Tong Wang
Spectral Vegetation Indices and Their Relationships With Nitrogen Stress Levels
(Under the direction of Dr. Chi N. Thai)

This thesis describes the development of statistical models to determine the relationships between vegetation indices and bush bean plant nitrogen stress levels thereby help early detection of bush bean nutritional stress. Two approaches, statistical analysis by SAS procedure and genetic algorithm (GA) were employed. Polynomial regression was used to fit data under all possible two-bands combinations. The best bands identified were 700nm, 710 nm, 720 nm, and 750 nm. The best model selected by both SAS and GA was second degree polynomial for RVI with an adjusted R^2 value of 0.9144. Comparison of these two approaches showed that SAS procedure could provide accurate result but was less efficient. GA integrated both statistical analysis and model selection and was more efficient.

INDEX WORDS: Spectral reflectance, Vegetation index, Genetic algorithm, Polynomial regression.

Spectral Vegetation Indices and Their Relationships

With Nitrogen Stress Levels

By

Tong Wang

B.S., Fujian Medicial University, 1992

A Thesis Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2002

Spectral Vegetation Indices and Their Relationships

With Nitrogen Stress Levels

By

Tong Wang

Approved:

Major Professor: Chi N. Thai

Committee: Khaled Rasheed

Ron McClendon

Electronic Version Approved:

Gordhan L. Patel
Dean of the Graduate School
The University of Georgia
August 2002

ACKNOWLEDGMENTS

I would like to thank Dr. Chi N. Thai, my major professor, for his continuous guidance, encouragement and financial support through this research.

I would also like to thank Dr. Ron McClendon and Dr. Khaled Rasheed for their invaluable advice and suggestion for this research and sitting on my thesis committee. Thanks for Dr. J Reeves for his advice on the statistical analysis. Thanks for my friends in AI center and BAE department for their helps.

Special thanks for my family for their support, encouragement and love during my graduate study.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Remote sensing has been used in agriculture to map the health status of crops for many years. Early detection of crop stress is important for environmental and economic concerns. Crop stresses are influenced by plant biochemicals such as chlorophyll, and these biochemical contents can be estimated from spectral reflectance characteristics of plants (Hatfield and Pinter, 1993). This mapping procedure consists two stages: image processing and spectral data analysis.

1.1 Image processing

Two types of remote sensing technologies have been used in acquiring spectral images. One type uses broadband spectral imaging from aircraft or satellites. This approach has been successfully used for years. But its sensitivity is limited by the relatively low spatial and spectral resolution of the satellite images (Moran et al., 1997). This technique is suitable to monitor forest damage and crops grown in large areas.

The other group is ground based, which use narrow spectral imaging bands with imaging cameras located just a few meters above the crop canopy (Carter and Miller, 1994). Higher spatial and spectral resolution can be achieved with this approach. Recent literature has shown that the narrow bands may be crucial for providing additional information with significant improvements over broad bands in quantifying agricultural crops.

A suitable multi-spectral imaging system is critical in order to apply spectroscopy technique at the plant canopy level. Recently, due to the advancement in optics and computer miniaturization, researchers have integrated both satellite-based and ground-based technologies. This combines the advantages of both narrow band and broad band spectroscopy techniques and encouraging results have been obtained.

Evans et al.(1998) developed a hyper-spectral imaging system based on a liquid crystal tunable filter (LCTF) (See Figure 1-1). This system, which is used in this study, uses LCTF to achieve variable and narrow band filtering of the reflected light. Images are taken with video cameras located a few meters from the plants, achieving maximum spatial and spectral resolution. Several applications of this system for evaluating plant biochemical data and stress status have been reported (Thai et al., 1998, 2000).

1.2 Spectral data analysis

To map multi-spectral images of plants to their health status, spectral data extracted from these images are usually analyzed based on absorption/reflectance or fluorescence. A vegetation index is calculated using reflectance values under two or more spectral wavelengths according to different mathematic formulae. Usually vegetation indices are more sensitive than the reflectance at a single wavelength.

To better understand the spectral reflectance characteristics of plants, a number of vegetation indices (VI) have been developed by remote sensing researchers. Vegetation indices such as NDVI, RVI have been widely used to quantify plant biochemical data and detect plant health status (see Wiegand et al., 1991; Thenkabail et al., 2000).

Previous remote sensing research has identified the best wavelengths in the visible-near infrared spectrum for plant nitrogen content: the visible 534-640 nm and far-red/NIR 680-750 nm wavelength ranges are found to be the most sensitive to plant nutritional stress. The red edge between 680 nm and 750 nm, a sharp change in leaf reflectance was identified and used for stress detection (Filella and Penuelas, 1994).

Different statistical models have been developed to determine the relationships between vegetation indices and plant biochemical data. Linear regression is commonly used to determine the basic relationships, while nonlinear models such as exponential models are applied to further improve fitting. Artificial Neural Networks have also been successfully used in spectral data analysis (Thai, et al., 1998).

When dealing with various vegetation indices, comparisons among statistical models are usually needed in order to evaluate their performance. Simple enumeration (exhaustive search) is effective for a small set of data, but an optimal search strategy needs to be derived for a large set of data in order to save time or even make the analysis possible.

Genetic Algorithm (GA), which imitates the process of natural selection and evolution, is an efficient search method. In GA, each individual is regarded as a potential solution for the current problem. GA works by generating a new generation through processes of evaluating, selecting, mutating and recombining of individuals. The evaluation and selection is based on the values of individual “fitness”. The optimal individual represents the best solution for the problem and will be eventually generated after a number of generations. GA is especially powerful when used in large, complex search space where exhaustive search is either difficult or impossible to do.

In our study, we use both exhaustive search performed by conventional statistical methods and genetic algorithm search to identify optimum statistical models as well as optimal wavelengths.

1. 3 Objectives of this study

Our study aimed to determine the optimal vegetation indices and wavelengths that could be used in spectral imaging to best characterize the plant nitrogen stress levels. More specifically, the objectives of our work were:

1. Develop statistical models to correlate the nitrogen stress levels of bush bean plant with vegetation indices based on different wavelengths. Identify the best model with the highest correlation using exhaustive search. The model development and the selection of the best model are performed in separate procedures.
2. Implement a genetic algorithm that integrates all functions performed by the statistical models described above. The model development and model selection are performed in one procedure.
3. Evaluate the performance of these two approaches based on runtime and accuracy.

This thesis is organized as follows: chapter 1 is literature review. A detailed description of the research procedure is presented in chapter 3 and chapter 4. Chapter 3 covers objective 1 mentioned above while Chapter 4 covers objective 2. A summary of our study is given in chapter 5. Chapter 6 briefly introduces our future research direction. Figure 1 – 2 illustrates the outline of this study.

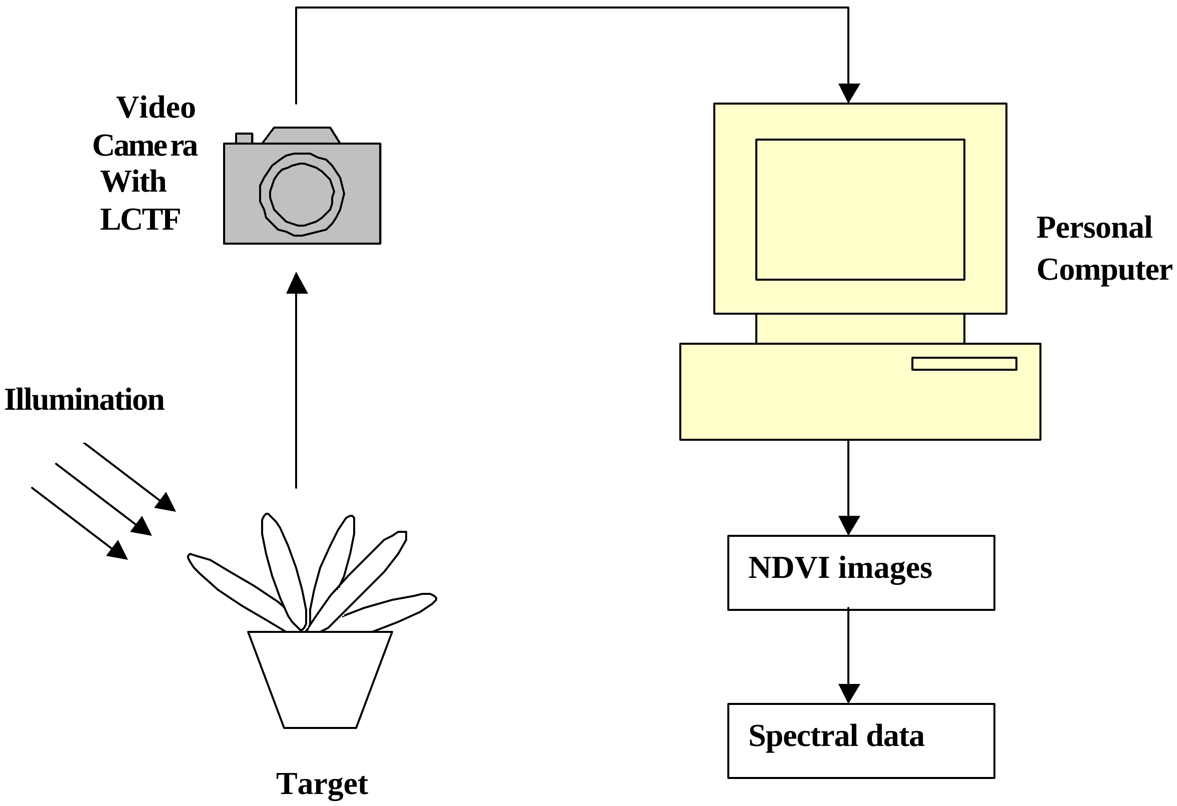


Figure 1-1 Spectral imaging system used in this research

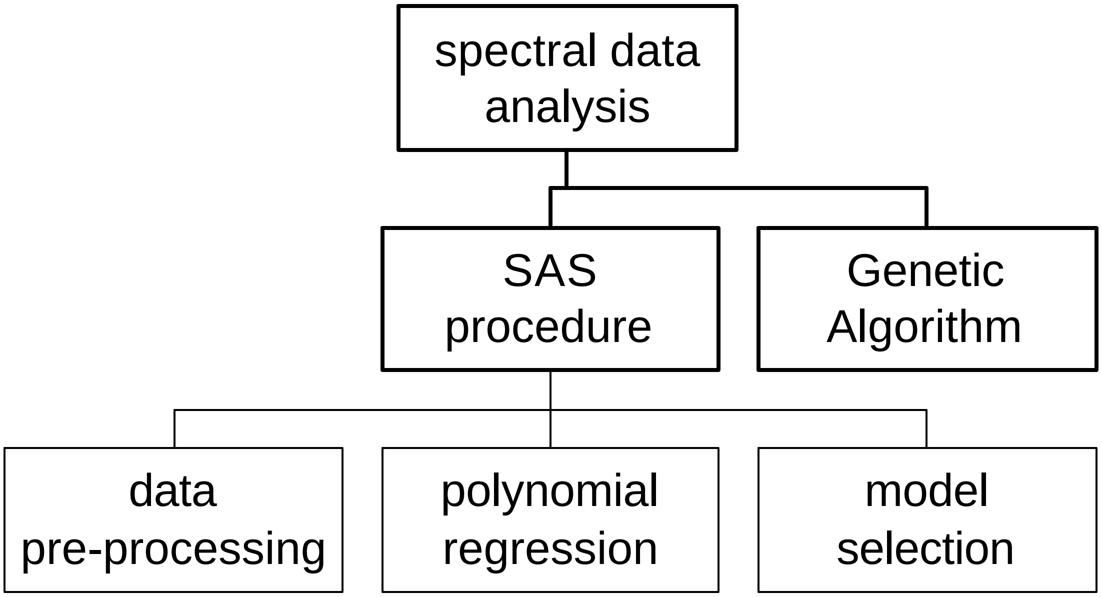


Figure 1-2 Research procedure outline

CHAPTER 2

LITERATURE REVIEW

This chapter introduces some remote sensing techniques for plant stress detection and described two spectral data analysis approaches used in our study: polynomial regression and genetic algorithm.

2.1 Research background

2.1.1 Research techniques for plant stress detection

Plants under stress usually do not function efficiently and certain biochemical changes can be detected using remote sensing technologies. Plant stresses are due to factors such as disease, nutrition deficiency and dehydration.

Two spectral methods are commonly used in remote sensing research: fluorescence and absorption/reflectance. It was found that the plant chlorophyll and blue-green fluorescence generally increases under stress. Thus red + far-red chlorophyll fluorescence and blue-green fluorescence can be used in plant stress detection. Fluorescence techniques are best suited for characterizing transient photosynthetic plant functions (Ning et al., 1995, Gitelson et al., 1999).

The reflectance technique focuses on the relationship between plant reflectance and plant photosynthetic function. Stress due to stress factors such as nitrogen deficiency reduces chlorophyll and as a result, modifies reflectance. This reflectance response is spectrally similar among agents of stress and plant species. Based on the proposition that the amount of reflectance is a function of the amount of the biochemical contents in the plant tissue, these biochemical contents can then be estimated by the measurement of reflectance. Reflectance techniques can measure more readily the permanent structures of the photosynthetic system such as chlorophyll and water content, which influence the survival and yield of the crop. A lot of successful research has been done with reflectance techniques, and good correlations between reflectance and many important biochemical variables such as biomass, nitrogen concentration and chlorophyll content have been found (Carter and Miller, 1994, Johnson and Billow 1996, Yoder and Pettigrew-Crosby, 1995).

2.1.2 Spectral characteristics important for stress detection

Agriculture remote sensing is commonly applied in the visible, near-infrared and thermal infrared portions of the spectrum for quantifying the biochemical contents of healthy and stressed plants. Green plants absorb most of the red light but very little near infrared light from sunshine for photosynthesis. Therefore the sensor above the crop receives very little red light reflected from the crop. On the other hand, most near infrared light is reflected. Conversely, plants in stress such as nitrogen deficiency will often have less chlorophyll and appear to be chloretic or yellow and can thus be detected by a decrease in red light absorbance and infrared light reflectance. Due to this important feature, the red edge is typically used for stress detection (Fillela and Penuelas, 1994).

To enhance the plant stress signal, the measured spectral reflectance data from two or more spectral wavelengths are computed into vegetation indices according to different mathematic formulae. Most vegetation indices use the red spectral band, which represents the chlorophyll level, and the near infrared (NIR) band, which represents the green vegetative biomass. These bands contain more than 90% of the information on a plant canopy. Many vegetation indices have been developed. Some very common used vegetation indices are listed in Table 2-1, among which the NDVI index is the most widely used index and has been found to perform best in several studies.

Table 2-1 Some commonly used vegetation indices and their formulae

Index	Formula
Ratio Vegetation Index (RVI)	NIR / RED
Normalized Difference Vegetation Index (NDVI)	(NIR –RED) / (NIR + RED)
Nitrogen Reflectance Index (NRI)	(NIR/GREEN) / (NIR/GREEN) ref

2.1.3 Optimum wavelengths and vegetation indices found by previous studies

Reflectance spectroscopy has been used to estimate nitrogen concentration and assist nitrogen management in agriculture and environmental research for many years. A number of optimal spectral bands and vegetation indices have been reported by previous studies.

Yoder and Pettigrew-Crosby (1995) reported that with first-difference transformations of *log (1/reflectance),* the spectral bands that correlated *log (1/reflectance)* highly with nitrogen concentration in the visible-near infrared (VIS-NIR) region were located at near 530-540 nm, 650 nm, 690 nm, 720-800 nm, 1200nm, 2070-2210 nm.

Claude and Pierre (1991) studied the relationship between leaf reflectance and leaf nitrogen concentration of broadleaf tree seedlings at the 400-800nm region and found that the highest correlations were measured in the red region of the spectrum at wavelengths 600 – 700nm.

In their studies of foliage, Johnson and Billow (1996) used NIR and visible diffuse reflectance spectral data scanned from 400 to 2498 nm. They identified the 2100-2350 nm region as the optimum wavelengths by regression analysis.

Studies by other researchers also showed different wavelength selection. These differences could be explained by many factors, including differences in water content, plant anatomy, and the concentration of cell constituents.

In addition to the search for the optimum wavelengths by reflectance measurement, many studies have been done to search for the best vegetation index.

Wanjura and Hatfield (1987) tested the sensitivity of three commonly used vegetation indices, RVI, NDVI and GVI (Greenness Vegetation Index), to crop biomass of four different species. It was reported that RVI was more sensitive to high levels of biomass and LAI (leaf area index). However, when crops were small, NDVI and GVI may be the best estimators of LAI and ground cover.

Lawrence and Ripple(1998) examined the use of seven types of vegetation indices for predicting vegetation cover in field studies and found that among the ratio-based vegetation indices, the simple ratio (RVI) and NDVI performed best under conditions of high substrate and vegetation heterogeneity.

Thenkabail et al. (2000) compared three types of vegetation indices (NDVI, Optimum Multiple Narrow Band Reflectance (OMNBR), and soil-adjusted vegetation indices) and recommended twelve types of narrow band NDVI predicators for crop variables. They also showed that OMNBR had the “over fitting” problem.

Studies by these researchers indicated that the performance of vegetation indices depended on which crop variable was to be estimated, the plant species, the atmospheric condition and optical properties of the soil background. Different vegetation indices should be selected for specific site studies.

2.2 Statistical analysis: polynomial regression

To determine the relationship between vegetation indices and an interested biochemical variable, it is usually necessary to perform regression or other statistical analysis.

Regression is a way to study the relationships among variables. Simple linear regression models with a log-transformed response variable have been traditionally used in vegetation index studies (Anderson et al., 1993; Chen and Cihlar, 1996). Some previous studies using stepwise linear regression already identified some optimum spectral bands and vegetation indices and established some sensitive predicators (Yoder and Pettigrew-Crosby, 1995; Lawrence and Ripple, 1998; Thenkabail et al., 2000).

Our study, however, adapted polynomial regression as the statistical analysis method.

A regression model contains a number of independent variables $X_1, X_2, \ldots, X_n$, which are used to explain or estimate some characteristics of the dependent variable. We can define the general linear regression model in terms of X variables as:

$$Yi = \boldsymbol{b}_0 + \boldsymbol{b}_1 X_{i1} + \boldsymbol{b}_2 X_{i2} + \ldots + \boldsymbol{b}_n\text{-}X_{in} + \boldsymbol{e}_i$$

Where:

$\boldsymbol{b}_0, \boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$ are parameters (coefficients) to be determined

ε_i are normal error terms

i = 1, 2, …, n

This general linear regression model with normal error terms encompasses a variety of situations. In general, the variables X_1, … X_n do not have to represent different independent variables, which is the case for polynomial regression models. They contain squared and higher-order terms of the independent variable, making the response function curvilinear. The following is a polynomial regression model with one independent variable:

$$Y_i = \boldsymbol{b}_0 + \boldsymbol{b}_1 X_i + \boldsymbol{b}_2 X_i^2 + \ldots + \boldsymbol{b}_{n-1} X_i^{n-1} + \boldsymbol{b}_n X_i^n + \boldsymbol{e}_i$$

The order of the independent variable is referred to be the degree of the polynomial regression model.

Polynomial regression models have two basic types of uses:

1. When the true response function is a polynomial function.
2. When the true response function is unknown (or complex) but a polynomial function is a good approximation to the true function.

The second type of use, where the polynomial function is employed as an approximation when the shape of the true response function is unknown, is very common.

To test the fitness of a polynomial regression model and compare the performance of different polynomial regression models, several statistical tests can be used, such as *coefficient of multiple determination, partial F-Test, prediction error sum of squares (PRESS)*. The following is the formula for the *coefficient of multiple determination*:

$R^2 = 1 – SSE/SST$

$$= 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

where:

$\hat{Y}$ is the estimated value of the response variable

$\overline{Y}$ is the mean (average) value of the response variable

The *adjusted coefficient of multiple determination*, denoted by $R_a{}^2$, adjusts R^2 by its degree of freedom. It can be used as one criteria for multiple regression model selection:

$$R_a{}^2 = 1 - \frac{n\text{-}1}{n\text{-}p} * \frac{SSE}{SST}$$

$$= 1 - \frac{n-1}{n-p} * \frac{\sum_{i=1}^{n}(Y_i - \hat{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

where *p* is the degree of freedom.

The use of polynomial models is not without drawbacks. Such models can be more expensive in degree of freedom than alternative nonlinear models or linear models with transformed variables (Neter et al., 1990). Fitting a polynomial regression with orders higher than three is rarely done as the interpretation of the coefficients becomes difficult and interpolation tends to become erratic.

Determining the relationship between vegetation indices and biochemical variables using polynomial regression has been reported by previous studies, and the results were generally better than simple linear regression models (Lawrence and Ripple, 1998).

2.3 Genetic Algorithm

Genetic algorithm (GA) is a powerful and widely applicable stochastic search and optimization technique. The GA search algorithm is inspired by genetic evolution and the process of natural selection.

For a particular problem, the genetic algorithm maintains a population of individuals for a generation. Each individual represents a potential solution to this problem at hand. Each individual is evaluated to give some measure of its “fitness”. Some individuals are transformed by means of genetic operations to form new individuals of the next generation. These new individuals are evaluated and transformed in a similar way. A new population is then formed by selecting the fittest individuals from the parent population and the offspring population. After several generations, the algorithm converges to the best individual, which generally represents an optimal solution to the problem. In summary, the population survive, breed, and change in a progression towards an optimal goal, similar to the natural selection and evolution process. Three typical operators in GA are called *selection, crossover and mutation*.

Selection refers to survival of the fittest. Individuals in the current population are selected for high fitness and put into a mating pool for further operations. Selection directs genetic algorithm toward promising regions in the search space.

Crossover is the recombination of two individuals (parents) to form new individuals (children). Children contain information from fit parents and are usually equally good or better than the parents.

Mutation randomly change the structure of an individual. It is a way to add new genetic material to a population. Its main purpose is to keep the population diverse during the search. Mutation is also used to help the GA avoid getting stuck at a local optimum.

A typical generational GA can be illustrated by the following figure:

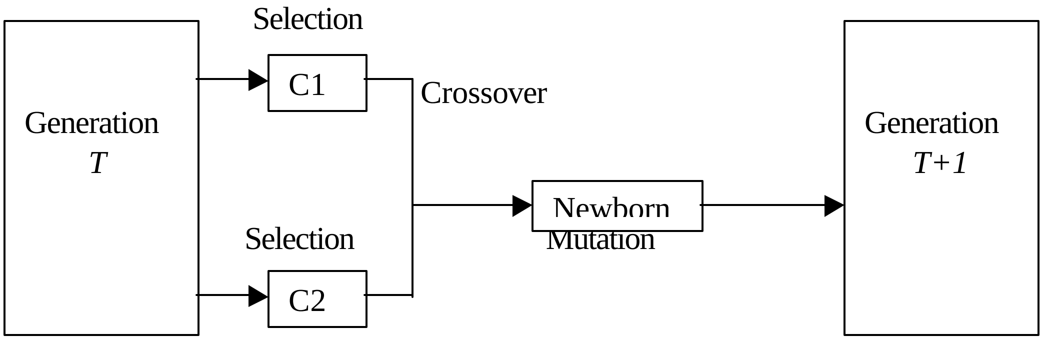


Figure 2-1 Generational GA working procedure.

The advantage of GA lies on its efficiency since it does not need to exhaust all possible search space to get the best solution. GA can handle discrete, continuous and mixed variable spaces and is relatively easy to implement. GA is also robust and less sensitive to noisy conditions.

On the other hand, GA is a heuristic search, that is, it is a simplification or educated guess that reduces or limits the search for solution. Heuristics do not guarantee the optimal solution, or even feasible solution and are often used with no theoretical guarantee. Therefore GA will not guarantee to find the true optimal solution. GA is also relatively slower due to genetic operations and searching from a population, and thereby not suitable for easy problems.

GA is especially suitable for problems with multiple search variables and large search space. Problems with multiple constraints that must be satisfied at the same time can usually be solved by GA. The time saving in GA compared with exhaustive search for large search space problems is remarkable. GA-based programs have been applied to problems in optimization, machine learning and evolving system modeling. Many successful applications of GA search have been developed in industrial engineering in the past decades. Examples include the travel salesman problem, which gives the shortest path among many cities in a given order, the bin pack problem, which consists of placing a number of objects into a number of bins such that the total weights of the object in each bin does not exceed its capacity and the number of bins used is minimized, the airline crew scheduling problem, which assigns a set of *m* flights that must be flown over a given time period to a set of *n* crews to minimize the operation cost. Genetic algorithm usually performs a lot better than conventional search strategies in these problems.

CHAPTER 3

STATISTICAL METHOD FOR DETERMINING OPTIMUM VEGETATION INDICES AND WAVELENGTHS

To detect plant stress caused by nitrogen deficient, an important issue is to choose a sensitive vegetation index and related optimal wavelengths. Under these two conditions a good correlation between nitrogen stress levels and vegetation index can be established with more accuracy. Therefore first we need to develop a correlation analysis between nitrogen stress levels and vegetation index.

In this chapter, after an introduction to the procedure of spectral data acquisition (image processing), we analyze these data using polynomial regression, develop statistical models, and select the optimum models by exhaustive search.

3.1 Materials and image processing

Bush bean plants (*Phaseolus vulgaris* L. var. Newport) were grown inside a greenhouse and in Perlite containers during the summer of 1997. There were 6 replicates of 4 treatments for a total of 24 plants. The plants were fed with a complete hydroponics nutrient solution except that nitrate and ammonium salt quantities were adjusted to provide 4 levels of nitrogen treatments: 30, 60, 90, 120 ppm (see Theisen et al., 1998 for cultural details). On day 47 after seeding, the plants were scanned by a multi-spectral imaging system from 695 nm to 795 nm in steps of 5 nm (from 760nm to 770 nm steps of 2 nm were used), under natural sunlight and inside the greenhouse (see Thai et al., 1998).

Our spectral imaging system interfaced a Liquid Crystal Tunable Filter into a monochrome video imaging system. This system implemented a scheme for leveling the system response across wavelengths in the face of varying illumination, filter transmittance, camera lens aperture setting, and gain response (see Evans et al., 1998). Spectralon standard targets from Labsphere were used in our study. The nominal reflectance factors of the steps in the multi-step Spectralon targets were 2% reflectance for Black target and 99% reflectance for White target. These reflectance factors were used later in computing the reflectance factor of each plant canopy. The 99% target was used for the "control white" region of interest (ROI) when we scanned from 695 nm to 795 nm. Our image scanning scheme was designed to maintain a given gray value for the "control white" ROI at all wavelengths by controlling the camera gain settings and the LCTF attenuation factor.

For each of the 24 plants, NDVI images were computed from the scanned image at 695 and 760 nm with the following equation (see Carter and Miller, 1994; Wiegand et al., 1991):

$$NDVI\,\mathrm{Im}\,age = \frac{(\mathrm{Im}\,age@760nm - \mathrm{Im}\,age@695nm)}{(\mathrm{Im}\,age@760nm + \mathrm{Im}\,age@695nm)} \qquad (1)$$

Figures 3-1 to 3-2 show the four NDVI images of bush bean plants under four different nitrogen treatments. One can see that the gray values of the plant canopy in these NDVI images became brighter with the increase of nitrogen concentration.

Next, we isolated the pixels of the plant canopy using a simple gray value thresholding scheme to form a mask that will be used to extract the plant canopy pixels at all other wavelengths scanned images. During this image processing step, the mean gray value for each plant canopy GVPlant$_i$ at each wavelength λ_i was collected. Mean gray values of ROIs positioned over the “control white” and “control black” standard targets were also collected and saved respectively as GVWhite$_i$ and GVBlack$_i$ for each wavelength λ_i. We defined RWhite$_i$, RBlack$_{i,}$ RPlant$_i$ as the corresponding reflectance values for the “white” target, “black” target and the plant at each wavelength $\lambda_{i.}$ Using calibration data from Labsphere for their standard targets, we could compute the values of RWhite$_i$ and RBlack$_i$ at any wavelength λ_i. Since the camera response was found to be linear (Evans et al., 1998), a correspondence could be made between the image gray values and the reflectance values of the objects involved. We could write the following equation for each wavelength λ_i:

$$\frac{GVPlant_i - GVBlack_i}{GVWhite_i - GVBlack_i} = \frac{RPlant_i - RBlack_i}{RWhite_i - RBlack_i} \tag{2}$$

Then compute the reflectance value RPlant$_i$ as following:

$$RPlant_i = \frac{(GVPlant_i - GVBlack_i)}{(GVWhite_i - GVBlack_i)} * (RWhite_i - RBlack_i) + RBlack_i \tag{3}$$

Using equation (3) we computed the percentage reflectance for 24 wavebands under 4 different nitrogen levels. Thus the following four vegetation index values can be calculated:

$$NDVI(nomalized\_difference) = \frac{R_1 - R_2}{R_1 + R_2} \tag{4}$$

$$RVI\ (ratio) = R_1 / R_2 \tag{5}$$

$$DVI\ (difference) = R_1 - R_2 \tag{6}$$

$$R\ (reflectance) = R_i \tag{7}$$

Here R$_1$ and R$_2$ corresponded respectively to the reflectance values under two different wavelengths ë$_1$ and ë$_2$.

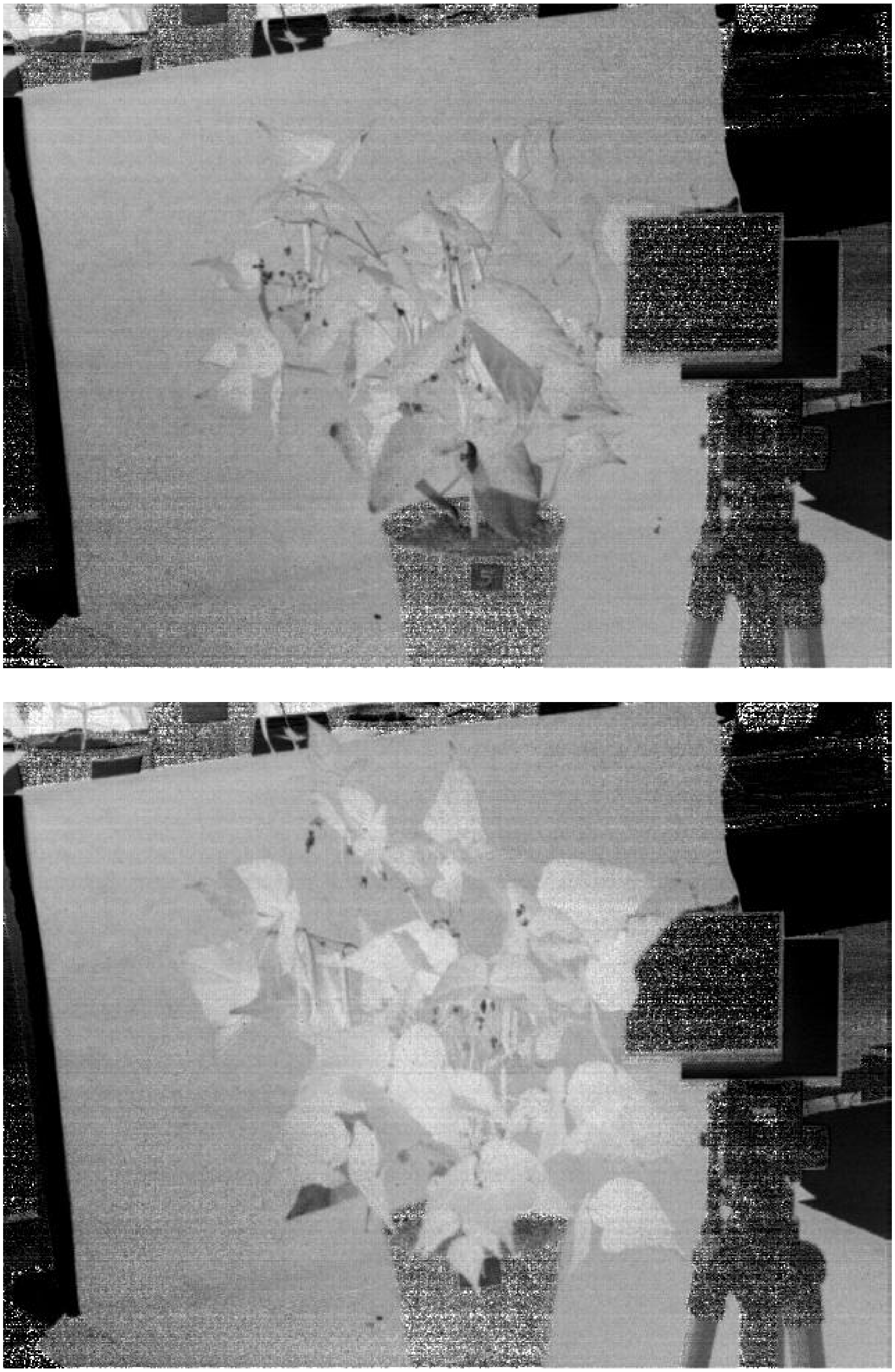


Figure 3-1 NDVI image of Bush Bean plants @ 695 nm, 760nm, with nitrogen treatments 30 ppm (top) and 60 ppm (bottom)

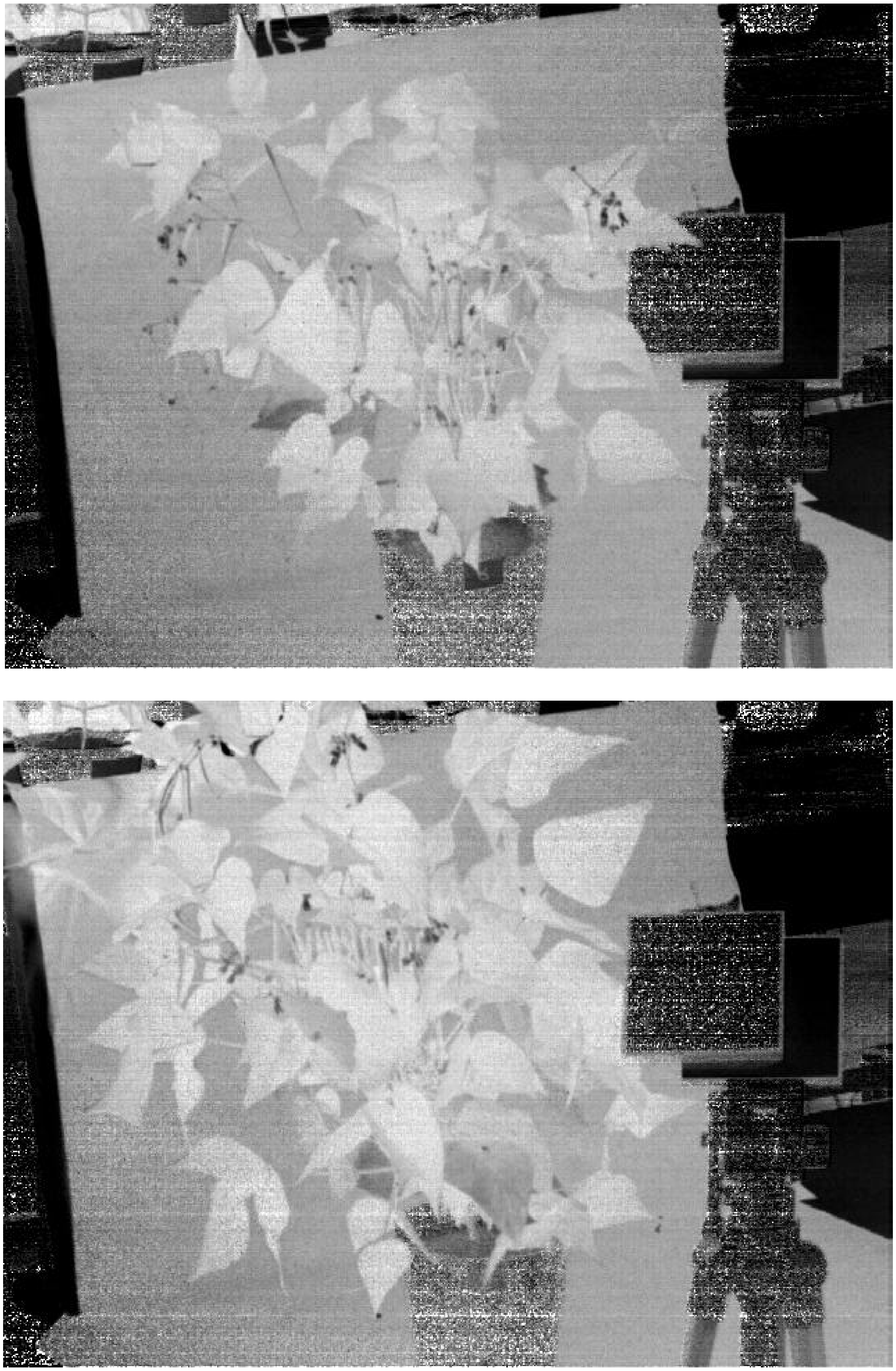


Figure 3-2 NDVI image of Bush Bean plants @ 695 nm, 760nm, with nitrogen treatments 90 ppm (top) and 120 ppm (bottom)

3.2 Statistical analysis and exhaustive search

3.2.1 Spectral reflectance characteristics

Before locating individual optimum wavelength, we plot wavelengths versus reflectance values under different nitrogen treatments as a first step (see Figure 3-3). The wavelength portion of 695-750 nm had higher change in reflectance per wavelength unit difference than the other portions. The plot showed the typical ramp of the plant "red edge", which is a region of high chlorophyll absorption in green vegetation. This is consistent with previous researches (Boochs et al., 1990; Fillela & Penuelas, 1994), and gives us a visual feeling of the location of the optimum wavelength.

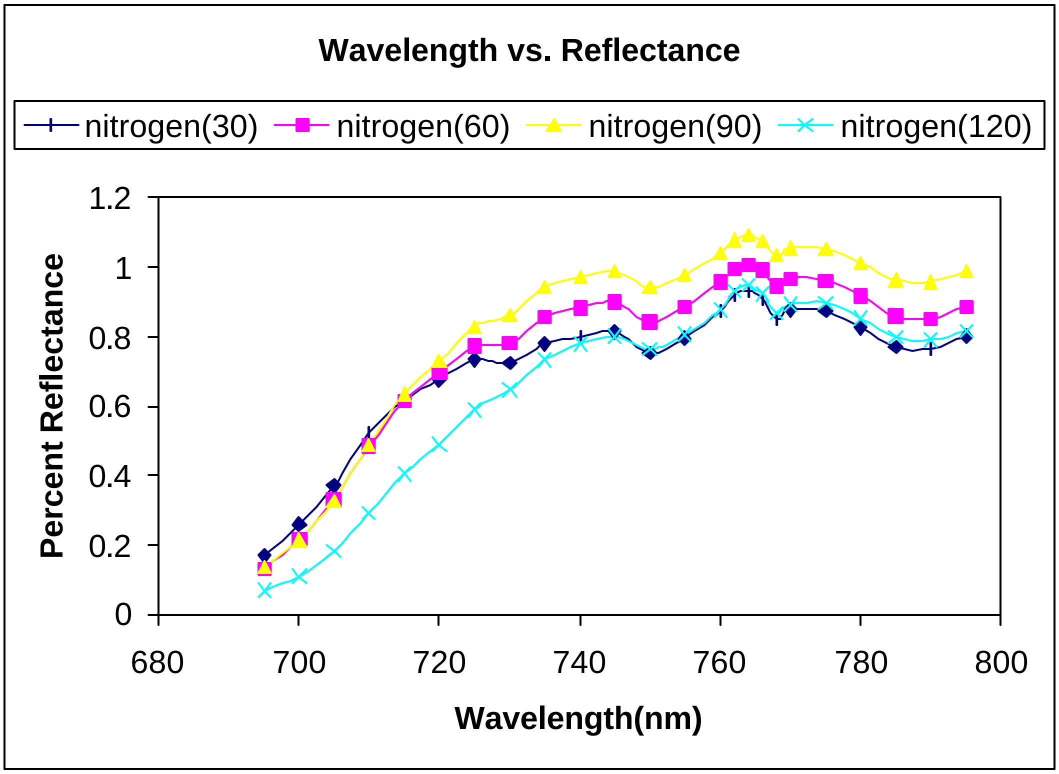


Figure 3-3 Plot of percent reflectance vs. wavelength for plant1 under four nitrogen treatments

3.2.2 Computation of vegetation indices

Statistical analysis need to be performed on individual band and index to identify the optimum wavelengths and vegetation indices. All possible two-bands wavelength combinations will be involved for each vegetation index. These combinations need to be computed before further correlation analysis.

Three distinct types of vegetation indices, NDVI, RVI and DVI, were computed from equations (4), (5), (6) using 24 spectral bands of percentage reflectance data. For each index, there are 276 (24 * 23 / 2) possible combinations of two different wavelengths, which were computed with a VBA (Visual Basic for Applications) macro program inside Microsoft Excel 2000.

For each vegetation index, we grouped combinations of all the four treatment levels and sorted them by wavelength. Thus we got 276 * 4 = 1104 sets of data, of which each set had 24 values (6 plants * 4 treatments) and represented the corresponding vegetable index values under four different nitrogen treatment levels. Polynomial regression analysis was then performed on each dataset. Table 3-1 shows the NDVI values of four nitrogen treatments for wavelength combination 695 nm and 760nm, and for each of the 6 plants tested.

Table 3-1 NDVI (695 nm, 760nm) values for different nitrogen treatments

wavelength	plant1	plant2	plant3	plant4	plant5	plant6	treatment
695, 760	0.676157	0.685291	0.673049	0.711778	0.69331	0.695768	30
695, 760	0.755439	0.787415	0.765097	0.766605	0.777407	0.770381	60
695, 760	0.765168	0.802903	0.793288	0.79669	0.840219	0.816067	90
695, 760	0.851377	0.814104	0.804599	0.809205	0.818001	0.848814	120

3.2.3 Polynomial regression and exhaustive search

Polynomial regression is a commonly used approximation method when the relationships between the independent variables and the dependent variables are uncertain. That is, the mechanism of the true response function is unknown. Visual examination of the plots of nitrogen treatments vs. different vegetation indices under wavelengths 695 nm and 750 nm from our dataset indicated that there was a non-linear relationship between nitrogen treatments and vegetation indices (see Figures 3-4 to Figure 3-7). We therefore fit the data with polynomial regression models. This was performed by using the software package Statistical Analysis System (SAS) version 8.0.

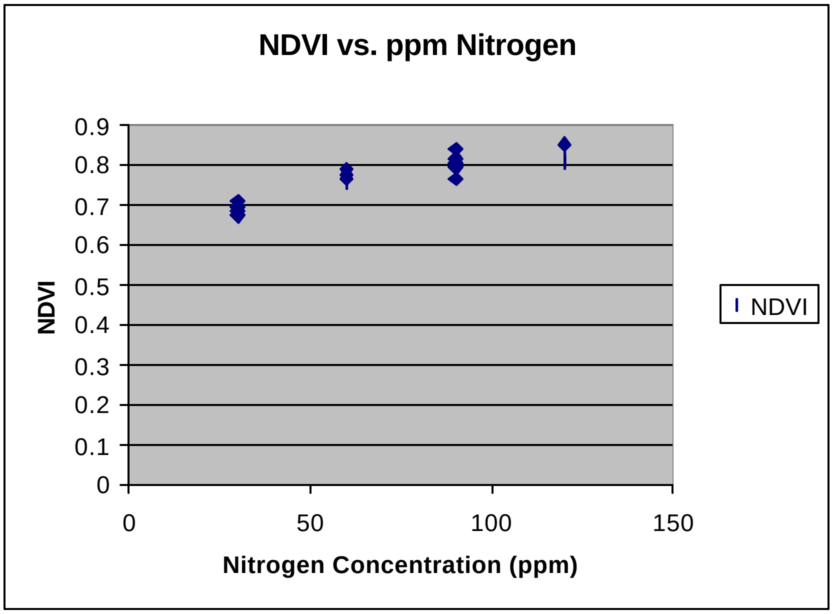


Figure 3-4 NDVI vs. ppm Nitrogen under wavelengths 695 nm, 760 nm

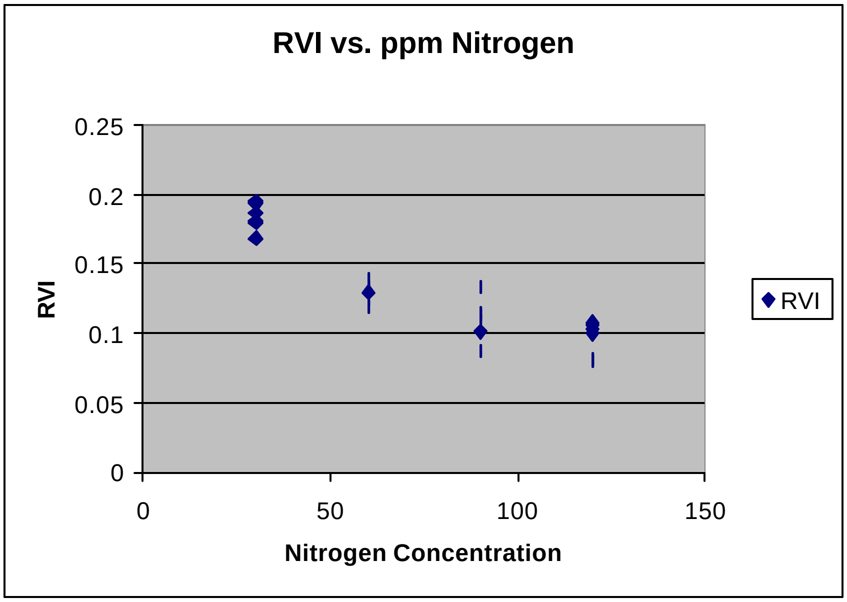


Figure 3-5 RVI vs. ppm Nitrogen under wavelengths 695 nm, 760 nm

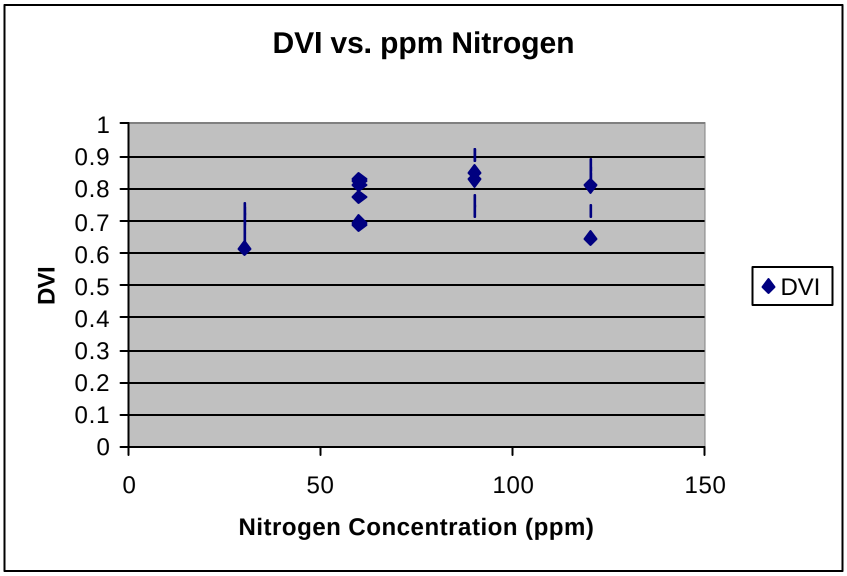


Figure 3-6 DVI vs. ppm Nitrogen under wavelengths 695 nm, 760 nm

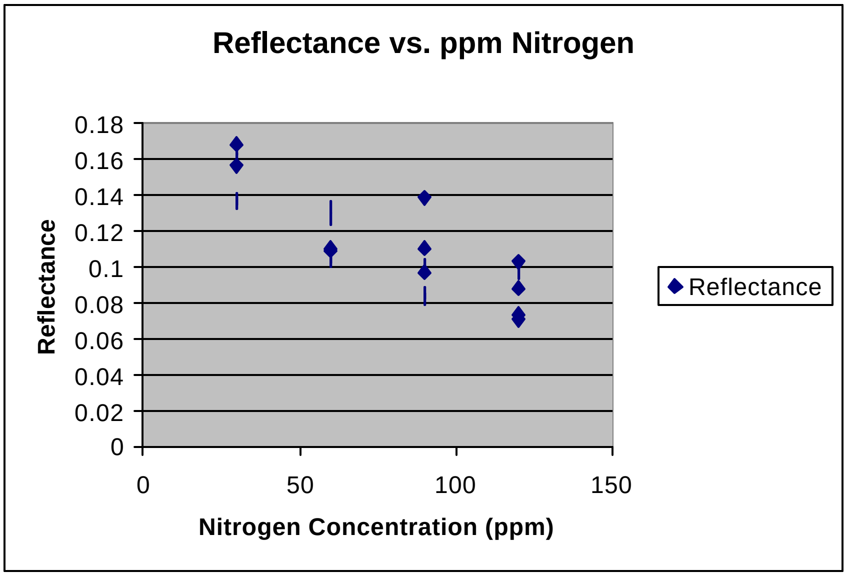


Figure 3-7 Reflectance vs. ppm Nitrogen under wavelength 695 nm

Polynomial regression models are often fitted with the hierarchical approach in which higher powers are introduced one at a time and tested for significance, and if a term of a higher order is included (say, x^3) then all terms of lower order (x and x^2) are also included. We started with first degree polynomial and increased the polynomial powers until the fourth degree.

The *Adjusted coefficient of multiple determination (adjusted R^2)* for all possible two-band vegetation indices were determined and sorted. The model selection criteria were set to both maximize adjusted R^2 and keep the P value significant (P value < 0.05). Residual and data plots were also used to guide regression analysis and model selection. Optimum bands and indices were obtained by comparison among the adjusted R^2 values of all significant models. This was essentially an exhaustive search procedure because it compared adjusted R^2 for all possible wavelength combination of four different vegetation indices. The best statistical model for each vegetation index with the corresponding adjusted R^2, wavelengths were listed in Table 3-2.

Table 3-2 Optimum spectral bands and regression models for nitrogen treatments against individual vegetation indices

Index	best bands (nm)	adjusted R^2	regression model
NDVI	710, 720	0.9096	$NDVI=0.087+0.02ppmN-0.67E\text{-}5ppmN^2$
RVI	700, 750	**0.9144**	$RVI = 0.29 - 0.003ppmN + 0.00001ppmN^2$
DVI	715, 750	0.8715	$DVI = 0.11 + 0.002\ ppmN$
Reflectance	700	0.7276	$Reflectance = 0.17 - 0.0007\ ppmN$

From Table 3-2, DVI and Reflectance were simple linear regressions, while the final models for NDVI and RVI included second-degree polynomials. These four final models were plotted in Figures 3-8 to 3-11.

All vegetation indices under the corresponding best wavelengths correlated well to nitrogen treatments. The adjusted R^2 values ranged from 0.7276 to 0.9144. NDVI and RVI performed better than the other two indices. This was consistent with Lawrence and Ripple's study (1998). The best wavelengths identified were: 700 nm, 710 nm, 715 nm, 720 nm, 750 nm, which were also close to those found by some other researchers (Tucker, 1979; Yoder and Pettigrew-Crosby, 1995; Thenkabail et al., 2000).

The overall results of this comprehensive analysis were illustrated in contour plots of the R^2 values for each wavelength pair (see Figures 3-12 to 3-14). An examination of these results for different vegetation indices showed a remarkable strong relationships region center at the red-NIR 695 nm to 750 nm.

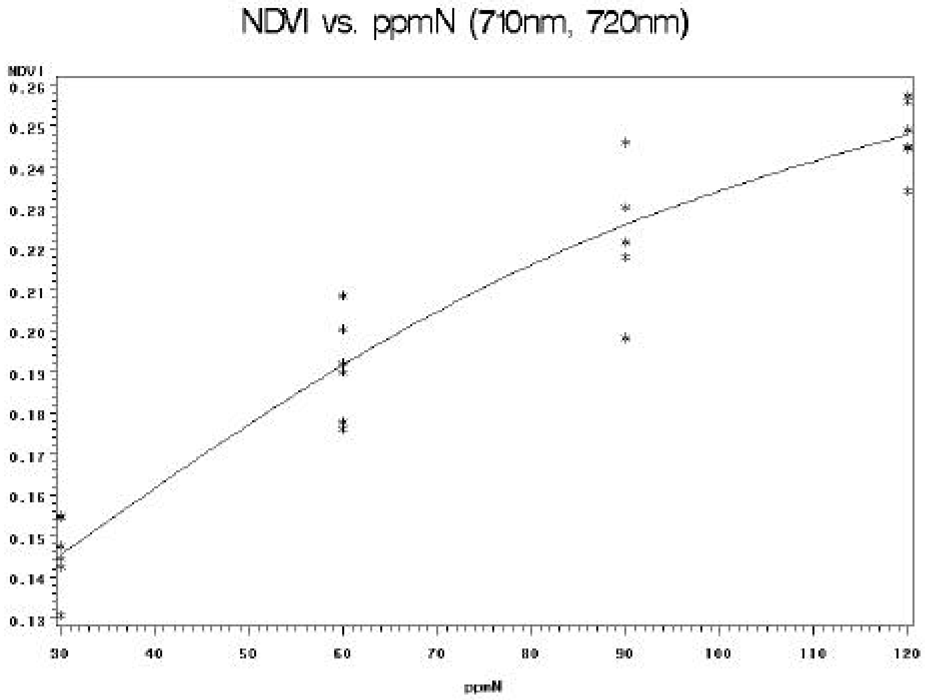


Figure 3-8 Relationship between NDVI and nitrogen concentration under wavelengths 710 nm and 720nm

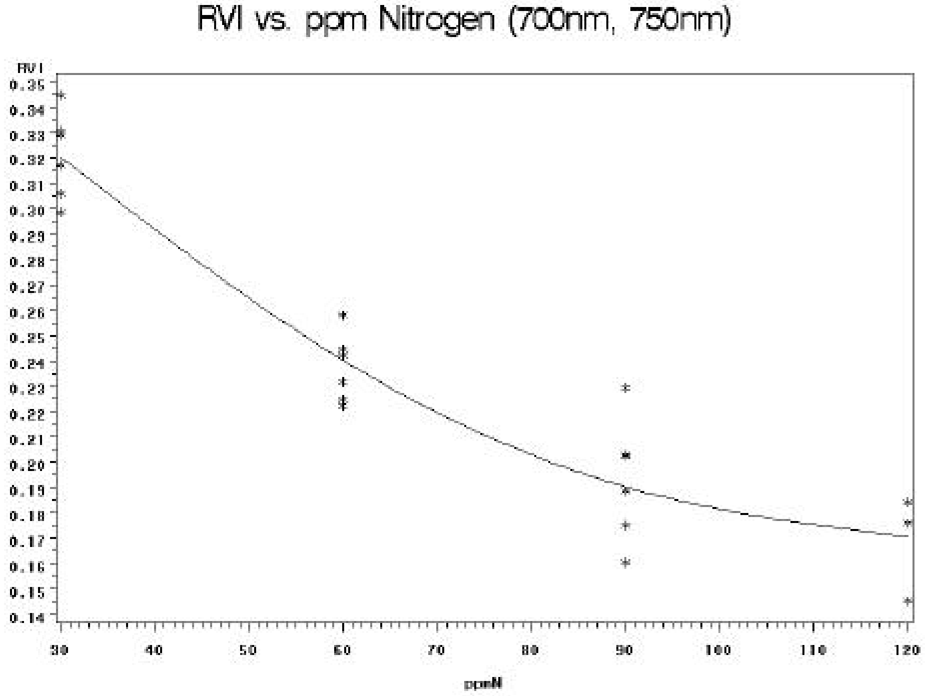


Figure 3-9 Relationship between RVI and nitrogen concentration under wavelengths 700 nm and 750nm

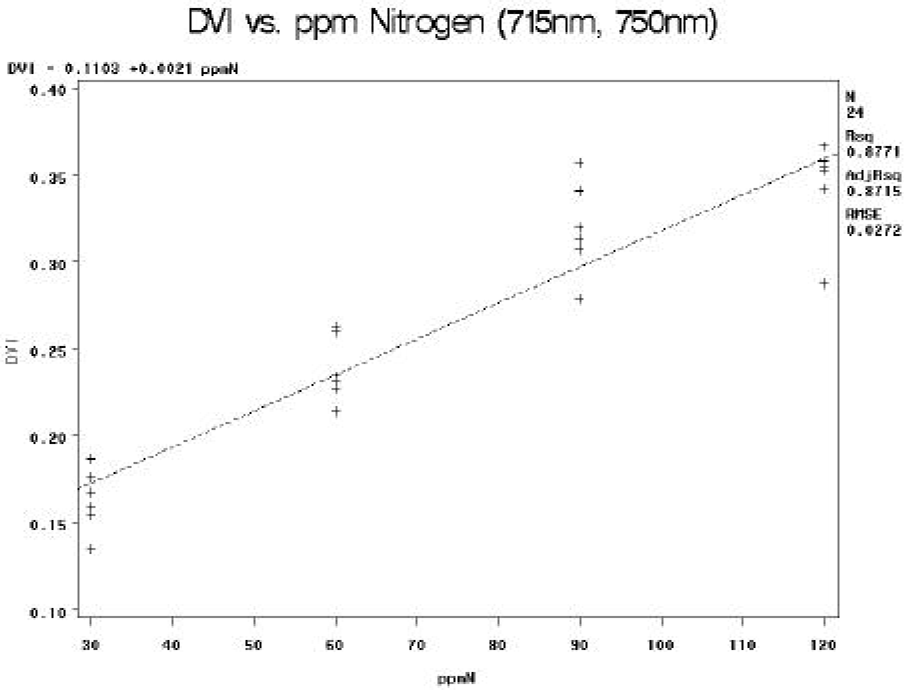


Figure 3-10 Relationship between DVI and nitrogen concentration under wavelengths 715 nm and 750nm

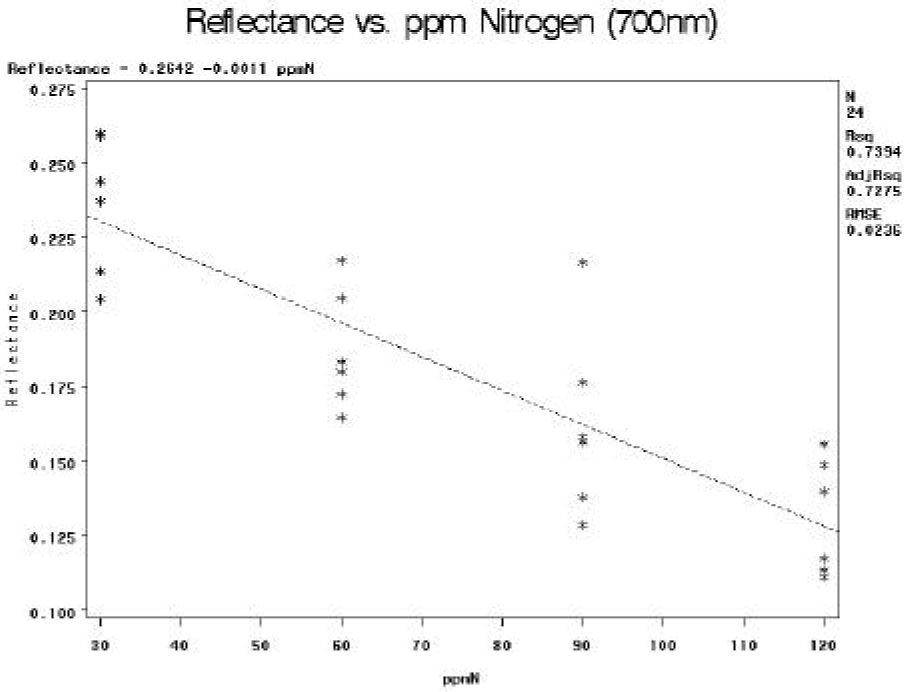


Figure 3-11 Relationship between reflectance and nitrogen concentration under wavelengths 700nm

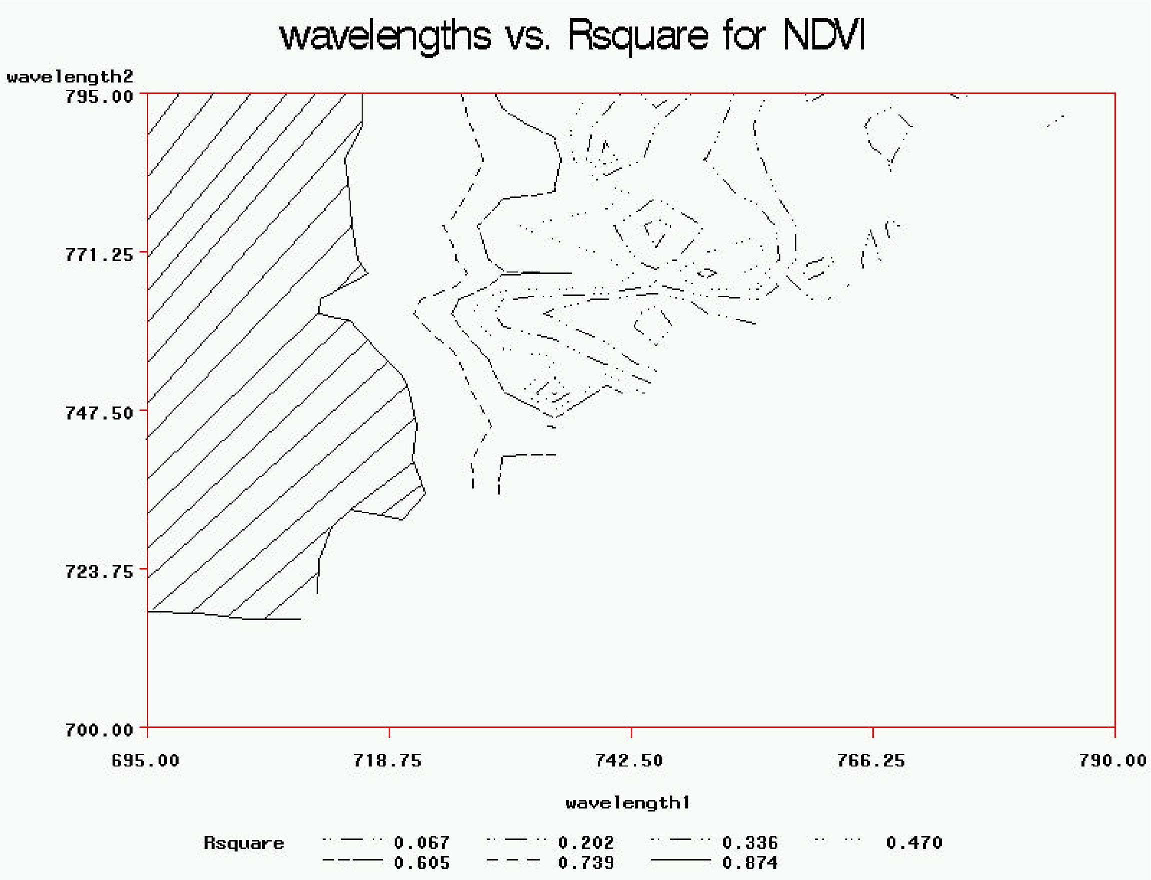


Figure 3-12 Contour plot showing the correlation (R^2) between nitrogen concentration and NDVI values calculated for 276 wavelength combinations. The area filled with solid line indicates the region with high R^2 values ($R^2 > 0.874$)

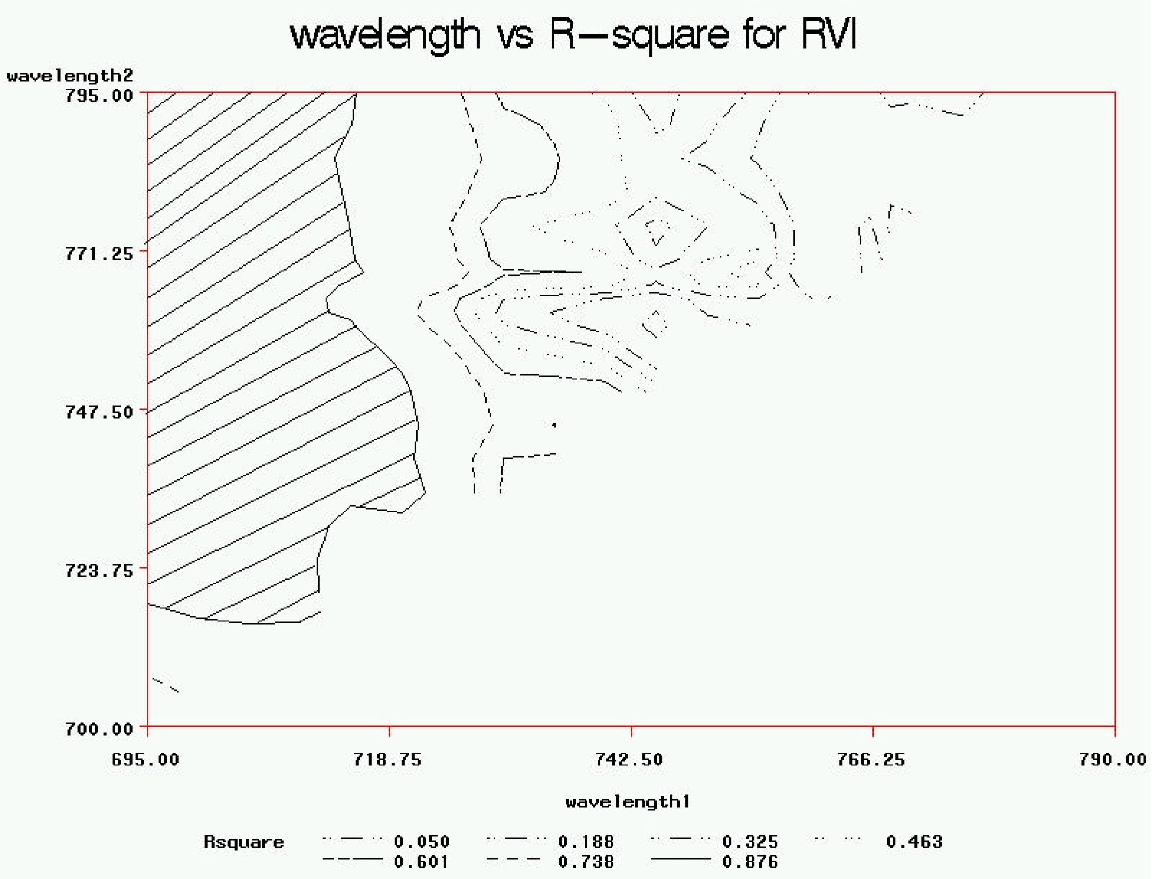


Figure 3-13 Contour plot showing the correlation (R^2) between nitrogen concentration and RVI values calculated for 276 wavelength combinations. The area filled with solid line indicates the region with high R^2 values ($R^2 > 0.876$)

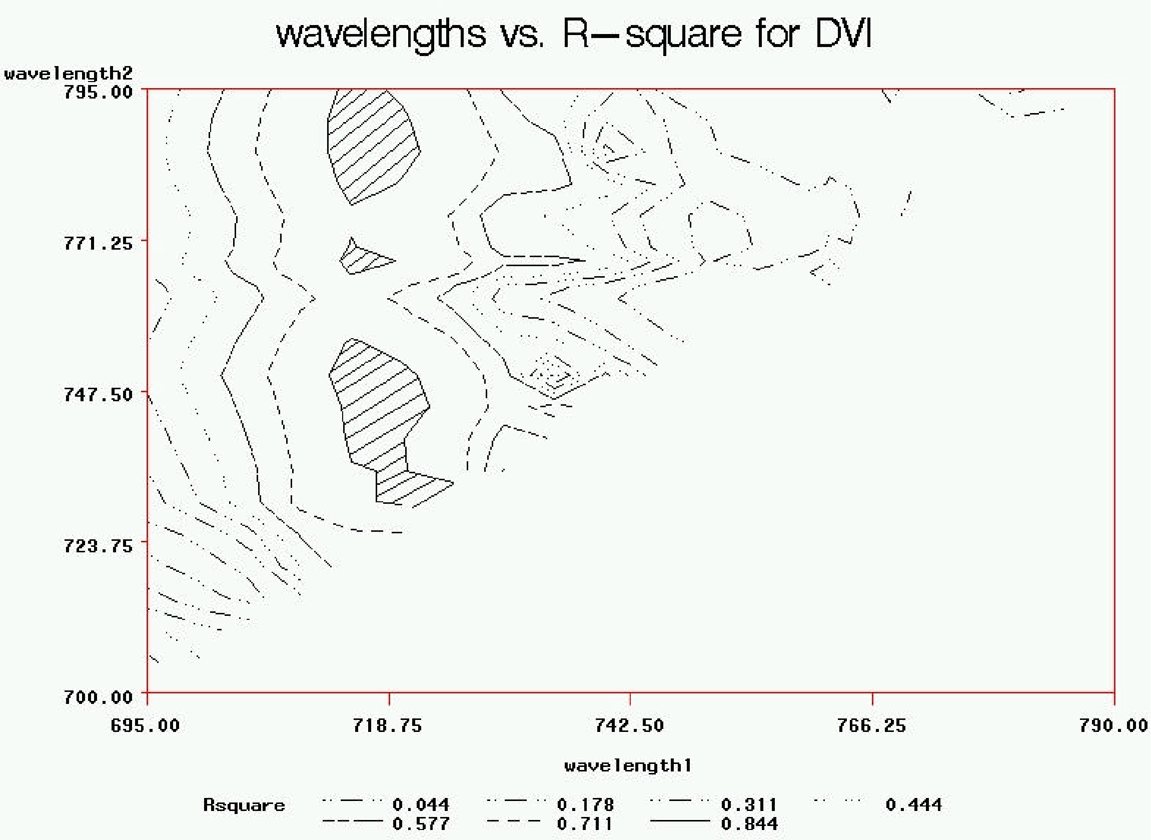


Figure 3-14 Contour plot showing the correlation (R^2) between nitrogen concentration and DVI values calculated for 276 wavelength combinations. The areas filled with solid line indicate the region with high R^2 values ($R^2 > 0.844$)

3.2.4 *log*-transformed polynomial regression

Regression models with a *log*-transformed response variable (*log*-transformed models) were used in many previous studies and were reported to perform better than non-transformed model (Yoder and Waring, 1994; Anderson et al., 1993). In our study, we also developed *log*-transformed polynomial regression models and selected the best model by exhaustive search to see if we could improve regression results.

Using the same spectral reflectance data, we first calculated the inverse-*log* of vegetation indices (*log 1/VI*), then each data set, which represents the *log*-transformed VI under four nitrogen treatments for a specific wavelength pair, was fitted with first, second, third and fourth degree polynomial models. The general regression formula was:

$$log\ (1/Y) = \boldsymbol{b}_0 + \boldsymbol{b}_1 X + \boldsymbol{b}_2 X^2 + \ldots + \boldsymbol{b}_{p-1} X^{p-1} + \boldsymbol{b}_p X^p + \boldsymbol{e}_i$$

Where p indicated the degree of polynomial. The model selection criteria and the exhaustive search procedure were the same as those described in section 3.2.3.

Table 3-3 shows the result for the *log*-transformed polynomial regression. The final regression model for NDVI under wavelengths 710 nm and 720 nm explained substantially more variation than the other three vegetation indices, with an adjusted R^2 of 0.9167. NDVI, RVI and DVI were second-degree polynomial regression. Reflectance was a simple linear regression. All models were statistically significant, and the adjusted *R2* values ranged from 0.7294 to 0.9167.

Comparing results of Table 3-2 and Table 3-3, we found that *log*-transformation could improve regression results for NDVI and DVI, but did not improve the performance of RVI. The optimum wavelengths and polynomial models selected by both approaches were similar except for RVI.

Table 3-3 Optimum spectral bands and regression models for nitrogen treatments against individual *log*-transformed vegetation indices

Index	best bands (nm)	adjusted R^2	regression model
NDVI	710, 720	**0.9167**	$log1/\mathrm{NDVI}=2.28-0.013\mathrm{ppmN}+0.00005\mathrm{ppmN}^2$
RVI	710, 720	0.9087	$log1/\mathrm{RVI}=0.17+0.04\mathrm{ppmN}-0.00001\mathrm{ppmN}^2$
DVI	715, 750	0.9138	$log1/\mathrm{DVI}=2.39-0.02\mathrm{ppmN}+0.00009\mathrm{ppmN}^2$
Reflectance	700	0.7294	$log1/\mathrm{Reflectance} = 1.27 + 0.006\mathrm{ppmN}$

CHAPTER 4

GENETIC ALGORITHM FOR DETERMINING OPTIMAL WAVELENGTHS AND VEGETATION INDICES

In this chapter, we applied genetic algorithm to perform statistical analysis and model selection. The problem formulation, GA working procedure, and experimental trials will be described in different subsections. We also discussed the experimental results and compared GA's performance with the SAS procedure obtained in the previous chapter based on accuracy and runtime.

4.1 GA search system architecture

To identify the best wavelengths, vegatation index and regression model, the complete spectral reflectance data were fed as input to the GA, from which some datasets were randomly chosen. Each dataset represented vegetation index values of different nitrogen treatments under a particular wavelength pair. Each dataset was then evaluated by the fitness function. The fitness function performed statistical analysis and returned a value as a measure of merit of the current regression model. The GA search was directed by the fitness selection and went toward the more optimal (fit) direction until it found the optimal solution. The overall GA search system architecture is illustrated in Figure 4-1.

In general, a genetic algorithm has the following components (Michalewicz, 1996): a genetic representation of solution to the problem, a way to create an initial population of solutions, an evaluation function rating solutions in terms of their fitness, genetic operators that alter the genetic composition of offsprings during reproduction and values for the parameters of genetic algorithm. We formatted our problem based on these components.

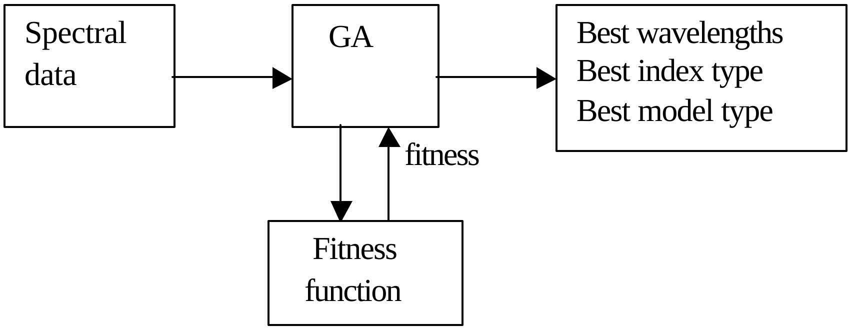


Figure 4-1 Outline of the Genetic Algorithm search process

4.2 Problem formulation

1. Representation

A vector of integers was used to represent a GA individual. Individuals were encoded by integer permutation. The goal of the GA was essentially to search for a permuted individual that best satisfied all the problem constraints. Each individual consisted of four variables. The first and the second represented the two wavelengths, the third represented vegetation index type, the fourth represented the regression model (i.e. the degree of the polynomial), as shown in Figure 4-2.

Gene	1	2	3	4
String	wavelength1	wavelength2	vegetation index	regression model

Figure 4-2 A GA representation for an individual.

The values of wavelength1 and wavelength 2 range from 1 to 24, which represent all 24 spectral bands used in our study. The values of vegetation indices range from 1 to 4, with each representing one type of vegetation index. The values of regression model range from 1 to 4, which represented first, second, third and fourth degree polynomial regression models.

An example of an individual could be {1, 12, 2, 3} where (1, 12) meant the first wavelength (695 nm) and the twelfth wavelength (750 nm), 2 meant RVI, and 3 meant third degree polynomial regression.

2. Fitness function

The fitness function took two wavelengths, one vegetation index type and one regression model type as input. It then selected all reflectance values (i.e. all 6 tested plants) under four nitrogen treatments by the two wavelengths. Depending on the type of the vegetation index, the vegetation index value could be computed by equations 4, 5 or 6. This gave a set of data which could be used for regression analysis. After computation similar to the polynomial regression procedure with the specified degree (Younger, 1979), the fitness function returned a value that represented the correlation between the nitrogen treatments and the vegetation index chosen. This value is based on the *adjusted coefficient of multiple determination* (adjusted R^2). Since the adjusted R^2 value varied from a very narrow range of 0 to 1, the selection pressure towards the optimum was small. This problem was called “poor scaling” (Goldberg, 1989). We multiplied the value of adjusted R^2 by 1000 as the measurement of fitness before the individual input was evaluated. The objective of GA was to maximize the thus modified adjusted R^2.

Note that another statistical test, the P value, which indicates whether a model is significant or not, should also be considered as a model selection criterion since a non-significant statistical model (P value > 0.05) was meaningless. But the computation of the P value was too complex to be implemented in our current system. Therefore we use only adjusted R^2 as the regression model selection criterion, i.e, the measure of merit of the fitness function.

3. Selection

In GA, selection means survival of the fittest. The effect of selection is to gradually bias the sampling procedure toward individuals whose fitness is estimated to be above average. Over time, the fitness of the population will become more and more optimal. Some commonly used selection types are Roulette wheel selection, Tournament selection, Rank and scaling. The Roulette wheel selection (Holland et al., 1986) was used in our study since it was relatively easy to implement. This method reproduced a new generation proportional to the fitness of each individual. A model roulette wheel was made to display fitness probabilities of individuals. The selection process was based on spinning the wheel the number of times equal to population size, each time selecting an individual for the new population. The drawback of this selection method was that early on there was a tendency for a few super individuals to dominate the selection process. Later on, when the population was largely converged, competition among individuals was less strong. Strategies such as fitness sharing, fitness scaling to be described later can be used to overcome this drawback.

4. Operators

   a. Crossover operator: Crossover operator is the major operator of GA (Goldberg et al., 1993). It changes the composition of offspring by exchanging and recombining genes of parents. There are several crossover operators such as point crossover, random crossover and uniform crossover. Figure 4-3 illustrates single point crossover used in our study, which involves cutting the chromosomes of the parents at a random point and exchanging the sub-chromosomes of parents.

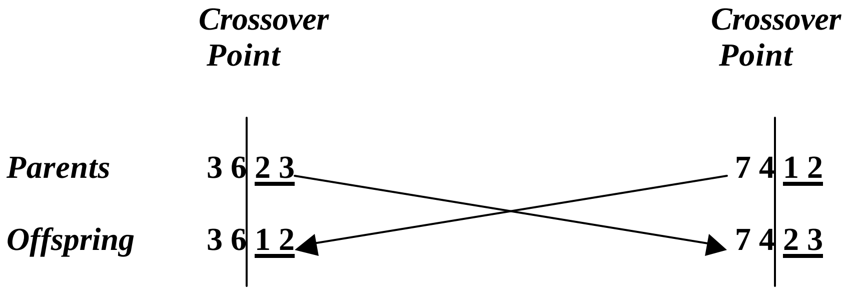


Figure 4-3 Single point crossover

   b. Mutation operator: the mutation operator mutates one or more genes in an individual and randomly changes the individual. It is used together with crossover to explore the entire solution space and to avoid local optimum. Mutation also prevents the loss of diversity in the population. It is usually used as a background operator to overcome some drawback of crossover. Some commonly used mutations are random mutation, uniform mutation, and boundary mutation. Figure 4-4 shows the random mutation operator used in our study.

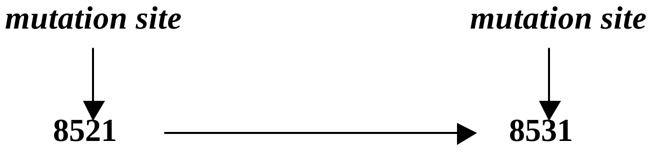


Figure 4-4 Random mutation operator

4. Repair

Using integer permutation encoding for individual representation, infeasible or illegal individuals can be generated during population initialization, crossover or mutation. Infeasible individual refers to the individual lying outside the feasible solution region of a given problem. Illegal individual refers to individual that does not represent a solution to a given problem (Gen and Cheng, 2000). For example, an individual {3, 3, 2, 1} was generated in our problem by integer permutation. This was an illegal individual and did not represent a solution to our problem because the first two numbers should represent two different wavelengths and should not be identical when vegetation index 2 (RVI) was used.

Integer permutation encoding also leads to redundant individuals. For example, individual {1, 2, 3, 2} and individual {2, 1, 3, 2} are essentially the same individual since wavelength pair 1-2 and 2-1 are identical for vegetation index value computation. They map into the same solution for our problem.

Repair techniques are usually adopted to solve the above problems. GA can be improved by adding a repair operator that applies certain constraints to individuals. These constraints convert an illegal individual to a legal individual and ensure that no redundant individual is generated. Some examples of repair operation are:

{3, 3, **2**, 1} → {3, 3, **4**, 1}

{1, 5, **4**, 2} → {1, 5, **2**, 2}

{**5, 1**, 2, 3} → {**1, 5**, 2, 3}

In the first two examples illegal individuals are converted to legal ones. In the third example a redundant individual is repaired to avoid duplication. After repair operation, only legal and unique individuals are kept, resulting in a reduced search space and better performance.

5. Special GA operations

To make the GA more robust and efficient, in addition to the conventional operations such as selection, crossover, mutation, three "safeguard" operations were used in our program to avoid local optimum and maintain population diversity.

a. Fitness scaling: this was done to keep appropriate levels of competition throughout a simulation. This operation applied linear scaling on the raw fitness value. We simply calculated the scaled fitness f' from the raw fitness f using a linear equation:

$$f' = af + b$$

The coefficients a and b could be calculated based on the number of expected copies desired for the best population member, which was denoted by c. They were chosen to enforce equality of the raw and the scaled average fitness and cause the maximum scaled fitness to be a specified multiple of the average fitness. In our study, *c* was set to1.8, *a* and *b* were calculated according to different mathematic formulae (see Goldberg, 1989 for calculation details).

On the early stage, fitness scaling prevented the early domination of extraordinary individuals. On the later stage, it encouraged a healthy competition among near equal individuals.

b. Fitness sharing: if a population contained identical individuals, only one of them received the fitness value calculated in the normal way, the others were assigned degraded fitness to reduce their reproductive abilities. In our program, we defined the fitness decreasing factor as the percentage of decreased fitness to be applied, which was provided by the user. In our study, the fitness decreasing factor was set to 10%. Identical individuals received degraded fitness base on the fitness decreasing factor.

   Fitness sharing helped to maintain the population diversity, and reduce the chance of a whole population being dominated by a single, relative superior individual.

c. Diversity restoration: this operation monitored the evolution process of GA. When it found that there was not any progress in the recent *n* generation (*n* was provided by the user), it automatically applied mutation on the population. This mechanism reduced the tendency for GA to get stuck at a local optimum.

6. Termination criterion

   GA converges when a target value was reached or certain convergence criterion was met. GA could also be stopped when the maximum number of evaluations has been exhausted. Two termination criteria were used to in our problem:

   a. The GA stopped when this condition was met: *average fitness / maximum fitness* > *t* in the population, where 0.95 < *t* < 1. In our experiment, *t* was set to 0.98. At this point the population loses diversity and practically converges to a single point in the search space.
   b. A target value (which is the best adjusted R^2 value found by the SAS procedure) is given to GA. The GA will not stop until it finds this target value.

7. Evaluation

   The evaluation of GA's performance is based on how long it takes to find the optimal solution and how good is the solution. Since CPU time depends on each individual computer's hardware and operating system, we use iteration number to measure the speed of convergence. One iteration means one call to the fitness function from a unique individual, which corresponds to one execution of a specific degree of polynomial regression in the SAS procedure. Using termination criterion *a* described above, the difference between the adjusted R^2 found by GA and the one provided by SAS is a good indicator for GA's performance. Using termination criterion *b,* iteration number for convergence can be used to evaluate the performance of GA.

4.3 Genetic algorithm working procedure

The genetic algorithm approach to determine optimum wavelengths and vegetation indices works as follows:

1. Generate an initial population of *N* random solutions. Set the generation number to *T*.
2. Select two solutions, P_1 and P_2, from the population using Roulette Wheel selection.
3. Combine P_1 and P_2 to form a new solution, *C*, using the single point crossover operator.
4. Mutate *C* randomly with the random mutation operator.
5. Make *C* legal and non-redundant by applying the repair approach.
6. Repeat steps 2 to 5 to generate a new offspring population of the same size *N*.
7. Sort generation *T* and its offspring by the fitness values.
8. Construct generation *T+1* by keeping *g%* solutions on the top of generation *T* and replace the remaining individuals with *(1-g)%* solutions of the top of the offspring population, where *g* is generation gap.
9. Repeat above steps until the termination criteria is satisfied.

4.4 Experimental procedure

1. Source code modification.

   The software package “GA-playground” written by Ariel Dolan (http://www.arieldolan.com/ofiles/gaa.html) was used to perform GA search. This package was implemented in JAVA. In addition to providing the fitness function, some modifications need to be made for our problem requirement. In the original GA-playground, the stopping criterion of GA was set to an exit value which must be provided by the user. We modified the source code and add another option so that the GA could converge when the termination criterion *a* described above was met, which did not require a user provided exit value. The code which calculated iteration number was rewritten so that only fitness function calls from unique individuals were counted. A repair operator was added to GA to avoid illegal and redundant individuals.

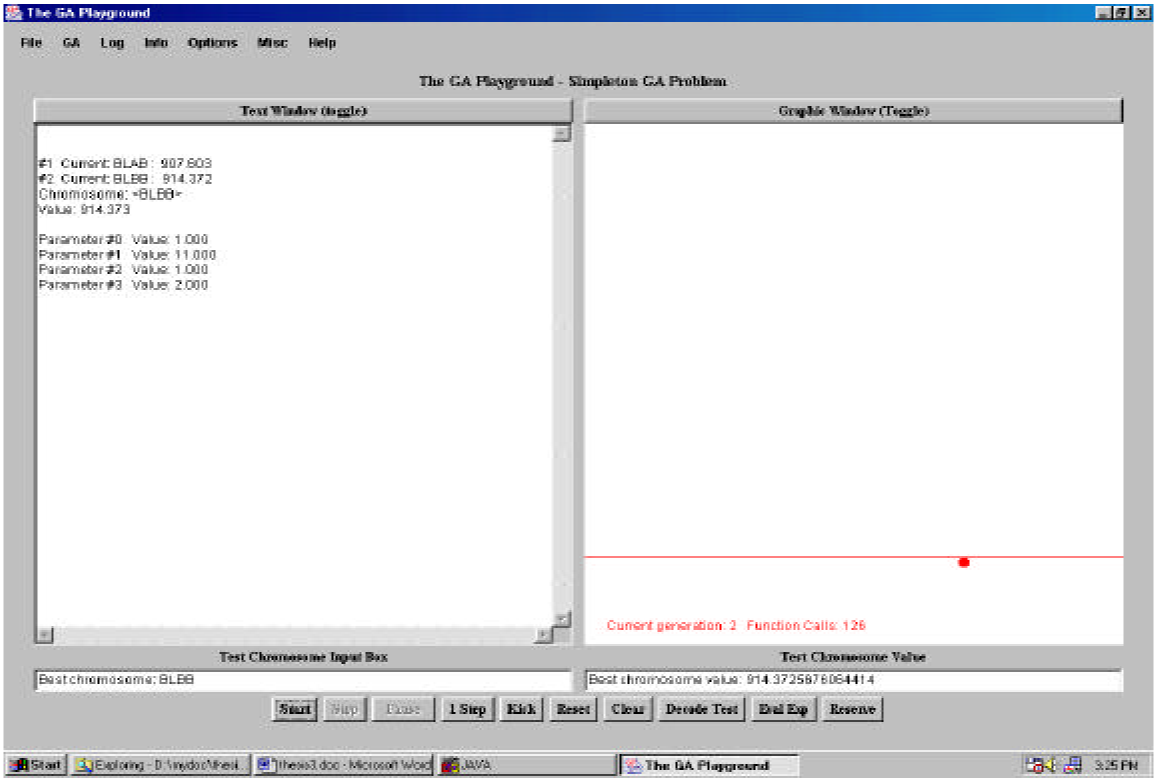
2. GA parameter setting

    To run the GA, a text file with GA parameters setting information must be provided to the GA-playground for GA initialization and configuration. An allele definition file that defines the range of each allele (variable) was also needed to generate the initial population.

    The user must set several parameters before using the GA-playground program (see Table 4-1). Generation gap represents the percentage of individuals to copy to the new generation *T+1* from the old generation *T*. Crossover rate and mutation rate represent the percentage of individuals undergo crossover and mutation. Parameter setting may be adjusted to get the best result. Table 4-1 lists our final parameter settings. The graphic user interface (GUI) of this program shows current search status (see Figure 4-5).

Table 4-1 GA parameter setting

GA type:	Generational GA with generation gap
Initial pool:	Randomly
Stop GA:	1. Until converge; 2. Target value
Population size:	100
Number of gene:	4
Objective:	Maximize
Crossover type:	Single point crossover
Mutation type:	Simply random
Selection type:	Roulette Wheel
Generation gap:	5-25%
Crossover rate:	1.0
Mutation rate:	0.02



```
#1  Current: BLAB :  907.603
#2  Current: BLBB :  914.372
Chromosome: <BLBB>
Value: 914.373

Parameter #0   Value: 1.000
Parameter #1   Value: 11.000
Parameter #2   Value: 1.000
Parameter #3   Value: 2.000
```

Current generation: 2   Function Calls: 126

Best chromosome: BLBB

Best chromosome value: 914.372587606441 4

Figure 4-5 The GA-playground user interface

4.5 Result and discussion

The experimental tests were run on a personal computer with a 166 MHz Pentium processor under Windows 2000 operating system.

We did two experiments using different termination criteria. In the first experiment, the target value (exit value) was not provided to the GA. GA stopped after convergence, that is, when the ratio *average fitness / maximum fitness* is greater than 0.98. The result found by GA was compared to the result found by the SAS procedure. In the second experiment, the target value (the best adjusted R^2 from SAS) was provided to GA. We evaluated GA's performance based on its convergence iterations.

1. Run GA until convergence

We run GA 100 times with each generation gap setting of 5%, 10%, 15%, 20% and 25%. The convergence criteria was set to: *average fitness / maximum fitness > 0.98*. Table 4-2 showed the result.

Table 4-2 GA experimental result using termination criteria: average fitness / maximum fitness > 0.98.

generation gap (%)	average iteration	percentage of true optimum	mean R^2 error	Best bands(nm)	best VI	best model
5	2228	94	0.000055	700, 750	RVI	2
10	2164	90	0.000132	700, 750	RVI	2
15	1354	81	0.000239	700, 750	RVI	2
20	851	60	0.000424	700, 750	RVI	2
25	691	37	0.000804	700, 750	RVI	2

2. Stop GA when reaching the target value (adjusted R^2 = 0.9144). We run GA 100 times. The generation gap was set to 20%. Table 4-3 lists the result.

Table 4-3 GA Experimental result using termination criteria: exit value = 0.9144

Experiment Number	iteration number	best bands (nm)	best VI	best model
1	740	700, 750	RVI	2
2	1046	700, 750	RVI	2
3	156	700, 750	RVI	2
4	555	700, 750	RVI	2
5	116	700, 750	RVI	2
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
99	317	700, 750	RVI	2
100	499	700, 750	RVI	2
Average iterations: 560				

Accuracy and runtime

From the result of experiment one, one can see that as the iteration number increased, GA approached the true optimal solution (See figure 4-6). Big generation gap led to fast convergence but less true optimal solutions percentage wise, while small generation gap led to relatively slow convergence but more true optimal solutions. This was consistent with De Jong's work (1975) that the nonoverlapping population model was best in most optimization studies.

We defined the term *mean absolute R^2 error* to measure the difference between the GA solution and the true optimal solution, which could be represented as:

$$mean\ absolute\ R^2\ error = \frac{\sum_{i=1}^{n} | SASsolution - GAsolution |}{n}$$

In our experiment, the minimal mean absolute R^2 error achieved by GA was 0.000055 with generation gap 5%. As the iteration number decreased, the mean absolute R^2 error increased, which indicated that there was tradeoff between runtime and accuracy.

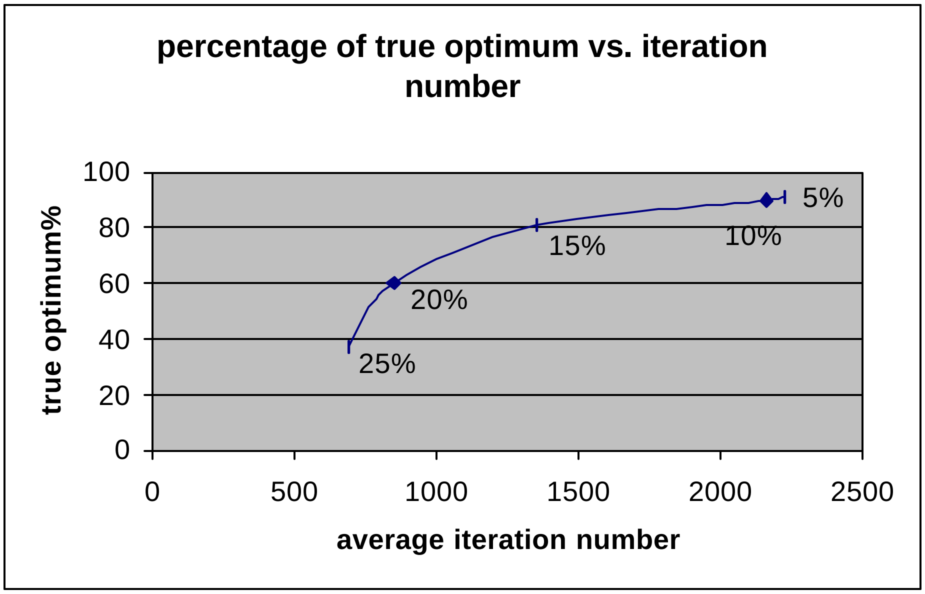


Figure 4-6 The effect of iteration number and generation gap on the percentage of true optimal solution

From the result of experiment two, the convergence number of iterations for GA ranged from 116-1959. The average convergence iteration was 560. In the polynomial regression procedure performed by SAS, where all possible two-bands vegetation indices were exhausted, the total iteration number (number of polynomial regressions executed) were 276 * 4*4 = 4416, where 276 represented the combinations of all wavelength pairs, the first 4 stood for the four types of vegetation indices, and the second 4 stood for four types of polynomial regression models. GA saves almost 85% of the time in the searching. From runtime point of view, GA outperforms SAS procedure.

From our experiments, GA identified the optimal wavelengths 700 nm and 750 nm. RVI was the best vegetation index found by GA. The final regression model for RVI was second degree polynomial:

$$\boldsymbol{RVI = 0.29 - 0.003ppmN + 0.00001ppmN^2}$$

Besides efficiency, another advantage of GA was simplicity. Unlike the regression procedure performed by SAS, which required additional computation to find all two-bands combinations for different vegetation indices before statistical analysis, no pre-computation was needed by GA since the individual encoding (integer permutation) of GA had automatically handled this. GA also performed statistical analysis and fitness selection at the same stage rather than in two separate stages, which was unavoidable in the SAS procedure. Therefore GA was simpler and more efficient since it integrated data pre-processing, statistical analysis and model selection in a single procedure.

One major limitation of the current GA implementation was that the fitness function was not complete. It used only adjusted R^2 as model selection criteria. This may lead to infeasible solutions. In experiment 1, GA may select an insignificant solution (third degree polynomial model) due to the lack of significant test in the fitness function. Further implementation of the fitness function that included a significant test will be needed to overcome this limitation.

CHAPTER 5

SUMMARY AND CONCLUSION

In this thesis, we developed two approaches to identify optimum spectral bands and vegetation indices that could best characterize the relationship between vegetation indices and the bush bean plant nitrogen treatments levels. The first one was an exhaustive search approach, using a SAS procedure of polynomial regression on individual bands. The other was an optimal stochastic search using genetic algorithm (GA).

Our study employed a hyper-spectral imaging system based on a liquid crystal tunable filter and involved 24 discrete spectral bands in the VIS-NIR spectrum and 4 types of vegetation indices (NDVI, RVI, DVI, Reflectance).

In the first approach, all possible two-bands vegetation indices under four nitrogen treatments with 6 plants in each treatment were computed. Polynomial regression was then performed on individual bands to find the relationship between vegetation indices and nitrogen treatments. The optimum spectral bands and vegetation indices were identified by exhaustive search. The result indicated that all vegetation indices correlated well with nitrogen treatments under the corresponding optimum spectral bands. The second degree polynomial regression model of RVI under wavelengths 700 nm and 750 nm performed best with an adjusted R^2 of 0.9144. Depending on the type of vegetation indices, other optimal spectral bands were also determined: 710 nm, 715 nm, and 720 nm.

In the second approach, genetic algorithm was used to search for the optimal solution. Integer permutation encoding was used for the representation of GA individuals. Single point crossover and random mutation were applied. A repair operator was also added to avoid illegal and redundant individuals. The fitness function performed polynomial regression analysis. Each individual was evaluated by the rescaled adjusted R^2 value. Two experiments were carried out with different termination criteria. In the first experiment, we ran GA until the convergence condition, *average fitness / maximum fitness > 0.98,* was met. GA achieved 0.000055 mean absolute R^2 error with generation gap 5%, and 94% of the solutions were true optimal solutions. As the generation gap increased, the convergence number of iterations decreased, but the mean absolute R^2 error increased. Therefore there was a tradeoff between runtime and accuracy. In the second experiment, we ran GA until it found the same best adjusted R^2 provided by the SAS procedure. GA found the best solution in significantly less iteration than the SAS procedure.

Comparing the performance of these two approaches, we may conclude that both can be used to fulfill our goal. The SAS procedure provided accurate result and true optimal solution, but it could not be performed in a single procedure therefore was not efficient and practical. The GA search combined data pre-processing, polynomial regression and regression model selection in one procedure therefore saved time. But it may select an insignificant model due to the limitation of the current fitness function implementation. There was a tradeoff between accuracy and runtime for these two approaches.

CHAPTER 6

FUTURE RESEARCH

Using current available data, we have identified the optimum wavelengths and vegetation indices, and established the statistical models that correlated nitrogen stress levels with vegetation indices. This is the calibration stage. Our future research direction will focus on validating current model and improving GA:

1. Test the established model by cross-validation if more data are available, that is, apply the model to some new data and evaluate the performance of the model.
2. Improve GA by adding *P* value as a model selection criterion. This ensures that the final model selected by GA is statistically significant.
3. After cross-validation test, use the established model to predict nitrogen concentration given the spectral data.

REFERENCES

Anderson, G. L., Hanson, J. D., and Hass, R. H. (1993), Evaluating Landsat thematic mapper derived vegetation indices for estimating above-ground biomass on semiarid rangelands. Remote Sensing Environment. 45:165-175.

Boochs, F., G. Kupfer, K. Dockter, and W. Kuhbauch. 1990. Shape of the red edge as vitality indicator for plants. Int. J. Remote Sensing 11(10): 1741-1753.

Carter, Gregory A. and Miller, L. R, 1994, Early detection of plant stress by digital imaging within narrow stress-sensitive wavebands, Remote Sensing Environment, 50:295-302.

Chen, J. M., and Cihlar, J. (1996), Retrieving leaf area index of boreal conifer forest using Landsat TM images. Remote Sensing Environment, 55:153-162.

Claude. S. and Pierre B., 1991, Determining leaf nitrogen concentration of broadleaf tree seedlings by reflectance measurements, Tree Physiology 8, 391-398.

De Jong, K. A and J. Sarma. 1993. Foundations of Genetic Algorithms 2, 19-28, Morgan Kaufmann, San Mateo.

De Jong, K. A. 1975, An analysis of the behavior of a class of genetic adaptive systems. Dissertation Abstracts International 36(10), 5140B (University Microfilms No. 76-9381).

Evans, M. D., Thai, N. C. and Jason, C. G., 1998, Computer control and calibration of a liquid crystal tunable filter, ASAE Paper No. 973142, ASAE, St. Joseph, MI.

Fillela, I. and J. Penuelas.1994. The Red Edge position and shape as indicators of plant chlorophyll content, biomass and hydric states. Int J. Remote Sensing 15(7): 1459-1470.

Freund, J. R. and Littell, C. R., 2000, SAS System for Regression, third edition, Wiley & Sons, Inc. New York.

Gen, M. and Cheng, R., 2000, Genetic algorithms and engineering optimization, page 4-5, Wiley –Interscience Publication, New York.

Gitelson, A. A, Buschmann, C. and Lichtenthaler, K. H., 1999, The chlorophyll fluorescence ratio F735/F700 as an accurate measure of the chlorophyll content in plants, Remote Sensing Environment, 69:296-302.

Goldberg, D. E. 1989. Genetic algorithm in search, optimization, and machine learning, page 76-79. MA: Addison-Wesley.

Goldberg, D. E., Deb, K., & Thierens, D. (1993). Toward a better understanding of mixing in genetic algorithms. Journal of The Society of Instrument and Control Engineers, 32 (1), 10-16.

Hatfield, J. L., and P. J. Pinter Jr. 1993. Remote sensing for crop protection. Crop Protection 12(6): 403-411.

Holland, J. H., Holyoak, K. J., & Thagard, P. R. (1986), Induction: processes of inference, learning, and discovery. Cambridge: MIT Press.

Johnson, F. L., and Billow, R. C. 1996, Spectrometric estimation of total nitrogen concentration in Douglas-fir foliage. Remote Sensing Environment, 53: 199-211.

Lawrence, L. R. and Ripple, J. W., 1998, Comparisons among vegetation indices and bandwise regression in a highly disturbed, heterogeneous landscape: Mount St. Helens, Washington, Remote Sensing Environment, 64:91-102.

Michalewicz, Z., Genetic Algorithm + Data Structure = Evolution Programs, 3rd edition, Springer-Verlag, New York, 1996.

Moran, M. S., Y. Inoue, and E. M. Barnes. 1997. Opportunities and limitations for image-based remote sensing in precision crop management. Remote Sensing Environment. 61: 319-346.

Neter, J., Kuter, H. M, and Nachtsheim, J. C, 1990, Applied linear statistical models, fourth edition, The McGraw-Hill Companies.

Ning, L., G. E. Edwardes, G. S. Strobel, L. S. Daley, and J. B. Callis. 1995. Imaging fluorometer to detect pathological and physiological change in plants. Applied Spectroscopy 49(10): 1381-1389.

Thai, Chi N, 2000, Influence of canopy types on spectral evaluation of stress in plants, ASAE Paper No. 003020, ASAE, Milwaukee, WI.

Thai, Chi N., M. D. Evans and Xiaode Deng, 1998, Visible and NIR imaging of bush beans grown under different nitrogen treatments, ASAE Paper No. 983074, ASAE, St. Joseph, MI.

Theisen, A. F., L. Jarrell, P. K. Kebalian, and A. Freedman. 1998. Remote detection of vegetation stress using sunlight-excited fluorescence. Proceedings of the First International Conference on Geospatial Information in Agriculture and Forestry, Lake Buena Vista, FL, June 1-3, 1998(Vol. II, pp. 547-552).

Thenkabail, S. P., Smith, B. R. and De Pauw, E., 2000, Hyperspectral vegetation indices and their relationships with agriculture crop characteristics, Remote Sensing Environment, 71:158-182.

Tucker, C. J., 1979, Red and photographic infrared linear combinations for monitoring vegetation. Remote Sensing Environment, 8:127-150.

Wanjura, F. D. and Hatfield. L. J., 1987, Sensitivity of spectral vegetation indices to crop biomass. Transactions of the ASAE, Vol. 30(3): 810-816.

Wiegand, C. J., Richardson, A. J., Escobar, D., E., and Gerbermann, A. H., 1991, Vegetation indices in crop assessments. Remote Sensing Environment, 35:105-119.

Yoder, J. B. and Pettigrew-Crosby, E. R, 1995, Predicting nitrogen and chlorophyll content and concentrations from reflectance spectra (400-2500 nm) at leaf and canopy scales, Remote Sensing Environment, 53:199-211.

Yoder, J. B., and Waring, R. H. (1994), The normalized difference vegetation index of small Douglas-fir canopies with varying chlorophyll concentrations, Remote Sensing Environment. 48:1-25.

Younger, M. S., Handbook for linear regression, Duxbury Press, 1979, 55-80.