WILD CAUGHT MOSQUITO SPECIES IDENTIFICATION USING IR SPECTROSCOPY AND CHEMOMETRICS

A thesis presented to the faculty of the Graduate School of Western Carolina University in partial fulfillment of the requirements for the degree of Masters of Science in Chemistry.

By

Harrison O'Neal Edmonds

Advisor: Dr. Scott W. Huffman
Associate Professor of Chemistry
Department of Chemistry & Physics

Committee Members: Dr. Carmen L. Huffman, Chemistry & Physics
Dr. Brian D. Byrd, Health and Human Sciences

April 2021

ACKNOWLEDGEMENTS

Throughout this project, there have been many setbacks, both large and small. I am thankful to many individuals for continuing to push me past these obstacles. I want to thank my parents Tracy and Kelley Edmonds, for encouraging my siblings and me to never quit from a very young age. I want to thank my lovely wife Kristyn for her continuous support throughout this project. I would also like to thank the giants whom shoulders I stand on Dr. Scott Huffman, Dr. Brian Byrd, and Dr. Carmen Huffman have been available every step of the way to answer any questions and provide direction to this project. Also, fellow students Lamyae Sroute, Bradley Guilliams, Connor Larmore, and Mark Rothermund have had a profound impact on this project's work, and I am very thankful for their intellectual input. I would also be remiss if I did not recognize the two people responsible for sparking my love for science. Dr. Alesia Jennings and Dr. Mike Bowman, thank you for being the type of educator I aspire to be one day. Last but certainly not least, much of this work would not have been possible if not for the financial support from the American Mosquito Control Association Research Fund (2018) and the North Carolina Biotechnological Center Biotechnology Innovation Grant (2018-BIG-6511).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BEI	Biodefense and Emerging Infections Research Resources Repository
CDC	Centers for Disease Control and Prevention
DNA	Deoxyribonucleic Acid
DOF	Degrees of Freedom
DP	Data Processing
ED	Euclidean Distance
FT	Fourier Transform
HDF	Hierarchical Data Format
IR	Infrared
JDX	Java Desktop For XWindows
MCD	Minimum Covariance Determinant
$MR4$	Malaria Research and Reference Reagent Resource Center
MSU	Michigan State University
NC	North Carolina
PCA	Principle Component Analysis
PLS	Partial Least Squares
$PLS-DA$	Partial Least Squares Discriminate Analysis
$PLS-R$	Partial Least Squares Regression
SPA	Software Publishers Association
WCU	Western Carolina University

ABSTRACT

WILD CAUGHT MOSQUITO SPECIES IDENTIFICATION USING IR SPECTROSCOPY AND CHEMOMETRICS

Harrison O'Neal Edmonds, M.S.,Chemistry

Western Carolina University (April 2021)

Advisor: Dr. Scott W. Huffman

At its current state, mosquito control is all but reliant on the work of the entomologist. The entomologist and other mosquito control personnel are society's first line of defense against harmful vector-borne diseases that have caused mosquitoes to be named the most deadly animal on earth. An accurate and rapid way of accessing a mosquito population is critical to combat mosquito-borne disease. Current methods of accessing an adult mosquito species rely almost exclusively on microscopic identification by highly trained personnel. This process is both very tedious and labor-intensive. This process is also subject to a series of operator and or laboratory errors. Therefore, there is a need for rapid and nondestructive adult mosquito species identification techniques that can be used on an ecologically, economically, and epidemiologically meaningful scale. Our current research aims to develop biochemical discrimination methods between multiple wild species of mosquitoes using infrared spectroscopy. Infrared spectroscopy is a sensitive, information-rich technique capable of detecting a wide range of molecular signals, ranging from subtle changes in protein secondary structure to transmembrane protein-lipid interactions. The resulting data, when coupled with numerical analysis (chemometric) methods such as principal component analysis, linear discriminate analysis, and partial least squares, may be used to classify mosquito species. Herein, we have applied Fourier transform infrared (FT-IR) microspectroscopy to identify a subset of wild mosquito species, including *Culex quinquefasciatus* and *Aedes triseriatus*), using a chemometric model trained by laboratory-reared mosquitoes of the same species. When

trained using laboratory-reared mosquitoes of varying ages, this method can yield up to 96.2% accuracy when predicting *Culex quinquefasciatus* and *Aedes triseriatus*. This method, which is rapid and easy to use, can decrease labor cost and time associated with species identification. Further development coupled with process automation may provide operationally practical methods for rapid species identification of many mosquitoes and other distinguishable mosquito features.

CHAPTER ONE: INTRODUCTION

Background

Mosquitoes are the most deadly animal on planet Earth with over 750,000 deaths per year caused by mosquito-transmitted diseases.[1] This public health issue is not just in tropical areas; cases of mosquito-borne illness in the United States are rising rapidly and becoming an increasing public health concern.[2,3] Although the West Nile virus is the most prevalent mosquito-borne disease in the United States, the most common in NC is La Crosse encephalitis. La Crosse encephalitis is a predominately pediatric disease occurring primarily in Western counties.[4] Currently, there are no vaccines for humans to protect against most mosquito-transmitted viruses. Thus, mosquito control agencies are our last line of defense against the transmission of mosquito-borne diseases. Entomologists working in mosquito control are responsible for the routine surveillance of mosquito populations in their area. Mosquito surveillance is typically completed by setting up surveillance traps such as the industry-standard CDC light trap, then identifying mosquitoes caught in said traps.[5]

Identification typically consists of microscopic identification of adult female mosquitoes using morphological signals, a very labor-intensive task.[6] Routine surveillance is often so expensive that smaller mosquito control agencies cannot afford it.[7–9] These agencies are also responsible for evaluating the associated public health risk factors from a given population of mosquitoes. After evaluation, the responsibility of the entomologist's and other mosquito control personnel's is to develop a plan to combat the risk associated with populations of specific mosquito species. This defense typically takes the form of population control using insecticides. If not used responsibly, these pesticides can have harmful effects on human health where pesticide exposure has been linked to a variety of cancers.[10] Pesticides also have a significant negative economic impact both inside and outside of the public health industry; current statistics predict that insecticides cause billions of US dollars per year in economic and environmental damage to fields such as public

health ($1.1 billion), pesticide resistance ($1.5 billion), and crop losses caused by pesticides ($1.4 billion).[11,12] Because of this, it is essential to use targeted insecticide application only to sites with populations of mosquitoes that are vector threats; this minimizes the harmful side effects of insecticide application. Careful record keeping and frequent surveillance with rapid turnaround are essential for targeted insecticide applications. However, microscopic species identification is very tedious, labor-intensive, and subject to operator errors. Therefore, to properly marshal public health resources, it is imperative that an accurate, rapid, and affordable way of surveying a given mosquito population's species is developed. The technique of using microspectroscopy to identify a mosquito's species has already been shown to differentiate between multiple species of laboratory-reared mosquito populations by Sroute et al.[7] The method is rapid and easy to use and has the potential to decrease both the cost of labor and time associated with species identification, making it an impressive tool for species surveillance work in mosquito control agencies. Multiple groups have even expanded on this work and used inferred spectroscopy to determine the age of the mosquitoes measured.[13,14] However, until now there has been no study's highlighting the use of laboratory-reared mosquitoes to train a database used to identify wild-caught mosquitoes. This aspect is essential to this research because the method's application in the mosquito control setting would be on the wild mosquito population for species surveillance. Thus, It is crucial to determine whether the method built by Sroute et al. will need to be adapted to study the wild-caught mosquitoes or be directly applied in mosquito control surveillance.

Vectorial Capacity Equation

Vectorial capacity is a model that uses the information provided by routine surveillance to predict the threat level of a given mosquito population from a public health perspective. The Vectorial capacity (V) is the total number of potentially infectious bites resulting from all the mosquitoes biting a single perfectly infectious human on a single day.[15] Meaning the higher Vectorial capacity, the higher the risk of the spread of the disease in question. The Vectorial capacity (V) can be

expressed using the equation:

$$V = \frac{ma^2p^n}{-\ln p} \tag{1}$$

where m is representative of the ratio of female mosquitoes to humans, a represents daily blood-feeding rate, p is a single mosquito's odds of survival through one day, and the parasite or virus's incubation period is characterized by n days.[15,16] On average, a human interacting with m mosquitoes per day would be bitten at the rate of ma^2; this value is squared due to two bites being required to transmit disease from infected to an uninfected host. For a mosquito to become infectious, it would have to survive the incubation period with the probability p^n. Then, the infectious mosquitoes live on average $\frac{1}{-ln(p)}$ days biting at a rate of a. The equation was developed initially for malaria but is now used to discuss many other mosquito-transmitted diseases. The Vectorial capacity equation estimates can be optimized with carefully surveillance of a mosquito population's age, since it relates to survival rate (p) and incubation rate (n). Resent work by Guilliams et al.[14] has shown infrared spectroscopy's ability to determine the age of a mosquito using the same technique as species identification. This discovery means that one spectroscopic measurement can provide a variety of information about a mosquito, which can then be used to epitomize a variety of parameters in the Vectorial capacity equation.

Infrared Spectroscopy

Infrared spectroscopy is a measurement tool commonly used to study the interactions of infrared light and molecules, and the resulting spectra act as a molecular fingerprint. The infrared region of the electromagnetic spectrum is from 12500-$10\,\mathrm{cm}^{-1}$, and is typically divided into three sub-regions: near-infrared region (12500-$4000\,\mathrm{cm}^{-1}$), mid-infrared region (4000-$400\,\mathrm{cm}^{-1}$) and far-infrared region (400-$10\,\mathrm{cm}^{-1}$). The mid-infrared region is commonly used to study chemical structures' fundamental vibrations because nearly all molecules have characteristic vibrations in this region.[17] For a molecule to absorb infrared light, the molecule must undergo a net change

in dipole moment in at least one of its vibrational modes. The maximum number of vibrational modes (n) for a given molecule can be calculated using the following equation for nonlinear molecules:

$$n = 3N - 6 \tag{2}$$

where N is the number of atoms in the structure. After a light particle or photon is absorbed for a particular vibrational mode the molecule undergoes a transition from ground vibrational energy level $(v = 0)$ to the first excited state $(v = 1)$. The energy difference between the two energy levels (ΔE) must be equal to the energy of the photon of IR light, which is calculated using:

$$\Delta E = hc\tilde{\nu} \tag{3}$$

where h is Plank's constant, c is the speed of light, and $\tilde{\nu}$ is the wavenumber of the light used to cause the excitation. $\tilde{\nu}$ can be calculated using the equation:

$$\tilde{\nu} = \frac{1}{2\pi c}\sqrt{\frac{k}{\mu}} \tag{4}$$

where k is the force or spring constant of the vibrational mode denoting the strength of the bonds involved in the vibration and μ is the reduced mass of the atoms involved in the vibration. This equation shows that vibrational modes involving heavy atoms with weaker chemical bonds tend to absorb lower-energy infrared light. By comparison, lighter atoms' vibrational modes with stronger chemical bonds tend to absorb higher energy infrared light. This ability to differentiate between vibrational modes using the exact wavenumber of infrared light causes infrared spectroscopy to be a robust tool to analyze a chemical's molecular structure.[18]

Beer's Law

The total amount of inferred light absorbed by the samples vibrational modes can be measured using absorbance (A), which can be expressed with the equation known as the Beer-Lambert-Bouguer Law (Beer's Law),[19]

$$A = \epsilon b C \tag{5}$$

where A is the absorbance at a specific wavelength, ϵ is the molar absorptivity for a particular substance, b is the pathlength, and *C* is the concentration of the substance. Beer's Law is additive, so when dealing with a mixture, Beer's Law can be written as:

$$A = \epsilon_1 b_1 C_1 + \epsilon_2 b_2 C_2 + \epsilon_n b_n C_n ... \tag{6}$$

where n is the number of components in the mixture. This assumption is valid unless the compounds chemically interfere with each other.

Infrared Microspectroscopy

Mosquitoes are complicated organisms with chemically distinct body parts and heterogeneous morphologic features. For this reason, a specific form of infrared spectroscopy was chosen: Fourier transform infrared (FT-IR) microspectroscopy. Using a microscope allows measurement of a particular area of a mosquito and allows for efficient user adjustments of the sample to assure the measurement's quality.

Data Processing

Since almost every compound has at least one IR vibrational mode, infrared spectra of mixtures can be very complicated and challenging to interpret, biological or otherwise. FT-IR spectroscopy is often paired with chemometric data processing (DP) methods and statistical analysis, which allows a computer to perform much of the data interpretation. This combination has been used

for many applications, including everything from street drug detection[20] to characterization of petroleum-based products.[21]

The data processing of any classification technique contains two main steps: training and validation. Before any classification technique begins, the user must randomly split the dataset into two portions, where one portion will be used in the model's training and the other in the model's validation. The validation set can then be used to evaluate the model's prediction accuracy.

Also, before classification, it is typical that spectra will undergo various pre-processing steps to decrease the impact of non-correlative information represented in the data. Typically these steps help reduce the impact of spectral variances caused by the instrument, environment, user, or sample geometry.

Cropping

Atmospheric water vapor and carbon dioxide can cause unwanted spectral variation because of changes in the environment where the spectra are measured, such as humidity. Therefore, it is typical that areas of the infrared spectrum that show a majority of the water vapor/carbon dioxide absorbance are removed before analysis.[17]

Normalization

Normalization's primary purpose is accounting for variations in sample thickness since a larger sample would effectively have a longer path length producing larger absorbance values. So it is essential to normalize the spectra, so the sample size is not a contributing factor to classification.[17,22]

Savitzky-Golay Smoothing and Second Derivative Function

When working with biological samples, it is common to have baseline spectral issues (e.g., sloping or oscillating baselines) due to the scattering of light away from the sample. Often a first or second derivative is used to de-emphasize this baseline fluctuation. This method has also shown the ability to resolve overlapping spectra bands.[17] In general, a Savitzky-Golay smoothing/ differentiation algorithm attempts to amplify important spectral information by limiting the impact of

spectral noise.[17,23]

Outlier Rejection

The final pre-processing step is to perform outlier rejection on the data to remove spectra flawed due to user error. Outlier rejection can take many different forms in various data analysis methods but often involves a distance measurement technique[24] to analyze each spectrum compared to the mean or reference spectrum. This experiment's outlier rejection uses the Mahalanobis distance to compare each spectrum to the others within its group (wild, laboratory-reared, *Aedes triseriatus*, etc.). The Mahalanobis distance specifically has been shown to be a valuable tool for spectroscopic outlier rejection[25,26] and provides a quick and efficient way to find flawed infrared spectra.

Data Analysis Tools

Euclidean Distance

Euclidean distance (ED) is a popular data analysis tool used to calculate the difference between two vectors (spectra).[27] In this study, this distance measurement technique is used to compare the mean of the data to each spectrum, which provides a metric that can be used to visualize spectral deviation from the mean. The equation used to calculate the Euclidean distance for spectra analysis is,[28]

$$ED_n = \sqrt{(R - S_n)^2} \tag{7}$$

where R is the reference or mean spectrum, S_n is the nth spectra in the dataset, and ED_n is the Euclidean distance value for the nth spectrum. Euclidean distance can be adapted and used as a tool for cluster analysis where, rather than finding the mean of all the data, the mean of a single group is compared to individual spectra. This process can be applied as a rapid way to identify significant spectral differences between two groups. An example (using artificially constructed data) of clear separation between two spectral groups using Euclidean distance is shown in Figure 1. In this example, a ED value of 0 means that the data did spectra did not match the reference,

while a $ED = 6$ means the spectra matched the reference very well. In Figure 1, the Euclidean distances of group 1 (red) are different from that of group 2 (blue). In Figure 1 the x-axis is the Euclidean distance value calculated from Equation 7, and the y-axis is the "Spectrum ID" or index of the nth spectra. When comparing plots of this nature, it is important to understand that the y-axis (Spectrum ID) separation is meaningless and only added to the plot for visualization purposes. The size of each group analyzed with ED will also be displayed in the figure caption of the score plots with *n* corresponding to the number of spectra in each group of spectra. Group 1 spectra all have $ED < 3$ while group 2 spectra all have $ED > 4$. This suggests the groups are distinct.

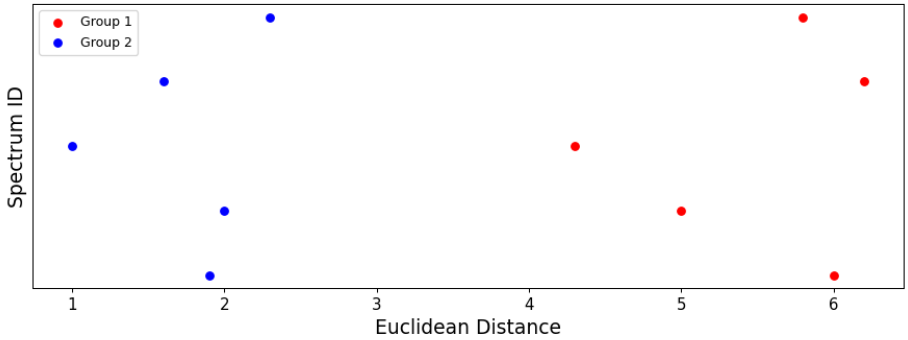


Figure 1. Artificially constructed plot showing clear euclidean distance separation between group 1 (red, $n = 5$) and group 2 (blue, $n = 5$).

Student T-Test

The ED and PCA analysis results can be analyzed group-wise with a student t-test to determine the statistical significance of Euclidean distance differences. The student t-test will provide a value either rejecting or failing to reject the null hypothesis. A null-hypothesis hypothesizes that there is no significant difference between two populations of data, and any observed difference is simply due to random error. The spectra in each group are denoted by the indices a and b is in the

equation below calculating the student's t-test value (t_{exp}):

$$t_{exp} = \frac{|a_{mean} - b_{mean}|}{\sqrt{\frac{a_{std}^2}{a_{size}} + \frac{b_{std}^2}{b_{size}}}} \tag{8}$$

where the indices mean, std and size refer to the average, standard deviation and number of samples/spectra in each group, respectively. The t_{exp} value is then compared to the corresponding t_{table} value based on the degrees of freedom (DOF). If $t_{exp} > t_{table}$ the null-hypothesis is rejected, and if $t_{exp} < t_{table}$ the null-hypothesis cannot be rejected. For example for the Euclidean distance representation shown in Figure 1, t_{exp} was found to be 4.966 which is larger than the t_{table} value of 2.776 ($DOF = 4$) meaning the null-hypothesis is rejected with 95% ($\alpha = .05$) confidence level.[29,30] Therefore confirming that the groups referenced in Figure 1 are significantly different.

Principle Component Analysis

Principle component analysis is a data analysis technique that uses machine learning to analyze a selection of spectra (A) and a given amount (nlv) of loading vectors (L) that are representative of significant spectral differences. Each loading vector is orthogonal to the next, meaning that ideally, there is little to no overlap of spectral information. These loading vectors can then be used to find score matrix values (S), quantifying a spectrum's similarity to corresponding loading vectors, and are often plotted on a Cartesian coordinate system (x, y, z, etc.). A schematic of this process is shown in Figure 2.[31]

The score matrix values scores are plotted on a Cartesian plane where each axis (x,y,z, etc.) is a group of score values. These types of graphs are referred to as score plots. An example of one such score plot is shown below in Figure 3 where the blue points represent the spectra of group 1 and the red points represent the spectra of group 2. The size of each group of spectra analyzed with PCA will be included in the figure caption of the score plots with *n* corresponding to the

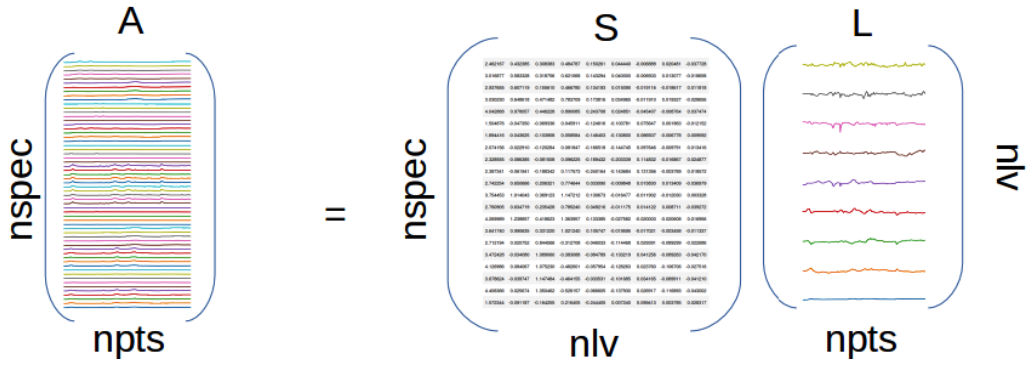


Figure 2. Schematic Representation of PCA Analysis.

number of spectra in each group. Figure 2 displays a clear example of what two distinct groups should look like using PCA analysis with the entirety of group 1 (blue) separated from the entirety of group 2 (red).

Partial Least Squares - Regression

Partial least squares regression (PLS-R) is used to compare information in two data matrices using a linear multivariate model. PLS-R is a robust tool that excels when dealing with noisy, collinear, and even incomplete datasets,[32,33] making it a valuable tool for complex statistical analysis. PLS-R can also be adapted to where C becomes categorical, and this is commonly referred to as Partial Least Squares Discriminant Analysis (PLS-DA). PLS-DA is especially useful for pattern recognition since the training data set allows the model to recognize distinct differences between each group assigned categorically. PLS-DA may be best understood with distinct steps. First, a regression model (B) is built using the training dataset ($A_{training}$) and the group assignment (C). This model (B) is designed to maximize the statistical difference between the categorically assigned groups. The equation to express this relationship is,

$$C = A_{training}B \tag{9}$$

where B, much like with PCA, is a matrix of loading vectors containing combinations of spectral

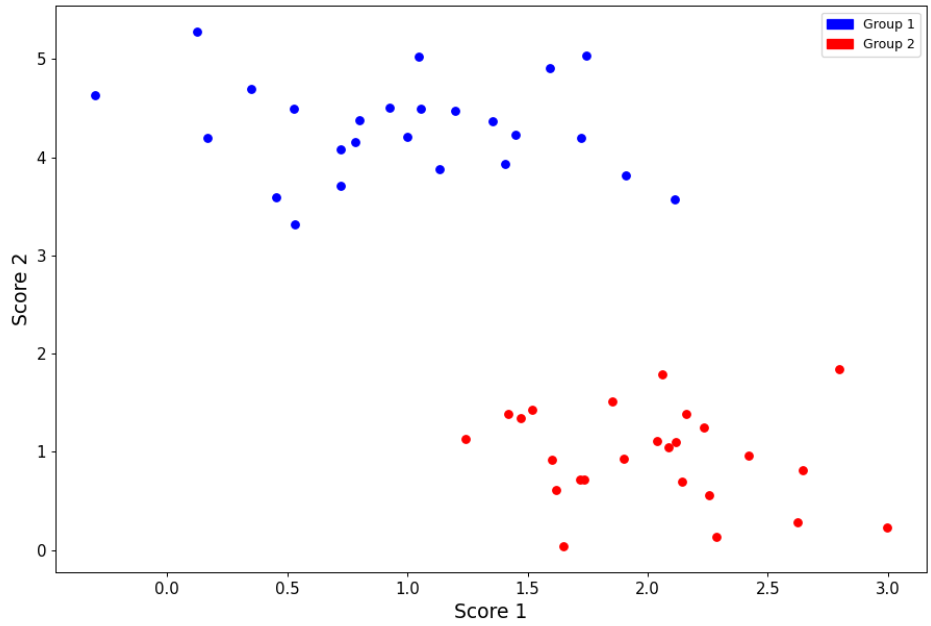


Figure 3. PCA score plot showing separation of hypothetical group 1 (blue, $n = 25$) and group 2 (red, $n = 25$).

features distinct to the assigned groups in the training data set (A). The matrix C is the group assignment of the training dataset, where 1 is assigned to samples in-group and -1 to samples out of the group. The second step of PLS-DA is to predict the group membership of the validation set with the trained model by applying the trained model to the validation dataset ($A_{validation}$); this produces matrix discrimination scores (P).

$$P = A_{validation}B \tag{10}$$

These discrimination scores (P) can then be used for sample classification. In this process, a threshold is typically chosen to optimize the in-group/out-group prediction performance. In PLS-DA, it may be necessary also to take careful note of the size of each group of spectra used in the analysis to optimize the performance further. Because of this, each group's size will be noted within the figure captions of every discrimination score plot with *N* corresponding to the in-group and *M* corresponding to the out-group. An artificially manufactured discrimination score plot is shown in Figure 4. Figure 4 displays blue points as part of artificially constructed group 1 (blue) and red points as part of artificially constructed group 2 (red). Similarly to the Euclidean distance plots, it is important to understand that the y-axis (Sample Number) separation for PLS-DA discrimination score plots is meaningless and only added to the plot for visualization purposes. In this example figure, all six samples were identified correctly and assigned to their correct groups according to the chosen threshold (0). This is evident in Figure 4 because all group 1 (blue) points have a value > 0 and all group group 2 (red) points have a value < 0.

Once a PLS-DA classification model was constructed, the accuracy (*A*) of the model was calculated using the equation:

$$A = \frac{T_P + T_N}{Total} \tag{11}$$

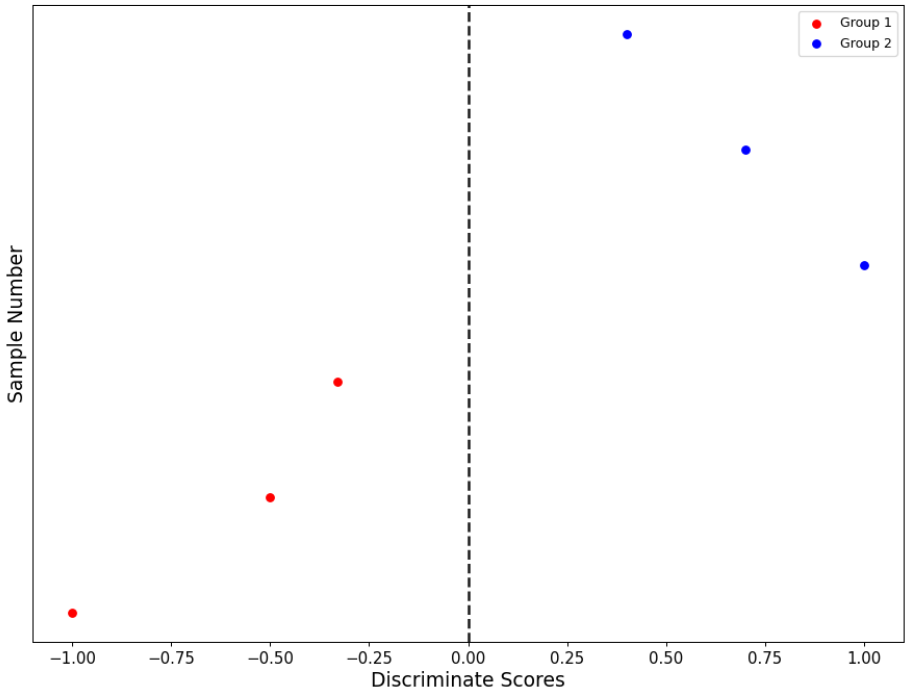


Figure 4. Example PLS-DA discrimination score plot showing separation of hypothetical group 1 (blue, $M = 3$) and group 2 (red, $N = 3$).

where T_P is the amount of true positive's or the number of spectra which were correctly predicted as part of the in-group (*N*) with a value greater than the threshold, T_N is amount of true negative's or the number spectra correctly predicted the out-group (*M*) with a value less than the threshold and $Total$ is the total number of samples in the study.

Research Objectives

This research aims to establish if laboratory-reared mosquitoes can be used in an FT-IR spectral library to train a chemometric model to predict a wild-caught mosquito species correctly. This problem is significantly more complex than previous species prediction studies like the one completed by Sroute et al.[7] The increased complexity of the wild mosquito spectra by a variety of factors including; (I) contaminants that the wild mosquitoes can pick up from its environment, (II) the diet of the wild mosquitoes when compared to that of the laboratory-reared, (III) the broader differences in the gene pool of wild-caught populations compared to that of a laboratory-reared colony, (IV) as well as the wide age variations within species that are represented with a random collection of wild samples. We hypothesize that IR spectra of laboratory-reared mosquitoes can be used to train a chemometric model to predict the species of wild-caught mosquitoes. This would be the optimal scenario because laboratory-reared colony's are typically readily available, and can be reliably aged and identified. Additionally we predict that, classification performance may suffer because of the spectral variation represented in wild mosquitoes' spectra but not spectra of their laboratory-reared counterparts. Suppose spectral differences are identified between wild and laboratory-reared mosquitoes. In that case, this project's secondary goal is to determine the cause of these differences and quantify their impact on species determination. This knowledge will then help to guide the optimization of the species identification method developed by Sroute et al.[7] for use in the mosquito control setting for wild species surveillance.

CHAPTER TWO: EXPERIMENTAL

Materials

Mosquito Samples

Approximately 100 female, *Aedes triseriatus* mosquitoes and 280 *Culex quinquefasciatus* were collected and stored in a $-80°C$ freezer. Wild *Aedes triseriatus* (63 mosquitoes, Age: Unknown) were collected using CDC light traps in areas near Western Carolina University. *Aedes triseriatus* (36 mosquitoes, strain: MSU, Age: 12-15 days) were pulled from a locally housed colony initiated by Michael Kaufman at Michigan State University(MSU) in 2018. This MSU strain has been maintained as a continuous lab-controlled colony housed on Western Carolina University's campus in the Vector-borne Infectious Disease Laboratory. *Culex quinquefasciatus* ($\approx$280 mosquitoes, strain: JBH, Age: (see Table 1)) were obtained from MR4/BEI resources. The colony was then established field collected samples collected at a pond north of Johannesburg, South Africa (Coordinates 26 66'S 27 50'E). The colony was contributed to MR4/BEI by A.J. Cornel.

A selection of laboratory-reared *Aedes triseriatus* (210 mosquitoes, strain: MSU, Age: (see Table 1)) of various ages was also pulled from its established colony. Rearing information for these mosquitoes was carefully monitored and thoroughly documented elsewhere,[14] but briefly, the mosquitoes were reared in cohorts according to age killed by freezing. The exact amount mosquitoes in each age group are listed in table 1. A selection of laboratory-reared *Culex quinquefasciatus* ($\approx$280 mosquitoes, strain: JBH, Age: (see Table 1)) were also collected and divided into two groups: young mosquitoes $<$ 1 week old (105), and old mosquitoes $\geq$ 2-weeks old (156) based on their respective initial holding times.

Instrumentation

All samples were measured using the FT-IR Microscope with the instrument parameters shown below in Table 2. All spectra were acquired at room temperature (20-23 C). A background spec-

Table 1. Mosquito Sample Ages.

Parameter	**Value**	**Uncertainty in age**
Culex quinquefasciatus	$<$ 1 week (n = 105) $\geq$ 2 weeks (n = 156)	$\Delta t \geq 7$ days
Aedes triseriatus	1 day (20-24 hrs) (n = 30) 2 days ($>$ 24 hrs, $<$ 30 hrs) (n = 30) 7 days (n = 30) 14 days (n = 30) 21 days (n = 30) 28 days (n = 30) 35 days (n = 30)	Δt = $\pm$ 1 day

trum was acquired every 10-15 samples to reduce the effects of constantly changing atmospheric and other instrumental conditions. Backgrounds were acquired somewhat subjectively based on the humidity of the environment and spectral signs of distortion like the presence of water vapor between $2000\,\mathrm{cm}^{-1}$ and $1800\,\mathrm{cm}^{-1}$. Each sample was measured, and the files were saved in both JDX[34] and SPA file formats. The SPA file format is used by the OMNIC^{TM} software and allows spectra to be viewed at the instrument. Simultaneously, all the JDX files are used to compile an HDF file[35] containing all spectral and meta-data, allowing the data to be efficiently processed using a personal computer. HDF file format also has the added benefit of being optimized to transfer large data sets efficiently.

Methods

Measurement Procedure

All measurements were completed using the ThermoNicolet^{TM} model Centaurus infrared microspectrometer. The hind leg of the mosquito (closest to the abdomen) was first removed, and the tibia (middle segment of the leg) was positioned to be measured using the infrared microscope. The location of the measurement on the mosquito is shown in Figure 5. The infrared microspectrometer also featured a camera that allowed easier focusing of both the IR and visible

Table 2. FT-IR Microscope Parameters.

Parameter	**Value**
Microscope Make & Model	ThermoNicolet^{TM} Centaurus
Bench Make & Model	ThermoNicolet^{TM} IS10
Software	OMNIC^{TM} version 9.8.372
Wavelength range	650-4000 cm^{-1}
Near/mid/far IR	Mid
Detector	MCT/A, Liquid Nitrogen Cooled
Beamsplitter	KBr
Blank	Air
Scans	64
Resolution	4 cm^{-1}

light onto the sample's surface, leading to better reproducibility. Thus, the camera was focused on each leg before analysis. This process also allowed the user to prevent possible spectral variation sources before they happened, such as a leg bent at an angle on the stage; this would deflect a large percent of the IR radiation away from the detector, causing spectral distortion. With the visible camera, this type of sample orientation issue is apparent and correctable.

Data Pre-Processing

The analysis of the resulting spectra was performed in-house using software written in the Python programming language. Before fitting the data, an optimized set of pre-processing procedures were performed on the spectra. Each spectrum was cropped to the range 1800 - 650 cm^{-1} to eliminate fitting the information poor region between the Amid I and C-H stretching regions. Also, each spectrum was normalized by setting the Amid I band height to an absorbance of one and the baseline at 1800 cm^{-1} to zero. Then lastly, A second derivative Savitzky-Golay algorithm using a window size of 25cm^{-1} and a second-degree polynomial was performed on each spectrum.

Outlier rejection

Principle Component Analysis (PCA) scores are used along with Mahalanobis distance in a Minimum Covariance Determinant (MCD)[36] to find the outlier spectra of each group (wild, laboratory-

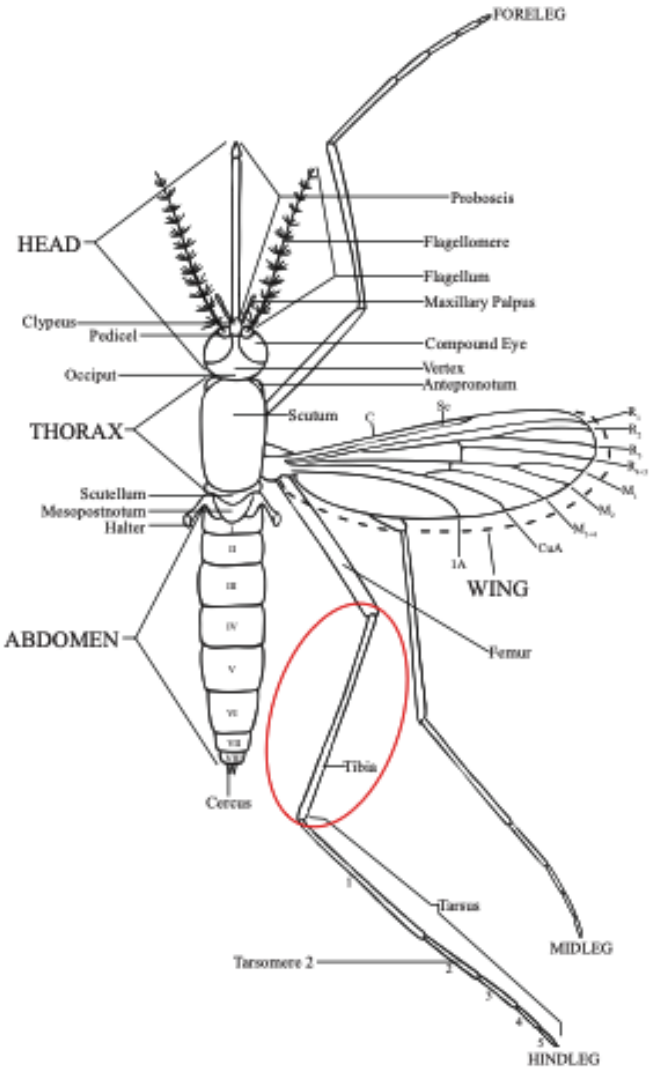


Figure 5. Mosquito anatomy with location of infrared measurement shown by a red outline.

reared, *Aedes triseriatus*, etc.) and remove them.

Data Analysis methods

Principle Component Analysis (PCA)

PCA was performed to compare the spectra of laboratory-reared mosquitoes to that of the wild mosquitoes within the same species. Four loading vectors and their corresponding score plots were analyzed, and the score plot with the most significant separation is presented in the Results and Discussion chapter.

Partial Least Squares - Discriminate Analysis (PLS-DA)

PLS-DA was also performed to compare the spectra of laboratory-reared mosquitoes to that of the wild mosquitoes within the same species, as well as for species classification of a selection of *Aedes triseriatus* and *Culex quinquefasciatus*. The analysis used 4 loading vectors to represent the data. When the data was split for the species analysis, the laboratory-reared mosquitoes were used in the training dataset, and the wild mosquitoes were used in the validation dataset. This separation was done to determine whether a model trained using laboratory-reared mosquitoes could correctly identify wild mosquito species. The data had approximately equal representation of *Aedes triseriatus* (*N*) and *Culex quinquefasciatus* (*M*).

CHAPTER THREE: RESULTS AND DISCUSSION

Aedes triseriatus Wild vs. Lab

Analyzing a Mosquito's IR Spectra

A mosquito is a complex heterogeneous mixture that gives a unique IR spectrum of proteins, nucleic acids, lipids, carbohydrates, and other smaller molecules to contribute to broad superpositioned bands. A mean *Aedes triseriatus* spectra is shown in Figure 6; this figure highlights many of the key regions associated with biological molecules.

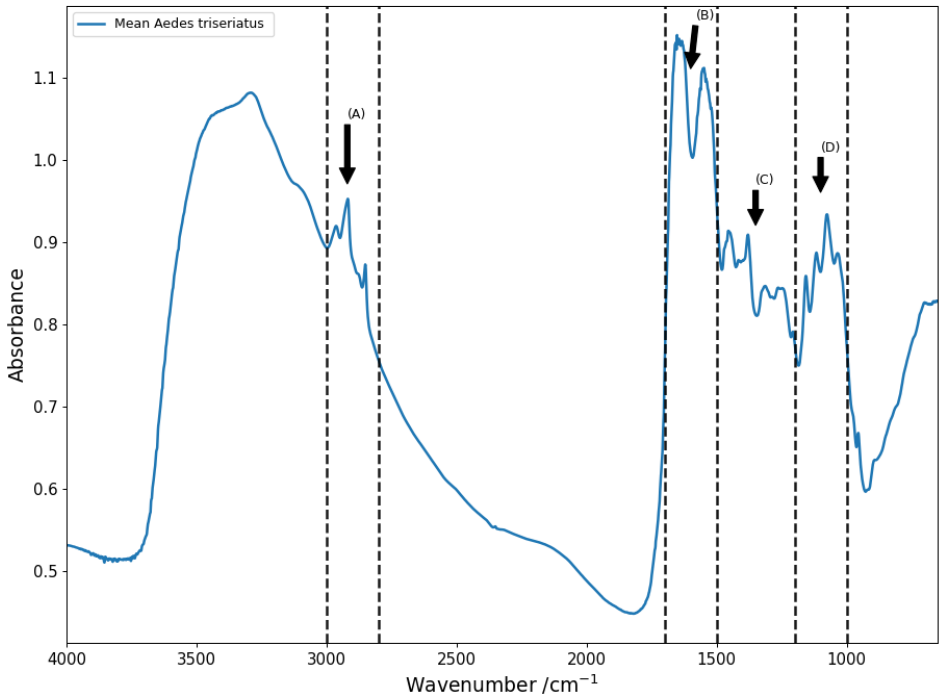


Figure 6. Mean IR spectrum of *Aedes triseriatus* mosquitoes, highlighting important peaks for species identification, **A**: CH_2 and CH_3 stretching (unsaturated lipids), **B**: carbonyl stretching and N-H deformation (Amide I and II), **C**: hydrocarbon bending (Deoxyribonucleic acid (DNA), lipids, protein, etc.), **D**: C-O-C stretching (chitin)

When visually inspecting these mosquitoes' infrared spectra, it can be challenging to identify any apparent differences. However, after pre-processing, slight differences in carbohydrate bands or Amid I and II bands have proven to be substantial enough to determine a variety of information about the mosquito, such as their species[7] or age.[14] Figure 7 shows mean spectra of all the laboratory-reared *Aedes triseriatus* (orange) to the mean spectra of the wild samples (blue) of the same species. Careful examination of Figure 7 reveals a decrease in absorbance for the laboratory-reared spectra through many of the key regions noted in Figure 6. Most notably the carbohydrate-based chitin bands from 1190 - $1000\,\mathrm{cm}^{-1}$ experiences a large decrease in absorbance for the laboratory-reared mean spectra. This absorbance difference could signal that the wild mosquitoes' average age is older than the average age of the laboratory-reared mosquitoes, causing the legs to be slightly larger or thicker on average and thus higher absorbance values.

Pre-Processing of Data

After the spectra were all measured, a selection of outliers was removed using the outlier rejection tool described in the corresponding methods section. Then the remaining spectra (n $\approx$ 100) were all cropped to the region $1800 - 650\,\mathrm{cm}^{-1}$. This pre-processing step results are shown with all the *Aedes triseriatus* spectra in Figure 8. Figure 8 also highlights the need for the normalization step with Amid I band absorbance values ranging from approximately 0.45 all the way to 1.65.

The spectra were then normalized following the procedure described in the Methods/Data Pre-Processing section. The results of this normalization process are shown in Figure 9. This pre-processing step causes the spectra to look much more uniform in nature when compared to the spectra in Figure 8. Figure 9 also displays the wide variation in absorbance values of the chitin bands from 1190 - $1000\,\mathrm{cm}^{-1}$ that may be caused by the variation in age of the wild samples.

The spectra were then processed using a second derivative Savitzky-Golay smoothing algorithm described in the Methods/Data Pre-Processing section. The results of this process are shown in Figure 10. Second derivative spectra are difficult to interpret visually, but the spectra do

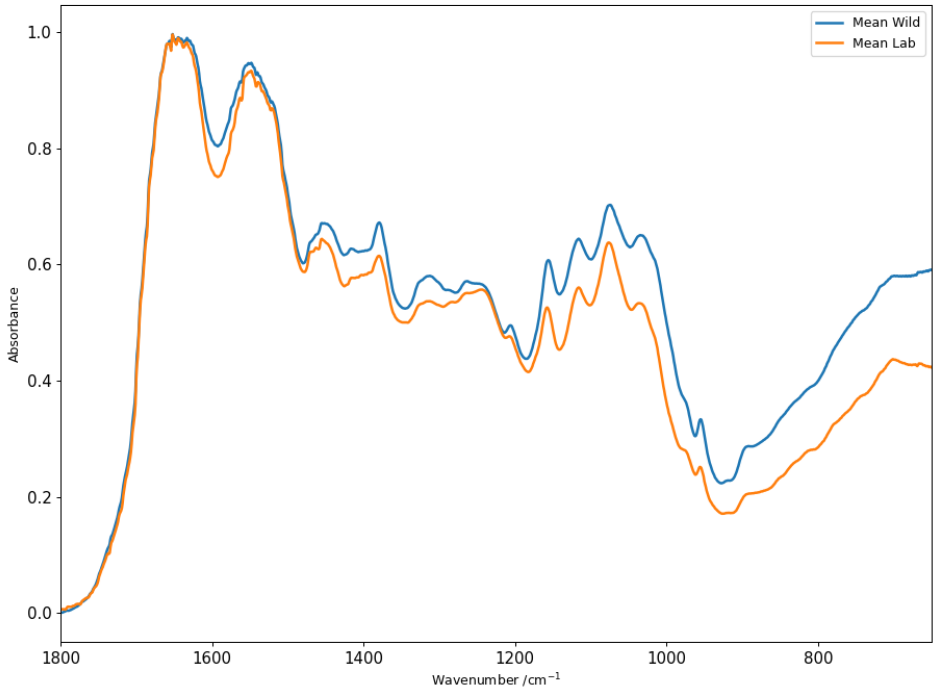


Figure 7. Mean laboratory-reared *Aedes triseriatus* spectra (orange) vs. mean wild *Aedes triseriatus* spectra (blue).

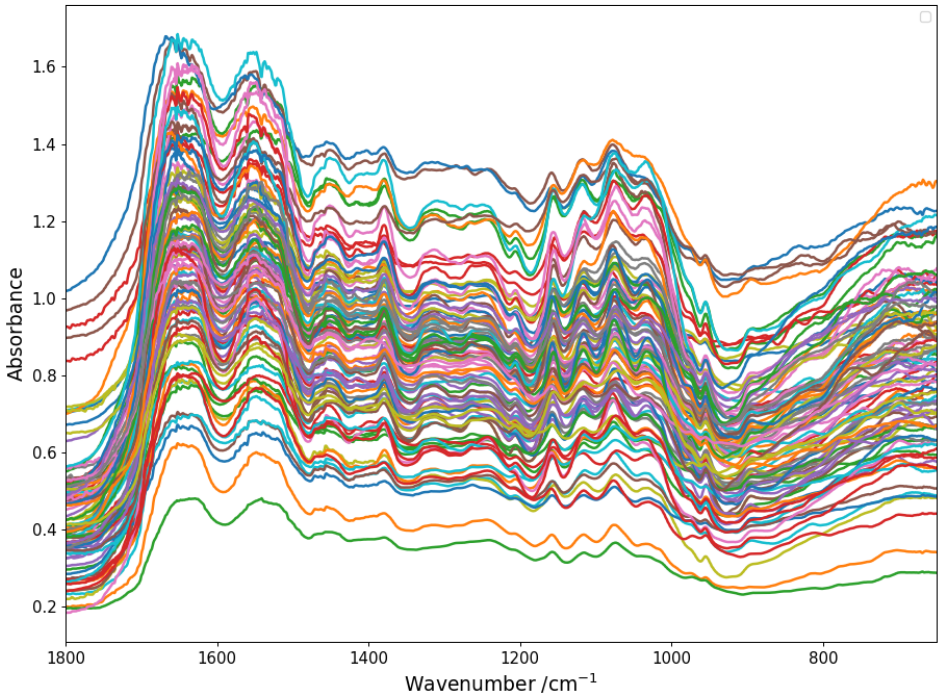


Figure 8. All cropped *Aedes triseriatus* FT-IR Spectra.

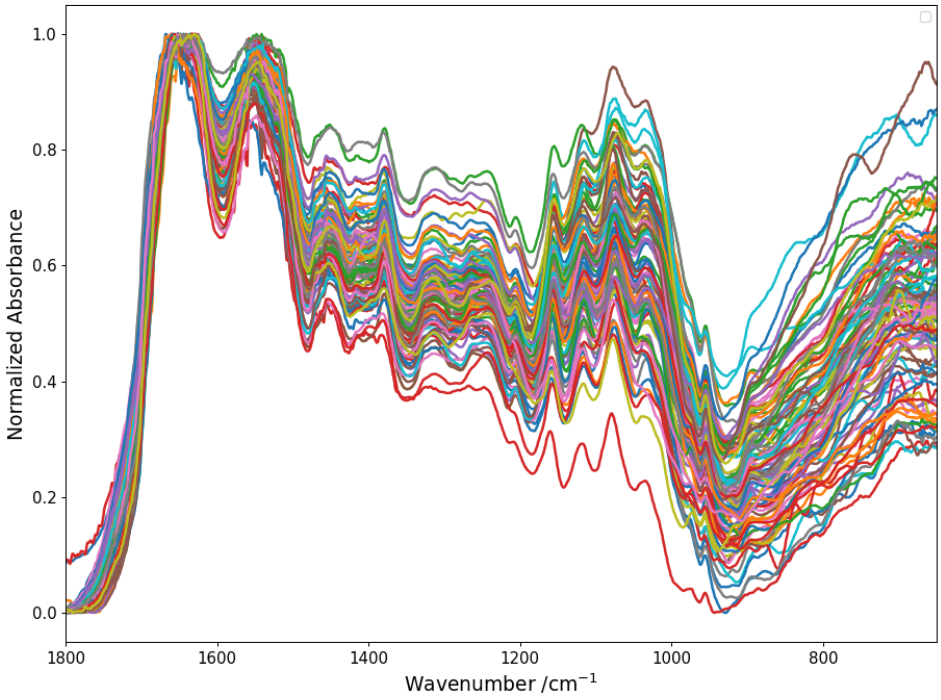


Figure 9. All normalized *Aedes triseriatus* FT-IR Spectra.

experience a lot of variation 900 - $650\,\mathrm{cm}^{-1}$, meaning these areas could be essential for spectral classification.

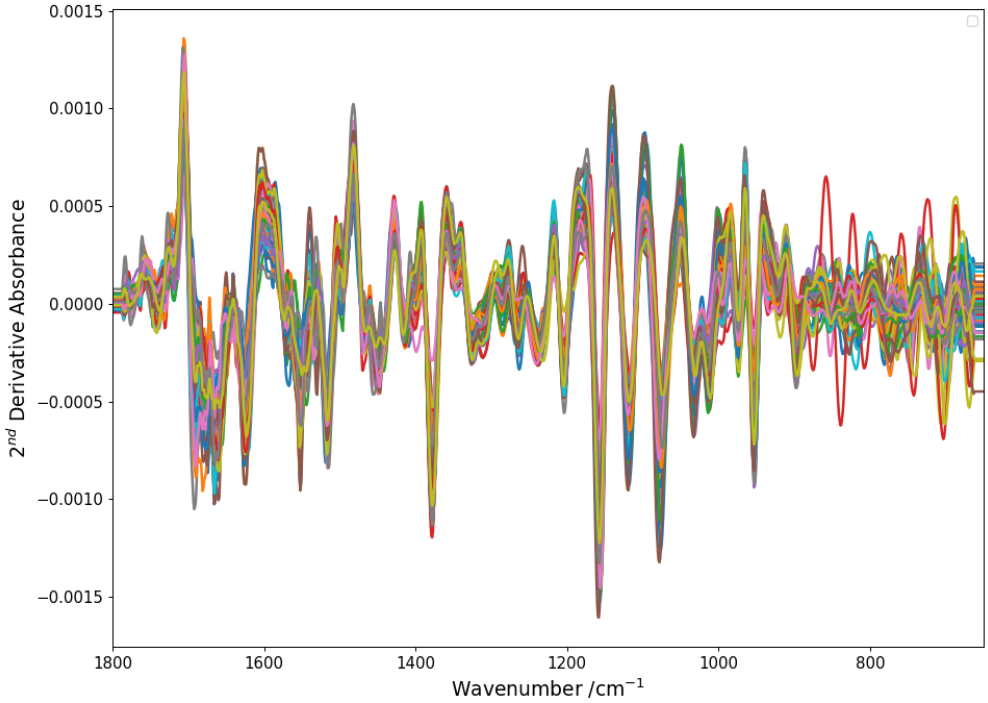


Figure 10. All second derivative Savitzky-Golay smoothed FT-IR Spectra.

Hypothesis

The first step of this study was to identify whether wild *Aedes triseriatus* spectra could be treated the same as laboratory-reared *Aedes triseriatus* spectra and identified using the same technique outlined by Sroute et al.[7] This approach would only be possible if the chosen classification technique was unable to separate wild from laboratory-reared *Aedes triseriatus* meaning both groups of spectra are effectively the same. If the selected classification technique was able to separate wild and laboratory-reared *Aedes triseriatus* spectra, further analysis would be required to de-

termine the source of the differentiation. Then steps should be taken to minimize the impact of whatever is found to be the source of the spectral separation between wild and laboratory-reared samples within the same species. Alternatively, if the source is undetermined, a calibration transfer could be performed to uniformly make the wild samples' spectra match that of the laboratory-reared samples within the same species.

Euclidean Distance

Euclidean distance was the first and least complex classification method used to identify any spectral separation between wild and laboratory-reared *Aedes triseriatus* spectra. The Euclidean distance of each pre-processed spectra was calculated according to the corresponding methods section and plotted in Figure 11. Figure 11 shows a clear separation in the x-direction between the wild and laboratory-reared *Aedes triseriatus* suggesting the laboratory-reared *Aedes triseriatus* spectra are more similar to each other than their wild counterparts.

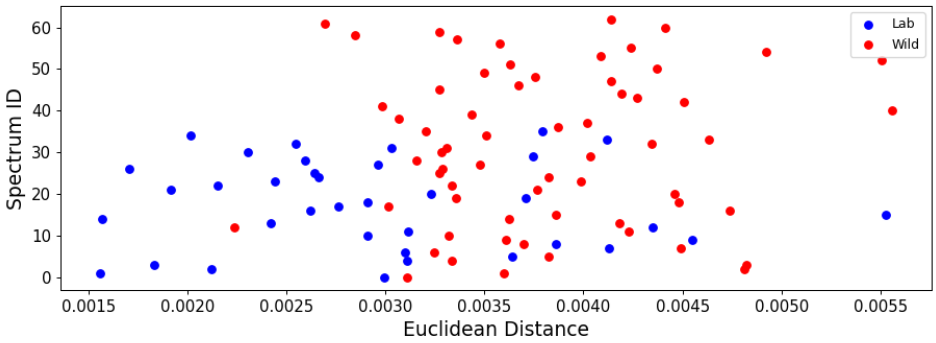


Figure 11. Euclidean distance Scatter plot for wild (red, $n = 63$) and lab reared *Aedes triseriatus* spectra (blue, $n = 36$).

A t-test was performed to test the null-hypothesis using these Euclidean distances. In this example, t_{exp} was found to be 4.966 which is larger than the t_{table} value of 2.042 ($DOF = 35$) meaning the null-hypothesis is rejected with a 95% ($\alpha = .05$) confidence level.[29,30] In the figure,

the Euclidean distances of laboratory-reared *Aedes triseriatus* (age 12-15 days) spectra are blue and wild-caught *Aedes triseriatus* spectra are red. The results of the t-test suggest that these two groups of spectra are significantly different. Thus the spectra should not be used as if they were the same for further species determination, even though both groups of spectra are indeed *Aedes triseriatus*. The results presented in Figure 11 and the subsequent t-test suggest that the wild samples have a spectral variable unrepresented in the laboratory-reared dataset (or vice versa) causing this clear separation between groups.

Principle Component Analysis (PCA)

Principal components analysis was then tested as one of the classification methods of choice in the field of chemometrics.[14] PCA was performed on the pre-processed spectra, and resulting loading vectors 1 and 2 are shown in Figure 12. These loading vectors give the user some insight into what bands cause the separation between groups. In Figure 12 is shown that most of the separation in score values will be caused by the variations in the spectral features from approximately $1700\text{- }920\,\mathrm{cm}^{-1}$. The scores for each spectra were calculated based on each loading vector, and Score 1 was plotted versus score 2, resulting in Figure 13. The blue points represent spectra of laboratory-reared *Aedes triseriatus*, and the red points represent spectra of wild *Aedes triseriatus*. Scores 1 and 2 were chosen somewhat subjectively for the score plot shown in Figure 13 because they resulted in the largest visual separation between groups of wild and laboratory-reared mosquitoes.

Figure 13 shows a clear separation between most of the wild and laboratory-reared groups of spectra. The PCA score plot in Figure 13 echoes the same results as the Euclidean distance plot shown in Figure 11 suggesting that the wild samples have a spectral variable that is unrepresented with the laboratory-reared dataset causing clear separation. A small cluster of approximately eight spectra of wild *Aedes triseriatus* with a score 1 value < 0 that were not as clearly separated into their wild group (red). This could be because these samples were the only wild samples collected that matched the age profile of the laboratory-reared *Aedes triseriatus* in this

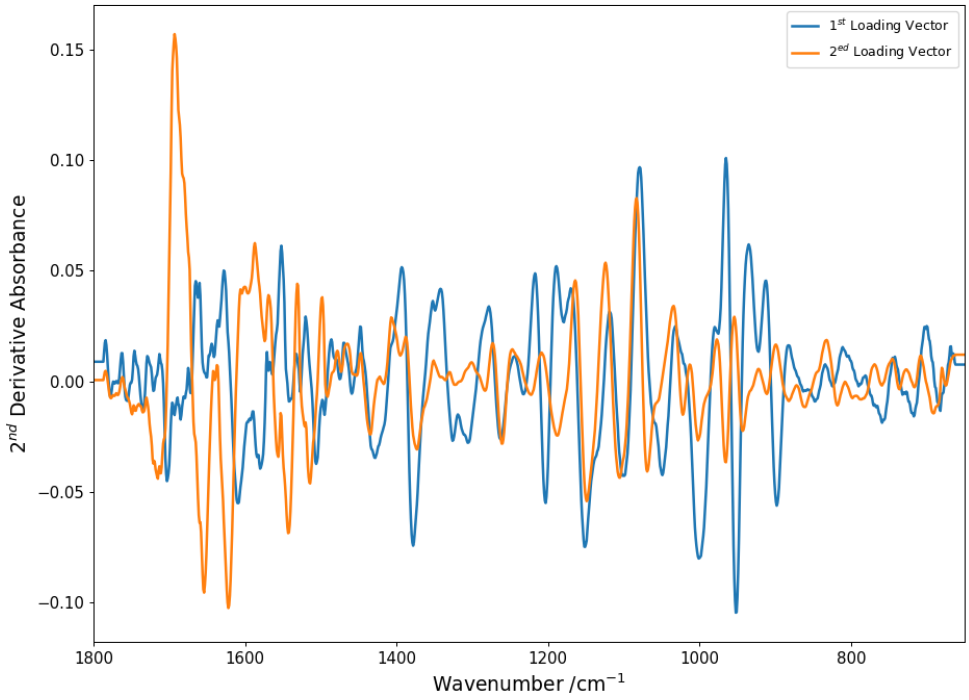


Figure 12. Loading vectors 1 and 2, used to calculate scores in the PCA score plot.

study (12-15 days). This assumption would also infer that one of the leading causes of the spectra separation between the wild and laboratory-reared samples is the spectral differences in sample age.

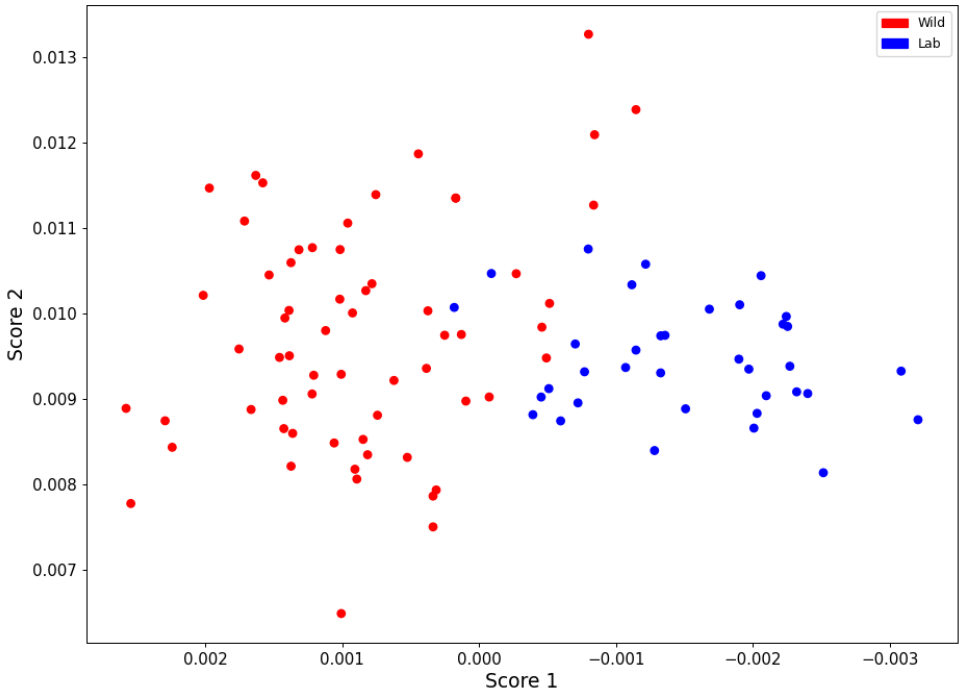


Figure 13. PCA score plot of *Aedes triseriatus* with laboratory-reared mosquitoes (blue, $n = 36$) and wild caught mosquitoes (red, $n = 63$).

Partial Least Squares - Discriminate Analysis (PLS-DA)

Since PLS-DA was the chosen classification method for species determination by Sroute et al.,[7] it was also performed on the wild and laboratory-reared *Aedes triseriatus*. PLS-DA was completed using the pre-processed spectra, and resulting discrimination scores were calculated based on loading vectors $1 - 4$. The discrimination scores are shown in Figure 14 where the red points

represent the wild mosquitoes' spectra, and the blue represents laboratory-reared mosquitoes' spectra. The classification accuracy resulting from the discrimination scores shown in Figure 14 was calculated to be 94.3%. Confirming the results proposed by the ED and PCA classification methods, suggesting wild *Aedes triseriatus* spectra have a variable that is inconsistent with the laboratory-reared dataset. This variable or variables causes clear separation, even when using a small amount of loading vectors to represent the data.

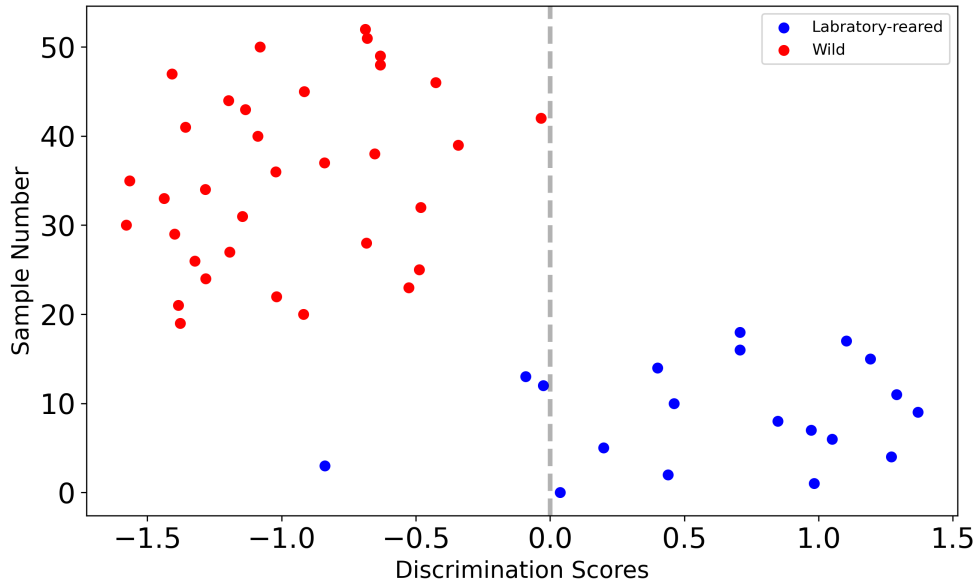


Figure 14. PLS-DA Discrimination score plot comparing wild (red, $M = 34$) to laboratory-reared (blue, $N = 19$) *Aedes triseriatus*.

Aedes triseriatus Wild vs. Lab with varying ages

Hypothesis

Since all three classification methods showed a clear separation between wild and laboratory-reared *Aedes triseriatus* spectra, the next goal was to find the cause of the distinction. Diet, spectral contaminants, genetic variations, age variations, and more could all help to distinguish the

wild *Aedes triseriatus* spectra from their laboratory-reared counterparts. Although, one variation source was of particular interest since the lower absorbance values were recognized in the chitin bands for in Figure 7 was the variation in age of wild samples when compared to that of laboratory-reared samples. To identify whether age variation was a contributing factor in the spectral separation between wild and laboratory-reared mosquitoes, a selection of *Aedes triseriatus* of varying ages was added to the training dataset. The goal being to imitate a wild mosquito collection of mosquitoes by evenly varying the amount mosquitoes represented in each age demographic within the laboratory-reared dataset. The specifics of how many mosquitoes were used from each age class can be found in Table 1.

This hypothesis immediately gained life when the mean spectra of *Aedes triseriatus* samples with the new age variation dataset was plotted with the mean wild *Aedes triseriatus* spectra in Figure 15. Figure 15 reveals that varying the age of the laboratory-reared dataset results in a mean spectra that is much more similar to that of the wild *Aedes triseriatus*. However, the mean laboratory-reared *Aedes triseriatus* spectrum without the age variation has noticeably lower absorbance values, especially in the chitin bands from 1190 - $1000\,\mathrm{cm}^{-1}$ when compared to the mean wild *Aedes triseriatus*. This suggests added demographics of mosquitoes represented in the laboratory-reared dataset increases the absorbance values sightly in the chitin region and caused the mean spectra to appear more similar by result. Figure 15 also indicates that the laboratory-reared dataset containing age variation does a better job of representing wild *Aedes triseriatus* spectra. To thoroughly test whether age is a factor contributing to the separation between wild and laboratory-reared spectra, all classification methods tested above were repeated using this augmented dataset.

Euclidean Distance

Each spectrum was pre-processed using the same pre-processing techniques referenced in the methods section. The Euclidean distance was calculated for each spectrum compared to the mean laboratory-reared spectra (including the new samples with varying ages), and the results were

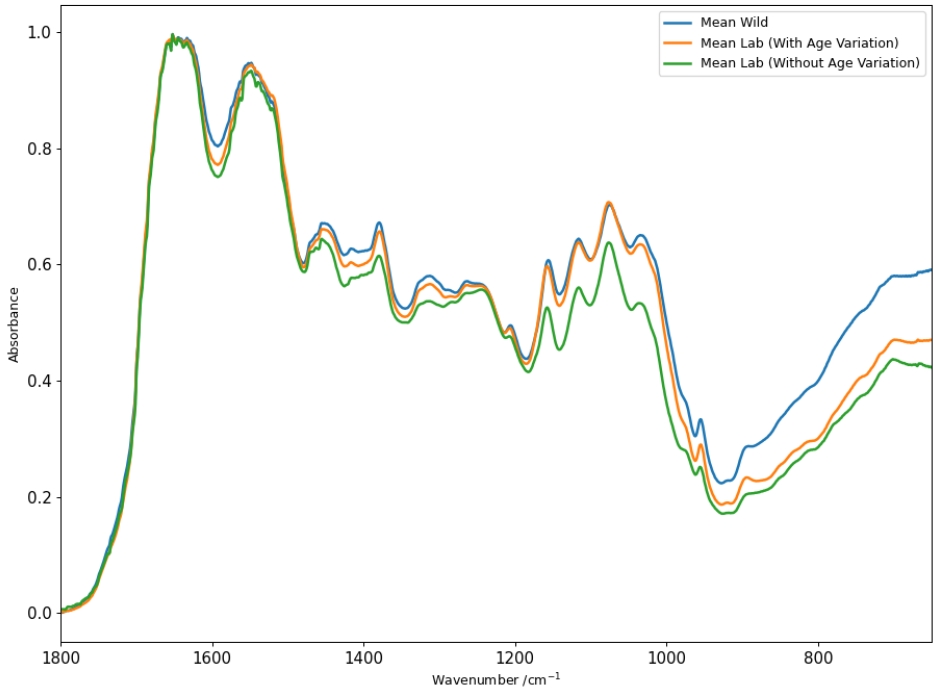


Figure 15. Mean laboratory-reared *Aedes triseriatus* spectra both with (orange) and without (green) varying ages vs. mean wild *Aedes triseriatus* spectra (blue).

plotted in Figure 16. Figure 16 shows much no signs of visual separation in the x-direction between groups of spectra. Therefore suggesting, the laboratory-reared *Aedes triseriatus* spectra are indistinguishable from their wild counterparts. This inference is backed up by the t-test calculated using these Euclidean distance scores. In this example, t_{exp} was found to be 0.07344 which is smaller than the t_{table} value of 1.960 ($DOF = 62$) meaning the null-hypothesis cannot be rejected with a 95% ($\alpha = .05$) confidence level[29,30] suggesting any variation in the data is simply due to varying sources of random error. Overall, the poorer separation shown in Figure 16 suggests that age variation is a large contributor to the separation between wild and laboratory-reared *Aedes triseriatus* spectra.

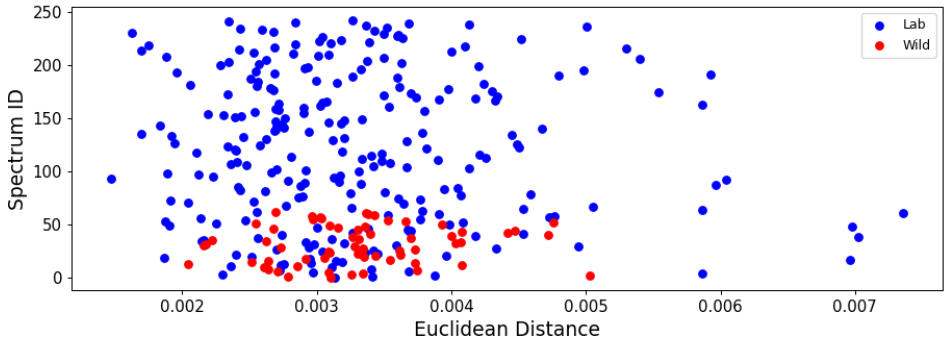


Figure 16. Euclidean distance Scatter plot for wild (red, $n = 63$) and lab reared *Aedes triseriatus* with varying ages spectra (blue, $n = 244$).

The Euclidean distances shown in Figure 16 also seem to vary much more with the laboratory-reared *Aedes triseriatus*. This could be caused by the smaller sample size of wild *Aedes triseriatus* spectra, and as more wild spectra are added to the dataset, the range of Euclidean distances would be expected to increase. Along that same line of thought, the laboratory-reared *Aedes triseriatus* spectra dataset may include age demographics that are not represented at all within the wild *Aedes triseriatus* dataset.

Principle Component Analysis (PCA)

PCA was performed on the pre-processed spectra and the scores for each spectra were calculated based on each loading vector, and score 1 was plotted versus score 2, resulting in Figure 17. In Figure 17 the blue points represent spectra of laboratory-reared *Aedes triseriatus* with age variation, and the red points represent spectra of wild *Aedes triseriatus*. The PCA plot seemed to confirm the Euclidean distance results, suggesting that the difference between the groups was vastly diminished when the laboratory-reared dataset was represented by *Aedes triseriatus* of varying ages. Therefore, age variation may be a significant contributor to the separation between wild and laboratory-reared *Aedes triseriatus* spectra.

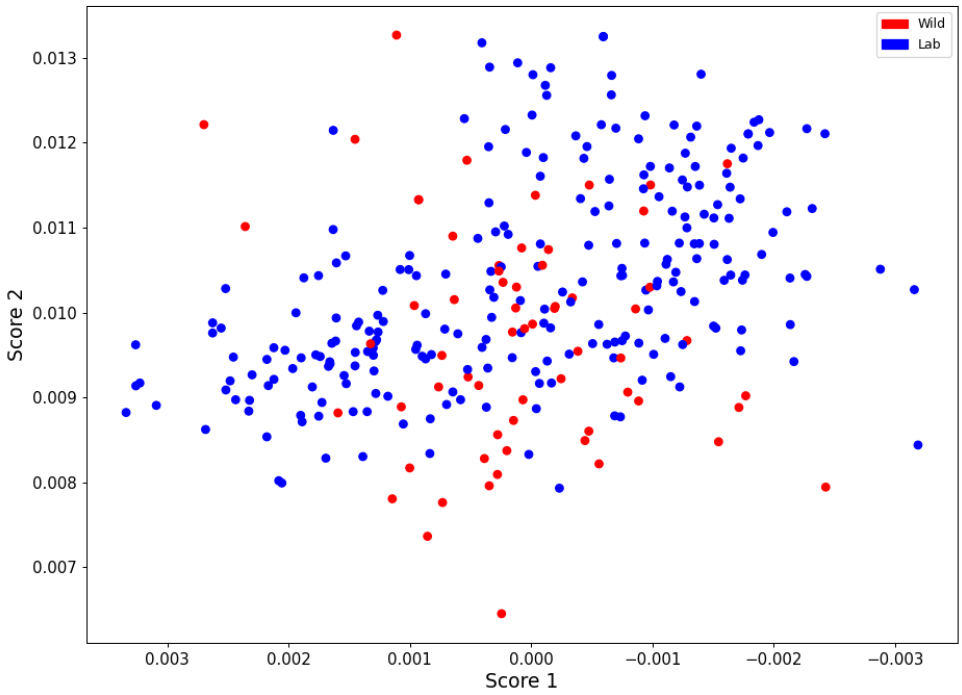


Figure 17. PCA score plot of *Aedes triseriatus* with laboratory-reared mosquitoes of varying age (blue, $n = 244$) and wild caught mosquitoes (red, $n = 63$).

Partial Least Squares - Discriminate Analysis (PLS-DA)

PLS-DA was performed comparing the pre-processed spectra of the laboratory-reared *Aedes triseriatus* with varying ages against the wild *Aedes triseriatus*. The resulting discrimination scores were then calculated based on loading vectors 1-4. The discrimination scores are shown in Figure 18 where the blue points represent the laboratory-reared *Aedes triseriatus* spectra with varying ages represented, and the red points represent wild *Aedes triseriatus* spectra. The classification accuracy resulting from the discrimination scores shown Figure 18 was calculated to be 87.2%. This value means that the PLS-DA models' accuracy experienced a 7.1% decrease, resulting from the addition laboratory-reared *Aedes triseriatus* spectra of varying ages. These results echo the findings of a PCA and ED; following the addition of the *Aedes triseriatus* spectra with varying ages, a decrease is seen in the classification model's ability to resolve wild and laboratory-reared samples. This finding suggests that age variation plays an important role in the spectra of mosquitoes, and needs to be accounted for when preparing the training sets for species discrimination.

Species Separation of Wild Caught *Aedes triseriatus*

Hypothesis

All three classification methods suggested that age variation contributed to the separation between wild and laboratory-reared *Aedes triseriatus* spectra. Since this research's overarching goal in the broadest sense is to build the most efficient/accurate tool for species determination of wild mosquito samples, the next step was to quantify age variation's impact on species determination. PLS-DA was chosen for this test as the primary classification method used for species determination by Sroute et al.[7] PLS-DA was performed using both the dataset containing and not containing the laboratory-reared age varying *Aedes triseriatus*. Both tests were classified against a selection of *Culex quinquefasciatus*, also of various ages, the specifics of the age distributions of the *Culex quinquefasciatus* samples can be found in Table 1.

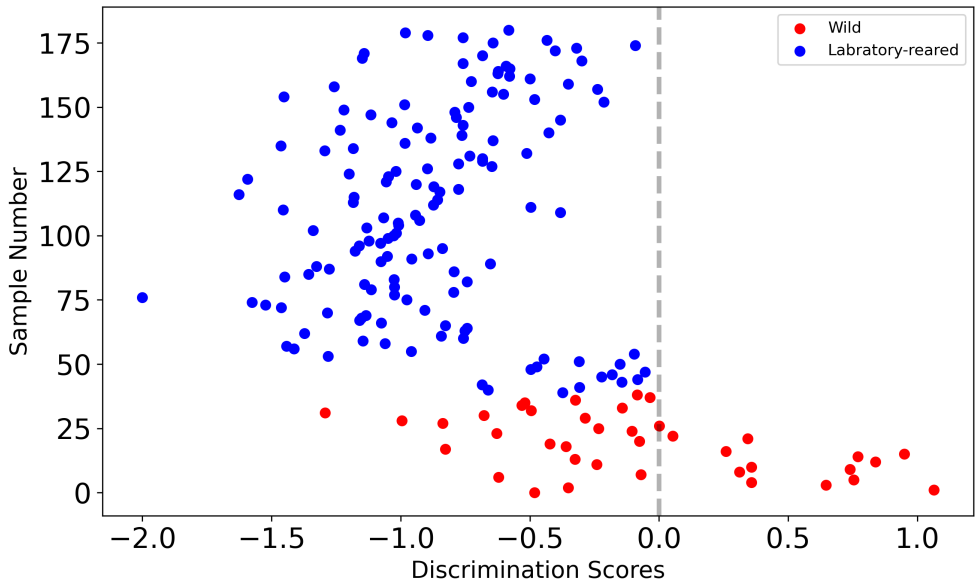
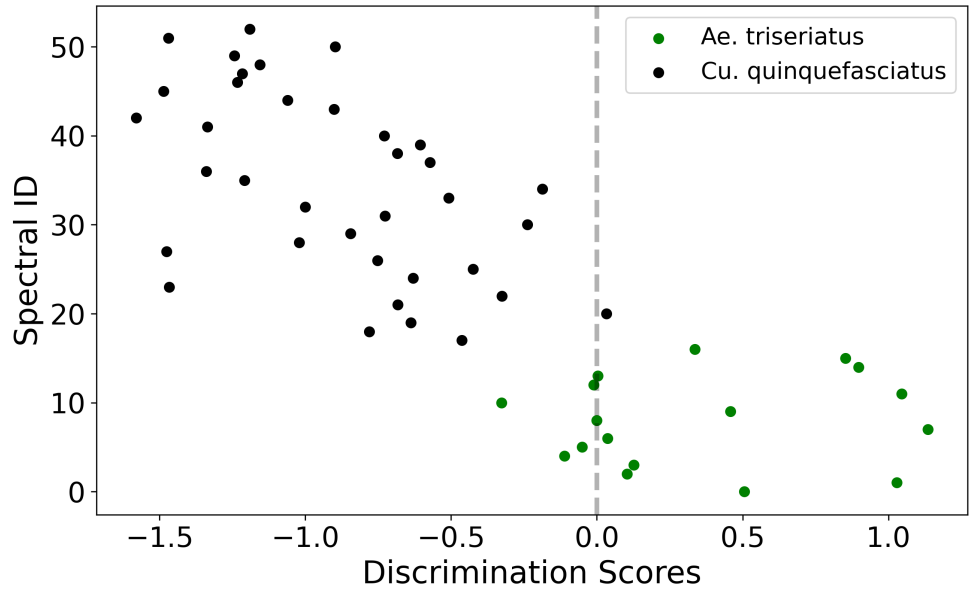


Figure 18. PLS-DA Discrimination score plot comparing wild (blue, $N = 39$) to laboratory-reared (red, $M = 142$) *Aedes triseriatus* (with age variation).

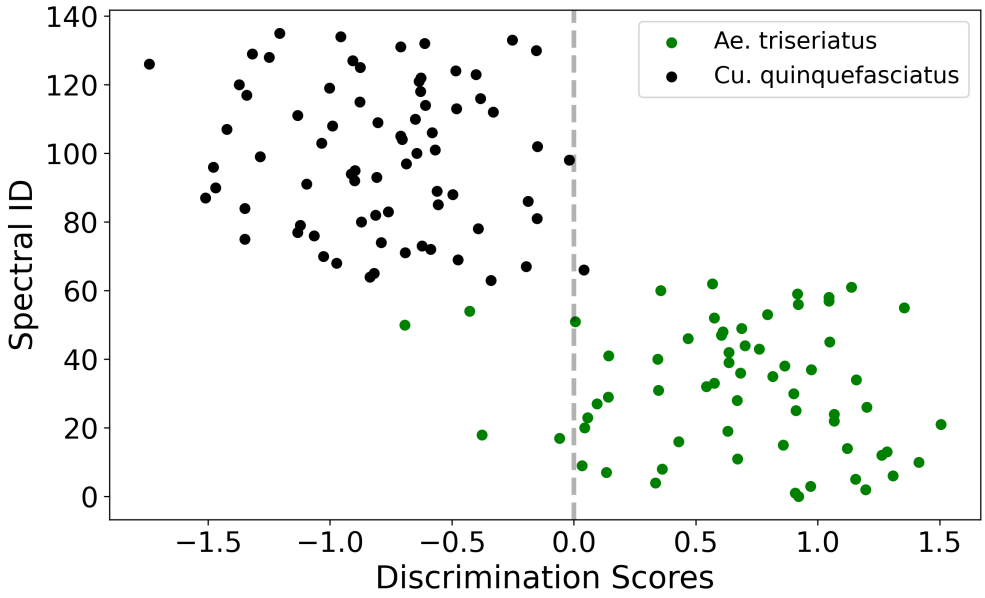
Partial Least Squares - Discriminate Analysis (PLS-DA) Results

PLS-DA was performed using the pre-processed spectra of wild *Aedes triseriatus* as the validation dataset and the laboratory-reared *Aedes triseriatus* as the training dataset. This process is done intentionally to test whether a model built entirely from laboratory-reared samples could be used to predict wild sample species. This would be the optimum scenario for applying the method built by Sroute et al.[7] to the mosquito control setting, requiring no wild mosquito collections to build a functional model. The *Aedes triseriatus* spectra were compared using PLS-DA against *Culex quinquefasciatus* of varying ages that were split in half randomly. The resulting discrimination scores were then calculated based on loading vectors 1-4. The discrimination scores are shown in Figure 19 where the green points represent the laboratory-reared *Culex quinquefasciatus* spectra with varying ages, and the black represent the *Aedes triseriatus* with (b) and without (a) age variation. Figure 19 shows that without (a) age variation species determination,

between *Aedes triseriatus* and *Culex quinquefasciatus*, can be completed with 91.3% accuracy. With (b) the addition of *Aedes triseriatus* samples of varying age an accuracy improvement is seen from 91.3% (a) to 96.2% (b). This accuracy improvement could directly result from the model containing age variations, improved ability to represent the spectra of wild *Aedes triseriatus* mosquitoes.



(a) Without age variation.



(b) With age variation.

Figure 19. PLS-DA discrimination plots showing species separation between *Aedes triseriatus* (black, $N = 20$ (a), $N = 63$ (b)) and *Culex quinquefasciatus* (green, $M = 15$ (a), $M = 71$ (b)), both with (b) and without (a) age variation.

CHAPTER FOUR: CONCLUSIONS AND FUTURE DIRECTIONS

Overall this study showcased that regardless of the classification method used, wild and laboratory-reared *Aedes triseriatus* spectra can be separated. Thus, proving that wild spectra have fundamental variables that impact their infrared spectra, such as diet, spectral contaminants (things the mosquitoes pick up), genetic variation differences, age variations, and more. Further analysis confirmed age variations as a root cause for some of the spectral differences between wild and laboratory-reared *Aedes triseriatus* spectra. This study proceeded to demonstrate the benefits of using a laboratory-reared dataset containing spectra of varying ages to build a chemometric model for species classification. When creating the classification database for wild species determination, the goal being to mimic the wild samples as closely as possible with the same species' laboratory-reared dataset. Ideally, this causes the spectral difference between species to stand out more prominently and promotes better prediction accuracy. When accounting for age variation of wild populations with the laboratory-reared dataset, there is an accuracy improvement from 91.3% (a) to 96.2% (b). This variable is only one of the possible causes of spectral difference between wild and laboratory-reared mosquitoes. The impact of other wild spectra variables on species determination should also be investigated in the future. With full knowledge of these variables, a database can be constructed of laboratory-reared samples that maximizes the classification method's performance.

Future experimentation should be done to (I) validate the performance of the PLS-DA model used for wild mosquito species determination, (II) further investigate the use of age variation in laboratory-reared mosquitoes and its impact on species determination of the wild counterparts, (III) increase the scale; increasing the dataset's size should produce more accurate and repeatable results (IV) expand, more than 3,000 mosquito species around the world to test this method on. Depending on the performance of PLS-DA with a larger dataset and or other species, new classification models should also be developed and optimized. While further testing is still required,

this project's completion has played a significant role in showing how far this new technology can go. The ability to use FT-IR microspectroscopy to identify wild-caught mosquito species is a valuable tool to mosquito control agencies worldwide that could save countless resources and labor.

REFERENCES

[1] Spielman, A. *Mosquito: The Story of Man's Deadliest Foe*; Hachette Books, 2001.

[2] Kindhauser, M. K.; Allen, T.; Frank, V.; Santhana, R. S.; Dye, C. Zika: The Origin and Spread of a Mosquito-Borne Virus. *Bulletin of the World Health Organization* **2016**, *94*, 675, DOI: https://doi.org/10.2471/BLT.16.171082.

[3] Gratz, N. G. Emerging and Resurging Vector-Borne Diseases. *Annual Review of Entomology* **1999**, *44*, 51–75, DOI: https://doi.org/10.1146/annurev.ento.44.1.51.

[4] Kappus, K. D.; Calisher, C. H.; Baron, R. C.; Davenport, J.; Francy, D. B.; Williams, R. M. La Crosse Virus Infection and Disease in Western North Carolina. *The American Journal of Tropical Medicine and Hygiene* **1982**, *31*, 556–560, DOI: https://doi.org/10.4269/ajtmh.1982.31.556.

[5] Sriwichai, P.; Karl, S.; Samung, Y.; Sumruayphol, S.; Kiattibutr, K.; Payakkapol, A.; Mueller, I.; Yan, G.; Cui, L.; Sattabongkot, J. Evaluation of CDC Light Traps for Mosquito Surveillance in a Malaria Endemic Area on the Thai-Myanmar Border. *Parasites & Vectors* **2015**, *8*, 1–10, DOI: https://doi.org/10.1186/s13071-015-1225-3.

[6] Lukacik, G.; Anand, M.; Shusas, E. J.; Howard, J. J.; Oliver, J.; Chen, H.; Backenson, P. B.; Kauffman, E. B.; Bernard, K. A.; Kramer, L. D., et al. West Nile Virus Surveillance in Mosquitoes in New York state, 2000–2004. *Journal of the American Mosquito Control Association* **2006**, *22*, 264–271, DOI: https://doi.org/10.2987/8756-971X(2006)22[264:WNVSIM]2.0.CO;2.

[7] Sroute, L. Mosquito Identification Using Infrared Spectroscopy and Chemometrics. M.Sc. thesis, Western Carolina University, 2018.

[8] McNeil, D. Tick and Mosquito Infections Spreading Rapidly, C.D.C. Finds. *New York Times* **2018**,

[9] Vazquez-Prokopec, G. M.; Chaves, L. F.; Ritchie, S. A.; Davis, J.; Kitron, U. Unforeseen Costs of Cutting Mosquito Surveillance Budgets. *PLoS Negl Trop Dis* **2010**, *4*, e858, DOI: https://doi.org/10.1371/journal.pntd.0000858.

[10] Bassil, K. L.; Vakil, C.; Sanborn, M.; Cole, D. C.; Kaur, J. S.; Kerr, K. J. Cancer Health Effects of Pesticides: Systematic Review. *Canadian Family Physician* **2007**, *53*, 1704–1711.

[11] Pimentel, D.; Burgess, M. *Integrated Pest Management*; Springer, 2014; pp 47–71, DOI: https://doi.org/10.1007/978-94-007-7796-5_2.

[12] Asidi, A.; N'Guessan, R.; Akogbeto, M.; Curtis, C.; Rowland, M. Loss of Household Protection from use of Insecticide-Treated Nets Against Pyrethroid-Resistant Mosquitoes, Benin. *Emerging Infectious Diseases* **2012**, *18*, 1101, DOI: https://doi.org/10.3201/eid1807.120218.

[13] González Jiménez, M.; Babayan, S. A.; Khazaeli, P.; et al., Prediction of Mosquito Species and Population Age Structure using Mid-Infrared Spectroscopy and Supervised Machine Learning [version 3; peer review: 2 approved]. *Wellcome Open Research* **2019**, *4:76*, DOI: http://dx.doi.org/10.12688/wellcomeopenres.15201.3.

[14] Guilliams, B. F. Mosquito Chronological Age Determination Using Mid-Infrared Spectroscopy and Chemometrics. M.Sc. thesis, Western Carolina University, 2020.

[15] Kramer, L. D.; Ciota, A. T. Dissecting Vectorial Capacity for Mosquito-Borne Viruses. *Current Opinion in Virology* **2015**, *15*, 112–118, DOI: https://doi.org/10.1016/j.coviro.2015.10.003.

[16] Brady, O. J.; Godfray, H. C. J.; Tatem, A. J.; Gething, P. W.; Cohen, J. M.; McKenzie, F. E.; Perkins, T. A.; Reiner, R. C.; Tusting, L. S.; Sinka, M. E., et al. Vectorial Capacity and Vector Control: Reconsidering Sensitivity to Parameters for Malaria Elimination. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **2016**, *110*, 107–117.

[17] Baker, M. J.; Trevisan, J.; Bassan, P.; Bhargava, R.; Butler, H. J.; Dorling, K. M.; Fielden, P. R.; Fogarty, S. W.; Fullwood, N. J.; Heys, K. A., et al. Using Fourier Transform IR Spectroscopy to Analyze Biological Materials. *Nature Protocols* **2014**, *9*, 1771, DOI: https://doi.org/10.1038/nprot.2014.110.

[18] Takamura, A.; Watanabe, K.; Akutsu, T.; Ozawa, T. Soft and Robust Identification of Body Fluid Using Fourier Transform Infrared Spectroscopy and Chemometric Strategies for Forensic Analysis. *Scientific reports* **2018**, *8*, 1–10.

[19] Beer, A.; Beer, P. Determination of the Absorption of Red Light in Colored Liquids. *Annalen der Physik und Chemie* **1852**, *86*, 78–88.

[20] Risoluti, R.; Materazzi, S.; Gregori, A.; Ripani, L. Early Detection of Emerging Street Drugs by Near Infrared Spectroscopy and Chemometrics. *Talanta* **2016**, *153*, 407–413, DOI: https://doi.org/10.1016/j.talanta.2016.02.044.

[21] Khanmohammadi, M.; Garmarudi, A. B.; de la Guardia, M. Characterization of Petroleum-Based Products by Infrared Spectroscopy and Chemometrics. *TrAC Trends in Analytical Chemistry* **2012**, *35*, 135–149, DOI: https://doi.org/10.1016/j.trac.2011.12.006.

[22] Zarnowiec, P.; Lechowicz, L.; Czerwonka, G.; Kaca, W. Fourier Transform Infrared Spectroscopy (FTIR) as a Tool for the Identification and Differentiation of Pathogenic Bacteria. *Current Medicinal Chemistry* **2015**, *22*, 1710–1718, DOI: https://doi.org/10.2174/0929867322666150311152800.

[23] Vivó-Truyols, G.; Schoenmakers, P. J. Automatic Delection of Optimal Savitzky- Golay Smoothing. *Analytical Chemistry* **2006**, *78*, 4598–4608, DOI: https://doi.org/10.2174/092986732266615031115280O.

[24] Kriegel, H.-P.; Kröger, P.; Zimek, A. *Outlier Detection Techniques*; Citeseer, 2010; Vol. 10; pp 1–76.

[25] Chen, B.; Zou, X.; Zhu, W. Eliminating Outlier Samples in Near-Infrared Model by Method of PCA-Mahalanobis Distance. *Journal of Jiangsu University (Natural Science Edition)* **2008**, *4*.

[26] Pierna, J. F.; Wahl, F.; De Noord, O.; Massart, D. Methods for Outlier Detection in Prediction. *Chemometrics and Intelligent Laboratory Systems* **2002**, *63*, 27–39, DOI: https://doi.org/10.1016/S0169-7439(02)00034-5.

[27] He, Q. A Review of Clustering Algorithms as Applied in IR. *Graduate School of Library and İnformation Science University of İllinois at Urbana-Champaign* **1999**, *6*, 1–33.

[28] Deborah, H.; Richard, N.; Hardeberg, J. Y. A Comprehensive Evaluation of Spectral Distance Functions and Metrics for Hyperspectral Image Processing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2015**, *8*, 3224–3234, DOI: https://doi.org/10.1109/JSTARS.2015.2403257.

[29] Harvey, D. *Analytical Chemistry 2.0*; McGraw-Hill Companies, 2000.

[30] CliffsNotes, Two Sample t test for Comparing Two Means. 2021; https://www.cliffsnotes.com/study-guides/statistics/univariate-inferential-tests/two-sample-t-test-for-comparing-two-means.

[31] Huffman, S. Chemometrics. 2019; https://agora.cs.wcu.edu/~huffman/chem455.html#org51a5774.

[32] Wold, S.; Sjöström, M.; Eriksson, L. PLS-Regression: A Basic Tool of Chemometrics. *Chemometrics and Intelligent Laboratory Systems* **2001**, *58*, 109–130, DOI: https://doi.org/10.1016/S0169-7439(01)00155-1.

[33] Barker, M.; Rayens, W. Partial Least Squares for Discrimination. *Journal of Chemometrics: A Journal of the Chemometrics Society* **2003**, *17*, 166–173, DOI: https://doi.org/10.1002/cem.785.

[34] McDonald, R. S.; Wilks Jr, P. A. JCAMP-DX: A Standard Form for Exchange of Infrared Spectra in Computer Readable Form. *Applied Spectroscopy* **1988**, *42*, 151–162, DOI: https://doi.org/10.1366/0003702884428734.

[35] Fortner, B. HDF: The hierarchical data format. *Dr Dobb's J Software Tools Prof Program* **1998**, *23*, 42.

[36] Hubert, M.; Debruyne, M. Minimum Covariance Determinant. *Wiley Interdisciplinary Reviews: Computational Statistics* **2010**, *2*, 36–43, DOI: https://doi.org/10.1002/wics.61.