2010

Statistical Analysis of Wastewater Remediation and Bio-Fuels Production of Algae

Jay D. Jones
Utah State University

Follow this and additional works at: https://digitalcommons.usu.edu/gradreports

Part of the Applied Mathematics Commons, Mathematics Commons, and the Statistics and Probability Commons

Recommended Citation
https://digitalcommons.usu.edu/gradreports/1206

This Report is brought to you for free and open access by the Graduate Studies at DigitalCommons@USU. It has been accepted for inclusion in All Graduate Plan B and other Reports by an authorized administrator of DigitalCommons@USU. For more information, please contact dylan.burns@usu.edu.
STATISTICAL ANALYSIS OF WASTEWATER REMEDIATION AND BIO-FUELS PRODUCTION OF ALGAE

By

Jay D. Jones

A thesis submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

Statistics

Approved:

D. Richard Cutler, Ph. D
Major Professor

Ronald C. Sims, Ph. D
Committee Member

Daniel C. Coster, Ph. D
Committee Member

UTAH STATE UNIVERSITY
Logan, Utah
2010
ABSTRACT

STATISTICAL ANALYSIS OF WASTEWATER REMEDIATION AND BIO-FUELS PRODUCTION OF ALGAE

By

Jay D. Jones, Master of Science
UTAH STATE UNIVERSITY, 2010

Major Professor: Dr. D. Richard Cutler
Department: Statistics

The Logan city wastewater treatment system consists of a series of seven large aerated ponds (460 acres) that biologically treats 15 million gallons per day of wastewater from Logan city and six other communities. Tighter regulations of allowed phosphorus levels in the effluent have recently been implemented due to environmental concerns of a downstream reservoir. The Biological Engineering program at Utah State University, the Bio-fuels Center, the Utah Water Research Laboratory (UWRL) and the city of Logan are working together to remediate the wastewater treatment system using microalgae. Algal growth requires the uptake of phosphorus. Thus, phosphorus in the effluent can be removed by encouraging algal growth and then removing the algae from the system. The harvested algae can then be used by the Bio-fuels Center to conduct research related to algal bio-fuels. The work reported here concerns the construction and analysis of experiments that were used by the Bio-fuels Center and UWRL to study growth and harvest methods for algae, and the design of a survey of the wastewater ponds to spatially characterize concentrations of algae. The experiments described herein demonstrate that, by encouraging algal growth in the ponds, phosphorous may be removed from the system and effective methods for growing and harvesting algae are proposed. Because a manually
sampled survey of the ponds is expensive and requires a great deal of resources, an alternative approach using aerial image data of the ponds was used to predict the manually sampled data of algae concentrations. Predicting algal concentrations using image data would mitigate the cost required of determining a location to harvest. Two prediction methods – linear models and random forests – are investigated and compared. The methods are compared on the basis of the amount of variability in the data they are able to explain. Random forests outperform linear models in predictive ability when using only the image data for prediction. The results of this project provide a proof of concept for the Bio-fuel Center, UWRL and Logan city initiative, that phosphorus can be removed from the effluent of the Logan lagoons by algae that can be grown and harvested efficiently.
1. Introduction

The Biological and Irrigation Engineering (BIE) department at Utah State University (USU) is involved in research funded by the state of Utah and the Utah Water Research Laboratory (UWRL) to develop economical and efficient methods of producing bio-fuels through wastewater reclamation. This comes as part of Utah’s Bio-fuels Center initiative. The center functions to conduct research, generate discoveries, and aid in the flow to commercialization through the USU Technology Commercialization Office (TCO). In order to accomplish these goals statistically defensible experiments, data and results must be used and implemented. The Bio-fuels project is specifically interested in creating bio-fuels from algae. The focus of this project is to ensure the statistical integrity of the data collected, and to use the most accurate and appropriate prediction and inference methods. This project aims to: 1. work with researchers of the Biological Engineering Program and Bio-fuels Center to provide statistically sound designs for laboratory scale experiments, and 2. Design and implement a survey that will allow for prediction while requiring minimal resources. To obtain the best predictions possible, two statistical methods of predicting are examined. Fulfilling these two objectives provides the Bio-fuels Center and the City of Logan with valuable information about their objective to utilize algal growth and recovery as the method of removing phosphorus from wastewater and producing bio-fuels from the harvested algae.

1.1. Experimental Design

Research is being done around the world to determine how to effectively produce renewable energy. One area of research that is of particular interest is using algae to create bio-fuels. This is the focus of the Biological Engineering program collaboration with the Bio-fuels
Center, UWRL, and the city of Logan. Using algae is appealing because it is abundant, relatively fast and easy to grow, and naturally occurs in many ecosystems. The lipids that algae produce when under stress can be converted into bio-diesel through transesterification. Algal bio-mass can be converted to bio-methane through anaerobic digestion. One drawback of using algae to produce bio-fuels is that, as compared to traditional agriculture, little is known about the conditions needed to optimize algal growth, especially if one is interested in knowing the necessary conditions to optimize lipid production. Conditions and equipment needed to optimize corn or wheat production are well known, whereas studies to optimize algal lipid production and extraction is relatively new. More understanding of algae is needed before bio-fuels from algae become a viable option as a renewable energy source. Sound statistical designs must be in place to accurately conclude what conditions and harvest methods optimize the yield of algae. Just as in agriculture, only significant improvements in production or harvesting methods can be made when appropriate statistical designs of experiments are in place.

The Biological Engineering program at USU and the Utah Bio-fuels Center are involved in the pursuit of optimizing conditions for algal growth. Several experiments have been carried out by the researchers. It is the purpose of this project to provide appropriate statistical designs for experiments within the logistical constraints faced by researchers. As examples of experiments that have been carried out, two are discussed in detail, their designs explained and the analysis put forth. These experiments were carried out by Erick Griffiths and Logan Christensen, both Master’s students from the BIE department at USU. While their experiments focus on different methods of algal growth, they have the same goal: to optimize algae production.
1.2. Logan Lagoons:

The Logan City environmental department in northern Utah currently uses a series of large aerated ponds (460 acres) and wetlands (240 acres) to biologically treat and polish wastewater produced by the city and surrounding communities. The treatment facility receives wastewater from Logan, North Logan, Hyde Park, Smithfield, River Heights, Providence and Nibley. The facility receives between 14 and 18 million gallons of wastewater daily. The influent is then split into two ponds referred to as A1 and A2. In the lagoons system, microbial activity and large aerators facilitate the reduction of biological and chemical oxygen demand and the removal of pathogens, solid materials, and excess nutrients. Water from ponds A1 and A2 flows into B1 and B2, respectively, and water from both B ponds flows into C, then D, then E, as shown in figure 1.

Figure 1: An aerial view of the lagoons with the top of the figure pointing east. The arrows indicate head gates to denote the flow of the ponds. (Griffiths 2009)
From pond E the water is discharged into wetlands before flowing into Cutler Reservoir. Wastewater that enters this facility has an average residence time in the lagoons of about 60 days before exiting the system. For the purposes of this project, the system as a whole is referred to as the Lagoons but specific ponds are referred to by name (i.e. A1, A2, ... etc.) when necessary.

Once the wastewater has been treated in the lagoons and subsequent wetlands it flows into Cutler reservoir, an open water body that serves recreation, fishing, and irrigation purposes. One challenge that Logan City’s environmental department is faced with is that the average phosphorus levels in the effluent of the treatment facility is approximately 4 mg/L, twice the current state regulation. These relatively high levels of phosphorus that enter Cutler reservoir provide conditions for algae to flourish. The problem created is eutrophication, where the algal presence competes for the oxygen available to fish during times that solar irradiation is not available. Eutrophication of Cutler reservoir would have long term effects on the eco-scape of Cutler reservoir. The state of Utah’s Department of Environmental Quality (DOE) is concerned about this environmental problem and would like to correct it before it becomes a larger problem. The solution proposed by the state is to implement stricter regulations on the allowed level of phosphorus in the effluent of the lagoons. This would require the phosphorus from the lagoons to be less than 1 mg/L (one-fourth of the current concentration). These regulations, if not met, could require the city to implement a chemical treatment plant as part of the wastewater treatment system. This would be an expensive solution to the problem, costing the city 100 – 200 million dollars which would raise the average household utility bill for wastewater treatment from eleven dollars a month to as much as ninety dollars a month (Issa Hamud, Director Logan City’s Environmental department).
An alternative solution to this problem comes from USU’s BIE department that is investigating sources of economical production of bio-fuels. It has been shown that algae have the potential to be a source of lipids that, through transesterification converted to bio-diesel (Chisti 2009, Sheenan et. al. 1998, Johnson 2009). Also that algal biomass itself, through anaerobic digestion, produces bio-methane (Chynoweth et. al. 1987, Sheenan et. al. 1998). The lagoons provide an abundant source of algae that can be harvested and used in researching bio-diesel and bio-methane production. Harvesting algae from the lagoons provides Logan City with a free stock of algae and nutrients while aiding the city in the removal of phosphorus from the system. Simply growing more algae will not necessarily lower the phosphorus levels of the effluent. Phosphorus is removed from the system during harvesting because as algae grow they use phosphorus as a nutrient. If the algae die before they are harvested, phosphorus is released back into the system. In order to remove the phosphorus, the algae need to be harvested from the lagoons. Harvesting the algae from the system removes the phosphorus. This solution creates a synergy between Logan City and USU.

The location of high and low algal concentrations must be known in order to effectively harvest algae from the lagoons. This project aims to determine spatial algal concentrations and provide a method for efficiently obtaining this information. The primary method of sampling has been to obtain grab samples taken at the head-gates of each pond. This type of sampling only provides researchers with information of what is being transferred from pond to pond and does not provide information of the spatial distribution of algae within the ponds. This project determined and implemented an appropriate method of sampling each pond in multiple locations to provide an understanding of algal concentrations throughout the lagoons system. The sampling method required to provide the information needed is time consuming and expensive. It
has been shown by Kloiber et. al. (2002) and Ramakrishna and Chang (2009) that measures of water clarity and algae activity can be predicted using aerial image data. This project generalizes these results to a wastewater lagoons system. Aerial photos of the lagoons are taken at or near the time of manual sampling. The aerial images and the manually sampled ground truth data are then used as a training data set from which statistical prediction methods can be implemented to determine algal concentrations from aerial images. This approach eliminates the need for manual sampling, except for model calibrations that may be necessary. Two statistical methods for prediction are examined in this project for their accuracy in predicting algal concentrations from aerial images. The two methods used for prediction in this project are linear models and random forests. A comparison of their respective accuracy in this application is examined to determine which method is best for making predictions of future observations. How well each method performed is discussed.

An accurate and easy to implement method for determining algal concentrations will provide information that is important in developing useful harvesting methods. Helping the Biofuels Center to plan and accomplish effective harvests will remove phosphorus in the effluent, making it possible for Logan City to meet state regulations and preserve the integrity of Cutler Reservoir.

2. Methods

2.1. Experimental Design

This project uses traditional experimental design techniques in collaboration with researchers in the Biological Engineering program at USU. Researchers associated with the Biofuels Center need to answer specific questions, so they plan experiments to provide the answers.
These research questions all revolve around the goals of how to best grow and harvest algae. Once meaningful factors and factor levels have been selected, logistics and the resources available to the experiment are considered. The goal is to find an experimental design that enables the researcher to answer their questions of interest under logistical constraints. An appropriate experimental design is important to be able to test for significant factors, interactions, and differences between factor levels within the logistic and resource constraints that include resources and time.

Specific experiments are used to illustrate different points of design. The primary software used in the analysis of the data collected from these experiments is Statistical Analysis Software (SAS Statistical Institute 2003), while some graphics and summary statistics were produced using R statistical package (R Core Development Team 2009). Within SAS, `proc mixed` is the procedure that is appropriate for mixed models analysis of variance of experimental data and it was used for all analyses. Where appropriate, `proc glm` was also used.

2.1.1. A factorial design with repeated measures

A selection of Erick Griffiths’s research involves a series of experiments carried out to determine what nitrogen source produces the most algal biomass in laboratory scale raceway reactors. In one such experiment, Griffiths wanted to look at algal growth over time with different sources of nitrogen. The primary goal of this experiment was to determine which nitrogen source produces the most biomass of algae native to the ponds under the same external conditions. A secondary goal of this experiment was to determine limiting factors for algal growth. This secondary goal was achieved through chemical analysis of the concentrations nutrients such as total soluble phosphorous, total soluble nitrogen, orthophosphate, and ammonia.
in the reactors. By monitoring the levels of these nutrients, it can be determined if there is nutrient limitation causing lower biomass yields. This experiment has several different responses. Each response is treated as an independent experiment, and analyzed separately.

In this experiment Griffiths tested the effects of three nitrogen sources and a control, on algal growth over time. The algae stock used in the experiment comes from the effluent of the lagoons which is then spiked with the different nitrogen sources. The control in the experiment is algae from the effluent of the Logan lagoons that did not receive nitrogen other than what was already in the water. Griffiths constructed eight reactors with a paddle wheel to keep the nutrients and algae well mixed for this experiment, (See figure 2). Each reactor was approximately 2 ft. long and 1 ft. wide. The reactors were split down the middle length ways to create two 0.5 ft wide channels, with the paddle wheel in one channel. The paddles rotated at a constant rate of ~10 rpm. Lighting was equally distributed by a series of florescent natural light plant growth bulbs. It was determined that all other variables that could potentially affect the results were held constant.

With the available resources and end goals considered the experiment was designed as a 4x12x2 factorial design with repeated measures taken on the experimental subjects, the reactors. The first factor is the nitrogen source with four levels: ammonium (NH₄⁺), nitrate (NO₃⁻), urea ((NH₂)₂CO), and the control. The second factor is time with twelve levels corresponding to the twelve time points that samples are drawn from each reactor. Each reactor is changing over time and is also being sampled over time so reactor becomes the subject of repeated measures. Each nitrogen source and time period is observed in two independent reactors which becomes the replication. The reactors were randomly assigned to the level of nitrogen source they would receive.
The experiment was conducted over the course of 12 days. Samples were drawn once a
day then analyzed for total suspended solids (TSS), ammonia (NH$_3$), orthophosphorus (PO$_4$),
total nitrogen, and total phosphorus. Each response is analyzed by the model:

$$y_{ijk} = \mu + N_{sourcei} + \text{Time}_j + (N_{source} \times \text{Time})_{ij} + \epsilon_{ijk}$$

where $y_{ijk}$ is the response for the $i^{th}$ level of $N_{source}$, at the $j^{th}$ level of time and the $k^{th}$ replicate.

Tukey's p-value adjustment is to correct for multiple comparisons.

![Figure 2: Reactors of Erick Griffiths's experiment testing the effect of different nitrogen sources in small scale raceway reactors (Griffiths 2009).](image)

2.1.2. A nested factorial design

A drawback of using algae to produce bio-fuels is that extracting the algae from the water
can be difficult and expensive. Traditional filters clog easily with algae and drying is either too
time consuming or not energy efficient. One solution is to grow the algae on a surface in water
that can be easily scrapped or removed for harvesting. Logan Christensen proposed a method of
growing algae using rotating photo-bioreactors (RPB) (Christensen 2010). This uses the concept
of rotating biological contactors, where the surface of a rotating drum's surface seeded with algae
then placed in a nutrient source. The drum is submerged about 40% and slowly rotates so that all
surfaces of the drum receive nutrients and light. Christensen was interested in testing eight
surface materials applied to the drum, or *substrates*, as an exploratory step of determining an appropriate substratum. The eight tested substrates are: acrylic fabric (Acr), cotton drop (CotDp), cotton rope (CotRp), flannel (Fln), jute, polyester (Pester), polypropylene (PP), and nylon. The result of interest was to determine which substrate allowed for the most biomass in $g/m^2$. The response was measured by removing the substrate from the RPB and scraping the algal bio-film from the substrate. The harvested algal bio-film is lyophilized then weighed in grams. The dried cell weight is divided by the surface area of the substrate and used as the response.

A system of nine RPBs was available for this experiment. Each RPB was 40 inches long with a 3 inch diameter and rotated at a constant speed of 4.8 rpm. The system of nine RPBs held nine self-contained reactors, each with a trough that holds eight liters when the reactor is approximately 40% submerged (See figure 3). The major focus of this experiment was to see how the algae grow on the substrates over time, so three harvests time periods were observed: 14, 22, and 26 days.

This experiment was designed so that repeated measures did not have to be taken. This is because each RPB contains all eight substrates being tested, which minimized error from sampling over time. The placement of the substrates on each RPB was randomly assigned. At each time point, three randomly selected reactors were harvested, meaning that all substrates on the reactor are removed, scraped, and the algae biomass is weighed. This means that each reactor serves as an experimental unit. So, with the available resources this experiment was a nested factorial design with three factors: substrate, time, and reactor. Substrate has eight levels, time has three levels and there are nine reactors, three per level of time. Substrate and time are fixed factors and reactor is a random factor, and reactor is nested within time.
The model used to analyze which factors and factor levels make significant changes in the harvestable algal biomass is:

\[ y_{ijk} = \mu + Substrate_i + Time_j + (Substrate \times Time)_{ij} + Reactor(Time)_{jk} + \varepsilon_{ijk} \]

where \( y_{ijk} \) is grams of algae per meter squared for the \( i^{th} \) substrate, at the \( j^{th} \) time and the \( k^{th} \) replication.

**Figure 3:** Rotating Photo-Bioreactors of Christensen’s experiment. This figure shows the randomization of the substrates before the experiment was started. (Christensen 2010)

2.2. Logan Lagoons Sampling

The cost and time required to determine algal concentrations of the Logan City wastewater lagoons needs to be minimized in order to make the harvest of algae as effective as possible. To this end, a survey of the ponds was implemented that would provide accurate measurements of surface variables throughout the lagoons. These surface variables include Secchi Depth Transparency (SDT), Total Suspended Solids (TSS), Chlorophyll-a (Chl-a), Optical Density (OD), Total Nitrogen (TN), Total Phosphorus (TP), Ammonia (NH₃), Orthophosphate (PO₄³⁻), Nitrate (NO₃⁻), Dissolved Oxygen (DO), pH, and temperature. These or a subset of these were measured at every location each time sampling took place. The first four variables identified (SDT, TSS, Chl-a, and OD) are considered surrogates of algal concentration.
because each of these variables are strongly correlated to the concentration of algae while not directly measuring algal concentration.

2.2.1. Materials

A brief explanation of how each variable was measured is given to provide a deeper understanding of the survey.

*Secchi Depth Transparency* (SDT) is a low-tech measurement of the turbidity of the water which is directly related to the algal concentrations of the wastewater lagoons. It is measured using a disk, eight inches in diameter that is separated into four quadrants by two perpendicular diameters. Quadrants diagonal from each other are the same color, either black or white. From the center of the disk there is a length of rope with notches indicating depth in inches. The disk is lowered in the water until the black and white quadrants are no longer distinguishable and that depth is recorded as the SDT in inches. (See figure 4) (Eaton et. al. 2005)

![Secchi Disk](image)

**Figure 4:** An example of a Secchi Disk that was used in the survey of the Lagoons. *Secchi Depth Transparency provides a quick and easy method of determining the turbidity of water.*

There are two ways to measure *Total Suspended Solids* (TSS): with a probe that can be lowered from the boat into the water about 3 – 4 inches, or a sample can be taken back to the lab
to be later analyzed. The probe uses infrared sensors to detect particulates and is then able to provide a measurement of TSS in mg/L. Samples that are taken back to the lab to be analyzed require that first a standard curve of known TSS samples are analyzed using a spectrophotometer, then a regression line is fit and used to predict TSS in mg/L of the samples collected. Only one TSS probe was available for sampling so only one sampling crew was able to use this method. Both sampling crews took samples to be analyzed in the lab for TSS. (Eaton et. al. 2005)

The measurement of chlorophyll-a (chl-a) is a time consuming and resource intensive process. Fifty ml samples are taken from each sampling location and are transported to the lab. Precautions are taken with these samples to make sure that they receive minimal light and that they stay as close to 4° C in temperature as possible. Once in the lab the samples are filtered to remove daphnia which eat the algae and destroy the chlorophyll. Next, the concentration of chlorophyll in the samples are compared to those of samples with known chlorophyll concentrations are measured. These data are used to create a calibration curve to correct the individual and machine bias of the procedure. The samples are then analyzed for chlorophyll concentration in mg/L as stated in “Protocol for chlorophyll determination”, a report by Biological Engineering student Katerine Nepal for the Bio-fuels Center (2009).

Optical Density (OD) is measured by a spectrophotometer after the sample has been placed in a vortex to ensure uniform mixture. The measurement recorded is the amount of absorbance of light at a specific wavelength set by the user and the value used for the survey data is 664 nm. (Eaton et. al. 2005)

Total Nitrogen, Total Phosphorus, Ammonia (NH₃), Orthophosphate (PO₄³⁻), and Nitrate (NO₃⁻) are all variables that are analyzed in the lab from samples taken from the lagoons. For
each of these variables different HACH kits are used to determine the respective chemical concentrations in mg/L.

*Dissolved Oxygen* (DO), *pH*, and *temperature* are all measured by probes from HACH. The data is recorded by the sampling crews from each location.

### 2.2.2. Manual Sampling

The purpose of manual sampling is to provide a training data set that can be matched to the aerial images. A useful training data set contains the entire range of data values that one would like to be able to predict. The ability to predict high and low concentration of algae is necessary for this project. A grid of sampling locations was created for each pond in order to maintain the statistical integrity of the sampling. The sampling locations were placed roughly equidistant from each other in each pond, although logistics didn’t allow for the equidistant grid to be consistent throughout all ponds. The restrictions came due to cables and heavy duty electrical wires strung between aerators in the lagoons. This was the largest problem in ponds A1 and A2 where the aerators themselves are placed in a grid. Alterations to the equidistant grid were made as necessary (See figure 5). All sampling locations were predetermined using Google earth, then the latitude and longitude coordinates were programmed into Garmin Rino 110 handheld GPS units. The accuracy of the GPS units depends on the availability of satellites on a given day. Days when manual sampling occurred, accuracy was within 10 – 20 ft. The GPS units are used to enable the researchers within 20 feet of the predetermined point. Once the researchers get within 20 feet of a sampling location they record their exact location in latitude and longitude as given by the GPS. The sampling points were located on aluminum row boats equipped with battery powered trolling motors (See figure 6). These motors are designed to minimally disturb
the water while providing propulsion. This is necessary to ensure that the samples obtained are not affected by the boat and represent as accurately as possible the concentrations of algae.

Sampling in a grid formation has statistical benefits. By grid sampling there is an even coverage of sampled locations throughout the Lagoons. This gives a spatial understanding how the ponds flow, and the mixing within the ponds. Obtaining data from locations that are equidistant from each other increases the accuracy of predictions between points. If points were sampled irregularly or randomly, there would be areas that have high predictive accuracy (where sampling locations are concentrated) and areas that have low predictive accuracy (where sampling locations are sparse). Predetermining sampling locations ensured the statistical integrity of the samples by requiring that the samples are taken at predetermined locations rather than allowing the persons doing the sampling to make potentially biased judgments about where to sample. The predetermined sampling locations on a grid ensure that a large range of data values can be observed.

**Figure 5 (left):** An aerial view of the Lagoons provided by Google earth with the sampling locations indicated by the points on the map. Notice ponds A1, and A2 do not have a regular grid as do the rest, this is due to the logistic constraints of the aerators. **Figure 6 (right):** The sampling crews for October’s manual sampling shown at the north east corner of pond E.

Manual sampling has taken place on three occasions. On June 23, 2009 the first sampling run took place as a practice where only SDT, TSS, and temperature were measured. This
sampling run was an opportunity for those taking samples to familiarize themselves with the GPS units, how to measure each variable, how to operate the instruments, and how to navigate the boats on the lagoons. For this practice sampling run, only pond B2 was sampled. It took approximately 2.5 hours to sample the entire pond, a total of 16 locations. No aerial images were taken along with this sampling.

The second sampling exercise took place on July 8 and 13. This was the first full sampling of the lagoons, a total of 98 locations were sampled (two points were omitted from pond A2 due to a loss of battery power during sampling). To accomplish this sampling, two boats equipped with the same instruments set out to collect the 100 samples. About half of the samples were collected on July 8th. Sampling had to be cut short due to winds in the early afternoon. The sampling was completed on the next available date: July 13th. The total time to complete the full sampling in July was about 12 hours. All surface variables mentioned above were measured during this sampling run. Due to the gap in time between the first half of sampling and the second half, special consideration was taken with this data. This was done by analyzing the as two separate data sets and as a whole. Aerial images were taken on July 8th from an unmanned aerial vehicle (UAV). Due to complications with the images taken from the UAV like sun glare and image borders not aligning correctly, satellite images from July 6th were obtained and used as the image data in the analysis.

On October 23 the third manual sampling took place, the second full sampling. A total of 100 locations were sampled. This sampling run started earlier in the morning, knowing that winds tended to pick up in the afternoon making it difficult to sample. All 100 locations were sampled in the same day by two boats in approximately 5 hours. The sampling time was decreased significantly by reducing the number of variables measured on the site, and by
relaxing the accuracy by which the sample locations was to be determined from within 20 feet of the predetermined location to being within 40 feet of the predetermined location. The variables measured for this sampling run were SDT, TSS, pH, Temp, Chl-a, NH3, and OD. Due to weather conditions on October 23, the UAV was not able to take images that day but flew and took images the following day October 24.

2.2.3. Aerial Image Data

Aerial images were taken at or near the time of each sampling run except for the first “practice run” in June. According to Kloiber et. al. (2002) and Ramakrisha and Chang (2009) predictive ability is far better if the images are taken within a day of sampling, with decreased predictive ability as the time between manual sampling and when the image is taken increases. They report that no reasonable predictions are able to be made if the time between sampling and the image being taken is more than ± seven days. With this in mind, the images would be ideally taken while sampling happening, but logistics, weather, and data availability have prevented this from being the case.

The UAV used in this project was designed by Austin Jensen (see figure 7), a research engineer at the UWRL. The UAV has built within it a 9 megapixel digital camera, altered to fit inside the UAV that takes pictures continuously on a set time interval. Every time the camera takes a picture the position and orientation of the UAV are recorded to help with the post processing of the images. The UAV flies overhead at approximately 1000 meters, and takes a series of images. Flying at this altitude with this camera gives a pixels resolution of about 0.25 m². These images are then stitched together using EnsoMOSAIC software, which geo-references the pixels. Geo-referencing assigns each pixel an easting and northing value from the Universal Transverse Mercator (UTM) coordinate system. The images are then converted into a data file.
which records each pixel location and the pixel intensities at three wavelengths: 450 \textit{nm}, 550 \textit{nm}, and 600 \textit{nm}, corresponding to blue, green and red respectively. From this point on pixel intensities of these wavelengths are referred to by their corresponding color. This data file, when coming from the UAV is huge: \(\sim 35\) million entries. The satellite image’s resolution is about 1 – 2 \(m^2\) resolution so it contained only \(\sim 350,000\) entries.

\textbf{Figure 7 (left):} The UAV used to collect image data for both July and October samplings. \textbf{Figure 8 (right):} Austin Jensen is shown aiding the take off of the UAV for the July sampling. \textit{Figures courtesy of UWRL.}

2.2.4. Cost Analysis:

At, or near the time of manual sampling aerial photos of the lagoons are taken. These photos provide image data that can be easily obtained in the future. The goal of this survey is to be able to predict the surrogates of algal concentration using only the image data. The image data can be obtained, compressed, stitched, and analyzed in a matter of hours whereas manually sampling these variables can take days to obtain the samples and then weeks analyzing them. The cost for obtaining the aerial data is about $800 whereas the labor of the manually sampling alone is about $950. The labor for analyzing the samples is another approximate $1400, and the materials to analyze the samples, assuming that the machines and instruments are available for use are no less than another $1050. So the manual sampling, in total, costs about $3400, but
more likely costs up to $5000 – $6000 for a full data set that is available in no less than two weeks. The aerial data potentially provides the same information in hours for about $800. It is important to note that while the manually sampling is necessary to predict the surrogates of algae concentration from the image data, it will be made unnecessary by utilizing aerial images and the latest prediction methods.

2.2.5. Data Analysis:

Manual sampling and the aerial images provide large amounts of data. It is the purpose of this project to effectively analyze and summarize these data in a meaningful way. Summaries of the manually sampled data are displayed by prediction maps of the lagoons with lines denoting the roads overlaid to provide context (see figures 15 and 19). These prediction maps are created using a spatial statistical smoothing technique called Kriging. This method of summarizing data fits a covariance structure to the data based on the variance between points a fixed distance apart. The fitting of the covariance structure is done by plotting the distances between sampling locations on the x-axis by the variance at those distances on the y-axis, where the variance, \( \gamma(h) \), is estimated by:

\[
\hat{\gamma}(h) = \frac{1}{2 \, n(h)} \sum_{i=1}^{n(h)} (z(x_i + h) - z(x_i))^2
\]

\( \hat{\gamma}(h) \) is the estimated variance, \( z \) is the datum at a particular location, \( h \) is the distance between ordered locations, and \( n(h) \) is the number of pairs at distance \( h \). The plot of the distances verses the estimated variances is the variogram of the data. The appropriate covariance is heuristically determined by examining the fit of different covariance structures to the empirical variogram. The covariance structures that are focused on for these predictions are Gaussian,
Exponential, and Spherical (Schabenberger & Gotway, 2005). Along with the prediction maps, standard deviation maps were computed to provide estimates of the errors associated with the predictions. Prediction maps and their associated standard deviation maps were created for each variable from both full samplings. These maps were created in R statistical software (R Core Development Team 2009) using the geoR package (Ribeiro and Diggle 2001). These prediction maps are also compared to box-plots of each variable by pond.

The relationship between the measured variables is very important to the researchers of the USTAR bio-fuels project. To investigate these relationships correlation matrices were created to display all possible combinations of variables (see appendix figures 37 – 38). The combination of variables that showed interesting patterns that warranted further investigations were then analyzed using traditional regression methods.

The image data presented a unique problem. The image data files themselves each contain about 35 million rows and 5 columns which correspond to the easting, northing, and then the pixel intensities for red, green, and blue. These data sets are very large and difficult to manipulate. The image data sets are rounded to the nearest 1, 2, 5, 10, 12.5, 20, 25, 33.3, 50, 100 meters by easting and northing, then the pixel intensities that shared the same easting and northing coordinates are averaged. This is done to make the data sets smaller and more manageable and to determine the effect of rounding to the entries on the predictive ability of the different methods used.

In order to add image values to the manually sampled data set, the latitude and longitude coordinates were converted to UTM coordinates using Geographic Translator 2.4.2 (GEOTRANS) software. The manually sampled data was then also rounded to the nearest 1, 2, 5, 10, 12.5, 20, 25, 33.3, 50, and 100 meters, and then paired to the image values that share the
same coordinates. The training data set is created with the manually sampled data matched to the image data. The training data set can then be used to predict the surrogates of algal concentration from the images.

There are two methods of prediction that are investigated in this project: traditional linear models and random forests. Linear models are used because of their ubiquity in research where predictions are needed to be made. Ramakrishna and Chang (2009) also showed that water clarity can be predicted by aerial images using linear models. Their study used satellite images to predict algal concentrations in the Chesapeake Bay. The models used in their study are investigated as well as many other models. Linear models have nice mathematical properties with easily interpretable results in the context of the problem. They are also easy to implement in almost any software. For these reasons linear models are investigated. Linear models that are to be used for prediction will have assumptions of the residuals verified. These assumptions are investigated to ensure the model’s ability for statistical inference.

This project also investigates models implemented by Ramakrishna and Chang and others created through an understanding of the lagoons and the relationships between predictor variables. The following models are examined for their usefulness in predicting algal concentration from rounded image data and other covariates:

\[
SDT.1* \log(SDT) = \beta_0 + \beta_1 Red + \beta_2 Green + \beta_3 Blue
\]

\[
SDT.2* \log(SDT) = \beta_0 + \beta_1 \left(\frac{Blue}{Red}\right) + \beta_2 Blue
\]

\[
SDT.3 SDT = \beta_0 + \beta_1 \left(\frac{Blue}{Red}\right) + \beta_2 Blue
\]

\[
SDT.4 SDT = \beta_0 + \beta_1 Red + \beta_2 Green + \beta_3 Blue
\]

\[
SDT.5 SDT = \beta_0 + \beta_1 Red + \beta_2 Green + \beta_3 Blue + \beta_4 Easting
\]

\[
SDT.6 SDT = \beta_0 + \beta_1 Red + \beta_2 Green + \beta_3 Blue + \gamma_4 Pond
\]
\text{SDT.7} \log(\text{SDT}) = \beta_0 + \beta_1 \text{Red} + \beta_2 \text{Green} + \beta_3 \text{Blue} + \beta_4 \text{Easting} + \gamma_5 \text{Pond}

\text{Chl-a.1*} \log(\text{Chl-a}) = \beta_0 + \beta_1 \text{Red} + \beta_2 \text{Green}

\text{Chl-a.2*} \log(\text{Chl-a}) = \beta_0 + \beta_1 \left(\frac{\text{Red}}{\text{Blue}}\right)^2 + \beta_2 \text{Green}

\text{Chl-a.3} \log(\text{Chl-a}) = \beta_0 + \beta_1 \text{Red} + \beta_2 \text{Green} + \beta_3 \text{Blue}

\text{Chl-a.4} \text{Chl-a} = \beta_0 + \beta_1 \text{Red} + \beta_2 \text{Green} + \beta_3 \text{Blue}

\text{Chl-a.5} \log(\text{Chl-a}) = \beta_0 + \beta_1 \text{Red} + \beta_2 \text{Green} + \beta_3 \text{Blue} + \beta_4 \text{Easting}

\text{Chl-a.6} \text{Chl-a} = \beta_0 + \beta_1 \text{Red} + \beta_2 \text{Green} + \beta_3 \text{Blue} + \beta_4 \text{Easting}

\text{Chl-a.7} \log(\text{Chl-a}) = \beta_0 + \beta_1 \text{Red} + \beta_2 \text{Green} + \beta_3 \text{Blue} + \gamma_4 \text{Pond}

\text{Chl-a.8} \log(\text{Chl-a}) = \beta_0 + \beta_1 \text{Red} + \beta_2 \text{Green} + \beta_3 \text{Blue} + \gamma_4 \text{Pond} + \beta_5 \text{Easting}

\text{OD.1} \log(\text{OD}) = \beta_0 + \beta_1 \text{Red} + \beta_2 \text{Green} + \beta_3 \text{Blue}

\text{OD.2} \text{OD} = \beta_0 + \beta_1 \text{Red} + \beta_2 \text{Green} + \beta_3 \text{Blue}

\text{OD.3} \log(\text{OD}) = \beta_0 + \beta_1 \text{Red} + \beta_2 \text{Green} + \beta_3 \text{Blue} + \beta_4 \text{Easting}

\text{OD.4} \text{OD} = \beta_0 + \beta_1 \text{Red} + \beta_2 \text{Green} + \beta_3 \text{Blue} + \beta_4 \text{Easting}

\text{OD.5} \log(\text{OD}) = \beta_0 + \beta_1 \text{Red} + \beta_2 \text{Green} + \beta_3 \text{Blue} + \gamma_4 \text{Pond}

\text{OD.6} \log(\text{OD}) = \beta_0 + \beta_1 \text{Red} + \beta_2 \text{Green} + \beta_3 \text{Blue} + \beta_4 \text{Easting} + \gamma_5 \text{Pond}

*denotes models used by Ramakrishna and Chang (2009)

Random forests (RF) is a powerful alternative to linear methods for prediction and classification, particularly when the relationships are non-linear or involve interactions among variables. To be familiar with RF, one needs an understanding how a RF is constructed and what the benefits and shortcomings of the methodology are. As the name suggests, a RF is a collection of trees and so the first step is an in-depth look at classification and regression trees.

A classification or regression tree is a series of binary divisions of the sample space of the predictor variables in the training data into increasingly homogeneous groups with respect to the
response variable. When the response is numerical and continuous regression trees are used and when the response variable is categorical (a classification), classification trees are used. The measure of homogeneity for a regression tree is the residual sum of squares (RSS), given by

$$RSS = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$  

Initially RSS is computed as

$$RSS = \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

where \(\bar{y}\) is the overall average of the response variable. Next, a predictor variable \(X_j\) with values \(x_{ij}\) and a cutoff \(c_j\) are selected to minimize

$$RSS = \sum_{\text{all } y_i \text{ such that } x_{ij} \leq c_j} (y_i - \bar{y}^{(1)})^2 + \sum_{\text{all } y_i \text{ such that } x_{ij} > c_j} (y_i - \bar{y}^{(2)})^2,$$

where \(\bar{y}^{(1)}\) is the average of all \(y_i\)'s for which \(x_{ij} \leq c_j\), and \(\bar{y}^{(2)}\) is the average of all \(y_i\)'s for which \(x_{ij} > c_j\).

The process continues until the RSS is not reduced by further splitting at which point the tree is said to be fully grown, and the final groups of data are called terminal nodes. Often there is a terminal node for each distinct combination of values of the predictor variables.

A fully grown tree usually over-fits the data because later divisions are just modeling noise in the data. To avoid over-fitting, regression and classification trees may be “pruned” to eliminate the lower branches. Pruning can be accomplished in a number of ways. For example, one can specify the smallest number of observations in a node or the smallest decrease in the RSS—called the complexity parameter—that is permitted for a split to take place.

An example of a regression tree is shown in figure 9. The regression tree is fit from the October image data rounded to the nearest 10 meters, to predict SDT. The tree has been pruned so that there are 5 terminal nodes.
In RF, many regression or classification (whichever is appropriate) trees are fit to randomly selected subsets of data and the predictions from those trees are combined to yield a more accurate prediction that may be obtained from any single tree. A bootstrap sample of the data is defined to be a sample of the same size as the dataset, drawn from the data with replacement. In a typical bootstrap sample, about 63% of the original data occur in the bootstrap sample one or more times. The remaining observation in the original dataset that do not occur in the bootstrap sample are said to be out-of-bag (OOB) with respect to that sample.

The RF algorithm begins by selecting many (the default is 500) bootstrap samples from the training data. To each bootstrap sample a regression or classification tree is fit with the restriction that at each node only a randomly selected subset of variables is available for splitting. For regression trees the default number of variables available is one third of the total number of variables; for classification trees it is the square root of the number of observations. The trees
from these bootstrap samples are fully grown. That is, no pruning of the trees takes place. For each tree, predictions are obtained for the observations that are out-of-bag for the bootstrap sample for that tree. Finally, a single predicted value is obtained for each observation by averaging the out-of-bag predictions, if the tree is a regression tree, or by voting the out-of-bag predictions in the case of a classification tree. The fact that only the out-of-bag predictions are combined to get the overall predictions ensures that RF does not over-fit and so the OOB error rate accurately estimates the true generalization error rate.

RFs are used for prediction in this application because they come with them some very attractive benefits:

- RFs are completely non-parametric. Thus:
  - There are no distributional assumptions (e.g., normality) on the predictor or response variables.
  - The relationship between the predictor variables and the response variables is not assumed to be linear and, indeed, may be highly non-linear.
  - Complex interactions among the effects of the predictor variables on the response variable may be modeled.
- RF does not over-fit the data.
- The RF algorithm is efficient in handling large data sets, with large numbers of predictor variables
- RF do not require the tuning of many model parameters
- The out-of-bag error rate for RF is an unbiased estimate of the generalized prediction error rate.
RF provide a measure of relative variable importance

RF outperforms other prediction methods in most applications. (Breiman & Cutler 2004, Cutler et. al. 2007, Cutler & Stevens 2006)

One drawback to RF is that, unlike linear models, they do not have a simple mathematical formula for how to compute the predictions. It is, in a sense, a “black box.” An input value of predictor variables is given, and a predicted response is output but the exact relationship between the predictor variable values and the predicted response variable value is not clear.

The following sets of predictor variables are used to predict all surrogates of algal concentration (SDT, Chl-a, and OD):

Set 1: Red, Green, Blue
Set 2: Red, Green, Blue, Easting
Set 3: Red, Green, Blue, Pond
Set 4: Red, Green, Blue, Easting, Pond

2.2.6. Comparison procedure

The purpose of this project is to provide the USTAR bio-fuels group with the most accurate and efficient method of determining algal concentrations in the lagoons. From the cost analysis section it is clear that it would be unreasonable and uneconomical to manually sample every time harvesting is to be done. The aerial image data combined with the manually sampled data allows us to determine an appropriate prediction method that best suits the lagoons. The methods of linear models and RF are compared on the basis of the percent of variability that is able to be explained by the method. As mentioned earlier RF has an unbiased estimate of the
generalized prediction error rate from the OOB observations. One minus this generalized prediction error rate is output when RF is run as the percent of variability explained by the prediction method. The percent of variability explained by the linear models will be taken as 100% times the adjusted r-square value.

Before the models and RF are compared, a preliminary analysis of how well each predictive method does with the different degrees of area averaging will take place. Once the amount of averaging has been determined, the linear models and RF will be compared. Models that contain the same predictor variables as a RF are compared to each other. To give an idea of how well predictions could be made using image data taken under the same conditions using a given prediction method, 95% confident intervals are constructed around the point estimates for each method. These confidence intervals are constructed from the appropriate F distributions using the R function CI.Rsq (R development core team, 2008).

The prediction method with the greatest predictive ability and most reasonable predictions will be investigated further to determined when this method can be appropriately and effectively used by the USTAR bio-fuels project to predict algal concentrations.

3. Results

3.1. Experimental Design:

3.1.1. A factorial design with repeated measures:

The task here was to help Erick Griffiths design and analyze an experiment that would be able to determine the effect of different nitrogen sources on algal growth and the limiting factors for that growth. The data received from this experiment contained 96 entries with no missing values. From this point each of the responses of the experiment are analyzed separately.
The response TSS, which measures algal growth, resulted in data that did not fit model assumptions of normality of the residuals so the log transformation of TSS was analyzed. This transformation fixed the problems of non-normality of the residuals while satisfying the assumption of constant variance of the residuals over the factor levels (Residual plots are included in the appendix, figures 1-3). The repeated measures of this experiment were taken into account by fitting the \( \text{ar}(1) \) covariance structure based on an AIC model selection criteria. The type III tests of the fixed effects for the model show that the interaction between Nsource and Time is significant. To help characterize this interaction a plot of the lsmeans over time for the different Nsources is produced, see figure 10.

**Figure 10:** The interaction plot of Griffiths’s experiment testing the effect of different nitrogen sources on algal growth over time. The blue line with U’s represents urea, the green N’s represents nitrate, the teal M’s represent the mean value at each time point, the red A’s represents ammonia, and the black C’s represents the control group. The y-axis of this graph is the log TSS, or log growth of algae, and the x-axis is the time in weeks where week 1 represents the initial concentrations of algae.
The significant interaction detected from the model is due to the leveling off of the log TSS from the control, and that it takes longer for ammonia to stabilize. This is leveling off of the control is due to the fact that the media of the control consisted of water from the effluent of the lagoons, which has less nutrients available so the growth could not continue at the same exponential rate.

From the interaction plot one can see that the estimated log TSS for algae grown using urea is consistently greater over time than algae grown using another Nsource. A look into the least squares means of Nsource allows us to determine if that difference is significant. The adjusted p-value comparing the log TSS from algae grown on urea vs. nitrate is 0.1984, indicating that urea is not significantly better than nitrate. Comparing urea and ammonia there is a significant difference (adjusted p-value = 0.0015). Comparing nitrate and ammonia there is also a significant difference (adjusted p-value = 0.0044). Urea, nitrate and ammonia are all significantly better than the control (adjusted p-values 0.0001, 0.0002, and 0.0020 respectively). The significance grouping of the Nsources can be summarized in table 1:

Table 1: Comparison of mean log(TSS) for the four nitrogen sources. Nitrogen sources that are not significantly different results are indicated in the table by sharing the same letter.

<table>
<thead>
<tr>
<th>Significance group</th>
<th>Mean: log(mg/L)</th>
<th>Nsource</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5.60</td>
<td>Urea</td>
</tr>
<tr>
<td>A</td>
<td>5.50</td>
<td>Nitrate</td>
</tr>
<tr>
<td>B</td>
<td>5.17</td>
<td>Ammonia</td>
</tr>
<tr>
<td>C</td>
<td>4.78</td>
<td>Control</td>
</tr>
</tbody>
</table>

These results suggest that future experiments under the same conditions can expect algae in media of urea or nitrate to grow more than algae grown in media of ammonia or the effluent of the lagoons.

The results of the chemical analysis are combined and briefly summarized. The model assumptions for each of the chemical analyses were roughly satisfied with the original data so no
transformations were needed. The interactions between the nitrogen source and time are all significant for the models of the individual chemical responses. The best way to characterize these interactions is through their respective interaction plots (appendix figures 4-7).

These interaction plots show an inverse pattern of the growth curve (shown by TSS in figure 10). As the algae start to grow they use the nutrients available, this decreases the chemical concentrations. It is interesting to note that the point at which the nutrients begin to level off (about week 5) is the same time that the growth of the algae begins to slow, and level off. This point of leveling off is more obvious in the ammonia, orthophosphate, and total phosphorus concentrations and less obvious in the total nitrogen concentrations. This indicates that the concentration of ammonia, orthophosphate and total phosphorus play large roles in limiting growth, while total nitrogen does not seem to limit growth till a few weeks later.

These results give a clear indication that the addition of supplementary nitrogen in the form of nitrate or urea to water taken from the effluent of the lagoons will improve algal growth and helps lower phosphorus to levels below the state of Utah’s proposed regulations.

3.1.2. A nested factorial design:

Christensen’s experiment investigated the effects of different solid supports for algal growth on a RPB over time. The experiment looked at 8 different substrates at three different time periods and there were three replicates. The data received for this experiment had 72 entries with no missing data. The model assumption for the residuals of normality was readily satisfied but after examining the plot of the residuals against the predicted values and the residuals against the factor levels it was determined that there was unequal variance across the groups of substrates with evidence of homoscedasticity. This is due to the lack of any growth on two of the
eight substrates: nylon, and polypropylene. These substrates were dropped from the analysis because the information needed about these substrates was qualitatively determined for this experiment and do not need to be formally compared to the other substrates because neither substrate yielded any harvestable algae. This allows us to proceed with the analysis, comparing the other substrates against each other. The assumption of homoscedasticity and constant variance were now satisfied, along with the normality of the residuals (Plots the residuals are included in the appendix, figures 8 – 10). The type III fixed effects for the model shows that the interaction between time and substrate is significant (p-value <0.0001). This requires the characterization of the interaction before further analysis can be done (see figure 11).

**Figure 11:** Mean value of biomass (in mg/m$^2$) for the different substrates overtime. The black line with 1’s represents the average of all of the different substrates over time.
The interaction between the substrate and time is difficult to characterize but it is from observations made during the experiment it was noticed that each substrate went through a period of growth and then when the substrate reach a certain point of biomass, parts of the algal bio-film began to slough off. Because the main goal of this experiment is to determine which substrate performs best after a set amount of time (about 20 days) the experiment could be broken up to three separate experiments, one for each time point. Specifically the time points from 22 and 26 days are analyzed as their own experiment with 1 factor, substrate, with 8 levels. For these separate analyses, polypropylene and Nylon were included in the models without a violation of the assumptions.

The data from harvesting on day 22 shows that the cotton rope has the greatest yield with an estimated 51.564 grams of algae per meter square. Cotton rope had significantly more algae than the other substrates tested. The results for the substrates after 22 days are summarized in table 2:

**Table 2: Mean biomass totals (in g/m² for different levels of substrate. Substrates that are not significantly different share the same letter. Data from Day 22.**

<table>
<thead>
<tr>
<th>Significance group</th>
<th>Mean: g/m²</th>
<th>Substrate</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>51.6</td>
<td>Cotton Rope</td>
</tr>
<tr>
<td>B</td>
<td>35.2</td>
<td>Flannel</td>
</tr>
<tr>
<td>C</td>
<td>29.1</td>
<td>Jute</td>
</tr>
<tr>
<td>C</td>
<td>29.0</td>
<td>Cotton Drop</td>
</tr>
<tr>
<td>C</td>
<td>27.9</td>
<td>Acrylic</td>
</tr>
<tr>
<td>D</td>
<td>10.9</td>
<td>Polyester</td>
</tr>
<tr>
<td>E</td>
<td>0.0</td>
<td>Polypropylene</td>
</tr>
<tr>
<td>E</td>
<td>0.0</td>
<td>Nylon</td>
</tr>
</tbody>
</table>

Data from day 26 shows a similar pattern. Cotton rope still has the highest yield and the rest are summarized in table 3:
Table 3: Mean biomass totals (in g/m$^2$) for different levels of substrate. Substrates that are not significantly different share the same letter. Data from Day 26.

<table>
<thead>
<tr>
<th>Significance group</th>
<th>Mean: g/m$^2$</th>
<th>Substrate</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>54.4</td>
<td>Cotton Rope</td>
</tr>
<tr>
<td>B</td>
<td>41.9</td>
<td>Jute</td>
</tr>
<tr>
<td>B</td>
<td>40.1</td>
<td>Flannel</td>
</tr>
<tr>
<td>B</td>
<td>37.7</td>
<td>Cotton Drop</td>
</tr>
<tr>
<td>C</td>
<td>20.9</td>
<td>Acrylic</td>
</tr>
<tr>
<td>D</td>
<td>8.1</td>
<td>Polyester</td>
</tr>
<tr>
<td>D</td>
<td>0.0</td>
<td>Polypropylene</td>
</tr>
<tr>
<td>D</td>
<td>0.0</td>
<td>Nylon</td>
</tr>
</tbody>
</table>

Johnson (2009) suggested that after about 10 days algal growth leaves the exponential growth phase and begins to stabilize by about 20 days. Johnson’s research used algae cultures grown on RPBs from a continually growing stock, whereas Christensen’s research is growing algae from the effluent of the Lagoons which has just been treated with chlorine. Christensen estimates that there is a 10 day lag phase where the algae recover from the chlorine shock and begin growing under the new conditions implemented by the RPB. This implies that the algae for these growth experiments leave the exponential growth phase at about 20 days and stabilize by 30 days. Harvesting should take place as soon as the algae leave the exponential growth part of the curve, so the original data is also examined to see if there is a significant difference in biomass on the RPB at 22 and 26 days. This is done using the least squares means from *proc mixed* in SAS. Comparing the mean biomass of time period 22 with time period 26 gives an estimate of -3.4054 mg/m$^2$ which is not statistically different from zero (Tukey’s adjusted p-value 0.1761). This implies that harvesting after 22 days would not give significantly different yields than harvesting after 26 days.

The results of this experiment have led Christensen to look closer into when the algae could most effectively be harvested from the RPBs. He has also begun looking into effective ways to harvest algae from cotton rope and has developed a mechanism that harvests the algae.
from the cotton rope and then rewraps it onto the RPB. This device has a provisional patent filed while more testing determines the need for a patent.

**Figure 12:** Reactors after 14 days (right before the first harvest). Notice the yellow rope (polypropylene) visibly contains no algae whereas some of the others already have a thick algal bio-film. (Christensen 2010)

### 3.2. Logan Lagoons

There have been two successful, complete sample surveys of the lagoons to date. The data collected has provided valuable information to the Bio-fuels Center about the ponds. Through these two complete surveys we have data to show us how well the ponds are mixed, where the most effective treatment is taking place, and where the nutrient rich areas are that can produce algae.

The data set obtained in July has 98 sampling locations with no missing data of the algal surrogate variables. There was some trouble with the dissolved oxygen, and temperature probe causing 18 missing values. Some of the samples used in the chemical analysis were lost causing 5 missing values, and only one of the sampling crews had a TSS probes causing 45 missing values. When the pixel locations of the IKONOS image were rounded to the nearest 1 and 2 meters there were not matches for every observation so the training data sets for these sizes of rounding areas contain only 10 and 25 observations, respectively.
The dataset from the October sampling contains 100 observations with only one missing value for TSS, OD and ammonia. The sample from location 47 was lost before measurements of these variables could be made. There are also 47 missing values of TSS from the probe because only one sampling crew had a probe. There were no problems in rounding the pixels of the image data and matching them to sampling locations.

3.2.1 Prediction maps from manual sampling:

Figures 13 – 20 show box-plots, variograms, prediction maps and standard deviation maps of SDT from July and October sampling data. The prediction maps of the other algal surrogates for both July and October are shown in the appendix figures 11 – 26. The prediction and standard deviation maps have an outline of where the roads between the ponds are. Figures 13 – 16 show graphs and prediction maps of SDT from the July Sampling data.

Figure 13: Distribution of SDT by pond from the July sampling. The y-axis shows SDT in inches.
Figure 14: Variogram of the July data for SDT. Distances between data points are shown in meters on the x-axis and the y-axis shows the variance. The estimated variance is shown as hollow dots for a given distance h and the fitted covariance structures are shown in red, blue and green corresponding to Gaussian, spherical, and exponential respectively. The exponential covariance structure was determined to best fit the July SD data.

Figure 15: Predicted values of SDT in inches, which are inversely related to algal concentration so lower values of SDT, indicated in the prediction map by blues and purples, denote areas of high concentrations of algae.
Figure 16: Standard deviation map for the prediction map in figure 15. Blue areas show where there is greater predictive accuracy. This indicates that our greatest predictive accuracy is in the areas that we have the highest concentrations of algae.

The sets of box-plots show highest concentrations of algae exist in the C ponds. The prediction maps confirm this and also give more precise information about the distribution of the concentrations of algae. The prediction map indicates increase algae concentrations at the south end of the C pond. Also, the prediction map shown and those in the appendix (figures 12 through 27) show increased algal concentrations on the east side of pond B2. These areas were, during sampling noted to have algal blooms, confirming these findings.
Figures 17 – 20 show graphs and predictions maps of SDT from the October sampling data.

**Figure 17:** Distribution of SDT by pond from the October sampling. The y-axis shows SDT in inches.

**Figure 18:** Variogram of the October data for SDT. This plot shows the distances between data point in meters on the x-axis and the y-axis shows the variance. The estimated variance is shown as hollow dots for a given distance $h$ and the fitted covariance structures are shown in red, blue and green corresponding to Gaussian, spherical, and exponential respectively. The covariance structure that was determined to best fit the July SD data was the Gaussian covariance structure.
**Figure 19:** Predicted values of SDT in inches, which are inversely related to algal concentration so lower values of SDT, indicated in the prediction map by blues and purples, denote areas of high concentrations of algae.

**Figure 20:** Standard deviation map for the prediction map in figure 19. Blue areas show where there is greater predictive accuracy.
The box-plots of the variables by pond from the October sampling data showed highest concentrations of algae in ponds A1, and A2. This is confirmed by the prediction maps that show that higher algae concentrations are found in A1, and A2, especially towards the influent on the east side of the ponds.

From the sampling data collected in July and in October there are some impressive differences between the results. The lagoons in the October sampling were much cooler. This cooling seemed to stabilize the ponds, in that all of the other measured variables were more consistent throughout individual ponds as compared to the July sampling. The variability for the October data was consistently lower than that of the July data as illustrated by smoother transitions in the prediction maps rather than the blotchy irregular pattern from the July data. It is also clear that the cooler temperature hampered algal growth as evident by the lack of algal blooms in the October sampling. This was noticed at the time of sampling and confirmed through the data collected in October. All algal surrogates from the October sampling showed lower concentrations of algae and overall less variability as the July data.

3.2.2. Rounding effects:

An important step in determining the predictive ability of the linear models and RF is knowing how large to make the pixels that are matched with the data from the sampled locations. This section explores the predictive ability of all of the linear models and RF with the intensities of red, green, and blue from pixels rounded to the nearest 1, 2, 5, 10, 12.5, 20, 25, 33.3, 50 and 100 meters. The adjusted $R^2$ values of each degree of rounding are displayed with those of the same model. For efficient comparisons of these results, the y-axis of the graphs have been set to (0,1) and the x-axis labeled with the degree of rounding. The results of the RFs are displayed
with box-plots created by 25 runs with the same sets of predictor variables for each surrogate of algal concentration and each degree of rounding. Each set of predictor variables is displayed separately (See figures 21 – 30).

Once it has been determined what size area of rounding is most appropriate, the results of the linear models (and their assumptions) are discussed in more detail.

The primary reason for averaging is that the GPS units used for the physical sampling of the ponds are accurate only to 10 – 20 feet while the image data for July was accurate to 1 – 2 meters and for October was accurate to 0.25 meters. Training data sets that rounded pixel and sampling locations to the nearest 1, 2, 5, 10, 12.5, 20, 25, 33.3, 50 and 100 meters were created to determine what degree of rounding should be used in the analysis. Pixels that share the same location after rounding were then averaged to create a single value for each wavelength. The image and sampling data sets are merged and then fit to every proposed linear model and random forest. As mentioned earlier, problems arose when merging the July image data that had been rounded to the nearest 1 and 2 meters because not every sampling location had a pixel that matched. This caused these data sets to have only 10 and 25 observations respectively. This means that the first two dots in the July linear model and RF plots are not reliable, and should therefore be disregarded.

Figures 21 – 30 show the percent of variability explained by each predictive method at every degree of rounding.
Figure 21: Each plot shows the adjusted $R^2$ of the denoted model by degree of rounding using the July training data. Because of the problems in matching the image data to the sampling data the first two adjusted $R^2$ should be disregarded.

Figure 22: RF for each surrogate of algal concentration and predictor variables red, green, and blue using the July training data. Because of the problems in matching the image data to the sampling data the first box plots should be disregarded.
Figure 23: RF for each surrogate of algal concentration and predictor variables red, green, blue, and easting using the July training data. Because of the problems in matching the image data to the sampling data the first box plots should be disregarded.

Figure 24: RF for each surrogate of algal concentration and predictor variables red, green, blue, and pond using the July training data. Because of the problems in matching the image data to the sampling data the first box plots should be disregarded.
Figure 25: RF for each surrogate of algal concentration and predictor variables red, green, blue, easting, and pond using the July training data. Because of the problems in matching the image data to the sampling data the first box plots should be disregarded.

Figures 26 – 30 show the percent of variability explained by each predictive method at every degree of rounding.

Figure 26: Each plot shows the adjusted $R^2$ of the denoted model by degree of rounding using the October training data.
Figure 27: RF for each surrogate of algal concentration and predictor variables red, green, and blue using the October training data.

Figure 28: RF for each surrogate of algal concentration and predictor variables red, green, blue, and easting using the October training data.
Figure 29: RF for each surrogate of algal concentration and predictor variables red, green, blue, and pond using the October training data.

Figure 30: RF for each surrogate of algal concentration and predictor variables red, green, blue, easting, and pond using the October training data.
These graphs show that rounding to only 1 or 2 meters is often too little and rounding to 50 and 100 meters is too much. Consistently through the models and RFs, rounding to the nearest 5, 10 and 12.5 meters allows the predictive method to explain the most variability in the data. These results allow us to proceed further into the investigation. From this point all predictive methods will be using image data that has been rounded to the nearest 10 meters.

3.3.3. Method Comparison

The two methods are evaluated so that the linear models that contain the same predictor variables as the RFs will be compared. The models and RF will be examined for the best predictive ability as measured by adjusted $R^2$ for the linear models and the percent of variability explained by the RFs. Part of this comparison will be seeing how well an algal surrogate is able to be predicted from the image data alone. These results come from looking at the first few linear models for each surrogate (ones containing only a combination of Red, Green, and Blue) as well as the first RF.

The following graphs show the percent of variability explained by predictive methods, linear models and RF. The adjusted $R^2$ and % variability explained are shown with 95% confidence intervals calculated from the appropriate F-distribution. The results for the RFs that share the same predictor variables as linear models are shown to the right of the results for the linear models. The results are organized by the algal surrogate that is being predicted. The July data is examined first.
Figure 31: Percentage of variability from each predictive method for SDT using the July training data rounded to the nearest 10 meters. The results for the linear models are shown in brown and the results for the RFs are shown in green.

Figure 32: Percentage of variability from each predictive method for Chl-a using the July training data rounded to the nearest 10 meters. The results for the linear models are shown in brown and the results for the RFs are shown in green.
July OD Data explained by Predictive Methods

![Graph showing percentage of variability from each predictive method for OD using the July training data rounded to the nearest 10 meters. The results for the linear models are shown in brown and the results for the RFs are shown in green.]

Figure 33: Percentage of variability from each predictive method for OD using the July training data rounded to the nearest 10 meters. The results for the linear models are shown in brown and the results for the RFs are shown in green.

For the data from July the predictive ability is generally low. The best predictive ability is from linear models is with SDT6 and SDT7 which include pond as a covariate. These models have adjusted R² of 0.6411 and 0.6458 respectively. The RFs 3 and 4 for SDT have the same predictor variables as SDT6 and SDT7 and explain only 56.44% and 56.61%. Predictive ability of any of the algal surrogates is pretty dismal when only using the image data. The best result comes from OD1 which has an adjusted R² of only 0.3392. With the July linear models, the models that included pond and easting as predictor variables no longer had significant coefficients from red, green, or blue, indicating that in the presence of easting or pond that the pixel intensity does not make much of a difference. This is confirmed by partial dependence plot produced by RF (as shown in figures 33 – 34 in the appendix).

Now, the results from the October data training data rounded to the nearest 10 meters is examined.
October SDT Data explained by Predictive Methods

Figure 34: Percentage of variability from each predictive method for SDT using the October training data rounded to the nearest 10 meters. The results for the linear models are shown in brown and the results for the RFs are shown in green.

October Chl-a Data explained by Predictive Methods

Figure 35: Percentage of variability from each predictive method for Chl-a using the October training data rounded to the nearest 10 meters. The results for the linear models are shown in brown and the results for the RFs are shown in green.
Figure 36: Percentage of variability from each predictive method for Chl-a using the October training data rounded to the nearest 10 meters. The results for the linear models are shown in brown and the results for the RFs are shown in green.

The sampling data from October shows much improvement in predictive ability over the July data. Out of all of the predictive methods the ones that explain the most variability in the data are SDT6, SDT7, closely trailed by RFs 3, 2, and 4 predicting SDT. These linear models have adjusted r-squares of 0.8041, 0.8103, followed by the RF which explain 76.25%, 77.30%, and 79.51% of the variability in the data, respectively. While the predictive ability for SDT is fairly good for when using covariates other than the image data, it is not significantly better than the predictive ability when using only the image data due to the overlapping 95% confidence intervals. The best example of this is SDT1 and RF 1 for predicting SDT. This linear model has an adjusted $R^2$ of 0.6363, and the RF explains 68.10% of the variability in the data. This amount of predictive ability is encouraging to the project, because it shows that predictions of algae concentrations can be made with some degree of accuracy for the Logan lagoons wastewater system with only the image data.
The same pattern of non-significant coefficients for the image data variables in the presence of the other covariates exists for the October linear models. This relative variable importance for the October data is confirmed by Figures 35 – 36 shown in the appendix.

An interesting note about the linear models from July and October is that the estimates of $\tilde{\beta}$ for red were roughly the same in most of the models. This may indicate that there is an overarching pattern of the image and the algal concentration that can be modeled and used to make prediction.

These results show the aerial image data and sampling data from October do a considerably better job of predicting algae concentrations than July’s data. There are a number of factors that are contributing to the lack of ability to predict algae concentrations using the available images for the July data. Probable causes of this lack of ability are but not limited to the following: The sampling took place in two parts with five days in between the two parts; The image data was taken July 6th and the first day of sampling took place July 8th and the second day July 13th; High temperatures in the ponds cause extreme variability in algae concentration that image data alone is not able to capture; Warmer times of the year are unsuitable for using aerial imaging to predict algal concentrations.

4. Conclusions

4.1. Experimental Design

An understanding of what provides ideal conditions for algal growth and harvest is needed to meet the goals of Logan City and the Bio-fuels Center project. Researchers of this field have the challenge of determining and engineering ways to remove algae and produce bio-fuels and do so efficiently enough to make it economical. Traditional agriculture has been
researched and fine tuned for hundreds if not thousands of years. For the use of algae in bio-fuel production to be a viable option the same sort of understanding needs to be obtained of how to effectively produce and harvest algae. To this end, Logan Christensen and Erick Griffiths set up experiments to investigate how to effectively grow and harvest algae.

Griffiths’s investigated the effects of additional nutrients added the effluent of the Lagoons. The results indicate that adding additional nitrogen increases algae production, and helps decrease phosphorous levels. The best nitrogen sources to add were found to be nitrate and urea, suggesting that in future tests and large scale raceway ponds either of these nitrogen sources could be added to give significantly improved results in both algae production and phosphorus removal.

The results from Christensen’s experiment offer a viable substrate option for RPBs to grow algae. This experiment was exploratory in nature, trying to narrow the scope of what type of substrates warrant further investigation in algae production, harvest-ability and reuse-ability. The results indicate natural substrates, in general allow for a more biomass than the synthetic materials. This is evident by in tables 2 and 3 where the natural materials: cotton rope, jute, flannel, and cotton drop; for the most part have significantly greater means than the synthetic materials: acrylic, polyester, polypropylene and nylon. It was concluded in this experiment that cotton rope while allowing for significantly more algal growth to occur also was far easier to harvest and had the added benefit of the same rope being reusable. Also, qualitatively compared to all other substrates, cotton rope was much easier to harvest the algae from and remained intact for future use. Christensen has already created a machine to automatically harvest the algae from cotton rope and wraps it back onto the RPB for continuous use (See figures 37 and 38). Future
experiment will focus on optimizing the effectiveness of RBPs while using cotton rope and the algae-spool harvester.

**Figure 37**(left): *The algae-spool harvester on a dry run. Figure 38** (right): *The algae-spool harvester is shown unwrapping, harvesting and re-spooling cotton rope. (Christensen 2010)*

**4.2. Logan Lagoons**

The goals of the Bio-fuels Center project and Logan City depend on the ability to effectively harvest algae from the Lagoons. This goal can be realized if there is the ability to predict concentrations of algae in the Lagoons. Determining the spatial distribution of algae concentrations is made much simpler with the use of aerial images. This project has shown that the ability to predict algal concentrations exists. Several models and methods for predicting algal surrogates were investigated to ensure that the most accurate predictions are made from the image data. The methods were compared using the image data and other covariates. RF consistently had greater predictive ability using only the image data, as measured by the percent of variability explained. The highest percent explained by a RF using only the image data was 68.10% while the best linear model only explained 63.63% (RF1 for SDT and SDT.1 using the October data). Linear models generally outperformed RF when covariates other than the image
data were included in the model, but how appropriate these models are is questionable due to the lack of significance of the image variables in these models. RF performed much better in prediction with only the image data. In the future, the effect of easting and pond will not be nearly as consistent as the effects of the image data, because the algal concentrations will change throughout the lagoons but the will generally look the same. Thus, by using aerial images and RF, along with qualitative ground observations the information needed to accomplish Logan City and Bio-fuels Center goals can be obtained.

Figure 38: Mosaic of the Logan lagoons taken by the UWRL's UAV team October 25, 2009.

5. References:


Appendix:
Figures 1-9 are additional graphs from Erick Griffiths's experiment:

Plot of Residuals against NSource (TSS)
Plot of Resid*NSource. Legend: A = 1 obs, B = 2 obs, etc.

Figure 1: Residuals against the factor levels of nitrogen. Nitrogen source 1 is the control, 2 is Ammonia, 3 is Urea, and 4 is Nitrate. The assumption of constant variance is satisfied because the range of the residuals for any one factor level is not 5 times greater than that of another.
Plots of Residuals against Predicted Values (TSS)
Plot of Resid*Pred. Legend: A = 1 obs, B = 2 obs, etc.

Figure 2: Residuals against the predicted values. The assumption of homoscedasticity is approximately satisfied because there is no alarming shape that indicates a violation.
Figure 3: Residuals against the factor levels of time. The assumption of constant variance is satisfied because the range of the residuals for any one factor level is not 5 times greater than that of another.
Figure 4: Interaction plot of the nitrogen sources over time for the response variable ammonia. The red line is for the ammonia nitrogen source, the blue line is for urea, the green line is for nitrate, the black line is for the control, and the teal line is for the mean of them all. Notice how the ammonia is nearly gone after 5 days, which is approximately when the TSS leaves its exponential growth phase and begins to level off.
Figure 5: Interaction plot of the nitrogen sources over time for the response variable orthophosphate. The red line is for the ammonia nitrogen source, the blue line is for urea, the green line is for nitrate, the black line is for the control, and the teal line is for the mean of them all. Notice how orthophosphate for ammonia, nitrate and urea nitrogen sources continues to decrease for 4 – 5 more days after the control orthophosphate levels stabilize and begin to increase. This indicates that the algae growth is phosphorous limited after about 5 days.
Figure 6: Interaction plot of the nitrogen sources over time for the response variable total nitrogen. The red line is for the ammonia nitrogen source, the blue line is for urea, the green line is for nitrate, the black line is for the control, and the teal line is for the mean of them all.
Figure 7: Interaction plot of the nitrogen sources over time for the response variable total phosphorus. The red line is for the ammonia nitrogen source, the blue line is for urea, the green line is for nitrate, the black line is for the control, and the teal line is for the mean of them all. Here the total phosphorus levels decrease till day 5 then more or less stabilize.
Figures 8-10 are residual plots from Logan Christensen’s experiment:

Plots of the Residuals against the predicted values
Plot of Residual*Pred. Legend: A = 1 obs, B = 2 obs, etc.

Figure 8: Plot of the residuals against the predicted values. The assumption of homoscedasticity is approximately satisfied because there is no alarming shape that indicates a violation.
Figure 9: Plot of the residuals against the factor levels of substrate. The assumption of constant variance is approximately satisfied because the range of the residuals for any one factor level is not 5 times greater than that of another.
Figure 10: Residuals against the factor levels of time. The assumption of constant variance is approximately satisfied because the range of the residuals for any one factor level is not 5 times greater than that of another.
Figure 11: Distribution of OD by pond from the July sampling data to show how the distribution changes through the lagoons. The y-axis shows OD. Notice the large degree of variability in the ponds where there were high concentrations of algae: ponds B2 and C.

Figure 12: Variogram of the July data for OD. This plot shows the distances between data point in meters on the x-axis and the y-axis shows the variance. The estimated variance is shown as hollow dots for a given distance h and the fitted covariance structures are shown in red, blue and green corresponding to Gaussian, spherical, and exponential respectively. The covariance structure that was determined to best fit the July OD data was the spherical covariance structure.
Figure 13: Predicted values of OD in July. OD is directly related to algal concentration so higher values of OD, indicated in the prediction map by oranges and yellow, denote areas of high concentrations of algae. It is clear in this prediction map that the highest concentrations are found in B2 and C ponds.

Figure 14: Standard deviation map of the July OD prediction map of figure 13. Blue areas show where there is greater predictive accuracy. This indicates that the covariance structure fit the data fairly well, giving no more than about 0.100 for standard deviation for each prediction point.
Figure 15: Distribution of Chl-a by pond from the July sampling data to show how distribution of SDT changes through the lagoons. The y-axis shows concentration of Chl-a in mg/L. Notice the large degree of variability in ponds B2 and C where high concentrations of algae were observed.

Figure 16: Variogram of July data for Chl-a. The distances between data point is shown on the x-axis and the y-axis shows the variance. The estimated variance is shown as hollow dots for a given distance h and the fitted covariance structures are shown in red, blue and green corresponding to Gaussian, spherical, and exponential respectively. The covariance structure that was determined to best fit the July Chl-a data was the exponential covariance structure.
Figure 17: Predicted values of Chl-a in mg/L for July. Chl-a is directly related to algal concentration so higher values of Chl-a, indicated in the prediction map by oranges and yellow, denote areas of high concentrations of algae. It is clear in this prediction map that the highest concentrations are found in B2 and C ponds.

Figure 18: This shows a standard deviation map of the prediction map of figure 17. Blue areas show where there is greater predictive accuracy. This indicates that the covariance structure fit the data pretty poor, with an average standard deviation in the predictions of more than 1000 mg/L.
Figure 19: This shows the distribution of OD by pond from the October sampling data to show how the distribution changes through the lagoons. The y-axis shows the OD. Note the generally decreasing trend of the OD in the lagoons.

Figure 20: Variogram of the October data for OD. The distances between data point are shown on the x-axis and the y-axis shows the variance. The estimated variance is shown as hollow dots for a given distance h and the fitted covariance structures are shown in red, blue and green corresponding to Gaussian, spherical, and exponential respectively. The covariance structure that was determined to best fit the October OD data was the exponential covariance structure.
Figure 21: This shows the predicted values of OD. OD is directly related to algal concentration so higher values of OD, indicated in the prediction map by orange and yellow. This prediction map shows the same general pattern as SDT and Chl-a.

Figure 22: This shows a standard deviation map of the October OD prediction map of figure 21. This plot shows that the standard deviation for each prediction point is no more than 0.020.
Figure 23: This shows the distribution of Chl-a by pond from the October sampling data to show how the distribution changes through the lagoons. The y-axis shows the concentration of Chl-a in mg/L.

Figure 24: This is the variogram of the October data for Chl-a. This plot shows the distances between data point in meters on the x-axis and the y-axis shows the variance. The estimated variance is shown as hollow dots for a given distance h and the fitted covariance structures are shown in red, blue and green corresponding to Gaussian, spherical, and exponential respectively. The covariance structure that was determined to best fit the October Chl-a data was the exponential covariance structure.
Figure 25: This shows the predicted values of Chl-a in mg/L for October. This prediction map shows more specifically where algae activity was taking place: on the edges of A1, A2, and B2.

Figure 26: This shows a standard deviation map of the October Chl-a prediction map of figure 25. The purple areas show where there is greater predictive accuracy.
Figure 27: These show normal quantile plots of the residuals from the July SDT models. Note: SDT3- SDT7 are approximately normally distributed, implying these models can be generalized for obtaining predictions of future image data.

Figure 28: These show normal quantile plots of the residuals from the July Chl-a models. Note: Chl-a3, and possibly Chl-a5, Chl-a7, and Chl-a8 (Chl-a7 and Chl-a8 with the exception of an outlier) are approximately normal, implying these models can be generalized for obtaining predictions of future image data.
**Figure 29:** These show normal quantile plots of the residuals from the July OD models. Note: OD1, OD3, OD5 and OD6 with the exception of a lower outlier are approximately normally distributed, implying these models can be generalized for obtaining predictions of future image data.

**Figure 30:** These show normal quantile plots of the residuals from the October SDT models. Note: All October SDT models are approximately normally distributed, implying these models can be generalized for obtaining predictions of future image data.
Figure 31: These show normal quantile plots of the residuals from the October Chl-a models. Note: None of the October Chl-a models satisfy the assumption of normality in the residuals, implying that these models cannot reliably be used in obtaining predictions of future image data.

Figure 32: These show normal quantile plots of the residual from the October OD models. Note: OD1 and OD3 (with the exception of 2 upper outliers) are approximately normally distributed, implying these models can be generalized for obtaining predictions of future image data.
Figure 33(left): This plot shows the relative importance of each predictor variable in RF #1 for predicting SDT using the July data, red being the most important and with decreasing importance green then blue. Variable importance is determined by how many time that variable is selected for a division in a tree in the RF. Figure 34 (right): This plot shows the relative variable importance of each predictor variable in RF #4 for predicting SDT using the July data. Comparing this plot with figure 33 (left), notice that pond is the most important.

Figure 35(left): This plot shows the relative importance of each predictor variable in RF #1 for predicting SDT using the October data, red being the most important and with decreasing importance green then blue. Figure 36 (right): This plot shows the relative variable importance of each predictor variable in RF #4 for predicting SDT using the October data. Comparing this plot with figure 35 (left), notice that pond and easting are the most important with a gap before the image data variables.
Figure 37: Pair wise plots of the July sampling and image data.
Figure 38: Pair wise plots of the October sampling and image data.