli blondeli* (TR, SP_TR2).

In Figure 6, species DE appears to be at a 90° angle to other species, particularly species PR, giving some indication that they are independent of one another. Also, LM and DI are mirror angles from one another, separated by 180 degrees, thus implying that they are negatively correlated based upon the sign of their respective PCA loadings. For figures 5 and 6, the separation of species and gender were not clear when viewing all 22 groups. However, as seen in Figure 7, plotting the data separately by individual species can discern some separation by gender.

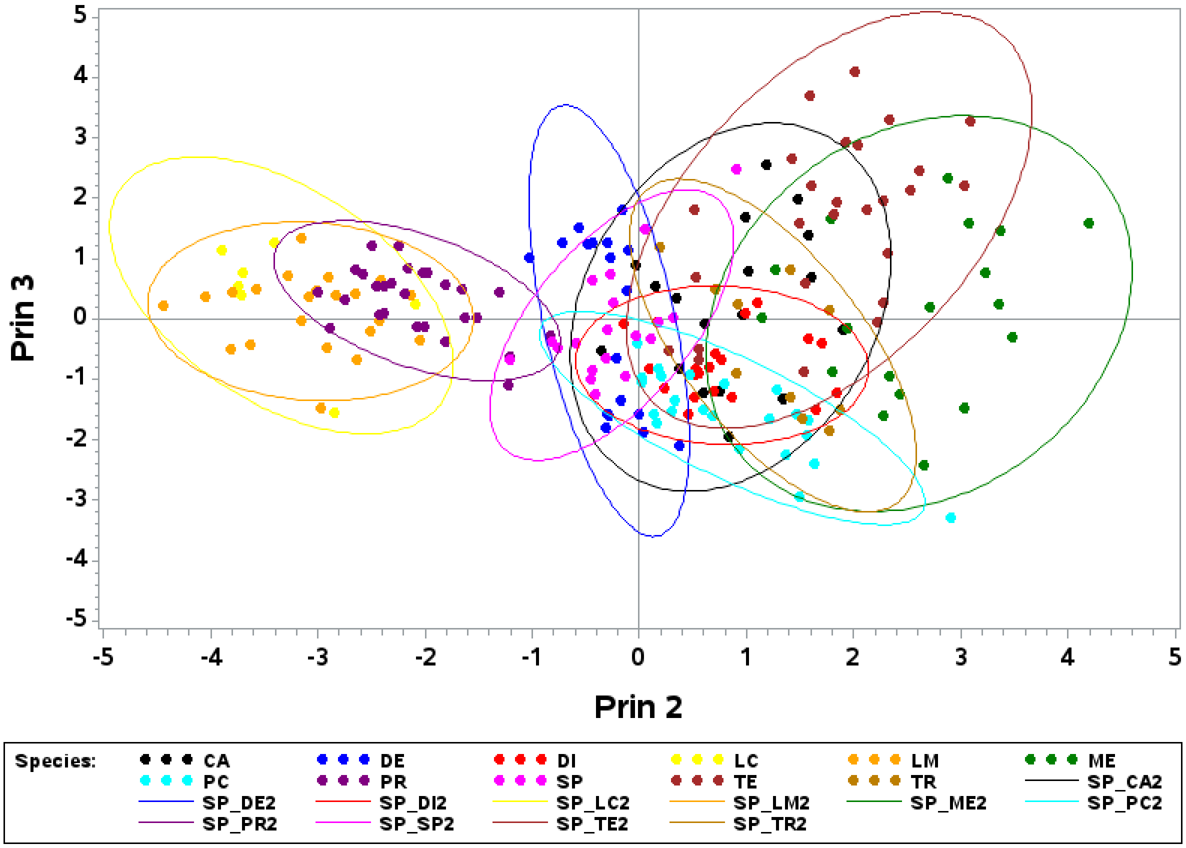

Figure 6. The 95% prediction ellipse displays PCA axis two vs PCA axis three. The points are the original data points projected into the PCA space. The ellipses are the 95% confidence interval assuming bivariate normality. The abbreviations represent the following species: *Callidium* sp. (CA, SP_CA2), *Desmocerus piperi* (DE, SP_DE2), *Dicerca tenebrica* (DI, SP_DI2), *Lucanus capreolus* (LC, SP_LC2), *Lucanus mazama* (LM, SP_LM2), *Melanophila atropurpurea* (ME, SP_ME2), *Buprestis lyrata* Casey (PC, SP_PC2), *Prionus californicus* (PR, SP_PR2), *Spondylis upiformis* (SP, SP_SP2), *Temnocheila chlorodia* (TE, SP_TE2), *Trachykele blondeli blondeli* (TR, SP_TR2).

In Figure 7, *Desmocerus piperi* (DE) indicates separation between male and female. The ellipse shapes are different indicating that males are better described by PCA axis 2 while the females are described by both PCA axes 2 and 3.

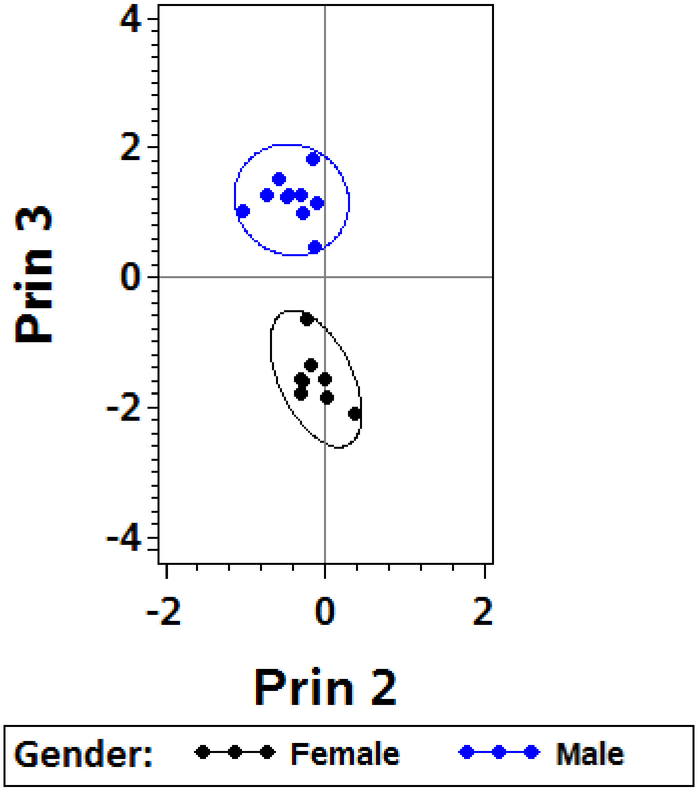

Figure 7. The 95% prediction intervals separating male and female of *Desmocerus piperi* (DE) when viewed by PCA axis two and three.

The PCA loadings for each variable by wavelength$(R_1, R_2, R_3, \ldots, R_{18})$, are plotted in Figure 8. The first PCA axis (red), primarily explains the overall variability through a positive loading value across the spectrum. The second PCA axis (green) explains data variability by providing an approximate inverse to the respective loadings of the third principal component (yellow). While the true meaning of these axes is purely speculative, the inverse behavior seen between PCA axes two and three and the relationship shown in Figure 8 may indicate some gender differentiation based on these axes.

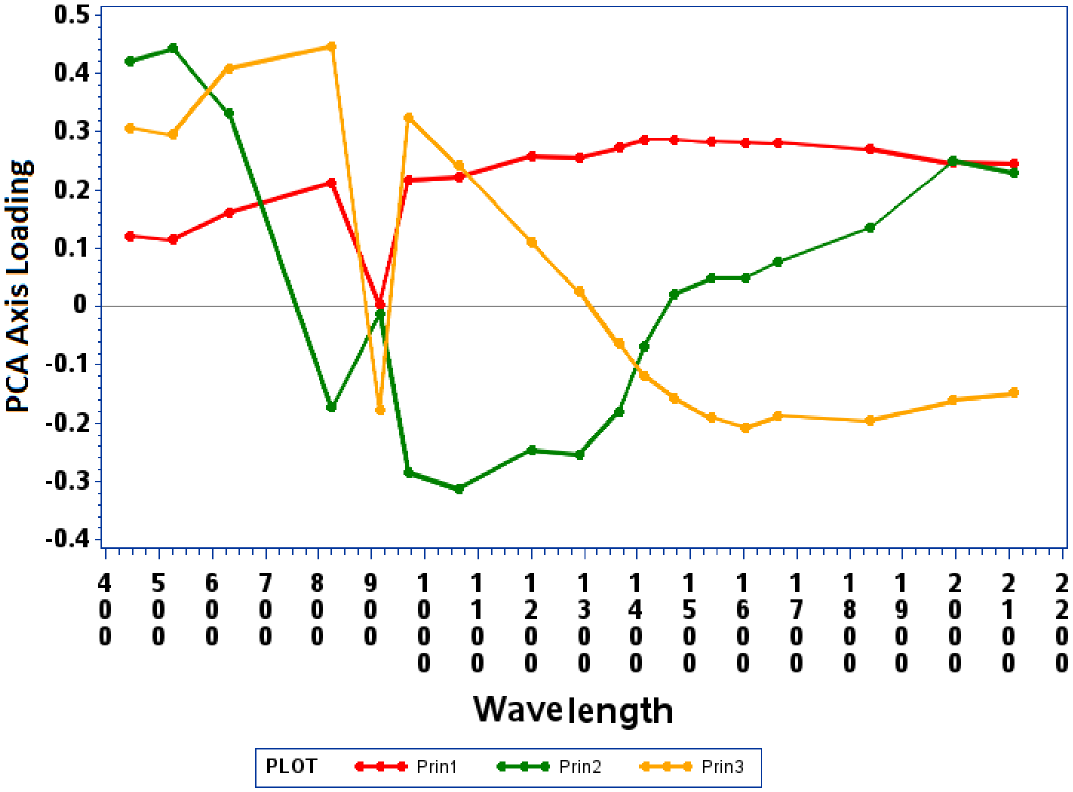

Figure 8. Principal component loadings by the wavelengths (R_1, R_2, … , R_{18}) is defined by principal component axis 1 (red) 2 (green) and 3 (yellow).

Figure 9 displays a heat map of the correlation matrix for the peak wavelength values. The values of correlation along the diagonal are one, or very close to one (white). This signifies that variables (peak wavelength values) close to one another are highly correlated. The lower correlation values observed betweenR_1, R_2, R_3, or rather the visual spectrum, (400 to 700 nm), verses R_4 through R_{16} does not correlate with the near infrared spectrum (800 – 1800 nm). It is unexpected, however, that the visual spectrum,R_1, R_2, R_3, is correlated to R_{17}, & R_{18}. The visual spectrum encompasses what humans can see with their naked eye, violet, blue, green, yellow, orange and red. Insects can sense a wider spectrum, outside of the human's capabilities, which range from ultraviolet (350 nm) to red (700 nm) (Stark and Tan 1982). The near infrared spectrum describes the bonds between molecules, which may indicate the composition of the chitin. The wavelengths 1654, 1560 and 1310 nm are known identifiers of beetle's chitin components (Liu et al. 2012). Chitin composes insects elytra, and the wavelengths that closely match are R_{15} (1654), and R_{13} (1560).

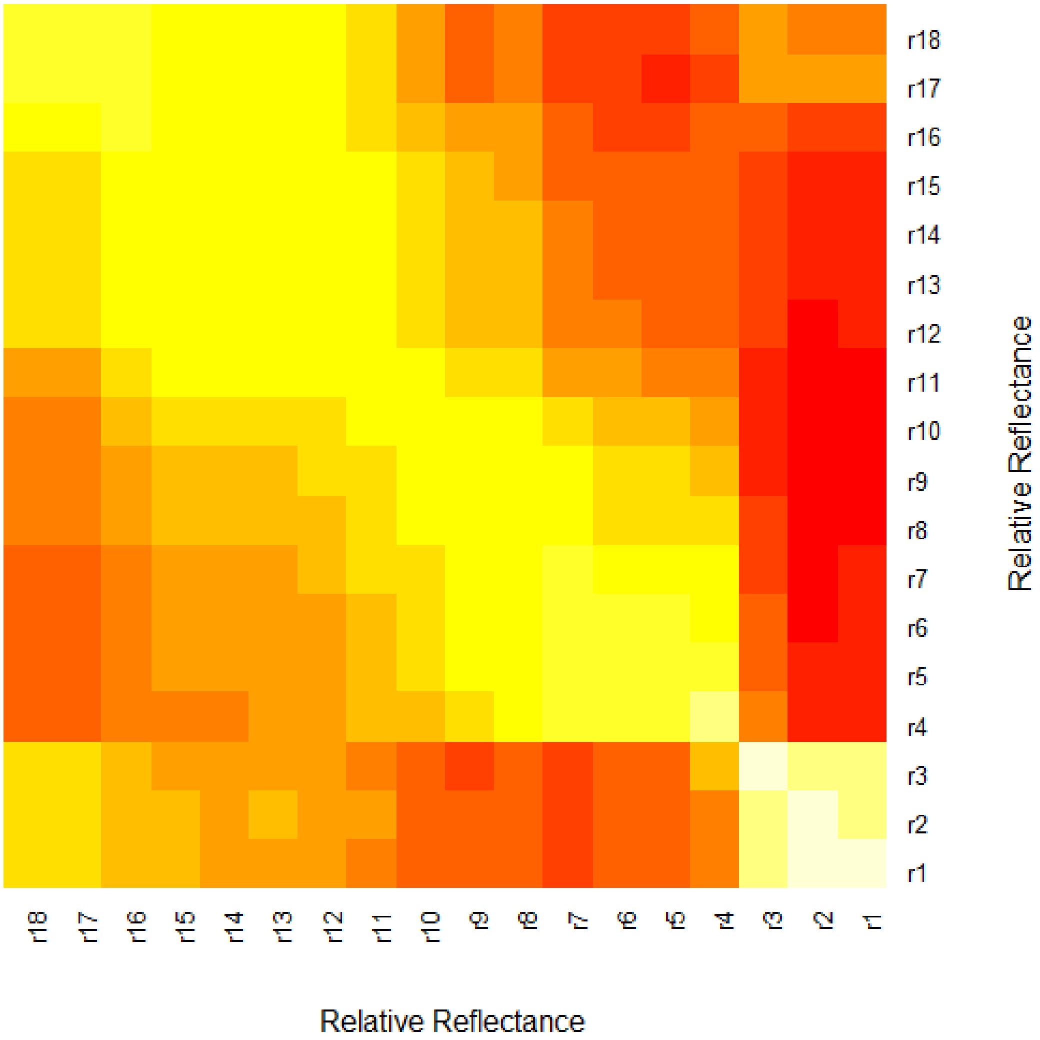

Figure 9. The heat map of the correlation matrix indicating the correlation between peak wavelength values. The wavelengths closely correlated to one another are yellow; while the lower correlation values are red. The color values are assigned based upon their z-score value.

The PCA analyses attempted to reduce the dimensions of the data while separating species and gender. While the genders were not always clearly separated from one another, the species do appear to separate. The classification of species however, required consideration of additional statistical techniques such as multivariate discriminant functions.

Multivariate Discriminate Analysis

Linear discriminant analysis

Linear discriminant analysis was used to classify each species based on the eighteen variables of $(R_1, R_2, R_3, \dots, R_{18})$ and the assumption of multivariate normality. This was completed using equations (4) and (5).

The multivariate normal discriminant analysis resulted in a misclassification rate of 4.14% of individuals incorrectly classified as the wrong species. The majority of the error originated from the comparison of species LC (*Lucanus capreolus*) to LM (*Lucanus mazama*), with a 27.27% misclassification rate. This misclassification might be attributed to LC having a small number of observations and the fact that LC and LM are taxonomically very similar. The misclassification between CA (*Callidium*) and TE (*Trachykele blondeli blondeli*) is thought to stem from the similar blue iridescent color they share and the low sample size of TE. The small misclassification rate between PC (*Buprestis lyrata*) and DI (*Dicerca tenebrica*) is thought to stem from their very similar elytra. The complete classification results from the multivariate normal discriminant analysis are provided in Table 3.

Table 3. Linear discriminant analysis misclassification results of individual species. The cells in the table contain two numbers, the top number is the number of individuals, and the bottom number is the percent classified of the specific species. The abbreviations represent the following species: *Callidium* sp. (CA), *Desmocerus piperi* (DE), *Dicerca tenebrica* (DI), *Lucanus capreolus* (LC), *Lucanus mazama* (LM), *Melanophila atropurpurea* (ME), *Buprestis lyrata* Casey (PC), *Prionus californicus* (PR), *Spondylis upiformis* (SP), *Temnocheila chlorodia* (TE), *Trachykele blondeli blondeli* (TR).

Number of Observations and Percent Classified into Species												
From Species	CA	DE	DI	LC	LM	ME	PC	PR	SP	TE	TR	Total
CA	17 94.4	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	1 5.6	0 0	18 100
DE	0 0	18 100	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	18 100
DI	0 0	0 0	19 95	0 0	0 0	0 0	1 5	0 0	0 0	0 0	0 0	20 100
LC	0 0	0 0	0 0	7 100	0 0	0 0	0 0	0 0	0 0	0 0	0 0	7 100
LM	0 0	0 0	0 0	6 27.3	16 72.7	0 0	0 0	0 0	0 0	0 0	0 0	22 100
ME	0 0	0 0	0 0	0 0	0 0	18 100	0 0	0 0	0 0	0 0	0 0	18 100
PC	0 0	0 0	0 0	0 0	0 0	0 0	24 100	0 0	0 0	0 0	0 0	24 100
PR	0 0	0 0	0 0	0 0	0 0	0 0	0 0	27 100	0 0	0 0	0 0	27 100
SP	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	19 100	0 0	0 0	19 100
TE	2 7.7	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	24 92.3	0 0	26 100
TR	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	10 100	10 100
Total	19 9.1	18 8.6	19 9.1	13 6.2	16 7.7	18 8.6	25 11.9	27 12.9	19 9.1	25 11.9	10 4.8	209 100
Priors	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	

In figure 10, the heat map of the linear discriminant function, the location of the highest (white) and lowest coefficients (red) of the original variables are at $R_{11}, R_{12}, R_{13}, \& \, R_{14}$. It can be inferred that the majority of the information provided by the discriminant function comes from these variables, or rather the near-infrared spectrum. One of the variables, R_{13}, contributing a higher loading is considered a wavelength identifying the chitin molecule particularly the amide II of N-H bond (Liu et al. 2012).

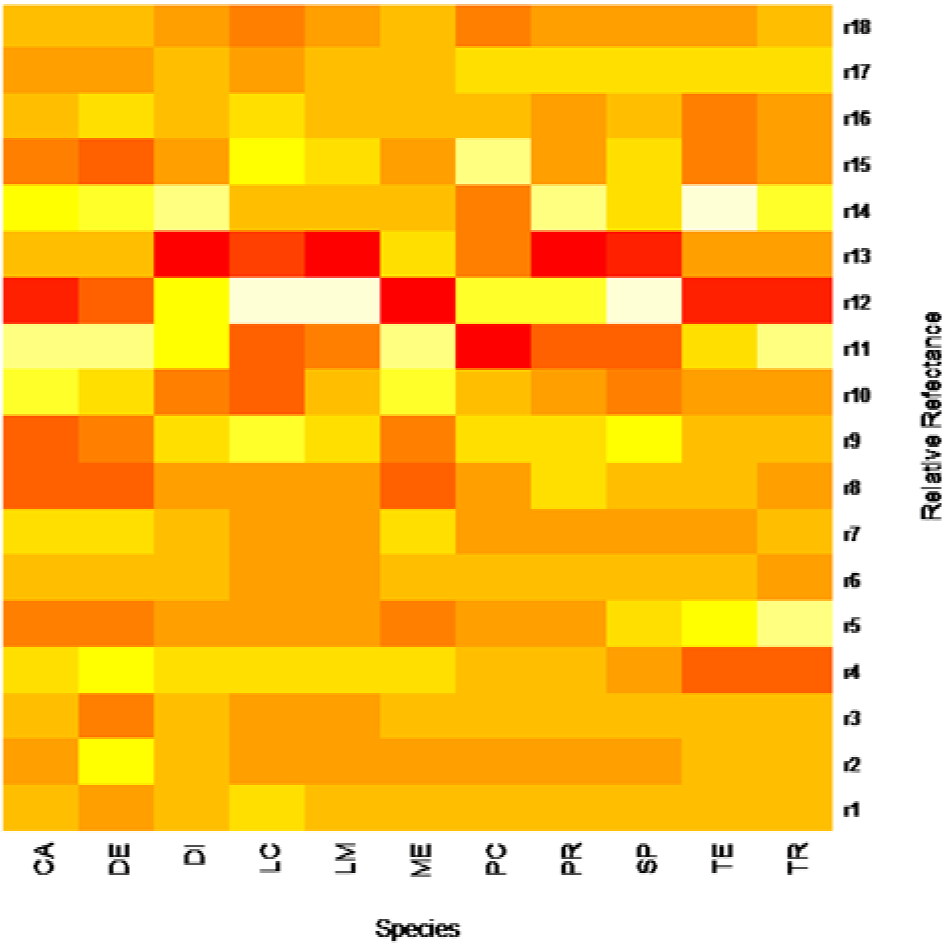

Figure 10. The Heat Map of the Linear Discriminant Function for individual Species. Correlation colors are assigned based upon their z-score value, with low z-score given red and high z-score given white or yellow. The abbreviations represent the following species: *Callidium* sp. (CA), *Desmocerus piperi* (DE), *Dicerca tenebrica* (DI), *Lucanus capreolus* (LC), *Lucanus mazama* (LM), *Melanophila atropurpurea* (ME), *Buprestis lyrata* Casey (PC), *Prionus californicus* (PR), *Spondylis upiformis* (SP), *Temnocheila chlorodia* (TE), *Trachykele blondeli blondeli* (TR).

The misclassification rate produced by the multivariate linear discriminant analysis is below 0.05, signifying that this model works well as a classification for the data set. However, the underlying distribution of the wavelengths are often skewed, and hence, the assumption of multivariate normality may not have been appropriate. Thus, additional statistical approaches were considered in an attempt to relax the assumptions of normality.

Uniform Bayesian Prior Discriminant Analysis

Bayesian priors utilize ‘K’ nearest neighbor for analysis. K defines the number of nearest points utilized for discriminating the species differences. K < 3 was considered too few points and K > 10 too many points. The local maxima of misclassification occurred at K = 6 and was chosen for subsequent analysis. At K = 6, the misclassification rate was 3.8% with the highest rate of misclassification occurring between LC and LM at 27.27%. The total misclassification rate is 3.8% which is below 0.05, signifying that the uniform prior provides a good classification for these data. The tabulated results of the rate of misclassification for the uniform prior discriminant analysis are given in Table 4.

Table 4. Uniform prior discriminate analysis misclassification of individual species. The cells in the table contain two numbers, the top number is the number of individuals, and the bottom number is the percent classified of the specific species. The abbreviations represent the following species: *Callidium* sp. (CA), *Desmocerus piperi* (DE), *Dicerca tenebrica* (DI), *Lucanus capreolus* (LC), *Lucanus mazama* (LM), *Melanophila atropurpurea* (ME), *Buprestis lyrata* Casey (PC), *Prionus californicus* (PR), *Spondylis upiformis* (SP), *Temnocheila chlorodia* (TE), *Trachykele blondeli blondeli* (TR).

Number of Observations and Percent Classified into Species												
From Species	CA	DE	DI	LC	LM	ME	PC	PR	SP	TE	TR	Total
CA	18 100	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	18 100
DE	0 0	18 100	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	18 100
DI	0 0	0 0	20 100	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	20 100
LC	0 0	0 0	0 0	7 100	0 0	0 0	0 0	0 0	0 0	0 0	0 0	7 100
LM	0 0	0 0	0 0	6 27.3	16 72.7	0 0	0 0	0 0	0 0	0 0	0 0	22 100
ME	0 0	0 0	0 0	0 0	0 0	18 100	0 0	0 0	0 0	0 0	0 0	18 100
PC	0 0	0 0	0 0	0 0	0 0	0 0	24 100	0 0	0 0	0 0	0 0	24 100
PR	0 0	0 0	0 0	2 7.4	0 0	0 0	0 0	25 92.6	0 0	0 0	0 0	27 100
SP	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	19 100	0 0	0 0	19 100
TE	2 7.7	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	24 92.3	0 0	26 100
TR	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	10 100	10 100
Total	20 9.6	18 8.6	20 9.6	15 7.2	16 7.7	18 8.6	24 11.5	25 11.9	19 9.1	24 11.5	10 4.8	209 100
Prior	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	

Proportional Bayesian Prior Discriminant Analysis

Given the knowledge of the Coleoptera order, it becomes apparent that the species are not equally abundant. Proportional priors assume that the collections found at the University of Idaho Entomological museum are proportional to species abundance in their habitat. Equations (7) and (9) were utilized for the proportional prior Bayesian discriminant analysis.

The species misclassification rate was calculated using non-parametric Kth nearest neighbor at K = 6. The value at K=6 was chosen for the location of a local maxima of the misclassification, and for consistency with the previous method, the uniform prior. The proportional prior discriminant analysis error rate was 5.2%. While this value is very close to the misclassification values obtained under uniform priors, it is the most accurate given our knowledge about Coleoptera. The species misclassification rates are somewhat consistent with the uniform prior analysis with regard to species CA, LC, LM, and TE. The ‘other’ species category received several individuals accounting for the highest rate of species misclassification. The results of misclassification rates by species using proportional prior discriminant analysis are given in Table 5.

Table 5. Proportional prior discriminate analysis misclassification of individual species. The cells in the table contain two numbers, the top number is the number of individuals, and the bottom number is the percent classified of the specific species. The abbreviations represent the following species: *Callidium* sp. (CA), *Desmocerus piperi* (DE), *Dicerca tenebrica* (DI), *Lucanus capreolus* (LC), *Lucanus mazama* (LM), *Melanophila atropurpurea* (ME), *Buprestis lyrata* Casey (PC), *Prionus californicus* (PR), *Spondylis upiformis* (SP), *Temnocheila chlorodia* (TE), *Trachykele blondeli blondeli* (TR).

Number of Observations and Percent Classified into Species												
From Species	CA	DE	DI	LC	LM	ME	PC	PR	SP	TE	TR	Other	Total
CA	18 100	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	18 100
DE	0 0	18 100	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	18 100
DI	0 0	0 0	20 100	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	20 100
LC	0 0	0 0	0 0	1 14.3	1 14.3	0 0	0 0	0 0	0 0	0 0	0 0	5 71.4	7 100
LM	0 0	0 0	0 0	0 0	19 86.4	0 0	0 0	0 0	0 0	0 0	0 0	3 13.6	22 100
ME	0 0	0 0	0 0	0 0	0 0	18 100	0 0	0 0	0 0	0 0	0 0	0 0	18 100
PC	0 0	0 0	0 0	0 0	0 0	0 0	24 100	0 0	0 0	0 0	0 0	0 0	24 100
PR	0 0	0 0	0 0	0 0	0 0	0 0	0 0	27 100	0 0	0 0	0 0	0 0	27 100
SP	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	19 100	0 0	0 0	0 0	19 100
TE	1 3.85	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	24 92.3	0 0	1 3.8	26 100
TR	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	10 100	0 0	10 100
Total	19 9.09	18 8.61	20 9.57	1 0.48	20 9.57	18 8.61	24 11.48	27 12.92	19 9.09	24 11.48	10 4.78	9 4.31	209 100

Validation

Internal Validation

Further exploration into the proportional prior method was completed using a bootstrap simulation technique for the purpose of model validation. The proportional prior was used because it more accurately described the underlying population. The bootstrap simulation was created using 5,000 separate samples selected with replacement and data splitting. The dataset was split with 60% of the data being used to construct the model, and the remaining 40% used for the purpose of model validation. Each selection generated a species misclassification rate based upon a proportional prior discriminant analysis.

The distribution of misclassification by the proportional prior discriminant analysis bootstrap is given in Figure 11. The distribution can be approximated with a normal curve that has a mean of 0.0348 and a standard deviation of 0.011. The standard deviation is rather low, indicating that a majority of the data is within a small range of the mean. The fifth percentile error is 0.025 and the ninety-fifth percentile is 0.067. The median is located at 0.0341, which indicates that the skewness is low. The low skewness is another indicator that the mean and median agree, and that the normal curve is a reasonable approximation of the data in this case. The range of misclassification is low in value, validating the use of the proportional prior for this data set.

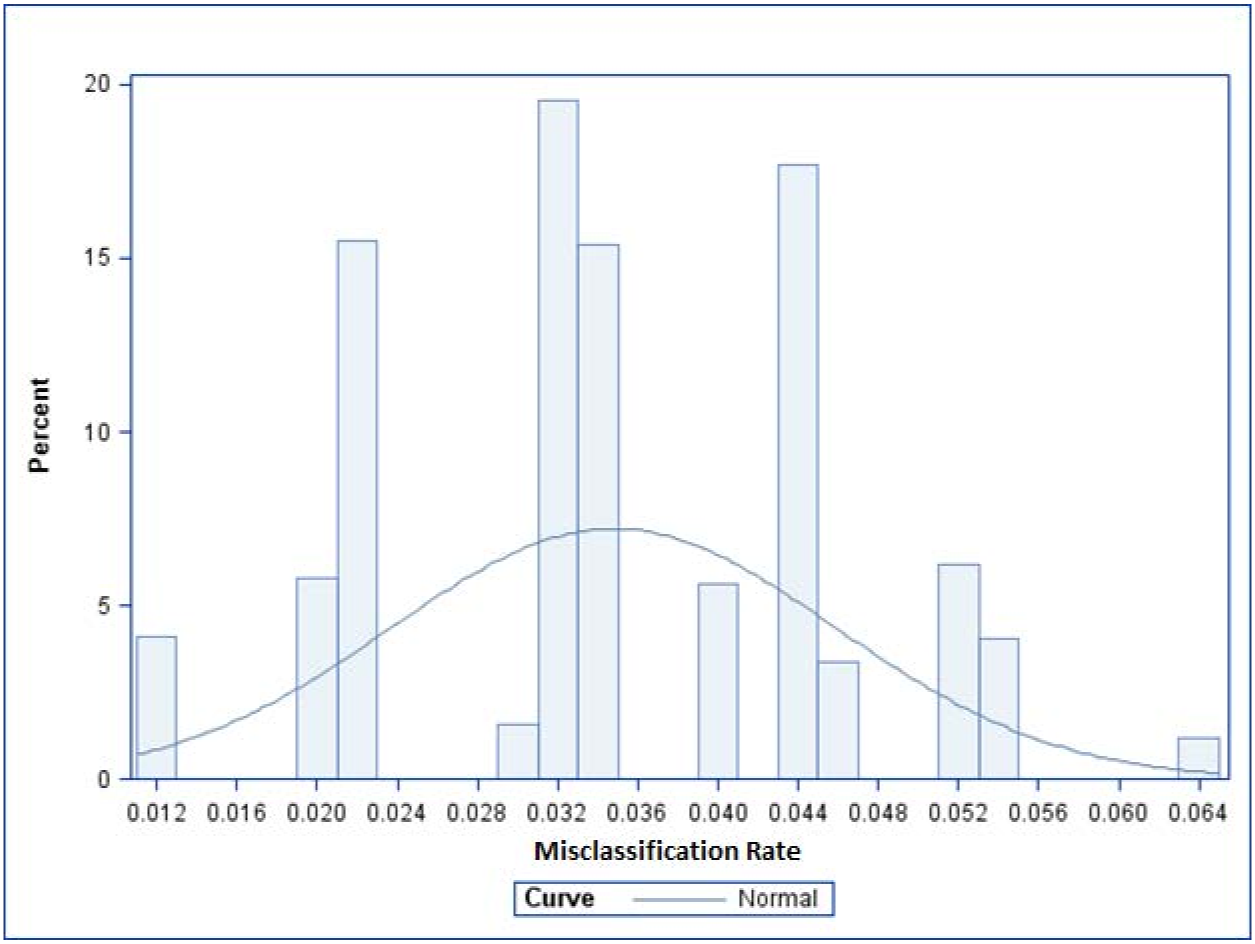

Figure 11. The distribution of species misclassification rate for the internal bootstrap is described using a normal approximation. Species misclassification rate has a mean of 0.0348 and a standard deviation of 0.011.

External Validation

External validation was performed on a new data set independent from the original database. The new data contained 187 insects of the previous taxonomic groups, not controlled for by location or year collected. When the 187 individuals were subjected to the proportional prior discriminate analysis, this resulted in a 4.28% observed external misclassification. When conducting a bootstrap simulation, the distribution of misclassification had a fifth percentile of

0% and the ninety-fifth percentile of 11.95%. The actual validation bootstrap distribution is given in Figure 12.

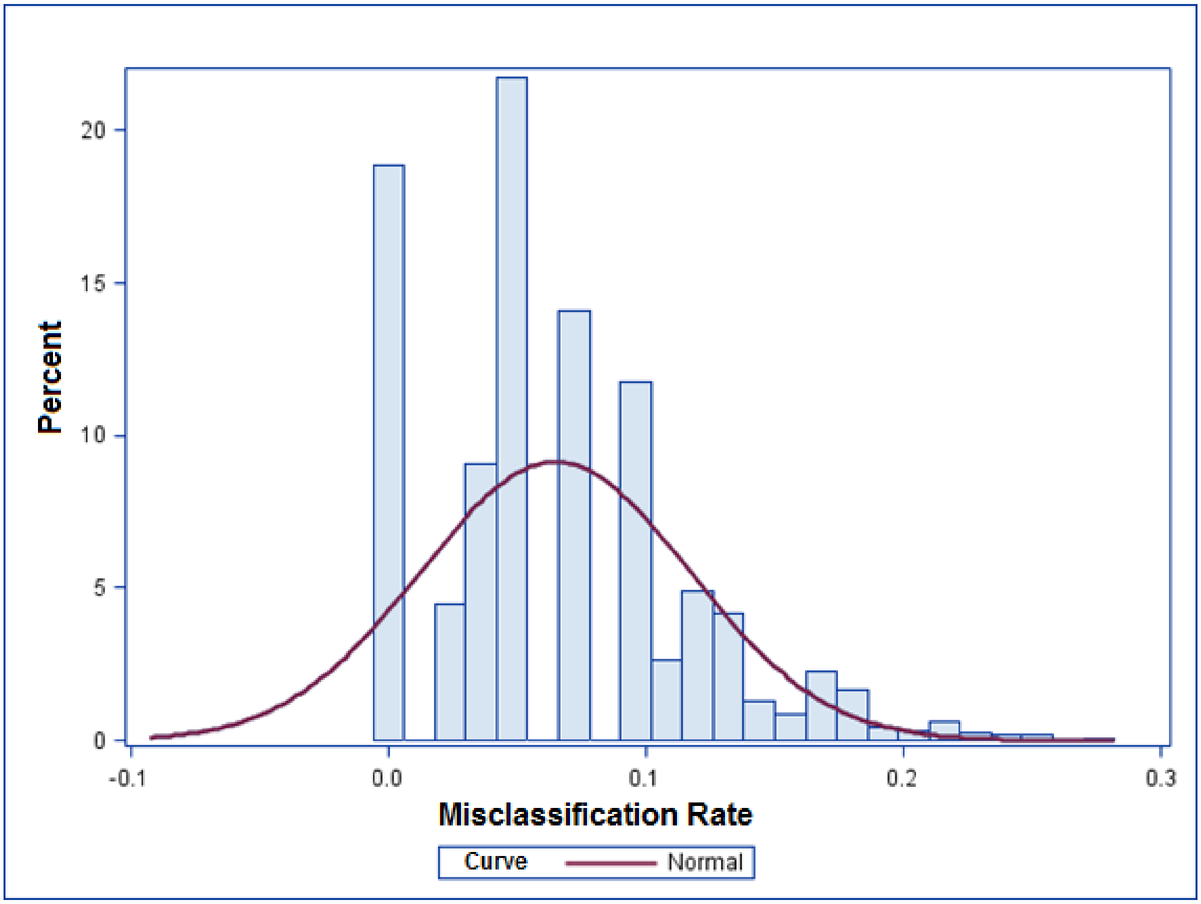

Figure 12. The distribution of species misclassification rate for the external bootstrap is described using a normal approximation. The mean species misclassification rate is 0.0646 or 6.46% and the standard deviation is 0.0278 or 2.7%.

The mean misclassification was 0.0646 and the median misclassification was 0.0455. The amount of skewness was 0.02 which is low in value. The misclassification rate between LC (*Lucanus capreolus*) and LM (*Lucanus mazama*) decreased to 14%, which might imply that the misclassification rate is dependent on sample size. The standard deviation is 0.027 which is low in value, so the data are centered near the mean. Overall, given that the specimens were not from a laboratory colony, geographically controlled or collected during the same year, the small misclassification rate of the proportional prior discriminant analysis provides an effective way to correctly classify these Coleoptera species.

Concluding Remarks

Replicated samples of eleven species of wood primarily boring beetles were selected from William Barr Entomology Museum at the University of Idaho for potential differentiation of their taxonomic group and gender based upon spectral reflectance readings. The methodology used for correctly identifying Coleoptera species typically relies on morphology of the individual species. In this study, however, spectroscopy on elytra composition of the insects was utilized for the purpose of separation of their species and gender. Specifically, the analyses focused on the visual and near-infrared spectrum to differentiate species and gender. Spectrometer readings generated for each species-gender group were fitted to normal distribution mixture models to identify multiple peak reflectance wavelengths of prominence for further statistical analyses. Principal component and discriminant analyses were subsequently used to assess the differentiation of taxonomic groups and genders based on spectral reflectance. The principal component ordination technique clearly grouped Coleoptera by taxonomic groups, while the linear discriminant analysis, under an assumption of multivariate normality, provided a distinct classification of taxonomic groups and provided a low rate of misclassification error. The assumption of normality was subsequently relaxed using a nonparametric nearest neighbor discriminant analysis, which resulted in highly accurate classification of Coleoptera species. Further internal and external validation of the nearest neighbor discriminant model confirmed the results of low species misclassification error rates.

Given the low error rates of misclassification, the multivariate statistical approaches outlined in this study are recommended for analysis of spectral reflectance in Coleoptera and other similar insect groups. However, it is noted that further research in this area should consider using a larger number of individual insects, as well as increasing the number of species analyzed. Also, extrapolation of results has to be practiced cautiously due to varying sensitivity of spectroscopy equipment. If practically feasible, utilizing insects from multiple museums is highly recommended. Incorporation of other Coleoptera attributes such as developmental stage, length, pheromones present, location and collection date might further improve the resolution of the classification techniques. Finally, it is recommended to obtain additional spectrometer readings in the ultraviolet spectrum because insects are sensitive to that spectrum and it may contain markings that are invisible to the human eye.

References

Choate, Paul. "Introduction to the Identification of Beetles (Coleoptera)." *Dichotomous Keys to Some Families of Florida Coleoptera.* (1999): 23-33.

Daly, Howell, and S.S. Balling. "Identification of Africanized Honeybees in the Western Hemisphere by Discriminant Analysis." *Journal of the Kansas Entomological Society*. 51.4 (1978): 857-869.

Dowell, Floyd, J.E. Throne, D. Wang, and J.E. Baker. "Identifying Stored-Grain Insects Using Near-Infrared Spectroscopy." *Journal of Economic Entomology*. 92.1 (1999): 165-169.

Efron, Bradley. "The 1977 Rietz Lecture Bootstrap Methods: Another Look at the Jackknife." *Annals of Satistics*. 7.1 (1979): 1-26.

Hinich, Melvin. "Dichotomous Variable Regression Coefficients and Discriminant Weights." *Political Methodology*. 6.1 (1979): 5-9.

Kawakami, Yasuko, K. Yamazaki, and K. Ohashi. "Geographical variations of elytral color polymorphism in Cheilomenes sexmaculata (Fabricius) (Coleoptera: Coccinellidae)." *Entomological Science*. 16. (2013): 235-242.

Lachenbruch, Peter, and M. Goldstein. "Discriminant Analysis." *Biometrics*. 35.1 (1979): 69-85.

Liu, Shaofang. J. Sun, L. Yu, C. Zhang, J. Bi, F. Zhu, M. Qu, C. Jiang, Q. Yang. "Extraction and Characterization of Chitin from the Beetle Holotrichia parallela Motschulsky." *Molecules*. 17. (2012): 4604-4611.

Piszter, Gabor, K. Kertesz, Vertesy, Z. Zofia, Laszlo, Z. Balint, L.P. Biro. "Color based discrimination of chitin–air nanocomposites in butterfly scales and their role in conspecific recognition." *Hungarian Natural History Museum*. (2010): 3(1). 78-83.

Rao, Radhakrishna. "The Use and Interpretation of Principal Component Analysis in Applied Research." *Indian Journal of Statistics*. 26.4 (1964): 329-358.

Royle, Andrew, and W.A. Link. "A General Class of Multinomial Mixture Models for Anuran Calling Survey Data." *Ecology*. 86.9 (2005): 2505-2512.

SAS online documentation 9.3 Copyright © 2012 SAS Institute Inc., Cary, NC, USA.

Seago, Ainsley, B. Parrish, J. Vigneron, and T.D. Schultz. "Gold bugs and beyond: a review of iridescence and structural colour mechanisms in beetles (Coleoptera)." *Journal of The Royal Society Interface*. 6. (2009): S165-S184.

Shafii, Bahman, W. J. Price, C. Holderman, C. Gidley, and P. J. Anders. 2010. Modeling fish length distribution using a mixture technique. Applied Statistics in Agriculture, W. Song and G. L. Gadbury (Eds). Kansas State University, Manhattan, Kansas, pp. 2-11.

Stark, Williams, and K.P. Tan. “Ultraviolet light: photosensitivity and other effects on the visual system.” *Photochem*. Photobiol. (1982): 371–380.

Vigneron, Jean Pol, M. Rassart, C. Vandenbem, V. Lousse, O. Deparis, L.P. Biró, László. D. Dedouaire, A. Cornet, P. Defrance. "Spectral filtering of visible light by the cuticle of metallic woodboring beetles and microfabrication of a matching bioinspired material," *Physical Review E*, 73, No. 4 (2006): 1-8.