CALIBRATION OF PERMITTIVITY SENSORS TO MEASURE CONTAMINANTS IN
WATER AND IN BIODIESEL FUEL

by

SARAH SHULTZ

A THESIS

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Biological and Agricultural Engineering
College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2009

Approved by:

Major Professor
Naiqian Zhang

Abstract

Four permittivity probes have been developed and tested to measure contaminants in water and in biodiesel fuel. An impedance meter was also used to measure the same contaminants. The pollutants measured in water were nitrate salts (KNO_3, $Ca(NO_3)_2$, and NH_4NO_3) and atrazine. The contaminants measured in biodiesel were water, glycerol, and glyceride. Each sensor measured the gain and phase of a sample with a known concentration of one of these pollutants.

The resulting signals were analyzed using stepwise regression, partial least squares regression, artificial neural network, and wavelet transformation followed by stepwise regression to predict the concentration of the contaminant using changes in the gain and phase data measured by the sensor. The same methods were used to predict the molecular weight of the nitrate salts. The reliability of the probes and the regression methods were compared using the coefficient of determination and the root mean square error. The frequencies selected using stepwise regression were studied to determine if any frequencies were more useful than others in detecting the contaminants.

The results showed that the probes were able to predict the concentration and the molecular weight of nitrates in water very accurately, with R^2-values as high as 1.00 for the training data and 0.999 for the validation data for both concentration predictions and molecular weight predictions. The atrazine measurements were somewhat promising, the training R^2-values were as high as 1.00 in some cases, but there were many low validation values, often below 0.400. The results for the biodiesel tests were also good; the highest training R^2-value was 1.00 and the highest validation R^2-value was 0.966.

Table of Contents

iv

List of Figures

List of Tables

Acknowledgements

I would first like to express my sincere appreciation for my major professor, Dr. Naiqian Zhang, for his help, support, and encouragement as I have worked on my Master's degree. I am very thankful to have learned so much by working with him.

I would like to thank my committee members, Dr. Gary Clark and Dr. Philip Barnes. Dr. Clark also provided valuable guidance as my undergraduate advisor, and Dr. Barnes has given me help on my research, especially with the atrazine tests.

Darrell Oard deserves many thanks for building the probes used in this research and the stands that were used to hold the probes in place. He found a solution for every problem that arose with the probes, and I am very grateful for his help.

I'd like to acknowledge the students in the Instrumentation and Control Lab for their help and friendship: Ning Tang, Yali Zhang, Ling Xue, Peng Li, Wei Han, Huiquan Zhang, Joseph Dvorak, Xu Wang, Alan Baurely, Matthew Worcester, and Daniel Bigham.

Finally, I wish to express my deep gratitude to my family for their love and support. I especially want to thank my parents, Lisa and Michael Shultz, and my fiancé, Patrick Rogers-Ostema, who have always been there when I need them.

CHAPTER 1 - Introduction

Water pollution is becoming an increasingly large problem in the United States. As of 2008, there were 43,868 watersheds, lakes, and wetlands reported to the Environmental Protection Agency (EPA) as impaired; 1,333 of which were in Kansas. Nitrates were the cause of 474 impairments, and atrazine was the cause of 147 of the impairments (US Environmental Protection Agency, 2009).

Nitrate and atrazine have numerous negative effects on water quality. The biggest source of nitrate pollution is fertilizer that runs off of fields during rain and irrigation. Excessive nitrates in water can lead to algae blooms and decreased oxygen. High levels of nitrates in drinking water can cause methemoglobinemia in infants (Self and Waskom, 2008). Atrazine, a common herbicide, is harmful to sensitive aquatic plants, amphibians, and some species of fish. It also causes a variety of health issues in humans, such as cardiovascular problems, congestion of the heart and lungs, and possibly cancer (National Safety Products Incorporated, n.d.).

Impurities are also a concern in biodiesel fuel. As biodiesel becomes more and more popular, it is important to have a method for quickly and accurately measuring the purity of the fuel. Water in biodiesel can cause corrosion of fuel tanks. Glycerol is the biggest by-product of biodiesel production, and it is also one of the most common contaminants. Glycerol in biodiesel creates problems with fuel storage and engine fouling. Glycerides are unreleased glycerol molecules that cause similar problems.

There are many methods used to detect water pollutants and biodiesel impurities. The most common methods involve chromatography, a procedure which is very expensive and must be done in a laboratory. Colorimetric analysis is also used, and it too requires expensive equipment.

Permittivity is a frequency-dependent measure of polarization and conduction in dielectric materials. The permittivity of a material such as water or biodiesel fuel changes when impurities are present. The molecules in the impurities have different polarities and conductive properties

than the molecules in the water or fuel. Therefore, the concentration of an impurity in a sample can be determined using a calibration model for variations in permittivity of the sample material. Permittivity is commonly used to measure properties of other dielectric materials such as moisture content in soils (Scholte et al., 2002).

Calibration can be achieved using regression. Regression describes how closely one set of variables is related to another set. Two sets of variables are needed to build a regression model: response variables and predictor variables. Once the model is created, it can be used to estimate the response variables given a set of predictor variables. Stepwise and partial least squares are linear regression methods. Artificial neural network is a popular nonlinear regression technique. Wavelet transformation is a way of compressing data, and it can be used as a preprocessing tool for regression.

The purpose of this study was to calibrate four different permittivity probes and an impedance meter to measure the concentration of water pollutants (potassium nitrate, calcium nitrate, ammonium nitrate, and atrazine) and biodiesel impurities (water, glycerol, and glyceride). The four probes were developed at the Instrumentation and Control Laboratory in the Biological and Agricultural Engineering Department at Kansas State University. These probes are portable and relatively inexpensive. This study also compared four different regression methods: stepwise regression, partial least squares regression, artificial neural network, and stepwise regression of wavelet transformed data.

CHAPTER 2 - Research Goals

The main objective of this study was to test the ability of four different permittivity probes and an impedance meter to detect pollutants in water and in biodiesel fuel. The permittivity probes were developed in the Instrumentation and Control Laboratory in the Biological and Agricultural Engineering Department at Kansas State University. These probes were designed to be durable, portable, and inexpensive. The probes were composed of parallel stainless steel plates with the even numbered plates electrically connected to one another and the odd numbered plates electrically connected to one another. The size of the plates and the number of plates was different for each probe.

 The specific goals of the research were:

1. To convert the permittivity measurement of each water or fuel sample to a measure of the concentration of contaminant in the sample using regression methods.

2. To distinguish between different nitrate salts (KNO_3, $Ca(NO_3)_2$, and NH_4NO_3) in water by converting the permittivity signals from the sample to a measure of the molecular weight of the nitrate salt in the water using regression methods.

3. To compare stepwise regression, partial least squares regression, artificial neural network, and stepwise regression of wavelet transformed data to determine which method could produce the most reliable prediction models.

4. To compare the four different probes and the impedance meter in terms of their ability to detect contaminants in water and biodiesel.

5. To determine if any frequencies are more significant than others in detecting contaminants in water and in biodiesel.

CHAPTER 3 - Literature Review

3.1 Water Pollution

Water quality is a major concern in the U.S. The agricultural industry is one of the biggest contributors to water pollution. Fertilizers and pesticides are applied to fields in large quantities to promote crop growth. Rainfall and unmanaged irrigation can wash these chemicals off the field, and they end up in streams, lakes, and wells where they harm the natural environment and contaminate drinking water.

3.1.1 Nitrates

Nitrates are a common pollutant in water. They come from a number of sources including fertilizers, feedlots, septic tanks, and municipal wastewater. Potassium nitrate, calcium nitrate, and ammonium nitrate are inorganic nitrates that can be found in fertilizers. Nitrates occur naturally in the environment when microorganisms break down plants and other organic materials. Nitrates are safe in small amounts, but human activities increase nitrate concentrations to unhealthy levels.

3.1.1.1 Negative Impacts

Nitrates can cause health issues in humans. A high intake of nitrates is not likely to cause anything more serious than gastric problems in adults. Infants, on the other hand, are very susceptible to nitrate poisoning. Methemoglobinemia, also called blue baby syndrome, is a disease that can occur when infants ingest water with high nitrate levels. The nitrates cause hemoglobin, an oxygen-carrying protein, to be converted to methemoglobin, which is not a good oxygen carrier. As a result, the child's brain does not receive enough oxygen. This can cause brain damage and, in severe cases, death (Self and Waskom, 2008).

Nitrates are also harmful to the environment, especially aquatic habitats. Nitrogen is a nutrient for plants and algae, and when the nitrate concentration in a body of water is elevated above the

4

normal level, algae grow very rapidly and prevent sunlight from penetrating through the water body. The absence of light increases the activity of some aerobic bacteria and the amount of oxygen in the water is greatly reduced by these bacteria, a process called eutrophication. Many fish and other aquatic organisms cannot survive with these low levels of oxygen.

3.1.1.2 Water Quality Standards

The maximum contaminant level goal (MCLG), which is the maximum amount of contaminant that is believed to be safe for human consumption, and the maximum contaminant level (MCL), which is the maximum amount of contaminant that is legally allowed to be in drinking water, are both 10 mg/L for nitrate. Levels below 10 mg/L are not believed to cause methemoglobinemia (Self and Waskom, 2008).

3.1.1.3 Detection Methods

There are EPA methods for measuring nitrate concentration in water. The first way, EPA Method 300.0, is by ion chromatography. The lowest concentration this method can detect is 0.4 mg/L, or 0.4 ppm (California State Water Resource Board, 2008). The second procedure for detecting nitrate is EPA Method 353.2, colorimetric automated cadmium reduction. This method can detect nitrate concentrations of as low as 0.05 mg/L, or 0.05 ppm (O'Dell, 1993). Both tests have two major problems: they are expensive and they require the use of several hazardous chemicals.

Ion selective electrodes (ISEs) are an inexpensive tool that can measure the concentration of nitrates and other ions in water in only a few minutes. ISEs function by converting the activity of an ion dissolved in solution to an electric potential. This electric potential can then be measured using a voltmeter. ISEs are usually capable of measuring ion concentration within a range of $\pm 3\%$ of the actual concentration value. There are some disadvantages of using ISEs. They contain gel that needs to be replaced periodically, they often need to be recalibrated after every test, and they are not very durable (Rundle, 2000).

3.1.2 Atrazine

Atrazine is a popular herbicide that is used on row crops such as corn and grain sorghum to kill broadleaf and grassy weeds. It is an organochlorine compound, and its molecular formula is $C_8H_{14}ClN_5$. The use of atrazine is widespread in the Midwestern U.S. In 2001, between 3,664,000 and 5,413,000 pounds of atrazine were applied to fields in Kansas (Natural Resources Defense Council, 2004).

Over the last twenty years, the use of atrazine has been challenged and even banned in some European countries because of problems associated with water contamination. In 2005, the European Union banned the use of atrazine. Germany and Italy, countries that produce millions of tons of corn each year, banned atrazine in 1991. Corn yield in these countries did not decline in comparison to the U.S., where atrazine was used on the crops, suggesting that a good yield is still possible without using atrazine. Some consideration has been given to banning atrazine in the U.S., but it has been met with a lot of resistance due to concerns about decrease in crop yield and increase in crop price (Ackerman, 2007).

3.1.2.1 Negative Impacts

Atrazine can be very dangerous to humans. It has been known to cause cardiovascular problems and reproductive problems in humans (Kansas Department of Health and Environment, 2004). Studies also suggest that atrazine can cause congestion of the heart, lungs, and kidneys and damage adrenal glands (National Safety Products Incorporated, n.d.). Research has been done that links repeated atrazine exposure to breast cancer in laboratory rats; however, there is no conclusive evidence that these results apply to humans (Wisconsin Department of Health Services, 2008).

Atrazine is harmful to aquatic habitats and starts causing serious problems at concentrations of 10 to 20 µg/L. Atrazine is moderately toxic to fish and highly toxic to aquatic invertebrates. It can cause disturbances in the reproductive and endocrine systems of aquatic organisms; this is most common in amphibians and largemouth bass. Atrazine is especially toxic to aquatic plants. These plants are an important part of the food chain and they provide cover that allows small fish

to hide from predators. When this vegetation is reduced, the whole community is affected (US Environmental Protection Agency, 2006).

3.1.2.2 Water Quality Standards

In 1992, the EPA set regulations for the amount of atrazine that could be present in drinking water. The maximum contaminant level goal (MCLG) was set at 3 µg/L because it was believed that a lifetime of exposure at this level would not cause health problems in humans.   The detection limit was set at 0.1 µg/L for drinking water (Kansas Department of Health and Environment, 2004).

3.1.2.3 Detection Methods

Detecting atrazine in water can be challenging. The conventional method of testing water for the presence of atrazine is through gas chromatography and mass spectrometry. These methods are effective, but they are time consuming and expensive. At-home test kits are also available. These kits are based on calorimetric immunoassay methods, and they use antibodies to measure atrazine. These tests are faster and less expensive, but they can only measure to a limit of about 0.5 µg/L and, therefore, do not meet the standards for drinking water (Mosiello et al., 1998).

3.2 Biodiesel Fuel

Biodiesel fuel is a renewable energy source composed of long-chain mono alkyl esters, also known as fatty acid methyl esters (FAME). Biodiesel is created from the transestrification of plant or animal fats. Soybean, corn, canola, cottonseed, sunflower, rapeseed, and beef tallow are common sources of fat that go into biodiesel production.

Biodiesel is a very promising alternative fuel source because it contains 2.5 to 3.5 units of energy for every one unit of fossil fuel energy that goes into producing it. Biodiesel is also a cleaner burning fuel than diesel; it reduces greenhouse gas emissions and tailpipe emissions when used in place of diesel, and it does not emit the carcinogenic fumes associated with diesel fuel. B20, a

blend of 20% biodiesel and 80% diesel, is becoming a very popular option because it can be used in traditional diesel engines with no modifications (NREL, 2009).

3.2.1 Water in Biodiesel

Water is used to separate the catalyst in the transestrification reaction from the biodiesel (Kim, et al., 2008). If the biodiesel is not dried correctly after this separation, it may be contaminated with water. Biodiesel can also be contaminated with water if it comes into contact with it during transportation and storage. Excessive amounts of water can cause corrosion of the fuel tank and allow for the growth of microorganisms (NREL, 2009).

According to ASTM standard D6751-08, biodiesel can contain at most 0.05% (500 ppm) water and sediment, on a volume basis. The standard for detecting water in biodiesel is ASTM method D-2709 (National Biodiesel Board, 2008).

3.2.2 Glycerol in Biodiesel

Pure glycerol, $C_3H_5(OH)_3$, is an odorless, colorless, sugar alcohol with many uses. It is added to food as a preservative and a sweetener; it can be found in hygiene products such as toothpaste, lotion, soap, and shampoo; and it is often added to medicines such as cough syrup.

Crude glycerol is the main by-product of biodiesel fuel production; for every 3.79 L of biodiesel produced about 0.35 kg of crude glycerol is also made (USDA, 2006). The crude glycerol contains a lot of impurities, but it can go through an expensive refining process and become pure glycerol.

The glycerol by-product is much denser than the biodiesel, and it can be removed using gravitational separation. According to the ASTM standard D6751-08, biodiesel can only contain 0.240 % (2,400 mg/L) total glycerol, which is the sum of free glycerol and glycerides, and 0.020% (200 mg/L) free glycerol (National Biodiesel Board, 2008). If the fuel contains glycerol at levels higher than 0.020%, storage tanks and system fuel filters can become clogged. Engine

fouling, a problem in which a spark plug becomes coated in fuel, causing short circuit, is another consequence of high glycerol levels in biodiesel (ADM, n.d.).

The most common method for detecting glycerol in biodiesel is gas chromatography. The industry standard for this process is set by ASTM D-6584. This procedure can be done by a lab for around $275 per test (Hoar, 2008). There are a number of chromatographers that can perform this test manufactured by PerkinElmer, Koehler, Thermo Fisher Scientific, and Agilent Technologies.

3.2.3 Glycerides in Biodiesel

Glycerides are esters formed from glycerol and fatty acids. Animal fats and vegetable oil are composed of glycerides. Glycerides are present in biodiesel if the transestrification reaction was not complete. Each glyceride contains a glycerol molecule that was not released in the reaction (Kim, et al., 2008). Glycerides are a problem in biodiesel because they can contaminate the engine and clog filters (NREL, 2009).

The amount of glyceride that biodiesel can contain is given by ASTM standard D6751-08, and is the difference between total glycerol and free glycerol, or 0.220% (2,200 mg/L) (National Biodiesel Board, 2008). The standard for removing glyceride from biodiesel is ASTM method D-6584, gas chromatography (NREL, 2009).

3.3 Permittivity

Permittivity describes how well a dielectric material can store an electric charge. It is a frequency-dependent measurement composed of two factors- polarization and conduction. Polarization is the ability of a material to store an electrical charge, and conduction is the movement of electrical charges through the material (Scholte et al., 2002). When an electric field is applied to a material, the polar molecules in that material tend to align themselves so that the positive end of the molecule points to the negative side of the electric field and the negative end of the molecule points to the positive side of the field (Figure 3.1). This is called polarization, and it decreases the effective electric field. If the force binding the atoms together

is too strong for the atoms to align themselves with the electric field, the energy absorbed from the field is dissipated. This is called dielectric relaxation (Topp et al., 2000).
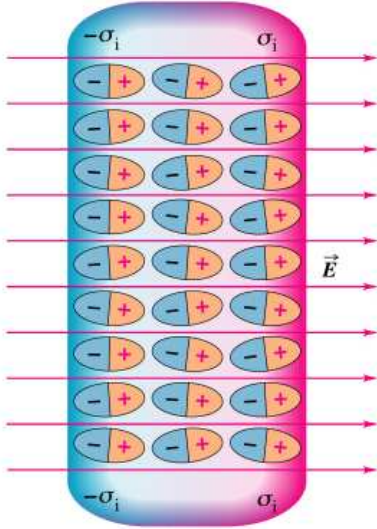
Figure 3.1: Polarization of molecules in an electric field (Becker, 2009)

Permittivity can be expressed as a complex variable. The definition of relative permittivity is:

$$\varepsilon^*_r(\omega) = \varepsilon'_r(\omega) - j\varepsilon''_r(\omega) \tag{3.1}$$

where ω = angular frequency,

$\varepsilon^*_r(\omega)$ = complex relative permittivity,

$\varepsilon'_r(\omega)$ = real part of permittivity,

$j = (-1)^{1/2}$, and

$\varepsilon''_r(\omega)$ = imaginary part of permittivity (Scholte et al., 2002).

The real part of permittivity is the amount of energy stored in the dielectric material from the alternating current field. It can be expressed as:

$$\varepsilon'_r(\omega) = \varepsilon'_e + \varepsilon'_d(\omega) \tag{3.2}$$

where ε'_e = apparent permittivity due to polarization of electrodes, and

$\varepsilon'_d(\omega)$ = frequency dependent permittivity (Scholte et al., 2002).

The imaginary component is the energy loss due to the AC electric field, also known as the relative loss factor. It is made up of two components.

$$\varepsilon"_r(\omega) = \left(\varepsilon"_d(\omega) + \frac{\kappa}{\omega\varepsilon_o}\right) \hspace{4cm} (3.3)$$

where $\varepsilon"_d(\omega)$ = relative dielectric loss,

κ = static electrical conductivity, and

ε_o = permittivity of free space = 8.854×10^{-12} F/m (Scholte et al., 2002).

Every dielectric material has a different conductivity and a different ability to polarize, and therefore a different permittivity. If a contaminant is added to a material, such as water, the contaminated water will have a different permittivity than pure water. If the concentration of the contaminant is increased, the permittivity will change again. For this reason, it is possible to detect changes in the chemical composition of a material by measuring the material's permittivity.

3.4 Regression Methods

Regression analysis is a technique for modeling a set of dependent variables, y-variables, using a set of independent variables, x-variables. The dependent variables are also referred to as the response variables, and the independent variables are called predictor variables. The y-variables can be predicted from the x-variables with a regression model.

It is useful to have two sets of data for the x- and y-variables for regression. One set of data, called the training dataset, is used to fit the model. Another set, called the validation set, is used to test the reliability of the model. Validating the model is important because the validation results show whether or not the model can be applied to other, similar, situations or if it is only valid for the dataset that created it. If the validation does not provide reasonable estimates of the response variables, it suggests that there may have been inconsistencies in data collection or that the original model is affected by over-fitting.

There are many methods that can be used to model the y-variables in terms of the x-variables. The methods that will be discussed in this section are: stepwise regression, partial least squares regression, artificial neural network, and stepwise regression with wavelet preprocessing.

A parameter is needed to judge the goodness-of-fit of the models. This makes it possible to determine how useful a prediction model is and also to compare one model to another. In this study, two statistical parameters were utilized to evaluate the goodness-of-fit.

The first parameter used was the coefficient of determination, or R^2-value. The coefficient of determination measures how well the regression line fits the data. It can also be thought of as the ratio of variation explained by the model to the total variation in the data. The procedure for calculating R^2 is shown in Equations 3.4 - 3.7.

$$S_{XX} = \Sigma(X_i^2) - \frac{(\Sigma X_i)^2}{n} \tag{3.4}$$

$$S_{YY} = \Sigma(Y_i^2) - \frac{(\Sigma Y_i)^2}{n} \tag{3.5}$$

$$S_{XY} = \Sigma(X_i Y_i) - \frac{(\Sigma X_i)\,(\Sigma Y_i)}{n} \tag{3.6}$$

$$R^2 = \frac{S_{XY}^2}{S_{XX} S_{YY}} \tag{3.7}$$

where, S_{XX} = is the sum of x deviations squared,

X_i = actual value,

Y_i = predicted value,

n = number of samples,

S_{YY} = the total sum of squares, and

S_{XY} = the sum of x deviations times y deviations (Ott and Longnecker, 2004).

The coefficient of determination is always between 0 and 1. A value of one means that the model explains the data perfectly and a value of 0 indicates that there is no fit. The R^2-value can also be explained in percentage. For example, having an R^2-value of 0.900 is equivalent to saying that the model explains 90% of the variability in the data. Because the coefficient of

determination can only fall within a set range of values, it is easy to decide what an acceptable value is, and 0.900 is usually considered a good value.

The second parameter used to compare goodness-of-fit of the models was the root mean square error (RMSE), also referred to as the residual standard deviation. The RMSE can be any positive value, it is not limited the way the R^2-value is. A large RMSE means that there is a lot of uncertainty in the model. It also indicates that variability in the x-variables is low. The formula for the root mean square error is shown below. It is more difficult to determine what an acceptable RMSE value is because this will change depending on the size of the data; however, RMSE values from models created using the same dataset can be compared.

$$RMSE = \left[\frac{\Sigma \, (Y_i - \hat{Y}_i)^2}{n - 2} \right]^{1/2} \tag{3.8}$$

where RMSE = root mean square error,

Y_i = the actual y-variable,

$\hat{Y}_i$ = the predicted y-variable,

n = the number of y-variables (Ott and Longnecker, 2004).

3.4.1 Stepwise Regression

Stepwise regression (SWR) is a combination of forward selection and backward elimination that is especially useful if multicollinearities exist in the data. Stepwise regression chooses significant x-variables from a large set of x-variables and uses them to predict the y-variables. The process begins with a model that has no x-variables in it. Then the program runs simple linear regression between the y-variable and each x-variable. The x-variable with the highest R^2-value enters the model first, followed by the x-variable which most increases the R^2-value. Predictors that improve the R^2-value continue to be added one-by-one, and the significance of the predictors is checked using an F-test. If a predictor is non-significant, which usually indicates that it is involved in a multicollinearity, that predictor is eliminated from the model. This process continues until all x-variables in the model are significant at a specified level and the model cannot be improved by adding more variables (Ott and Longnecker, 2001).

The formula used to determine if a variable j is significant enough to be added to a model with p variables is as follows:

$$F_j^+ = \max_j \left[\frac{RSS_p - RSS_{p+j}}{S^2_{p+j}} \right] > F_{in} \qquad (3.9)$$

where, S^2_{p+j} = the variance of the model with variables $p+j$,

RSS_p = the residual sum of squares of the model with p variables,

RSS_{p+j} = the residual sum of square of the model with variables $p+j$, and

F_{in} = a specified comparison value (Marengo et al., 2008).

The model that is created by stepwise regression is composed of x-variables and regression coefficients, as shown below.

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \qquad (3.10)$$

where, X_1 through X_n = x-variables,

β_1 through β_n = regression coefficients which are zero for x-variables not included in the model,

β_0 = the intercept of the line, and

$\hat{Y}$ = the predicted y-value (Ott and Longnecker, 2001).

One of the main benefits of stepwise regression is that it eliminates x-variables that are not significant; making it possible to see which x-variables were useful. In future data collection, it could be possible for data to only be collected for the significant variables. This can reduce time and cost in data collecting.

3.4.2 Partial Least Squares Regression

Partial least squares regression (PLS) is a good prediction method when there are a large number of correlated x-variables (Numerical Algorithms Group, 2007). PLS combines principal component regression (PCR) with multiple linear regression (MLR). PCR finds factors that maximize the variance of the independent variables, and MLR finds a variable to maximize correlation between the independent and dependent variables. PLS is an improvement over MLR and PCR because it uses information from both the x-variables and the y-variables to form

a model. As a result, PLS finds factors that maximize variance and correlation (Wise et al., 2006).

PLS selects "orthogonal linear combinations of predictors" from the predictor data that explain variance in the x-variables as well as the y-variables. These combinations are called factors, and the factors are used to calculate the PLS model. The model is initially calculated using a large number of factors. The number of factors that are actually needed is then estimated using cross-validation. After the number of factors is determined, the model is fit with this number of factors using linear regression. Finally, the model is used to estimate the y-variables with the given x-variables (Numerical Algorithms Group, 2007).

The models that are generated by PLS are:

$$X = TP + E \tag{3.11}$$

and

$$Y = UQ + F \tag{3.12}$$

where X is a matrix of the predictor variables,

Y is a matrix of the response variables,

T and U are the scores of X and Y,

P and Q = loadings,

E and F = error terms (Beebe and Kowalski, 1987).

3.4.3 Artificial Neural Network

An artificial neural network (ANN) is a computational model composed of artificial neurons that imitate the workings of the biological nervous system. The neurons are connected to each other by weights. Neural networks can be used as data modeling tools because they can find patterns and relationships in data. The neural networks discussed and used in this paper are feed-forward, back propagation neural networks.

The main advantage of ANNs over various statistical regression methods is that they are nonlinear, and so they can be used to model nonlinear patterns. The functions performed by the

15

artificial neurons are preformed in parallel, as opposed to sequentially, as the other regression techniques discussed in this chapter are.

A neural network has three layers: the input layer, the hidden layer, and the output layer (Jiang, et al., 2004). This is illustrated in Figure 3.2. The number of neurons, which are found in the hidden layer of the network, can be specified based on the complexity of the data being analyzed. More neurons make the program run slower but also can improve the model-building ability of the network.

The inputs are the x-variables. The network looks for patterns and relationships in inputs and uses them to predict the outputs. Once the outputs are predicted, the predictions are compared to the target values, the actual y-variables. The weights between neurons are adjusted based on the error in the predictions. This process continues until the error becomes very small, the exact value varies depending on the situation and what the user believes is allowable (Tokar and Johnson, 1999).
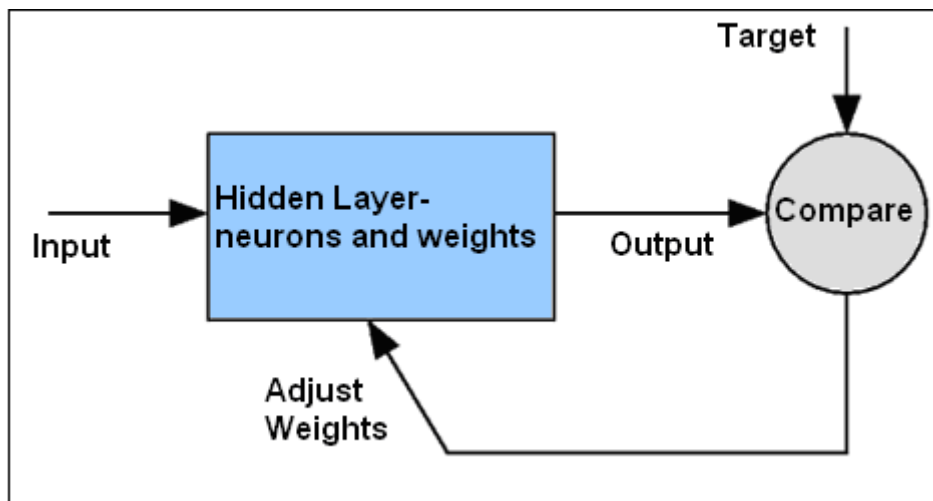
Figure 3.2: Artificial neural network layers (The Mathworks, 2008b)

3.4.4 Wavelet Transformation

Similar to Fourier transformation, wavelet transformation is a way to compress data. The difference is that the wavelet transform localizes both the time and frequency domains, whereas

the Fourier transform only localizes the frequency domain. The wavelet compression is a useful tool for signal preprocessing and de-noising (Leger and Wentzell, 2004).

The signal is passed through two filters: a high-pass filter and a low-pass filter. The coefficients produced by these filters are referred to as the detail coefficients and the approximation coefficients respectively. The number of coefficients in each group is equal to half the length of the original signal. The approximation coefficients can be put through the filters repeatedly, until there is only one coefficient left. This is illustrated in Figure 3.3. The number of times the signal is passed through the filters is called the level. Each level is divided into sections called wavebands. Each waveband contains information within a certain frequency range.
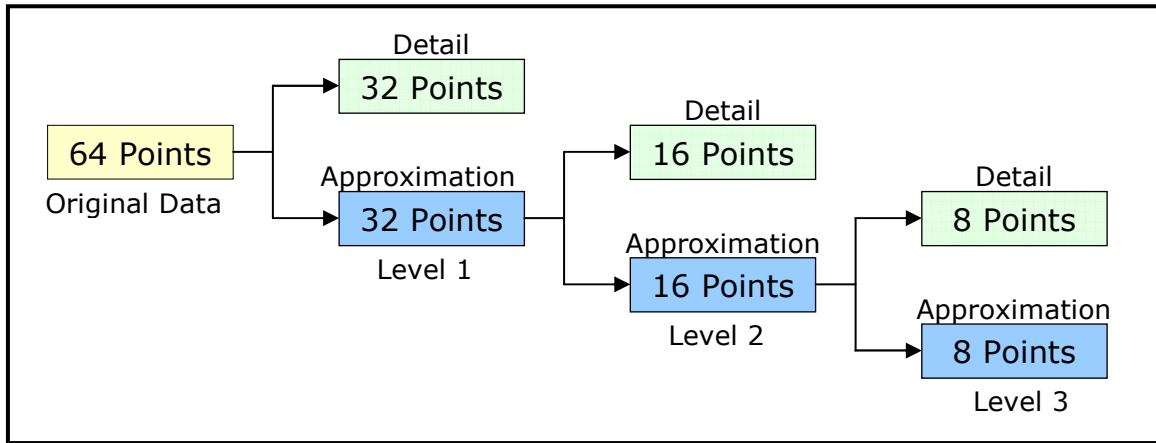
Figure 3.3: Wavelet decompression, adapted from (The MathWorks, 2008d)

The equations for the approximation coefficients and the detail coefficients are as follows.

$$A(j,k) = < f(n), \Phi_{j,k}(n) > = \sum_{n=0}^{N-1} f(n) \times \Phi^{*}_{j,k}(n) \qquad (3.13)$$

$$D(j,k) = < f(n), \varphi_{j,k}(n) > = \sum_{n=0}^{N-1} f(n) \times \varphi^{*}_{j,k}(n) \qquad (3.14)$$

$$\Phi_{j,k}(n) = s_0^{j/2} \Phi(s_0^{j} \cdot n - k) \qquad (3.15)$$

$$\varphi_{j,k}(n) = s_0^{j/2} \varphi(s_0^{j} \cdot n - k) \qquad (3.16)$$

where, *A (j,k)* = approximation coefficient,

   D (j,k) = detail coefficient,

$f(n)$ = signal of length N,

$\Phi_{j,k}(n)$ = a shifted scaling function, and

$\varphi_{j,k}(n)$ = a shifted wavelet function

j = discretized version of the scaling parameter

k = discretized version of the shifting parameter (Ge et al, 2007).

The simplest filter is called the Haar transform. The Haar transform is based on a step function. It replaces two steps with one wider step and one wavelet. The wide step is calculated by taking the average of the two original steps and the wavelet measures the difference between the two original steps. The Haar wavelet function is defined below (Nievergelt, 1999).

$$\Psi_{[0,1[} = \varphi_{[0,\frac{1}{2}[} - \varphi_{[\frac{1}{2},1[} \tag{3.17}$$

This is shown graphically in Figure 3.4. For every number r,

$$\Psi_{[0,1[}(r) = \begin{cases} 1 & \text{if } 0 \le r < \frac{1}{2}, \\ -1 & \text{if } \frac{1}{2} \le r < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{3.18}$$
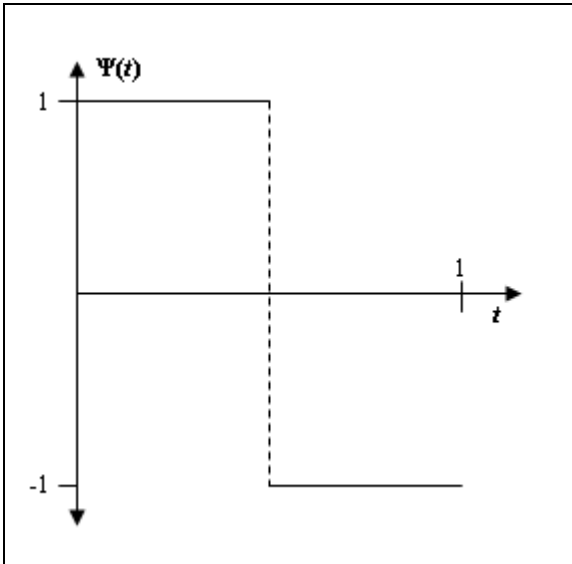
Figure 3.4: The Haar wavelet, adapted from (Nievergelt, 1999)

Transforming a signal using wavelet decomposition is an effective way of de-correlating data so that regression analysis can be done without any multi-collinearity issues (Ge et al, 2007). Wavelet compression of a signal followed by stepwise regression is especially helpful because

18

when stepwise selects a waveband to use in the model, the waveband that it selected is associated with several frequencies and frequency bands of different widths. The information collected at each of these frequencies and frequency bands can be thought of as significant for prediction.

A tiling diagram is a good way to represent the chosen frequencies. Figure 3.5 is a tiling diagram for a hypothetical dataset with 16 points. The darkened rectangles correspond to wavebands that are selected through the above mentioned procedure to be included in the model. The numbers written inside the rectangles indicate the number of data points included in each waveband. At level zero, each waveband only contains one data point, and is equivalent to the original data.
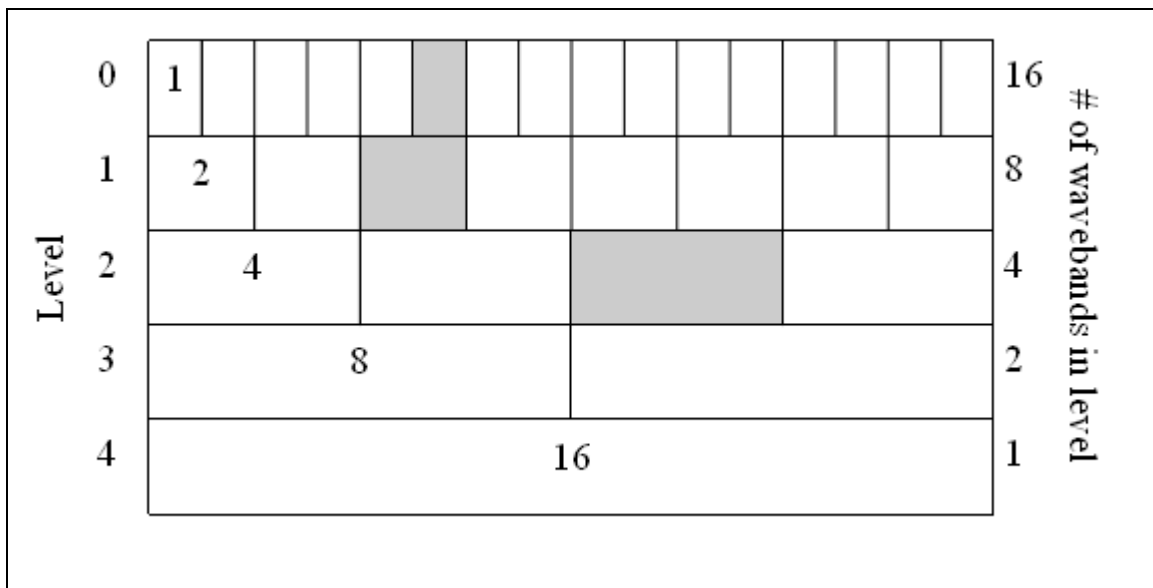
Figure 3.5: An example tiling diagram

Wavelet analysis has been used to analyze near-infrared reflectance spectroscopy. In a study done to relate the spectroscopic reflectance signals of 270 soil samples to the clay content of each sample, the regression of wavelet transformed data was an improvement over tradition regression. The study found that the R^2-value for wavelet transformed data was 0.99, whereas it was only 0.79 for non-preprocessed data (Ge et al., 2007).

CHAPTER 4 - Materials and Methods

4.1 Control Boxes

Two different control boxes were used with each of the four probes. The boxes will be referred to as the "old control box" and the "new control box." Each control box was composed of a signal generator, a microcontroller, and a gain and phase detector. The signal generator was used to generate sinusoidal signals which were sent through a probe. The gain and phase detector was used to measure the difference in gain and phase between the signal that went into the probe and the signal that left the probe (Tang, 2009).

4.1.1 The Old Control Box

The old control box measured the gain and the phase of the samples at 635 different frequencies ranging from 50 Hz to 120 MHz. This control box was programmed to test the gain and phase at each frequency three times. The old control box had to be connected to a computer to store the data because it had no internal memory. It did not have its own power supply, so it was connected to a BK Precision Triple Output DC Power Supply 1660 which supplied it with 10 V of power.

4.1.2 The New Control Box

The new control box measured the gain and phase of the samples at 524 frequencies ranging from 200 Hz to 400 MHz. It also had an input for a thermocouple, so the temperature of the samples could be monitored. Like the old control box, the new control box took a measurement at each frequency three times. The new box had its own memory storage, but for these experiments it was connected to a computer for data storage. This box also had its own power supply.

4.2 Probes

All of the experiments were conducted with four probes developed by researchers in the Instrumentation and Control Laboratory of the Biological and Agricultural Engineering Department of Kansas State University and with an impedance meter from Agilent Technologies.

The probes made at Kansas State University were constructed from a series of parallel stainless steel plates. The odd numbered plates were electrically connected with each other, and so were the even numbered plates. The probes were immersed in a sample where the gain and phase shift were measured at a few hundred different frequencies, determined by the control box. These probes varied by size and by the number of plates.

There were two general designs for the probes. The smallest probes, the "2 cm probe" and the "2.5 cm probe", both had plastic spacers between the metal plates to hold them in position. The cable that connected the probe to the control box was a 50-52 coaxial cable with shield. The larger probes, the "5 cm" and "7.5 cm probes", were constructed out of plastic, food storage containers. The plates were screwed into the bottom of the container and the plates were wired together inside of the container. The 50-52 coaxial cable exited from the top of the containers.

4.2.1 The 2 cm Probe

The smallest probe, shown in Figure 4.1, had 6 plates that were 2 cm wide, 3 cm tall, and 1 mm thick. The spacing between the plates was 1 mm. The area of each plate was 6 cm^2. This probe was designed to measure biodiesel fuel, and was built to these dimensions so that it could fit inside a fuel tank. This width also allowed the probe to be inserted directly into the glass bottles in which the solutions were kept. Figure 4.2 shows the gain and phase of 1 L of distilled water measured by the 2 cm probe.
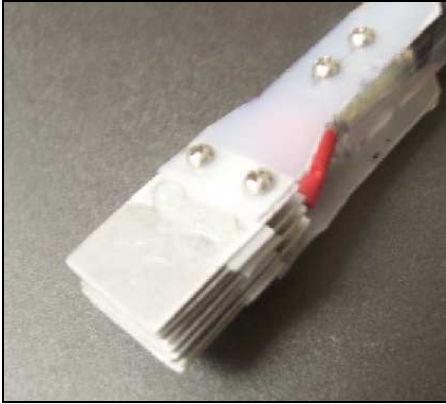
Figure 4.1: The 2 cm probe
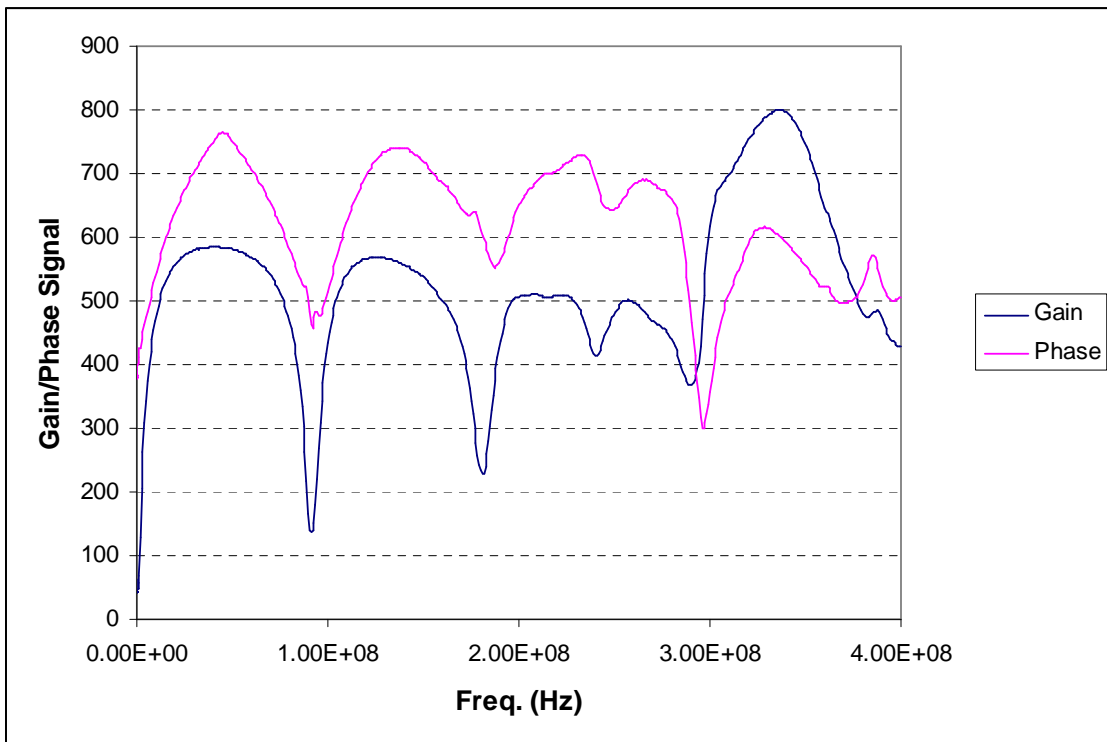
Figure 4.2: Signal for distilled water from the 2 cm probe and the new control box

4.2.2 The 2.5 cm Probe

The 2.5 cm probe is shown in Figure 4.3. It was made of 12 plates that were 2.5 cm wide, 3.8 cm tall, and 1 mm thick. The distance between the plates was 1.5mm. The area of each plate was 9.5 cm^2. Figure 4.4 shows the gain and phase of 1 L of distilled water measured by the 2.5 cm probe.
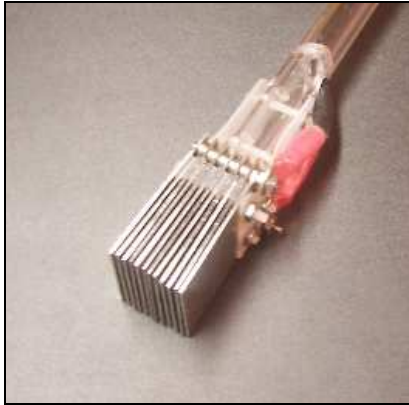
Figure 4.3: The 2.5 cm probe
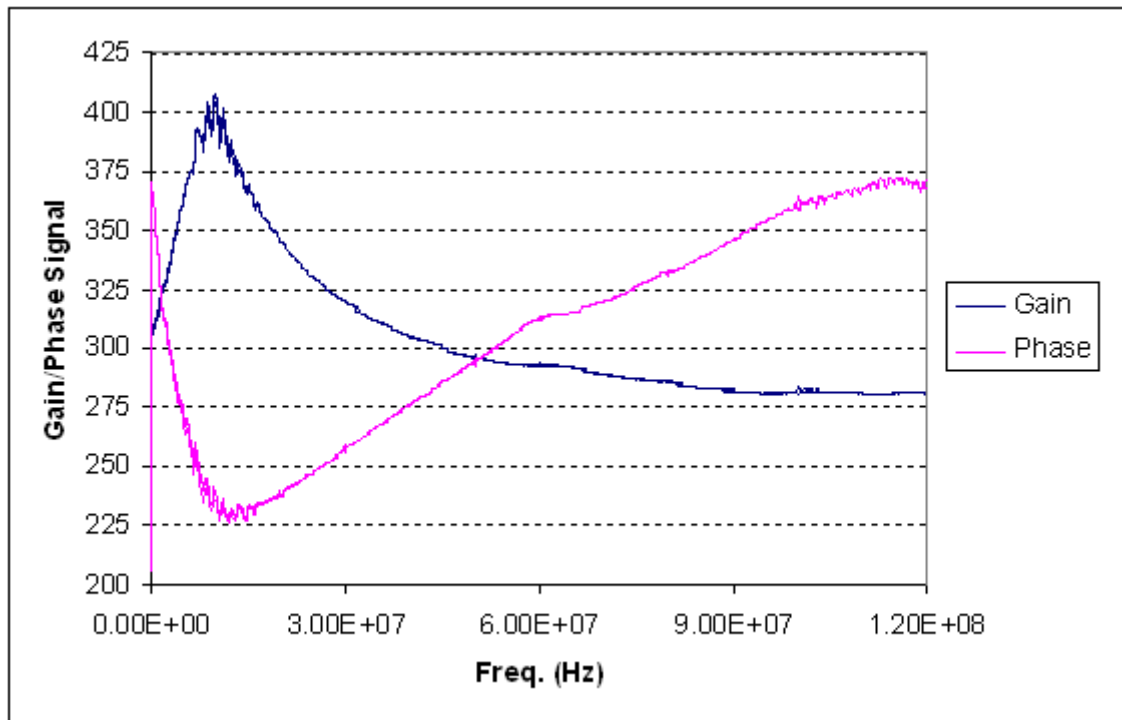
Figure 4.4: Signal for distilled water from the 2.5 cm probe and old control box

4.2.3 The 5 and 7.5 cm Probes

The two larger probes were constructed almost identical to one another. They each had 6 plates that were 5 or 7.5 cm wide, 3.5 cm tall and 5 mm thick. The spacing between the plates was 5 mm. The area of the 5 cm plates was 17.5 cm^2 and the area of the 7.5 cm plates was 26.3 cm^2. These probes fit into a rectangular container to measure the solution. The 7.5 cm probe is shown in Figure 4.5 and Figure 4.6 shows the gain and phase of 1 L of distilled water measured by the 5 cm probe.
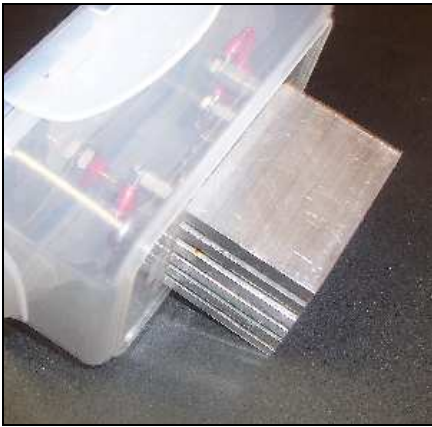
Figure 4.5: The 7.5 cm probe
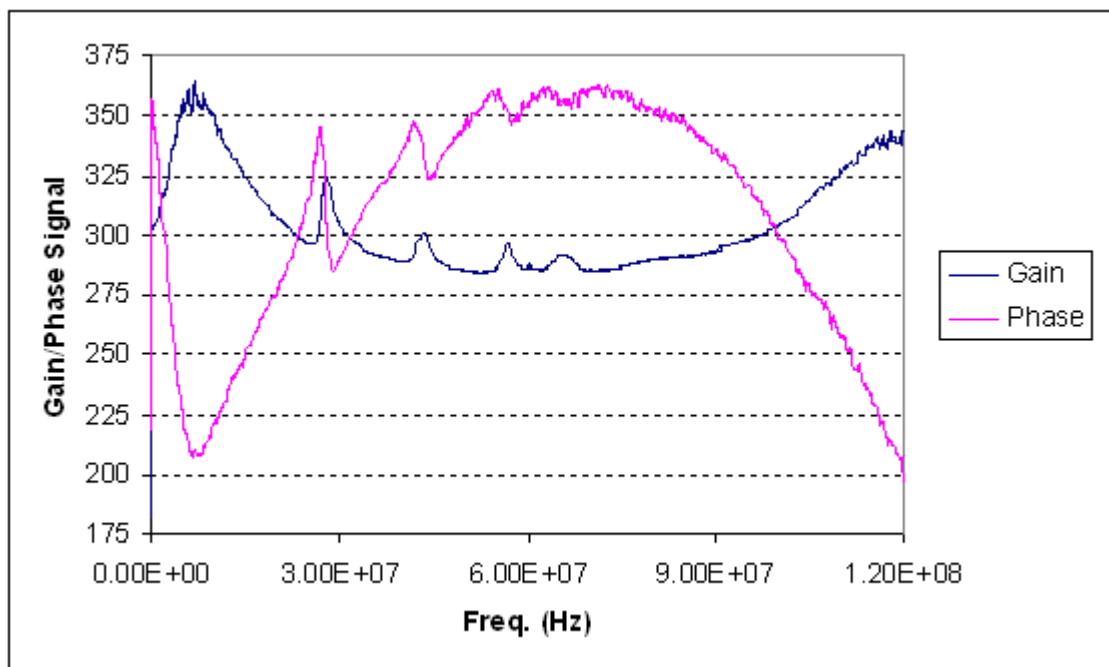
Figure 4.6: Signal for distilled water from the 5 cm probe and old control box

4.3 Impedance Meter

The impedance meter used in this research was an Agilent E4991A RF Impedance/Material Analyzer, shown in Figure 4.7. The impedance meter was set to measure ε', the real part of permittivity, and ε", the imaginary part of permittivity, at 596 frequencies between 1 MHz and 120 MHz in increments of 0.2 MHz. The method that it used to measure permittivity is called the capacitance method (Agilent, 2005). Measurements made in 1 L of distilled water are shown in Figure 4.8.

Figure 4.7: Agilent E4991A RF Impedance/Material Analyzer (Agilent, 2009)
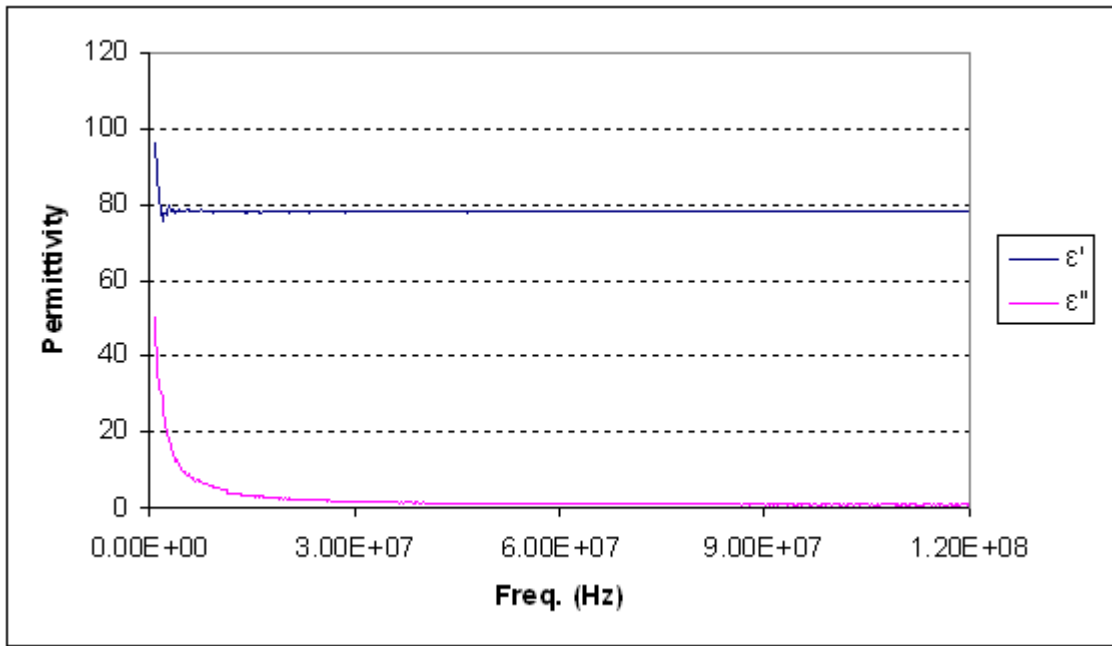
Figure 4.8: Signal for distilled water with the impedance meter

4.4 Water Sample Preparation

4.4.1 Nitrate Salts

The salts studied in this experiment were potassium nitrate (KNO_3), calcium nitrate ($Ca(NO_3)_2$), and ammonium nitrate (NH_4NO_3). Solutions of water and salt were made at concentrations of 0-200 μmol/L in 50 μmol /L increments and at concentrations of 300-1,000 μmol /L in increments of 100 μmol /L. Two sets of 0-200 μmol/L solutions were prepared, a training set and a validation set, making a total of 54 samples.

These samples were made from stock solutions of 0.01M KNO_3, 0.005M $Ca(NO_3)_2$, and 0.01M NH_4NO_3. The stock solutions were prepared by weighing the appropriate amount of dry salt, shown in Table 4.1, placing it in a 1 L volumetric flask, and adding distilled water to make the total volume 1 L. The salts were all water soluble and dissolved readily at room temperature.

Table 4.1: Molecular weights of salts

Salt	Molecular Weight (g/mol)	Mass added to prepare stock solution (g)
KNO_3	101.1	1.01
$Ca(NO_3)_2$	164.00	0.82
NH_4NO_3	80.04	0.80

To make the individual samples, the necessary amount of stock solution was measured into a 1 L volumetric flask using an Eppendorf Repeater Plus Pipettor. Distilled water was added to make the total volume of the sample 1 L. Table 4.2 shows the amount of stock solution needed to make each concentration. The contents of the flask were mixed thoroughly, and the samples were stored in 1 L glass bottles from Fisher Scientific. The flask was rinsed with distilled water three times after each sample was prepared to make sure that none of the samples contaminated the others.

Table 4.2: Volume of stock solutions used to make 1 L nitrate samples

Sample Concentration (µmol /L)	Volume of Stock Solution (mL)	Concentration of KNO_3 (mg/L)	Concentration of $Ca(NO_3)_2$ (mg/L)	Concentration of NH_4NO_3 (mg/L)
0	0	0	0	0
50	5	5.06	8.20	4.00
100	10	10.11	16.40	8.00
150	15	15.17	24.60	12.01
200	20	20.22	32.80	16.01
300	30	30.33	49.20	24.01
400	40	40.44	65.60	32.02
500	50	50.55	82.00	40.02
600	60	60.66	98.40	48.02
700	70	70.77	114.80	56.03
800	80	80.88	131.20	64.03
900	90	90.99	147.60	72.04
1,000	100	101.10	164.00	80.04

4.4.2 Atrazine

Solutions of water and atrazine were prepared at concentrations of 0-12 mg/L in increments of 2 mg/L. Two samples were made for each concentration, one for training and one for validation; there were a total of 14 atrazine solutions. The individual atrazine solutions were prepared from a 1,000 mg/L stock solution. The stock solution was composed of 1.187 g (1 mL) atrazine and 999 mL of distilled water. The individual samples were prepared from the stock solution using the same procedure as the nitrate salts. The amount of stock solution used to make each sample is shown in Table 4.3.

Table 4.3: Volume of stock solution used to make 1 L atrazine samples

Concentration of Sample (mg/L)	Volume of Stock Solution (mL)
0	0
2	2
4	4
6	6
8	8
10	10
12	12

4.5 Testing Procedure

4.5.1 One Liter Water Samples

The nitrate salts were tested in three groups: the training set of 0-200 μmol/L concentrations, the validation set of 0-200 μmol/L concentrations, and the set 300-1,000 μmol/L concentrations. The atrazine samples were the fourth group. The samples in each group were assigned a number, and a list of these numbers in a random order was produced using the RAND function in Microsoft Excel. This list of numbers was the order in which the samples in each group were tested. The purpose of the random order was to ensure that the patterns seen from sample to sample were due to the differences in chemical composition in the samples and not to uncontrolled changes of other factors, such as room temperature, from one test time to another. The room temperature was recorded each day before testing began to be sure that it did not greatly fluctuate.

For each test, the control box was connected to the computer. Microsoft HyperTerminal 5.1 was used to collect the data from the control box. HyperTerminal saved the data for each test as a .txt file. These files were opened in Microsoft Excel so that the frequency could be plotted against the gain and phase measurements.

4.5.1.1 The 2 cm Probe

A stand was built to keep the position of the 2 cm probe, relative to the bottle of solution, constant for each test (Figure 4.9). The stand held the probe in place. A shelf for the bottle to sit on was located 26 cm below the probe. The shelf could be slid in and out of the stand, allowing the bottle to be changed without moving the probe. The shelf had a circular depression in the center of it, exactly the same diameter as the bottom of the bottles, ensuring that every bottle was placed in the same position.

Figure 4.9: Testing with the 2 cm probe and new control box

Before the samples were tested with the 2 cm probe, the amount of solution in each bottle was adjusted to be exactly 900 mL so that the probe would be immersed to exactly the same depth for each test. After each test was done, the probe was dipped in distilled water and dried with pressurized air.

When the 2 cm probe was used to measure the 1 L samples, there were significant inconsistencies in the signals it measured, especially at higher frequencies. Experiments were done to investigate a number of possible causes, including bumping of the probe when it was removed from a sample, movement of the plates due to the pressurized air that was used to clean the probe, and the size of the container that held the solution. It was theorized that the 1 L glass bottles were too small, and that the probe needed to be surrounded by more solution to reduce the boundary effect. The boundary effect results from the interference of the electric field by the bottles in which the samples was measured. The glass might react to the electric field, or its permittivity may interfere with the measurement of the sample's permittivity. The tests were duplicated in a 10 gallon (37.9 L) glass aquarium to see if this would improve the results. The procedure for the tests done in the aquarium is described in Section 4.5.2.

4.5.1.2 The 2.5 cm Probe

A stand was also built to hold the 2.5 cm probe in place, as shown in Figure 4.10. The stand held the sample in place below the probe. Unlike the 2 cm probe, the 2.5 cm probe was removed from the stand after each test so that it could be cleaned. A ring clamp on the probe ensured that it always slid back into its holder on the stand at the same height.
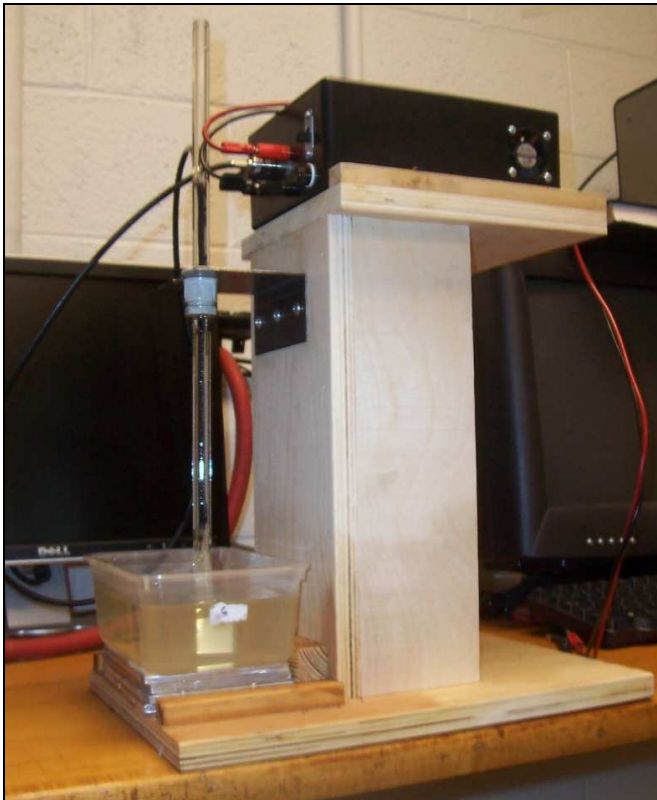
Figure 4.10: Testing with the 2.5 cm probe and old control box

Each sample was poured into a rectangular plastic container to be tested. A line was drawn on the container to which the water was always filled. The purpose of the line was to make sure that the probe was always immersed to the same depth. After each test the probe and the container were washed thoroughly using dish soap and a brush. They were dried with pressurized air. This was done to ensure that one sample did not contaminate another and that the water from washing the equipment did not mix with any of the samples.

4.5.1.3 The 5 and 7.5 cm Probes

The testing procedures for the 5 cm probe and the 7.5 cm probe were identical. For each test, the sample was poured into the plastic container used for the tests with the 2.5 cm probe. Another fill line was drawn for these probes. The container sat on a board with the control box, as shown in Figure 4.11. The board had wooden guides on it to line up the container and the control box so that their positions remained the same for every test. The probes fit securely into a rectangular hole in the fiber glass lid that covered the container.
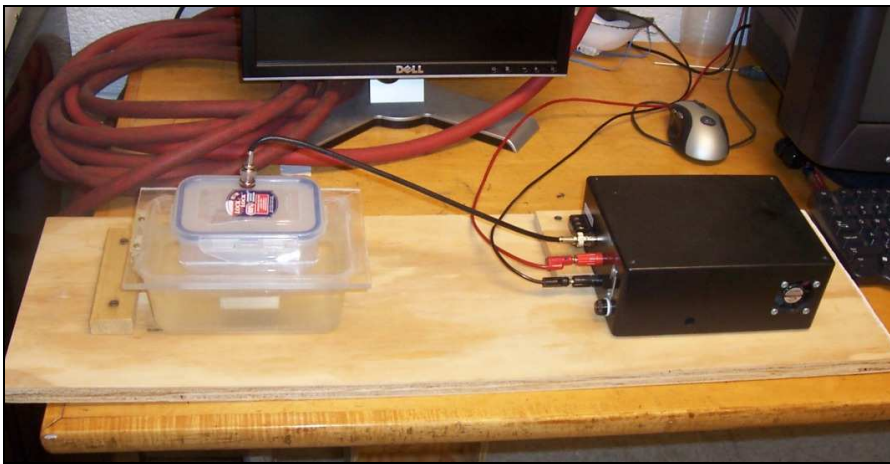
Figure 4.11: Testing with the 5 cm probe and old control box

The coaxial cable connecting the probe to the control box remained attached to the control box between tests, but had to be removed from the probe after each test so that the probe could be cleaned. After each sample was tested with the 5 cm probe and the 7.5 cm probe, the plastic container, the lid, and the probes were washed and dried with the same procedure used for the 2.5 cm probe.

4.5.1.4 Impedance Meter

A stand was also built so that the probe on the impedance meter remained in a fixed position for each sample measured. This is illustrated in Figure 4.12. Before tests could be performed, the machine had to be calibrated for open, short, and 50Ω load using a 16195B cal kit. The calibration was done to remove error and make the measurements as accurate as possible.
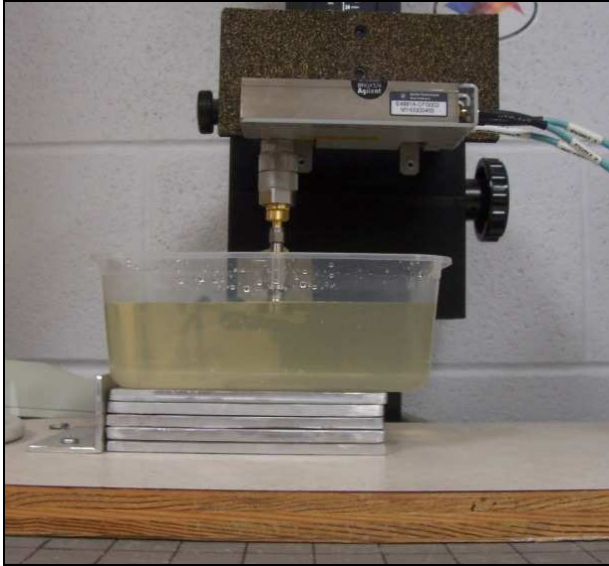
Figure 4.12: Testing with the impedance meter

The plastic containers were also used to hold the samples for the impedance meter tests. Once again, the container was always filled to a specific line to keep the probe depth constant throughout the tests. The impedance meter was connected to a computer, and the data was collected using 85070E software from Agilent Technologies. Between tests, the probe was dipped into deionized water and dried with a lint-free cloth.

4.5.2 Testing Procedure for 36 L Water Samples

To reduce the influence of the boundary effect, the water was also tested in a 37.9 L aquarium. The 2, 2.5, and 7.5 cm probes were used with the new control box for these tests. A special top was built for the aquarium that held each of the probes in a fixed position, as shown in Figure 4.13. The center of each probe was 12.5 cm from the wall of the aquarium and from the center of the adjacent probe.

Figure 4.13: Testing the 36 L samples and new control box

Because of the large amount of water needed for these tests, individual samples were not used. Instead, for each contaminant, the aquarium was filled with 36 L of clean, distilled water. The water was measured with each of the three probes twice, once for training and once for validation, and then a specific amount of contaminant was added from a stock solution. Each stock solution was 1 L, and they were prepared at concentrations of 0.15 M for KNO_3 and NH_4NO_3, 0.075 M for $Ca(NO_3)_2$, and 4,000 mg/L for atrazine. The amount of contaminant used to make each stock solution is shown in Table 4.4. Table 4.5 shows the amount of nitrate stock solution added for each concentration, and Table 4.6 shows the amount of atrazine stock solution added for each test.

Table 4.4: Amount of contaminant used to make stock solutions for 36 L tests

Contaminant	Stock Solution Concentration	Mass Added (g)
KNO_3	0.15 M	15.165
NH_4NO_3	0.075 M	12.300
$Ca(NO_3)_2$	0.15 M	12.006
Atrazine	4,000 mg/L	4.748

33

Table 4.5: Dilution of 36 L nitrate solutions

Concentration (μmol/L)	Total Water (L)	Volume Stock Solution Added (mL)
0	36.0	0
50	36.0	12.0
100	36.0	12.0
150	36.0	12.0
200	36.1	12.0
300	36.2	24.2
400	36.2	24.2
500	36.3	24.5
600	36.4	24.5
700	36.6	25.0
800	36.8	25.2
900	37.0	25.5
1000	37.2	26.0

Table 4.6: Dilution of 36 L atrazine solutions

Concentration (mg/L)	Total Water (L)	Volume Stock Solution Added (mL)
0	36.0	0
2	36.1	18.0
4	36.1	18.0
6	36.2	18.0
8	36.2	18.2
10	36.3	18.2
12	36.4	18.2

All three probes remained in the water/contaminant solution during the entire series of tests for a particular contaminant. The contents of the aquarium were stirred carefully after each addition of stock solution to make sure that the newly added contaminant was evenly distributed throughout the aquarium. After all the tests were done for one contaminant, the probes were cleaned with dish soap and water and dried with pressurized air.

4.5.3 Testing Procedure for Biodiesel Samples

Biodiesel fuel was tested with water, glycerol, and glyceride as contaminants. The testing procedure for the biodiesel samples was similar to the procedure for the 36 L water samples. The aquarium was filled with 36 L of clean soybean biodiesel at the beginning of each series of

tests, measured twice with each of the three probes, and then contaminant was added. Eleven concentrations were measured for each of the contaminants from 0 ppm to two times the maximum limit allowed by the ASTM standards, in evenly distributed increments, and ten concentrations were measured from two times the limit to ten times the limit, also in evenly distributed increments. Thus, a total of twenty-one concentrations were measured for each contaminant. Table 4.7 shows the concentrations of each contaminant that were measured. The row in bold is the ASTM limit for each of the impurities.

Table 4.7: Concentrations of contaminants measured in biodiesel tests

Test Number	Water (ppm volume basis)	Glycerin (mg/L)	Glyceride (mg/L)
1	0	0	0
2	100	40	440
3	200	80	880
4	300	120	1320
5	400	160	1760
6	**500**	**200**	**2200**
7	600	240	2640
8	700	280	3080
9	800	320	3520
10	900	360	3960
11	1000	400	4400
12	1400	560	6160
13	1800	720	7920
14	2200	880	9680
15	2600	1040	11440
16	3000	1200	13200
17	3400	1360	14960
18	3800	1520	16720
19	4200	1680	18480
20	4600	1840	20240
21	5000	2000	22000

To calculate the amount of contaminant that needed to be added to the biodiesel for each test, the glycerin and glyceride limits had to be converted to a volume basis using the density of each substance. The volume of contaminant added is shown in Table 4.8. To make sure that the contaminants were evenly distributed throughout the aquarium, approximately 2 L of biodiesel was drained from the aquarium to a smaller container, and the contaminant was added to the

smaller container. The container was vigorously shaken, and then the contents of the container were stirred back into the aquarium.

Table 4.8: Amount of contaminant added for each biodiesel test

Test No.	Water (mL)	Glycerin (mL)	Glyceride (mL)
1	0	0	0
2	3.60	1.00	15.0
3	3.60	1.00	15.0
4	3.60	1.00	15.0
5	3.60	1.00	15.0
6	3.60	1.00	15.0
7	3.60	1.00	15.0
8	3.60	1.00	15.0
9	3.60	1.00	15.0
10	3.60	1.00	15.0
11	3.60	1.00	15.0
12	14.5	4.00	60.0
13	14.5	4.00	60.0
14	14.5	4.00	60.0
15	14.5	4.00	60.0
16	14.5	4.00	60.0
17	14.5	4.00	60.0
18	14.5	4.00	60.0
19	14.5	4.00	60.0
20	14.5	4.00	60.0
21	14.5	4.00	60.0

Experiments were done to determine if the probes were able to detect more than one impurity at a time. The first contaminant was added, and when all 21 concentrations of that contaminant had been measured, the second contaminant was added. The 21 concentrations of the second contaminant, with the first contaminant at its maximum level, were measured, and then the same procedure was repeated for the third contaminant. The order in which the contaminants were added for each series of tests is given in Table 4.9. Only the first series of tests is discussed in this thesis.

Table 4.9: Order that contaminants were measured for each biodiesel test series

Test Series 1	Test Series 2	Test Series 3
Water	Glycerin	Glyceride
Glycerin	Glyceride	Glycerin
Glyceride	Water	Water

4.5.4 Temperature Experiments

The temperature of the room in which the testing was done could not be kept constant. The room temperature fluctuated very little, not more than 3ºC; however, tests still needed to be done to determine if these small fluctuations would affect the results. To study the temperature effect, 1 L samples of distilled water, 100 µmol/L of each nitrate salts, and 200 µmol/L of each nitrate salt were put in the refrigerator overnight. The samples were removed from the refrigerator in the morning, and measurements were taken approximately every 20 minutes with the 5 and 7.5 cm probes until the sample reached room temperature.

These experiments showed that varying the temperature in the range of 3ºC to 23ºC affected the results very little. Figure 4.14 shows the gain signals from distilled water as its temperature increased. These measurements were taken with the 5 cm probe and the old control box. Figure 4.15 shows the phase signals from these tests.
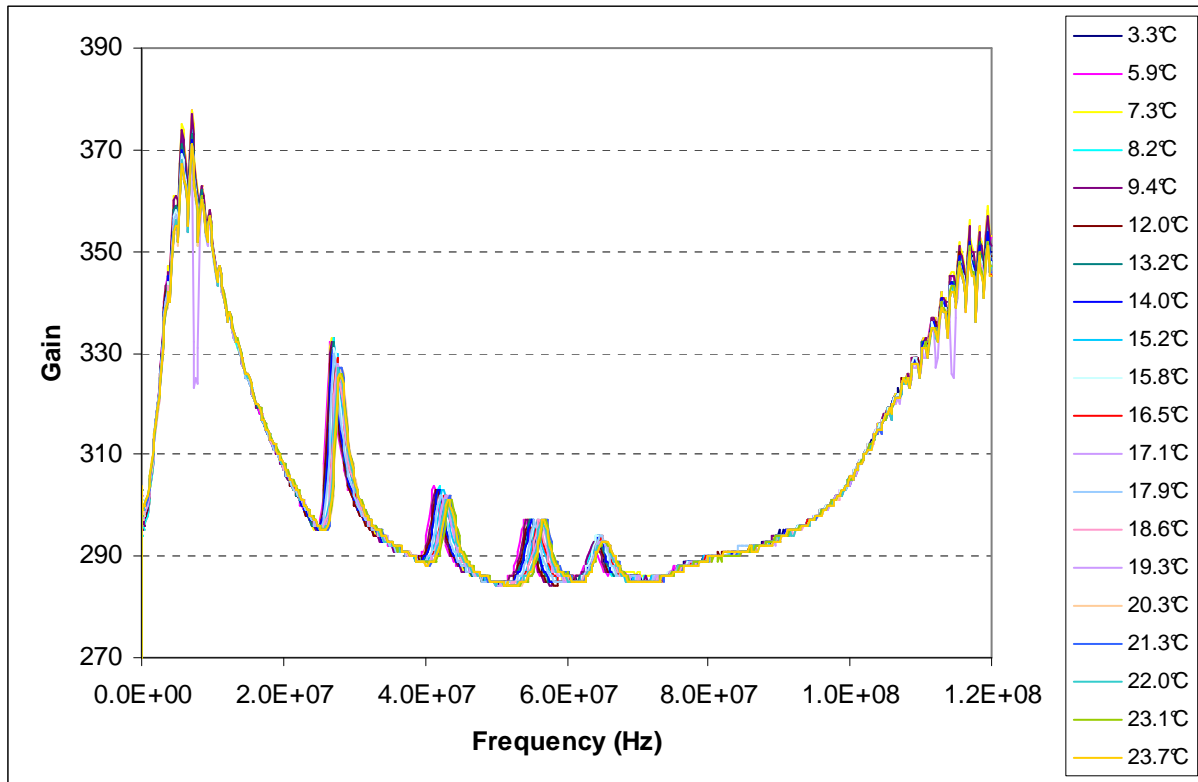
Figure 4.14: Gain signals with increasing temperature from 1 L of distilled water with the 5 cm probe and the old control box
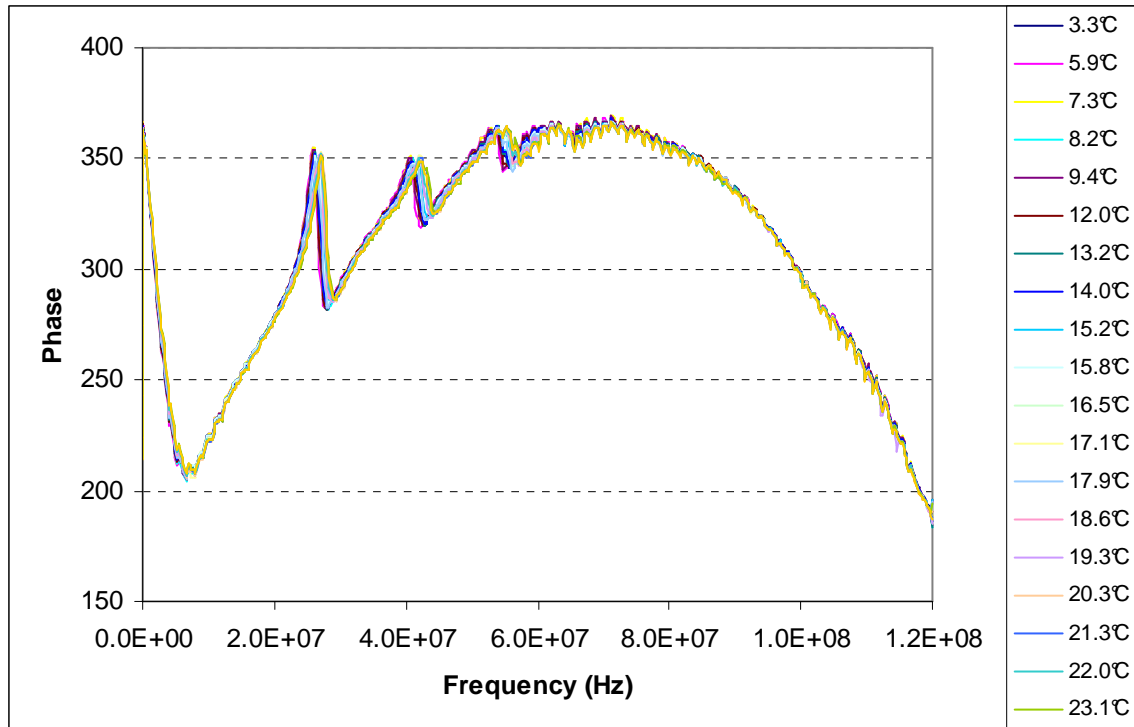
Figure 4.15: Phase signals with increasing temperature from 1 L of distilled water with the 5 cm probe and the old control box

4.6 Data Analysis

Four different programs were written using MatLab 7.6.0 (The Mathworks, 2008a) and various toolboxes for MatLab. The regression methods used were stepwise regression, partial least squares regression, artificial neural network, and stepwise regression of wavelet transformed data. The gain and/or phase data were the x-variables and the concentration or molecular weight values were the y-variables.

The probes developed at K-State measured the gain and phase of every sample three times at each frequency. These three signals were averaged, and the regression programs were run on the average values. Only one reading was taken by the impedance meter. For each group of samples, regression was done using just the gain signal, just the phase signal, and the gain and phase signals together. The same thing was done for the impedance meter with the real and imaginary permittivity signals. Regression was done on the nitrate samples for just the low concentrations (0-200 µmol/L), just the high concentrations (300-1,000 µmol/L), and for all of

38

the concentrations together. This was also done for the fuel tests; the concentrations up to two times the ASTM standard were considered the low concentrations, and the concentrations that were higher than twice the ASTM standard were considered the high concentrations. As a result, there were nine subsets of data from each sensor for each regression method. The results from these subsets were analyzed both individually and as a group to show the overall trend for a particular probe or regression technique.

Each group of samples consisted of a training data set, used to create the regression model, and a validation data set that was used to verify the model. Because only one set of high concentration nitrate salts was made for the 1 L samples, the odd numbered concentrations (300, 500, 700, and 900 µmol/L) were used as the training set, and the even numbered concentrations (400, 600, 800, and 1,000 µmol/L) were used as the validation set. For the 36 L nitrate tests, all of the odd numbered concentrations (50, 150, 300, 500, 700, and 900 µmol/L) were used as the training data and the even numbered concentrations (0, 100, 200, 400, 600, 800, and 1,000 µmol/L) were used as the validation data. Alternating concentrations were also used for the biodiesel tests. The data was analyzed without including 0 µmol/L in the validation data, and it was found that excluding 0 µmol/L from regression did not change the results very much.

4.6.1 Stepwise Regression

For this research, stepwise regression was done using MatLab 7.6.0 (The Mathworks, 2008a) and the Statistics Toolbox 7.0 (The Mathworks, 2008c). Figure 4.16 is a flowchart of the program. A copy of the complete program can be found in Appendix B.

The program begins by loading the training and validation data, as well as a list of the frequencies that correspond to each x-variable. Next, a stepwise model is made with the training data using the function stepwisefit. The p-value that is needed to add a variable into the model is 0.05, a commonly used value, and the p-value needed to remove a variable is 0.10. A list of the frequencies at which the x-variables were significant enough to be included in the model is stored. After this, the model is used to predict the y-variables from the training and validation x-variables. The predictions are stored, and the R^2-values and RMSE values are calculated and stored. Finally, a plot is made that shows the actual y-values on the x-axis and the predicted y-values on the y-axis. A perfect fit line is drawn through the actual values.
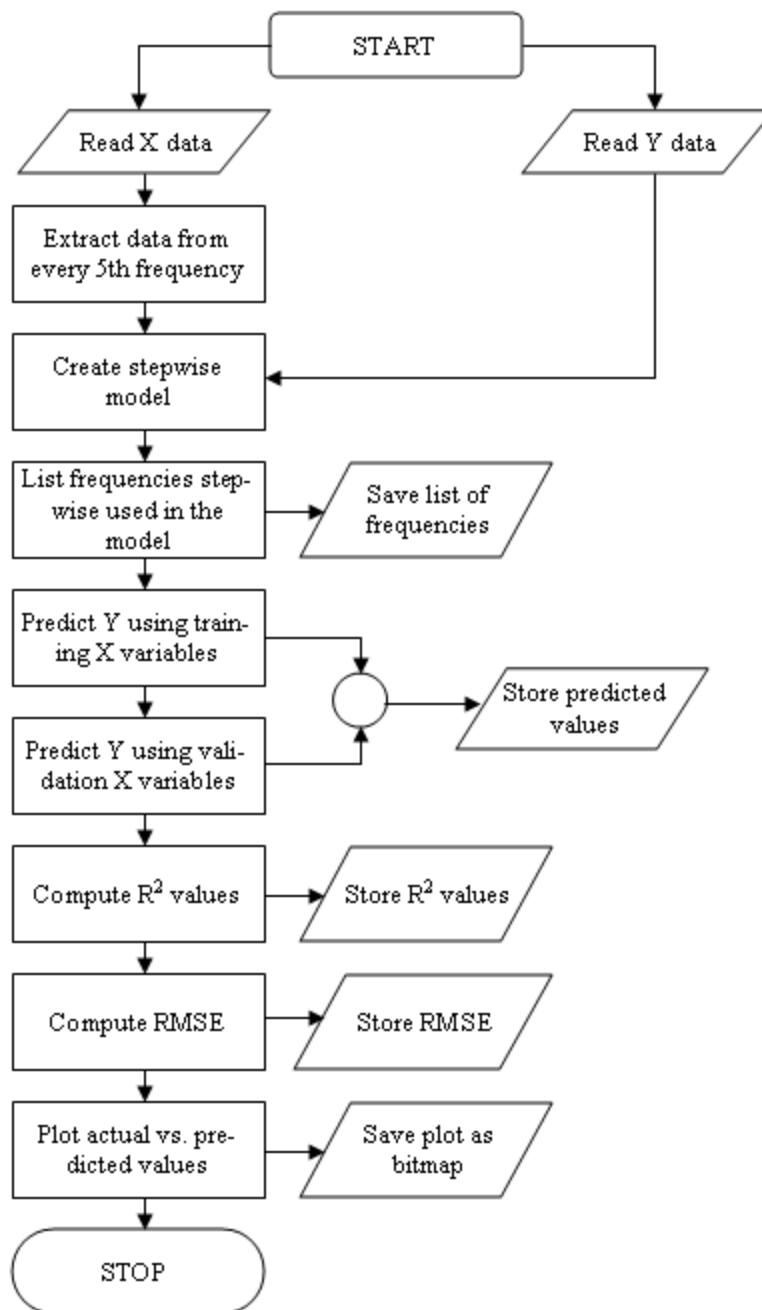
Figure 4.16: A flowchart of the stepwise regression program

4.6.2 Partial Least Squares Regression

The program that was used for PLS regression was written using PLS Toolbox 4.0 (Eigenvector Research Inc., 2006) and MatLab 7.6.0 (The Mathworks, 2008a). A flowchart of the program is shown in Figure 4.17, and a copy of the program is provided in Appendix C. The program begins by cross-validating the data with leave-one-out cross validation using the crossval function. The predictive residual error sum of squares, or PRESS, statistic is calculated, along with the cumulative PRESS (CUMPRESS), which is the sum of columns in the PRESS matrix. The formula for calculating the PRESS is:

$$PRESS = \Sigma \; (Y_{pred} - Y_{act})^{\,2} \tag{4.1}$$

$$\text{where } Y_{pred} = \text{the predicted Y-value, and}$$
$$Y_{act} = \text{the actual Y-value.}$$

The program then selects the value at which CUMPRESS is smallest; this is the number of factors that explain the most variation. The data is mean-centered with the preprocess function, which means that all of the columns are adjusted to have a mean of zero, and then a model is created using the training data and the number of factors previously determined with the pls function.

The model uses the training x-variables to predict the training y-variables. The same thing is done for the validation data. Once the predictions have been made, the program computes the R^2-values and the RMSE values for both training and validation data. Finally, a plot is made that shows the actual y-values on the x-axis and the predicted y-values on the y-axis. A perfect fit line is drawn through the actual values.
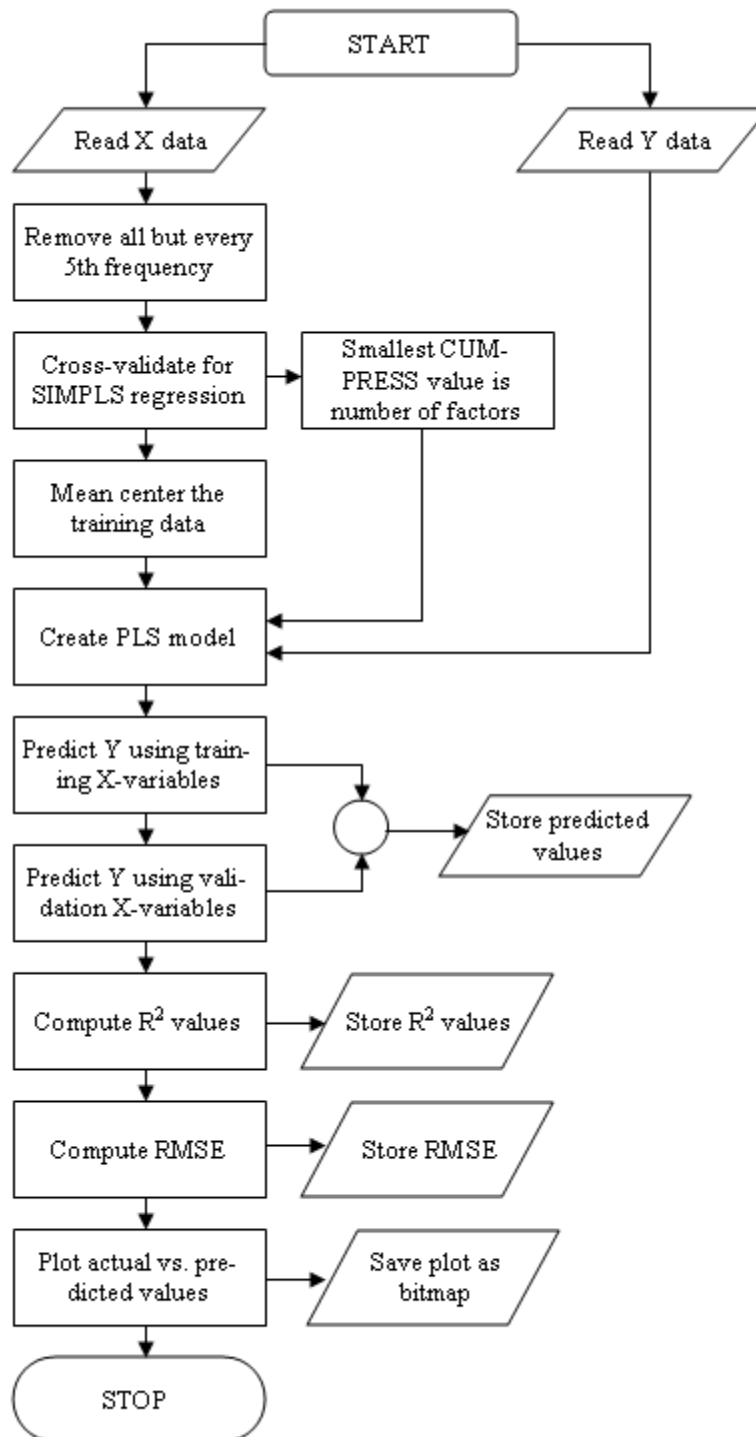
Figure 4.17: A flowchart of the PLS program

4.6.3 Artificial Neural Network

The neural network program in this research was written using MatLab 7.6.0 (The Mathworks, 2008a) and the Neural Network Toolbox 6.0.1 (The Mathworks, 2008b). A flowchart for the program is shown in Figure 4.18. A copy of the program can be found in Appendix D.

The program begins by removing all rows with constant values, using the removeconstantrows function. If a row has a constant value all the way across, there was no variation in the gain or phase measurement from one concentration to another at the frequency corresponding to the constant row. Because there is no variability in the x-variables at frequencies where this occurs, this data will not be useful in model building.

Next, the data is rescaled using the mapminmax function, so that the smallest value in a row is -1 and the largest value is 1. A feed-forward back propagation network is created with ten hidden neurons, a common number to use for a dataset of this size, using the newff function. Then, from the rescaled data, the training dataset is divided into three subsets- a training subset, a validation subset, and a test subset. The dividevec function randomly separates the data so that 60% of the samples are in the training subset, 20% are in the validation subset, and 20% are in the test subset.

In the following step, the train function is used to create a model with the network and the training data subset. The model is tested and adjusted with the validation and test subsets. This is done to avoid over-fitting. Once the model has been adjusted, the entire training data set and the entire validation data set are each fed into the model and the model predicts the outputs. After the predictions are made, the program computes the R^2-values and the RMSE values for both training and validation data. Finally, a plot is made that shows the actual y-values on the x-axis and the predicted y-values on the x-axis. A perfect fit line is drawn through the actual values.
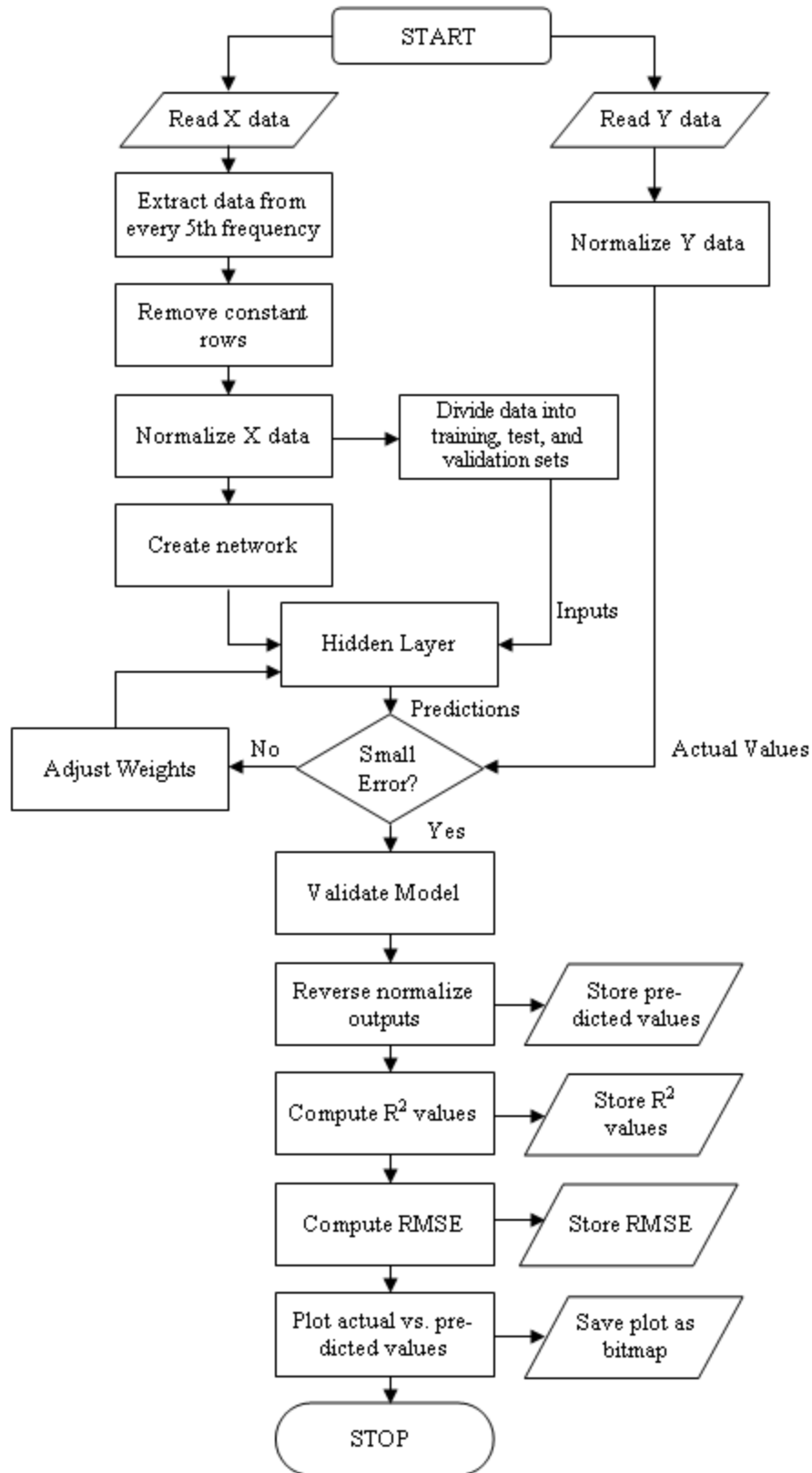
Figure 4.18: A flowchart for the artificial neural network program

4.6.4 Wavelet Transformation

The wavelet program was written using MatLab 7.6.0 (The Mathworks, 2008a) and the Wavelet Toolbox 4.2 (The Mathworks, 2008d). A flowchart of the program is shown in Figure 4.19 and the complete program is provided in Appendix E.

The program begins by determining the highest level transform that can be performed given the number of x-variables present. This is calculated by rounding log (length of x data)/log (2) down to the nearest whole number. Data from five different levels will be used to make the model, so the lowest level is found by subtracting four from the highest level. Once the levels have been determined, the x-data is transformed with a Haar wavelet and the appropriate levels using the wavedec function. This compresses the data from a certain range of frequencies together into an approximation and a detail coefficient.

Stepwise regression is run on the transformed training data using the function stepwisefit. The model made by stepwise is applied to the wave-transformed training and validation x-variables to predict the Y-variables. The R^2-values and the RMSE values are calculated and stored. A plot is made that shows the actual y-values on the x-axis and the predicted y-values on the y-axis. A perfect fit line is drawn through the actual values.

A list of the data points that were selected by stepwise is made and converted into a table of the levels and wavebands corresponding to each point. This information is used to make a tiling diagram. The tiling diagram is constructed using the annotation function. The wavebands that were selected are colored in, and the transform levels and frequencies are added to the diagram.
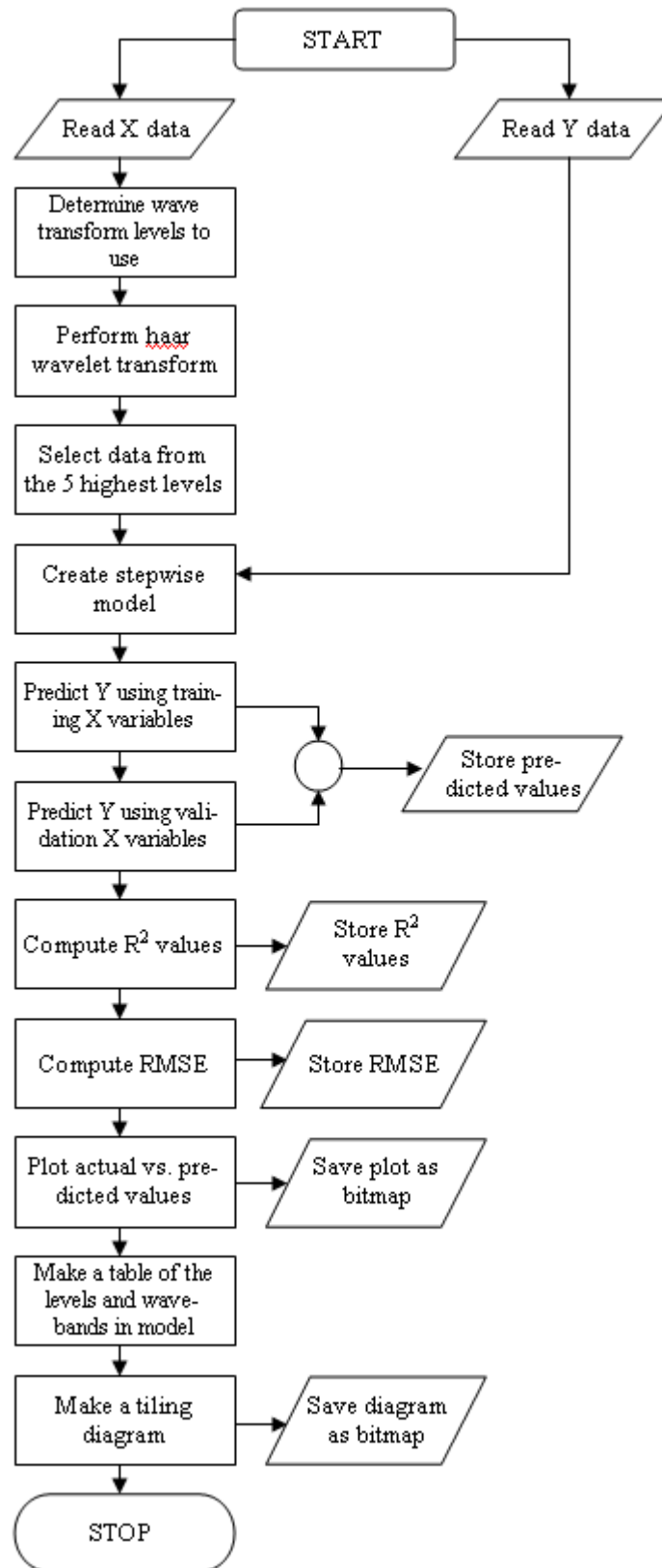
Figure 4.19: A flowchart of the wavelet program

CHAPTER 5 - Results and Discussion

5.1 Nitrate Salts and Water

The results for predicting nitrate concentration and molecular weight in water were promising. There were many high R^2-values from these tests. These values are given in tables that have the validation R^2-values above 0.900 bold and in blue. For tables in which there were no R^2-values above 0.900, the highest validation value is bolded. Tables are also given for the RMSE values.

5.1.1 Concentration

Overall, the probes and regression methods used in this research proved to be reliable for measuring nitrate concentration in water. The highest R^2-values obtained for training and validation were both 1.00, and the lowest RMSE values for both training and validation were 0.00 µmol/L. This means that for some of the tests that were completed, the regression model was able to explain 100% of the variation in the data.

5.1.1.1 One Liter Samples

The R^2-values for predicting the concentration of nitrate salt in the 1 L samples are given in Table 5.1, and the RMSE values are given in Table 5.2. The highest training R^2-value for the 1 L samples was 1.00, and the highest validation R^2-value was 0.988. The lowest training RMSE value was 0.00 µmol/L, and the lowest validation RMSE value was 22.0 µmol/L.

5.1.1.1.1 Comparison of Probes

For the 1 L samples, the 5 cm probe and the 7.5 cm probe were the most accurate, based on R^2- and RMSE values, of measuring the nitrate concentration. When both the high and low concentrations were used to build the models, the lowest validation R^2-value for each of these probes was around 0.858, meaning that 85% of the variation in the validation data was explained by the model made using the training data. Both the 5 and 7.5 cm probes had training R^2-values above 0.900 for approximately 90% of the data subsets. The validation R^2-values for the 5 cm

probe were above 0.900 for 58% of the data subsets, and the validation R^2-values for the 7.5 cm probe were above 0.900 for 69% of the data subsets.

The 2 cm probe did not do a good job of predicting concentration with the 1 L samples. Most of the training R^2-values for this probe were good, many above 0.90, but the average validation R^2-value for this probe was only 0.171 and there were no validation R^2-values above 0.900. Therefore, the variations in the data were not consistent from the training set to the validation set for this probe.

The 2.5 cm probe and the impedance meter did not perform as well as the larger probes, but they predicted nitrate concentration more accurately than the 2 cm probe. Both of these sensors had training R^2-values that were mainly above 0.900, but the 2.5 cm probe only had validation R^2-values above 0.900 for 33% of the regression models. The impedance meter had validation R^2-values above 0.900 for 64% of the regression models.

5.1.1.1.2 Comparison of Regression Techniques
The partial least squares regression models were able to explain more variation in the 1 L nitrate concentration data than the models from the other regression methods. The models made using the real part of permittivity from the impedance meter with only the 0-200 µmol/L concentrations were unreliable for all of the regression methods. When these measurements are excluded from the data, the lowest training R^2-value from PLS was 0.917 and the lowest validation R^2-value was 0.814. For the training data, the PLS R^2-values were above 0.900 89% of the time and the validation R^2-values were above 0.900 64% of the time, more often than for any other regression method studied.

Neural network proved to be the least effective method in predicting nitrate concentration. The average training R^2-value from ANN was only 0.802, even without the poor results from the 2 cm probe, whereas the average values for the other techniques were all above 0.900. The training R^2-values for this method were only above 0.900 49% of the time, and the validation R^2-values were above 0.900 11% of the time.

49

Traditional stepwise regression and wavelet transform followed by stepwise regression were comparable in their abilities to detect nitrates in water. Both had average training R^2-values around 0.960 and average validation values around 0.840. The RMSE values were lower, overall, for traditional stepwise regression than for stepwise regression on wavelet transformed data.

Table 5.1: R^2-values for predicting concentration of 1 L nitrate samples

		SWR		PLS		ANN		Wavelet	
		Training	Validation	Training	Validation	Training	Validation	Training	Validation
2 cm Probe	All Gain	0.938	0.153	0.974	0.323	0.969	0.136	0.918	0.172
	All Phase	0.747	0.350	0.965	0.265	0.610	0.004	0.761	0.255
	All Gain+Phase	0.937	0.114	0.975	0.340	0.216	0.193	0.605	0.550
	Low Gain	0.707	0.106	0.651	0.208	0.013	0.122	0.553	0.144
	Low Phase	0.367	0.094	0.764	0.191	0.001	0.050	0.528	0.170
	Low Gain+Phase	1.000	0.055	1.000	0.155	0.351	0.008	0.330	0.110
	High Gain	1.000	0.446	0.994	0.001	0.662	0.000	0.561	0.585
	High Phase	0.852	0.002	0.513	0.027	0.360	0.003	0.917	0.312
	High Gain+Phase	1.000	0.492	0.507	0.031	0.258	0.002	0.000	0.000
2.5 cm Probe	All Gain	**0.997**	**0.901**	**1.000**	**0.966**	0.885	0.818	**0.982**	**0.960**
	All Phase	**0.979**	**0.945**	**0.996**	**0.910**	0.907	0.710	0.992	0.893
	All Gain+Phase	**1.000**	**0.923**	0.999	**0.970**	0.934	0.848	**0.976**	**0.941**
	Low Gain	0.983	0.831	**0.999**	**0.965**	0.336	0.007	0.913	0.867
	Low Phase	**0.862**	**0.979**	0.997	0.869	0.262	0.039	0.970	0.610
	Low Gain+Phase	**1.000**	**0.921**	0.999	0.906	0.669	0.228	0.901	0.060
	High Gain	1.000	0.855	0.927	0.884	0.772	0.008	0.961	0.893
	High Phase	0.959	0.841	0.980	0.843	0.574	0.391	0.953	0.668
	High Gain+Phase	0.959	0.841	0.917	0.852	0.807	0.776	0.936	0.842
5 cm Probe	All Gain	**0.991**	**0.970**	**0.995**	**0.978**	0.879	0.858	**0.982**	**0.980**
	All Phase	**0.995**	**0.977**	**0.981**	**0.984**	0.967	0.894	**0.989**	**0.976**
	All Gain+Phase	**0.995**	**0.964**	**0.985**	**0.977**	0.979	0.886	**0.986**	**0.972**
	Low Gain	0.971	0.521	**0.937**	**0.969**	0.838	0.385	0.974	0.195
	Low Phase	**0.952**	**0.905**	**0.924**	**0.964**	0.950	0.852	0.949	0.387
	Low Gain+Phase	1.000	0.034	**0.931**	**0.970**	0.891	0.141	0.957	0.762
	High Gain	0.998	0.591	**0.940**	**0.926**	0.053	0.056	**0.952**	**0.947**
	High Phase	**0.987**	**0.937**	**0.956**	**0.944**	0.900	0.885	**0.978**	**0.941**
	High Gain+Phase	1.000	0.881	**0.952**	**0.940**	0.947	0.882	**0.955**	**0.939**
7.5 cm Probe	All Gain	**0.987**	**0.964**	**1.000**	**0.956**	0.973	0.884	**0.989**	**0.968**
	All Phase	**1.000**	**0.924**	**0.994**	**0.961**	0.971	0.877	**0.987**	**0.949**
	All Gain+Phase	**0.992**	**0.972**	**0.994**	**0.959**	0.934	0.859	**0.985**	**0.968**
	Low Gain	**0.934**	**0.979**	**0.929**	**0.971**	0.383	0.163	**0.938**	**0.951**
	Low Phase	**0.925**	**0.964**	**0.917**	**0.976**	0.920	0.713	**0.927**	**0.978**
	Low Gain+Phase	**0.934**	**0.979**	**0.923**	**0.975**	0.619	0.386	**0.938**	**0.951**
	High Gain	0.987	0.722	**0.999**	**0.961**	0.872	0.660	**0.967**	**0.947**
	High Phase	**0.979**	**0.934**	**0.945**	**0.926**	0.992	0.832	**0.962**	**0.958**
	High Gain+Phase	0.986	0.879	**0.943**	**0.922**	0.932	0.828	**0.962**	**0.960**
Impedance Meter	All Real	1.000	0.692	0.995	0.827	0.906	0.773	**0.975**	**0.952**
	All Imaginary	**1.000**	**0.945**	**0.991**	**0.979**	**0.989**	**0.985**	**0.979**	**0.977**
	All Real+Imaginary	**1.000**	**0.969**	**0.991**	**0.956**	0.976	0.884	**0.981**	**0.985**
	Low Real	0.497	0.066	0.487	0.061	0.019	0.335	0.857	0.632
	Low Imaginary	**0.951**	**0.953**	**0.936**	**0.988**	**0.940**	**0.934**	**0.965**	**0.966**
	Low Real+Imaginary	**1.000**	**0.961**	**0.956**	**0.959**	**0.971**	**0.939**	0.993	0.823
	High Real	0.970	0.763	0.987	0.814	0.987	0.796	**1.000**	**0.914**
	High Imaginary	**1.000**	**0.901**	**1.000**	**0.941**	**0.941**	**0.934**	**0.988**	**0.900**
	High Real+Imaginary	0.987	0.885	**0.956**	**0.951**	**0.989**	**0.936**	**0.969**	**0.953**

Table 5.2: RMSE values for predicting concentration of 1 L nitrate samples (µmol/L)

		SWR		PLS		ANN		Wavelet	
		Training	Validation	Training	Validation	Training	Validation	Training	Validation
2 cm Probe	All Gain	76.39	392.53	49.24	353.73	56.13	409.51	87.61	414.12
	All Phase	154.08	289.23	57.55	394.38	195.46	436.27	149.47	311.22
	All Gain+Phase	76.66	494.84	48.60	351.66	338.30	402.24	192.32	240.21
	Low Gain	41.08	208.44	44.89	192.93	95.25	112.03	50.79	192.29
	Low Phase	60.44	136.17	36.92	167.55	102.38	91.69	52.19	166.87
	Low Gain+Phase	0.00	325.00	0.14	271.53	65.34	103.35	62.18	98.66
	High Gain	0.00	230.70	18.65	340.71	159.09	334.32	162.22	180.77
	High Phase	94.29	362.97	170.99	324.29	207.58	314.98	70.52	264.20
	High Gain+Phase	0.00	195.44	172.00	324.67	384.66	318.19	701.43	804.98
2.5 cm Probe	All Gain	15.42	115.07	0.05	79.07	114.74	150.96	40.69	100.93
	All Phase	44.38	103.64	18.64	110.42	101.60	229.59	27.35	107.05
	All Gain+Phase	0.00	111.60	9.55	65.27	82.08	150.78	47.78	91.36
	Low Gain	9.97	70.58	2.62	49.64	63.37	105.84	22.38	58.68
	Low Phase	28.23	57.99	4.43	50.41	72.17	105.29	13.12	57.40
	Low Gain+Phase	0.00	51.75	2.71	49.74	44.21	74.43	23.92	113.28
	High Gain	0.81	136.71	66.33	86.11	128.71	352.89	48.63	85.54
	High Phase	49.43	114.20	34.42	100.59	215.07	235.09	53.38	145.80
	High Gain+Phase	49.43	114.20	70.51	102.70	108.98	154.06	62.04	102.68
5 cm Probe	All Gain	28.39	70.80	22.36	54.09	131.33	134.42	40.89	68.66
	All Phase	21.85	59.30	42.15	64.55	66.53	129.17	31.92	41.75
	All Gain+Phase	21.21	92.72	36.89	64.36	46.49	134.63	36.02	71.74
	Low Gain	12.87	109.65	19.00	77.22	35.68	69.00	12.28	134.47
	Low Phase	16.60	88.15	20.94	68.58	17.63	30.67	17.11	155.29
	Low Gain+Phase	0.00	234.74	20.01	71.42	28.22	104.51	15.68	124.05
	High Gain	10.94	181.45	60.18	75.97	269.25	308.52	53.93	61.63
	High Phase	28.38	71.56	51.54	60.86	80.27	112.52	35.93	60.07
	High Gain+Phase	0.00	125.29	53.68	64.49	62.16	147.66	51.81	62.42
7.5 cm Probe	All Gain	34.36	76.08	1.26	93.88	51.04	132.27	32.68	85.39
	All Phase	0.54	102.30	23.37	85.28	53.90	140.01	34.47	69.88
	All Gain+Phase	26.98	76.65	24.34	85.70	81.39	133.28	37.42	65.94
	Low Gain	19.57	76.26	20.19	73.50	68.10	99.06	18.98	86.42
	Low Phase	20.81	70.80	21.85	71.36	24.16	61.82	20.55	81.76
	Low Gain+Phase	19.57	76.26	21.03	72.20	49.12	63.67	18.98	86.42
	High Gain	28.31	137.65	7.18	50.49	98.73	204.01	44.36	67.17
	High Phase	35.33	70.66	57.44	72.23	22.61	186.95	48.02	51.09
	High Gain+Phase	28.95	93.56	58.63	74.92	67.17	203.79	47.60	51.23
Impedance Meter	All Real	0.98	201.81	21.19	158.38	98.35	205.14	48.33	79.03
	All Imaginary	0.00	91.36	28.22	59.92	35.18	57.97	44.37	43.68
	All Real+Imaginary	0.00	84.55	29.29	127.16	49.93	129.07	42.50	48.99
	Low Real	53.86	184.65	54.38	171.68	123.50	140.49	28.75	89.02
	Low Imaginary	16.85	28.32	19.23	57.80	19.65	25.72	14.29	67.27
	Low Real+Imaginary	0.00	24.59	15.84	41.69	13.93	22.00	6.19	36.92
	High Real	42.51	130.53	27.68	112.75	38.05	200.99	0.01	86.28
	High Imaginary	1.49	108.30	2.04	73.72	71.76	83.05	26.81	81.01
	High Real+Imaginary	27.58	90.94	51.41	54.37	27.90	87.49	43.17	57.94

5.1.1.2 Thirty-six Liter Samples

Overall, the predictions were more accurate with the 36 L samples than with the 1 L samples. Table 5.3 shows the R^2-values for predicting the concentration of the 36 L nitrate salts, and Table 5.4 shows the RMSE values. The highest training R^2-value for the 36 L tests were 1.00, and the highest validation R^2-value was 0.999. The lowest training and validation RMSE values for these tests were 0.00 µmol/L.

5.1.1.2.1 Comparison of Probes

The 7.5 cm probe performed the best in the 36 L nitrate concentration tests. All but two of the training R^2-values for this probe were at or above 0.900, and many of these values were 1.00. The average validation R^2-value for this probe was 0.899, and the validation R^2-values were 0.900 or above for 70% of the regression models.

The second best performing probe was the 2.5 cm probe. The 2.5 cm probe and the 7.5 cm probe had similar RMSE values; the training RMSE values tended to be a little bit lower for the 2.5 cm probe than the 7.5 cm probe. The training R^2-values for the 2.5 cm probe were above 0.900 for 83% of the regression models, and the validation R^2-values were above 0.900 for 58% of the regression models.

Once again, the 2 cm probe provided the least reliable results. The training R^2-value for this probe was only above 0.900 half of the time, and the validation R^2-values were only above 0.900 for 11% of the regression models. This probe also had the highest RMSE values of the three probes for both training and validation.

5.1.1.2.2 Comparison of Regression Techniques

PLS was the most effective regression method for predicting the concentration of nitrate salts in water for the 36 Lsamples. This method had the lowest RMSE values, overall, and the highest R^2-values. The training R^2-values for PLS were above 0.900 for 100% of the data, and the validation R^2-values were above 0.900 for 78% of the data. The average training R^2-value for PLS was 0.993, and the average validation R^2-value was 0.909.

Stepwise regression was also a promising method for predicting nitrate concentration. The RMSE values for stepwise were only a little bit higher than the values for PLS. The training R^2-values for stepwise were above 0.900 85% of the time and the validation R^2-values were above 0.900 52% of the time.

Neural network and wavelet were the two least reliable methods. The RMSE values for these methods were, on average, much higher than the RMSE values for PLS and stepwise. The R^2-values for these methods were not very high. The average training R^2-values for neural network and wavelet were 0.813 and 0.744, respectively. The average validation R^2-values for these methods were both around 0.600, and neither method had very many R^2-values above 0.900.

Table 5.3: R^2-values for predicting concentration of 36 L nitrate samples

		SWR		PLS		ANN		Wavelet	
		Training	Validation	Training	Validation	Training	Validation	Training	Validation
2 cm Probe	All Gain	0.997	0.743	1.000	0.825	0.931	0.887	0.969	0.839
	All Phase	0.994	0.880	**0.999**	**0.909**	0.854	0.854	0.161	0.185
	All Gain+Phase	0.894	0.686	1.000	0.881	0.971	0.728	0.130	0.165
	Low Gain	0.555	0.055	1.000	0.665	0.813	0.705	0.000	0.000
	Low Phase	0.784	0.412	0.997	0.641	0.407	0.574	0.465	0.236
	Low Gain+Phase	0.963	0.684	0.999	0.626	1.000	0.126	0.397	0.185
	High Gain	0.976	0.574	**0.999**	**0.944**	0.745	0.365	0.508	0.483
	High Phase	0.707	0.241	**0.999**	**0.945**	0.479	0.426	0.000	0.000
	High Gain+Phase	0.945	0.661	**1.000**	**0.915**	0.615	0.829	0.000	0.000
2.5 cm Probe	All Gain	**1.000**	**0.954**	**0.998**	**0.958**	0.664	0.397	**0.997**	**0.973**
	All Phase	**1.000**	**0.967**	**1.000**	**0.964**	**0.929**	**0.937**	**0.993**	**0.941**
	All Gain+Phase	**1.000**	**0.987**	**0.999**	**0.965**	**0.957**	**0.920**	**0.995**	**0.955**
	Low Gain	0.956	0.689	**0.972**	**0.926**	0.522	0.713	0.824	0.889
	Low Phase	0.999	0.656	0.939	0.762	1.000	0.145	0.969	0.871
	Low Gain+Phase	0.992	0.554	**0.915**	**0.914**	0.002	0.174	0.901	0.803
	High Gain	**1.000**	**0.973**	**1.000**	**0.937**	0.898	0.702	**0.999**	**0.934**
	High Phase	**1.000**	**0.951**	**1.000**	**0.949**	0.831	0.740	1.000	0.863
	High Gain+Phase	**1.000**	**0.977**	**1.000**	**0.950**	0.984	0.779	**0.997**	**0.971**
7.5 cm Probe	All Gain	**1.000**	**0.992**	**1.000**	**0.995**	0.985	0.880	**0.997**	**0.993**
	All Phase	**1.000**	**0.995**	**1.000**	**0.996**	**0.980**	**0.960**	**0.997**	**0.995**
	All Gain+Phase	**1.000**	**0.996**	**1.000**	**0.999**	**0.920**	**0.920**	**0.999**	**0.994**
	Low Gain	1.000	0.783	**1.000**	**0.968**	0.574	0.206	0.966	0.878
	Low Phase	**0.999**	**0.921**	**1.000**	**0.943**	1.000	0.702	0.952	0.610
	Low Gain+Phase	**1.000**	**0.923**	**1.000**	**0.975**	1.000	0.587	0.864	0.785
	High Gain	**1.000**	**0.985**	**1.000**	**0.996**	0.984	0.815	**1.000**	**0.994**
	High Phase	**1.000**	**0.993**	**1.000**	**0.997**	0.937	0.792	**1.000**	**0.970**
	High Gain+Phase	**1.000**	**0.994**	**1.000**	**0.999**	0.964	0.829	**1.000**	**0.999**

Table 5.4: RMSE values for predicting concentration of 36 L nitrate samples (µmol/L)

		SWR		PLS		ANN		Wavelet	
		Training	Validation	Training	Validation	Training	Validation	Training	Validation
2 cm Probe	All Gain	180.38	245.65	177.56	202.45	163.19	228.94	316.86	311.67
	All Phase	150.33	200.16	164.53	219.30	190.45	180.88	250.29	356.11
	All Gain+Phase	149.70	242.70	164.19	188.61	206.54	237.55	305.19	319.92
	Low Gain	59.32	62.47	3.90	44.87	60.21	71.59	131.56	131.56
	Low Phase	59.76	56.21	0.81	37.41	37.04	46.11	61.26	74.03
	Low Gain+Phase	5.37	96.85	3.22	38.39	83.60	97.48	62.86	59.05
	High Gain	269.52	296.73	240.83	241.67	266.00	272.19	289.46	294.71
	High Phase	257.96	277.99	240.29	228.40	404.57	496.57	295.36	266.87
	High Gain+Phase	259.18	262.02	240.88	237.28	299.21	325.62	701.43	804.98
2.5 cm Probe	All Gain	151.03	164.45	151.80	172.08	165.27	185.19	155.54	186.58
	All Phase	151.60	167.06	152.37	171.13	203.17	172.55	163.97	209.73
	All Gain+Phase	152.32	171.58	152.28	180.47	155.55	138.29	155.19	168.19
	Low Gain	11.33	36.96	13.45	22.97	26.48	58.40	24.33	30.92
	Low Phase	9.06	22.00	17.46	27.88	16.60	44.29	12.95	30.20
	Low Gain+Phase	0.01	22.89	13.86	23.37	35.45	80.25	9.29	19.00
	High Gain	240.83	270.45	240.76	287.76	239.94	349.77	242.88	274.55
	High Phase	240.83	351.08	240.33	278.73	269.61	318.28	245.13	293.76
	High Gain+Phase	240.83	275.66	239.57	287.26	238.09	284.44	238.19	272.19
7.5 cm Probe	All Gain	152.32	153.32	152.44	161.59	140.40	167.50	149.63	152.13
	All Phase	152.13	168.85	152.30	171.27	556.19	500.55	149.43	200.86
	All Gain+Phase	152.32	172.73	151.54	169.69	149.64	196.07	147.60	165.20
	Low Gain	7.09	32.48	0.29	13.10	26.94	50.70	16.49	18.95
	Low Phase	4.73	20.40	0.26	23.50	36.03	42.93	12.16	33.02
	Low Gain+Phase	1.16	24.58	1.68	17.00	100.38	99.30	15.10	18.57
	High Gain	240.83	245.65	240.50	276.12	240.37	393.87	238.20	265.76
	High Phase	240.83	262.14	240.83	302.11	238.41	205.27	243.00	315.64
	High Gain+Phase	240.83	258.64	240.78	280.16	242.75	231.82	241.23	272.96

5.1.1.3 Comparison of Sample Size

Overall, the concentration estimations were better for the 36 L samples than for the 1 L samples. The validation R^2-value was higher for the 36 L samples 75% of the time, and the validation RMSE value was lower for the 36 L samples 50% of the time. Table 5.5 shows a comparison of the R^2-values for each sample size, and Table 5.6 shows a comparison of the RMSE values for each sample size. For these tables, the column with the higher validation R^2-value or the lower validation RMSE value is highlighted.

The higher R^2-values for the 36 L samples might be partially due to the fact that the tests were done in order of increasing concentration for each salt, and not randomly as the 1 L tests were done. The differences could also be explained by the different control boxes used. The 1 L tests for the 2 cm probe were done with the new control box; all other 1 L tests were done with the old control box. All of the 36 L tests were done using the new control box, and the wider frequency

range that the new control box measured may account for the higher R^2-values for the 36 L tests. The results for the 2 cm probe, which used the same control box to measure both sample sizes, were better for the larger sample size. This suggests that the control box was not the only factor contributing to the higher R^2-values for the 36 L samples, but for the tests performed for this thesis, it cannot be concluded that the larger sample size was responsible for the better results for the 36 L tests.

Table 5.5: Comparison of 36 L and 1 L R^2-values for predicting nitrate concentration

			Stepwise Regression				PLS				Neural Network				Wavelet			
			36 L		1 L		36 L		1 L		36 L		1 L		36 L		1 L	
			Train	Valid.	Train	Valid.	Train	Valid.	Train	Valid.	Train	Valid.	Train	Valid.	Train	Valid.	Train	Valid.
2 cm Probe	All	Gain	1.00	0.74	0.94	0.15	1.00	0.82	0.97	0.32	0.93	0.89	0.97	0.14	0.97	0.84	0.92	0.17
		Phase	0.99	0.88	0.75	0.35	1.00	0.91	0.96	0.26	0.85	0.85	0.61	0.00	0.16	0.19	0.76	0.26
		G+P	0.89	0.69	0.94	0.11	1.00	0.88	0.97	0.34	0.97	0.73	0.22	0.19	0.13	0.16	0.61	0.55
	Low	Gain	0.56	0.05	0.71	0.11	1.00	0.67	0.65	0.21	0.81	0.71	0.01	0.12	0.00	0.00	0.55	0.14
		Phase	0.78	0.41	0.37	0.09	1.00	0.64	0.76	0.19	0.41	0.57	0.00	0.05	0.47	0.24	0.53	0.17
		G+P	0.96	0.68	1.00	0.05	1.00	0.63	1.00	0.15	1.00	0.13	0.35	0.01	0.40	0.19	0.33	0.11
	High	Gain	0.98	0.57	1.00	0.45	1.00	0.94	0.99	0.00	0.75	0.36	0.66	0.00	0.51	0.48	0.56	0.59
		Phase	0.71	0.24	0.85	0.00	1.00	0.94	0.51	0.03	0.48	0.43	0.36	0.00	0.00	0.00	0.92	0.31
		G+P	0.94	0.66	1.00	0.49	1.00	0.91	0.51	0.03	0.62	0.83	0.26	0.00	0.00	0.00	0.00	0.00
2.5 cm Probe	All	Gain	1.00	0.95	1.00	0.90	1.00	0.96	1.00	0.97	0.66	0.40	0.88	0.82	1.00	0.97	0.98	0.96
		Phase	1.00	0.97	0.98	0.94	1.00	0.96	1.00	0.91	0.93	0.94	0.91	0.71	0.99	0.94	0.99	0.89
		G+P	1.00	0.99	1.00	0.92	1.00	0.96	1.00	0.97	0.96	0.92	0.93	0.85	1.00	0.96	0.98	0.94
	Low	Gain	0.96	0.69	0.98	0.83	0.97	0.93	1.00	0.96	0.52	0.71	0.34	0.01	0.82	0.89	0.91	0.87
		Phase	1.00	0.66	0.86	0.98	0.94	0.76	1.00	0.87	1.00	0.15	0.26	0.04	0.97	0.87	0.97	0.61
		G+P	0.99	0.55	1.00	0.92	0.91	0.91	1.00	0.91	0.00	0.17	0.67	0.23	0.90	0.80	0.90	0.06
	High	Gain	1.00	0.97	1.00	0.86	1.00	0.94	0.93	0.88	0.90	0.70	0.77	0.01	1.00	0.93	0.96	0.89
		Phase	1.00	0.95	0.96	0.84	1.00	0.95	0.98	0.84	0.83	0.74	0.57	0.39	1.00	0.86	0.95	0.67
		G+P	1.00	0.98	0.96	0.84	1.00	0.95	0.92	0.85	0.98	0.78	0.81	0.78	1.00	0.97	0.94	0.84
7.5 cm Probe	All	Gain	1.00	0.99	0.99	0.96	1.00	0.99	1.00	0.96	0.99	0.88	0.97	0.88	1.00	0.99	0.99	0.97
		Phase	1.00	0.99	1.00	0.92	1.00	1.00	0.99	0.96	0.98	0.96	0.97	0.88	1.00	1.00	0.99	0.95
		G+P	1.00	1.00	0.99	0.97	1.00	1.00	0.99	0.96	0.92	0.92	0.93	0.86	1.00	0.99	0.99	0.97
	Low	Gain	1.00	0.78	0.93	0.98	1.00	0.97	0.93	0.97	0.57	0.21	0.38	0.16	0.97	0.88	0.94	0.95
		Phase	1.00	0.92	0.92	0.96	1.00	0.94	0.92	0.98	1.00	0.70	0.92	0.71	0.95	0.61	0.93	0.98
		G+P	1.00	0.92	0.93	0.98	1.00	0.98	0.92	0.98	1.00	0.59	0.62	0.39	0.86	0.78	0.94	0.95
	High	Gain	1.00	0.98	0.99	0.72	1.00	1.00	1.00	0.96	0.98	0.82	0.87	0.66	1.00	0.99	0.97	0.95
		Phase	1.00	0.99	0.98	0.93	1.00	1.00	0.95	0.93	0.94	0.79	0.99	0.83	1.00	0.97	0.96	0.96
		G+P	1.00	0.99	0.99	0.88	1.00	1.00	0.94	0.92	0.96	0.83	0.93	0.83	1.00	1.00	0.96	0.96

Table 5.6: Comparison of 36 L and 1 L RMSE values for predicting nitrate concentration (µmol/L)

			Stepwise Regression				PLS				Neural Network				Wavelet			
			36 L		1 L		36 L		1 L		36 L		1 L		36 L		1 L	
			Train	Valid.	Train	Valid.	Train	Valid.	Train	Valid.	Train	Valid.	Train	Valid.	Train	Valid.	Train	Valid.
2 cm Probe	All	Gain	180	246	76	393	178	202	49	354	163	229	56	410	317	312	88	414
		Phase	150	200	154	289	165	219	58	394	190	181	195	436	250	356	149	311
		G+P	150	243	77	495	164	189	49	352	207	238	338	402	305	320	192	240
	Low	Gain	59	62	41	208	4	45	45	193	60	72	95	112	132	132	51	192
		Phase	60	56	60	136	1	37	37	168	37	46	102	92	61	74	52	167
		G+P	5	97	0	325	3	38	0	272	84	97	65	103	63	59	62	99
	High	Gain	270	297	0	231	241	242	19	341	266	272	159	334	289	295	162	181
		Phase	258	278	94	363	240	228	171	324	405	497	208	315	295	267	71	264
		G+P	259	262	0	195	241	237	172	325	299	326	385	318	701	805	701	805
2.5 cm Probe	All	Gain	151	164	15	115	152	172	0	79	165	185	115	151	156	187	41	101
		Phase	152	167	44	104	152	171	19	110	203	173	102	230	164	210	27	107
		G+P	152	172	0	112	152	180	10	65	156	138	82	151	155	168	48	91
	Low	Gain	11	37	10	71	13	23	3	50	26	58	63	106	24	31	22	59
		Phase	9	22	28	58	17	28	4	50	17	44	72	105	13	30	13	57
		G+P	0	23	0	52	14	23	3	50	35	80	44	74	9	19	24	113
	High	Gain	241	270	1	137	241	288	66	86	240	350	129	353	243	275	49	86
		Phase	241	351	49	114	240	279	34	101	270	318	215	235	245	294	53	146
		G+P	241	276	49	114	240	287	71	103	238	284	109	154	238	272	62	103
7.5 cm Probe	All	Gain	152	153	34	76	152	162	1	94	140	168	51	132	150	152	33	85
		Phase	152	169	1	102	152	171	23	85	556	501	54	140	149	201	34	70
		G+P	152	173	27	77	152	170	24	86	150	196	81	133	148	165	37	66
	Low	Gain	7	32	20	76	0	13	20	74	27	51	68	99	16	19	19	86
		Phase	5	20	21	71	0	23	22	71	36	43	24	62	12	33	21	82
		G+P	1	25	20	76	2	17	21	72	100	99	49	64	15	19	19	86
	High	Gain	241	246	28	138	241	276	7	50	240	394	99	204	238	266	44	67
		Phase	241	262	35	71	241	302	57	72	238	205	23	187	243	316	48	51
		G+P	241	259	29	94	241	280	59	75	243	232	67	204	241	273	48	51

5.1.2 Molecular Weight

The results of the tests showed that the probes and the impedance meter had little ability to distinguish between the molecular weights of the three nitrates for the 1 L tests. This could have been caused by the boundary effect. Many of the validation R^2-values for these tests were below 0.100. The results were better for the 36 L tests. There were many training R^2-values of 1.00, and the highest validation R^2-value was 0.999.

5.1.2.1 One Liter Samples

The tests done to predict the molecular weight of the nitrate salts with the 1 L samples did not have good results. The R^2-values were always low for the validation data, and were usually low for training data. There was not a lot of difference in prediction ability between the different probes or the different regression methods.

Table 5.7 shows the R^2-values for predicting the molecular weight of the 1 L nitrate salts, and Table 5.8 shows the RMSE values. The highest validation R^2-value for the 1 L samples was 0.413, which came from the 2.5 cm probe and neural network. The lowest validation R^2-values were around zero for each probe and each regression technique. The RMSE values were relatively high for all of the probes and regression methods, most falling between 70.0 g/mol and 100.0 g/mol.

Stepwise was unable to create predictions for several of the datasets because there were no variables in the data significant enough to enter the model. This happened for both the traditional stepwise regression and the stepwise regression on the wavelet transformed data. There were a few situations where stepwise found a small number of significant data points, usually around five, and built a model with good training R^2-values, but the validation R^2-values for these models were always very low, usually below 0.2.

Table 5.7: R^2-values for predicting molecular weight of 1 L nitrate samples

		SWR		PLS		ANN		Wavelet	
		Training	Validation	Training	Validation	Training	Validation	Training	Validation
2 cm Probe	All Gain			0.267	0.009	0.008	0.134		
	All Phase			0.027	0.000	0.345	0.003		
	All Gain+Phase			0.039	0.000	0.000	0.038		
	Low Gain			0.163	0.007	0.462	0.089	0.927	0.086
	Low Phase			0.246	0.059	0.001	0.037		
	Low Gain+Phase			1.000	0.002	0.000	0.055		
	High Gain	1.000	0.082	0.727	0.083	0.962	0.183		
	High Phase	0.983	0.360	0.684	0.037	0.333	0.031	0.544	0.169
	High Gain+Phase	1.000	0.016	0.702	0.048	0.653	0.228	0.451	0.116
2.5 cm Probe	All Gain	0.154	0.007	0.014	0.006	0.028	0.005		
	All Phase	0.686	0.133	0.015	0.005	0.059	0.000		
	All Gain+Phase	0.511	0.191	0.016	0.006	0.307	0.010		
	Low Gain			0.038	0.039	0.082	0.041		
	Low Phase	0.995	0.018	0.048	0.005	0.033	0.064		
	Low Gain+Phase	0.966	0.001	0.056	0.010	0.103	0.008		
	High Gain			0.039	0.026	0.000	0.000		
	High Phase			0.017	0.032	0.007	0.031		
	High Gain+Phase			0.028	0.030	**0.024**	**0.413**		
5 cm Probe	All Gain			0.011	0.005	0.298	0.121		
	All Phase			0.012	0.010	0.524	0.229		
	All Gain+Phase			0.012	0.008	0.058	0.008		
	Low Gain	0.748	0.063	0.045	0.000	0.066	0.158	0.417	0.028
	Low Phase	0.764	0.000	0.061	0.007	0.105	0.006	0.450	0.027
	Low Gain+Phase	1.000	0.144	0.056	0.003	0.170	0.089	0.384	0.098
	High Gain	0.940	0.045	0.194	0.206	0.095	0.115		
	High Phase	1.000	0.234	0.186	0.253	0.432	0.014		
	High Gain+Phase	1.000	0.156	0.189	0.241	0.052	0.168		
7.5 cm Probe	All Gain			0.009	0.003	0.015	0.030		
	All Phase			0.009	0.004	0.027	0.008		
	All Gain+Phase			0.009	0.004	0.000	0.002		
	Low Gain			0.023	0.006	0.004	0.019		
	Low Phase			0.036	0.005	0.085	0.000	0.492	0.063
	Low Gain+Phase			0.031	0.005	0.124	0.112		
	High Gain			0.036	0.039	0.019	0.002		
	High Phase			0.032	0.035	0.004	0.002		
	High Gain+Phase			0.033	0.036	0.018	0.124		
Impedance Meter	All Real			0.037	0.001	0.269	0.011		
	All Imaginary			0.007	0.002	0.005	0.002		
	All Real+Imaginary			0.008	0.002	0.270	0.025		
	Low Real	1.000	0.089	0.990	0.055	0.080	0.001	0.683	0.010
	Low Imaginary			0.023	0.004	0.000	0.004		
	Low Real+Imaginary	1.000	0.173	0.977	0.124	0.614	0.000		
	High Real			0.936	0.235	0.201	0.164	0.978	0.020
	High Imaginary			0.997	0.071	0.014	0.003		
	High Real+Imaginary			0.364	0.003	0.188	0.136		

Empty cells indicate that stepwise did not find any data points that met the specified level of significance and no model was made.

Table 5.8: RMSE values for predicting molecular weight of 1 L nitrate samples (g/mol)

		SWR		PLS		ANN		Wavelet	
		Training	Validation	Training	Validation	Training	Validation	Training	Validation
2 cm Probe	All Gain			61.52	95.72	80.42	92.40		
	All Phase			70.88	72.99	58.73	80.32		
	All Gain+Phase			70.46	73.24	94.88	102.70		
	Low Gain			67.95	93.77	55.21	101.86	20.11	355.21
	Low Phase			64.52	93.29	92.72	103.36		
	Low Gain+Phase			0.86	274.19	83.73	94.17		
	High Gain	0.00	167.32	39.61	86.43	16.80	74.04		
	High Phase	9.99	82.69	42.57	91.63	66.19	93.48	51.15	82.70
	High Gain+Phase	0.00	139.92	41.38	88.50	47.56	78.75	56.15	90.28
2.5 cm Probe	All Gain	66.12	75.70	71.39	71.81	73.71	89.10		
	All Phase	40.30	96.65	71.34	71.94	57.22	122.53		
	All Gain+Phase	50.27	77.75	71.32	71.91	60.57	88.03		
	Low Gain			72.88	73.31	84.39	89.80		
	Low Phase	5.04	121.78	72.47	76.49	95.54	99.30		
	Low Gain+Phase	13.68	122.73	72.17	75.95	85.38	106.07		
	High Gain			74.28	75.24	73.81	102.66		
	High Phase			75.13	74.97	114.32	112.95		
	High Gain+Phase			74.70	74.91	102.13	88.38		
5 cm Probe	All Gain			71.49	71.84	73.04	70.75		
	All Phase			71.44	71.56	50.38	69.01		
	All Gain+Phase			71.45	71.65	79.70	100.50		
	Low Gain	37.32	226.36	72.59	77.43	81.35	77.84	56.73	237.06
	Low Phase	36.07	205.79	71.99	76.24	75.37	104.90	55.10	182.71
	Low Gain+Phase	0.00	420.34	72.18	76.74	86.95	113.20	58.33	113.74
	High Gain	18.61	116.76	68.03	68.57	77.42	75.60		
	High Phase	0.00	181.25	68.34	67.40	69.39	107.30		
	High Gain+Phase	0.00	154.96	68.24	67.68	96.13	98.35		
7.5 cm Probe	All Gain			71.54	71.91	86.37	98.73		
	All Phase			71.56	71.84	85.36	79.84		
	All Gain+Phase			71.55	71.86	92.03	106.75		
	Low Gain			73.43	75.05	89.18	90.20		
	Low Phase			72.95	75.69	72.73	81.53	52.94	89.31
	Low Gain+Phase			73.13	75.46	75.05	101.51		
	High Gain			74.41	74.57	90.51	99.69		
	High Phase			74.56	74.70	92.56	100.23		
	High Gain+Phase			74.51	74.66	85.96	85.06		
Impedance Meter	All Real			70.52	75.20	64.06	77.19		
	All Imaginary			71.62	71.93	83.79	83.57		
	All Real+Imaginary			71.59	71.98	62.13	76.08		
	Low Real	0.00	119.64	7.33	210.68	116.39	117.25	41.84	126.23
	Low Imaginary			73.42	74.94	79.52	78.98		
	Low Real+Imaginary	0.00	83.71	11.30	463.28	46.74	86.91		
	High Real			19.15	80.46	76.71	102.62	11.22	126.51
	High Imaginary			4.09	98.74	95.54	94.04		
	High Real+Imaginary			60.43	84.12	88.42	103.25		

Empty cells indicate that stepwise did not find any data points that met the specified level of significance and no model was made.

5.1.2.2 Thirty-six Liter Samples

The molecular weight predictions were much better for the 36 L samples than for the 1 L samples. This might be because there was more sample around between the probe and the sides of the container, reducing the effect that the container had on the electric field. There were several instances where the training R^2-value was 1.00 and the validation R^2-value was 0.90 or higher. Table 5.9 shows the R^2-values for the 36 L nitrate molecular weight tests, and Table 5.10 shows the RMSE values.

5.1.2.2.3 Comparison of Probes

The 2 cm probe and the 7.5 cm probe did the best job of predicting molecular weight for the 36 L samples. Both probes had average training R^2-values above 0.9 and average validation R^2-values above 0.85. Overall, the R^2-values tended to be a little bit higher for the 7.5 cm probe. The 7.5 cm probe had training R^2-values above 0.900 for 89% of the regression models, and validation R^2-values above 0.900 for 75% of the regression models. The RMSE values were also lower for the 7.5 cm probe for both training and validation.

The 2.5 cm probe did not predict molecular weight as well. The training and validation R^2-values were above 0.900 for less than 30% of the datasets. The average training R^2-value for this probe was only 0.500, and the average validation R^2-value was 0.454. Despite the low R^2-values for this probe, its RMSE values were comparatively good. This probe had the lowest average validation RMSE of any of the probes.

5.1.2.2.4 Comparison of Regression Methods

The highest R^2-values for predicting the molecular weight of nitrate came from PLS and stepwise regression. Both methods had average training R^2-values around 0.85, and the average validation R^2-value for PLS was 0.807, about 0.017 higher than for stepwise. The training R^2-value was above 0.900 for 78% of the datasets used for each of these methods. The validation R^2-values were above 0.900 for 70% and 60% of the datasets for PLS and stepwise, respectively. The RMSE values were also very similar for these two methods, averaging around 55 g/mol for training and 65 g/mol for validation.

Neural network and wavelet did not perform as well for predicting the molecular weight of nitrate. Both methods had training RMSE values that averaged over 60 g/mol, and validation RMSE values that averaged over 70 g/mol. Neural network did the worst overall. The training R^2-value for this method was over 0.900 for only 44% of the tests, and the validation R^2-value was over 0.900 30% of the time.

Table 5.9: R^2-values for predicting molecular weight of 36 L nitrate samples

		SWR		PLS		ANN		Wavelet	
		Training	Validation	Training	Validation	Training	Validation	Training	Validation
2 cm Probe	All Gain	**0.972**	**0.910**	**0.979**	**0.951**	0.803	0.874	**0.940**	**0.917**
	All Phase	**0.999**	**0.910**	**1.000**	**0.962**	0.129	0.129	0.932	0.822
	All Gain+Phase	**0.990**	**0.935**	**0.987**	**0.964**	**1.000**	**1.000**	0.917	0.889
	Low Gain	0.932	0.899	**0.999**	**0.944**	0.816	0.761	0.827	0.757
	Low Phase	**0.998**	**0.933**	0.943	0.795	**0.993**	**0.913**	0.708	0.538
	Low Gain+Phase	0.993	0.875	0.973	0.767	0.969	0.849	0.927	0.687
	High Gain	**0.997**	**0.974**	**0.999**	**0.984**	**1.000**	**0.998**	**0.991**	**0.971**
	High Phase	0.955	0.863	**1.000**	**0.963**	**1.000**	**1.000**	0.859	0.784
	High Gain+Phase	**0.995**	**0.980**	**1.000**	**0.979**	**1.000**	**1.000**	0.624	0.705
2.5 cm Probe	All Gain	0.375	0.124	0.329	0.236	0.338	0.233	0.255	0.160
	All Phase	0.298	0.282	0.300	0.245	0.001	0.000	0.325	0.259
	All Gain+Phase	0.375	0.124	0.277	0.280	0.038	0.045	0.000	0.000
	Low Gain	**0.995**	**0.926**	**1.000**	**0.940**	0.838	0.818	0.910	0.734
	Low Phase	**1.000**	**0.941**	**1.000**	**0.979**	**0.820**	**0.904**	0.955	0.714
	Low Gain+Phase	1.000	0.875	**1.000**	**0.940**	0.969	0.853	0.942	0.851
	High Gain	0.356	0.356	0.345	0.348	0.230	0.281	0.359	0.359
	High Phase	0.352	0.350	0.324	0.349	0.076	0.246	0.320	0.313
	High Gain+Phase	0.359	0.359	0.351	0.358	0.235	0.235	0.337	0.315
7.5 cm Probe	All Gain	**1.000**	**0.992**	**1.000**	**0.992**	**0.932**	**0.901**	**0.999**	**0.988**
	All Phase	**1.000**	**0.994**	0.999	0.989	0.145	0.092	**0.988**	**0.902**
	All Gain+Phase	**1.000**	**0.989**	**1.000**	**0.998**	0.993	0.884	**0.998**	**0.995**
	Low Gain	0.991	0.878	**1.000**	**0.973**	0.914	0.754	**0.953**	**0.950**
	Low Phase	**0.996**	**0.937**	**1.000**	**0.915**	0.827	0.748	0.974	0.820
	Low Gain+Phase	**1.000**	**0.939**	**1.000**	**0.961**	0.040	0.023	**0.960**	**0.945**
	High Gain	**1.000**	**0.992**	**1.000**	**0.995**	0.997	0.635	**0.999**	**0.998**
	High Phase	**1.000**	**0.994**	**1.000**	**0.990**	0.893	0.689	**0.992**	**0.951**
	High Gain+Phase	**1.000**	**0.998**	**1.000**	**0.999**	0.992	0.873	**1.000**	**0.999**

Table 5.10: RMSE values for predicting molecular weight of 36 L nitrate samples (g/mol)

		SWR		PLS		ANN		Wavelet	
		Training	Validation	Training	Validation	Training	Validation	Training	Validation
2 cm Probe	All Gain	84.99	82.27	83.20	80.45	82.99	83.16	82.92	81.42
	All Phase	82.62	90.03	82.58	86.82	94.52	95.61	81.90	77.52
	All Gain+Phase	84.19	83.10	83.85	81.88	83.49	87.64	86.19	86.62
	Low Gain	19.41	25.31	1.67	17.82	0.65	36.42	30.94	39.96
	Low Phase	3.42	19.64	17.70	34.91	29.81	43.61	40.18	52.69
	Low Gain+Phase	6.40	26.43	12.12	38.41	8.14	17.23	20.12	42.77
	High Gain	129.81	128.71	130.22	129.06	129.59	129.34	131.43	126.16
	High Phase	131.23	125.99	130.55	137.27	129.16	127.30	129.89	129.52
	High Gain+Phase	130.16	136.64	130.29	132.26	130.83	129.71	124.42	121.94
2.5 cm Probe	All Gain	56.95	77.79	69.05	78.86	76.77	85.70	70.97	77.36
	All Phase	53.81	67.21	55.61	63.36	81.78	78.13	43.64	58.04
	All Gain+Phase	56.95	77.79	50.95	54.18	73.78	88.35	161.44	161.44
	Low Gain	5.02	23.37	1.13	18.78	31.69	35.03	22.26	40.60
	Low Phase	1.24	19.07	1.52	10.99	37.09	32.06	15.76	42.65
	Low Gain+Phase	0.00	27.22	0.28	18.50	13.24	29.47	17.86	32.31
	High Gain	79.54	78.15	84.64	82.12	96.03	95.15	79.79	79.46
	High Phase	76.30	75.04	75.43	79.93	102.37	96.34	85.18	87.42
	High Gain+Phase	81.20	79.35	83.70	86.13	117.42	106.65	76.46	69.67
7.5 cm Probe	All Gain	43.45	47.69	49.94	57.27	57.10	65.49	56.02	50.54
	All Phase	56.50	54.52	51.51	62.74	69.79	68.08	52.12	49.92
	All Gain+Phase	43.45	47.69	49.19	59.82	115.72	117.98	57.59	68.70
	Low Gain	0.30	30.96	0.02	29.06	25.93	78.25	26.48	47.38
	Low Phase	0.00	112.56	0.03	28.88	21.41	36.08	12.52	42.50
	Low Gain+Phase	0.00	49.76	0.66	28.85	6.23	44.30	6.41	31.13
	High Gain	78.39	96.78	80.64	97.22	76.36	105.85	79.11	93.41
	High Phase	86.26	85.09	69.00	84.84	78.93	68.66	81.21	91.21
	High Gain+Phase	87.67	84.20	79.57	95.71	106.31	101.35	79.11	93.41

5.1.2.3 Comparison of Sample Sizes

The molecular weight predictions from the 36 L samples were much better than the predictions from the 1 L samples. The R^2-value was higher for the 36 L samples in 98% of the tests, and the RMSE was lower for the 36 L samples in 75% of the tests. Table 5.11 lists the R^2-values for predicting the molecular weight for each sample size, and Table 5.12 lists the RMSE values. For each table, the sample size that gave the highest validation R^2-value or the lowest validation RMSE value is highlighted.

As discussed in Section 5.1.1.3, the procedures for testing the 1 L tests and the 36 L tests were not identical. The nitrate salts in the 1 L tests were measured randomly, whereas the nitrate salts in the 36 L tests were measured in order of increasing concentration. The control boxes that were used were not the same for the different sample sizes. Either one of these factors could

contribute to the improved results for the 36 L samples. The reduced boundary effect with the larger sample size could also explain the improved results.

Table 5.11: Comparison of 36 L and 1 L R^2-values for predicting nitrate molecular weight

			Stepwise Regression				PLS				Neural Network				Wavelet			
			36 L		1 L		36 L		1 L		36 L		1 L		36 L		1 L	
			Train	Valid.	Train	Valid.	Train	Valid.	Train	Valid.	Train	Valid.	Train	Valid.	Train	Valid.	Train	Valid.
2 cm Probe	All	Gain	0.97	0.91			0.98	0.95	0.27	0.01	0.80	0.87	0.01	0.13	0.94	0.92		
		Phase	1.00	0.91			1.00	0.96	0.03	0.00	0.13	0.13	0.35	0.00	0.93	0.82		
		G+P	0.99	0.93			0.99	0.96	0.04	0.00	1.00	1.00	0.00	0.04	0.92	0.89		
	Low	Gain	0.93	0.90			1.00	0.94	0.16	0.01	0.82	0.76	0.46	0.09	0.83	0.76	0.93	0.09
		Phase	1.00	0.93			0.94	0.80	0.25	0.06	0.99	0.91	0.00	0.04	0.71	0.54		
		G+P	0.99	0.87			0.97	0.77	1.00	0.00	0.97	0.85	0.00	0.06	0.93	0.69		
	High	Gain	1.00	0.97	1.00	0.08	1.00	0.98	0.73	0.08	1.00	1.00	0.96	0.18	0.99	0.97		
		Phase	0.95	0.86	0.98	0.36	1.00	0.96	0.68	0.04	1.00	1.00	0.33	0.03	0.86	0.78	0.54	0.17
		G+P	0.99	0.98	1.00	0.02	1.00	0.98	0.70	0.05	1.00	1.00	0.65	0.23	0.62	0.71	0.45	0.12
2.5 cm Probe	All	Gain	0.37	0.12	0.15	0.01	0.33	0.24	0.01	0.01	0.34	0.23	0.03	0.01	0.26	0.16		
		Phase	0.30	0.28	0.69	0.13	0.30	0.24	0.01	0.00	0.00	0.00	0.06	0.00	0.32	0.26		
		G+P	0.37	0.12	0.51	0.19	0.28	0.28	0.02	0.01	0.04	0.04	0.31	0.01	0.00	0.00		
	Low	Gain	1.00	0.93			1.00	0.94	0.04	0.04	0.84	0.82	0.08	0.04	0.91	0.73		
		Phase	1.00	0.94	1.00	0.02	1.00	0.98	0.05	0.00	0.82	0.90	0.03	0.06	0.96	0.71		
		G+P	1.00	0.88	0.97	0.00	1.00	0.94	0.06	0.01	0.97	0.85	0.10	0.01	0.94	0.85		
	High	Gain	0.36	0.36			0.35	0.35	0.04	0.03	0.23	0.28	0.00	0.00	0.36	0.36		
		Phase	0.35	0.35			0.32	0.35	0.02	0.03	0.08	0.25	0.01	0.03	0.32	0.31		
		G+P	0.36	0.36			0.35	0.36	0.03	0.03	0.23	0.24	0.02	0.41	0.34	0.32		
7.5 cm Probe	All	Gain	1.00	0.99			1.00	0.99	0.01	0.00	0.93	0.90	0.02	0.03	1.00	0.99		
		Phase	1.00	0.99			1.00	0.99	0.01	0.00	0.14	0.09	0.03	0.01	0.99	0.90		
		G+P	1.00	0.99			1.00	1.00	0.01	0.00	0.99	0.88	0.00	0.00	1.00	0.99		
	Low	Gain	0.99	0.88			1.00	0.97	0.02	0.01	0.91	0.75	0.00	0.02	0.95	0.95		
		Phase	1.00	0.94			1.00	0.91	0.04	0.00	0.83	0.75	0.09	0.00	0.97	0.82	0.49	0.06
		G+P	1.00	0.94			1.00	0.96	0.03	0.01	0.04	0.02	0.12	0.11	0.96	0.94		
	High	Gain	1.00	0.99			1.00	1.00	0.04	0.04	1.00	0.64	0.02	0.00	1.00	1.00		
		Phase	1.00	0.99			1.00	0.99	0.03	0.04	0.89	0.69	0.00	0.00	0.99	0.95		
		G+P	1.00	1.00			1.00	1.00	0.03	0.04	0.99	0.87	0.02	0.12	1.00	1.00		

Empty cells indicate that stepwise did not find any data points that met the specified level of significance and no model was made.

Table 5.12: Comparison of 36 L and 1 L RMSE values for predicting nitrate molecular weight (g/mol)

| | | | Stepwise Regression | | | | PLS | | | | Neural Network | | | | Wavelet | | | |
| | | | 36 L | | 1 L | | 36 L | | 1 L | | 36 L | | 1 L | | 36 L | | 1 L | |
			Train	Valid.	Train	Valid.	Train	Valid.	Train	Valid.	Train	Valid.	Train	Valid.	Train	Valid.	Train	Valid.
2 cm Probe	All	Gain	85	82			83	80	62	96	83	83	80	92	83	81		
		Phase	83	90			83	87	71	73	95	96	59	80	82	78		
		G+P	84	83			84	82	70	73	83	88	95	103	86	87		
	Low	Gain	19	25			2	18	68	94	1	36	55	102	31	40	20	355
		Phase	3	20			18	35	65	93	30	44	93	103	40	53		
		G+P	6	26			12	38	1	274	8	17	84	94	20	43		
	High	Gain	130	129	0	167	130	129	40	86	130	129	17	74	131	126		
		Phase	131	126	10	83	131	137	43	92	129	127	66	93	130	130	51	83
		G+P	130	137	0	140	130	132	41	89	131	130	48	79	124	122	56	90
2.5 cm Probe	All	Gain	57	78	66	76	69	79	71	72	77	86	74	89	71	77		
		Phase	54	67	40	97	56	63	71	72	82	78	57	123	44	58		
		G+P	57	78	50	78	51	54	71	72	74	88	61	88	161	161		
	Low	Gain	5	23			1	19	73	73	32	35	84	90	22	41		
		Phase	1	19	5	122	2	11	72	76	37	32	96	99	16	43		
		G+P	0	27	14	123	0	18	72	76	13	29	85	106	18	32		
	High	Gain	80	78			85	82	74	75	96	95	74	103	80	79		
		Phase	76	75			75	80	75	75	102	96	114	113	85	87		
		G+P	81	79			84	86	75	75	117	107	102	88	76	70		
7.5 cm Probe	All	Gain	43	48			50	57	72	72	57	65	86	99	56	51		
		Phase	57	55			52	63	72	72	70	68	85	80	52	50		
		G+P	43	48			49	60	72	72	116	118	92	107	58	69		
	Low	Gain	0	31			0	29	73	75	26	78	89	90	26	47		
		Phase	0	113			0	29	73	76	21	36	73	82	13	42	53	89
		G+P	0	50			1	29	73	75	6	44	75	102	6	31		
	High	Gain	78	97			81	97	74	75	76	106	91	100	79	93		
		Phase	86	85			69	85	75	75	79	69	93	100	81	91		
		G+P	88	84			80	96	75	75	106	101	86	85	79	93		

Empty cells indicate that stepwise did not find any data points that met the specified level of significance and no model was made.

5.2 Atrazine and Water

The sensors studied were able to detect concentration changes in atrazine. The highest training R^2-value for predicting atrazine concentration was 1.00, and the highest validation R^2-value was 0.979. The R^2-values for detecting atrazine are shown in Table 5.13, in which the highest validation R^2-value is bolded, and Table 5.15, in which all R^2-values above 0.900 are bolded. The RMSE values are given in Table 5.14 and Table 5.16.

5.2.1 One Liter Samples

Overall, the probes were not able to predict the concentration of atrazine as accurately as they were able to predict the concentration of nitrates. Only stepwise regression had an average training R^2-value above 0.900. There were several high training R^2-values, but almost all of them corresponded to very low validation R^2-values. These R^2-values for predicting atrazine concentration are summarized in Table 5.13, and the RMSE values are summarized in Table 5.14.

5.2.1.1 Comparison of Probes

For the 1 L atrazine samples, none of the probes consistently had higher R^2-values than the others. The 2.5 cm probe had the highest average training R^2-value, but it had one of the lower average validation R^2-values. The 2.5 cm probe also had the highest percentage of training R^2-values above 0.900, but the high training R^2-values are meaningless because the corresponding validation R^2-values were so low, usually below 0.100.

All four probes and the impedance meter had at least one model with a training R^2-value of 1.00, but the highest validation R^2-value for the models with a training R^2-value of 1.00 was 0.214. The highest validation R^2-value, 0.477, was obtained from the 5 cm probe. The lowest validation R^2-value for most of the probes was 0.000.

The training RMSE values were, for the most part, lowest for the 2.5 cm probe, and the validation RMSE values tended to be lowest for the 7.5 cm probe. The RMSE values, both training and validation, were highest for the 2 and 5 cm probes.

5.2.1.2 Comparison of Regression Methods

None of the regression methods provided accurate atrazine predictions. Stepwise regression had the highest training R^2-values, many were 1.00, and the lowest training RMSE values, most were close to 0.00 mg/L, but it had terrible validation values. The average validation R^2-value for stepwise was only 0.079.

The highest validation R^2-values came from PLS, but the average value was only 0.169. Neural network was, once again, the least reliable prediction method. The average training R^2-value for neural network was 0.067. Stepwise regression on the wavelet transformed data was not a reliable prediction method either. For some of the datasets, stepwise was not able to build a model using the wavelet transformed data. When it could build a model, the R^2-values were very low, mostly below 0.100, and the RMSE values were high.

Table 5.13: R^2-values for predicting concentration of 1 L atrazine samples

		SWR		PLS		ANN		Wavelet	
		Training	Validation	Training	Validation	Training	Validation	Training	Validation
2 cm Probe	Gain	0.670	0.007	0.109	0.002	0.378	0.039		
	Phase	1.000	0.018	0.994	0.001	0.109	0.010	1.000	0.047
	Gain+Phase	0.670	0.007	0.995	0.000	0.789	0.008		
2.5 cm Probe	Gain	1.000	0.124	1.000	0.058	0.131	0.063	0.965	0.025
	Phase	1.000	0.058	0.829	0.010	0.291	0.063	1.000	0.000
	Gain+Phase	1.000	0.072	1.000	0.055	0.368	0.006	0.965	0.025
5 cm Probe	Gain	0.974	0.160	0.924	0.202	0.077	0.058	0.713	0.016
	Phase	1.000	0.345	0.288	0.204	0.345	0.003	0.587	0.143
	Gain+Phase	1.000	0.010	0.492	0.338	**0.784**	**0.477**	0.713	0.016
7.5 cm Probe	Gain	1.000	0.080	0.395	0.259	0.912	0.380	0.918	0.247
	Phase	1.000	0.007	0.250	0.284	0.320	0.069		
	Gain+Phase	1.000	0.004	0.267	0.281	0.067	0.053		
Impedance Meter	Real	0.631	0.040	0.539	0.280	0.618	0.000	0.964	0.068
	Imaginary	1.000	0.214	0.447	0.310	0.138	0.000		
	Real+Imaginary	0.631	0.040	0.550	0.254	0.786	0.023	1.000	0.000

Empty cells indicate that stepwise did not find any data points that met the specified level of significance and no model was made.

Table 5.14: RMSE values for predicting concentration of 1 L atrazine samples (mg/L)

		SWR		PLS		ANN		Wavelet	
		Training	Validation	Training	Validation	Training	Validation	Training	Validation
2 cm Probe	Gain	2.72	5.74	4.47	5.15	7.79	7.10		
	Phase	0.00	6.90	0.38	7.02	5.84	7.07	0.03	7.51
	Gain+Phase	2.72	5.74	0.34	6.42	2.42	5.46		
2.5 cm Probe	Gain	0.00	6.75	0.09	6.92	5.08	7.24	0.88	7.36
	Phase	0.00	7.39	1.96	5.71	7.34	4.95	0.00	6.14
	Gain+Phase	0.00	5.63	0.05	6.87	4.67	6.30	0.88	7.36
5 cm Probe	Gain	0.76	5.95	1.30	8.09	6.44	5.03	2.54	5.20
	Phase	0.00	9.51	4.00	6.81	6.81	6.07	3.04	10.17
	Gain+Phase	0.00	5.22	3.37	7.21	2.68	9.88	2.54	5.20
7.5 cm Probe	Gain	0.00	7.72	3.68	5.01	1.79	4.73	1.36	7.84
	Phase	0.00	7.69	4.10	4.69	4.42	5.86		
	Gain+Phase	0.00	7.24	4.05	4.71	5.30	5.11		
Impedance Meter	Real	2.88	5.19	3.21	4.60	3.24	7.10	0.90	10.60
	Imaginary	0.00	5.38	3.52	5.16	5.33	6.17		
	Real+Imaginary	2.88	5.19	3.18	4.60	2.66	8.22	0.00	6.71

Empty cells indicate that stepwise did not find any data points that met the specified level of significance and no model was made.

5.2.2 Thirty-six Liter Samples

The results for predicting atrazine concentration were not as good as the results for predicting nitrate concentration with the 36 L samples, but they were better than the results from the 1 L atrazine tests. Overall, the 2 cm probe was the best probe for atrazine prediction and wavelet transformation followed by stepwise regression was the best regression method. Table 5.15 shows the R^2-values for predicting the concentration of the 36 L atrazine samples, and Table 5.16 shows the RMSE values.

5.2.2.1 Comparison of Probes

The 2 cm probe did the best job of predicting the concentration of atrazine in the 36 L samples. The RMSE values for both training and validation were lowest, on average, for this probe. It also had the highest average training R^2-value, 0.837, and the highest average validation R^2-value, 0.693. The training R^2-values for this probe were over 0.900 for 66.7% of the datasets,

and the validation R^2-values were above 0.900 for 41.7% of the tests, both higher than for any other probe.

The 7.5 cm probe had a few very high R^2-values. All of the R^2-values for PLS and stepwise were 1.00 with this probe, but the neural network R^2-values for this probe were extremely low. The 7.5 cm probe also had very poor values for validation. The highest validation R^2-value for the 7.5 cm probe was only 0.829 and the average was 0.558. The average R^2-values were similar for the 2.5 cm probe, 0.701 for training and 0.595 for validation. The 2.5 cm probe did not have as many high validation R^2-values as the other two probes, but the lowest validation R^2-value for this probe was 0.187, about ten times higher than the lowest values for the other probes.

5.2.2.2 Comparison of Regression Methods

PLS and stepwise regression with wavelet preprocessing were the best regression methods for the 36 L atrazine tests. The average training R^2-value was higher for wavelet, 0.942 compared to 0.817. Wavelet also had two more training R^2-values above 0.900 than PLS did. The average validation R^2-value was slightly higher for PLS than for wavelet, 0.729 compared to 0.708, but wavelet had three validation R^2-values above 0.900 whereas PLS only had two. The training RMSE values were slightly lower, overall, for wavelet than for PLS, but the validation RMSE values were lower for PLS.

Stepwise regression had the highest R^2-values and lowest RMSE values for training, but the validation values for stepwise were not as good as they were for PLS and wavelet. The average validation R^2-value for stepwise was only 0.611. The neural network results were the worst, overall. It had the highest RMSE values and the lowest R^2-values. The average training R^2-value for neural network was only 0.298.

Table 5.15: R²-values for predicting concentration of 36 L atrazine samples

		SWR		PLS		ANN		Wavelet	
		Training	Validation	Training	Validation	Training	Validation	Training	Validation
2 cm Probe	Gain	**0.986**	**0.979**	**0.985**	**0.932**	0.002	0.094	0.834	0.339
	Phase	1.000	0.676	0.894	0.773	0.965	0.800	0.972	0.729
	Gain+Phase	**1.000**	**0.934**	**0.981**	**0.914**	0.419	0.180	**0.999**	**0.964**
2.5 cm Probe	Gain	1.000	0.418	0.490	0.511	0.467	0.892	**1.000**	**0.948**
	Phase	1.000	0.187	0.518	0.559	0.701	0.715	0.970	0.618
	Gain+Phase	1.000	0.446	0.490	0.518	0.033	0.385	**0.738**	**0.942**
7.5 cm Probe	Gain	1.000	0.633	1.000	0.750	0.018	0.041	0.998	0.787
	Phase	1.000	0.649	1.000	0.829	0.044	0.100	0.971	0.254
	Gain+Phase	1.000	0.581	1.000	0.778	0.032	0.504	0.998	0.787

Table 5.16: RMSE values for predicting concentration of 36 L atrazine samples (mg/L)

		SWR		PLS		ANN		Wavelet	
		Training	Validation	Training	Validation	Training	Validation	Training	Validation
2 cm Probe	Gain	0.553	0.829	0.587	1.670	6.639	6.195	1.927	5.337
	Phase	0.000	3.378	1.539	2.302	1.027	2.427	0.797	2.630
	Gain+Phase	0.000	1.367	0.653	1.997	4.056	5.487	0.123	1.041
2.5 cm Probe	Gain	0.000	3.778	3.381	3.313	4.243	2.375	0.006	1.176
	Phase	0.002	5.952	3.285	3.152	2.809	3.685	0.818	3.784
	Gain+Phase	0.000	9.790	3.381	3.289	5.117	4.006	2.421	2.600
7.5 cm Probe	Gain	0.000	2.941	0.088	3.167	5.186	5.981	0.226	3.158
	Phase	0.005	4.391	0.037	2.191	6.559	6.400	0.808	4.630
	Gain+Phase	0.000	3.394	0.072	2.749	6.392	6.550	0.226	3.158

5.3 Biodiesel Samples

The three probes studied with the biodiesel samples were able to measure the concentration of the impurities in the fuel. Overall, the results for the biodiesel tests were not as good as the water and nitrate tests, but there were still some good results achieved for the fuel tests. For each of the contaminants, the highest training R^2-value achieved was 1.00. The highest validation R^2-values were all above 0.960. Tables are given that show the R^2-values and RMSE values for predicting the concentration of each contaminant. For the R^2-value tables, the validation R^2-values above 0.800 are bolded and in blue.

5.3.1 Water

The results from the tests done with biodiesel and water showed that the probes had some ability to detect the concentration of the water. The first three tests that were done for the 2 cm probe were not accurate, and the signals for these three tests were drastically different from one another. This was the result of a problem with the screws in the probe. The screws that held the plates together had become loose, and the plates were no longer electrically connected as they should have been. Once the screws were tightened, the measurements from this probe were much more consistent. The first three concentrations of water measured were not included in the data during regression because of this problem. The R^2-values and RMSE values for predicting the concentration of water in biodiesel are given in Table 5.17 and Table 5.18, respectively.

5.3.1.1 Comparison of Probes

There was not a lot of variability in the ability of the three probes to detect the concentration of water in biodiesel for the training data. All had average training R^2-values between 0.800 and 0.850, and the R^2-values for the three probes was above 0.900 50% to 75% of the time. The 2.5 cm probe tended to have lower training RMSE values than the other two probes.

In terms of the validation data, the 2.5 cm probe performed the best. This probe had an average validation R^2-value of 0.619 and had R^2-values above 0.900 for 25 % of the data, higher than the other two probes. The 2.5 cm probe also had lower validation RMSE values than the other probes; the average was 805 ppm, compared to 1,270 ppm for the 7.5 cm probe and 1,540 ppm for the 2 cm probe.

5.3.1.2 Comparison of Regression Methods

Stepwise regression was the best regression method for the training data; the average training R^2-value was 0.984 and the training R^2-values were above 0.900 for 93% of the tests. Stepwise regression also had the lowest average RMSE value for training. The validation results for the stepwise models were not very good, however. The average validation R^2-value was 0.454 and the validation R^2-values were only above 0.900 19% of the time.

The best validation results, in terms of R^2-values, came from stepwise regression of wavelet transformed data. The average validation R^2-value for wavelet was 0.550, slightly higher than for traditional stepwise regression, and the validation R^2-value was above 0.900 19% of the time. This method also had good training R^2-values, the average was 0.956. In terms of validation RMSE values, partial least squares regression had lower values, on average, than any other method.

The validation R^2-values were comparable for PLS and neural network, both averaging just over 0.400. The RMSE values were higher, overall, for neural network.

Table 5.17: R^2-values for predicting concentration of water in biodiesel

		SWR		PLS		ANN		Wavelet	
		Training	Validation	Training	Validation	Training	Validation	Training	Validation
2 cm Probe	All Gain	**0.986**	**0.908**	0.867	0.225	0.989	0.418	0.992	0.324
	All Phase	0.999	0.347	0.800	0.345	0.996	0.594	0.837	0.424
	All Gain+Phase	0.974	0.454	0.877	0.278	0.805	0.174	0.947	0.540
	Low Gain	1.000	0.545	0.740	0.606	1.000	0.763	**1.000**	**0.992**
	Low Phase	1.000	0.195	0.796	0.434	1.000	0.604	1.000	0.238
	Low Gain+Phase	1.000	0.292	0.788	0.495	1.000	0.605	**0.916**	**0.996**
	High Gain	0.837	0.195	0.663	0.014	0.486	0.134	1.000	0.142
	High Phase	0.810	0.001	0.157	0.164	0.074	0.023	0.876	0.090
	High Gain+Phase	1.000	0.392	0.321	0.094	0.045	0.105		
2.5 cm Probe	All Gain	**0.985**	**0.883**	**0.996**	**0.845**	**0.958**	**0.836**	**0.995**	**0.939**
	All Phase	1.000	0.732	**0.991**	**0.846**	**0.971**	**0.918**	**0.970**	**0.861**
	All Gain+Phase	**0.985**	**0.883**	**0.997**	**0.898**	0.043	0.082	**0.984**	**0.870**
	Low Gain	**1.000**	**0.922**	0.548	0.127	0.253	0.560	1.000	0.161
	Low Phase	0.999	0.358	**0.996**	**0.922**	0.610	0.037	0.910	0.666
	Low Gain+Phase	**1.000**	**0.922**	0.533	0.140	0.013	0.007	0.825	0.327
	High Gain	**1.000**	**0.955**	0.996	0.684	0.005	0.000	**1.000**	**0.989**
	High Phase	**1.000**	**0.913**	0.999	0.313	0.010	0.140	**0.979**	**0.800**
	High Gain+Phase	1.000	0.194	1.000	0.711	**0.912**	**0.890**	**1.000**	**0.944**
7.5 cm Probe	All Gain	0.982	0.032	**0.980**	**0.910**	0.093	0.038	0.953	0.538
	All Phase	1.000	0.273	0.914	0.649	0.333	0.260	0.817	0.708
	All Gain+Phase	1.000	0.103	0.942	0.690	0.867	0.381	0.975	0.616
	Low Gain	1.000	0.191	0.872	0.717	**0.991**	**0.941**	0.987	0.549
	Low Phase	1.000	0.697	**0.912**	**0.950**	**0.976**	**0.962**	1.000	0.526
	Low Gain+Phase	1.000	0.423	**0.907**	**0.827**	0.385	0.341	0.896	0.553
	High Gain	1.000	0.059	0.934	0.083	**0.593**	**0.803**	1.000	0.013
	High Phase	1.000	0.351	0.811	0.069	0.009	0.365	1.000	0.080
	High Gain+Phase	1.000	0.026	0.846	0.070	0.360	0.012	1.000	0.425

Table 5.18: RMSE values for predicting concentration of water in biodiesel (ppm)

		SWR		PLS		ANN		Wavelet	
		Training	Validation	Training	Validation	Training	Validation	Training	Validation
2 cm Probe	All Gain	213.05	567.34	658.99	1530.18	225.72	1271.30	157.43	1951.16
	All Phase	66.10	2770.56	807.03	1356.08	116.73	1238.37	729.49	1269.17
	All Gain+Phase	289.03	1407.12	632.44	1450.43	1369.93	1899.29	414.03	1184.44
	Low Gain	3.13	279.14	161.11	274.77	0.00	171.55	5.87	85.22
	Low Phase	0.53	390.58	142.94	275.51	0.01	243.14	1.48	359.72
	Low Gain+Phase	0.00	836.66	145.57	271.67	0.04	242.38	91.41	156.10
	High Gain	590.03	2504.02	847.91	1518.54	1660.15	2826.77	0.53	5516.29
	High Phase	637.35	6758.71	1341.36	1462.20	1929.32	1765.32	515.17	4827.47
	High Gain+Phase	1.94	1401.86	1203.66	1462.71	1936.31	2188.98		
2.5 cm Probe	All Gain	221.25	631.96	120.75	807.99	421.35	1353.61	124.13	461.00
	All Phase	0.00	907.33	175.03	803.46	368.37	518.98	318.87	718.87
	All Gain+Phase	221.25	631.96	92.09	636.00	2687.47	2464.55	232.23	686.34
	Low Gain	0.00	288.03	281.24	401.52	387.18	275.78	0.06	593.40
	Low Phase	12.41	311.35	25.19	127.79	490.01	549.95	125.18	368.93
	Low Gain+Phase	0.00	288.03	285.96	416.55	472.39	459.88	175.24	479.15
	High Gain	0.00	410.07	86.62	1127.40	1930.52	2022.15	0.12	492.04
	High Phase	0.00	730.19	35.01	1475.34	2035.08	2299.19	214.16	985.41
	High Gain+Phase	0.00	1851.90	23.77	1117.54	442.67	893.15	0.11	374.65
7.5 cm Probe	All Gain	243.69	2121.59	255.79	563.90	1797.91	2041.82	394.78	2059.06
	All Phase	1.76	2858.16	537.21	1015.06	1738.75	1728.56	780.49	1454.58
	All Gain+Phase	0.00	1831.69	440.00	956.36	704.86	2009.99	289.12	1444.60
	Low Gain	0.00	508.53	149.58	219.18	42.76	174.42	48.02	306.15
	Low Phase	0.09	307.36	123.83	118.82	82.52	185.78	0.02	314.80
	Low Gain+Phase	0.01	421.57	127.34	177.93	690.79	628.12	134.66	563.55
	High Gain	0.00	1889.83	376.63	1551.38	1045.97	2864.72	1.57	1891.37
	High Phase	0.48	1376.38	634.60	1496.29	2203.94	1266.91	1.21	1722.50
	High Gain+Phase	0.00	2498.00	572.72	1506.14	1352.59	2234.30	0.56	1416.00

5.3.2 Glycerol

The three probes were mostly equal in their ability to detect changes in glycerol concentration. Partial least squares was the best regression method for predicting the glycerol concentration. Table 5.19 shows the R^2-values for predicting the concentration of glycerol in biodiesel, and Table 5.20 shows the RMSE values.

5.3.2.1 Comparison of Probes

The three probes were fairly equivalent in their ability to predict the concentration of glycerol in biodiesel and 5,000 ppm water. The average training R^2-values ranged from 0.836 to 0.887. The 2 cm probe was the lowest, and the 7.5 cm probe was the highest. The average validation R^2-values ranged from 0.560 to 0.663; the 2 cm probe was the lowest and the 2.5 cm probe was the highest. These probes also had similar percentages of R^2-values above 0.900. For training they

ranged from 58% for the 2 cm probe to 69% for the 2.5 cm probe. For validation, the percent of R^2-values above 0.900 ranged from 8% for the 2 cm probe to 17% for the 7.5 cm probe.

The average training RMSE values for the probes were also very similar; the 2.5 cm probe had the lowest value of 109.1 mg/L, followed by the 7.5 cm probe at 112.9 mg/L, and the 2 cm probe at 129.6 mg/L. The average validation RMSE values for the 2.5 and 7.5 cm probes were comparable, 334.7 mg/L and 326.9 mg/L, respectively, but the 2 cm probe had a much higher average validation RMSE, 410.0 mg/L.

5.3.2.2 Comparison of Regression Techniques

Partial least squares was the most reliable method for predicting the concentration of glycerol in biodiesel. This method had the second highest average training R^2-value, 0.934, and the second lowest average training RMSE value, 58.47 mg/L. PLS was the best by far in terms of validation. The average validation R^2-value was 0.705, much higher than for any other method, and the validation R^2-values were above 0.900 33% of the time. Wavelet transform followed by stepwise regression was the method with the second highest percentage of validation R^2-values above 0.900, only 7.4%. PLS also had much lower validation RMSE values than the other regression methods, the average was 258.2 mg/L.

The other three regression methods had similar results for predicting the glycerol concentration. Stepwise regression had the highest training R^2-values, 0.995 on average, and the lowest training RMSE values, 17.11 mg/L on average, but the validation results for stepwise were no than for neural network or stepwise regression of wavelet transformed data.

Table 5.19: R^2-values for predicting concentration of glycerol in biodiesel

		SWR		PLS		ANN		Wavelet	
		Training	Validation	Training	Validation	Training	Validation	Training	Validation
2 cm Probe	All Gain	1.000	0.859	0.968	0.676	0.996	0.759	0.951	0.707
	All Phase	0.896	0.690	0.997	0.464	0.721	0.595	0.991	0.436
	All Gain+Phase	1.000	0.655	0.981	0.617	0.430	0.316	0.837	0.724
	Low Gain	1.000	0.440	0.779	0.149	0.000	0.239	0.804	0.090
	Low Phase	1.000	0.430	0.771	0.341	0.039	0.660	0.950	0.084
	Low Gain+Phase	1.000	0.249	0.788	0.227	0.743	0.221	0.830	0.326
	High Gain	1.000	0.800	0.971	0.918	0.287	0.342	0.820	0.783
	High Phase	0.998	0.840	0.980	0.966	0.981	0.860	1.000	0.722
	High Gain+Phase	1.000	0.350	0.977	0.924	0.617	0.790	1.000	0.893
2.5 cm Probe	All Gain	1.000	0.688	1.000	0.923	0.969	0.800	0.991	0.890
	All Phase	1.000	0.713	1.000	0.925	0.839	0.583	0.984	0.883
	All Gain+Phase	1.000	0.857	0.999	0.954	0.664	0.770	0.794	0.813
	Low Gain	1.000	0.366	0.716	0.539	0.585	0.000	1.000	0.457
	Low Phase	1.000	0.850	0.705	0.509	0.927	0.571	0.808	0.907
	Low Gain+Phase	1.000	0.883	0.720	0.544	0.051	0.018	1.000	0.750
	High Gain	1.000	0.360	0.981	0.776	0.929	0.641	0.945	0.507
	High Phase	1.000	0.403	0.991	0.795	0.748	0.216	1.000	0.858
	High Gain+Phase	1.000	0.763	0.967	0.762	0.575	0.805	0.991	0.778
7.5 cm Probe	All Gain	0.988	0.254	0.999	0.906	0.730	0.405	0.819	0.572
	All Phase	0.998	0.790	1.000	0.936	0.779	0.569	1.000	0.910
	All Gain+Phase	0.988	0.254	0.998	0.942	0.798	0.541	0.893	0.731
	Low Gain	1.000	0.719	0.978	0.480	0.103	0.201	1.000	0.500
	Low Phase	1.000	0.601	0.997	0.697	0.421	0.626	0.954	0.029
	Low Gain+Phase	1.000	0.306	0.986	0.522	0.886	0.206	0.840	0.009
	High Gain	1.000	0.915	0.963	0.896	0.752	0.858	0.999	0.596
	High Phase	1.000	0.496	0.998	0.863	0.278	0.938	1.000	0.783
	High Gain+Phase	1.000	0.549	0.999	0.781	0.780	0.803	1.000	0.816

Table 5.20: RMSE values for predicting concentration of glycerol in biodiesel (mg/L)

		SWR		PLS		ANN		Wavelet	
		Training	Validation	Training	Validation	Training	Validation	Training	Validation
2 cm Probe	All Gain	0.00	501.12	129.74	398.83	50.73	364.99	161.49	378.08
	All Phase	236.17	391.12	37.45	554.31	434.93	508.04	67.40	730.00
	All Gain+Phase	8.87	416.76	99.64	446.56	593.79	589.18	295.33	379.38
	Low Gain	0.02	118.61	78.67	143.41	215.24	193.40	74.17	163.29
	Low Phase	0.29	251.09	80.14	121.70	226.60	87.97	37.53	222.90
	Low Gain+Phase	0.01	159.64	77.08	134.32	88.98	179.97	69.02	128.47
	High Gain	0.00	672.51	100.19	577.67	544.21	581.41	247.71	808.32
	High Phase	27.57	722.89	82.08	393.78	92.14	283.49	0.22	467.70
	High Gain+Phase	0.00	1537.44	88.09	380.85	419.80	390.87	0.40	381.73
2.5 cm Probe	All Gain	0.01	511.48	0.59	289.16	140.52	352.95	67.55	266.63
	All Phase	0.00	868.39	4.39	217.44	312.16	520.53	93.50	433.30
	All Gain+Phase	0.00	556.08	22.58	191.29	545.27	394.41	331.97	385.89
	Low Gain	0.00	220.00	89.15	110.39	162.43	188.21	0.36	113.06
	Low Phase	0.00	59.22	90.96	109.17	46.71	134.16	73.36	195.24
	Low Gain+Phase	0.00	53.27	88.47	107.35	231.22	241.30	0.02	130.21
	High Gain	0.30	518.85	81.43	345.42	235.77	397.64	136.96	465.03
	High Phase	0.00	721.48	55.82	327.09	354.47	610.96	0.04	337.76
	High Gain+Phase	0.03	301.41	105.90	367.90	597.27	615.48	56.81	392.69
7.5 cm Probe	All Gain	78.91	687.22	22.84	211.79	449.44	602.47	310.47	541.23
	All Phase	31.36	333.70	0.85	202.77	356.84	568.83	5.07	234.17
	All Gain+Phase	78.91	687.22	29.52	168.93	389.89	661.05	239.45	378.84
	Low Gain	0.10	115.28	25.04	110.86	235.10	249.26	0.26	107.01
	Low Phase	0.03	219.78	8.75	81.83	174.96	124.34	35.72	207.93
	Low Gain+Phase	0.01	190.76	19.82	105.30	58.12	174.69	66.94	195.57
	High Gain	0.00	427.82	112.48	270.05	398.18	392.04	15.36	402.82
	High Phase	0.00	489.96	27.95	270.91	512.76	682.11	11.98	289.27
	High Gain+Phase	0.00	465.97	19.02	332.50	349.36	276.59	0.21	308.17

5.3.3 Glyceride

The three probes were roughly equally capable of measuring changes in the glyceride concentration in the biodiesel. Partial least squares was the best regression method, and the only method that had good results for validation. The R^2-values for predicting the concentration of glyceride in biodiesel are shown in Table 5.21, and the RMSE values are given in Table 5.22.

5.3.3.1 Comparison of Probes

The range of R^2-values for the three probes was very small for predicting the concentration of glyceride in biodiesel. The average training R^2-values ranged from 0.756 for the 2.5 cm probe to 0.799 for the 2 cm probe. The average validation R^2-values ranged from 0.434 for the 2.5 cm probe to 0.569. The 2 cm probe had the highest percentage of R^2-values above 0.900, 56% for training and 22% for validation. The 2 cm probe also had the lowest average training RMSE

value, 2,073.1 mg/L; however, the 2.5 cm probe had the lowest average validation RMSE, 4,898.3 mg/L.

5.3.3.2 Comparison of Regression Techniques

PLS was the only method with good validation results. The average validation R^2-value for PLS was 0.601, higher than for any other method, and the average validation RMSE value for PLS was lower than it was for the other methods, 4,108.4 mg/L.

Once again, stepwise had the highest training R^2-values and lowest validation RMSE values, but had poor validation values. The average training R^2-value for stepwise was 0.987 and the training R^2-values for stepwise were above 0.900 for 93% of the datasets. The average validation R^2-value for stepwise was only 0.425. Neither stepwise regression of wavelet transformed data nor neural network were very successful in predicting the glyceride concentration. Wavelet had a decent average R^2-value, 0.867, but wavelet and neural network both had validation R^2-values that were only just above 0.400.

Table 5.21: R^2-values for predicting concentration of glyceride in biodiesel

		SWR		PLS		ANN		Wavelet	
		Training	Validation	Training	Validation	Training	Validation	Training	Validation
2 cm Probe	All Gain	0.998	0.304	0.990	0.756	0.948	0.652	0.916	0.486
	All Phase	1.000	0.347	0.990	0.500	0.846	0.554	0.611	0.490
	All Gain+Phase	1.000	0.625	0.995	0.575	0.580	0.071	0.958	0.501
	Low Gain	1.000	0.478	**0.998**	**0.978**	0.705	0.252	**0.998**	**0.916**
	Low Phase	1.000	0.674	0.966	0.737	0.003	0.202	0.749	0.280
	Low Gain+Phase	1.000	0.775	**0.973**	**0.928**	0.004	0.375	**0.998**	**0.916**
	High Gain	1.000	0.082	0.615	0.029	0.480	0.370	0.873	0.288
	High Phase	**1.000**	**0.851**	**0.637**	**0.992**	0.240	0.703	**0.772**	**0.983**
	High Gain+Phase	**1.000**	**0.920**	**0.646**	**0.956**	0.440	0.501	0.821	0.418
2.5 cm Probe	All Gain	**0.994**	**0.929**	0.875	0.749	0.686	0.566	0.908	0.703
	All Phase	**1.000**	**0.825**	1.000	0.411	0.435	0.704	0.795	0.786
	All Gain+Phase	**1.000**	**0.946**	0.993	0.622	0.009	0.056	0.908	0.703
	Low Gain	1.000	0.183	0.841	0.688	0.150	0.471		
	Low Phase	1.000	0.429	0.805	0.541	0.154	0.004		
	Low Gain+Phase	1.000	0.215	0.851	0.391	0.828	0.555	1.000	0.437
	High Gain	1.000	0.093	0.488	0.298	0.220	0.308		
	High Phase	0.944	0.170	0.905	0.251	0.073	0.357	0.830	0.173
	High Gain+Phase	1.000	0.233	0.548	0.333	0.704	0.137	1.000	0.040
7.5 cm Probe	All Gain	0.997	0.065	1.000	0.357	0.095	0.114	0.937	0.441
	All Phase	0.842	0.076	0.979	0.644	0.909	0.125	0.552	0.152
	All Gain+Phase	1.000	0.070	1.000	0.491	0.045	0.077	0.455	0.494
	Low Gain	1.000	0.130	0.770	0.673	**0.643**	**0.835**	1.000	0.254
	Low Phase	0.884	0.680	0.852	0.691	0.967	0.528	0.847	0.650
	Low Gain+Phase	1.000	0.130	0.843	0.706	0.084	0.523	1.000	0.301
	High Gain	1.000	0.721	0.941	0.493	0.006	0.613	1.000	0.016
	High Phase	1.000	0.092	**0.880**	**0.832**	**0.246**	**0.838**	**1.000**	**0.893**
	High Gain+Phase	**1.000**	**0.881**	**0.922**	**0.970**	0.497	0.000	0.875	0.132

Table 5.22: RMSE values for predicting concentration of glyceride in biodiesel (mg/L)

		SWR		PLS		ANN		Wavelet	
		Training	Validation	Training	Validation	Training	Validation	Training	Validation
2 cm Probe	All Gain	336.92	6490.40	799.18	3909.95	2600.42	8634.15	2328.63	6225.25
	All Phase	0.32	6441.03	807.44	6303.54	3348.85	8300.25	5010.51	5853.49
	All Gain+Phase	0.01	5574.55	549.56	5539.96	5583.88	8530.70	1642.65	6244.25
	Low Gain	0.00	1260.21	78.03	423.93	1184.58	1684.43	81.18	1644.85
	Low Phase	0.78	1490.05	337.81	1008.26	2637.81	2828.99	922.26	1866.81
	Low Gain+Phase	0.71	1368.20	301.87	655.93	3101.61	2895.65	81.18	1644.85
	High Gain	0.00	10885.47	3989.78	8629.90	10417.79	10840.70	2290.85	7123.10
	High Phase	1.70	7938.58	3874.23	4549.01	6616.65	7218.44	3066.22	2751.80
	High Gain+Phase	0.00	10478.20	3822.96	6019.59	6094.70	9342.64	2719.96	6848.71
2.5 cm Probe	All Gain	618.63	2859.31	2841.50	4224.28	4541.80	7410.23	2439.51	4031.91
	All Phase	129.14	4557.29	22.82	7075.89	6463.95	4136.76	3636.28	3940.35
	All Gain+Phase	0.29	2402.41	669.64	5331.72	9786.36	8167.29	2439.51	4031.91
	Low Gain	0.00	2688.23	733.81	1797.56	1821.77	1612.61		
	Low Phase	5.36	1395.07	812.10	1467.33	2056.20	2347.05		
	Low Gain+Phase	0.00	1649.57	709.41	1692.03	998.21	1367.15	1.53	3613.20
	High Gain	0.00	10657.33	4598.91	6539.93	5796.09	10508.56		
	High Phase	1515.56	6212.81	1979.59	6004.18	7570.30	5222.87	2646.42	6059.05
	High Gain+Phase	0.00	5982.85	4319.41	6388.00	9366.08	8303.18	19.10	11966.62
7.5 cm Probe	All Gain	470.36	10189.65	0.01	6337.60	8285.88	10158.88	2011.59	7856.14
	All Phase	3192.24	10908.39	1157.09	4662.13	2484.22	10728.78	5377.21	7236.97
	All Gain+Phase	0.04	11121.32	48.86	5485.68	13774.14	11943.67	5931.38	5287.59
	Low Gain	0.00	6673.84	883.24	1030.89	1178.40	1550.74	1.30	2367.19
	Low Phase	626.14	1159.07	707.50	1101.29	437.31	2255.54	720.50	1252.14
	Low Gain+Phase	0.00	6673.84	728.58	1045.99	2208.32	3134.47	10.36	3384.61
	High Gain	0.00	12732.15	1555.17	5016.29	9844.12	10545.21	11.13	7658.10
	High Phase	5.32	8100.77	2228.59	4236.52	5840.48	4087.69	5.08	10731.67
	High Gain+Phase	0.00	3028.75	1799.04	4448.63	4922.12	7255.55	2269.96	6216.08

5.4 Most Significant Frequencies from Stepwise

5.4.1 Nitrate Salts and Water

5.4.1.1 Concentration

Stepwise regression selected a large number of frequencies to build regression models to predict the concentration of nitrates in water. The frequency that was selected the most to predict nitrate concentration from the old control box was 1,000,000 Hz, and 17,600,000 and 301,600,000 Hz were selected most often from the new control box.

5.4.1.1.1 One Liter Samples

A complete table of the frequencies that stepwise used for each sensor to predict the concentration of nitrate in the 1 L water samples is given in Appendix F. In Figure 5.1, the total

number of frequencies measured by the old control box is divided evenly into twenty groups, each containing 31 or 32 frequencies. The horizontal bars next to these frequency ranges indicate the number of frequencies stepwise selected from each of these ranges to build regression models for predicting the concentration of nitrate in the 1 L samples. The selected frequencies were distributed fairly evenly from 50 Hz to 120 MHz. The frequency range that had the most frequencies selected from it was the 56,600,000 to 62,800,000 Hz range. The smallest number of frequencies was selected from the 24,800,00 to 31,000,000 Hz range. The individual frequency that occurs most often in nitrate prediction with the old control box and the impedance meter was 1,000,000 which was selected 4 times.

The nitrate tests with the 2 cm probe were done using the new control box, so the range of frequencies at which measurements were taken was different. A complete table of the frequencies selected with the 2 cm probe and the new control box is given in Appendix G. The frequency that was selected the most often for the 2 cm probe was 181,600,000 Hz. It was selected five times.
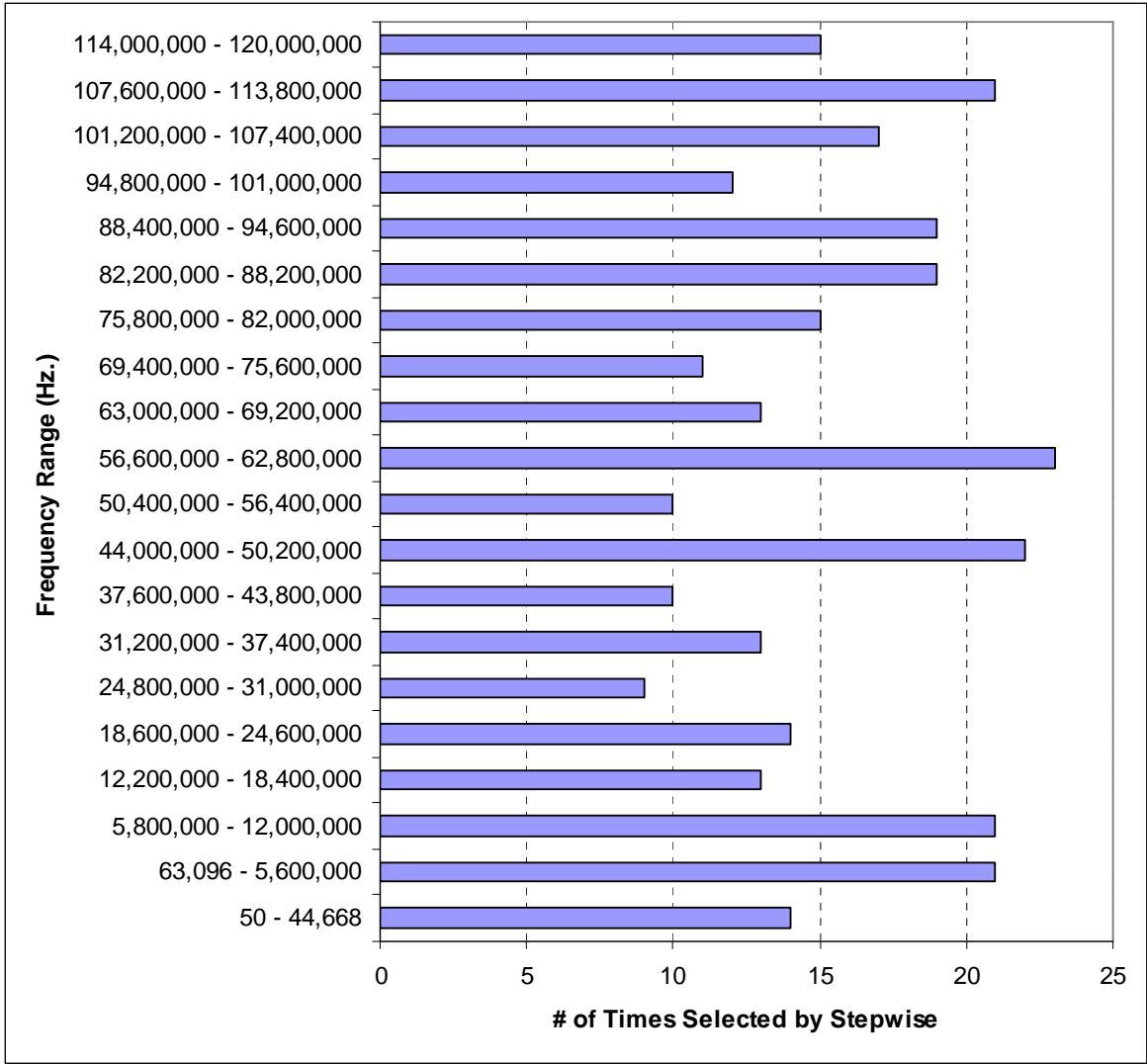
Figure 5.1: Frequencies selected by stepwise for predicting the concentration of 1 L nitrate samples with the 2.5-7.5 cm probes and the old control box or with the impedance meter

5.4.1.1.2 Thirty-six Liter Samples

A complete list of the frequencies selected to predict nitrate concentration for the 36 L samples is shown in Appendix H, and Figure 5.2 shows how many frequencies from each range were selected by stepwise. The frequencies that were selected most often with the 36 L samples were 17,600,000 and 301,600,000 Hz, both of which were used seven times. The frequency range from which the most frequencies were selected was the 2,400,000 to 22,400,000 Hz range. The smallest number of frequencies was selected from the 274,400,000 to 294,400,000 Hz range.

Overall, a lot more frequencies were selected from the lower frequencies; 185 were selected from the lower half and only 111 were selected from the upper half.
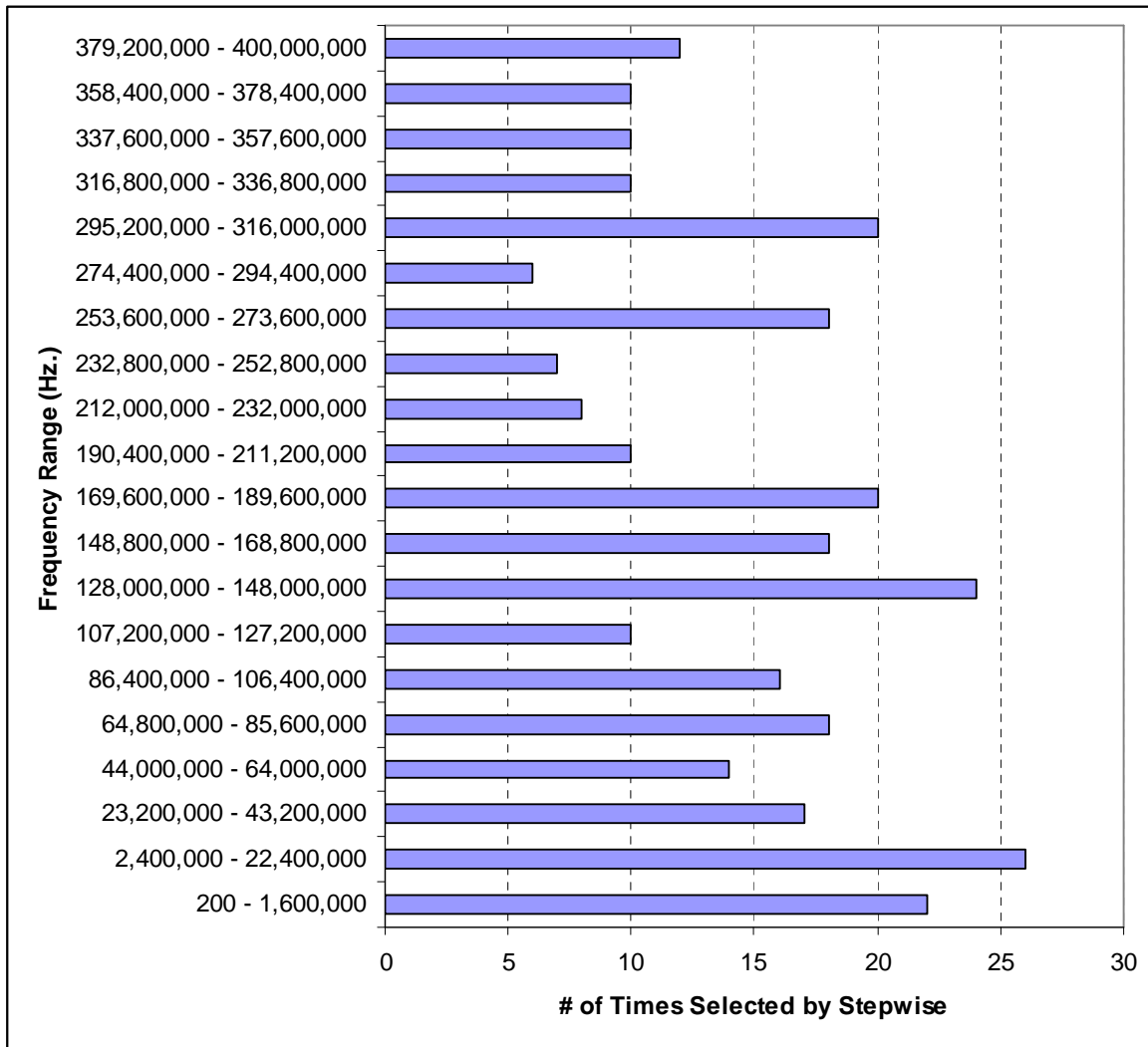
Figure 5.2: Frequencies selected by stepwise for predicting the concentration of 36 L nitrate samples with the 2, 2.5, and 7.5 cm probes and the new control box

5.4.1.2 Molecular Weight

The number of frequencies selected by stepwise to predict molecular weight was a lot smaller than the number used to predict concentration. The frequency used most often from the old control box was 110.2 MHZ, and 2,674.4 MHz was used most often from the new control box.

5.4.1.2.1 One Liter Samples

A complete list of the frequencies selected for the 1 L samples with the impedance meter and the old control box can be found in Appendix I, and a list of the frequencies selected for the 2 cm probe and the new control box can be found in Appendix J. Eighty-six frequencies were selected to create stepwise regression models to predict the molecular weight of the 1 L nitrate samples with the old control box. The frequency that was selected most was 110.2 MHz, and it was selected five times. Higher frequencies were selected to make these predictions. Almost twice as many frequencies were selected above 56.4 MHz, the middle frequency for the old control box, than below 56.4 MHz.

Twenty-three frequencies were selected to predict the molecular weight of the 1 L nitrate solutions with the 2 cm probe and the new control box. Only one of these frequencies was used more than once, 1,413 Hz. The frequencies that were selected were distributed fairly evenly throughout the entire frequency range.

5.4.1.2.2 Thirty-six Liter Samples

The 90 frequencies that stepwise used to predict the molecular weight of the 36 L nitrate samples can be found in Appendix K. The most used frequency, 321,600,00 Hz, was selected six times. As with the 1 L samples, a very large proportion of high frequencies was used to make the stepwise regression models. The middle frequency in the range measured by the new control box is 189,600,000 Hz. Approximately 1.5 times more frequencies were selected above this value than below it.

5.4.2 Atrazine and Water

5.4.2.1 One Liter Samples

The stepwise regression program did not select any frequencies more than twice for the 1 L atrazine samples with the old control box and impedance meter. A table of the frequencies that it did select is given in Appendix L. Thirty-three different frequencies were used, and seven of these were used twice. There was no particular pattern in the frequencies that were selected.

Only four frequencies were selected to predict the 1 L atrazine concentration with the 2 cm probe and the new control box, and they are shown in Appendix M. One of the frequencies, 137,600,000 Hz, was used twice. This was also the largest selected frequency.

5.4.2.2 Thirty-six Liter Samples

Appendix N gives a table of the 35 frequencies selected to predict the concentration of atrazine for the 36 L samples. The frequency that was selected most often was 245,600,000 Hz, which was selected three times. None of the frequencies selected were above the middle frequency, 189,600,000 Hz. Only three frequencies above 1,464.8 MHz were selected. There was no pattern to the frequencies selected when the old control box was used, but in both experiments that used the new control box, which can measure the gain and phase at higher frequencies that the old control box, the frequencies selected by stepwise to measure atrazine concentration were low.

5.4.3 Impurities in Biodiesel Fuel

All of the frequencies used by stepwise to predict the concentration of impurities in biodiesel can be found in Appendix O. One hundred and forty frequencies were used. The frequency that was selected the most often was 69,600,000 Hz, which was used seven times. There was not much of a pattern to the frequencies that were selected. Frequencies below 189,600,000 Hz, the middle frequency, were selected 155 times, and frequencies above 189,600,000 Hz were selected 157 times.

5.5 Most Significant Wavebands

5.5.1 Nitrate Salts and Water

5.5.1.1 Concentration

For both control boxes and sample sizes, the wavebands used the most to predict the nitrate concentration in water contained information from lower frequencies. The first waveband of the fifth level was used many times by stepwise for nitrate concentration prediction.

5.5.1.1.1 One Liter Samples

Stepwise regression used several wavebands to predict the concentration of nitrate salts in the 1 L water samples. Figure 5.3 is a tiling diagram that summarizes the wavebands that were selected and how many times they were selected. Data from the new control box and the impedance meter are not included in this diagram because they collected data at different frequencies and therefore have different numbers of wavebands in each level. A complete table of the levels and wavebands that were selected by stepwise for each sensor and each data subset can be found in Appendix P.
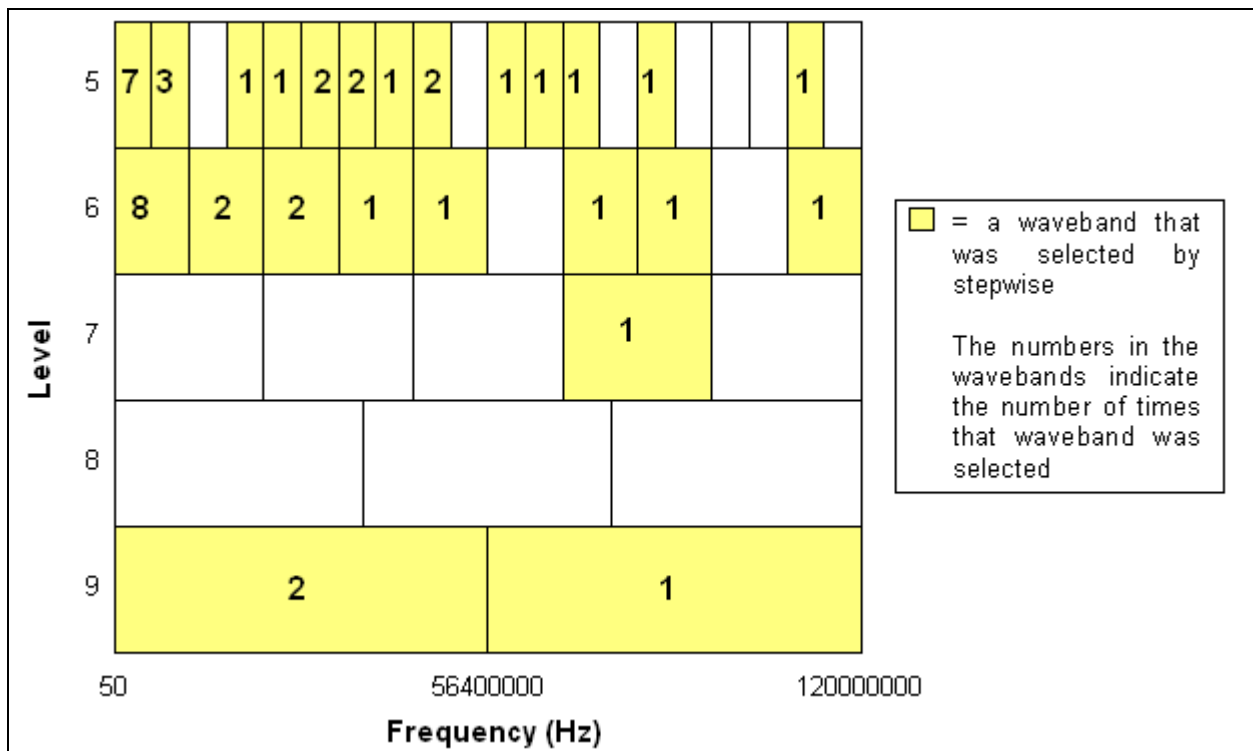
Figure 5.3: Wavebands used to predict 1 L nitrate concentration with 2.5, 5, and 7.5 cm probes and old control box

Many of the wavebands that were used were selected for more than one probe or for both the gain and phase data for the same probe. The two that were selected most often were the first waveband of the fifth level, which contains information about data from frequencies between 50 and 44,668 Hz, and the first waveband of the sixth level, which contains information about frequencies between 50 and 5,600,000 Hz.

5.5.1.1.2 Thirty-six Liter Samples

Stepwise regression also used several wavebands to predict the concentration of the 36 L nitrate samples. A detailed list of the wavebands that were selected by stepwise regression for each probe and each data subset can be found in Appendix Q. Figure 5.4 shows how many times each waveband was selected. The first and seventh wavebands from the fifth level were used the most often. These wavebands represent data from frequencies ranging from 200 to 5,600,000 Hz and 128,000,000 to 152,800,000 Hz, respectively.
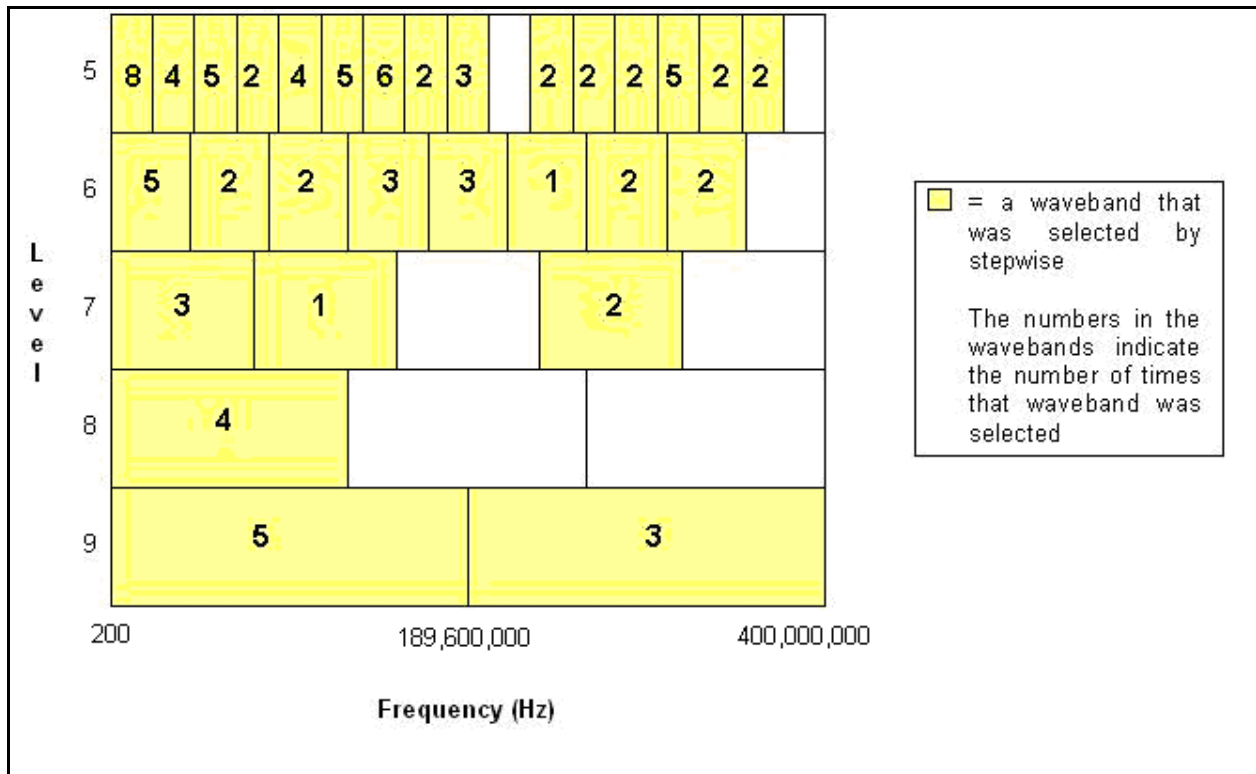
Figure 5.4: Wavebands used to predict 36 L nitrate concentration with 2, 2.5, and 7.5 cm probes and new control box

5.5.1.2 Molecular Weight

A fewer number of wavebands were selected to predict molecular weight of nitrates than to predict the concentration of nitrates. Only four wavebands were selected from the data for the 1 L samples.

84

5.5.1.2.1 One Liter Samples

Very few wavebands were used to predict the molecular weight of the nitrates in the 1 L samples because the stepwise program was often unable to find wavebands that were significant enough to build a model with. Figure 5.5 shows the wavebands that were chosen: the tenth and 15th wavebands from the fifth level, the fourth waveband from the seventh level, and the first waveband from the ninth level.
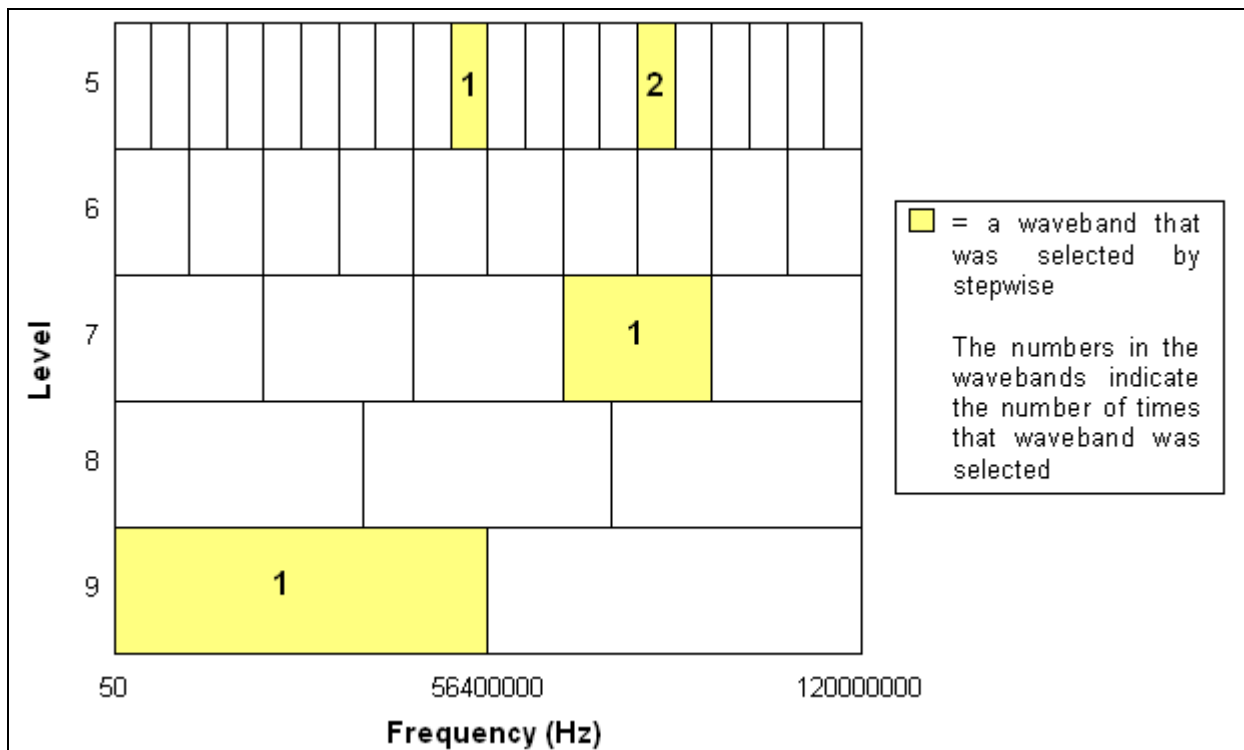
Figure 5.5: Wavebands used to predict 1 L nitrate molecular weight with 2.5, 5, and 7.5 cm probes and old control box

5.5.1.2.2 Thirty-six Liter Samples

Stepwise regression selected more wavebands to predict the molecular weight of the 36 L samples than of the 1 L samples. A table of the wavebands selected by stepwise for the 36 L samples can be found in Appendix R. The tiling diagram in Figure 5.6 shows that the distribution of wavebands selected for predicting the molecular weight of the 36 L samples is different than it was for predicting the concentration and that a lot fewer wavebands were selected for molecular weight. For the concentration predictions, almost all of the wavebands

corresponding to lower frequencies were selected for the models. For the molecular weight predictions, the lowest and highest wavebands were ignored, and bands in the middle were selected.

The waveband that was selected most frequently was the 14th band of the fifth level. The 13th and 15th bands on either side of it were also selected several times. The frequency range included in these three wavebands is 276,800,000 to 350,400,000 Hz. The fifth and sixth wavebands in the sixth level were selected four times each. Together these two wavebands range from 166,400,000 to 260,000,000 Hz. Therefore, there is no overlap in frequency in the bands used most often in the 5th level and the 6th level.
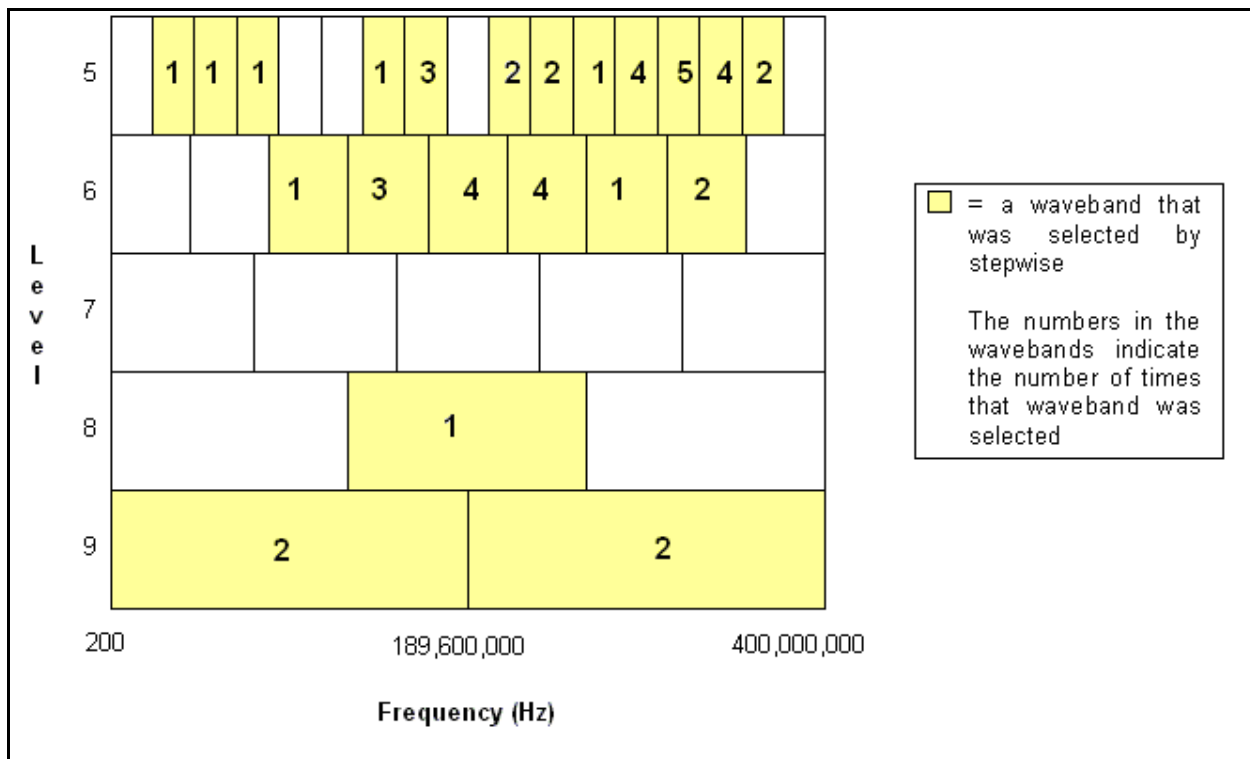
Figure 5.6: Wavebands used to predict 36 L nitrate molecular weight with the 2, 2.5, and 7.5 cm probes and the new control box

5.5.2 Atrazine and Water

Stepwise also selected very few wavebands to predict the concentration of atrazine in both the 1 L samples and in the 36 L samples. Appendix S lists the wavebands that were used for each samples size.

5.5.2.1 One Liter Samples

As illustrated in Figure 5.7, none of the wavebands selected by stepwise for the 1 L atrazine samples were used for more than one data subset. The wavebands all corresponded to lower frequencies; the highest waveband chosen was the seventh waveband of the sixth level, which contained information from the frequencies ranging from 69.4 to 82.0 MHz. There were no higher frequencies used.
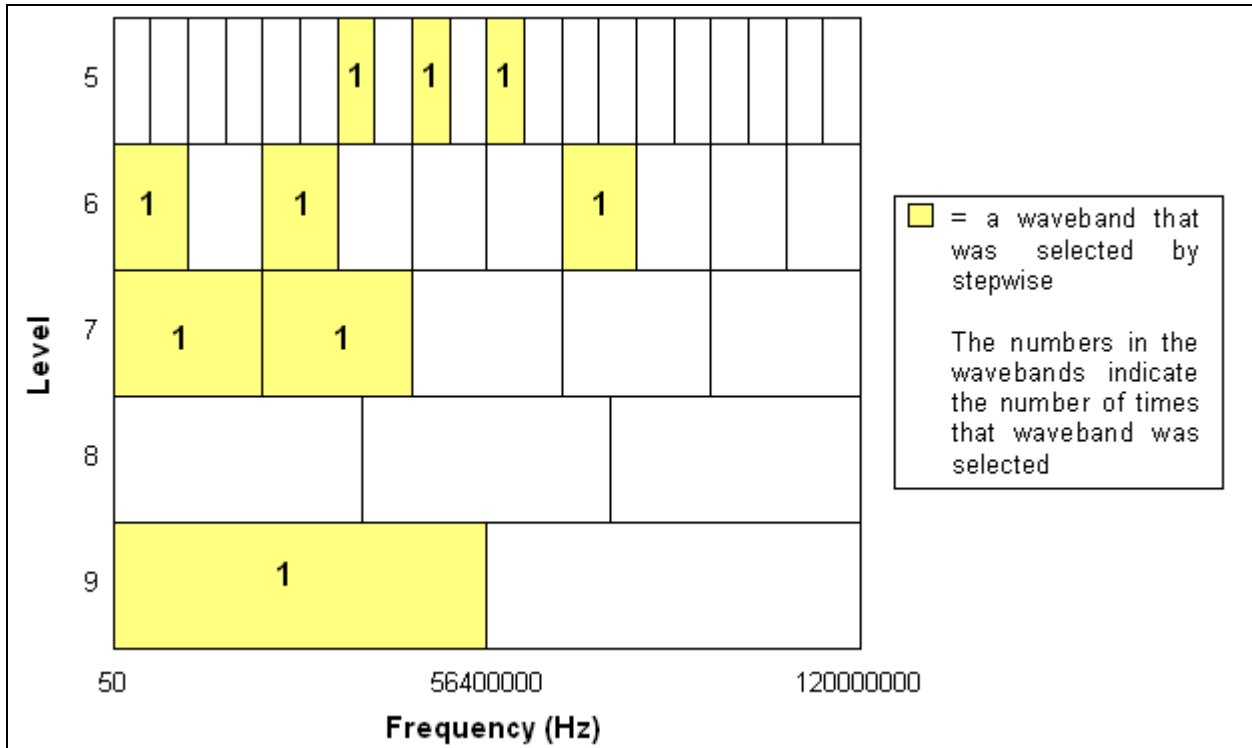
Figure 5.7: Wavebands used to predict 1 L atrazine concentration with the 2.5, 5, and 7.5 cm probes and the old control box

5.5.2.2 Thirty-six Liter Samples

Stepwise also selected wavebands corresponding to lower frequencies for the 36 L samples. This is summarized in the tiling diagram in Figure 5.8. The only waveband that was used for more than one subset of data was the first band of the seventh level. This waveband contains information from the frequency range 200 to 64,000,000 Hz. The seventh level was the highest level from which stepwise selected wavebands, no bands were chosen from levels eight or nine.
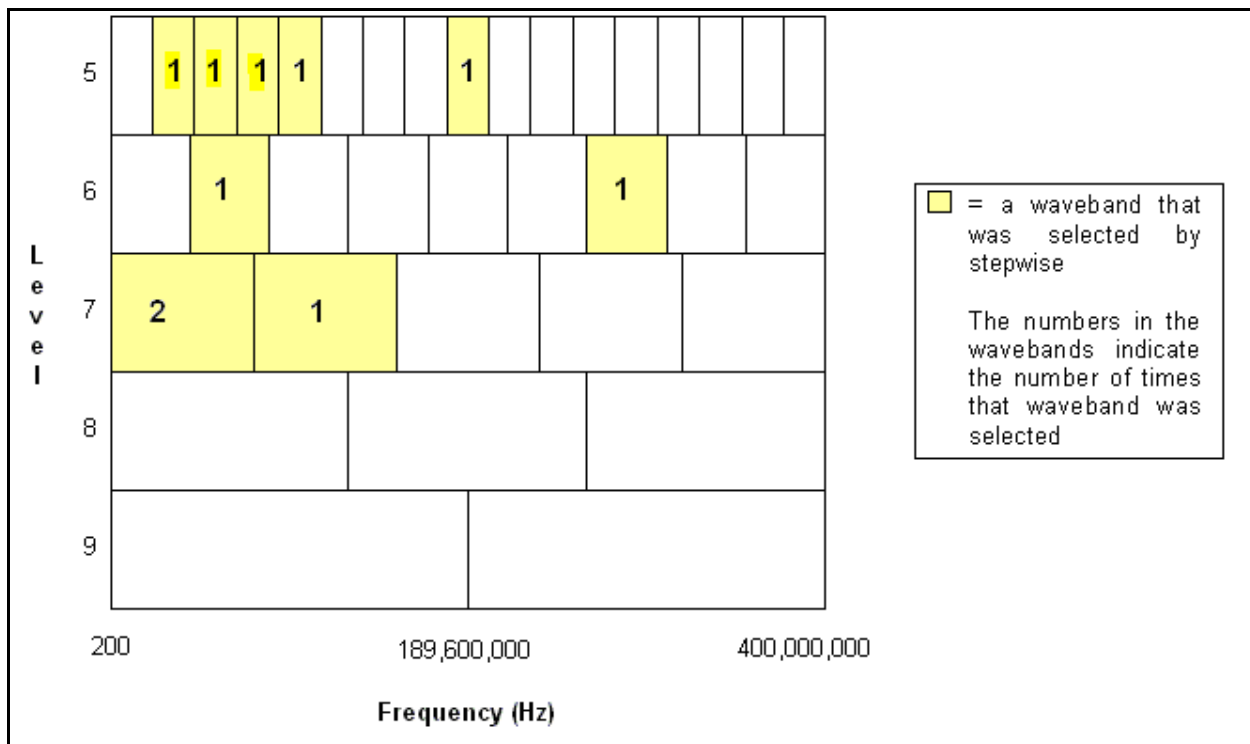
Figure 5.8: Wavebands used to predict 36 L atrazine concentration with the 2, 2.5, and 7.5 cm probes and the new control box

5.5.3 Biodiesel Impurities

All but four of the wavebands that were available to stepwise were used to build at least one regression model to predict the concentration of contaminants in biodiesel. The four wavebands that were not used were the highest wavebands for levels five through eight, the wavebands that correspond to the highest frequency ranges. This can be seen in Figure 5.9. A complete list of the wavebands used to build prediction models for each contaminant is located in Appendix T.

The waveband that is used most often, nine times, is the sixth waveband of the sixth level. This waveband contains information about the data recorded at frequencies of 213,600,000 to 260,000,000 Hz. The second most used waveband, which was selected seven times, is the tenth waveband of the fifth level which corresponds to data from the frequency range 202,400,000 to 227,200,000 Hz.
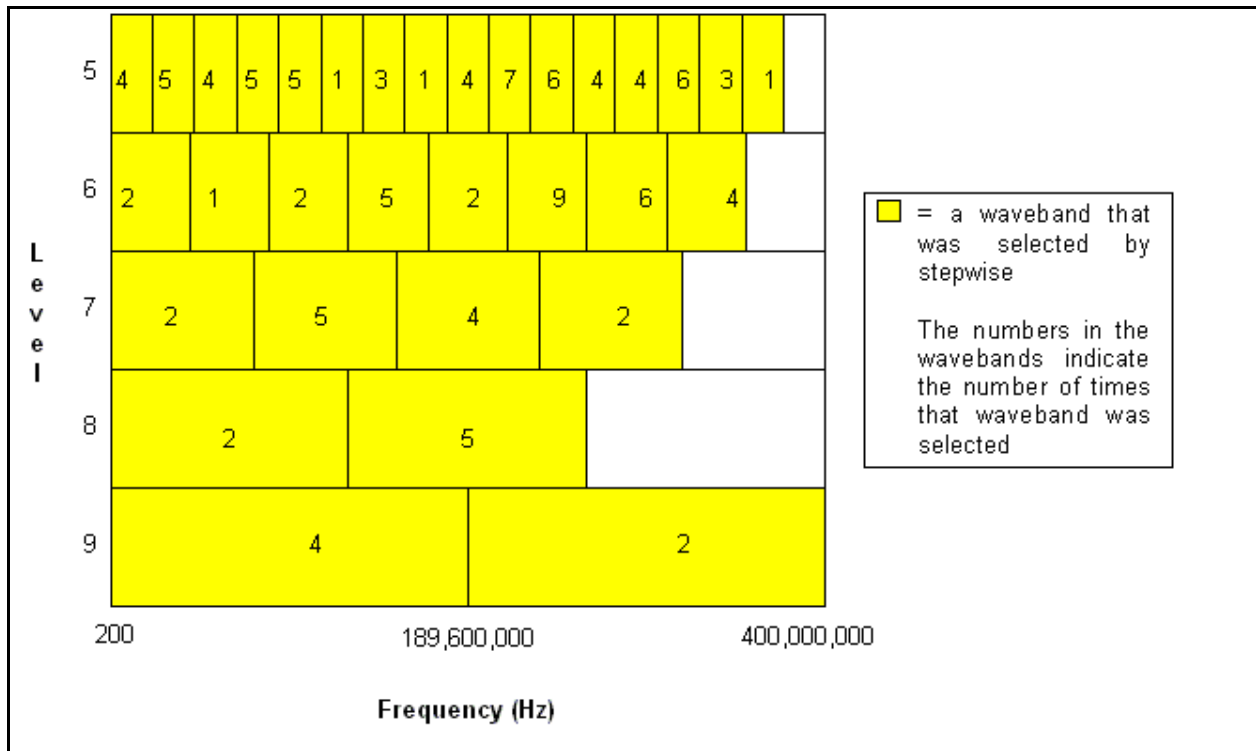
Figure 5.9: Wavebands used to predict the 36 L biodiesel impurities with the 2, 2.5, and 7.5 cm probes and the new control box

CHAPTER 6 - Conclusions

6.1 Water Tests

6.1.1 Probes

All of the probes tested in this study were able to detect changes in nitrate concentration to a reasonable degree of accuracy. The 2 cm probe required a larger sample size in order to detect these changes. Based on the tests done in this study, the 7.5 cm probe was able to detect nitrates in water better than the other probes. This probe almost always had the highest R^2-values for predicting the concentration of nitrate in both the 1 L water samples and the 36 L water samples.

The probes in this study were not able to accurately predict molecular weight of nitrates in a small sample of water. The results were greatly improved with the 36 L water samples; and the 7.5 cm probe was also the best at predicting the molecular weight of the nitrates. The 2 cm probe was also capable of measuring differences in molecular weight; however, the 2.5 cm probe was not as reliable as the others.

Atrazine was not detected as accurately as nitrates with these probes. For the 1 L atrazine samples, none of the probes especially outperformed the others. For the 36 L samples, the 2 cm probe detected the change in atrazine concentration much better than the other probes. The R^2-values and the RMSE values were consistently better for this probe than for any other probe.

The results for the three probes tested were much better for the 36 L sample size than for the 1 L samples.

6.1.2 Regression Techniques

Partial least squares was the most reliable regression method used to detect contaminants in water samples. PLS did not always have training R^2-values as high as stepwise regression, but it almost always had higher validation R^2-values than the other methods. PLS also had lower

RMSE values than the other methods did. Neural network was the least reliable of the methods studied. It usually had low R^2-values for both training and validation, and high RMSE values.

6.1.3 Important Frequencies

There were not any major trends in the frequencies and wavebands that stepwise used to make regression models. The frequencies used to predict nitrate concentration were often in the lower half of the frequency range, and the frequencies used to predict nitrate molecular weight were often in the upper half of the frequency range. This may be the reason that the molecular weight predictions were so much better for the 36 L samples; they were all tested using the new control box which measures the gain and phase at much higher frequencies than the old control box.

6.2 Biodiesel Tests

6.2.1 Probes

The three probes that were used to test the biodiesel fuel were all fairly comparable in their ability to measure the concentration of the contaminant in the fuel. The results, in terms of R^2- and RMSE values, indicate that the probes are not as capable of detecting changes in biodiesel impurity levels as they are in detecting water impurity levels. This could be a result of the settling of the impurities to the bottom of the aquarium due to their high densities.

6.2.2 Regression Techniques

PLS was the best regression method for predicting the concentration of glycerin and glyceride in biodiesel, and stepwise regression of wavelet transformed data predicted the concentration of water most accurately. Neural network typically produced the worst results.

6.2.3 Important Frequencies

Stepwise selected frequencies that were spread fairly evenly throughout the total frequency range, suggesting that higher and lower frequencies are equally significant in biodiesel

contaminant concentration prediction. The frequency that was used most often in biodiesel contaminant prediction was 69,600,000 Hz, which is one of the lower frequencies measured by the new control box measures data.

A large assortment of wavebands was used to generate the stepwise regression models. Not all of the wavebands corresponded to low frequencies; however the wavebands that contained information from the highest frequencies for levels five through eight were the only wavebands not selected. This could also indicate that lower frequencies are more useful in predicting biodiesel contaminants.

CHAPTER 7 - Recommendations for Future Research

The main goal of this experiment was to produce a portable, durable, and inexpensive probe to measure water and biodiesel contaminants. The experiments done for this thesis indicate that the probes studied do have some ability to measure contaminant concentration and to distinguish between different contaminants. More research will need to be conducted to determine how well these probes will work in real life situations. The following is a list of suggestions:

1. Tests should be conducted on samples that contain more than one contaminant.

2. A more in-depth study of container size and shape should be done. A variety of container sizes and shapes should be tested to determine which size and shape gives the most accurate results. Containers composed of different materials should also be compared because it is possible that the permittivity of the container material could affect the permittivity measurement of the sample. The container might also interfere with the electric field.

3. Independent 36 L samples should be prepared and tested in a random order to determine if testing the samples in order of increasing concentration falsely improves the results.

4. The tests should be repeated using only the frequencies that stepwise selected as significant and these results should be compared to the results obtained using data from the entire frequency range.

5. A better method for mixing the contaminants with the biodiesel should be developed and tested.

References

Ackerman, F. (2007). The economics of atrazine. *The International Journal of Occupational and Environmental Health, 13*(4), 441-449. Retrieved May 19, 2009 from http://ase.tufts.edu/gdae/Pubs/rp/EconAtrazine.pdf

ADM. (n.d.). *Biodiesel Technical Information.* Retrieved September 2, 2009 from http://www.biodiesel.org/pdf_files/Biodiesel_Technical_Manual.pdf

Agilent Technologies. (2009). *Agilent E4991A RF Impedance/Material Analyzer data sheet.* Retrieved August 20, 2009 from http://cp.literature.agilent.com/litweb/pdf/5980-1233E.pdf

Agilent Technologies. (2005). *Agilent E4991A RF Impedance/Material Analyzer operation manual, (7th ed.)*

Becker, J. F. (2009). *Capacitors.* Retrieved August 31, 2009 from http://www.physics.sjsu.edu/becker/physics51/capacitors.htm

Beebe, R.K. and Kowalski, B.R. (1987). An introduction to multivariate calibration and analysis. *Analytical Chemistry, 59*(17)

California State Water Resources Control Board. (2008). *Groundwater information sheet: nitrate/nitrite.* Retrieved September 1, 2009 from http://www.swrcb.ca.gov/gama/docs/coc_nitrate.pdf

Eigenvector Research, Inc. (2006). PLS_Toolbox (Version 4.0) [Software]. Wenatchee: Eigenvector Research, Inc.

Ge, Y, Morgan, C.L.S., Thomasson, J.A., and Waiser, T. (2007). A new perspective to near-infrared reflectance spectroscopy: A wavelet approach. *Transactions of the American Society of Agricultural and Biological Engineers 50*(1), 303-311

Hoar, P. (2008). *State W&M quality assurance.* Retrieved September 4, 2009 from http://www.biodieselconference.org/2008/post/secure/_xlI10oO/41%20Regulartoy%20Hoar.pdf

Jiang, D., Yang, X., Clinton, N. and Wang, N. (2004). An artificial neural network model for estimating crop yields using remotely sensed information. *International Journal of Remote Sensing, 25*(9), 1723-1732

Johnson, B., Whitford, F., Hahn, L., Flanke, D., Frankenberger, J., Janssen, C., and Bailey, T. (2006). *Atrazine and drinking water: Understanding the needs of farmers and citizens.* Retrieved May 19, 2009 from http://www.btny.purdue.edu/Pubs/PPP/PPP-66.pdf

Kansas Department of Health and Environment. (2004). *Primary drinking water regulations. Article 15a.* Retrieved May 19, 2009 from http://www.kdheks.gov/pws/regs/drinking_water_regs.pdf

Kim, M., Salley, S.O., and Ng, K.Y.S. (2008). Transestrification of glycerides using a heterogeneous resin catalyst combined with a homogeneous catalyst. *Energy Fuels, 22*(6), 3594-3599.

Leger, M.N. and Wentzell,P.D. (2004). Maximum likelihood principal components regression on wavelet-compressed data. *Applied Spectrometry, 58*(7), 855-862

Marengo, E., Longo, V., Robotti, E., Bobba, M., Gosetti, F., Zerbinati, O., and Martino, S.D. (2008). Development of calibration models for quality control in the production of ethylene/propylene copolymers by FTIR spectroscopy, multivariate statistical tools, and artificial neural networks. *Journal of Applied Polymer Science, 109*(6) , 3975-3982

Mosiello, L., Cremisini, C., Segre, L., Chiavarini, S., and Spanò, M. (1998). Dipstick
immunoassay format for atrazine and terbuthylazine analysis in water samples. *Journal
of Agricultural and Food Chemistry, 46*(9), 3847-3851

National Biodiesel Board. (2008). *Specification for biodiesel (B100)- ASTM D6751-08.*
Retrieved September 2, 2009 from
http://www.biodiesel.org/pdf_files/fuelfactsheets/BDSpec.PDF

National Safety Products Incorporated. (n.d.). *Consumer fact sheet on atrazine.* Retrieved May
19, 2009 from http://testproducts.com/water/consumer_fact_sheet_on_atrazine.htm

Natural Resources Defense Council. (2004). *EPA won't restrict toxic herbicide atrazine, despite
health threat.* Retrieved May 19, 2009 from
http://www.nrdc.org/health/pesticides/natrazine.asp

Nievergelt, Y. (1999). *Wavelets made easy.* Boston: Birkhäuser.

NREL. (2009). *Biodiesel handling and use guide, 4th ed.* Retrieved September 2, 2009 from
http://www.nrel.gov/vehiclesandfuels/pdfs/43672.pdf

Numerical Algorithms Group. (2007). Partial least squares. *NAGNews, 64.* Retrieved August
12, 2009 from http://www.nag.com/IndustryArticles/partialleastsquares.pdf

O'Dell, J.W. (1993). *Method 353.2: Determination of nitrate-nitrite nitrogen by automated
colorimetry (Revision 2.0).* Retrieved September 4, 2009 from
http://www.epa.gov/waterscience/methods/method/files/353_2.pdf

Ott, R.L and Longnecker, M.T. (2004). *A first course in statistical methods.* Belmont:
Thomson Learning, Inc.

Ott, R.L and Longnecker, M.T. (2001). *An introduction to statistical methods and data analysis, 5th edition.* Belmont: Cencage Learning.

Rundle, C. (2009). *A beginners guide to ion-selective electrode measurements.* Retrieved December 17, 2009 from http://www.nico2000.net/Book/Guide1.html

Self, J.R., and Waskom, R.M. (2008). *Nitrates in drinking water.* Retrieved May 19, 2009 from http://www.ext.colostate.edu/PUBS/crops/00517.html

Scholte, J.W., Shang, J.Q., and Rowe, R.K. (2002). Improved complex permittivity measurement and data processing technique for soil-water systems. *Geotechnical Journal, 25*(2), 187-198

Tang, N. (2009). *A real-time control system for a frequency response-based permittivity sensor.* Kansas State University.

The MathWorks, Inc. (2008a). MatLab 7.6.0 (R2008a) [Software]. Natick: The MathWorks, Inc.

The Mathworks, Inc. (2008b). Neural network toolbox (Version 6.0.1) [Software]. Natick: The Mathworks, Inc.

The Mathworks, Inc. (2008c). Statistics toolbox (Version 7.2) [Software]. Natick: The MathWorks, Inc.

The Mathworks, Inc. (2008d). Wavelet toolbox (Version 4.2) [Software]. Natick: The Mathworks, Inc.

Tokar, A.S. and Johnson, P.A. (1999). Rainfall-runoff modeling using artificial neural networks. *Journal of Hydrological Engineering 4*(3), 232-239

Topp, G.C., Zegelin, S., White, I. (2000). Impacts of the real and imaginary components of relative permittivity on time domain reflectometry measurements in soils. *Soil Science Society of America Journal, 64*(4), 1244-1252

USDA. (2006). Characterization of crude glycerol from biodiesel production from multiple feedstocks. *Biodiesel Tech Notes, 3*(3). Retrieved September 2, 2009 from http://www.uiweb.uidaho.edu/bioenergy/NewsReleases/Technote06.pdf

US Environmental Protection Agency. (2009). *National summary of impaired waters and TMDL information*. Retrieved September 9, 2009 from http://iaspub.epa.gov/waters10/attains_nation_cy.control?p_report_type=T

US Environmental Protection Agency. (2006). *Decision documents for atrazine*. Retrieved September 9, 2009 from http://www.epa.gov/oppsrrd1/REDs/atrazine_combined_docs.pdf

Van Gerpen, J.H. (2008). *Biodiesel quality.* Retrieved September 4, 2009 from http://www.deq.state.mt.us/Energy/bioenergy/Biodiesel_Production_Educ_Presentations/ 07Montana_Quality_Jan2008_JVP.pdf

Wisconsin Department of Health Services. (2008). *Atrazine.* Retrieved May 19, 2009 from http://dhs.wisconsin.gov/eh/chemfs/fs/atrazine.htm

Wise, B.M., Shaver, J.M., Gallagher, N.B., Windig, W., Bro, R., and Koch, R.S. (2006). *PLS_Toolbox 4.0 Manual*. Wenatchee: Eigenvector Research, Inc.

Appendix A - MatLab Regression Program

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Program:       Stepwise Regression & PLS & Neural Network & Wavelet    %
% Description:  To Predict Concentration of Nitrate Salts in Water       %
% Functions:     Uses Graph_Function, Stepwise_Function, PLS_Function,   %
%                Neural_Function, and Wave_Function                      %
% Probe:         7.5 cm (Big)                                            %
% Date:          3/6/09                                                  %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

clear all;
clc;

load '7_5 cm Probe Data';

Probe='7.5 cm';
Predict='Concentration';
Pred='Conc';


%%%%%%%%%%%%%%%
%  Graph Data  %
%%%%%%%%%%%%%%%

Graph_Function(freq,GAa,'G',Probe,Predict,'T',1);
Graph_Function(freq,PAa,'P',Probe,Predict,'T',2);
Graph_Function(freq,GAVa,'G',Probe,Predict,'V',3);
Graph_Function(freq,PAVa,'P',Probe,Predict,'V',4);


%%%%%%%%%%%%%%%%%%%%%%%
%  Stepwise Analysis  %
%%%%%%%%%%%%%%%%%%%%%%%

[FreqGA,R2GA,RMSEGA,PredictedGA]=Stepwise_Function(ConcT,ConcV,GA,GAV,...
    Probe,'Gain',Predict,'All',freq,5);
[FreqPA,R2PA,RMSEPA,PredictedPA]=Stepwise_Function(ConcT,ConcV,PA,PAV,...
    Probe,'Phase',Predict,'All',freq,6);
[FreqGPA,R2GPA,RMSEGPA,PredictedGPA]=Stepwise_Function(ConcT,ConcV,GPA,...
    GPAV,Probe,'Gain+Phase',Predict,'All',freq,7);
[FreqGL,R2GL,RMSEGL,PredictedGL]=Stepwise_Function(ConcL,ConcL,GL,GLV,...
    Probe,'Gain',Predict,'Low',freq,8);
[FreqPL,R2PL,RMSEPL,PredictedPL]=Stepwise_Function(ConcL,ConcL,PL,PLV,...
    Probe,'Phase',Predict,'Low',freq,9);
[FreqGPL,R2GPL,RMSEGPL,PredictedGPL]=Stepwise_Function(ConcL,ConcL,GPL,...
    GPLV,Probe,'Gain+Phase',Predict,'Low',freq,10);
[FreqGH,R2GH,RMSEGH,PredictedGH]=Stepwise_Function(ConcTH,ConcVH,GH,...
    GHV,Probe,'Gain',Predict,'High',freq,11);
[FreqPH,R2PH,RMSEPH,PredictedPH]=Stepwise_Function(ConcTH,ConcVH,PH,...
    PHV,Probe,'Phase',Predict,'High',freq,12);
[FreqGPH,R2GPH,RMSEGPH,PredictedGPH]=Stepwise_Function(ConcTH,ConcVH,...
    GPH,GPHV,Probe,'Gain+Phase',Predict,'High',freq,13);

%Make a Table of all R-Square Values for Stepwise
RSqrValue=[R2GA; R2PA; R2GPA; R2GL; R2PL; R2GPL; R2GH; R2PH; R2GPH];
```

```
%Make a Table of all RMSEs for Stepwise
RMSEValue=[RMSEGA; RMSEPA; RMSEGPA; RMSEGL; RMSEPL; RMSEGPL; ...
    RMSEGH; RMSEPH; RMSEGPH];

%Make a Table of all Predictions for Stepwise
PredictedValue=[PredictedGA PredictedPA PredictedGPA PredictedGL ...
    PredictedPL PredictedGPL PredictedGH PredictedPH PredictedGPH];

%Make Table of Frequencies Used in Each Test
L=1:60;
Freqs=zeros(60,9);
Freqs(L,1)=[sort(FreqGA);zeros(60-size(FreqGA,1),1)];
Freqs(L,2)=[sort(FreqPA);zeros(60-size(FreqPA,1),1)];
Freqs(L,3)=[sort(FreqGPA);zeros(60-size(FreqGPA,1),1)];
Freqs(L,4)=[sort(FreqGL);zeros(60-size(FreqGL,1),1)];
Freqs(L,5)=[sort(FreqPL);zeros(60-size(FreqPL,1),1)];
Freqs(L,6)=[sort(FreqGPL);zeros(60-size(FreqGPL,1),1)];
Freqs(L,7)=[sort(FreqGH);zeros(60-size(FreqGH,1),1)];
Freqs(L,8)=[sort(FreqPH);zeros(60-size(FreqPH,1),1)];
Freqs(L,9)=[sort(FreqGPH);zeros(60-size(FreqGPH,1),1)];

FreqList{1,1}=[Probe ' Probe- Frequencies Used to Predict Salt '...
    Predict ' by Stepwise'];
z={'All Gain','All Phase','All Gain+Phase','Low Gain','Low Phase',...
    'Low Gain+Phase','High Gain','High Phase','High Gain+Phase'};
for m=1:9
    FreqList{2,m}=z{1,m};
end

for q=1:60
    for s=1:9;
FreqList{q+2,s}=Freqs(q,s);
    end
end

%Make Table of All Frequencies Used in Stepwise
FreqAll=[FreqGA;FreqPA;FreqGPA;FreqGL;FreqPL;FreqGPL;FreqGH;FreqPH;FreqGPH];
FreqAll=sort(FreqAll,'ascend');


%%%%%%%%%%%%%%%%%
% PLS Analysis  %
%%%%%%%%%%%%%%%%%

[R2GA,RMSEGA,PredictedGA]=PLS_Function(ConcT,ConcV,GA,GAV,Probe,...
    'Gain',Predict,'All',14);
[R2PA,RMSEPA,PredictedPA]=PLS_Function(ConcT,ConcV,PA,PAV,Probe,...
    'Phase',Predict,'All',15);
[R2GPA,RMSEGPA,PredictedGPA]=PLS_Function(ConcT,ConcV,GPA,GPAV,Probe,...
    'Gain+Phase',Predict,'All',16);
[R2GL,RMSEGL,PredictedGL]=PLS_Function(ConcL,ConcL,GL,GLV,Probe,...
    'Gain',Predict,'Low',17);
[R2PL,RMSEPL,PredictedPL]=PLS_Function(ConcL,ConcL,PL,PLV,Probe,...
    'Phase',Predict,'Low',18);
[R2GPL,RMSEGPL,PredictedGPL]=PLS_Function(ConcL,ConcL,GPL,GPLV,Probe,...
    'Gain+Phase',Predict,'Low',19);
```

```matlab
[R2GH,RMSEGH,PredictedGH]=PLS_Function(ConcTH,ConcVH,GH,GHV,Probe,...
    'Gain',Predict,'High',20);
[R2PH,RMSEPH,PredictedPH]=PLS_Function(ConcTH,ConcVH,PH,PHV,Probe,...
    'Phase',Predict,'High',21);
[R2GPH,RMSEGPH,PredictedGPH]=PLS_Function(ConcTH,ConcVH,GPH,GPHV,...
    Probe,'Gain+Phase',Predict,'High',22);

%Add PLS R-Sqrs to Table
RSqrValue(:,3:4)=[R2GA; R2PA; R2GPA; R2GL; R2PL; R2GPL; R2GH; R2PH; R2GPH];

%Add PLS RMSEs to Table
RMSEValue(:,3:4)=[RMSEGA; RMSEPA; RMSEGPA; RMSEGL; RMSEPL;...
    RMSEGPL; RMSEGH; RMSEPH; RMSEGPH];

%Add PLS Predictions to Table
PredictedValue(:,19:36)=[PredictedGA PredictedPA PredictedGPA...
    PredictedGL PredictedPL PredictedGPL PredictedGH PredictedPH...
    PredictedGPH];


%%%%%%%%%%%%%%%%%%%%
%  Neural Network  %
%%%%%%%%%%%%%%%%%%%%

[R2GA,RMSEGA,PredictedGA]=Neural_Function(ConcT,ConcV,GA,GAV,Probe,...
    'Gain',Predict,'All',23);
[R2PA,RMSEPA,PredictedPA]=Neural_Function(ConcT,ConcV,PA,PAV,Probe,...
    'Phase',Predict,'All',24);
[R2GPA,RMSEGPA,PredictedGPA]=Neural_Function(ConcT,ConcV,GPA,GPAV,...
    Probe,'Gain+Phase',Predict,'All',25);
[R2GL,RMSEGL,PredictedGL]=Neural_Function(ConcL,ConcL,GL,GLV,Probe,...
    'Gain',Predict,'Low',26);
[R2PL,RMSEPL,PredictedPL]=Neural_Function(ConcL,ConcL,PL,PLV,Probe,...
    'Phase',Predict,'Low',27);
[R2GPL,RMSEGPL,PredictedGPL]=Neural_Function(ConcL,ConcL,GPL,GPLV,...
    Probe,'Gain+Phase',Predict,'Low',28);
[R2GH,RMSEGH,PredictedGH]=Neural_Function(ConcTH,ConcVH,GH,GHV,...
    Probe,'Gain',Predict,'High',29);
[R2PH,RMSEPH,PredictedPH]=Neural_Function(ConcTH,ConcVH,PH,PHV,...
    Probe,'Phase',Predict,'High',30);
[R2GPH,RMSEGPH,PredictedGPH]=Neural_Function(ConcTH,ConcVH,GPH,GPHV,...
    Probe,'Gain+Phase',Predict,'High',31);

%Add ANN R-Sqrs to Table
RSqrValue(:,5:6)=[R2GA; R2PA; R2GPA; R2GL; R2PL; R2GPL; R2GH; R2PH; R2GPH];

%Add ANN RMSEs to Table
RMSEValue(:,5:6)=[RMSEGA; RMSEPA; RMSEGPA; RMSEGL; RMSEPL; ...
    RMSEGPL; RMSEGH; RMSEPH; RMSEGPH];

%Add ANN Predictions to Table
PredictedValue(:,37:54)=[PredictedGA PredictedPA PredictedGPA ...
    PredictedGL PredictedPL PredictedGPL PredictedGH PredictedPH ...
    PredictedGPH];
```

```
%%%%%%%%%%%%%%%%%%%%%
% Wavelet Transform %
%%%%%%%%%%%%%%%%%%%%%


[R2GA,RMSEGA,PredictedGA,SeGA,LGA]=Wave_Function(ConcT,ConcV,GA,GAV,...
    Probe,'Gain',Predict,'All',freq,32);
[R2PA,RMSEPA,PredictedPA,SePA,LPA]=Wave_Function(ConcT,ConcT,PA,PAV,...
    Probe,'Phase',Predict,'All',freq,34);
[R2GPA,RMSEGPA,PredictedGPA,SeGPA,LGPA]=Wave_Function(ConcT,ConcV,GPA,...
    GPAV,Probe,'Gain+Phase',Predict,'All',freq,36);
[R2GL,RMSEGL,PredictedGL,SeGL,LGL]=Wave_Function(ConcL,ConcL,GL,GLV,...
    Probe,'Gain',Predict,'Low',freq,38);
[R2PL,RMSEPL,PredictedPL,SePL,LPL]=Wave_Function(ConcL,ConcL,PL,PLV,...
    Probe,'Phase',Predict,'Low',freq,40);
[R2GPL,RMSEGPL,PredictedGPL,SeGPL,LGPL]=Wave_Function(ConcL,ConcL,GPL,...
    GPLV,Probe,'Gain+Phase',Predict,'Low',freq,42);
[R2GH,RMSEGH,PredictedGH,SeGH,LGH]=Wave_Function(ConcTH,ConcVH,GH,...
    GHV,Probe,'Gain',Predict,'High',freq,44);
[R2PH,RMSEPH,PredictedPH,SePH,LPH]=Wave_Function(ConcTH,ConcVH,PH,...
    PHV,Probe,'Phase',Predict,'High',freq,46);
[R2GPH,RMSEGPH,PredictedGPH,SeGPH,LGPH]=Wave_Function(ConcTH,ConcVH,...
    GPH,GPHV,Probe,'Gain+Phase',Predict,'High',freq,48);

%Add Wavelet R-Sqrs to Table
RSqrValue(:,7:8)=[R2GA; R2PA; R2GPA; R2GL; R2PL; R2GPL; R2GH; R2PH; R2GPH];

%Add Wavelet RMSEs to Table
RMSEValue(:,7:8)=[RMSEGA; RMSEPA; RMSEGPA; RMSEGL; RMSEPL; ...
    RMSEGPL; RMSEGH; RMSEPH; RMSEGPH];

%Add Wavelet Predictions to Table
PredictedValue(:,55:72)=[PredictedGA PredictedPA PredictedGPA ...
    PredictedGL PredictedPL PredictedGPL PredictedGH PredictedPH ...
    PredictedGPH];

%Make Table of Selected Wavebands
Bands{1,1} = [Probe ' Probe- Predicting ' Predict ...
    ' of Salts- Wavebands Selected From Wavelet Transform'];
Bands{2,2} = 'Level';
Bands{2,3} = 'Waveband';
Bands{3,1} = 'All Gain';

Bands1=SeGA;
Bands{length(Bands1)+3,1} = 'All Phase';
Bands1((length(Bands1)+1):(length(Bands1)+LPA),:)=SePA;
Bands{length(Bands1)+3,1} = 'All Gain+Phase';
Bands1((length(Bands1)+1):(length(Bands1)+LGPA),:)=SeGPA;
Bands{length(Bands1)+3,1} = 'Low Gain';
Bands1((length(Bands1)+1):(length(Bands1)+LGL),:)=SeGL;
Bands{length(Bands1)+3,1} = 'Low Phase';
Bands1((length(Bands1)+1):(length(Bands1)+LPL),:)=SePL;
Bands{length(Bands1)+3,1} = 'Low Gain+Phase';
Bands1((length(Bands1)+1):(length(Bands1)+LGPL),:)=SeGPL;
Bands{length(Bands1)+3,1} = 'High Gain';
Bands1((length(Bands1)+1):(length(Bands1)+LGH),:)=SeGH;
Bands{length(Bands1)+3,1} = 'High Phase';
Bands1((length(Bands1)+1):(length(Bands1)+LPH),:)=SePH;
```

102

```
Bands{length(Bands1)+3,1} = 'High Gain+Phase';
Bands1((length(Bands1)+1):(length(Bands1)+LGPH),:)=SeGPH;

for n=1:length(Bands1)
    for m=1:2
    Bands{n+2,m+1}=Bands1(n,m);
    end
end

%Make Table of All RMSEs
RMSE{1,1} = [Probe ' Probe- Predicting ' Predict ' of Salts- RMSE'];
RMSE{2,2} = 'SWR';
RMSE{2,4} = 'PLS';
RMSE{2,6} = 'ANN';
RMSE{2,8} = 'Wavelet';
RMSE{3,2} = 'Training';
RMSE{3,3} = 'Validation';
RMSE{3,4} = 'Training';
RMSE{3,5} = 'Validation';
RMSE{3,6} = 'Training';
RMSE{3,7} = 'Validation';
RMSE{3,8} = 'Training';
RMSE{3,9} = 'Validation';
RMSE{4,1} = 'All Gain';
RMSE{5,1} = 'All Phase';
RMSE{6,1} = 'All Gain+Phase';
RMSE{7,1} = 'Low Gain';
RMSE{8,1} = 'Low Phase';
RMSE{9,1} = 'Low Gain+Phase';
RMSE{10,1} = 'High Gain';
RMSE{11,1} = 'High Phase';
RMSE{12,1} = 'High Gain+Phase';
for N=1:9
    for M=1:8
        RMSE{N+3,M+1}=RMSEValue(N,M);
    end
end

%Make Table of all R-Sqr Values
RSqr=RMSE;
RSqr{1,1} = [Probe ' Probe- Predicting ' Predict ' of Salts- R^2 Values'];

for N=1:9
    for M=1:8
        RSqr{N+3,M+1}=RSqrValue(N,M);
    end
end

%Make Table of all Prediction Values
z={'All Gain',' ','All Phase',' ','All Gain+Phase',' ','Low Gain',...
    ' ','Low Phase',' ','Low Gain+Phase',' ','High Gain',' ',...
    'High Phase',' ','High Gain+Phase',' '};
Prediction{1,1} = [Probe ' Probe- Predicting ' Predict...
    ' of Salts- Predictions'];
Prediction{2,1} = 'Stepwise Regression';
Prediction{2,19} = 'PLS';
Prediction{2,37} = 'Neural Network';
```

```
Prediction{2,55} = 'Wavelet';

for n=1:18
    Prediction{3,n}=z{1,n};
    Prediction{3,n+18}=z{1,n};
    Prediction{3,n+36}=z{1,n};
    Prediction{3,n+54}=z{1,n};
end

for n=1:36
    Prediction{4,2*n-1}='Training';
    Prediction{4,2*n}='Validation';
end

for N=1:72
    for M=1:27
        Prediction{M+3,N}=PredictedValue(M,N);
    end
end

%Write Tables to Excel
xlswrite([Probe '-' Pred '- R-Sqr Values.xls'], RSqr);
xlswrite([Probe '-' Pred '- RMSE Values.xls'], RMSE);
xlswrite([Probe '-' Pred '- Predicted Values.xls'],Prediction);
xlswrite([Probe '-' Pred '- All Frequencies Used.xls'], FreqAll);
xlswrite([Probe '-' Pred '- Wavebands.xls'], Bands);
xlswrite([Probe '-' Pred '- Frequencies Used for Each Test.xls'], FreqList);
```

Appendix B - Stepwise Program

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Program:   Stepwise Regression Function                            %
% Toolbox:   Mathworks Statistics Toolbox 7.0                        %
% Purpose:   To predict concentration or molecular weight of pollutants  %
%            in water samples                                        %
% By:        Sarah Shultz                                            %
% Date:      3/6/09                                                  %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

function [FreqList,R2,RMSE,Predicted]=Stepwise_Function(YT,...
    YV,XT,XV,Probe,Data,Prediction,Concentration,...
    freq,FigNum)
    %Inputs-
        %YT = training Y-variables
        %YV = validation Y-variables
        %XT = training X-variables
        %XV = validation X-variables
        %Probe = probe that data was collected from
        %Data = gain, phase, or gain and phase together
        %Prediction = concentration or molecular weight
        %Concentration = range of concentrations included
        %freq = list of frequencies corresponding to the X-variables
        %FigNum = number to assign to plot

    %Outputs-
        %FreqList = list of significant frequencies
        %R2 = R-squared values
        %RMSE = RMSE value
        %Predicted = Y-values predicted using stepwise model

%List of Frequencies
if size(XT,2)>size(freq,2);
    allfreq=[freq;freq]';
else allfreq=freq';
end

%Remove Frequencies from Data
R=1:5:length(XT);
XT=XT(:,R);
XV=XV(:,R);
allfreq=allfreq(:,R);


%Stepwise Model
[b,se,pval,inmodel,stats] = stepwisefit(XT,YT,...
    'penter',0.05,'premove',0.10);

%Make list of frequencies stepwise selects
X=allfreq.*inmodel;
Freq=X(X ~= 0);
FreqList=Freq';
```

```
%Training
T=XT';
n=size(XT,1);
for N=1:n
    Q(1:length(b),N)=b.*inmodel'.*T(1:length(b),N);
end
T=sum(Q)+stats.intercept;
T=T';
clear 'Q';

%Validate
V=XV';
n=size(XV,1);
    for N=1:n
     Q(1:length(b),N)=b.*inmodel'.*V(1:length(b),N);
    end
V=sum(Q)+stats.intercept;
V=V';
Predicted=[[T; zeros(27-size(T,1),1)] [V; zeros(27-size(V,1),1)]];

%Calculate R-Square Values
Sxy=sum(T.*YT)-((sum(T)*sum(YT))/length(T));
Sxx=sum(T.^2)-(sum(T)^2/length(T));
Syy=sum(YT.^2)-(sum(YT)^2/length(YT));
R2T=(Sxy/sqrt(Sxx*Syy))^2;

Sxy=sum(V.*YV)-((sum(V)*sum(YV))/length(V));
Sxx=sum(V.^2)-(sum(V)^2/length(V));
Syy=sum(YV.^2)-(sum(YV)^2/length(YV));
R2V=(Sxy/sqrt(Sxx*Syy))^2;

R2=[R2T R2V];

%Calculate RMSE for Training & Validation
SSET=sum((YT-T).^2);
RMSET=sqrt(SSET/(size(YT,1)-2));

SSEV=sum((YV-V).^2);
RMSEV=sqrt(SSEV/(size(YV,1)-2));

RMSE=[RMSET RMSEV];

%Title of Plot
Title1=[Probe ' Probe- Predicting ' Prediction ' with Stepwise Analysis'];
Title2=[Concentration ' ' Data];
Title3=['Training R^2=' mat2str(R2T,3) ', Validation R^2=' mat2str(R2V,3)];
Title4=['Training RMSE=' mat2str(RMSET,3)...
    ', Validation RMSE=' mat2str(RMSEV,3)];

%Plot of actual vs. predicted values
h=figure(FigNum);
scatter(YT,T,'bo')
hold('all');
scatter(YV,V,'go')
plot([YV YT],[YV YT],'r')
xlabel(['Actual ' Prediction]);
```

106

```
ylabel(['Predicted ' Prediction]);
title({Title1;Title2;Title3;Title4}, 'fontsize', 13);
grid on;
legend('Training','Validation','Perfect Fit','Location','NorthWest')
z=[Prediction '- ' mat2str(FigNum) '- ' Probe '-  Stepwise.bmp'];

%Save plot as a bitmap
saveas(h,z)
```

Appendix C - PLS Program

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Program:   PLS Function                                              %
% Toolbox:   Eigenvector PLS Toolbox 4.0                               %
% Purpose:   To predict concentration or molecular weight of pollutants %
%            in water samples                                          %
% By:        Sarah Shultz                                              %
% Date:      3/6/09                                                    %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

function [R2,RMSE,Predicted]=PLS_Function(YT,YV,XT,...
    XV,Probe,Data,Prediction,Concentration,FigNum)
    %Inputs-
        %YT = training Y-variables
        %YV = validation Y-variables
        %XT = training X-variables
        %XV = validation X-variables
        %Probe = probe that data was collected from
        %Data = gain, phase, or gain and phase together
        %Prediction = concentration or molecular weight
        %Concentration = range of concentrations included
        %FigNum = number to assign to plot

    %Outputs-
        %R2 = R-squared values
        %RMSE = RMSE values
        %Predicted = Y-values predicted using PLS model

%Remove Frequencies from Data
R=1:5:length(XT);
XT=XT(:,R);
XV=XV(:,R);

%Preprocess Data
max_pc=size(XT,1); %Number of Y-variables
[press,cumpress]=crossval(XT,YT,'sim',{'loo'},max_pc);
    %sim = SIMPLS algorithm
    %loo = leave-one-out cross-validation

min_train_pc=find(cumpress==min(cumpress));
    %find minimum CUMPRESS value

%Model options
options.name='options';
options.display='off';
options.plots='none';
options.outputversion=3;
options.preprocessing={preprocess('meancenter') preprocess('meancenter')};
    %center columns to have zero mean
options.algorithm='sim';
    %Use SIMPLS algorithm
options.blockdetails='standard';

%Create model with training data
```

```
model=pls(XT,YT,min_train_pc(1,1),options); %calibration model

%Prediction using the training data
pred=pls(XT,model,options);
predYT=pred.pred{2};

%Prediction using validation data
pred=pls(XV,YV,model,options);
predYV=pred.pred{2};

%Put prediction values together into one matrix to save
Predicted=[[predYT; zeros(27-size(predYT,1),1)] ...
    [predYV; zeros(27-size(predYV,1),1)]];

%Calculate R-Square values for training & validation
Sxy=sum(predYT.*YT)-((sum(predYT)*sum(YT))/length(predYT));
Sxx=sum(predYT.^2)-(sum(predYT)^2/length(predYT));
Syy=sum(YT.^2)-(sum(YT)^2/length(YT));
R2T=(Sxy/sqrt(Sxx*Syy))^2;

Sxy=sum(predYV.*YV)-((sum(predYV)*sum(YV))/length(predYV));
Sxx=sum(predYV.^2)-(sum(predYV)^2/length(predYV));
Syy=sum(YV.^2)-(sum(YV)^2/length(YV));
R2V=(Sxy/sqrt(Sxx*Syy))^2;

R2=[R2T R2V];


%Calculate RMSE for Training & Validation
SSET=sum((YT-predYT).^2);
RMSET=sqrt(SSET/(size(YT,1)-2));

SSEV=sum((YV-V).^2);
RMSEV=sqrt(SSEV/(size(predYV,1)-2));

RMSE=[RMSET RMSEV];


%Title of Plot
Title1=[Probe ' Probe- Predicting ' Prediction ' with PLS'];
Title2=[Concentration ' ' Data ];
Title3=['Training R^2=' mat2str(R2T,3) ', Validation R^2=' mat2str(R2V,3)];
Title4=['Training RMSE=' mat2str(RMSET,3) ...
    ', Validation RMSE=' mat2str(RMSEV,3)];

%Plot of actual vs. predicted values
h=figure(FigNum);
plot(YT,predYT,'bo');
hold on;
plot(YV,predYV,'go');
plot([YV YT],[YV YT],'r');
xlabel(['Actual ' Prediction]);
ylabel(['Predicted ' Prediction]);
title({Title1;Title2;Title3;Title4}, 'fontsize', 13);
grid on;
legend('Training','Validation','Perfect Fit','Location','NorthWest')
hold off;
```

```
%Save plot as a bitmap
z=[Prediction '- ' mat2str(FigNum) '- ' Probe '-  PLS.bmp'];
saveas(h,z)
```

Appendix D - ANN Program

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Program:   Neural Network Function                                    %
% Toolbox:   Mathworks Neural Network Toolbox 6.0.1                     %
% Purpose:   To predict concentration or molecular weight of pollutants %
%            in water samples                                           %
% By:        Sarah Shultz                                               %
% Date:      3/6/09                                                     %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

function [R2,RMSE,Predicted]=Neural_Function(YT,YV,...
    XT,XV,Probe,Data,Prediction,Concentration,FigNum)
    %Inputs-
        %YT = training Y-variables
        %YV = validation Y-variables
        %XT = training X-variables
        %XV = validation X-variables
        %Probe = probe that data was collected from
        %Data = gain, phase, or gain and phase together
        %Prediction = concentration or molecular weight
        %Concentration = range of concentrations included
        %FigNum = number to assign to plot

    %Outputs-
        %R2 = R-squared values
        %RMSE = RMSE values
        %Predicted = Y-values predicted using the ANN model

%Remove Frequencies from Data
R=1:5:length(XT);
XT=XT(:,R);
XV=XV(:,R);

%Transpose matrices to be in correct order
XT = XT';
XV = XV';
YT = YT';
YV = YV';

%Remove Constant Rows
[pT1,PS] = removeconstantrows(XT);
pV1 = XV(PS.keep,:);

%Normalize Inputs and Targets
[normInputT] = mapminmax(pT1);
[normInputV] = mapminmax(pV1);
[normYT,ts] = mapminmax(YT);
[normYV] = mapminmax(YV);

%Create Network
numHiddenNeurons = 10;  %Adjust as desired
numOutputs = size(YT,1);
```

```
net = newff(minmax(normInputT),[numHiddenNeurons,numOutputs]);

%Divide up Samples
testPercent = .2;
validatePercent = .2;
[trainSamples,validateSamples,testSamples] = ...
    dividevec(normInputT,normYT,testPercent,validatePercent);

%Train Network
[net] = train(net,trainSamples.P,trainSamples.T,[],[],...
    validateSamples,testSamples);

 %Simulate Network
[normTrainOutput] = sim(net,trainSamples.P,[],[],trainSamples.T);
[normTrainOutputT] = sim(net,normInputT,[],[],normYT);
[normValidateOutputV] = sim(net,normInputV,[],[],normYV);

%Reverse Normalize Outputs
trainOutput = mapminmax('reverse',normTrainOutput,ts);
trainOutputT = mapminmax('reverse',normTrainOutputT,ts);
validateOutputV = mapminmax('reverse',normValidateOutputV,ts);
Predicted=[[trainOutputT'; zeros(27-size(trainOutputT,2),1)]...
    [validateOutputV'; zeros(27-size(validateOutputV,2),1)]];

%Calculate R-Sqr Value for Training & Validation
Sxy=sum(trainOutputT.*YT)-((sum(trainOutputT)*sum(YT))...
    /length(trainOutputT));
Sxx=sum(trainOutputT.^2)-(sum(trainOutputT)^2/length(trainOutputT));
Syy=sum(YT.^2)-(sum(YT)^2/length(YT));
R2T=(Sxy/sqrt(Sxx*Syy))^2;

Sxy=sum(validateOutputV.*YV)-((sum(validateOutputV)*sum(YV))...
    /length(validateOutputV));
Sxx=sum(validateOutputV.^2)-(sum(validateOutputV)^2/...
    length(validateOutputV));
Syy=sum(YV.^2)-(sum(YV)^2/length(YV));
R2V=(Sxy/sqrt(Sxx*Syy))^2;

R2=[R2T R2V];

%Calculate RMSE for Training & Validation
SSET=sum((YT- trainOutputT').^2);
RMSET=sqrt(SSET/(size(YT,1)-2));

SSEV=sum((YV- trainOutputV').^2);
RMSEV=sqrt(SSEV/(size(YV,1)-2));

RMSE=[RMSET RMSEV];

%Title of Plot
Title1=[Probe ' Probe- Predicting ' Prediction ' with Neural Network'];
Title2=[Concentration ' ' Data];
Title3=['Training R^2=' mat2str(R2T,3) ', Validation R^2=' mat2str(R2V,3)];
Title4=['Training RMSE=' mat2str(RMSET,3)...
    ', Validation RMSE=' mat2str(RMSEV,3)];
```

```
%Plot of actual vs. predicted values
h=figure(FigNum);
scatter(YT,trainOutputT,'bo')
hold('all');
scatter(YV,validateOutputV,'go')
plot([YT YV],[YT YV],'r')
xlabel(['Actual ' Prediction]);
ylabel(['Predicted ' Prediction]);
title({Title1;Title2;Title3;Title4}, 'fontsize', 13);
legend('Training','Validation','Perfect Fit','Location','NorthWest')
grid on;

%Save plot as a bitmap
z=[Prediction '- ' mat2str(FigNum) '- ' Probe '- Neural.bmp'];
saveas(h,z)
```

Appendix E - Wavelet Program

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Program:   Wavelet Function                                           %
% Toolbox:   Mathworks Wavelet Toolbox 4.3                              %
% Purpose:   To predict concentration or molecular weight of pollutants %
%            in water samples                                           %
% By:        Sarah Shultz                                               %
% Date:      3/6/09                                                     %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

function [R2,RMSE,Predicted,Bands,Length]=Wave_Function(YT,YV,...
    XT1,XV1,Probe,Data,Prediction,Concentration,freq,FigNum)
      %Inputs-
        %YT = training Y-variables
        %YV = validation Y-variables
        %XT1 = training X-variables
        %XV1 = validation X-variables
        %Probe = probe that data was collected from
        %Data = gain, phase, or gain and phase together
        %Prediction = concentration or molecular weight
        %Concentration = range of concentrations included
        %freq = list of frequencies
        %FigNum = number to assign to plot


    %Outputs-
        %R2 = R-squared values
        %RMSE = RMSE values
        %Predicted = Y-values predicted using PLS model
        %Bands = levels & bands selected by stepwise
        %Length = number of bands selected



 %Label concentration or molecular weight
 if strcmp(Prediction,'Molecular Weight')==1
    K='MW';
 else K='Conc';
 end

 %Determine the level of transform depending on data size
 Level2=floor(log(length(XT1))/log(2));
 Level1=Level2-4;

 %Transform training and validation data
n=size(XT1,1);
for N=1:n
    [C(:,N),L(:,N)] = wavedec(XT1(N,:),Level2,'haar');
end

n=size(XV1,1);
for N=1:n
```

```
        [CV(:,N),LV(:,N)] = wavedec(XV1(N,:),Level2,'haar');
end


XT=C';
XV=CV';
L=L';
L=L(1,:);


%Select Only 5 of the Levels
J=1:sum(L(1,1:6),2);
XT=XT(:,J);
XV=XV(:,J);


%Stepwise Model
[b,se,pval,inmodel,stats] = stepwisefit(XT,YT,...
    'penter',0.05,'premove',0.10);


%XT
T=XT';
n=size(XT,1);
for N=1:n
    Q(1:length(b),N)=b.*inmodel'.*T(1:length(b),N);
end
T=sum(Q)+stats.intercept;
T=T';
clear 'Q';


% Validate
V=XV';
n=size(XV,1);
    for N=1:n
    Q(1:length(b),N)=b.*inmodel'.*V(1:length(b),N);
    end
V=sum(Q)+stats.intercept;
V=V';
Predicted=[[T; zeros(27-size(T,1),1)] [V; zeros(27-size(V,1),1)]];


%Calculate R-Square Values
Sxy=sum(T.*YT)-((sum(T)*sum(YT))/length(T));
Sxx=sum(T.^2)-(sum(T)^2/length(T));
Syy=sum(YT.^2)-(sum(YT)^2/length(YT));
R2T=(Sxy/sqrt(Sxx*Syy))^2;


Sxy=sum(V.*YV)-((sum(V)*sum(YV))/length(V));
Sxx=sum(V.^2)-(sum(V)^2/length(V));
Syy=sum(YV.^2)-(sum(YV)^2/length(YV));
R2V=(Sxy/sqrt(Sxx*Syy))^2;


R2=[R2T R2V];


%Calculate RMSE for Training & Validation
SSET=sum((YT-T).^2);
RMSET=sqrt(SSET/(size(YT,1)-2));


SSEV=sum((YV-V).^2);
```

```matlab
RMSEV=sqrt(SSEV/(size(YV,1)-2));


RMSE=[RMSET RMSEV];


%Make Sure Correct Labels are on Plot
if inmodel(1,1:size(inmodel,2))==0
    R2T=0;
    R2V=0;
    RMSET=9999;
    RMSEV=9999;
else
end;


%Title of Plot
Title1=[Probe ' Probe- Level ' mat2str(Level1) ' to ' ...
    mat2str(Level2) ' Haar Wavelet Transform'];
Title2=['Predicting ' Prediction ' with Wavelet Transform and Stepwise'];
Title3=[Concentration ' ' Data];
Title4=['Training R^2=' mat2str(R2T,3) ', Validation R^2=' mat2str(R2V,3)];
Title5=['Training RMSE=' mat2str(RMSET,3)...
    ', Validation RMSE=' mat2str(RMSEV,3)];


%Plot actual vs. predicted values
h=figure(FigNum);
scatter(YT,T,'bo')
hold('all');
scatter(YV,V,'go')
plot([YV YT],[YV YT],'r')
xlabel(['Actual ' Prediction]);
ylabel(['Predicted ' Prediction]);
title({Title1;Title2;Title3;Title4;Title5}, 'fontsize', 13);
grid on;
legend('Training','Validation','Perfect Fit','Location','NorthWest')


%Save the Plot
z=[mat2str(FigNum) '- ' K '-  ' Probe ' Probe- Level ' mat2str(Level1)...
    ' to ' mat2str(Level2) ' Predict.bmp'];
%saveas(h,z)


if inmodel(1,1:size(inmodel,2))==0
    Bands=[0,0];
    R2=[0 0];
    RMSE=[9999 9999];
    clear 'Predicted';
    Predicted=[zeros(27,1) zeros(27,1)];
    Length=1;
else
%Selected Frequencies
Selected=inmodel.*(1:length(inmodel));
Selected=Selected(Selected ~= 0);


%Make Table of Levels and Wavebands Selected
for M=1:size(Selected,2)
        switch logical(true)
            case Selected(1,M)<=L(1,1)
```

```
                Bands(M,1:2)=[Level2 Selected(1,M)];
            case Selected(1,M)>L(1,1) && Selected(1,M)<=sum(L(1,1:2))
                Bands(M,1:2)=[Level2 (Selected(1,M)-L(1,1))];
            case Selected(1,M)>sum(L(1,1:2)) && Selected(1,M)<=sum(L(1,1:3))
                    Bands(M,1:2)=[(Level2-1) ...
                        (Selected(1,M)-sum(L(1,1:2)))];
            case Selected(1,M)>sum(L(1,1:3)) && Selected(1,M)<=sum(L(1,1:4))
                    Bands(M,1:2)=[(Level2-2) (Selected(1,M)-...
                        sum(L(1,1:3)))];
            case Selected(1,M)>sum(L(1,1:4)) && Selected(1,M)<=sum(L(1,1:5))
                        Bands(M,1:2)=[(Level2-3) (Selected(1,M)-...
                            sum(L(1,1:4)))];
            case Selected(1,M)>sum(L(1,1:5)) && Selected(1,M)<=sum(L(1,1:6))
                            Bands(M,1:2)=[(Level2-4) (Selected(1,M)-...
                                sum(L(1,1:5)))];
        end
end


Length=length(Selected);


clear 'h';


%Make Tiling Diagram using Rectangle Annotations
%figure(FigNum+1)=figure;
h=figure(FigNum+1);


Z=L(1,6)/mean(L(1,6:7));
for M=0:.15:.6
    annotation(figure(FigNum+1),'rectangle',[0.1 0.15+M Z .15]);
end


for N=0:(L(1,1)-2)
        annotation(figure(FigNum+1),'rectangle',[0.1+N*Z/L(1,1) ...
            0.15 Z/L(1,1) .15]);
end


for N=0:(L(1,3)-2)
        annotation(figure(FigNum+1),'rectangle',[0.1+N*Z/L(1,3) ...
            0.3 Z/L(1,3) .15]);
end


for N=0:(L(1,4)-2)
        annotation(figure(FigNum+1),'rectangle',[0.1+N*Z/L(1,4) ...
            0.45 Z/L(1,4) .15]);
end


for N=0:(L(1,5)-2)
        annotation(figure(FigNum+1),'rectangle',[0.1+N*Z/L(1,5) ...
            0.6 Z/L(1,5) .15]);
end


for N=0:(L(1,6)-2)
        annotation(figure(FigNum+1),'rectangle',[0.1+N*Z/L(1,6) ...
            0.75 Z/L(1,6) .15]);
end
```

117

```matlab
%Convert Table of Selected Wavebands to the Vector needed to Color the
%Correct Bands in the Tiling Diagram
M=0:.15:.6;
for N=1:size(Bands,1)
    Color(N,:)=[(0.1+(Bands(N,2)-1)*Z/L(1,(length(L)-(Bands(N,1)))))...
        (0.15+M(1,(length(L)-Bands(N,1)-1)))...
        (Z/L(1,length(L)-(Bands(N,1)))) .15];
end

%Color the Rectangles
for N=1:size(Color,1)
    annotation(figure(FigNum+1),'rectangle','FaceColor',[0 1 .6],...
        'Position',Color(N,:));
end

%Add Level Numbers Along the Side of the Tiling Diagram
Q=Level2:-1:Level1;
for N=1:length(Q)
    annotation(figure(FigNum+1),'textbox','String',mat2str(Q(1,N)),...
        'HorizontalAlignment','right','VerticalAlignment','middle',...
        'FitBoxToText','off','Position',[.1 .15+M(1,N) 0 .15]);
end

%Add Chart Title and Label Axes
annotation(figure(FigNum+1),'textbox','String',{'L','e','v','e','l'},...
    'HorizontalAlignment','center','VerticalAlignment','middle',...
    'FontWeight','bold','FitBoxToText','on','LineStyle','none',...
    'Position',[.03 .15 0 .75]);
annotation(figure(FigNum+1),'textbox','String',{'Frequency (Hz)'},...
    'HorizontalAlignment','center','FitBoxToText','off','FontWeight',...
    'bold','LineStyle','none','Position',[.1 .05 Z 0]);
annotation(figure(FigNum+1),'textbox','String',...
    {['Tiling Chart- Predicting ' K ' with ' Probe ' Probe '...
    Concentration ' ' Data]},'HorizontalAlignment','center',...
    'VerticalAlignment','top','fontsize',12,'FitBoxToText','off',...
    'FontWeight','bold','LineStyle','none','Position',[.1 1 Z 0]);

%Add Highest and Lowest Frequencies to Table
annotation(figure(FigNum+1),'textbox','String',mat2str(freq(1,1)),...
    'HorizontalAlignment','center','LineStyle','none','Position',...
    [.1 .14 0 0]);
annotation(figure(FigNum+1),'textbox','String',...
    mat2str(freq(length(freq),1)),'HorizontalAlignment','center',...
    'LineStyle','none','Position',[.1+Z .14 0 0]);

%Add Additional Frequencies Depending on whether Gain+Phase data is used or
%just Gain or Phase
if strcmp(Data,'Gain+Phase')==1
    annotation(figure(FigNum+1),'textbox','String',mat2str(freq(1,1)),...
        'HorizontalAlignment','center','LineStyle','none','Position',...
        [.1+(Z/2) .14 0 0]);
    annotation(figure(FigNum+1),'textbox','String',{'Gain'},...
        'HorizontalAlignment','center','LineStyle','none','FontWeight',...
        'demi','Position',[.1 .1 Z/2 0]);
    annotation(figure(FigNum+1),'textbox','String',{'Phase'},...
```

```matlab
            'HorizontalAlignment','center','LineStyle','none','FontWeight',...
            'demi','Position',[.1+(Z/2) .1 Z/2 0]);
    else
        annotation(figure(FigNum+1),'textbox','String',mat2str(freq(floor...
            (length(freq)/2),1)),'HorizontalAlignment','center','LineStyle',...
            'none','Position',[.1+(Z/2) .14 0 0]);
    end


    %Save Tiling Chart
    z=[mat2str(FigNum+1) '- ' K '-  ' Probe ' Probe- Level ' ...
        mat2str(Level1) ' to ' mat2str(Level2) ' Tiling.bmp'];
    %saveas(h,z)
    end;
```

Appendix F - Frequencies Selected by Stepwise to Predict 1 L Nitrate Concentration with Old Control Box or Impedance Meter

Frequency (Hz)	# of Times Selected	Frequency (Hz)	# of Times Selected	Frequency (Hz)	# of Times Selected
50	1	19,800,000	2	69,800,000	1
300	3	21,800,000	1	70,200,000	2
550	1	22,000,000	1	71,200,000	1
800	3	22,200,000	1	74,200,000	1
1,413	3	23,000,000	2	76,000,000	1
7,943	3	26,200,000	1	76,800,000	1
44,668	2	27,200,000	1	77,200,000	2
251,189	3	28,000,000	1	78,200,000	1
1,000,000	4	28,200,000	2	79,200,000	1
1,200,000	1	31,200,000	1	80,200,000	1
2,000,000	2	32,000,000	1	80,800,000	1
2,200,000	2	35,200,000	1	81,200,000	1
3,200,000	1	37,200,000	1	83,200,000	2
3,800,000	1	42,200,000	1	84,000,000	1
4,000,000	2	43,200,000	1	84,200,000	2
4,200,000	1	46,000,000	1	84,800,000	1
5,000,000	1	48,800,000	1	89,000,000	1
5,800,000	2	50,000,000	1	89,200,000	1
6,000,000	2	50,800,000	1	92,000,000	2
6,200,000	1	52,000,000	1	94,200,000	1
7,000,000	1	52,200,000	1	99,800,000	1
8,000,000	1	53,200,000	1	103,800,000	1
8,200,000	1	55,800,000	1	105,000,000	1
9,200,000	2	56,000,000	2	108,000,000	1
9,800,000	1	57,000,000	2	108,200,000	1
10,000,000	1	57,800,000	1	110,000,000	1
10,200,000	2	58,800,000	1	111,000,000	1
10,800,000	1	59,200,000	1	111,200,000	2
11,200,000	3	59,800,000	1	113,200,000	1
11,800,000	1	60,200,000	2	115,200,000	2
12,200,000	1	60,800,000	1	116,800,000	1
13,800,000	1	61,200,000	1	118,200,000	2
14,000,000	1	62,000,000	1		
16,000,000	1	62,200,000	2		
16,200,000	1	64,800,000	1		
18,000,000	1	65,200,000	1		
18,200,000	1	68,000,000	1		
19,000,000	1	69,000,000	1		
19,200,000	1	69,200,000	1		

Appendix G - Frequencies Selected by Stepwise to Predict 1 L Nitrate Concentration with New Control Box and 2 cm Probe

Frequency (Hz)	# of Times Selected
5,600,000	1
9,600,000	1
9,600,000	1
13,600,000	1
18,400,000	1
26,400,000	1
37,600,000	1
61,600,000	2
65,600,000	2
109,600,000	2
117,600,000	1
129,600,000	1
133,600,000	1
134,400,000	1
158,400,000	1
162,400,000	1
174,400,000	1
181,600,000	5
186,400,000	1
189,600,000	1
201,600,000	1
202,400,000	1
205,600,000	1
229,600,000	1
241,600,000	2
245,600,000	1
246,400,000	2
257,600,000	1
261,600,000	1
273,600,000	1
309,600,000	1
313,600,000	1
322,400,000	1
326,400,000	1
330,400,000	1
345,600,000	1
357,600,000	1
365,600,000	1
366,400,000	1
377,600,000	2
381,600,000	1
398,400,000	1

Appendix H – Frequencies Selected by Stepwise to Predict 36 L Nitrate Concentration with New Control Box

Frequency (Hz)	# of Times Selected	Frequency (Hz)	# of Times Selected	Frequency (Hz)	# of Times Selected
200	6	102,400,000	2	241,600,000	1
400	1	105,600,000	4	246,400,000	1
1,413	2	109,600,000	2	249,600,000	1
7,943	1	118,400,000	1	253,600,000	3
44,668	3	121,600,000	2	254,400,000	1
63,096	2	122,400,000	1	257,600,000	2
251,189	1	125,600,000	4	261,600,000	1
1,600,000	6	129,600,000	2	262,400,000	1
2,400,000	1	130,400,000	2	265,600,000	3
5,600,000	1	133,600,000	6	266,400,000	1
6,400,000	1	134,400,000	1	269,600,000	5
9,600,000	5	137,600,000	3	273,600,000	1
10,400,000	2	138,400,000	2	277,600,000	2
13,600,000	2	141,600,000	2	289,600,000	3
14,400,000	3	142,400,000	2	293,600,000	1
17,600,000	7	145,600,000	4	297,600,000	2
18,400,000	1	153,600,000	1	298,400,000	1
21,600,000	3	157,600,000	4	301,600,000	7
25,600,000	4	157,600,000	3	305,600,000	3
26,400,000	2	158,400,000	1	309,600,000	3
29,600,000	1	161,600,000	2	313,600,000	4
33,600,000	2	165,600,000	3	317,600,000	3
34,400,000	1	165,600,000	2	321,600,000	2
37,600,000	2	165,600,000	1	325,600,000	1
38,400,000	1	166,400,000	1	326,400,000	1
41,600,000	3	169,600,000	4	329,600,000	1
42,400,000	1	170,400,000	1	333,600,000	2
45,600,000	3	173,600,000	5	337,600,000	3
46,400,000	1	177,600,000	1	338,400,000	1
50,400,000	1	181,600,000	3	341,600,000	1
53,600,000	4	186,400,000	2	342,400,000	1
57,600,000	4	189,600,000	4	349,600,000	2
61,600,000	1	190,400,000	1	357,600,000	2
65,600,000	2	193,600,000	2	358,400,000	1
66,400,000	1	194,400,000	1	361,600,000	1
69,600,000	3	197,600,000	2	366,400,000	3
70,400,000	1	198,400,000	1	369,600,000	1
73,600,000	4	205,600,000	1	373,600,000	2
77,600,000	1	209,600,000	2	377,600,000	1
78,400,000	1	213,600,000	2	378,400,000	1
81,600,000	3	217,600,000	1	381,600,000	1
85,600,000	2	221,600,000	1	385,600,000	2
86,400,000	1	225,600,000	3	386,400,000	1
89,600,000	1	230,400,000	1	389,600,000	1
94,400,000	1	233,600,000	1	393,600,000	3
97,600,000	5	234,400,000	1	397,600,000	3
101,600,000	2	237,600,000	2	398,400,000	1

Appendix I – Frequencies Selected by Stepwise to Predict 1 L Nitrate Molecular Weight with Old Control Box or Impedance Meter

Frequency (Hz)	# of Times Selected	Frequency (Hz)	# of Times Selected
50	1	79,200,000	1
1,000,000	1	80,000,000	2
1,200,000	1	80,200,000	1
3,000,000	1	85,200,000	2
5,200,000	1	86,200,000	1
10,000,000	2	87,200,000	2
10,800,000	1	88,200,000	1
12,200,000	1	89,000,000	1
12,800,000	1	89,200,000	1
13,800,000	1	90,200,000	1
15,200,000	2	90,800,000	1
20,200,000	1	92,000,000	1
23,000,000	2	92,200,000	1
24,000,000	1	95,200,000	3
29,800,000	1	100,200,000	1
37,200,000	1	102,200,000	2
40,800,000	1	103,000,000	1
43,200,000	1	103,200,000	1
44,200,000	2	104,200,000	1
46,200,000	1	105,200,000	1
47,200,000	2	106,200,000	2
50,200,000	4	108,200,000	2
60,200,000	1	109,000,000	1
61,000,000	1	109,200,000	1
61,800,000	1	110,200,000	5
62,800,000	1	113,200,000	1
63,000,000	1	114,200,000	2
63,200,000	1	115,200,000	2
66,200,000	1	118,200,000	1
68,000,000	1		
69,800,000	1		
75,200,000	3		

Appendix J – Frequencies Selected by Stepwise to Predict 1 L Nitrate Molecular Weight with New Control Box and 2 cm Probe

Frequency (Hz)	# of Times Selected
1,413	2
17,600,000	1
18,400,000	1
37,600,000	1
49,600,000	1
53,600,000	1
57,600,000	1
65,600,000	1
126,400,000	1
133,600,000	1
134,400,000	1
165,600,000	1
205,600,000	1
221,600,000	1
242,400,000	1
245,600,000	1
261,600,000	1
273,600,000	1
297,600,000	1
346,400,000	1
350,400,000	1
369,600,000	1
381,600,000	1

Appendix K - Frequencies Selected by Stepwise to Predict 36 L Nitrate Molecular Weight with New Control Box

Frequency (Hz)	# of Times Selected	Frequency (Hz)	# of Times Selected
200	1	189,600,000	2
44,668	1	190,400,000	1
251,189	1	193,600,000	1
1,600,000	1	197,600,000	2
17,600,000	1	201,600,000	1
21,600,000	1	205,600,000	1
29,600,000	1	206,400,000	1
33,600,000	1	209,600,000	2
37,600,000	1	217,600,000	1
41,600,000	1	225,600,000	1
53,600,000	1	229,600,000	4
57,600,000	1	233,600,000	3
58,400,000	2	237,600,000	2
61,600,000	2	241,600,000	3
73,600,000	1	245,600,000	1
77,600,000	2	249,600,000	2
85,600,000	2	253,600,000	3
97,600,000	1	254,400,000	1
98,400,000	1	257,600,000	2
105,600,000	1	265,600,000	2
106,400,000	1	277,600,000	5
109,600,000	1	289,600,000	2
113,600,000	1	293,600,000	3
118,400,000	1	305,600,000	3
121,600,000	1	309,600,000	4
125,600,000	1	313,600,000	2
126,400,000	1	317,600,000	4
129,600,000	3	321,600,000	6
133,600,000	2	322,400,000	1
137,600,000	4	325,600,000	3
141,600,000	1	329,600,000	1
142,400,000	1	333,600,000	1
145,600,000	2	334,400,000	1
149,600,000	1	338,400,000	2
157,600,000	1	341,600,000	3
158,400,000	1	349,600,000	1
161,600,000	2	350,400,000	1
165,600,000	1	353,600,000	1
169,600,000	5	357,600,000	1
173,600,000	1	365,600,000	2
177,600,000	1	381,600,000	1
178,400,000	1	385,600,000	3
181,600,000	1	386,400,000	2
182,400,000	1	389,600,000	2
185,600,000	1	394,400,000	1

Appendix L – Frequencies Selected by Stepwise to Predict 1 L Atrazine Concentration with Old Control Box or Impedance Meter

Frequency (Hz)	# of Times Selected
1,413	1
7,943	1
1,200,000	2
3,000,000	1
3,200,000	1
9,000,000	2
16,200,000	2
19,200,000	1
21,200,000	2
27,200,000	1
34,200,000	2
36,200,000	1
37,200,000	1
43,200,000	1
46,200,000	1
49,200,000	1
51,200,000	1
55,200,000	1
58,200,000	1
59,200,000	2
61,200,000	1
64,000,000	1
71,200,000	1
72,200,000	1
77,200,000	1
79,000,000	1
88,000,000	1
88,200,000	1
89,200,000	1
93,200,000	2
95,200,000	1
101,000,000	1
119,200,000	1

Appendix M - Frequencies Selected by Stepwise to Predict 1 L Atrazine Concentration with New Control Box and 2 cm Probe

Frequency (Hz)	# of Times Used
37,600,000	1
57,600,000	1
125,600,000	1
137,600,000	2

Appendix N – Frequencies Selected by Stepwise to Predict 36 L Atrazine Concentration with New Control Box

Frequency (Hz)	# of Times Selected
1,413	1
7,943	2
251,189	1
1,600,000	1
25,600,000	1
26,400,000	1
65,600,000	1
101,600,000	1
117,600,000	1
137,600,000	1
161,600,000	2
173,600,000	1
185,600,000	2
193,600,000	1
194,400,000	1
217,600,000	1
229,600,000	1
245,600,000	3
273,600,000	2
277,600,000	1
281,600,000	1
285,600,000	1
305,600,000	1
313,600,000	1
317,600,000	1
321,600,000	1
333,600,000	1
341,600,000	1
346,400,000	1
357,600,000	2
369,600,000	1
370,400,000	1
385,600,000	1
390,400,000	1
393,600,000	1

Appendix O – Frequencies Selected by Stepwise to Predict the Concentration of Impurities in Biodiesel with New Control Box

Frequency (Hz)	# of Times Selected	Frequency (Hz)	# of Times Selected	Frequency (Hz)	# of Times Selected
200	1	126,400,000	1	273,600,000	5
400	2	129,600,000	2	277,600,000	3
1,413	2	134,400,000	1	281,600,000	3
1,995	2	137,600,000	3	281,600,000	2
11,220	1	141,600,000	2	281,600,000	1
44,668	5	145,600,000	3	285,600,000	5
63,096	1	146,400,000	2	286,400,000	1
251,189	3	149,600,000	2	289,600,000	1
1,600,000	2	153,600,000	3	290,400,000	1
2,400,000	1	154,400,000	2	293,600,000	2
5,600,000	3	157,600,000	3	297,600,000	2
9,600,000	5	158,400,000	1	301,600,000	2
17,600,000	3	161,600,000	5	302,400,000	1
21,600,000	3	165,600,000	1	305,600,000	2
25,600,000	1	166,400,000	2	309,600,000	4
30,400,000	1	169,600,000	2	313,600,000	4
33,600,000	2	173,600,000	3	317,600,000	2
37,600,000	1	174,400,000	1	321,600,000	2
38,400,000	1	177,600,000	2	322,400,000	1
41,600,000	2	181,600,000	4	325,600,000	2
45,600,000	2	185,600,000	5	329,600,000	4
46,400,000	1	189,600,000	4	333,600,000	1
49,600,000	2	193,600,000	3	337,600,000	1
53,600,000	1	197,600,000	1	338,400,000	1
57,600,000	2	201,600,000	4	341,600,000	3
61,600,000	2	201,600,000	1	342,400,000	1
65,600,000	3	205,600,000	3	345,600,000	3
69,600,000	7	209,600,000	5	346,400,000	1
70,400,000	2	210,400,000	1	349,600,000	4
73,600,000	3	213,600,000	1	350,400,000	1
74,400,000	1	214,400,000	2	353,600,000	4
77,600,000	2	217,600,000	2	354,400,000	1
81,600,000	1	218,400,000	2	357,600,000	4
85,600,000	2	221,600,000	3	358,400,000	1
86,400,000	1	229,600,000	2	361,600,000	1
89,600,000	4	233,600,000	2	365,600,000	1
97,600,000	2	237,600,000	3	369,600,000	3
101,600,000	5	241,600,000	1	370,400,000	2
105,600,000	2	245,600,000	1	373,600,000	5
109,600,000	1	249,600,000	4	378,400,000	2
110,400,000	1	253,600,000	1	381,600,000	1
113,600,000	2	253,600,000	2	382,400,000	1
114,400,000	1	257,600,000	4	385,600,000	4
117,600,000	3	261,600,000	2	394,400,000	1
121,600,000	2	265,600,000	3	397,600,000	3
122,400,000	3	269,600,000	2	398,400,000	1
125,600,000	1	270,400,000	1		

Appendix P – Wavebands Used to Predict Concentration of 1 L Nitrate Samples with Old Control Box

	2.5 cm Probe		5 cm Probe		7.5 cm Probe		Impedance Meter	
	Level	Waveband	Level	Waveband	Level	Waveband	Level	Waveband
All Gain	6	1	5	1	6	5	9	1
	6	8			5	1	5	6
	5	1			5	4	5	18
	5	10			5	9		
All Phase	9	2	5	2	7	4	5	4
	9	1	5	7	5	2	5	14
	6	1	5	13				
	6	3						
	6	10						
	5	1						
	5	10						
	5	16						
All Gain+Phase	10	1	10	1	8	3	8	2
	9	2	6	12	6	8	7	2
	6	8			6	11	6	14
	6	10						
	6	11						
Low Gain	6	1	9	1	6	1	8	1
	5	1	5	8			7	5
	5	19	5	14				
Low Phase	6	1	6	1	6	1	5	1
	6	2	6	7			5	15
	5	6						
	5	12						
Low Gain+Phase	9	1	6	1	6	1	7	1
	8	4	6	17			6	9
	6	1					6	11
	6	19					6	13
							6	16
High Gain	5	1	5	1	6	3	8	1
					6	4	8	2
							7	4
							7	5
							5	1
							5	10
							5	11
							5	13
							5	18
							5	19
High Phase	6	1	5	2	6	2	6	6
	5	8	5	7			5	12
High Gain+Phase	9	2	6	12	6	12	6	15
	6	11						

Appendix Q – Wavebands Used to Predict Concentration of 36 L Nitrate Samples with New Control Box

	2 cm Probe		2.5 cm Probe		7.5 cm Probe	
	Level	Waveband	Level	Waveband	Level	Waveband
All Gain	6	2	6	3	5	1
	5	3	6	4	5	2
	5	6	6	6	5	7
	5	7	5	1	5	9
	5	8	5	5	5	14
	5	9	5	11		
	5	13	5	14		
	5	14				
All Phase	5	12	6	5	9	1
			6	7	7	1
			5	1	6	1
			5	2	6	3
			5	3	5	1
			5	6	5	5
			5	8		
			5	12		
All Gain+Phase	8	4	10	1	8	3
			10	1	7	5
			8	4	6	1
			7	8	6	2
			6	4	6	6
			6	5	6	16
			6	12		
			6	14		
Low Gain	0	0	9	1	9	1
					8	1
					5	7
Low Phase	9	1	7	4	8	1
			6	1	6	1
			5	1		
Low Gain+Phase	7	6	6	9	8	3
			6	11		
High Gain	9	2	6	8	9	1
	5	14	5	1	7	1
			5	3	7	2
			5	4	6	1
			5	11	6	4
			5	13	6	5
					5	1
					5	2
					5	3
					5	4
					5	5
					5	6
					5	7
					5	9
					5	16

	2 cm Probe		2.5 cm Probe		7.5 cm Probe	
	Level	Waveband	Level	Waveband	Level	Waveband
High Phase			9	2	9	2
			8	1	8	1
			7	1	6	5
			7	4	5	2
			6	1	5	5
			6	2	5	6
			6	4	5	7
			6	7	5	15
			6	8	5	16
			5	1		
			5	3		
			5	6		
			5	7		
			5	14		
			5	15		
High Gain+Phase			10	1	7	1
			10	1	7	5
			7	3	6	1
			6	2	6	9

Appendix R – Wavebands Used to Predict Molecular Weight of 36 L Nitrate Samples with New Control Box

	2 cm Probe		2.5 cm Probe		7.5 cm Probe	
	Level	Waveband	Level	Waveband	Level	Waveband
All Gain	6 6 5 5 5 5	4 5 7 11 13 14	50	20	5	14
All Phase	6 6 5 5	4 5 13 14	5	16	5 5	13 15
All Gain+Phase	10 6 6 6 6	2 4 6 8 13	6	5	6 6	13 15
Low Gain	6 6 5	6 8 10	9 6 6 5 5	1 3 7 14 15	6	6
Low Phase	5 5 5	10 11 15	6 5 5	5 4 12	9 6 6	2 4 8
Low Gain+Phase	10 8 7 7	1 4 7 8	10 8 8 7 6 6 6	1 2 4 8 11 15 16	6 6	6 16
High Gain	6 5 5 5 5	6 8 14 15 16	5	3	8	2
High Phase	9 9 6 5 5	1 2 5 8 13	6	6	5	8
High Gain+Phase	8 7	4 6	7	3	9	2

Appendix S – Wavebands Used to Predict Concentration of Atrazine in Water

Table 7.1: Wavebands used to predict atrazine concentration in 1 L water samples

	2.5 cm Probe		5 cm Probe		7.5 cm Probe		Impedance Meter	
	Level	Waveband	Level	Waveband	Level	Waveband	Level	Waveband
Gain	7	1	7	2	5	11	5	4
							5	7
Phase	9	1	6	1				
	6	3						
	6	7						
	5	7						
	5	9						
Gain+Phase	7	1	7	2			10	1
							7	4
							6	3
							6	7
							6	18

Table 7.2: Wavebands used to predict atrazine concentration in 36 L samples

	2 cm Probe		2.5 cm Probe		7.5 cm Probe	
	Level	Waveband	Level	Waveband	Level	Waveband
Gain	6	2	7	1	7	2
			5	3	0	0
Phase	7	1	7	1	5	5
	6	7			5	9
	5	2				
	5	4				
Gain+Phase	6	2	8	3	7	2
			7	7		
			6	4		

Appendix T – Wavebands Used to Predict Impurities in Biodiesel

Table 7.3: Wavebands used to predict the concentration of water in 36 L biodiesel samples

	2 cm Probe		2.5 cm Probe		7.5 cm Probe	
	Level	Waveband	Level	Waveband	Level	Waveband
All Gain	6	7	6	1	6	3
	5	11	6	6	5	3
	5	14	5	2	5	9
			5	10		
All Phase	9	1	6	5	7	2
	5	13	5	16		
All Gain+Phase	7	2	7	6	7	2
	6	7	6	1	7	6
					6	11
Low Gain	7	2	7	3	5	6
	5	12	6	6	5	13
			5	1		
			5	10		
Low Phase	8	2	9	2	5	3
	5	15			5	4
					5	5
					5	10
Low Gain+Phase	7	2	7	3	6	9
High Gain	6	7	6	7	7	3
	5	9	5	2	6	6
	5	14	5	9	5	14
High Phase	5	15	7	1	7	2
					5	1
					5	4
High Gain+Phase			7	2	10	1
			6	1	8	1
			6	12	6	16

Table 7.4: Wavebands used to predict the concentration of glycerol in 36 L biodiesel samples

	2 cm Probe		2.5 cm Probe		7.5 cm Probe	
	Level	Waveband	Level	Waveband	Level	Waveband
All Gain	8	1	8	1	8	2
	6	8	5	3	5	14
	5	10	5	12		
All Phase	7	2	6	3	7	2
	6	6	6	7	6	1
	6	7	5	4	6	8
	5	13	5	10	5	4
					5	11
					5	14
All Gain+Phase	6	8	8	1	10	2
					8	3
Low Gain	6	4	9	2	8	2
			8	2	5	1
			5	2	5	5
			5	11	5	9
Low Phase	6	6	7	1	5	7
	5	5			5	10
Low Gain+Phase	10	2	7	6	6	10
			7	8		
			6	3		
			6	6		
High Gain	6	8	6	8	6	4
					5	11
High Phase	5	8	7	3	7	3
	5	11	6	4	5	12
	5	12	6	6		
High Gain+Phase	8	3	7	7	8	1
	7	3			7	7
	6	12			6	4

Table 7.5: Wavebands used to predict the concentration of glyceride in 36 L biodiesel samples

	2 cm Probe		2.5 cm Probe		7.5 cm Probe	
	Level	Waveband	Level	Waveband	Level	Waveband
All Gain	9 5	1 7	6 6	4 5	8 6 5	2 4 14
All Phase	6	6	6 5	7 5	5	11
All Gain+Phase	7 6 6	2 10 14	6 6	4 5	6	5
Low Gain	9 6	1 7			9 5 5 5	1 3 13 15
Low Phase	5	1			6	6
Low Gain+Phase	9 6	1 7	8 7 7 7	3 1 7 8	7 6 6	1 8 10
High Gain	5	7			7 5 5	4 2 5
High Phase	7	4	5	2	6 5 5	2 4 10
High Gain+Phase	6	5	7 6 6	3 6 9	7	4