

Application of Machine Learning Methods and Airborne Hyperspectral Remote Sensing for Crop Yield Estimation

Yoji Uno

Department of Bioresource Engineering
Macdonald Campus of McGill University

August, 2003

A thesis submitted to the Faculty of Graduate Studies and Research in partial
fulfillment of the requirements of the degree of Masters of Science.

© Yoji Uno, 2003



Library and
Archives Canada

Bibliothèque et
Archives Canada

Published Heritage
Branch

Direction du
Patrimoine de l'édition

395 Wellington Street
Ottawa ON K1A 0N4
Canada

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

ISBN: 0-612-98755-8

Our file Notre référence

ISBN: 0-612-98755-8

NOTICE:

The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protègent cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.


Canada

Abstract

Accurate description of within-field crop yield variability is one of the greatest concerns in precision agriculture. This study investigated the potential of developing in-season crop yield forecasting and mapping systems based on interpretation of airborne hyperspectral remote sensing imagery by machine learning algorithms. The data used for this study was obtained over a corn (*Zea mays* L.) field in eastern Canada.

The experimental plots were set up at the Emile A. Lods Agronomy Research Center, Montréal, Québec. Corn was grown under the twelve combinations of three nitrogen application rates (60, 120, and 250 kg N /ha), and four weed control strategies (Broad leaf weed, Grass weed, Broad leaf and grass weed control, and no weed control). The images of the experimental field were taken with a Compact Airborne Spectrographic Imager (CASI) at three times (June 30 for early growth stage, August 5 for tassel stage, and Aug 25 for mature stage) during the year 2000 growing season.

Two machine learning algorithms, Artificial Neural Networks (ANN) and Decision Tree (DT) were evaluated. The performance of ANNs was compared with four conventional modeling methods, namely, Normalized Difference Vegetation Index (NDVI), Simple Ratio (SR), Photochemical Reflectance Index (PRI), and Stepwise Multiple Linear Regression (SMLR) models. Principal Component Analysis (PCA) was also used to reduce the dimensionality of the 72-band hyperspectral imagery. The results showed that much higher performance was obtained with ANNs than with three VI-based methods, although no clear difference was observed between SMLR and ANNs.

For the DT algorithms, two different aspects, (i) DT as a classification method, and (ii) DT as a feature selection tool, were explored in this study. The performance of the algorithms was evaluated, as compared with conventional reclassification methods (for performance in image classification) and PCA (for performance as a feature selection tool). The results demonstrated that the performance of the DT is comparable to that of the conventional methods for yield classifications. However, it did not seem to be

sufficient for practical purposes. As a feature selection tool, the DT algorithms performed better than PCA. However, the fact that the DTs require a large number of training samples needs to be addressed before this algorithm can be applied as an operative technology.

Résumé

La représentation exacte de la variabilité du rendement du maïs dans les champs est l'un des problèmes majeurs de l'agriculture de précision. Cette étude s'intéresse au développement de méthodes de prévision de rendement et de systèmes de cartographie basés sur l'interprétation d'imageries hyperspectrales aériennes, interprétation qui sera effectuée par des algorithmes d'apprentissage automatique. Les données utilisées pour cette étude ont été obtenues à partir d'un champ de maïs (*Zea mays* L.) de l'est du Canada.

Les parcelles employées pour les expérimentations ont été mises en place au Centre de Recherche Agronomique Emile A. Lods de Montréal (Québec). Le maïs a été cultivé sous diverses conditions: trois taux d'application d'azote (60,120 et 250 kg N/ha) et quatre stratégies de contrôle des mauvaises herbes (contrôle des dicotylédones, des graminées, contrôle des deux à la fois et aucun contrôle), soit douze environnements différents. Les images du champ expérimental ont été prises à l'aide d'un ISCA (Imageur Spectrographique Compact Aéroporté) à trois reprises au cours de l'année 2000: le 30 juin pour le début de la phase de développement, le 5 août pour la floraison et le 25 août à maturité.

Deux algorithmes d'apprentissage automatique, un RNA (Réseau de Neurones Artificiel) et un AD (Arbre Décisionnel), ont été évalués. Les performances du RNA ont été comparées avec quatre méthodes de modélisation conventionnelles: l'IVDN (Indice de la végétation par différence normalisée), le RS (Rapport Simple), l'indice de réflectance PRI et l'analyse de régression linéaire par degrés. L'ACP (Analyse en Composante Principale) a également été utilisée afin de réduire la dimensionnalité de l'imagerie hyperspectrale qui comptait soixante-douze bandes de fréquences. Les résultats ont montré que les RNA avaient des performances bien supérieures à celle des trois méthodes basées sur les index de végétation. En revanche, aucune différence flagrante n'a été observée entre la SMLR et les RNA.

Les AD ont été abordés sous deux aspects différents au cours de cette étude: (i) les AD en tant que méthode de classification (ii) les AD en tant qu'outil de sélection des caractéristiques des bandes de fréquences. Les performances des algorithmes ont été évaluées en comparaison avec des méthodes conventionnelles de reclassification (pour la classification d'images) et l'ACP (pour la sélection des caractéristiques des bandes de fréquences). Les résultats ont démontré que les performances des AD sont comparables à celles des méthodes conventionnelles pour la classification de rendement. Les AD n'ont cependant pas semblé suffisants dans la pratique. En tant qu'outil de sélection des caractéristiques des bandes de fréquences, les algorithmes d'AD ont surclassé l'ACP. Mais le problème que posent les AD du fait du nombre important d'échantillons d'étalonnage qu'ils nécessitent devra être résolu avant que cet algorithme puisse être utilisé sur le terrain.

Acknowledgement

First, I would like to express my appreciation to my thesis supervisor, Dr. Shiv. O. Prasher, for instruction and encouragement through this thesis work, and for the opportunity to work on this remote sensing project.

I would also thank Dr. Robert. B. Bonell, co-organizer of the remote sensing project, Dr. Rene Lacroix, for precious advices regarding machine learning algorithms and artificial intelligence, and Dr. Ian B. Strachan for the advice on spectral measurement and analysis.

I also acknowledge all the graduate students, staff and summer students, who worked together during this period, especially Dr. Pradeep. K. Goel and Mr. Yousef Karimi, Dr. R. M. Patel, Mr. Kenton Ollivierre, and Mr. Khaldoun El-Dirani, who spent the entire summer at the research farm to execute the field work, Dr. Chun-Chieh Yang for advice on ANNs, Mr. Marc-David Andrade for assistance in design of the irrigation system, Dr. Carlos Costa for consultation on statistical analysis, and Mr. Peter Alvo and Dr. Georges T. Dodds for editing my thesis.

Finally, I would also like to thank all the members of Canadian GEOIDE (Geomatics for Informed Decision) project, especially, Dr. A. A. Viau, Université Laval, project leader, and Dr. John Miller at York University for hyperspectral image acquisition and preprocessing.

Contribution of Authors

The following statement from the Guidelines for Thesis Preparation was included, based on the regulations of the Faculty of Graduate Studies and Research of McGill University.

As an alternative to the traditional thesis format, the dissertation can consist of a collection of papers of which the student is an author or co-author. These papers must have a cohesive, unitary character making them a report of a single program of research. The structure for the manuscript-based thesis must conform to the following:

- 1. Candidates have the option of including, as part of the thesis, the text of one or more papers submitted, or to be submitted, for publication, or the clearly-duplicated text (not the reprints) of one or more published papers. These texts must conform to the "Guidelines for Thesis Preparation" with respect to font size, line spacing and margin sizes and must be bound together as an integral part of the thesis. (Reprints of published papers can be included in the appendices at the end of the thesis.)*
- 2. The thesis must be more than a collection of manuscripts. All components must be integrated into a cohesive unit with a logical progression from one chapter to the next. In order to ensure that the thesis has continuity, connecting texts that provide logical bridges between the different papers are mandatory.*
- 3. The thesis must conform to all other requirements of the "Guidelines for Thesis Preparation" in addition to the manuscripts.*

The thesis must include the following:

- (a) a table of contents;*
 - (b) an abstract in English and French;*
 - (c) an introduction which clearly states the rationale and objectives of the research;*
 - (d) a comprehensive review of the literature (in addition to that covered in the introduction to each paper);*
 - (e) a final conclusion and summary;*
- 4. As manuscripts for publication are frequently very concise documents, where appropriate, additional material must be provided (e.g., in appendices) in sufficient*

detail to allow a clear and precise judgment to be made of the importance and originality of the research reported in the thesis.

5. In general, when co-authored papers are included in a thesis the candidate must have made a substantial contribution to all papers included in the thesis. In addition, the candidate is required to make an explicit statement in the thesis as to who contributed to such work and to what extent. This statement should appear in a single section entitled "Contributions of Authors" as a preface to the thesis. The supervisor must attest to the accuracy of this statement at the doctoral oral defense. Since the task of the examiners is made more difficult in these cases, it is in the candidate's interest to clearly specify the responsibilities of all the authors of the co-authored papers.

Manuscripts based on the thesis:

1. Y. Uno, S. O. Prasher, P. K. Goel, Y. Karimi, and A. A. Viau. Artificial neural networks to predict corn yield from Compact Airborne Spectrographic imager (CASI) data (Under preparation).
2. Y. Uno, S. O. Prasher, P. K. Goel, Y. Karimi, and A. A. Viau. Use of classification tree and compact airborne spectrographic imager (CASI) for corn yield estimation (Under preparation).

All data analysis and manuscript preparation for these papers were conducted by the candidate, Yoji Uno, under the supervision of S. O. Prasher, Professor of Bioresource Engineering, McGill University.

Dr. P. K. Goel and Mr. Y. Karimi of the Department of Bioresource Engineering, McGill University, contributed to the field experiments. Design of the experimental plots and data acquisition were all conducted with these two co-authors.

Professor A. A. Viau, Professor at Université Laval, was the leader of the GEOIDE (Geomatics for Informed Decisions) project, which funded this research work.

Table of contents

CHAPTER 1- INTRODUCTION	1
1.1 Problem statement	1
1.2 Objectives	3
1.3 Scope	3
1.4 Thesis organization	4
1.5 References	5
 CHAPTER 2 - LITERATURE REVIEW	 7
2.1 Introduction	7
2.2 Spectral signature of vegetation	7
2.3 Crop yield prediction	8
2.3.1 Estimating crop yield from physical environmental factors	9
2.3.1.1 Statistical analysis	9
2.3.1.2 Mechanistic crop growth modeling	9
2.3.2 Use of chlorophyll meters for yield estimation	11
2.3.3 Concluding remarks	11
2.4 Use of remotely sensed data for yield estimation	12
2.4.1 Sensors and platforms	12
2.4.2 Model development based on remotely sensed data	13
2.4.2.1 Regression models and vegetation indices	14
2.4.2.2 Coupling remotely sensed data with crop growth modeling	15
2.4.3 Concluding remarks.....	16
2.5 Machine learning algorithm and image interpretation	17
2.5.1 Artificial neural networks	18
2.5.2 Decision tree estimation algorithms	20
2.5.3 The Clementine Data mining System	22
2.6 Summary	23
2.7 References	25
 Figures and Tables	 34
 PREFACE TO CHAPTER 3	 45
 CHAPTER 3 – ARTIFICIAL NEURAL NETWORKS TO PREDICT CORN YIELD FROM COMPACT AIRBORNE SPECTROGRAPHIC IMAGERY DATA	
3.1 Abstract	46
3.2 Introduction	47

3.3 Methodology	49
3.3.1 Experimental design and image acquisition	49
3.3.2 Data Manipulation	50
3.3.3 Principal component analysis	51
3.3.4 Model development	52
3.3.5 Vegetation indices	52
3.3.6 Stepwise multiple linear regression models	53
3.3.7 Artificial neural networks	53
3.3.8 Performance analysis	55
3.4 Results and Discussion	56
3.5 Conclusions	60
3.6 References	62
 Figures and Tables	66
 PREFACE TO CHAPTER 4	79
 CHAPTER 4 - USE OF CLASSIFICATION TREE AND COMPACT AIRBORNE SPECTROGRAPHIC INTEGRATION IMAGER (CASI) FOR CORN YIELD ESTIMATION	
4.1 Abstract	80
4.2 Introduction	81
4.3 Methodology	82
4.3.1 Image acquisition and data preparation	82
4.3.2 Principal component analysis	83
4.3.3 C5.0 decision tree algorithm	83
4.3.4 Crop yield classification	84
4.3.4.1 Determination of decision boundaries	84
4.3.4.2 Model development	85
4.3.4.3 Performance analysis	86
4.3.5 Feature band selection	86
4.4 Results and Discussion	87
4.5 Conclusions	90
4.6 References	91
 Figures and Tables	94
 CHAPTER 5 - SUMMARY AND CONCLUSIONS	108
5.1 Summary	108
5.2 Conclusion	109

CHAPTER 6 - RECOMMENDATIONS FOR FURTHER RESEARCH111

6.1 Increasing the generality of models111

6.2 Further exploration of machine learning algorithms112

6.3 References113

APPENDICES

List of Figures

Number	Titles	Pages
Figure 2.1	Schematics of STICS mechanistic crop growth model	34
Figure 2.2	A McCulloch and Pitts model	35
Figure 2.3	Basic taxonomy of artificial neural network architecture	36
Figure 2.4	Relationships between decision boundaries and network structures in feed-forward networks	37
Figure 2.5	An example of tree representation for the human decision making process	38
Figure 2.6	An example of a hybrid decision tree classifier	39
Figure 2.7	An example of univariate and multivariate decision tree classifiers	40
Figure 2.8	Geometric interpretation of univariate and multivariate decision tree classifiers	41
Figure 2.9	A schematic of data mining process	42
Figure 3.1	Seven different modeling strategies taken for yield prediction	66
Figure 3.2	The ANN network structure used in this study	67
Figure 3.3	Performance of five different models for yield prediction (calibration)	68
Figure 3.4	Performance of five different models for yield prediction (validation)	69
Figure 3.5	Results of the ten-fold cross validation obtained with ANN model with 71 input variables	70
Figure 3.6	Difference of the model performance between original dataset and reduced dataset (calibration)	71
Figure 3.7	Difference of the model performance between original dataset and reduced dataset (validation)	72

Figure 4.1	An example of tree representation for the human decision making process	94
Figure 4.2	Distribution of yield samples	95
Figure 4.3	Three different classification strategies for yield classification	96
Figure 4.4	Three different input strategies for the performance analysis of C5.0 algorithm as a spectral band selection tool	97
Figure 4.5	Deviation of the overall classification accuracies (classification into four-levels) obtained with ten different validation datasets	98
Figure 4.6	Deviation of the overall classification accuracies (classification into two-levels) obtained with ten different validation datasets	99
Figure 4.7	Performance of ANN models with three different input strategies	100
Figure 4.8	Structure of a developed classification tree	101

List of tables

Number	Titles	Pages
Table 2.1	Symbols used in the schematic (Figure 2.1) for STICS	43
Table 2.2	Equations of commonly used vegetation indices and their references	44
Table 3.1	CASI specification and data processing	73
Table 3.2	Prediction accuracies obtained with seven different modeling strategies	74
Table 3.3	Results of ten-fold cross validation for the ANN model with 71 input variables	75
Table 3.4	Structure of the developed ANNs	76
Table 3.5	Equations of the developed SMLR models and VI-based linear models	77
Table 3.6	Eigenvalues of five principal components (PC) used for the model development	78
Table 4.1	Overall classification accuracies obtained with different classification and input strategies	102
Table 4.2	A confusion matrix obtained by C5.0 classifier with 71 input variables	103
Table 4.3	A confusion matrix obtained by C5.0 classifier with five principal components	104
Table 4.4	A confusion matrix obtained with performance standard (reclassification from ANN prediction values)	105
Table 4.5	Confusion matrices obtained with the second classification strategy (classification into two yield levels)	106
Table 4.6	Performance of the ANN models with three different input strategies	107

List of Symbols

ANNs = Artificial Neural Networks
ARVI = Atmospherically Resistant vegetation index
AVDIF = Average Difference
AVHRR = Advanced Very High Resolution Radiometer
CCD = Charged Coupled Device
CASI = Compact Airborne Spectrographic Imager
CRT = Classification and Regression Tree
DT = Decision Tree
fPAR = fraction Photosynthetic Active Radiation
FOV = Field of View
GNDVI = Green Normalized Difference Vegetation Index
GPS = Global Positioning System
ha = hectare
kg = kilogram
LAI = Leaf Area Index
m = meter
MLR = Multiple Linear Regression
nm = nanometer
N = Nitrogen
NDVI = Normalized Difference Vegetation index
NOAA = National Oceanic and Atmospheric Administration
NVI = Normalized Vegetation Index
PC = Principal Component
PCA = Principal Component Analysis
PE = Processing Element
PRI = Photochemical Reflectance Index
PVI = Perpendicular Vegetation Index
R = Coefficient of correlation
 R^2 = correlation of determination

RMSE = Root Mean Square Error
SAIL = Scattering by Arbitrarily Inclined Leaves
SAVI = Soil Adjusted Vegetation Index
SAR = Synthetic Aperture Radar
SMLR = Stepwise Multiple Linear Regression
SPAD = Speciality Products Agricultural Division
STICS = Simulateur mulTIdisciplinaire pour les Cultures Standard
SUCROS = Simple and Universal Crop Growth Simulator
SR = Simple Ratio Index
TSAVI = Transformed Soil-Adjusted Vegetation Index
VI = Vegetation Index
VRT = Variable Rate Technology
WDVI = Weighted Differences Vegetation Index
°C = Degree Celsius

Chapter 1 – INTRODUCTION

1.1 Problem statement

Description of within-field crop yield variability, or yield mapping, is currently one of the most commonly practiced methods in precision agriculture. Indeed, the yield map is not only regarded as a simple indicator of the productivity or fertility of a soil, it is also a useful diagnostic tool to identify various environmental factors that create the within-field variability (Reitz et al., 1996; Stafford et al., 1996). Since crop yield can be regarded to be an effective sensor of field conditions, various environmental factors such as water or nutrient deficiency are potentially detectable from the yield map. The information obtained from a yield map can be used to guide variable rate technologies (VRT) used in site-specific application of fertilizer and pesticide, as well as an aid in the design of irrigation and drainage systems, and windbreaks (Swinton et al., 1998)

The usefulness of the yield mapping in precision farming is clearly demonstrated by the increasing number of commercial crop yield monitoring systems, while some economic risks are also indicated due to their high initial cost in purchasing equipments and training the technicians (Swinton et al., 1998). According to Swinton et al. (1998), net profitability obtained from the introduction of crop yield monitoring systems largely depends on the crop types and the environmental conditions. However, effective use of these technologies could produce high profits for the high-value crops, such as for potato and sugar beet, and in some water-fed environments, such as exist in the southwestern United States (Earl et al., 1996; Swinton et al., 1998).

Use of remote sensing technologies is currently recognized to be the next generation of technical innovations that have the potential to refine the quality of within-field yield mapping technologies. Above all, the ability of realizing in-season yield mapping, or yield prediction, is the most important advantage (Yang et al., 200, 2001), in contrast to the conventional post-harvest yield analysis that apply combine-mounted yield monitoring systems. Some of the limitations of post-harvest analysis, such as difficulty in identifying seasonally changing factors, such as precipitation, temperature, and sunshine, could be surmounted using remote sensing technologies on a real-time basis (Swinton et al., 1998).

Although the potential benefits obtained from the real-time yield mapping systems are obviously high, commercialization of the concept still requires a lot of technical breakthroughs. One of the current problems is the difficulty in identifying the most appropriate vegetation index (VI) in a specific environment (Barrett and Curtis, 1999). Although recent scientific researches have established some basic relationships between various VIs and crop conditions or yield, there is still no clear guideline for selecting the best VI in a specific condition, because the spectral signature of canopy reflectance is always influenced by environmental conditions such as soil types and weather conditions (Rondeaux et al., 1996). Another problem is that the large volume of spectral information acquired with the latest sensors exceeds the capability of conventional VI-based methods, because most of the vegetation indices are calculated from simple combinations of several spectral wavelengths. New methodologies are currently needed to extract the whole potential of this large amount of information (Barrett and Curtis, 1999).

Recent studies show that coupling mechanistic crop growth models with remotely sensed information has the potential to become another effective method for estimating yield since the theoretical background of this approach is sometimes advantageous for the application of the models to different crop types and environments. However, the need to estimate crop and field parameters still prevents the wide application of models to practical situations, especially, when spatial variability of the field is high. In fact, it is documented that application of this approach to large-scale investigations is still unrealistic since it becomes extremely labor-intensive (Guérif and Duke, 1998). Complexity of the models is also indicated as another serious problem of this approach (Varcoe, 1990).

Use of machine learning technologies, especially artificial neural networks (ANNs) and decision tree (DT) estimation algorithms, could be effective alternatives for the creation of yield maps and for the development of yield forecasting system, along with the conventional empirical and mechanistic approaches. Above all, one of the most important characteristics of machine learning algorithms, “ability to learn”, offers various advantages in the development of flexible and adaptable image interpretation systems (Mather, 1999; Kimes et al., 1998; Atkinson and Tatnall,

1997). Such algorithms can be used to describe intricate non-linear relationship between spectral information and crop yield without human intervention through training.

Although the ability of currently available machine learning systems is still much lower than the ability of human brains in recognizing various features of images efficiently, the goal of making “intelligent image recognition systems” is not limited towards the development of a practical image analysis tool, but is also a motivating scientific challenge. Evaluating the performance of various machine learning algorithms for yield mapping and forecasting is currently an important issue in precision agriculture.

1.2 Objectives

Although the ultimate goal of this project is the development of in-season crop yield mapping and forecasting systems, based on hyperspectral remote sensing, the work involved an assessment of the performance of machine learning algorithms in yield estimation based on hyperspectral data sets. Two types of algorithm were investigated: Artificial Neural Networks (ANNs) and Decision Trees (DT). The hyperspectral data was obtained from an airborne sensor over a corn field in eastern Canada. The performance of the machine learning algorithms was assessed, as compared with conventional modeling methods, such as multiple linear regression models, and some VI-based models. Principal component analysis (PCA) was also used to reduce the large number of spectral bands obtained with the hyperspectral sensor.

1.3 Scope

In this study, the experiment was conducted in a corn field, in eastern Canada. The experimental plots had different nitrogen application rates and weed control strategies to simulate the various crop growth scenarios that occur in practice. However, it should be noted that the field layout was only set up for one year since this study focused on the performance of machine learning algorithms rather than on predictions of seasonal variations in crop yield. Therefore, further study is

recommended in different years with various field conditions and crop types to demonstrate the generality of the methods.

1.4 Thesis organization

This thesis consists of six chapters. The introduction, Chapter 1, introduces the importance, objectives, and scope of this study. The literature review, Chapter 2, presents background information relevant to this study. Recent research efforts on crop yield estimation, remote sensing, and machine learning algorithms, are presented in this chapter. A series of experiments, analyses, and results conducted for this thesis work are summarized in Chapters 3 and 4. These two chapters are formatted for paper submission. In chapter 3, the study focused on the performance of the ANN approach to image analysis, while in chapter 4, the study was focused on the exploration of decision tree estimation algorithms. Chapter 5 contains summary and conclusions. Finally, guidelines for further research are presented in Chapter 6.

1.5 References

- Atkinson, P. M. and A. R. L. Tatnall. 1997. Neural networks in remote sensing. *International journal of remote sensing* 18(4): 699-709
- Barrett E. C. and L. F. Curtis OBE. 1999. *Introduction to environmental remote sensing*. Stanley Thornes Publishers Ltd., Cheltenham, UK
- Earl, R., Wheeler P. N., Blackmore B. S. and R. J. Godwin. 1996. Precision farming – The management of variability. *Landwards* 51(4):18-23
- Guérif M. and C. L. Duke 2000. Adjustment procedures of a crop model to the site specific characteristics of soil and crop using remotely sensing data assimilation. *Agriculture, ecosystems and environment* 81 (1): 57-69
- Kimes, D. S., R. F. Nelson, M. T. Manry and A. K. Fung. 1998. Attributes of neural networks for extracting continuous vegetation variables from optical and radar measurements. *International journal of remote sensing* 19(14): 2639-2663
- Mather, P. M. 2000. *Computer processing of remotely-sensed images: an introduction*. John Wiley & Son Ltd., Chichester, UK
- Reitz, P., and H.D. Kutzbach. 1996. Investigations on a particular yield mapping system for combine harvesters. *Computers and Electronics in Agriculture* 14(2-3): 137-150
- Rondeaux, G., M. Steven, F. Baret. 1996. Optimization of soil-adjusted vegetation indices. *Remote sensing of environment* 55(2): 95-107
- Stafford, J. V., B. Ambler, R. M. Lark and J. Catt. 1996. Mapping and interpreting the yield variation in cereal crops. *Computers and electronics in agriculture* 14(2-3): 101- 119
- Swinton, S. M. and J. Lowenberg- DeBoer. 1998. Evaluating the profitability of site-specific farming. *Journal of production agriculture*. 11(2):439-446
- Varcoe, V. J. 1990. A note on the computer simulation of crop growth in agricultural land evaluation. *Soil use and management* 6(3):157-163
- Yang, C., J. H. Everitt, J. M. Bradford and D. E. Escobar. 2000. Mapping grain sorghum growth and yield variations using airborne multispectral digital imagery. *Transaction of the ASAE* 43(6): 1927-1938.

Yang, C., J. M. Bradford and C. L. Wiegand. 2001. Airborne multispectral imagery for mapping variable growing conditions and yields of cotton, grain sorghum, and corn. *Transaction of the ASAE* 44(6): 1983-1994

Chapter 2 - LITERATURE REVIEW

2.1 Introduction

This literature review presents background information pertaining to crop yield prediction, remote sensing, and machine learning methods. The review consists of two main parts: (i) literature review on crop yield prediction and remote sensing as applied to yield prediction, and (ii) literature review on machine learning algorithms and image interpretation. In the first part, various methodologies for yield prediction are introduced. Problems of conventional methodologies and the usefulness of remote sensing systems for yield prediction are discussed. In the second part, basic concepts and recent publications of two commonly used machine learning algorithms, ANNs and DT, are presented. A simple description of the data mining software used in this study (*Clementine Data mining systems, SPSS Inc.*) is also covered in this part.

2.2 Spectral signature of vegetation

A plant leaf does not absorb all wavelengths uniformly since it consists of biochemical and chemical substances with different absorption peaks. It is documented that various pigments - such as chlorophyll-a and b, anthocyanin, α and β -carotenoids, lutein, violaxanthin - the physical structure of leaves and their water content are the main factors determining the spectral signature (Zwiggelaar, 1998; Campbell, 2002).

An interesting point is that these components change during the season, depending on the phenophase, species, and nutrient conditions. The spectral signatures of leaves can therefore be utilized to monitor various plant conditions, such as nutrient and water deficiency, and to identify crop types and weeds.

At the canopy level, the absorption peaks of the chemical substances generally become unclear, since reflectance from canopy is strongly influenced by leaf size, plant density, number of layers, orientation of leaves, and other environmental factors such as soil reflectance and light angle

(Campbell, 2002). However, some of the biochemical properties are still detectable from the canopy reflectance. Indeed, the canopy reflectance spectrum has been used to identify crop types (Serpico et al., 1996) and weeds (Goel et al., 2002a, 2002b; Lamb et al., 1999; Lamb and Weedon, 1998; Lass et al., 1996), to detect water and nutrient stress (Goel et al., 2002a, 2002b; Lelong et al., 1998; Shibayama et al., 1993; Strachen et al., 2002), and to estimate crop phenology (Raillyan and Korobov, 1993; Boissard et al., 1993), leaf area index (Aparicio et al., 2002; Rastogi et al., 2000; Shibayama and Akiyama, 1989) and plant biomass (Serrano et al., 2000).

2.3 Crop yield prediction

Predicting crop yield before the harvest is one of the greatest concerns in agriculture, since variations in crop yield from year to year impact international trade, food supply, and market prices (Hayes and Decker 1998). Early prediction of crop yield on the global and regional scales offers useful information to policy planners. Appropriate recognition of crop productivity is essential for sound land use planning and economic policy (Hayes and Decker 1996). At the field-scale, crop yield information helps farmers to make quick decisions for upcoming situations, such as the choice of alternative crops and whether to abandon a crop at an early stage of growth. More recently, assessment of crop productivity at the within-field level has become an important issue in precision farming (Stafford 2000; Yang et al., 2001a). Describing the within-field variability of crop yield on a real-time basis offers precious information for VRTs (Stafford, 2000; Yang et al., 2001a).

The relationships between various environmental factors - typically meteorological information and soil parameters - and crop yield, has been the most common approach to predicting crop yields in past years (Varcoe, et al. 1990; Drummond et al., 2003). More recently, relationships between chlorophyll content in crop leaves and grain yield have been explored using SPAD chlorophyll meters (Daughtry et al., 2000; Costa et al., 2001; Smeal and Zhang, 1994; Blackmer and Schepers, 1995; Piekielek et al., 1995). Prediction of crop yield based on remotely sensed data has not yet become the norm.

2.3.1 Estimating grain yield from physical environmental factors

2.3.1.1 Statistical analysis

Tremendous efforts have been made to establish empirical relationship between crop productivity and various environmental factors (Drummond et al., 2003). Many soil properties (cation exchange capacity (CEC), pH, organic matter, phosphorous, calcium, magnesium, and potassium), soil characteristics (texture, soil types, and top soil depth), and climatic information (rainfall, temperature, and radiation from the sun) have been explored using linear and non-linear regression analysis, and ANNs (Drummond, et al., 2003; Kitchen et al., 1999; Kravchenko and Bullock, 2000).

Although these research efforts have established some basic relationships between environmental factors and soil fertility, the methods are generally regarded to be unrealistic for practical purposes since the monitoring of all these parameters over the growing seasons in highly variable field conditions is too labour-intensive. There also seem to be limitations in describing the complicated relationships between these environmental factors and crop yield by using simple empirical equations.

2.3.1.2 Mechanistic crop growth modeling

From the scientific point of view, elucidating the process of carbon assimilation and grain production is a challenge due to its complexity and due to its economic importance (Shibayama et al., 1991). Mechanistic crop growth models simulate the process of the carbon assimilation by using various environmental factors such as climatic information, management practice, and soil characteristics. In the mechanistic crop growth models, various empirically and physically based relationships between physical environments and crop and soil conditions are integrated to simulate crop growth and estimate grain yield (Figure 2.1 and Table 2.1).

Although some variations exist among crop growth models, mechanistic models usually consist of several different modules that simulate water and nitrogen availability in soil, root growth, Leaf Area Index (LAI), stem development, biomass, development stage, and nitrogen and water content of plants, from which grain yield is estimated. Commonly used inputs are: temperature, rainfall, and solar radiation (climatic information); fertilizer application, plant population density, irrigation, and tillage (management practices); and, soil types, depth, field capacity, and soil organic matter (as soil characteristics). Since collecting all these field parameters is difficult in many cases, the models normally work with default values. However, calibration of models with all of these field parameters is necessary to obtain reliable and accurate predictions.

Over the years, many models have been developed for different crop types and locations. Most of the mechanistic models are actually crop-specific: SOYGRO for legumes (Wilkerson et al., 1983), CERES-Maize (Ritchie et al., 1989) for corn, and CERES-wheat (Ritchie et al., 1985) for wheat. However, some models such as SUCROS (Simple and Universal Crop growth Simulator; Spitters et al., 1989) and STICS (Simulateur multIdisciplinaire pour les Cultures Standard; Brisson et al., 1998) are designed for various crop types through optimization.

Mechanistic crop growth models are advantageous in that interpretability of the models is generally high due to the theoretical background. This characteristic is useful for customization and localization of the models as well as for theoretical studies (Clevers, 1997). It is also advantageous that mechanistic crop growth models normally require less data for calibration, compared with the statistical approach. However, the number of field parameters required in order to obtain reliable and accurate simulations, still seemed to be large, especially in fields with high spatial variability. Complexity of the models is another problem because it often requires a certain amount of time for training. Indeed, the models often become black-box models for non-specialists due to this complexity (Varcoe, 1990). It should be also noted that limitations exist to completely simulate the intricate relationships between physical environments and crop growth in a mechanistic way.

2.3.2 *Use of chlorophyll meters for yield prediction*

Whereas mechanistic models generally use physical environmental factors to simulate crop growth, efforts to directly correlate chlorophyll content of leaves and crop productivities by using SPAD chlorophyll meters are also found in the recent literature (Blackmer and Schepers, 1995; Costa et al., 2001; Smeal et al., 1994). Since chlorophyll is one of the most effective indicators of the intensity of photosynthetic activities, this approach can simplify model development. In fact, past studies show that quite high correlations ($R^2=0.8$ or higher) exist between SPAD readings and grain yields, even though the results largely depend on the plant development stage, genotypes, and field conditions (Blackmer and Schepers, 1995; Costa et al., 2001; Smeal et al., 1994). The SPAD chlorophyll meter also yields a non-destructive and efficient measure of leaf chlorophyll content, and consequently reduces the time-consuming sampling of field and crop parameters and their analysis. However, it should be noted that accurate estimation of crop yield at the field scale still requires a large number of sample chlorophyll measurements due to the high variation in the plants and leaves. As mentioned above, generality of the results is still unclear, since the results largely depend on the plant species, genotypes, and environmental conditions (Daughtry et al., 2000).

2.3.3 *Concluding remarks*

Over the decades, many efforts have been made to develop methods of predicting crop yield. Before remote sensing techniques were introduced, crop yield was estimated from soil qualities, management practices, crop conditions, and meteorological data. Statistical and mechanistic approaches have been used to simulate crop growth and finally estimate grain yield. More recently, it is documented that the SPAD chlorophyll meter is an effective tool for the estimation of crop yield.

Although these studies showed that crop yield is somewhat predictable, some limitations have also been indicated, especially, in terms of labor-intensity of the methods. In many cases, these approaches are not realistic for large agricultural fields, where high spatial variability is expected. New methodologies, which effectively collect various field and crop parameters simultaneously at

large scale, are currently in demand. Remote sensing is one of the most effective alternatives due to the wide field of view (FOV).

2.4 Use of remote sensing for crop yield estimation

2.4.1 *Sensors and platforms*

The type of sensor and platform used to gather data are primary considerations in the development of yield forecasting or mapping systems. Many kinds of sensors and platforms have been explored for this purpose.

In past years, satellite platforms have been used to gather data for yield forecasting, since image acquisition over many years is easier, or less expensive, than airborne platform. In particular, the National Oceanic and Atmospheric Administration (NOAA) Advanced Very High Resolution Radiometer (AVHRR) images have been intensively explored for global-scale yield forecasting due to the short revisit period. Research on maize production in the United States Corn Belt (Hayes and Decker, 1996 and 1998), millet and sorghum yields in Niger (Masselli et al., 2000), corn yield estimation and drought monitoring in Southern Africa (Unganai and Kogan, 1998), and wheat yield estimation in India (Manjunath et al., 2002) have been conducted with satellite imagery since the 1990s.

In general, these studies have demonstrated the high potential of satellite imagery in yield forecasting. However, application of satellite images for field-scale investigations is still unrealistic due to the current limitation of satellite systems: coarse spatial resolution, longer revisit times, and strong effect of weather conditions.

While satellite images have been used mainly for global- or regional-scale yield forecasting, most of field-scale studies have been conducted with aerial digital and film photography. In the early studies, aerial film photography was the most common method due to the low cost and relatively high spatial resolution (Arnold et al., 1985; Brown et al., 1994; Curran, 1985; Plant et al., 2000).

However, more recent work gradually focused on the use of aerial digital imaging systems, such as multi- and hyper-spectral scanners and Charged Coupled Device (CCD) cameras (Lamb and Weedon, 1998; Lamb et al., 1999; Yang et al., 2000, 2001; Shanahan et al., 2001).

Many advantages have been indicated for aerial digital imaging systems. However, one of the most important advantages is that the turn-around time is much shorter than film photography, since it does not require film development. Indeed, this characteristic is quite helpful for the development of real-time crop monitoring systems and in-season yield forecasting systems (Yang et al., 2000; Chen et al., 2000). Another advantage is that digital imaging systems can directly incorporate the images into computer systems without scanning the developed images, which often produced large errors due to the differences in scanner settings (Lamb and Brown, 2001; Plant et al., 2000; Yang et al., 2000; Chen et al., 2000).

Although aerial digital imaging systems are still costly for the agricultural purposes, recent developments in sensor technology and information systems should result in their eventual application. In particular, airborne hyperspectral scanners that provide information in tens to hundreds of spectral bands simultaneously, may offer new opportunities in the monitoring of crop and field conditions.

2.4.2 Model development based on remotely sensed data

Many methodologies have been developed to incorporate spectral information into yield estimation models. However, these approaches can be generally categorized into two types: (1) methods which directly correlate the spectral information to crop yield using regression models and vegetation indices; and (2) methods which estimate various crop parameters, such as LAI and biomass, from remotely sensed data, which are then used to calibrate the mechanistic crop growth models.

2.4.2.1 Regression models and vegetation indices

The simplest way to estimate crop yield from spectral information is to use linear regression models, and to correlate the radiance and/or reflectance of specific wavelengths with crop yield (Ball and Frazier, 1993; Tucker, 1979; Yang et al., 2000, 2001b). Since the absorption peaks of chlorophyll are in the red and green regions, and since the cuticle on leaf surface strongly reflects in the near-infrared region (Campbell, 2002), brightness values of these spectral regions are highly correlated with the vigor of crops or strength of photosynthetic activity, and therefore with the crop yields.

However, one problem of this method is that the generality of the models is extremely low, since radiation from canopy is a complicated function of various canopy characteristics such as leaf size, layout, and soil background (Campbell, 2002). Moreover, brightness values of some wavelengths are strongly influenced by many other environmental factors such as atmospheric absorption and light angle (Lillesand and Kiefer, 2000; Barrett and Curtis, 1999).

Introduction of vegetation indices (VI) helps to overcome these problems to some extent, since the ratio or difference of two or more wavebands is taken to calculate these values. However, the performance of the methods still depends on the environmental conditions existing at the measurement sites. Various factors, such as soil background effect and atmospheric disturbance, have been noted to be potential sources of noise (Barrett and Curtis, 1999; Huete, 1988; Rondeaux et al., 1996).

Many vegetation indices, such as the perpendicular vegetation index (PVI), the soil-adjusted vegetation index (SAVI), the transformed soil-adjusted vegetation index (TSAVI), and the atmospherically resistant vegetation index (ARVI) have been suggested to remove these various noise effects (Huete, 1988; Rondeaux et al., 1996; Shanahan et al., 2001; Wiegand et al., 1991). Functions of wavebands involving also those in the green region, such as the green normalized vegetation index (GNDVI), the green/NIR ratio, and the photochemical reflectance index (PRI) have been reported to be effective in the estimation of crop yields as well as in monitoring crop conditions (Aparicio et al., 2000; Gitelson et al., 1996; Shanahan et al., 2001; Strachan et al., 2002).

However, these indices have not been shown to be applicable to a wide range of crops and environmental conditions.

Some studies indicate that the cumulative NDVI or Simple Ratio (SR) index over a growing season, can improve the prediction accuracy, since grain yield is normally represented by the accumulated photosynthetic activity over the growing season (Hayes and Decker, 1996 and 1998; Masseli et al., 2000; Serrano et al., 2000; Wiegand et al., 1991). As a more advanced approach, Clevers (1997) reported that the integration of the fraction photosynthetic active radiation (fPAR), which is one of the indicators of the intensity of radiation available for photosynthesis, is useful for estimating crop yield. However, high sensitivity to environmental factors is still a serious problem for all of these approaches.

The equations of commonly used vegetation indices and references are summarized in Table 2.2.

2.4.2.2 Coupling remotely sensed data with crop growth modeling

Whereas the previously mentioned approach (use of regression models and VIs) directly correlates spectral information and crop yield on the basis of empirical models, integration of remotely sensed information and mechanistic crop growth is normally conducted in two steps: (1) estimation of some crop parameters (mostly LAI) from remotely sensed information, and (2) recalibration of the crop growth models with these estimated crop parameters. For the recalibration, within-season model calibration is dominant in current research (Moran et al., 1997).

Clevers et al. (1994) used LAI extracted from remotely-sensed images, to improve the performance of the SUCROS crop growth model. The LAI was estimated from the weighted difference vegetation index (WDVI) on the basis of two radiative transfer models, SAIL (Scattering by Arbitrarily Inclined Leaves) and the PROSPECT leaf optical properties model. This study was conducted on a sugar beet field in the Netherlands.

Bouman et al. (1999) also improved the performance of SUCROS crop growth model using remotely sensed data over sugar beet, potato, and winter wheat fields in the Netherlands. An interesting point in this study is that the ERS satellite Synthetic Aperture Radar (SAR) was used in model calibration.

Guérif and Duke (1998) coupled the SUCROS and SAIL models to estimate sugar beet yield in Northern France. In this study, new values of field parameters were also estimated from canopy reflectance by using inversion techniques. The results showed that the renewed field parameters helped improving model performance.

Matthew et al. (2000) recalibrated the CROPGRO-soybean model in three different soybean fields in Iowa, U.S. For the calibration, leaf weights estimated from the NVI (normalized vegetation index), were used. Aerial photography was used in this study.

In all of these studies, the performance of crop growth models was generally improved by incorporating remotely sensed data. For instance, Matthew et al. (2000) improved the prediction accuracies (correlation of determination, R^2) from 0.47 to 0.68, 0.39 to 0.57, and 0.04 to 0.22, at the three different experimental sites. Clevers et al. (1994) also decreased the prediction errors from 6.6 tons/ha (8.6%) to 3.0 tons/ha (4.1%). However, the applicability of the method to large-scale investigations is not yet clear due to the greater spatial variability (Clevers et al., 1994; Guérif and Duke, 1998; Moulin et al., 1998). Another problem is that a greater proportion of prediction error results from the process of LAI retrieval from remotely sensed images, rather during than the simulation of crop growth itself (Guérif and Duke, 1998, 2000).

2.4.3 Concluding remarks

Two different approaches have been taken to use spectral information to estimate crop yield. In the first approach, spectral information is directly correlated with crop yield by using regression models and VIs. The advantage of this approach is that the method is quite simple. However, the performance largely depends on the experimental conditions due to its high sensitivity to

environmental factors which are potential sources of noise (Barrett and Curtis, 1999; Huete, 1988; Rondeaux et al., 1996; Aparicio et al., 2000). Due to this high sensitivity, selection of the most appropriate VI or wavelengths for model development must always be based on past experience and heuristic trial and error, which often becomes time-consuming. It should be noted that model development should normally be conducted with large amounts of past data, including many years of yield data and images.

In the second approach, spectral information is used to estimate crop growth parameters (mainly LAI) to recalibrate mechanistic crop growth models. This approach is not only useful for theoretical studies, but it is also advantageous for localization or customization of models, since optimization of the model is clearer than the first approach. However, applicability of this approach to large-scale investigations has not been clearly demonstrated, especially when spatial variability of the field is high (Clevers et al., 1994; Guérif and Duke, 1998; Moulin et al., 1998). Complexity of the models is another disadvantage in cases when non-specialists operate the models, since the procedure becomes a black-box methodology (Varcoe, 1990). Finally, it should be noted that major improvements in model performance cannot be expected unless the LAI estimation is improved (Guérif and Duke, 1998, 2000).

2.5 Machine learning and image interpretation

Computer-based image interpretation is one of the most intensively explored topics in recent studies of remote sensing. Although the best method for the analysis and interpretation of remotely sensed images is still regarded to be the eyes of trained technicians (Mather, 2000), the overwhelming amount of data that can be acquired by the latest remote sensing systems, makes it difficult for the skilled technicians to analyze and interpret the images in a short time. According to Barrett and Curtis (1999), some of the earth resource satellites currently being built would generate more than 500 million data bits of information per second, and further increases are expected in the near future.

The use of machine learning algorithms is currently regarded to be a key issue in the development of computer-based image interpretation techniques. The flexibility of these algorithms and their ability to incorporate ancillary information into the image classification process with relatively simple operations has been noted (Mather, 2000; Atkinson and Tatnall, 1997). Artificial neural networks (ANNs) and decision trees (DT) have received particular attention in recent studies. The following two sections will present the basic concept and taxonomies of these two types of machine-learning tools, as well as recent applications of these methods in the remote sensing community.

2.5.1 Artificial neural networks

An Artificial Neural Network (ANN) is a computational model that mimics the human nervous system and decision-making process (Weisse and Kulikowski, 1991). Since the first introduction of the perceptron, a simple computational model of the human neuron, by McCulloch and Pitts in 1943 (Figure 2.2, Mair et al., 2000), various new architectures (connection patterns of perceptrons) and learning algorithms have been developed to achieve more flexible and complicated decision making systems (figure 2.3).

Although there are many ANNs architectures, they can be categorized into two main groups, feed-forward networks and recurrent networks, based on the connection patterns used (Jain et al., 1996). The feed-forward network structure is the most commonly used network architecture due to its simplicity and lower requirements in computing power and memory (Atkinson and Tatnall, 1997; Jain et al., 1996).

As shown in figure 2.3, one of the most important characteristics of the feed-forward network architecture is that the connection of the perceptrons is unidirectional, in contrast with the feed-back connections found in recurrent/feedback networks (Jain et al., 1996; Kime et al., 1998). Since no feed-back process exists, this architecture normally requires less memory for model development, but is less dynamic (Jain et al., 1996). While the most primitive feed-forward network architecture, single-layer perceptron with threshold function, can only be applied to

linearly discriminative data, the multi-layer perceptron with back-propagation algorithm and sigmoid activation function generally makes the decision boundary more complex and smooth (Figure 2.4, Jain et al., 1996). Indeed, the recent wide application of ANNs to image interpretation was made feasible by the advent of this architecture (Atkinson and Tatnall, 1997; Kimes et al., 1998). Successful applications of feed-forward networks can be found for classification of land uses (Paola and Schowengerdt, 1995; Kanellopoulos and Wilkinson, 1997; Murai and Omatu, 1997; Atkinson et al., 1997; Bernard et al., 1997) and clouds (Lee et al., 1990). In precision agriculture, they have been applied to the classification of crop type (Serpico et al., 1996) and nitrogen and weed stress detection (Goel et al., 2003).

The ANN was mainly thought of as a classification method in the past. However, recent studies show that the ANN has the potential to be developed into a prediction tool (Atkinson and Tatnall, 1997; Kimes et al., 1998), since it can describe the non-linear relationships between inputs and target attributes. Indeed, successful applications have already been reported for surface water quality assessment (Keiner et al., 1998; Gross et al., 1999; Zhang et al., 2002), soil moisture retrieval (Chang et al., 2000; Del Frate et al., 2003), biomass retrieval (Jin and Liu, 1997), yield prediction (Simpson, 1994), and chlorophyll estimation (Keiner et al., 1998).

One of the most important characteristics of ANNs is the ability to learn (Mather, 2000). For image interpretation, various intricate non-linear relationships between spectral information and target attributes can be analyzed without human intervention. ANNs do not normally require assumptions regarding sample distribution and data types, since they are based on a non-parametric approach to model development (Mather, 2000). Ancillary information from different sources can be incorporated into a single ANN model due to this characteristic (Simpson, 1994; Atkinson and Tatnall, 1997). However, some limitations have been indicated. Among them are: (i) an ANN model usually requires high computing power, especially for the training process (Mather, 2000); (ii) the developed model is difficult to interpret (Mair et al., 2000); (iii) it usually takes a long time to determine the optimal network structure in terms of number of hidden layers and PEs (Kimes et al., 1998); and (iv) compared with parametric methods, a larger numbers of samples is usually required for the training stage.

Recent advances in computer technology and the development of optimization techniques for network architecture (SPSS Inc., 2001; Jiang et al., 1994), have offered solutions to some of the limitations of ANNs. However, low interpretability of the models and the requirement of large numbers of training samples, are, unfortunately, difficult to solve.

The details of the feed-forward ANN algorithms used in this study are given in Chapter 3.

2.5.2 Decision tree estimation algorithms

Rule induction technique, generally known as decision tree, is another type of machine learning method, which is mainly used for medical and business purposes. Although the tree-representation of the human decision-making process (figure 2.5) is simple and old, one of the problems was the development of algorithms which effectively determine the optimal tree structure (Swain and Hauska, 1977). Over the years, various new algorithms have been suggested in the Artificial Intelligence (AI) community (Weiss and Kulikowski, 1991; Michalski et al., 1998; Friedl and Brodley, 1997).

As a basic taxonomy, decision tree algorithms can be categorized into two main groups, homogeneous and hybrid decision tree algorithms, and further, the homogeneous decision tree algorithms can be categorized into two subgroups, univariate and multivariate decision tree algorithms (Figure 2.6 and 2.7, Brodley and Utgoff, 1995; Friedl and Brodley, 1997; Zhou and Chen, 2002). However, the univariate decision tree classifier is currently the most commonly used due to its simplicity.

As shown in figures 2.6 and 2.7, one of the important characteristics of homogeneous decision tree algorithms (univariate and multivariate decision tree) is that a single algorithm is applied to all the nodes in the tree. This is different from the hybrid decision tree classifier, in which different algorithms can be applied to the nodes of a single tree (Friedl and Brodley, 1997). The main characteristic of univariate decision tree classifiers is that the decision boundary in each node is

determined by only one variable, while the multivariate decision tree algorithm can use more than one input variables (Figure 2.7, Brodley and Utgoff, 1995; Friedl and Brodley, 1997).

In general, the decision boundary of the univariate algorithm is not as smooth or flexible as that of multivariate or hybrid decision tree algorithms, since the boundary of univariate tree is always determined by the combinations of smaller rectangular decision boundaries (Figure 2.8, Brodley and Utgoff, 1995). However, the simplicity of the algorithms is often advantageous in that the algorithm generally requires less computational power and memory, and that the interpretability of the developed models is higher than those of the multivariate and hybrid classifiers. The Classification and Regression Tree (CRT), ID3, and C4.5 algorithms (SPSS Inc., 2001; Quinlan, 1993) are univariate algorithms.

As mentioned above, the main areas of application of decision trees have been medicine and business. However, recent studies have shown decision trees can be useful in interpretation of remotely sensed images. Indeed, successful applications have already been reported for land use classification (Defries et al., 1998 and 2000; Friedl and Brodley, 1997; Friedl et al., 1999; Hansen et al., 1996 and 2000; McIver and Friedl, 2002; Swain and Hauska, 1977) and vegetation survey (Simard et al., 2000). In the area of precision agriculture, Goel et al. (2003) tested the performance of the CRT algorithm in classifying spectral images of nitrogen and weed stress in corn plots. Yang et al. (2002, 2003) also applied the CRT algorithm to the classification of different tillage and residue management strategies, and to fertilizer application strategies.

Similar to other machine learning algorithms, one of the most important advantages of decision trees is flexibility. The non-parametric approach that may be taken is often advantageous compared with the conventional classification methods such as linear discriminants and clustering techniques, since it can be used in the classification of samples with non-Gaussian distributions (Hansen et al., 1996). As is the case with ANNs, decision trees also have the ability to incorporate ancillary information, which may also include non-numerical values. This is an important advantage when multiple data sources are used in model development. An advantage which is unique to the univariate decision tree, is that the interpretability of the developed model is high. Since the decision tree model is represented by a set of explicit tree-structured rules, it is easy to identify the

importance of each input variables, in contrast to such as the ANN (Mair et al., 2000; Hansen et al., 1996). The disadvantages of decision trees are: (i) inherent limitations with simple tree representation (Weiss and Kulikowski, 1991); (ii) no backtracking process after the tree is established (Mair et al., 2000; Weiss and Kulikowski, 1991), and (iii) requirement of large numbers of samples for training.

The details of the rule induction process for the univariate decision tree algorithm are summarized in Chapter 4.

2.5.3 *The Clementine Data Mining System*

The *Clementine Data Mining System* (SPSS Inc.) is a decision support system in which various machine learning algorithms, statistical techniques, and visualization techniques are integrated into one user-friendly interface. Although this data mining software system is most commonly used for medical and business purposes at the present time (Bose and Mahapatra, 2001; Liu Sheng et al., 2000), it may prove to be an effective tool when vast amount of spectral information obtained with latest sensors is to be analyzed.

Data mining systems are generally defined as interactive decision support systems, in which users and the systems constantly exchange the information (data) to find the best solution for a certain problem (SPSS Inc., 2001, Figure 2.9). In this meaning, data mining is not simply a methodology for model development or machine learning, but is regarded as an integrated software system for the purpose of data analysis. Indeed, data mining systems normally consist of various components, such as data loading (input node), record management (sorting, selecting and removing some specific records), graphical analysis, modeling and statistical analysis, and final presentation (output node).

One of the advantages of data mining systems is that more than one analytical procedure can be performed in only one pallet, owing to the integrity of the software. For instance, three different procedures, PCA, DT, and ANNs, can easily be combined into one modeling process without

performing each separately. Conventionally, several different software packages, such as spreadsheet, statistical software package, and machine learning software, must be employed to conduct the same analysis. Each time one analysis is finished, records must normally be rearranged for the next operation.

Another advantage of data mining systems is that each component of the system, in particular the machine learning algorithms and statistical procedures, is designed for non-expert users, whereas conventional machine learning and statistical software packages are rather designed for experts. For example, *Clementine Data Mining Systems* features many new algorithms to automatically find the optimum architectures of ANNs, such as number of hidden-layers and PEs (SPSS Inc., 2001). These procedures were conventionally conducted based on expert's experience, and regarded to be complicated and time-consuming for non-expert users.

In this study, all the modeling processes were basically conducted with the *Clementine Data Mining System*. However, it should be noted that some other statistical software packages and spreadsheet softwares were also used to verify the results and produce graphical analyses and presentations, since some of the graphical presentations and statistical analyses were not sufficiently supported by the *Clementine Data Mining Systems*.

2.6 Summary

Accurate prediction of crop yield before harvest is one of the greatest concerns in agriculture due to the economic importance and scientific interest. Conventionally, crop yield prediction was conducted without using spectral information. Crop growth was simulated with models incorporating physical environmental factors (agro-meteorological information, soil quality and management practice) and data from chlorophyll meters. However, these methods were quite labor-intensive and generally unrealistic for practical purposes.

Recent studies show that the use of remote sensing systems to arrive at yield estimates has a high potential since they are one of the most effective methods of data acquisition in agricultural fields.

In general, two different approaches have been taken to use remotely sensed images for crop yield estimation. The most common and simplest method is to use regression models and VIs, and directly associate the spectral information and crop yield. However, the high sensitivity of VIs to environmental factors is a serious constraint in the development of highly generalized models.

Recent studies also show that coupling mechanistic crop growth models with remote sensing systems is an effective alternative for yield estimation. However, the unavailability of field parameters often becomes a constraint for accurate prediction, especially when the spatial variability of the field is high. The complexity of the models is also a disadvantage when non-experts operate the models.

Machine learning algorithms are currently regarded to be key technologies for the development of effective image interpretation systems. In particular, artificial neural networks (ANNs) and decision trees (DTs) are being intensively explored as methods for classification and prediction in agricultural applications due to their flexibility and ability to incorporate a variety of types of ancillary information.

The ANNs and DTs may provide effective alternatives in the development of yield mapping and forecasting systems based on the remotely sensed information.

2.7 References

- Arnold, G. W., P. G. Ozanne, K. A. Galbraith, F. Dandridge. 1985. The capweed content of pastures in south-west Western Australia. *Australian journal of Experimental Agriculture* 25: 117-123
- Aparicio, N., D. Villegas, J. Casadesus, J. L. Araus and C. Royo. 2000. Spectral vegetation indices as nondestructive tools for determining durum wheat yield. *Agronomy Journal* 92 (1): 83-91.
- Aparicio, N., D. Villegas, J. L. Araus, J. Casadesus, and C. Royo. 2002. Relationship between growth traits and spectral vegetation indices in durum wheat. *Crop science* 42: 1547-1555
- Atkinson, P. M. and A. R. L. Tatnall. 1997. Neural networks in remote sensing. *International journal of remote sensing* 18(4): 699-709
- Atkinson, P. M., M. E. J. Cutler, and H. Lewis. 1997. Mapping sub-pixel proportional land cover with AVHRR imagery. *International journal of remote sensing* 18(4) 917-935
- Barrett E. C. and L. F. Curtis. 1999. *Introduction to environmental remote sensing*. Stanley Thornes Publishers Ltd., Cheltenham, UK
- Ball, S. T. and B. E. Frasier. 1993. Evaluating the association between wheat yield and remotely-sensed data. *Cereal research communications* 21(2-3): 213-219
- Bernard, A. C. G. G. Wilkinson, and I. Kanellopoulos. 1997. Training strategies for neural network soft classification of remotely-sensed imagery. *International journal of remote sensing* 18(8): 1851-1856
- Blackmer, T. M. and J. S. Schepers. 1995. Use of a chlorophyll meter to monitor nitrogen status and schedule fertigation for corn. *Journal of production agriculture* 8: 56-60
- Boissard, P., J. –G. Pointel, P. Huet. 1993. Reflectance, green leaf area index and ear hydric status of wheat from anthesis until maturity. *International journal of remote sensing* 14: 2173-2729
- Bose, I., and R. K. Mahapatra. 2001. Business data mining – a machine learning perspective. *Information management* 39: 211-225
- Bouman, B. A. M., D. W. G. van Kraalingen, W. Stol, and H. J. C. van Leeuwen. 1999. An agroecological modeling approach to explain ERS SAR radar backscatter of agricultural crop. *Remote sensing of environment* 67(2): 137-146
- Brisson, N., B. Marry, D. Ripoche, M. H. Jeuffory, F. Ruget, B. Nicoullaud, P. Gate, F. Devienne-Barret, R. Antonioletti, C. Durr, G. Richard, N. Beaudoin, S. Recous, X. Tayot, D. Plenet., P.

- Cellier, J. M. Machet, J. M. Meynard, and R. Delécolle. 1998. STICS: a generic model for the simulation of crops and their water and nitrogen balance. 1. Theory and parameterization applied to wheat and corn. *Agronomie* 18: 311-346
- Brodley, C. E., P. E. Utgoff. 1995. Multivariate decision tree. *Machine learning* 19: 45-77
- Brown, R. B., J. -P. G. A. Steckler, G. W. Anderson. 1994. Remote sensing for identification of weeds in no-till corn. *Transaction of the ASAE* 37(1): 297-302
- Campbell, J. B. 2002. *Introduction to remote sensing*. Guilford Press. New York.
- Chang, D.-H. and S. Islam. 2000. Estimation of soil physical properties using remote sensing and artificial neural network. *Remote sensing of environment* 74(3): 534-544
- Chen, F., D. E. Kissel, L. T. West, and W. Adkins. 2000. Field-scale mapping of surface soil organic carbon using remotely sensed imagery. *Soil science society of America journal* 64: 746-753
- Clevers, J. G. P. W., C. Büker, H. J. C. van Leeuwen, and B. A. M. Bouman. 1994. Framework for monitoring crop growth by combining directional and spectral remote sensing information. *Remote sensing of environment* 50(2): 161-170
- Clevers, J. G. P. W. 1997. A simplified approach for yield prediction of sugar beet based on optical remote sensing data. *Remote Sensing of Environment* 61(2): 221-228(1997)
- Costa, C., L. M. Dwyer, P. Dutilleul, D. W. Stewart, B. L. Ma, and D. L. Smith. 2001. Inter-relationships of applied nitrogen, SPAD, and yield of leafy and non-leafy maize genotypes. *Journal of plant nutrition* 24(8): 1173-1194
- Curran, P. J. 1985. Aerial photography for the assessment of crop condition: a review. *Applied Geography* 5: 347-360
- Daughtry, C. S. T., C. L. Walthall, M. S. Kim, E. Brown de Colstoun, and J. E. McMurtrey III. Estimating corn leaf chlorophyll concentration from leaf and canopy reflectance. *Remote sensing of environment* 74(2): 229-239
- De Fries, R. S., M. Hansen, J. R. G. Townshend, and R. Sohlberg 1998. Global land cover classifications at 8km spatial resolution: the use of training data derived from Landsat imagery in decision tree classifiers. *International journal of remote sensing* 19(16): 3141-3168

- De Fries, R. S. and J. C-W Chan. 2000. Multiple criteria for evaluating machine learning algorithms for land cover classification from satellite data. *Remote sensing of environment* 74(3): 503-515
- Del Frate, F., P. Ferrazzoli and G. Schiavon. 2003. Retrieving soil moisture and agricultural variables by microwave radiometry using neural networks. *Remote sensing of environment* 84(2): 174-183
- Drummond, S. T., K. A. Sudduth, A. Joshi, S. J. Birrell, N. R. Kitchen. 2003. Statistical and neural methods for site-specific yield prediction. *Transaction of the ASAE* 46(1): 5-14
- Friedl, M. A. and C. E. Brodley. 1997. Decision tree classification of land cover from remotely sensed data. *Remote sensing of environment* 61(3): 399-409
- Friedl, M. A., C. E. Brodley, and A. H. Strahler. 1999. Maximizing land cover classification accuracies produced by decision trees at continental to global scales. *IEEE transactions on geoscience and remote sensing* 37(2): 969-977
- Gitelson, A. A., Y. J. Kaufmann, and M. N. Merzlyak. 1996. Use of green channel in remote sensing of global vegetation from EOS-MODIS. *Remote sensing of environment* 58(3): 289-298
- Goel, P. K., S. O. Prasher, R. M. Patel, D. L. Smith, and A. DiTommaso. 2002a. Use of airborne multi-spectral imagery for weed detection in field crops. *Transaction of the ASAE* 45(2): 443-449
- Goel, P. K., S. O. Prasher, J. A. Landry, R. M. Patel, R. B. Bonnell, A. A. Viau, J. R. Miller 2002b. Potential of airborne hyperspectral remote sensing to detect nitrogen and weed infestation. *Computer and Electronics in agriculture* 38(2): 99-124
- Goel, P. K., S. O. Prasher, R. M. Patel, J. A. Landry, R. B. Bonnell, and A. A. Viau 2003. Classification of hyperspectral data by decision trees and artificial neural networks to identify weed stress and nitrogen status of corn. *Computers and Electronics in Agriculture* 39(2) 67-93
- Gross, L., S. Thiria and R. Frouin. 1999. Applying artificial neural network methodology to ocean color remote sensing. *Ecological modeling* 120 (2-3); 237-246
- Guérif M. and C. L. Duke 1998. Calibration of the SUCRO emergence and early growth module for sugar beet using optical remote sensing data assimilation. *European journal of agronomy* 9 (2-3):127-136

- Guérif M. and C. L. Duke 2000. Adjustment procedures of a crop model to the site specific characteristics of soil and crop using remotely sensing data assimilation. *Agriculture, ecosystems and environment* 81 (1): 57-69
- Hansen, M. Dubayah, R. and R. Defries. 1996. Classification trees: an alternative to traditional land cover classification. *International journal of remote sensing* 17(5): 1075-1081
- Hansen, M. C., R. S. Defries, J. R. G. Townshend, and R. Shohlberg. 2000. Global land cover classification at 1km spatial resolution using a classification tree approach. *International journal of remote sensing* 21(6): 1331-1364
- Hayes, M. J. and W. L. Decker. 1996. Using NOAA AVHRR data to estimate maize production in the United States Corn Belt. *International Journal of Remote Sensing* 17(16): 3189-3200
- Hayes, M. J. and W. L. Decker. 1998. Using satellite and real-time weather data to predict maize production. *International Journal of Biometeorology* 42(1): 10-15
- Huete, A. R. 1988. A soil-adjusted vegetation index (SAVI). *Remote sensing of environment* 25: 295-309
- Jain, A. K., J. Mao, and K. M. Mohiuddin. 1996. Artificial neural networks: A tutorial. *Computer* 29(3): 31-44
- Jiang, X., M. -S. Chen, M. T. Manry, M. S. Dawson, and A. K. Fung. 1994. Analysis and optimization of neural networks for remote sensing. *Remote sensing review* 9: 97-114
- Jin, Y.-Q., and C. Liu. 1997. Biomass retrieval from high-dimensional active/passive remote sensing data by using artificial neural networks. *International journal of remote sensing* 18(4): 971-979
- Kanellopoulos, I. and G. G. Wilkinson. 1997. Strategies and best practice for neural network image interpretation. *International journal of remote sensing* 18(4): 711-725
- Keiner L. E. and X. Yan. 1998. A neural network model for estimating sea surface chlorophyll and sediments from thematic mapper imagery. *Remote sensing of environment* 66(2): 153-165
- Kimes, D. S., R. F. Nelson, M. T. Manry and A. K. Fung. 1998. Attributes of neural networks for extracting continuous vegetation variables from optical and radar measurements. *International journal of remote sensing* 19(14): 2639-2663
- Kitchen, N. R., K. A. Sudduth, and S. T. Drummond. 1999. Soil electrical conductivity as a crop productivity measure for claypan soils. *Journal of production agriculture* 12(4): 607-617

- Kravchenko, A. N., and D. G. Bullock. 2000. Correlation of corn and soybean grain yield with topography and soil properties. *Agronomy journal* 92(1): 75-83
- Lamb, D. W., and M. Weedon. 1998. Evaluating the accuracy of mapping weeds in fallow fields using airborne digital imaging: *Panicum effusum* in oil seed rape stubble. *Weed research* 38: 443-451
- Lamb, D. W., M. N. Weedon, and L. J. Rew. 1999. Evaluating the accuracy of mapping weeds in seedling crops using airborne digital imaging: *Avena spp.* in seedling triticale. *Weed research* 39: 481-492
- Lamb, D. W. and R. B. Brown. 2001. Remote sensing and mapping of weeds in crops. *Journal of agricultural engineering research* 78(2): 117-125
- Lass, L. W., H. W. Carson, and R. H. Callihan. 1996. Detection of *Yellow Starthistle* (*Centaurea solstitialis*) and common St. Johnswort (*Hypericum perforatum*) with multispectral digital imagery. *Weed technology* 10: 466-474
- Lee, J., R. C. Weger, S. K. Sengupta, and R. M. Welch. 1990. A neural network approach to cloud classification. *IEEE Transactions on Geoscience and Remote Sensing* 28 (5) Page(s): 846 -855
- Lelong, Camille C. D., P. C. Pinet, and H. Poilvé. 1998. Hyperspectral imaging and stress mapping in agriculture: A case study on wheat in Beauce (France). *Remote sensing of environment* 66(2): 179-191
- Lillesand, T. M. and R. W. Kiefer. 2000. *Remote sensing and image interpretation*. John Wiley and Sons, Inc., New York
- Liu Sheng, O. R., C. –P. Wei, P. J. –H. Hu, N. Chang. 2000. Automated learning of patient image retrieval knowledge: neural networks versus inductive decision trees. *Decision support systems* 30: 105-124
- Mair, C., G. Kadoda, M. Lefley, K. Phalp, C. Schofield, M. Shepperd, and S. Webster. 2000. An investigation of machine learning based prediction systems. *The journal of systems and software* 53(1): 23-29
- Manjunath, K. R., M. B. Potdar, and N. L. Purohit. 2002. Large area operational wheat model development and validation based on spectral and meteorological data. *International journal of remote sensing* 23(15): 3023-3038

- Maselli, F., S. Romanelli, L. Bottai and G. Maracchi. 2000. Processing of CAC NDVI data for yield forecasting in the Sahelian region. *International Journal of Remote sensing* 21(18): 3509-3523.
- Mather, P. M. 2000. *Computer processing of remotely-sensed images: an introduction*. John Wiley & Son Ltd., Chichester, UK
- Mathew, S. S., J. O. Paz, and W. D. Batchelor. 2000. Integrating remotely sensed images to improve spatial crop model calibration. *ASAE paper No. 00-3039*. Milwaukee, Wisconsin: ASAE
- McIver, D. K. and M. A. Friedl. 2002. Using prior probabilities in decision-tree classification of remotely sensed data. *Remote sensing of environment* 81(2-3): 253-261
- Michalski, R. S., Bratko I., and M. Kubat. 1998. *Machine learning and data mining: Methods and applications*. John Wiley & Son Ltd., Chichester, West Sussex, England
- Moran, S. M., Y. Inoue and E. N. Barnes. 1997. Opportunities and limitations for image-based remote sensing in precision crop management. *Remote sensing of environment* 61(3): 319-346
- Moulin, S., A. Bondeau, and R. Delecolle. 1998. Combining agricultural crop models and satellite observations: from field to regional scales. *International journal of remote sensing* 19(6): 1021-1036
- Murai, H. and S. Omatu. 1997. Remote sensing image analysis using a neural network and knowledge-based processing. *International journal of remote sensing* 18(4): 811-828
- Paola, J. D. and R. A. Schowengerdt. 1995. A detailed comparison of backpropagation neural network and maximum-likelihood classifiers for urban land use classification. *IEEE transaction on geoscience and remote sensing* 33(4): 981-996
- Piekielek, W. P., R. H. Fox, J. D. Toth, and K. E. Macneal. 1995. Use of a chlorophyll meter at the early dent stage of corn to evaluate nitrogen sufficiency. *Agronomy Journal* 87: 403-408.
- Plant, R. E., D. S. Munk, B. R. Roberts, R. L. Vargas, D. W. Rains, R. L. Travis and R. B. Hutmacher. 2000. Relationship between remotely sensed reflectance data and cotton growth and yield. *Transaction of the ASAE* 43(3): 535-546
- Quinlan, R. J. 1993. *C4.5: Programs for Machine learning*. M. Kaufmann Publisher Inc., San Mateo, CA.
- Railyan, V. Y., and R. M. Korobov. 1993. Red edge structure of canopy reflectance spectra of triticale. *Remote sensing of environment* 46(2): 173-182

- Rastogi, A., N. Kalra, P. K. Agarwal, S. K. Sharma, R. C. Harit, R. R. Navalgund, and V. K. Dadhwal. 2000. Estimation of wheat leaf area index from IRS LISS-III data using Price model. *International journal of remote sensing* 21(15): 2943-2949
- Ritchie, J. T., and S. Otter. 1985. Description and performance of CERES-Wheat: A user-oriented wheat yield model. *USDA-ARS, ARS-38*, p159-175
- Ritchie, J. T., U. Singh, D. C. Godwin, and T. Hunt. 1989. *A User's guide to CERES maize-v2.10*. International fertilizer development centre, Muscle Shoals, AL.
- Rondeaux, G., M. Steven, F. Baret. 1996. Optimization of soil-adjusted vegetation indices. *Remote sensing of environment* 55(2): 95-107
- Serpico S. B., L. Bruzzone, and F. Roli. 1996. An experimental comparison of neural and statistical non-parametric algorithms for supervised classification of remote-sensing images. *Pattern recognition letters* 17(13): 1331-1341
- Serrano, L., I. Filella and J. Penuelas. 2000. Remote sensing of biomass and yield of winter wheat under different nitrogen supplies. *Crop Science* 40(3): 723-731.
- Shanahan, J. F., J. S. Schepers, D. D. Francis, G. E. Varvel, W. W. Wilhelm, J. M. Tringe, M. K. Schlemmer and D. J. Major. 2001. Use of remote-sensing imagery to estimate corn grain yield. *Agronomy journal* 93: 583-589
- Shibayama, M. and T. Akiyama. 1989. Seasonal visible, near-infrared and mid-infrared spectra of rice canopies in relation to LAI and above-ground dry phytomass. *Remote sensing of environment* 27: 119-127
- Shibayama, M. and T. Akiyama. 1991. Estimating grain yield of maturing rice canopies using high spectral resolution reflectance measurements. *Remote sensing of environment* 36: 45-53
- Shibayama, M., W. Takahashi, S. Morinaga, and T. Akiyama. 1993. Canopy water deficit detection in paddy rice using a high resolution field spectroradiometer. *Remote sensing of environment* 45(2): 117-126.
- Simard, M. Saatchi, S. S. and G. De Grandi. 2000. The use of decision tree and multiscale texture for classification of JERS-1 SAR data over tropical forest. *IEEE Transaction on Geoscience and remote sensing* 38(5): 2310-2321
- Simpson, G. 1994. Crop yield prediction using a CMAC neural network. In *Proceedings of the Society of Photo-Optical Instrumentation Engineers* 2315, 160-171. Bellingham, WA: The International Society for Optical Engineering.

- Smeal, D. and H. Zhang 1994. Chlorophyll meter evaluation for nitrogen management in corn. *Communications in soil science and plant analysis*. 25: 1495-1503
- Spitters, C. J. T., H. van Keulen, and D. W. G. van Kraalingen. 1989. A simple and universal crop growth simulator: SUCRO87. *In Simulation and Systems Management in Crop Protection*, edited by R. Rabbinge, S. A. Ward, and H. H. van Laar. Simulation Monographs 32 (Wageningen: Pudoc), p147-181
- SPSS Inc. 2001. *Clementine Version 6.0 User's Guide*. SPSS Inc., Chicago, IL
- Stafford, J. V. 2000. Implementing precision agriculture in the 21st century. *Journal of Agricultural Engineering Research* 76: 267-275
- Strachan, I. B., E. Pattey and J. B. Boisvert. 2002. Impact of nitrogen and environmental conditions on corn as detected by hyperspectral reflectance. *Remote sensing of environment* 80(2):213-224
- Swain, P. H. and H. Hauska. 1977. The decision tree classifier: design and potential. *IEEE Transaction on geoscience electronics* GE-15(3): 142-147
- Tucker, C. J. 1979. Red and photographic infrared linear combinations for monitoring vegetation. *Remote sensing of environment* 8: 127-150
- Unganai, L. and F. N. Kogan. 1998. Drought monitoring and corn yield estimation in southern Africa from AVHRR data. *Remote sensing of environment* 63(3): 219-232
- Varcoe, V. J. 1990. A note on the computer simulation of crop growth in agricultural land evaluation. *Soil use and management* 6(3):157-163
- Weiss, S. M. and C. A. Kulikowski. 1991. *Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems*. M. Kaufmann Publishers Inc., San Mateo, Calif.
- Wilkerson, G. G., J. W. Jones, K. J. Boote, K. T. Ingram, and J. W. Mishoe. 1983. Modeling soybean growth for crop management. *Transaction of the ASAE* 26: 63-73
- Wiegand, C. L., A. J. Richardson, D. E. Escobar, and A. H. Gerbermann. 1991. Vegetation indices in crop assessments. *Remote sensing of environment* 35: 105-119
- Yang, C., J. H. Everitt, J. M. Bradford and D. E. Escobar. 2000. Mapping grain sorghum growth and yield variations using airborne multispectral digital imagery. *Transaction of the ASAE* 43(6): 1927-1938.

- Yang, C., J. H. Everitt, J. M. Bradford. 2001a. Comparisons of uniform and variable rate nitrogen and phosphorous fertilizer application for grain sorghum. *Transaction of the ASAE* 44(2): 201-209
- Yang, C., J. M. Bradford and C. L. Wiegand. 2001b. Airborne multispectral imagery for mapping variable growing conditions and yields of cotton, grain sorghum, and corn. *Transaction of the ASAE* 44(6): 1983-1994
- Yang, C. –C., S. O. Prasher, J. Whalen, and P. K. Goel. 2002. Use of hyperspectral imagery for identification of different fertilization methods with decision-tree technology. *Biosystems Engineering* 83(3): 291-298
- Yang, C. –C., S. O. Prasher, P. Enright, C. Madramootoo, M. Burgess, P. K. Goel, and I. Callum. 2003. Application of decision tree technology for image classification using remote sensing data. *Agricultural systems* 76: 1101-1117
- Zhang, Y., J. Pulliainen, S. Koponen and M. Hallikainen. 2002. Application of an empirical neural network to surface water quality estimation in the Gulf of Finland using combined optical data and microwave data. *Remote sensing of environment* 81(2-3): 327-336
- Zhou, Z. -H., Z. -Q. Chen. 2002. Hybrid decision tree. *Knowledge-based systems* 15: 515-528
- Zwiggelaar, R. 1998. A review of spectral properties of plants and their potential use for crop/weed discrimination in row-crop. *Crop protection* 17(3): 189-206

Figure 2.1 Schematics of STICS mechanistic crop growth model (Source: Brisson et al., 1998).
All the symbols are summarized in Table 2.1.

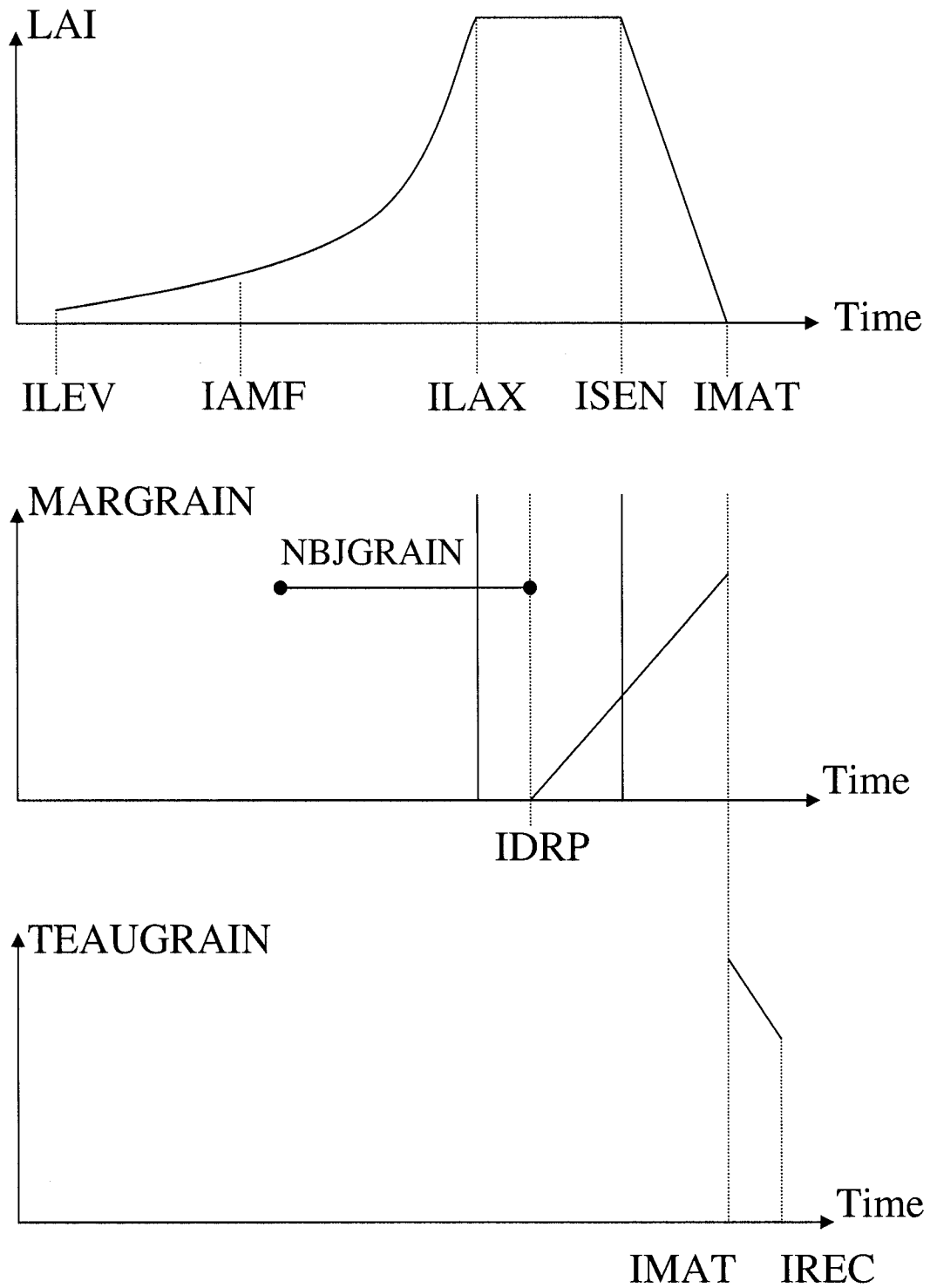


Figure 2.2 A McCulloch and Pitts model. A weighted sum of inputs is transferred by threshold function in perceptron, and Boolean values are returned as outputs. The threshold function can be replaced by several different functions, such as piecewise linear, sigmoid, and Gaussian functions to make more smooth decision boundary. Perceptrons can be connected to each other for developing various network architectures (Source: Mair et al., 2000)

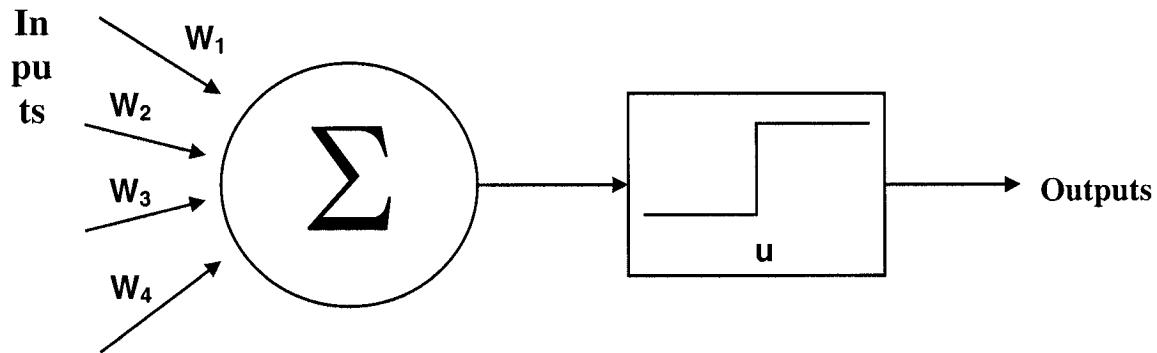


Figure 2.3 Basic taxonomy of artificial neural network architectures.
(Source: Jain et al., 1996)

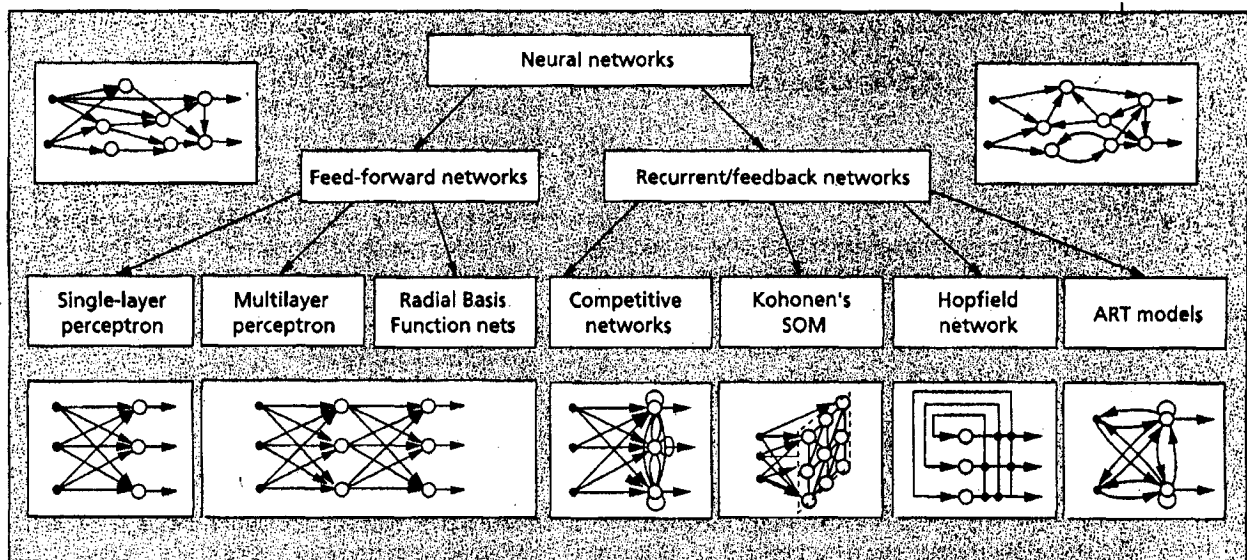


Figure 2.4 Relationships between decision boundaries and network structure in feed-forward networks. (Source: Jain et al., 1996)

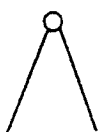
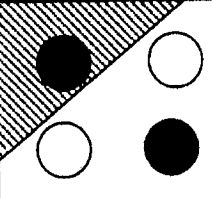
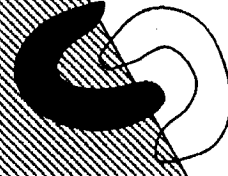
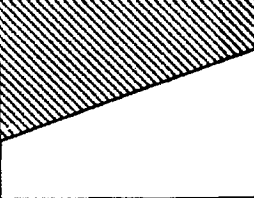
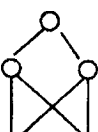
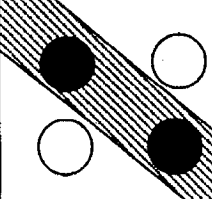
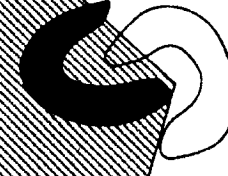
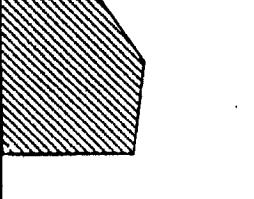

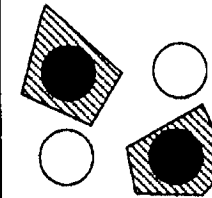
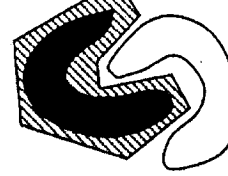
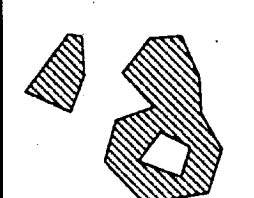
Structure	Description of decision regions	Exclusive-OR problem	Classes with meshed regions	General region shapes
 Single layer	Half plane bounded by hyperplane			
 Two layer	Arbitrary (complexity limited by number of hidden units)			
 Three layer	Arbitrary (complexity limited by number of hidden units)			

Figure 2.5 An example of tree representation for the human decision making process.

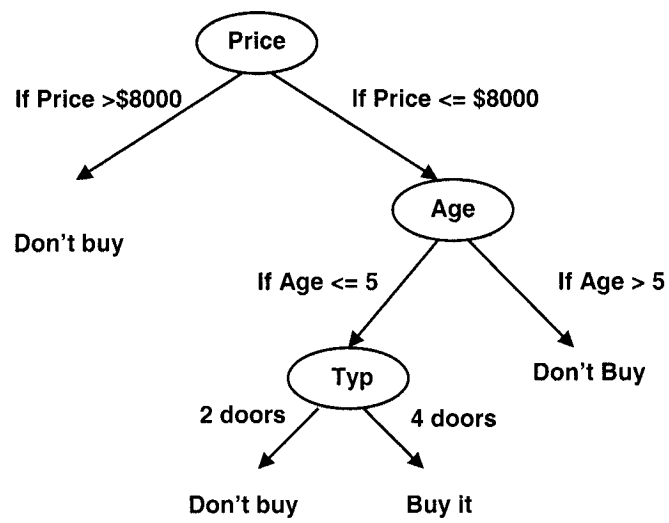


Figure 2.6 An example of a hybrid decision tree classifier. Different classifiers such as K-means, C5.0, maximum-likelihood classifier (MLC), linear discriminant function (LDF), can be incorporated into single tree.

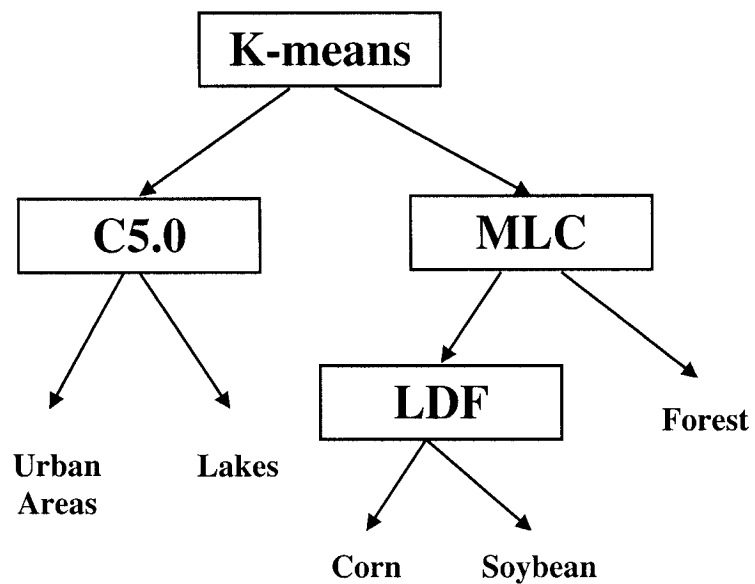


Figure 2.7 An example of univariate and multivariate decision tree classifiers. For the multivariate classifier, more than one variables can be used for each node.

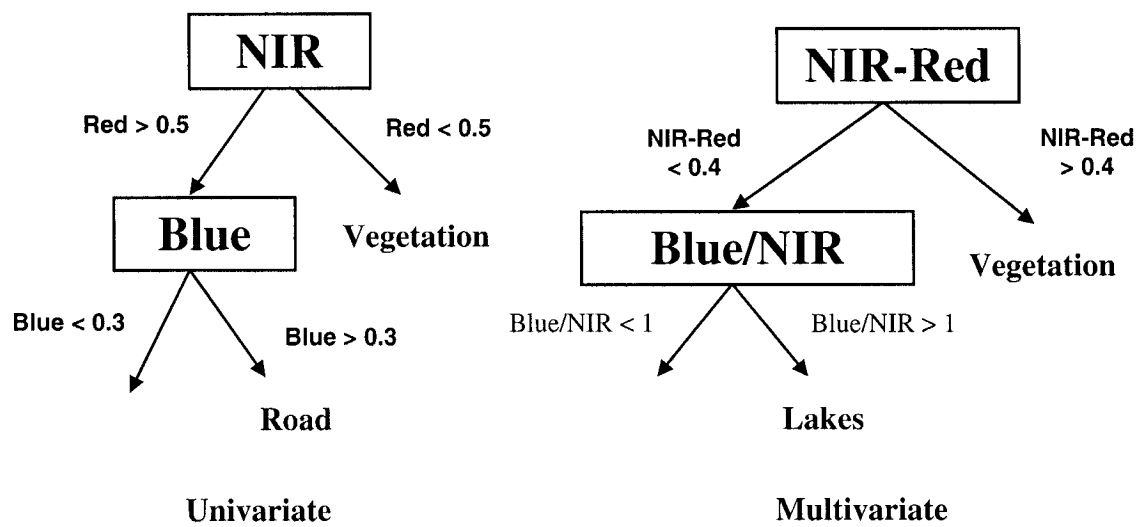


Figure 2.8 Geometric interpretation of univariate and multivariate decision tree classifiers. The decision boundary for multivariate classifiers in this example is made with:

If $X+Y < 8$, then \bigcirc

If $X+Y > 8$, then $+$

Source: Brodley and Utgoff, 1995)

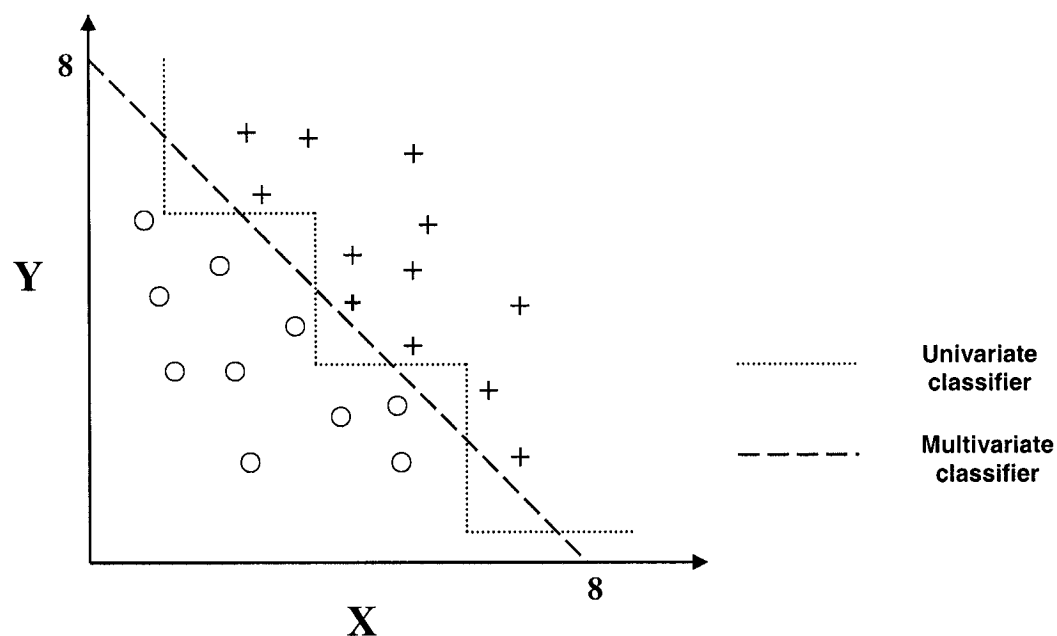


Figure 2.9 A schematic of data mining process.

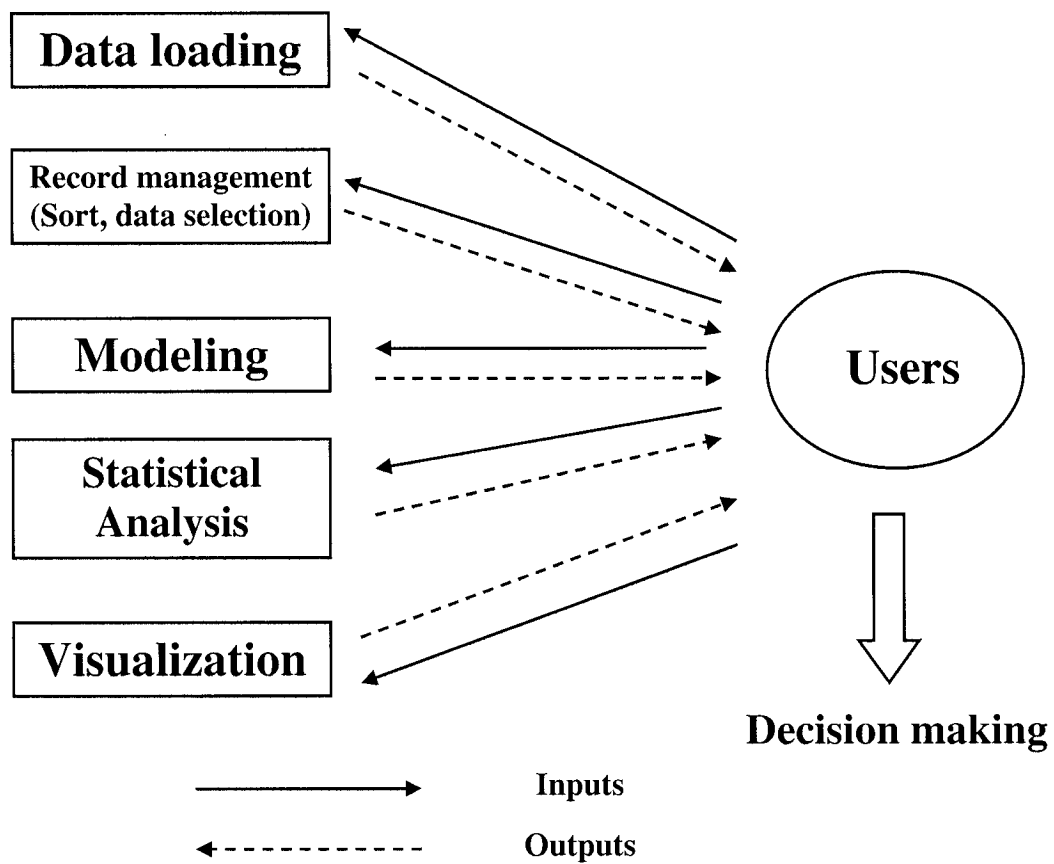


Table 2.1 Symbols used in the schematic (Figure 2.1) for STICS
(Source:Brissson et al., 1998).

Symbols	Description	Units
IAMF	Days of the stage AMF: maximal acceleration of leaf growth, end of juvenile phase	Days
IDRP:	Days of the stage DRP: beginning of grain filling	Days
ILAX	Days of the stage LAX: maximal leaf area index	Days
ILEV	Days of the stage LEV: emergence	Days
IMAT	Days of the stage MAT: Physiological maturity	Days
IREC	Days of the stage REC: harvest	Days
ISEN	Days of the stage SEN: beginning of net senescence	Days
LAI	Leaf Area Index	[m ² leaves m ⁻² soil]
MAGRAIN	Dry matter of grains	[gm ⁻²]
NBJGRAIN	Period when to compute NBGRAINS	[number of days before IDRP]
TEAUGRAIN	Water content of the grain	[g water g fresh grain]

Table 2.2 Equations of commonly used vegetation indices and their references.

VI _s	Equations	references
NDVI	$(NIR - RED)/(NIR + RED)$	
SR	NIR/RED	
PVI	$\sqrt{(Red_{soil} - Red)^2 + (IR_{soil} - IR_{veg})^2}$	Tucker (1979)
SAVI	$[(NIR - RED)/(NIR + RED + L)] \times (1 + L)$	Heute (1988)
TSAVI	$a(NIR - aRED - b)/[RED + a(NIR - b) + 0.08(1 + a^2)]$	Rondeaux et al. (1996)
ARVI	$(NIR - RB)/(NIR + RB)$ where $RB = RED - \gamma(BLUE - RED)$	Rondeaux et al. (1996)
GNDVI	$(NIR - GREEN)/(NIR + GREEN)$	Shanahan et al. (2001)
PRI	$(R_{570} - R_{531})/(R_{570} + R_{531})$	Strachan et al. (2002)
WDVI	$NIR - (C \times RED)$ where C=slope of the (soil-specific) soil line, or ratio between NIR and RED reflectance of soil	Clevers (1997)

Preface to Chapter 3

The literature review indicated that crop yield may be approximated from spectral information. However, the technical and economical limitations of the methods, currently in use, do not permit their application in precision agriculture.

Machine learning algorithms can be effective alternatives for the development of yield prediction models due to their flexibility and simplicity. In the next chapter (Chapter 3), the potential of Artificial Neural Networks (ANNs), one of the most commonly used machine learning methods, is explored, and compared with several conventional VI-based methods (NDVI, SR, and PRI) and Stepwise Multiple Regression (SMLR) models.

Successful application of the methodology could contribute to the development of in-season yield mapping or forecasting systems in precision agriculture.

The research paper based on this chapter

Y. Uno, S. O. Prasher, P. K. Goel, Y. Karimi, and A. A. Viau. Artificial neural networks to predict corn yield from Compact Airborne Spectrographic imager (CASI) data (Under preparation)

Chapter 3

ARTIFICIAL NEURAL NETWORKS TO PREDICT CORN YIELD FROM COMPACT AIRBORNE SPECTROGRAPHIC IMAGER (CASI) DATA

Y. Uno^a, S. O. Prasher^a, P. K. Goel^a, Y. Karimi^a, and A. Viau^b

^a Department of Bioresource Engineering, Macdonald Campus of McGill University, 21111 Lakeshore Rd., Ste-Anne-de-Bellevue, Quebec, Canada H9X 3V9, E-mail: shiv.prasher@mcgill.ca

^b Faculté de Foresterie et de Géomatique, Pavillion Louis-Jacques-Casault, Université Laval, Québec, Canada G1K 7P4

3.1 Abstract

In the light of recent advances in spectral imaging technology, highly flexible modeling methods must be developed to estimate various soil and crop parameters for precision farming from airborne hyperspectral imagery. The potential of artificial neural networks (ANNs) for the development of in-season yield mapping and forecasting systems was examined. Hyperspectral images of corn (*Zea mays* L.) plots in Eastern Canada, subjected to different fertilization rates and various weed management protocols, were acquired by a compact airborne spectral imager (CASI). Statistical and ANN approaches were used as well as various vegetation indices to develop yield prediction models. Principal component analysis (PCA) was used to reduce the number of input variables. Greater prediction accuracy (about 20% validation RMSE) was obtained with an ANN model than with either of the three conventional empirical models based on normalized difference vegetation index (NDVI), simple ratio (SR), or photochemical reflectance index (PRI). No clear difference was observed between ANNs and stepwise multiple linear regression model (SMLR). Although the high potential usefulness of ANNs was confirmed, particularly in the creation of yield maps, further investigations are needed before their application at the field scale can be

generalized.

Keyword. ANNs, Hyperspectral remote sensing, Precision agriculture, Crop yield, Corn, CASI.

3.2 Introduction

The creation of accurate yield maps is an essential component of the successful implementation of precision farming, as it offers useful information to variable rate technologies (VRTs; Stafford et al., 1996 and 2000; Reitz et al., 1996). Although the benefits obtained through yield mapping depend largely on the crop and environmental conditions (Earl et al., 1996; Swinton et al., 1998), its usefulness has been demonstrated with the recent commercialization of tractor-mounted crop yield-monitoring systems, now being used extensively by farmers to achieve uniform yields under highly variable field conditions.

Since remote sensing systems are capable of acquiring information over a large area within a very short period of time, these offer great advantages over tractor-mounted yield monitoring units in the creation of yield maps. More importantly, airborne digital imaging systems can provide real-time information on the condition of the crop and allow estimates of crop yield to be made long before the actual harvest. Consequently, such systems show great potential in assessing the impact of seasonally changeable factors (e.g., precipitation, temperature, and sunshine) in limiting crop growth (Yang et al., 2000 and 2001; Swinton et al., 1998). Moreover, introduction of hyperspectral sensors, capable of simultaneously gathering and recording spectral information in hundreds of wavebands, has the prospect to further revolutionize the application potential of remote sensing.

Over the years, a number of vegetation indices (VIs) have been developed by combining two or more wavebands in ratios and/or differences, to highlight various crop conditions. However, one of the problems in applying VIs to crop yield estimation is the difficulty in choosing the most appropriate vegetation index in a specific situation (Barrett and Curtis 1999; Osborne et al. 2002).

In fact, various environmental factors, such as background effects and crop canopy conditions, have been shown to be potential sources of noise, which affect the spectral reflectance in canopy level (Aparicio et al., 2000; Plant et al., 2000; Shanahan et al., 2001; Strachan et al., 2002). Ironically, these difficulties, to identify the most useful wavelengths or VIs under specific environmental conditions, have been heightened with the recent proliferation of large volume of data available from hyperspectral and broadband sensors. Sensitivity of vegetation indices and tapping the full potential of large quantities of spectral information acquired with the latest sensors are currently the most important impediments to successfully applying remote sensing technologies to precision farming.

Recent studies have shown that multivariate analytical techniques can prove quite useful in the interpretation of various forms of remotely-sensed data. Due to its great adaptability, stepwise multiple linear regression (SMLR) is one of the most commonly used methods to develop empirical models from large datasets, as has been done for a number of canopy-level crop condition parameters (Osborne et al., 2002; Shibayama et al., 1991). However, limitations to this technique exist as (i) it is based on the assumption that a linear relationship exists between input and target variables (Bethea et al., 1995), (ii) the assumption that samples follow a normal distribution (Bethea et al., 1995), and because (iii) in some cases model performance tends to be low because the method is extremely adaptable.

Machine learning algorithms, typically artificial neural networks (ANNs), have generated a strong interest in their potential effectiveness in estimating various field and crop conditions from remotely sensed images. The ability of ANNs to associate complicated spectral information with target attributes without any constraints for sample distribution (Mather, 2000), make them ideal for describing the intricate and complex non-linear relationships which exist between canopy-level spectral signatures and various crop conditions (Kimes et al., 1998; Lillesand and Keifer, 2000). Although in early studies ANNs were mostly used to classify data, the method has also shown a great potential for predicting continuous variables (Atkinson et al., 1997; Kimes et al., 1998). In fact, successful applications have already been reported for surface water quality assessment (Keiner et al., 1998; Gross et al., 1999; Zhang et al., 2002), soil moisture estimation

(Chang et al., 2000; Del Frate et al., 2003), biomass estimation (Jin and Liu, 1997), and yield prediction (Simpson, 1994).

This study sought to assess the potential of ANNs and hyperspectral aerial remote sensing for the development of field-scale yield estimation systems for corn. The performance of ANN models was evaluated using spectral values obtained from a Compact Airborne Spectral Imager (CASI) and compared with various conventional empirical methods, such as normalized difference vegetation index (NDVI), simple ratio (SR), photochemical reflectance index (PRI), and multiple linear regression (MLR) model. Advantages and disadvantages associated with ANNs models are discussed in the context of evaluating the feasibility of developing a yield mapping and forecasting system. Principal component analysis (PCA) was also introduced to avoid the risk of overfit, which is one of the most serious problems for ANN and SMLR model development.

3.3 Methodology

3.3.1 Experimental Design and Image Acquisition

The experiment was conducted at the Emile A. Lods Agronomy Research Center on the Macdonald Campus of McGill University, Sainte-Anne-de-Bellevue, Quebec, Canada. To simulate various crop growth scenarios, a corn (*Zea mays* L. cv. hybrid DK389BTY) crop was grown in forty-eight test plots (20 m x 20 m) under various weed management strategies and nitrogen fertilization rates. The two-factor experiment was laid out in split-plot design with three nitrogen fertilisation treatments (60, 120, 250 kg N/ha) and four weed control strategies, in quadruplicate. The weed treatments were: no weed control, control of grasses, control of broadleaf, and full weed control. Hyperspectral imagery was obtained with a Compact Airborne Spectrographic Imager (CASI) in 72 wavebands (spectral range 408 to 947 nm) at a spatial resolution of 2 m x 2 m. Images were acquired three times during the 2000 growing season: (i) at the early growth stage (30 days after planting, June 30), (ii) at the tasseling stage (66 days after planting, August 5), and (iii) at the fully mature stage (86 days after planting, Aug 25). However,

only the image obtained from the second flight (August 5) was used in this study, since this study was focused on the performance of machine learning techniques rather than on the physiological aspect of crop and crop canopy reflectance. An earlier study at this site (Goel et al., 2002, 2003) also had shown that the greatest correlation between spectral reflectance values and crop yield occurred at the tasseling stage. The raw radiance values, measured by the CASI sensor, were converted into spectral reflectance values through a series of pre-processing techniques, and images were also corrected for geometric distortions (Table 3.1).

Four different subplots (1 m x 1 m), representative of conditions prevailing in each treatment plot (20m x 20m), were selected as the sampling sites for grain yield measurements. Ten cobs were collected from each subplot, oven-dried at 70°C for 48 hours and, based on the crop density, grain yield was expressed in kg ha⁻¹.

3.3.2 Data Manipulation

Final corrected images were imported into the ENVI software (ENVI 3.1, Research System, Inc., Boulder, Colorado, USA) and reflectance values at four randomly selected points per treatment plot, corresponding with the yield sampling sites, were estimated. Spectral reflectance values were thus estimated from a total of 192 pixels. Reflectance values for wavelengths, ranging from 408 nm to 947 nm, were obtained from each pixel. These data were divided into two separate sets, one for calibration and another for the validation of models. Some 144 samples (75% of the samples) were randomly selected for calibration, and the remaining 48 samples (25% of the samples) were used for validation. For the performance analysis of ANN models with 71 input variables, a ten-fold cross validation procedure was also conducted to obtain more reliable results, since ANN are extremely sensitive to overfitting.

3.3.3 Principal component analysis (PCA)

One of the problems with ANN, and also with SMLR models, is that the methods are extremely adaptable. This means that spectral analysis by ANN must always take into account the potential problem of overfitting (SPSS Inc., 2001). The risk of overfitting arises when large numbers of independent variables are handled with a small number of samples. One of the solutions is testing the reliability or robustness of the developed models by using a validation dataset. However, another effective method is to reduce the number of input variables by removing the unnecessary or redundant information. This approach is more suitable for hyperspectral image analysis because it often contains large amounts of redundant information. In this study, principal component analysis (PCA) was used as a data reduction technique.

PCA is a data reduction, or data compression technique based on linear transformation. In this method, a new dataset (principal components) with k -variables is created from the original dataset with k -variables. Since these transformed variables (principal components) are ordered in terms of variance size, the number of variables can be reduced by removing the lower-level components without any remarkable loss of information (Ceballos and Bottino 1997; Manly 1994). A number of papers report the use of PCA for spectral analysis, including satellite-based remotely sensed images (Ceballos and Botino, 1997; Ricotta et al., 1999; Galvão et al., 2001) and hyperspectral aerial imagery (Blackburn and Milton, 1997).

In this study, factor scores, calculated from the top five principal components, were used as the input variables for two different modeling methods, ANNs and SMLR. PCA was carried out using the default option of the *Clementine Data Mining Systems* (SPSS Inc.), but this default mode does not conduct any factor rotation (SPSS Inc., 2001). The variance of each principal component was determined through the expert output options of the software. Results obtained with the reduced and non-reduced datasets were compared with each other to assess the performance of PCA.

3.3.4 Model development

Modeling processes were conducted in two main steps, data reduction and modeling (Figure 3.1). Three different vegetation indices, NDVI, SR and PRI, and four different combinations of data reduction and modeling methods were tested: (i) ANN model with 71 input variables (ANN-1), (ii) ANN model with 5 principal components (ANN-2), (iii) SMLR model with 71 input variables (SMLR-1), (iv) SMLR model with 5 principal components (SMLR-2) (Figure 1). All the models were developed using the *Clementine Data Mining Systems (SPSS Inc.)*, except for the regression analysis with the VIs, which was done with *Microsoft Excel (Microsoft Corp.)*.

3.3.5 Vegetation indices (VI)

Three commonly used vegetation indices, NDVI, SR, and PRI, were examined in this study, before ANN and SMLR models were developed.

$$NDVI = \frac{NIR_{900} - R_{680}}{NIR_{900} + R_{680}} \quad (1)$$

$$SR = \frac{NIR_{900}}{R_{680}} \quad (2)$$

$$PRI = \frac{G_{570} - G_{531}}{G_{570} + G_{531}} \quad (3)$$

where G_{λ} , NIR_{λ} , and R_{λ} are, respectively, the reflectance values in the green, infra-red, red, and yellow at the indicated wavelengths λ (nm). The spectral reflectance values at the required wavelength, centred at λ nm, were estimated by averaging spectral values from the two closest wavelengths obtained with CASI, assuming that a simple linear relationship existed between these two values.

3.3.6 Stepwise multiple linear regression (SMLR) models

As previously mentioned, SMLR is one of the most commonly used multivariate analytical techniques in remote sensing due to its flexibility. The process of SMLR consists of two main steps, variable selection and modeling. First, the importance of each input variable is evaluated by using a coefficient of determination (R^2), and then highly prioritized variables are added one-by-one to the multiple linear regression model. Each time one specific variable is added to the model, the significance of all the other variables is re-tested to evaluate their contribution to the model. If some variables are no longer significant at this stage, they are removed from the model. Normally, F-values are used to assess the significance of each variable (Bethea, 1995). The linear equation can be described as follows:

$$Yield = a_0 + a_1r_1 + a_2r_2 + a_3r_3 + + a_nr_n \quad (4)$$

where *Yield* is grain yield (kg ha^{-1}), $r_1, r_2, r_3...r_n$ are the spectral reflectance values at wavelengths 1 through n , and $a_1, a_2, a_3...a_n$ are regression coefficients.

Two different input strategies were taken for the development of SMLR models in this study. In the first strategy, all seventy-one spectral bands were directly incorporated into a SMLR model. In the second strategy, factor scores, acquired from five principal components, were used as the input variables. The stepwise criteria were of $P \leq 0.05$ for entry and $P > 0.10$ for removal.

3.3.7 Artificial neural networks

An artificial neural network (ANN) is a computational model which mimics the human nervous system and decision-making process (Jain et al., 1996). Although some technical difficulties, such as the low interpretability of the developed models (Mair et al., 2000), the complexity involved in optimizing the model structure (Mair et al., 2000), and the high processing power required for the training process, once made the intensive application of this techniques difficult, recent

improvements in computing power and learning algorithms has increased the applicability of the method in various fields. In fact, ANNs, given their great adaptability, are now regarded as an essential tool for image interpretation and development of prediction models from remotely sensed data.

Although ANN algorithms, implemented in *Clementine Data Mining System*, are based on the multi-layer feed-forward network architecture (Figure 3.2) with a back-propagation learning algorithm, and radial basis function networks, various new features were available to simplify and render operations more user-friendly. This software package's ability to automatically adjust the optimum number of processing elements (PEs) and network connections is one of its most useful characteristics. In the past, finding the optimum network structure has been one of the most time-consuming processes in the development of ANN models. Indeed, many ANN models developed in the past were based on previously obtained heuristic results, in which the optimum number of PEs had been determined. Although the *Clementine Data Mining System* offers four different strategies, quick, dynamic, multiple, and prune, to determine the optimal number of PEs, the “prune” option was adopted in this analysis since this option normally produces the highest performance while training time tends to be longer than other options (Integral solutions Ltd., 1998; SPSS Inc., 2001). In this option, a large network structure is constructed at the initial stage, and then unnecessary network connections are removed one-by-one to find the optimum network structure (SPSS Inc., 2001). The number of hidden-layers is fixed to one, unless the expert option is used (SPSS Inc., 2001). The “prevent overtraining” option was also used to avoid overtraining. With this option 50% of samples were randomly selected for training, and the remaining 50 % used for testing.

Two input strategies were employed: (i) using all seventy-one spectral bands as input variables, or (ii) using factor scores acquired from five principal components as input variables.

3.3.8 Performance analysis

Three statistical parameters were used for the performance analysis: the correlation coefficient (R), root mean square error (RMSE), and average difference (AVDIF). Correlation coefficients were calculated for three VI-based models and two SMLR models, solely for the calibration, since it seemed to be the most commonly reported parameter in previous reports of yield prediction (Aparicio et al., 2000; Osborne et al., 2002; Yang et al., 2000, 2001). However, this statistical parameter was not applied to the ANN models, as they are not based on the linear regression theory (Weiss and Kulikowski, 1991).

RMSE is one of the most commonly used statistical parameters, which represents the average difference between estimated and observed values. In this study, RMSE was calculated both on a percentage and Kg ha⁻¹ basis.

$$\text{RMSE } [\%] = (100 / \bar{O}) \sqrt{\frac{\sum (Pi - Oi)^2}{n}} \quad (5)$$

$$\text{RMSE } [\text{kg/ha}] = \sqrt{\frac{(Pi - Oi)^2}{n}} \quad (6)$$

where Pi is predicted yield, Oi is observed yield, \bar{O} is mean yield, and then i is the number of the yield estimate

AVDIF was used as it can be regarded as a better evaluation method for yield prediction at the farm level, whereas RMSE is a better estimator for the yield mapping. AVDIF was only calculated for validation. The values were presented with a percentage and kg/ha.

AVDIF is defined as the following equation:

$$\text{AVDIF} [\%] = (100/\bar{O}) \frac{\sum (P_i - O_i)}{n} \quad (7)$$

$$\text{AVDIF} [\text{kg/ha}] = \frac{\sum (P_i - O_i)}{n} \quad (8)$$

Visual analysis, plots of predicted *vs.* observed values, were also made to better understand model performance.

Considering the risk of overfit, a ten-fold cross validation procedure was conducted for the ANN model with 71 input variables (See 3.3.2). In this procedure, the original dataset (192 samples) was first randomly divided into 10 subgroups (i.e. Groups-A to J, 19 or 20 samples par group), and nine out of ten subgroups (i.e. Groups A-I) were selected for calibration, and the remaining subgroup (i.e. Group J) was kept for validation. In the next step, nine subgroups with different combinations (i.e. Groups-A to H and J) were selected from the original ten subgroups for calibration, and the remaining subgroup (i.e. Group I) was used for validation. After repeating the same calibration and validation processes with ten different combinations, the results (RMSEs and AVDIFs) obtained with these ten different validation datasets were summarized by calculating the mean value and 95% confidence interval (CI). It should be noted that the calibration and validation dataset were independent throughout this procedure. A graphical analysis was also conducted to better understand the model performance.

3.4 Results and Discussion

Different approaches were adopted to develop corn yield prediction models. Various statistical parameters, summarizing the performance of various yield prediction models for calibration and validation datasets, are presented in Table 3.2. The lowest RMSE for a calibration dataset was obtained with the SMLR-1 model (RMSE= 969.14 kg ha⁻¹, 16.68%), while the lowest RMSE for validation dataset was obtained with ANN-1 model (1092.54 kg ha⁻¹, 19.69%). Comparison

between observed and predicted yield for different models are also presented in Figures 3.3 and 3.4. These graphs clearly demonstrate that ANN and SMLR model performed better than all three VI-based models.

Although the difference between ANN-1 and SMLR-1 was generally small, graphical analysis (Figures 3.3 and 3.4) showed that the ANN-1 model produced larger prediction errors at high observed crop yield levels ($>7500 \text{ kg ha}^{-1}$) than did the SMLR-1 model (Figure 3.3). In fact, maximum estimated yield value (7368 kg ha^{-1}), generated with the ANN-1 model, was much lower than the highest observed yield (8664 kg ha^{-1}) in calibration, whereas this difference was much less for the SMLR-1 model (Figure 3.3). It is likely that the optimum network structure of the ANNs was slightly biased to a specific yield level ($O_i \approx 5000 \text{ kg ha}^{-1}$), because the number of training samples in this range was much larger than in other ranges ($O_i < 5000 \text{ kg ha}^{-1}$ and $O_i > 8000 \text{ kg ha}^{-1}$).

The results of 10-fold cross validation for the ANN-1 model are summarized in Figure 3.5 and Table 3.3. The mean values (19.11% for RMSE and -0.84% for AVDIF) and 95% confidence intervals of RMSEs and AVDIFs (16.94-21.28% for RMSE and -4.28-2.59% for AVDIF) obtained from 10 different validation datasets, showed that the prediction accuracies acquired with the validation dataset of 48 samples, which were already presented above (RMSE=19.69% and AVDIF=-1.01%), were quite reasonable.

Although the number of input variables (71 variables) used in this study generally seemed to be too large for ANN model, as compared to the number of training samples (144 records), these results showed that the risk of overfit was quite low in this particular case. This was probably the case because hyperspectral imagery normally includes large amounts of redundant information, in which most adjacent spectral bands are highly correlated to one another. It is also possible that the “pruning” option, which reduced input variables to 20 bands during the training (Table 3.4), might have contributed to developing a robust model. The basic structure of the developed ANN models, including selected wavelengths, is summarized in Table 3.4. For the ANN-2 model no input variables were removed as a result of pruning.

The equations of the two SMLR models are presented in Table 3.5. Although five bands were selected out of seventy-one input variables for the SMLR-1 model (Table 3.5), no clear trend was observed in wavelengths selected by the ANN-1 model (Table 3.4). The fifth principal component was rejected as a result of SMLR analysis for SMLR-2 model (Table 3.5), although all five principal components were used for ANN-2 model (Table 3.4).

Although the performance of VI-based models was generally lower than that of ANN models or SMLR models, a fairly high performance was obtained with the PRI (Table 3.2). In fact, the correlation coefficient obtained with PRI for the calibration dataset ($R=0.66$) was much higher than the results obtained with NDVI ($R=0.38$) or SR ($R=0.38$). Aparicio et al. (2000), on the other hand, found SR and NDVI to outperform PRI in the prediction of durum wheat yield in the Mediterranean region, which is not the case in this study.

One of the interesting points is that correlations obtained with NDVI and SR seemed to be quite low compared to previous works (Shanahan et al., 2001; Yang et al., 2001). However, these low correlations could possibly be the result of spectral reflectance values used to calculate the NDVI and SR being extracted from single pixel data, without standardization process, or by calculating mean reflectance values in the region of interest. In fact, a much higher correlation ($R=0.75$) had already been observed between NDVI and grain yield by Goel et al. (2003), using the same image. In their study, the mean yields of four subplots and mean spectral reflectance values of the corresponding area were computed for the model development. The strong noise from various environment factors, such as background soil effects and the existence of gaps in the canopy, could have produced these low prediction accuracies. Indeed, it is generally recognized that NDVI is quite sensitive to these kinds of environmental factors, so that various standardizing process are normally required to achieve the highest performance (Masseli et al., 2000). The positioning errors in global positioning system (GPS) could be another factor in this low performance. However, the effect of these positioning errors seemed to be limited, considering the results obtained with ANNs and some other methods, which showed much higher performance than NDVI- or SR-based models.

Past research has shown that some improvements could be expected by introducing curvilinear equations (Yang et al., 2001), taking multiple observations in the growing season (Plant et al., 2000; Serrano et al., 2000), and using geostatistical standardization to remove various environmental effects (Hayes and Decker, 1996, 1998; Masseli et al., 2000). However, it should be recognized that this kind of pre-processing normally requires a large amount of time and extra cost. It is also important that geostatistical standardization process not only reduces the number of samples available for model development, but it also decreases the spatial resolution of the images, which is essential for yield mapping. These negative factors could be crucial constraints, when application to precision agriculture is considered, since it often requires a high spatial resolution, and time critical action.

AVDIFs obtained for validation showed that the prediction errors at a farm level were quite low for all seven modeling strategies (Table 3.2). Indeed, all errors (maximum -3.42% by SMLR-1 model) generally appeared to be within acceptable levels in terms of agronomical importance. The results clearly indicate that the method used in this study could potentially be used to estimate crop yield at a field level using remotely sensed observations. However, it should be noted that some skepticism still exists with NDVI- and SR-based models due to the low correlations in calibration.

The performance of two PCA-based methods showed that any decrease in prediction accuracies, caused by PCA, seemed to be quite low (Table 3.2). In fact, the differences in RMSEs between two ANN models (2.26% for calibration and 2.37% for validation) and two SMLR models (1.50% for calibration and 0.70% for validation) generally seemed to be acceptable in terms of agronomical importance (Table 3.2). Considering the fact that the number of input variables was reduced from seventy-one spectral bands into only five principal components, the benefit of using PCA appeared to be quite large, especially in a situation where the number of samples for the modeling is limited. Graphical analysis also showed that the differences of the performance between reduced and original dataset were not large (Figures 3.6 and 3.7).

Eigenvalues of the top five principal components are summarized in Table 3.6. Based on these

statistics, 94% of the variance included in the 71 input variables, could be explained by these five principal components.

3.5 Conclusions

This study explored the potential of aerial hyperspectral spectral measurements to develop in-season field-scale yield prediction and mapping systems for corn. To simulate different crop growth scenarios, corn was grown under different weed management strategies and nitrogen fertilization rates. Statistical as well as ANN approaches were adopted to develop yield prediction models. PCA was also adopted to reduce the large amount of redundant information in hyperspectral imagery, and also tackle the problem of overfit. ANNs were quite efficient in capturing the complex relationship between crop yield and spectral reflectance values. Although the differences between ANN and SMLR models were not clear in this study, the higher performance of ANNs, compared to three VI-based methods, showed that ANNs could be effectively used to estimate the crop yield. While ANNs would be particularly useful in creating yield maps, no clear differences compared to other methods were observed for yield estimation at farm level. The usefulness of ANNs would be greater in a situation where selecting a VI has not been done before and could become a time-consuming process. However, it should be noted that further improvement is still required if this method is to be applied in a practical situation. In fact, the expected prediction errors of approximately 20% (in RMSE for validation) still seemed to be too large for the creation of yield map in precision agriculture, although the result of AVDIF (-1.01% for validation) seemed to be in acceptable level for the yield estimation at a farm level. Using a low-pass filter as a pre- and post-classification technique, incorporating ancillary data, and combining multiple data sources such as broadband sensors and radar images (Moran et al., 1997 and 2002) could be helpful on increasing the model performance.

A relatively high performance was observed with the PRI, compared to the NDVI or SR. However, the generality of this high performance for PRI was not clear, since no previous reports were found that PRI has the higher correlation with crop yield than NDVI and SR. More work is required to

identify the environmental and bio-physical factors which contribute to these differences.

This study also demonstrated that PCA was a useful data reduction technique for hyperspectral remote sensing. Although the performance of the models developed with reduced datasets were generally lower than that of models based on the full dataset, the benefits of using PCA were obvious, considering that the number of input variables was reduced to only five principal components.

3.6 References

- Aparicio, N., D. Villegas, J. Casadesus, J. L. Araus and C. Royo. 2000. Spectral vegetation indices as nondestructive tools for determining durum wheat yield. *Agronomy Journal* 92 (1): 83-91.
- Atkinson, P. M. and A. R. L. Tatnall. 1997. Neural networks in remote sensing. *International Journal of Remote Sensing* 18(4): 699-709
- Barrett E. C. and L. F. Curtis. 1999. *Introduction to environmental remote sensing*. Stanley Thornes Publishers Ltd., Cheltenham, UK
- Bethea, R. M. 1995. *Statistical methods for engineers and scientists*. M. Dekker Inc., New York
- Blackburn, G. A. and E. J. Milton. 1997. An ecological survey of deciduous woodlands using airborne remote sensing and geographical information system (GIS). *International Journal of Remote Sensing* 18(9):1919-1935
- Ceballos, J. C. and M. J. Bottino. 1997. Technical note: The discrimination of scenes by principal components analysis of multi-spectral imagery. *International Journal of Remote Sensing* 18(11): 2437-2449
- Chang, D.-H. and S. Islam. 2000. Estimation of soil physical properties using remote sensing and artificial neural network. *Remote Sensing of the Environment* 74(3): 534-544
- Earl, R., Wheeler P. N., Blackmore B. S. and R. J. Godwin. 1996. Precision farming – The management of variability. *Landwards* 51(4):18-23
- Del Frate, F., P. Ferrazzoli and G. Schiavon. 2003. Retrieving soil moisture and agricultural variables by microwave radiometry using neural networks. *Remote Sensing of the Environment* 84(2): 174-183
- Galvão, L. S. Pizarro M. A. and J. C. N. Epiphanio, 2001. Variations in reflectance of tropical soils: Spectral-chemical composition relationships from AVIRIS data. *Remote Sensing of the Environment* 75(2): 245-255
- Goel, P. K., S. O. Prahser, J. –A. Landry, R. M. Patel, R. B. Bonnell, A. A. Viau, and J. R. Miller. 2002. Potential of airborne hyperspectral remote sensing to detect nitrogen deficiency and weed infestation in corn. *Computers and Electronics in Agriculture* 38(2): 99-124
- Goel, P. K., S. O. Prahser, J. –A. Landry, R. M. Patel, and A. A. Viau. 2003. Estimation of crop

- biophysical parameters through airborne and field hyperspectral remote sensing. *Transaction of the ASAE* (in press)
- Gross, L., S. Thiria and R. Frouin. 1999. Applying artificial neural network methodology to ocean color remote sensing. *Ecological Modeling* 120 (2-3): 237-246
- Hayes, M. J. and W. L. Decker. 1996. Using NOAA AVHRR data to estimate maize production in the United States Corn Belt. *International Journal of Remote Sensing* 17(16): 3189-3200
- Hayes, M. J. and W. L. Decker. 1998. Using satellite and real-time weather data to predict maize production. *International Journal of Biometeorology* 42(1): 10-15
- Integral Solutions Ltd. 1998. *Clementine Reference Manual*. Hampshire, UK: Integral Solutions Limited.
- Jin, Y.-Q., and C. Liu. 1997. Biomass retrieval from high-dimensional active/passive remote sensing data by using artificial neural networks. *International Journal of Remote Sensing* 18(4): 971-979
- Jain, A. K., J. Mao, and K. M. Mohiuddin. 1996. Artificial neural networks: A tutorial. *Computer* 29(3): 31-44
- Keiner L. E. and X. Yan. 1998. A neural network model for estimating sea surface chlorophyll and sediments from thematic mapper imagery. *Remote Sensing of the Environment* 66(2): 153-165
- Kimes, D. S., R. F. Nelson, M. T. Manly and A. K. Fung. 1998. Attributes of neural networks for extracting continuous vegetation variables from optical and radar measurements. *International Journal of Remote Sensing* 19(14): 2639-2663
- Lillesand, T. M. and R. W. Kiefer. 2000. *Remote sensing and image interpretation*. John Wiley and Sons, Inc., New York
- Mair, C., G. Kadoda, M. Lefley, K. Phalp, C. Schofield, M. Shepperd, and S. Webster. 2000. An investigation of machine learning based prediction systems. *Journal of Systems and Software* 53(1): 23-29
- Manly, B. F. J. 1994. *Multivariate statistical methods: A primer*. Chapman & Hall, London, UK
- Maselli, F., S. Romanelli, L. Bottai and G. Maracchi. 2000. Processing of CAC NDVI data for yield forecasting in the Sahelian region. *International Journal of Remote Sensing* 21(18):

- 3509-3523.
- Mather, P. M. 2000. *Computer processing of remotely-sensed images: an introduction*. John Wiley & Son Ltd., Chichester, UK
- Moran, S. M. Hymer D. C. Qi J. and Y. Kerr. 2002. Comparison of ESR-2 SAR and Landsat TM imagery for monitoring agricultural crop and soil conditions. *Remote Sensing of the Environment* 79: 243-252
- Moran, S. M. Vidal A. Troufleau D. Qi J. Clarke T. R. Pinter Jr. P. J. Mitchell T.A. Inoue Y. and C. M. U. Neale 1997. Combining multifrequency microwave and optical data for crop management. *Remote Sensing of the Environment* 61: 96-109
- Osborne, S. L., J. S. Schepers, D. D. Francis and M. R. Schlemmer. 2002. Use of spectral radiance to estimate in-season biomass and grain yield in nitrogen- and water-stressed corn. *Crop Science* 42: 165-171
- Plant, R. E., D. S. Munk, B. R. Roberts, R. L. Vargas, D. W. Rains, R. L. Travis and R. B. Hutmacher. 2000. Relationship between remotely sensed reflectance data and cotton growth and yield. *Transaction of the ASAE* 43(3): 535-546
- Ricotta, C. and G. C. Avena. 1999. The influence of principal component analysis on the spatial structure of a multispectral dataset. *International Journal of Remote Sensing* 20(17): 3367-3376
- Reitz, P., and H.D. Kutzbach. 1996. Investigations on a particular yield mapping system for combine harvesters. *Computers and Electronics in Agriculture* 14(2-3): 137-150
- Serrano, L., I. Filella and J. Penelas. 2000. Remote sensing of biomass and yield of winter wheat under different nitrogen supplies. *Crop Science* 40(3): 723-731.
- Shanahan, J. F., J. S. Schepers, D. D. Francis, G. E. Varvel, W. W. Wilhelm, J. M. Tringe, M. K. Schlemmer and D. J. Major. 2001. Use of remote-sensing imagery to estimate corn grain yield. *Agronomy Journal* 93: 583-589
- Shibayama, M. and T. Akiyama. 1991. Estimating grain yield of maturing rice canopies using high spectral resolution reflectance measurements. *Remote Sensing of the Environment* 36: 45-53
- Simpson, G. 1994. Crop yield prediction using a CMAC neural network. In *Proceedings of the Society of Photo-Optical Instrumentation Engineers* 2315: 160-171.

- SPSS Inc. 2001. *Clementine Version 6.0 User's Guide*. SPSS Inc., Chicago, IL
- Stafford, J. V., B. Ambler, R. M. Lark and J. Catt. 1996. Mapping and interpreting the yield variation in cereal crops. *Computers and Electronics in Agriculture* 14(2-3): 101- 119
- Stafford, J. V. 2000. Implementing precision agriculture in the 21st century. *Journal of Agricultural Engineering Research* 76: 267-275
- Strachan, I. B., E. Pattey and J. B. Boisvert. 2002. Impact of nitrogen and environmental conditions on corn as detected by hyperspectral reflectance. *Remote Sensing of the Environment* 80(2):213-224
- Swinton, S. M. and J. Lowenberg- DeBoer. 1998. Evaluating the profitability of site-specific farming. *Journal of Production Agriculture*. 11(2):439-446
- Weiss, S. M. and C. A. Kulikowski. 1991. *Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems*. M. Kaufmann Publishers Inc., San Mateo, Calif.
- Yang, C., J. H. Everitt, J. M. Bradford and D. E. Escobar. 2000. Mapping grain sorghum growth and yield variations using airborne multispectral digital imagery. *Transactions of the ASAE* 43(6): 1927-1938.
- Yang, C., J. M. Bradford and C. L. Wiegand. 2001. Airborne multispectral imagery for mapping variable growing conditions and yields of cotton, grain sorghum, and corn. *Transactions of the ASAE* 44(6): 1983-1994
- Zhang, Y., J. Pulliainen, S. Koponen and M. Hallikainen. 2002. Application of an empirical neural network to surface water quality estimation in the Gulf of Finland using combined optical data and microwave data. *Remote Sensing of the Environment* 81(2-3): 327-336

Figure 3.1: The three major vegetation indices and four different combinations of data reduction and modeling techniques were tested in this study.

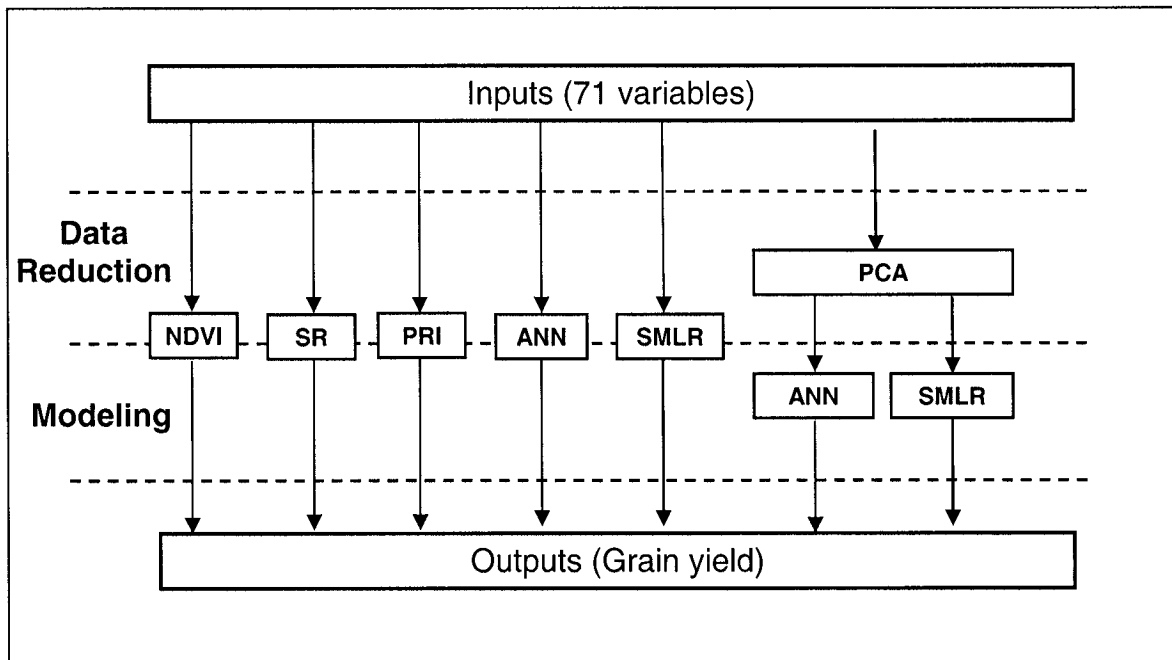


Figure 3.2 The network structure used in this study was based on multi-layer feed-forward networks with back-propagation learning algorithm. However, the optimum number of processing elements (PEs) and network connections was automatically determined by “prune” option of *Clementine data mining system*, in which network connections are reduced one-by-one from relatively large sized network. The number of hidden-layers is normally fixed with one in this option, unless expert option is used. The ovals represent PEs.

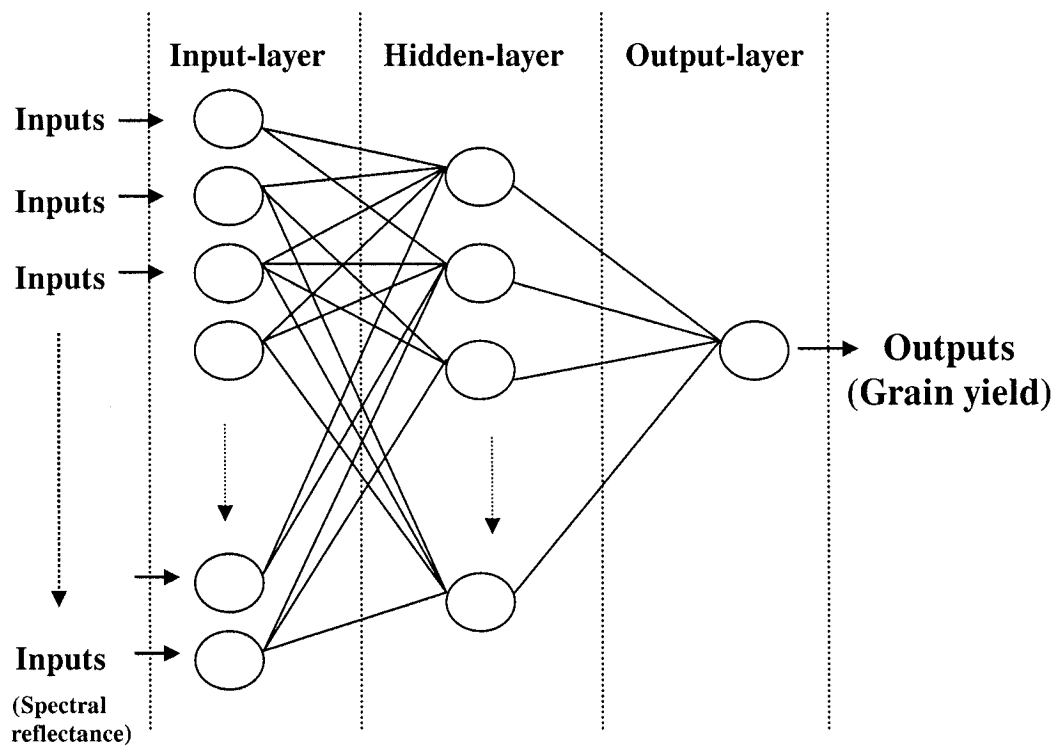


Figure 3.3 Performance of five different models for calibration dataset. (A) SMLR model with 71 input variables (B) ANN model with 71 input variables (C) NDVI (D) SR (E) PRI.

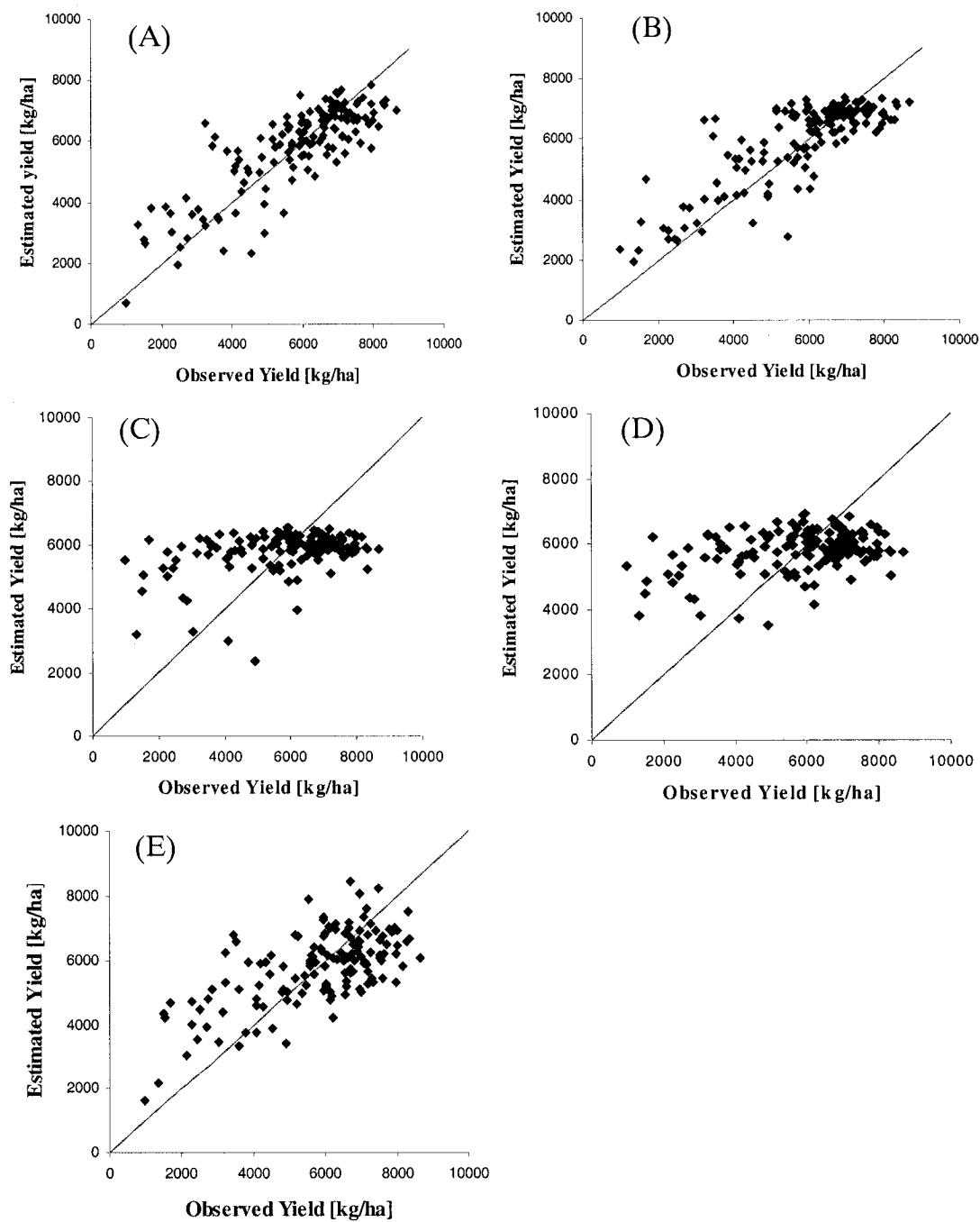


Figure 3.4 Performance of five different models for validation dataset. (A) SMLR model with 71 input variables (B) ANN model with 71 input variables (C) NDVI (D) SR (E) PRI.

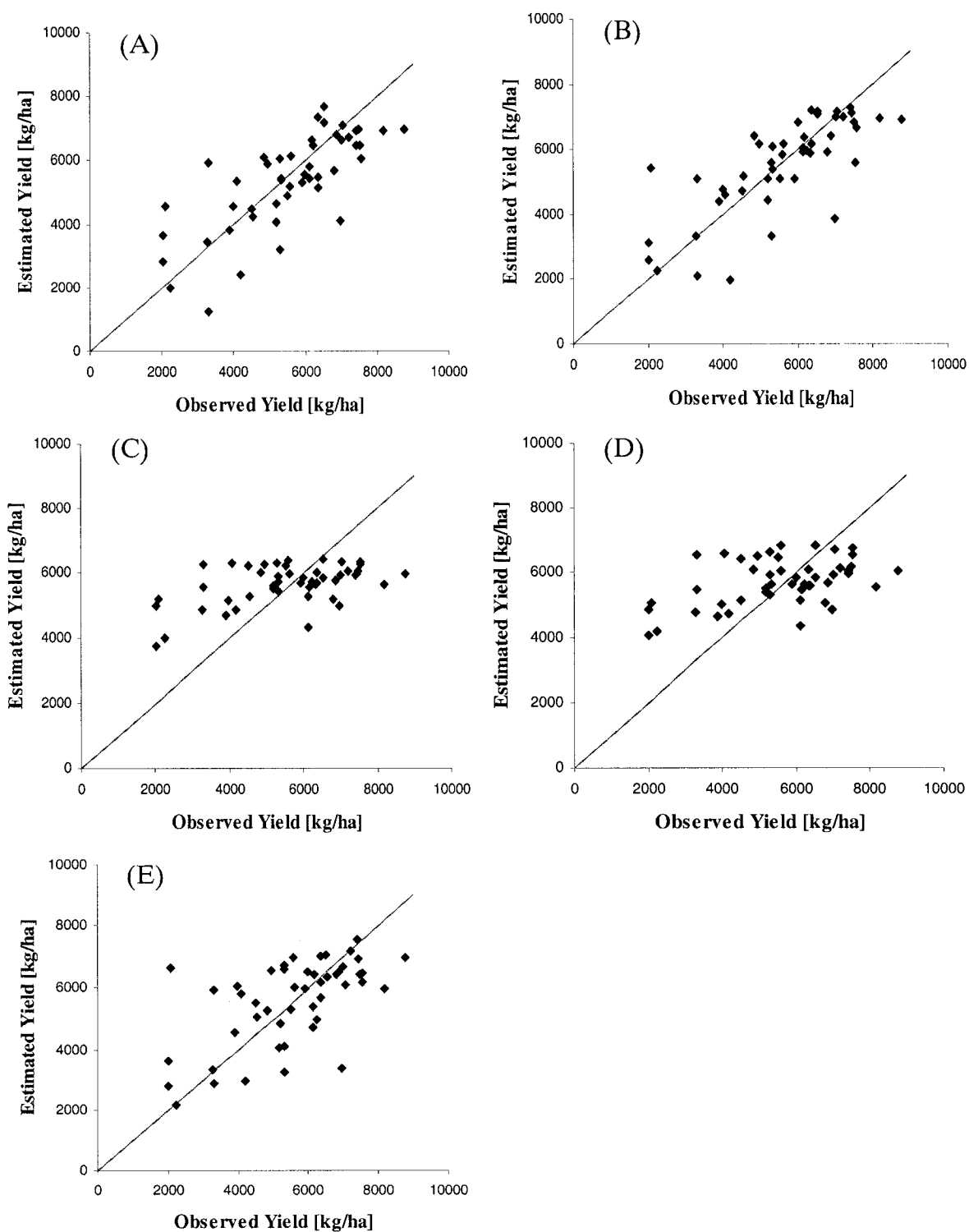


Figure 3.5 Results of ten-fold cross validation obtained with ANN model with 71 input variables.

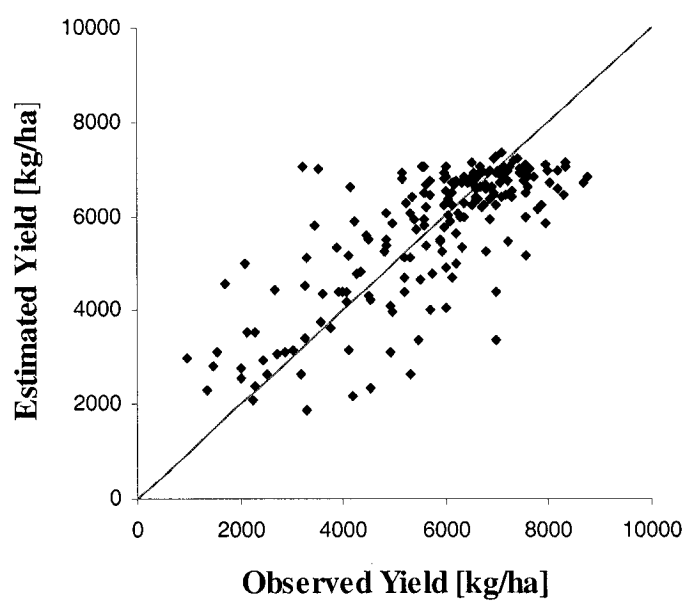


Figure 3.6 Difference of model performance between original dataset and reduced dataset. (A) ANN model with 71 input variables (B) ANN model with five principal components (C) SMLR model with 71 input variables (D) SMLR model with five principal components. All the figures were obtained with calibration dataset.

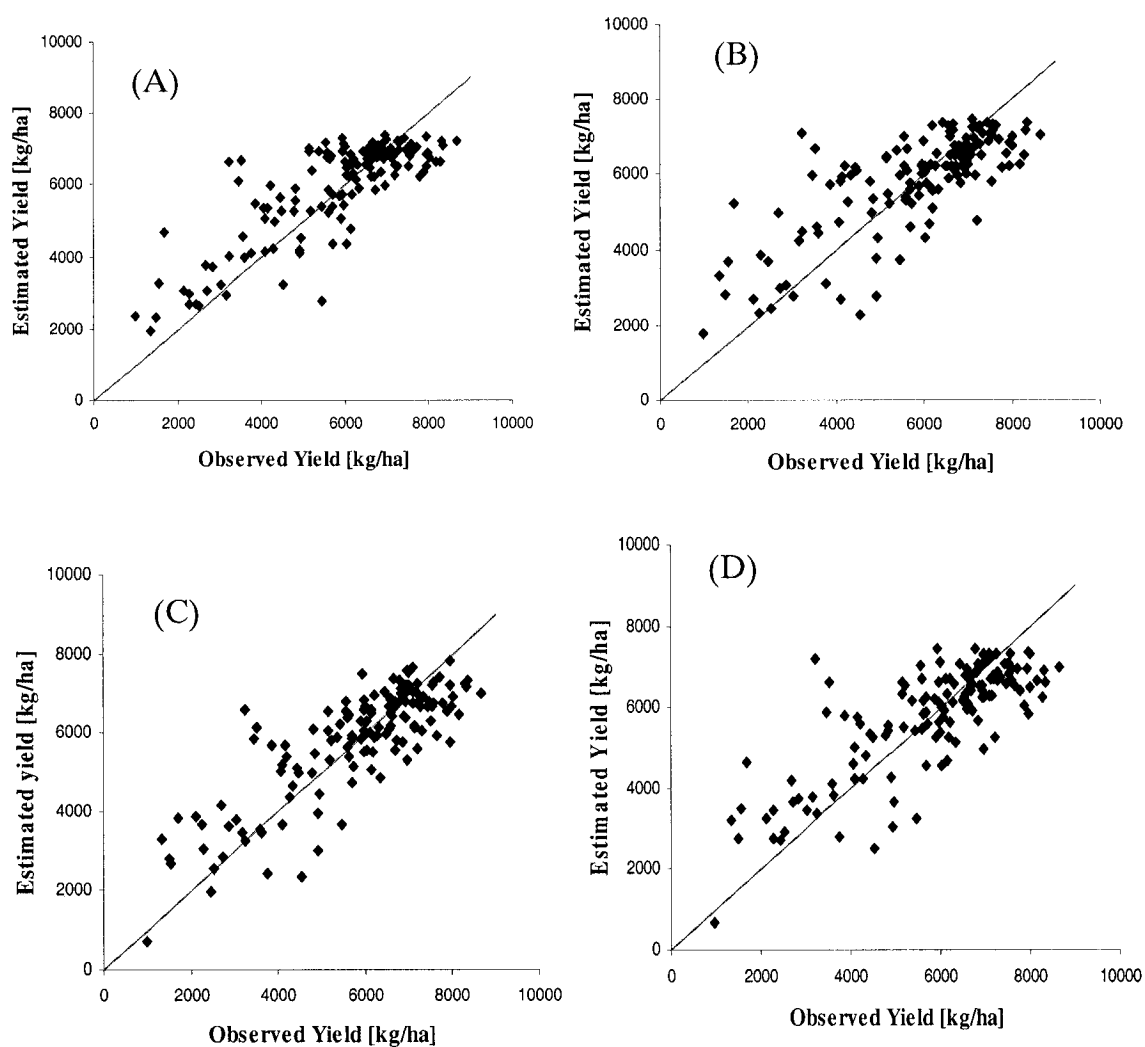


Figure 3.7 Difference of model performance between original dataset and reduced dataset. (A) ANN model with 71 input variables (B) ANN model with five principal components (C) SMLR model with 71 input variables (D) SMLR model with five principal components. All the figures were obtained with validation dataset.

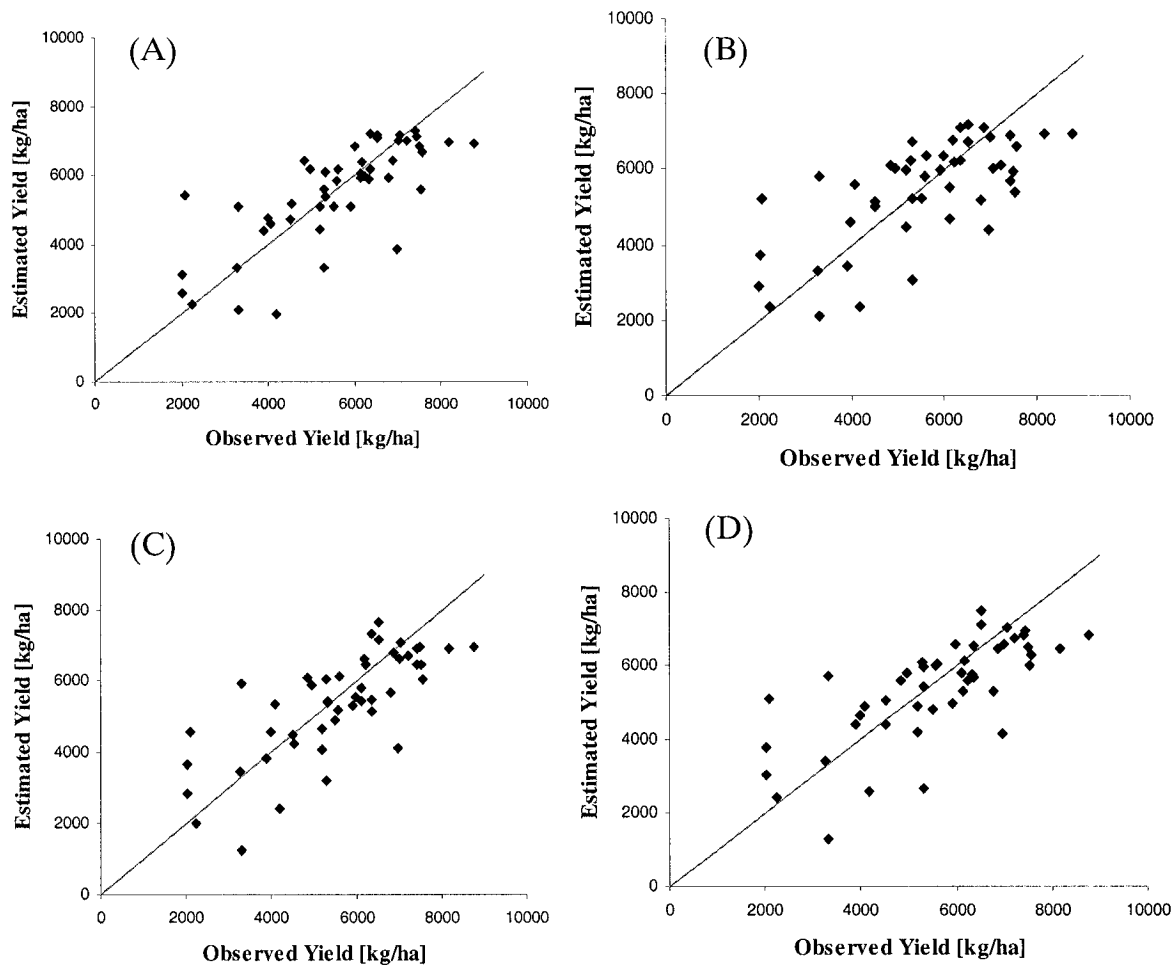


Table 3.1 CASI specification and data processing (Source: Goel et al., 2002)

Type of sensor	Pushbroom imager
Field of view	37.8°
Wavelength range	407 to 949 nm
Number of wavebands	72
Sampling rate	405 (spatial direction)
Spectral resolution	7.5 nm
Spatial resolution	2 m x 2 m
Noise floor	1.4 DN
S/N ratio	420:1 peak
a. June 30th, 2000 b. August 5th, 2000 c. August 25th, 2000	a. Heading: 150.732 North, Altitude above sea level: 1148 m, Time: 18:22, Cloud free b. Heading: 150.859 North, Altitude above sea level: 1130 m, Time: 15:30, Cloud free c. Heading: 331.225 North, Altitude above sea level: 1152 m, Time: 14:58, Cloud free
Data processing	
a. Radiometric and atmospheric corrections	Data collected from CASI were processed to at-sensor radiance using calibration coefficients determined in the laboratory by CRESTech (Center for Research in Earth and Space Technology). The CAM5S atmospheric correction model (O'Neill et al., 1997) was used to transform at-sensor radiance to ground-reflectance. Further, spectrally-flat uniform areas in each image (asphalt, bare soil and concrete surfaces) were used to do flat field adjustments in the spectral regions affected, residually by atmospheric absorption features for improved reflectance image data cubes.
b. Geometric corrections, geo-referencing, and image co-registration	Images were corrected for the aircraft movements (yaw, pitch, and roll) using GPS data onboard the aircraft, then rectified to UTM geographic coordinates. Further, white targets at the corners of the field were used for precise correction and error assessment. The estimated RMSE (root mean square error) was about 0.5 pixel.

Table 3.2 Prediction accuracies obtained with seven different modeling strategies. Three statistical parameters, correlation coefficient (R), root mean square error (RMSE), and average difference (AVDIF) were used. The numbers are presented on kilogram per hectare (kg/ha) and percentage.

Methods	Calibration			Validation			
	R	RMSE		RMSE		AVDIF	
		(kg/ha)	(%)	(kg/ha)	(%)	(kg/ha)	(%)
NDVI	0.38	1585.69	27.31	1430.57	25.77	94.76	1.71
SR	0.38	1589.94	27.38	1465.72	26.40	146.94	2.65
PRI	0.66	1287.03	22.16	1367.17	24.63	-9.09	-0.16
ANN-1	N.A.	981.48	16.90	1092.54	19.69	-56.32	-1.01
SMLR-1	0.83	969.14	16.68	1129.65	20.35	-189.67	-3.42
ANN-2	N.A.	1112.88	19.16	1224.70	22.06	-96.81	-1.74
SMLR-2	0.79	1055.66	18.18	1168.60	21.05	-183.87	-3.31

Table 3.3 Results of ten-fold cross validation for the ANN model with 71 input variables. Average value and 95% confidence interval (CI) of RMSEs and AVDIFs were calculated from ten different validation datasets.

	RMSE	AVDIF
Average	1097.55 kg/ha (19.11%)	-48.40 kg/ha (-0.84%)
95% CI	972.99 ~ 1222.11 kg/ha (16.94 ~ 21.28%)	245.57 ~ 148.77 kg/ha (-4.28% ~ 2.59%)

Table 3.4 Structure of the developed ANNs. The selected wavelengths were ordered with the relative importance of the each band for the output variables (Yield). This order was observed with “sensitivity analysis” option of Clementine data mining system.

Input methods	Number of PEs			Selected inputs
	Input Layer	Hidden Layer #1	Output Layer	
71 inputs	20	2	1	708.97, 701.36, 572.82, 762.35, 716.58, 655.83, 430.95, 739.45, 565.35, 475.53, 550.29, 423.53, 535.29, 557.79, 460.65, 724.20, 595.39, 633.13, 670.99, 678.57 (nm)
5 PCs	5	2	1	PC-1, PC-2, PC-5, PC-3, PC-4

Table 3.5 Equations of the developed SMLR models and VI-based linear models. $R(\alpha \text{ nm})$ shows reflectance value at $\alpha \text{ nm}$. (PC n) shows the principal component of n th level.

Input methods	Equation
71 inputs	Yield [kg/ha] = $1371.443688 \cdot R(482.98\text{nm}) - 4249.325871 \cdot R(602.93\text{nm}) + 3759.589996 \cdot R(655.83\text{nm}) + 305.581612 \cdot R(900.66\text{nm}) - 209.21606 \cdot R(931.58\text{nm}) + 3816.357964$
5 PCs	Yield [kg/ha] = $400.118518 \cdot (\text{PC}4) + 595.421 \cdot (\text{PC}3) - 440.630547 \cdot (\text{PC}2) - 1088.118413 \cdot (\text{PC}1) + 5696.886788$
NDVI	Yield [kg/ha] = $31278 \cdot \text{NDVI} - 22340$
SR	Yield [kg/ha] = $206.46 \cdot \text{SR} + 1759.5$
PRI	Yield [kg/ha] = $-108646 \cdot \text{PRI} - 5812.5$

Table 3.6 Eigenvalues of five principal components (PC) used for the model development. The variance of each PC and cumulative variance from the top were calculated.

<i>Component</i>	<i>Total</i>	<i>Variance (%)</i>	<i>Cumulative (%)</i>
1	32.276	45.459	45.459
2	23.978	33.772	79.231
3	7.926	11.164	90.395
4	1.880	2.648	93.043
5	.689	.970	94.013

Preface to Chapter 4

Although the previous analysis (Chapter 3) demonstrated that ANN modeling may be an effective tool to estimate crop yield, it is one of the several machine learning algorithms that can be used for this purpose. Over the years, many different concepts have been suggested to develop effective machine learning algorithms in the Artificial Intelligence (AI) community, and most of these methods can potentially be applied to image interpretation, and consequently for the development of yield prediction models from remotely sensed images.

Decision tree (DT) estimation algorithm is one of these machine learning methods, which is most commonly used in business and medical applications at the present time. Although some research work has already been done to evaluate the applicability of this algorithm for image interpretation, more efforts need to be made in the area of precision agriculture.

In the next chapter (Chapter 4), the possibility of using decision trees for the creation of field-scale yield maps from hyperspectral imagery is explored. The performance of DTs at two tasks is evaluated: (1) DT as an image classification tool, or for the classification of crop productivity, and (2) DT as a feature band selection tool.

The research paper based on this chapter

Y. Uno, S. O. Prasher, P. K. Goel, Y. Karimi, and A. A. Viau. Use of classification tree and Compact Airborne Spectrographic Imager (CASI) for corn yield estimation (Under preparation)

Chapter 4

USE OF CLASSIFICATION TREE AND COMPACT AIRBORNE SPECTROGRAPHIC IMAGER (CASI) FOR CORN YIELD ESTIMATION

Y. Uno^a, S. O. Prasher^a, P. K. Goel^a, Y. Karimi^a, and A. A. Viau^b

^a Department of Agricultural and Biosystems Engineering, Macdonald Campus of McGill University
Ste-Anne-de-Bellevue, Quebec, Canada H9X 3V9, E-mail: shiv.prasher@mcgill.ca

^b Faculté de Foresterie et de Géomatique, Pavillion Louis-Jacques-Casault, Université Laval, Québec, Canada G1K 7P4

4.1 Abstract

The creation of yield maps using remotely sensed images is currently one of the challenges in the development of precision crop management. This study evaluates the potential of a decision tree estimation algorithm to classify hyperspectral images of a corn (*Zea mays* L.) field, acquired from an airborne spectral imager (CASI), into yield categories. The images were acquired over corn experimental plots in eastern Canada, where crops were grown in different nitrogen application rates and weed control strategies. The results showed that the performance of the algorithm in terms of overall classification accuracies was comparable to a conventional classification method, but still seemed to be low for practical purposes. The results also demonstrated that the potential of the algorithm as a feature band selection tool was high. However, further investigation is still necessary to explore the reliability and stability of the developed models.

Keyword: Yield classification, Corn, Remote sensing, Decision tree, Feature band selection

4.2 Introduction

Describing the within-field variability of crop yield is one of the most important issues in precision agriculture, since it offers a variety of useful information for the recently developed variable rate technologies (VRT) (Reitz et al., 1996; Stafford, 2000). Recent research has shown that the information provided by airborne digital imaging systems can be used to create yield maps based on interpretation of crop conditions in real-time (Yang et al., 2001).

Machine learning algorithms are currently regarded as one of the keys to the successful application of remote sensing in precision agriculture, since the complicated and time-consuming process of the image interpretation can be done automated (Mather, 2000). Indeed, application of such algorithms can significantly reduce the time required for analyzing the spectral information and developing models, as well as reduce the time required to train skilled technicians. Complicated spectral information, which used to be difficult to analyze, even by highly skilled technicians, can now be processed with relatively simple operations. The use of machine learning algorithms for image interpretation has recently been increasing in order to handle the much larger data sets supplied by the latest hyperspectral and broadband sensors (Goel et al., 2003a).

The decision tree (DT) estimation algorithm, one of the most commonly used machine learning algorithms along with artificial neural networks (ANNs), can be an effective alternative for classifying the remotely sensed images to be used in precision agriculture. One of the advantages is that this classification tool does not require normally distributed data, contrary to most conventional classifiers, such as linear discriminant and maximum likelihood classifiers (Friedl et al., 1997; Hansen et al., 1996). Moreover, the expression of the induced explicit rules in the form of a classification tree is often helpful in clarifying the model structure and the classification process. This is perceived to be an advantage over the “black box” situation of another machine learning method, such as ANN (Mair et al., 2000; Debuse and Rayward-Smith, 1997). Indeed, past research have shown that the clarity of decision tree models is helpful in identifying the importance of input variables on target attributes, and consequently in reducing the dimensionality of remotely sensed data, or feature band selection (Hansen et al., 1996; De Fries et al., 1998; Simard et al., 2000)

Although there are inherent limitations in attempting to represent the human decision making process by simple tree structures (Weiss and Kulikowski, 1991), and due to the absence of backtracking processes once the tree is established (Mair et al., 2000; Weiss and Kulikowski, 1991), recent research has shown that the DT is one of the effective classification methods for remotely sensed images. Many successful applications have been reported for land use classification using various satellite images (Hansen et al., 1996; Friedl and Brodley, 1997; Defries and Chan, 2000; Simard et al., 2000; Rogan et al., 2002). In precision agriculture, Goel et al. (2003a) tested the performance of a classification and regression tree (CRT) in classifying plots cropped with corn into categories representing the nitrogen and weed stresses that were set up in a split-plot experiment. Yang et al. (2002a and 2003) also used CRT algorithm to classify different tillage practices and residue management strategies, and fertilizer application strategies.

The goal of this study was to assess the potential of the decision tree classifier for the development of a field-scale yield mapping system based on hyperspectral images of a corn field, acquired with a compact airborne spectrographic imager (CASI). The potential of the decision tree classifier to reduce the dimensionality of the hyperspectral dataset was also explored by analyzing the structure of the developed decision tree model.

4.3 Methodology

4.3.1 Image acquisition and data preparation

Spectral information for the analysis was obtained from a CASI sensor. Images were taken over corn (*Zea mays* L. cv. Hybrid DK398BTY) experimental plots at the Emile A. Lods Agronomy Research Center of the Macdonald campus of McGill University, Sainte-Anne-de-Bellevue, Quebec, Canada. The experimental site consisted of forty-eight test plots (20 m x 20 m) encompassing three nitrogen treatments and four weed control strategies. The CASI image was obtained with a spatial resolution of 2m x 2m and a spectral range of 408 to 947 nm. Although the image acquisition was made three times during the growing season of year 2000, the image obtained on August 5 (66 days after planting, tassel stage) was used in this study since this season

had the highest correlation between spectral information and crop yield in the previous studies (Goel et al., 2003b; Uno et al., 2003). Spectral reflectance values were extracted from 192 pixels, which corresponded to the sampling sites for the grain yield measurements. Although seventy-two reflectance values were obtained from each pixel, the spectral band at 949 nm was removed because of the high noise level. The complete details of the experimental design and data extraction are given in Goel et al. (2002) and Uno et al. (2003).

4.3.2 Principal component analysis (PCA)

Machine learning algorithms generally require a large number of training samples for model development. The risk of overfitting the data tends to be quite high when there are many input variables. Although most of the recent decision tree estimation algorithms support pruning algorithms to reduce the risk of overfit (Mingers, 1989; Quinlan, 1993), caution is still required during tree development. In this study, principal component analysis was introduced to reduce the number of input variables, since past research has shown that PCA is an effective method for reducing the dimensionality of hyperspectral imagery, which includes a significant amount of redundant information (Yang et al., 2002b; Uno et al., 2003).

4.3.3 The C5.0 decision tree estimation algorithm

The decision tree is a concept of decision-making systems, in which the human decision making process is mimicked by a tree-form representation (Figure 4.1). Although the concept of decision tree is relatively old, one of the constraints to practical application was the difficulty in developing effective algorithms to induce the optimal tree structure (Swain and Hauska, 1977). Indeed, a tremendous amount of effort has been made to develop more accurate and effective decision tree estimation algorithms in the artificial intelligence (AI) community (Weiss and Kulikowski 1991; Michalski et al., 1998; Friedl and Brodley, 1997). The C5.0 decision tree estimation algorithm, a commercial successor of the ID3 and C4.5 algorithms (Quinlan, 1993), is one of the most

commonly used univariate decision tree algorithms, along with the Classification and Regression Tree (C&RT) algorithm (De Fries and Chan, 2000).

As is the case for many such algorithms, the process of rule induction is conducted with simple iterations of partitioning and evaluation of the homogeneity in the partitioned subsets (Figure 4.1). First, an original dataset is randomly split into more than two sub-groups. The impurity of the subgroups is then measured by one of several mathematical equations, called evaluation functions. At this stage, all possible combinations of splitting are tested to find the combination that maximizes the reduction of impurity in the subgroups. Once this original dataset is divided into these subgroups, each split subgroup is randomly partitioned into more than two sub-subgroups by the same procedure. By repeating these partitioning processes, the original group is finally divided into completely homogeneous subgroups that consist of only one attribute (Friedl and Brodley, 1997, Weiss and Kulikowski, 1991). Although many metrics have been developed to measure the impurity of the divided subgroups (Quinlan, 1993; De Fries and Chan, 2000; SAS Institute Inc., 1998; Weiss and Kulikowski, 1991), the “entropy function” is used in C5.0 algorithm.

Once the partitioning process is completed, the leaf nodes are usually cut off one by one from the lowest hierarchal position, since the tree, established at this stage, occasionally becomes over-fitted with various noises and errors in the input values. Many algorithms have been suggested to perform this pruning process (Mingers, 1989). However, the C5.0 algorithm carries out this pruning process based on the comparative error rates of the pruned and unpruned trees (Salzberg, 1994). Details of the pruning algorithms are given in Mingers (1989) and Quinlan (1993).

4.3.4 Crop yield classification

4.3.4.1 Determination of classification boundaries

All the crop yield data (numerical values in kg/ha) were converted into categorical values (i.e. “high” and “low”) at the initial stage of the analysis. This procedure was conducted simply because the C5.0 algorithm does not allow for numerical outputs. However, it should be emphasized that

estimation of the yield levels in categorical values still offers a large amount of useful information in precision agriculture, although estimating the numerical value would have been ideal.

One of the problems in classifying a continuous variable, however, is that it is sometimes difficult to determine an appropriate classification boundary. In fact, the determination of the boundary should always be based on the clear purpose of classification as well as on the distribution of samples. In this study, two different yield-based classification strategies were used, divided according to the agronomic importance (Figure 4.2).

For the first strategy, the crop yields were simply categorized into four yield levels with intervals of 2000 [kg/ha]: (1) Very low: less than 3000 kg/ha; (2) Low: 3000 kg/ha – 5000 kg/ha; (3) High: 5000 kg/ha – 7000 kg/ha; (4) Very high: greater than 7000 kg/ha. For the second strategy, samples were divided into two categories, “low” and “normal”, since the distribution of yields was skewed (Figure 4.2). Twenty-five percent of the samples had yields less than 4828.58 kg/ha and were categorized as “low yield”. The remaining 75 % (144 samples) were categorized into “normal”. This decision boundary was determined on the assumption that crops represented by the lower 25% of the samples (Figure 4.2) could be regarded as growing in somewhat poor field conditions while the remaining 75% were regarded as growing in normal conditions or good conditions. It should be noted that detection of these “low yield levels” would offer useful information to diagnose various adverse field conditions, such as water, nitrogen, and weed stresses.

4.3.4.2 Model development

For the model development, two different input strategies were taken, in addition to a conventional reclassification method (Figure 4.3). For the first input strategy, all seventy-one spectral bands were used as input variables. For the second input strategy, factor scores obtained from the top five principal components were used as the input variables. The performance of these DT models was evaluated by comparing them to the results obtained from an ANN model. This was based on a previous study (Uno et al., 2003) that the ANN model achieved much higher performance than

other conventional methods, such as NDVI, SR, and PRI based models. The model development in this study was all conducted with *Clementine Data Mining System (SPSS Inc.)*

Although the *Clementine data mining system* offers many options to implement the optimum training process, the defaults were used except “generality” option. Although the exact mechanism of this “generality” option was not clearly mentioned, it is reported that it decreases the risk of overfit, whereas the “accuracy” option tries to develop the most accurate tree based on the training dataset, which usually results in poor performance during the validation step (SPSS Inc., 2001).

4.3.4.3 Performance analysis

The performance analysis was done by ten-fold cross validation, since the number of samples used in this study (192 records) was small compared to the number of input variables (71 inputs). The samples were first divided into ten independent subsets (group A to J, 19 or 20 records for each group), and nine out of ten subgroups (ex. Group A to I) were selected for the training and the remaining one subset (group J) was kept for the validation. After evaluating the performance of the model, nine different subgroups (ex. Group A to H and J) were picked for training, and the one unseen subgroup (group I) was later used for the validation. By repeating this process, all the ten combinations for training and validation were tested one by one. It should be noted that the training and validation datasets were always completely independent. Medians and 25% upper- and lower-quartiles were computed from the classification accuracies obtained with ten independent validation datasets. Finally, results obtained with all ten validation datasets were summarized into one confusion matrix.

4.3.5 Feature band selection

The following three steps were taken to evaluate the performance of the C5.0 algorithm as a feature band selection tool: (i) Identification of the important spectral bands by browsing the developed decision tree model, (ii) Development of an ANN model by using the identified spectral bands as

input variables, and (iii) Performance analysis of the developed model, as compared to ANN models with two different input strategies (Figure 4.4).

For the identification of the important spectral bands, six wavelengths were selected from the higher leaf nodes in the developed tree, since the nodes in the higher hierarchical position usually bear more useful information than do the lower nodes. The performance of the ANN model with these six inputs was evaluated by comparing it to (i) ANNs with all 71 input variables directly obtained from the CASI image, and (ii) ANNs whose input variables were obtained from five principal components, extracted from the CASI image (Figure 4.4). Further information on these two methods is given in Uno et al. (2003).

Root Mean Square Error (RMSE) was used to evaluate the performance of the ANN models. RMSE is one of the common statistical parameter for the analysis of model performance, and it represents the expected difference between observed and estimated values (Yang et al., 1997). The equation to calculate RMSE is:

$$\text{RMSE} [\%] = (100 / \bar{O}) \sqrt{\frac{\sum (P_i - O_i)^2}{n}} \quad (1)$$

Finally, a graphical analysis was conducted to compare these different input strategies in ANN models.

4.4 Results and Discussion

Overall classification accuracies, obtained with the three yield classification strategies, are summarized in Table 4.1. Median values and 25% upper and lower quartiles were calculated from ten different validations. For the first classification strategy (classification into four yield levels), the median value obtained with C5.0 classifiers (52.63% for ANNs with 71 input variables and 42.11% for ANNs with five principal components) was lower than ANN reclassification method

(57.89%). Graphical analysis (Figures 4.5) also showed that the performance of C5.0 was slightly lower than ANN reclassification method. However, the significance of these differences was not clear, since the classification accuracies varied widely depending on the selection of calibration and validation datasets. For the second classification strategy (classification into two yield levels), overall classification accuracy, obtained with the original 71 input variables (89.47%), was even higher than ANN reclassification method (87.24%) in terms of median values (table 4.1). Again, the significance of the difference was not clear due to the large deviation among the ten validation datasets (Figure 4.6).

One of the interesting points, observed in the first classification strategy, is that most of the misclassifications were made between adjacent yield levels (Table 4.2, 4.3, and 4.4). Indeed, more than 90% of the misclassifications (approximately 93% for C5.0 classifier with 71 input variables, 91% for C5.0 classifier with 5 principal components, and 96% for ANN reclassification method) were actually made between adjacent yield levels. One of the most important reasons for this seemed to be that samples, with characteristics close to the decision boundaries, are quite difficult to be classified into the appropriate category. As a matter of fact, differentiating two samples, located on either side of decision boundary, is sometimes a hard task, based on the limited amount of information obtained from the spectral signature.

From practical point of view, these high misclassification rates between adjacent yield levels imply that the risk of misclassification was relatively low, as compared to the impression which overall accuracies seem to offer. However, it should be noted that the performance still seemed to be unsatisfactory for application to precision agriculture. For example, a high misclassification rate between “Low” and “High”, made by the C5.0 classifier with five principal components (Table 4.3), would cause crucial mistakes in field operations if the information was used for VRTs, since the “Low” yield levels can be generally recognized as stressed areas (Figure 4.2) which require special treatment such as additional fertilizer and pesticide applications.

Confusion matrices obtained with second classification strategy (classification into two yield levels) were presented in Table 4.5. One of the important points was that misclassification rate in the low productivity areas (less than 4828.58 kg/ha in observed values) was quite high (36.1% for

C5.0 with 71 input variables, 51.1% for C5.0 with 5 PCs, and 29.8% for ANN reclassification method), compared to the overall accuracies. From the practical point of view, these misclassification rates indicate that more than 3 out of 10 stressed areas could be ignored or remain undetected if the C5.0 algorithm was used to create a yield map in this specific environment. In particular, the misclassification rate, made by the C5.0 classifier with five principal components (51.1%), was too high for practical purposes.

Although there are various reasons for a high misclassification rate in the low productivity areas, one of the most important reasons seemed to be that the developed models were biased towards classification in the high productivity areas, since the number of training samples were largely different between these two sub-groups (75% of the samples were categorized into “normal” yield level in this study). Indeed, it is reported that classification for the small sub-groups tend to be ignored during the optimization process of the C5.0 algorithm, since the algorithm normally tries to maximize the classification accuracies based on the overall classification accuracy (Friedl and Brodley, 1997; Hansen et al., 2000). Reducing the number of training samples in the larger subgroups can be helpful in producing higher classification accuracies in the low productivity areas.

Prediction accuracies of ANN models obtained with three different input strategies, (1) 71 input variables (ANNs-71inputs) (2) six bands selected by C5.0 algorithm (C5.0-ANNs), and (3) five principal components (PCA-ANNs) are summarized in Table 4.6. Although ANNs with 71 input variables exhibited the best performance, the differences among these three input strategies were generally small. Graphical analysis (Figure 4.7) also showed some minor differences among these input strategies. From the practical point of view, however, the loss of information caused by these data reduction processes seemed to be small enough, considering the benefit obtained from the reduction of dimensionality in spectral information.

The structure of the developed decision tree is presented in Figure 4.8. Six wavelengths (693.76, 792.96, 542.79, 453.21, 716.58, and 869.80 nm) were identified as the important wavelengths based on this structure. However, it should be noted that generality of the results was not clear,

since the structure of the developed tree could change depending on the way the training samples are selected.

4.5 Conclusion

This study evaluated the potential of the C5.0 decision tree algorithm for detecting within-field crop productivity based on hyperspectral imagery. The results showed that the performance of the C5.0 classifier was comparable to that of the conventional reclassification method. However, further improvements are still required to apply the method to precision agriculture. Further research should aim to increase the classification accuracy, reduce the risk of misclassification, and find appropriate classification boundaries. This study also demonstrated that the potential of C5.0 algorithm as a feature subset selection tool is high. The decrease in prediction accuracies with the reduced dataset with C5.0 algorithm was small (decrease in RMSE was 0.72% for calibration and 0.17% for validation), compared to the original dataset. It should also be noted that the prediction accuracies obtained with C5.0-ANN model (RMSE=17.62% for calibration and 19.86% for validation) were even higher than those of the PCA-ANN model (RMSE=19.16% for calibration and 22.06% for validation). However, further exploration is still required to evaluate the stability of the developed tree structures. This involves: (1) estimating the minimum number of samples to obtain a stable decision tree structure, and (2) evaluating the effect of sample quality, typically inter-correlation of spectral bands, on the stability of tree structure.

From a practical point of view, application of the C5.0 algorithms to feature extraction seemed to be somehow limited in comparison to PCA, since C5.0 is basically a supervised learning algorithm which requires output variables for training. This means that C5.0 may not be applicable to the situation where too few training samples are available.

4.6 References

- Debus, J. C. W., and V. J. Rayward-Smith, 1997. Feature subset selection within a simulated annealing data mining algorithm. *Journal of intelligent information systems* 9:57-81
- De Fries, R. S., M. Hansen, J. R. G. Townshend, and R. Sohlberg 1998. Global land cover classifications at 8km spatial resolution: the use of training data derived from Landsat imagery in decision tree classifiers. *International journal of remote sensing* 19(16): 3141-3168
- De Fries, R. S. and J. C-W Chan. 2000. Multiple criteria for evaluating machine learning algorithms for land cover classification from satellite data. *Remote sensing of environment* 74(3): 503-515
- Friedl, M. A. and C. E. Brodley. 1997. Decision tree classification of land cover from remotely sensed data. *Remote sensing of environment* 61(3): 399-409
- Goel, P. K., S. O. Prasher, J. A. Landry, R. M. Patel, R. B. Bonnell, A. A. Viau, J. R. Miller 2002. Potential of airborne hyperspectral remote sensing to detect nitrogen and weed infestation. *Computer and Electronics in agriculture* 38(2): 99-124
- Goel, P. K., S. O. Prasher, R. M. Patel, J. A. Landry, R. B. Bonnell, and A. A. Viau 2003a. Classification of hyperspectral data by decision trees and artificial neural networks to identify weed stress and nitrogen status of corn. *Computers and Electronics in Agriculture* 39(2) 67-93
- Goel, P. K., S. O. Prasher, J. -A. Landry, R. M. Patel, and A. A. Viau. 2003b. Estimation of crop biophysical parameters through airborne and field hyperspectral remote sensing. *Transaction of the ASAE* (In press)
- Hansen, M. Dubayah, R. and R. Defries. 1996. Classification trees: an alternative to traditional land cover classification. *International journal of remote sensing* 17(5): 1075-1081
- Mair, C., G. Kadoda, M. Lefley, K. Phalp, C. Schofield, M. Shepperd, and S. Webster. 2000. An investigation of machine learning based prediction systems. *The journal of systems and software* 53(1): 23-29
- Mather, P. M. 2000. *Computer processing of remotely-sensed images: an introduction*. John Wiley & Son Ltd., Chichester, UK

- Michalski, R. S., Bratko I., and M. Kubat. 1998. *Machine learning and data mining: Methods and applications*. John Wiley & Son Ltd., Chichester, West Sussex, England
- Mingers, J. 1989. An empirical comparison of pruning methods for decision tree induction. *Machine learning* 4: 227-243
- Quinlan, R. J. 1993. *C4.5: Programs for Machine learning*. M. Kaufmann Publisher Inc., San Mateo, CA.
- Reitz, P., and H.D. Kutzbach. 1996. Investigations on a particular yield mapping system for combine harvesters. *Computers and Electronics in Agriculture* 14(2-3): 137-150
- Rogan, J., J. Franklin, D. A. Roberts. 2002. A comparison of methods for monitoring multitemporal vegetation change using Thematic Mapper imagery. *Remote sensing of environment* 80(1): 143-156
- SAS Institute Inc. 1998. *Data Mining Primer: Overview of applications and methods*. SAS Institute Inc. Cary, NC, USA
- Salzberg S. L. 1994. Book Review: C4.5: Program for machine learning by J. Ross Quinlan. Morgan Kaufman Publishers, Inc., 1993. *Machine learning* 16: 253-240
- Simard, M. Saatchi, S. S. and G. De Grandi. 2000. The use of decision tree and multiscale texture for classification of JERS-1 SAR data over tropical forest. *IEEE Transaction on Geoscience and remote sensing* 38(5): 2310-2321
- SPSS Inc. 2001. *Clementine Version 6.0 User's Guide*. SPSS Inc., Chicago, IL
- Stafford, J. V. 2000. Implementing precision agriculture in the 21st century. *Journal of Agricultural Engineering Research* 76: 267-275
- Swain, P. H. and H. Hauska. 1977. The decision tree classifier: design and potential. *IEEE Transaction on geoscience electronics* GE-15(3): 142-147
- Uno, Y., S. O. Prasher, P. K. Goel, Y. Karimi, and A. A. Viau. 2003. Artificial neural networks to predict corn yield from Compact Airborne Spectrographic Imager (CASI) data. (Under preparation)
- Weiss, S. M. and C. A. Kulikowsk. 1991. *Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems*. M. Kaufmann Publishers Inc., San Mateo, Calif.
- Yang, C. -C., S. O. Prasher, G. R. Mehuys, and N. K. Patni. 1997. Application of artificial neural networks for simulation of soil temperature. *Transaction of the ASAE* 40(3): 649-656

- Yang, C. –C., S. O. Prasher, J. Whalen, and P. K. Goel. 2002a. Use of hyperspectral imagery for identification of different fertilization methods with decision-tree technology. *Biosystems Engineering* 83(3): 291-298
- Yang, C. –C., S. O. Prasher, P. Enright, C. Madramootoo, M. Burgess, P. K. Goel, and I. Callum. 2003. Application of decision tree technology for image classification using remote sensing data. *Agricultural systems* 76: 1101-1117
- Yang, C., J. M. Bradford and C. L. Wiegand. 2001. Airborne multispectral imagery for mapping variable growing conditions and yields of cotton, grain sorghum, and corn. *Transaction of the ASAE* 44(6): 1983-1994
- Yang, C., J. H. Everitt, and J. M. Bradford. 2002b. Airborne hyperspectral imaging and yield monitoring of grain sorghum yield variability. ASAE Paper No. 02-1079. Chicago Illinois: ASAE

Figure 4.1 An example of tree representation for the human decision making process. C5.0 is one of many algorithms, which estimate the optimum tree structure from the instances.

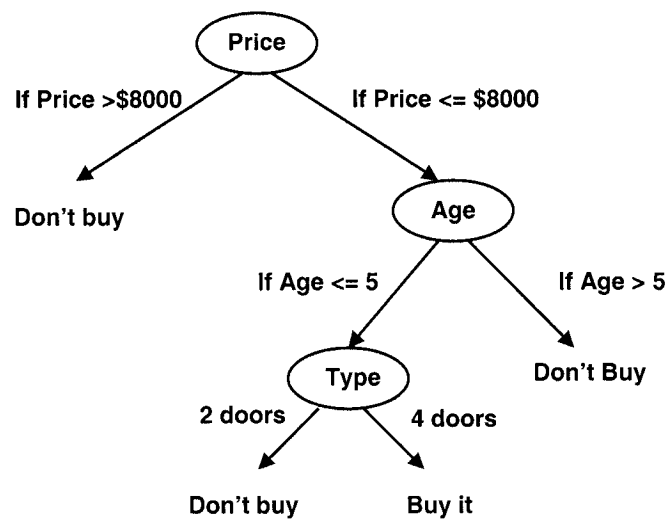


Figure 4.2 Distribution of yield samples. Two different classification strategies were taken based on this distribution in this study.

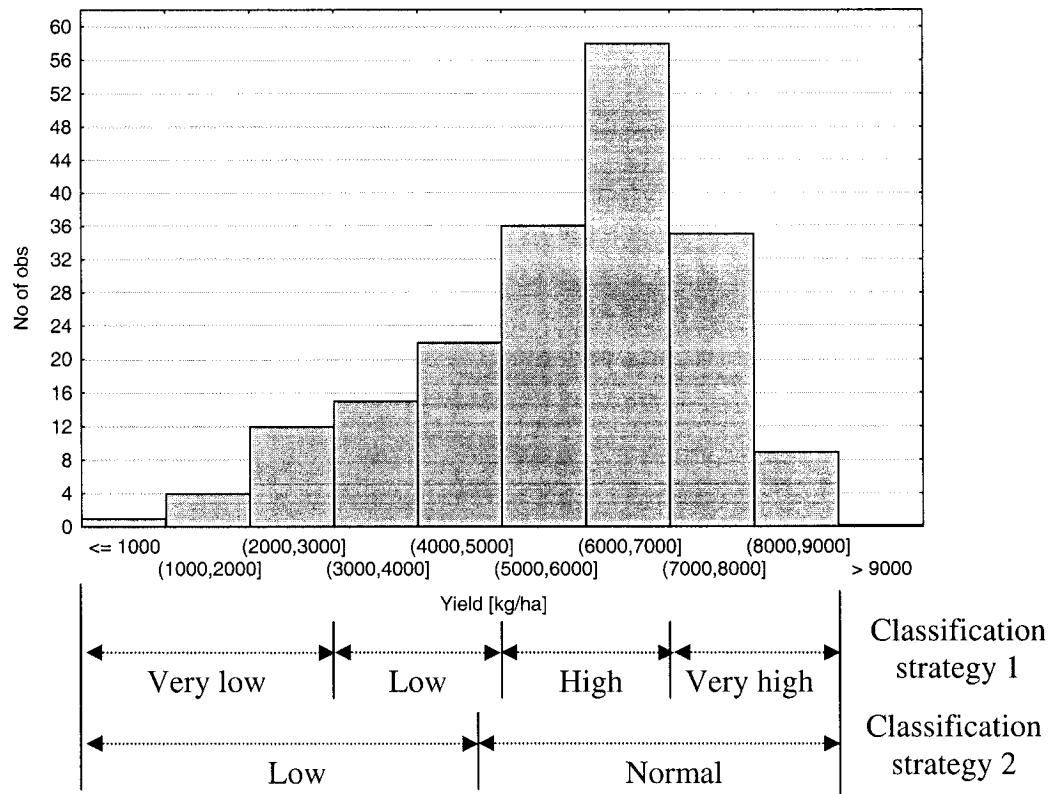


Figure 4.3 Two different classification strategies and performance standard made by reclassification process were tested to evaluate the performance of C5.0 algorithm in this study.

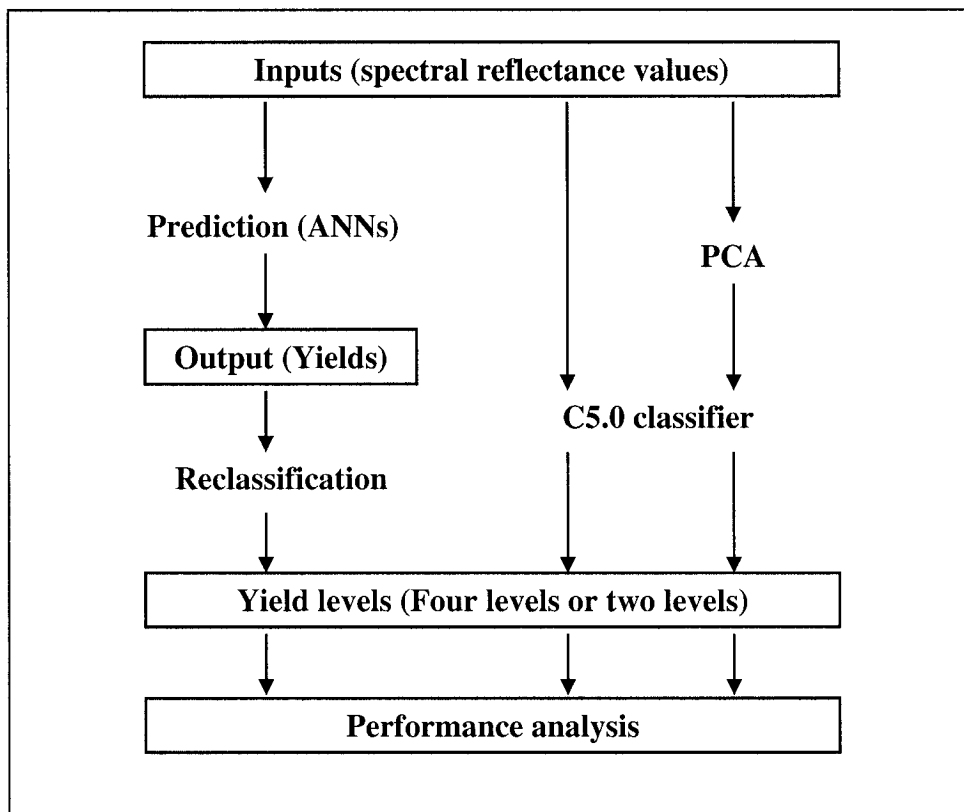


Figure 4.4 Three different input strategies were taken to evaluate the performance of C5.0 algorithm as a spectral band selection tool.

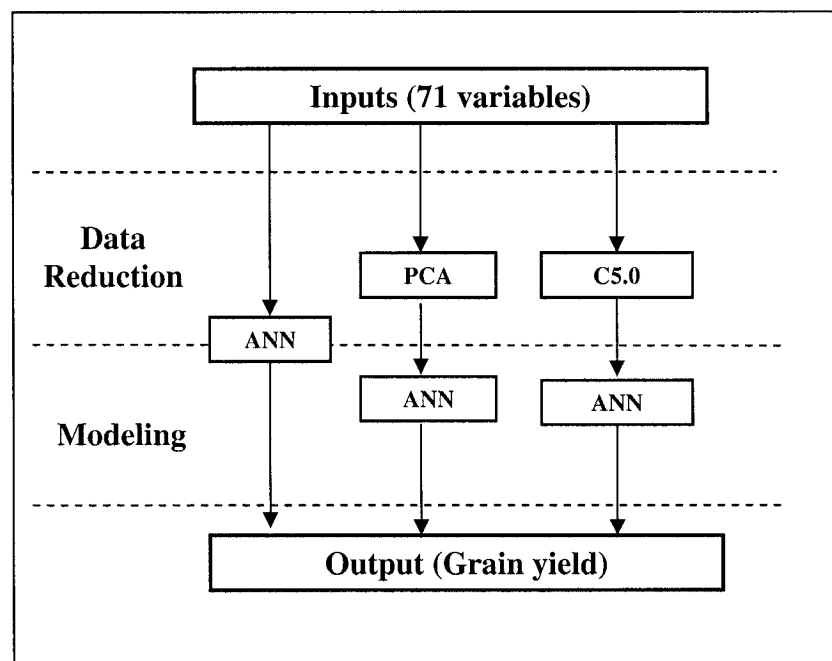


Figure 4.5 Deviation of the overall classification accuracies (classification into four-levels) obtained with ten different validation dataset. Vertical axis shows the classification accuracy on percentage.

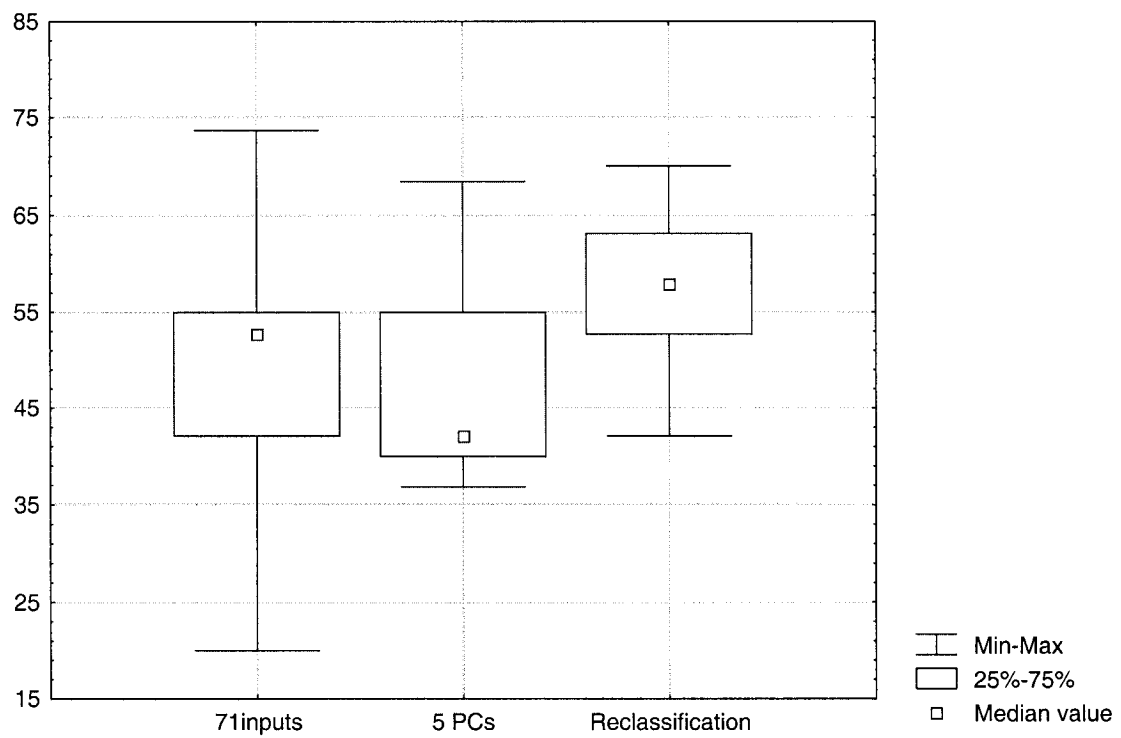


Figure 4.6 Deviation of the overall classification accuracies (classification into two-levels) obtained with ten different validation dataset. Vertical axis shows the classification accuracy on percentage.

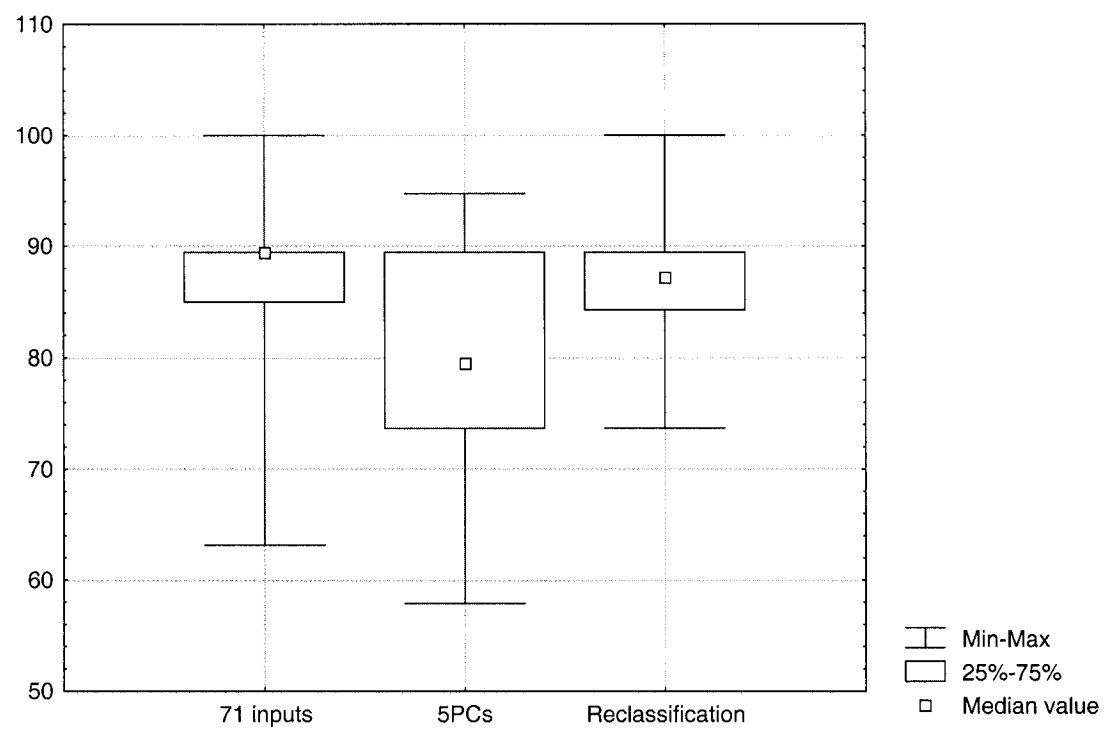


Figure 4.7 Performance of ANN models with three different input strategies. (A) ANNs with 71 input variables, (B) ANNs with six spectral bands selected by C5.0 algorithm (C) ANNs with five principal components. Figures on the left show the results for calibration, and figures on the right show the results for validation.

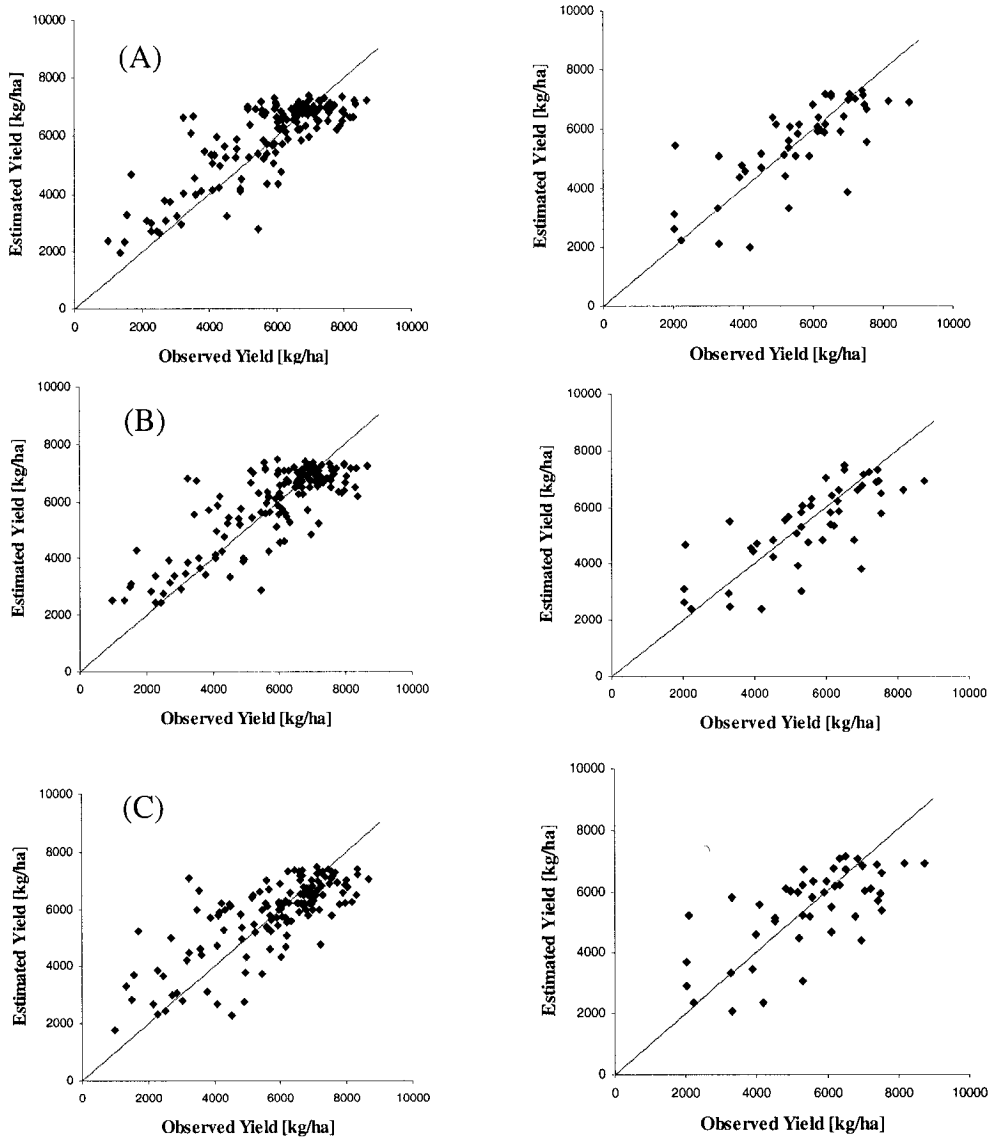


Table 4.1 Overall classification accuracies obtained with different classification and input strategies. Median values and inter quartile ranges (IQR) were calculated from ten different validation datasets.

Classification accuracies (%)						
Input strategy	71 input variables		5 principal components		ANN-Reclassification	
	Median	IQR	Median	IQR	Median	IQR
4 levels	52.63	42.11~55.00	42.11	40.00~55.00	57.89	52.63~63.16
2 levels	89.47	85.00~89.47	79.48	73.68~89.47	87.24	84.21~89.47

Table 4.2 A confusion matrix obtained by C5.0 classifier with 71 input variables. All the results obtained with ten different validation datasets were summarized in this one confusion matrix.

		Estimated yield level			
		Very Low	Low	High	Very High
Observed Yield level	Very Low	9	7	1	0
	Low	11	18	8	0
	High	3	8	59	23
	Very High	0	2	27	16

Table 4.3 A confusion matrix obtained by C5.0 classifier with five principal components. All the results obtained with ten different validation datasets were summarized in this one confusion matrix.

		Estimated yield level			
		Very Low	Low	High	Very High
Observed Yield level	Very Low	7	7	3	0
	Low	7	4	23	3
	High	3	7	69	14
	Very High	0	0	35	10

Table 4.4 A confusion matrix obtained with performance standard (reclassification from ANN prediction values). All the results obtained with ten different validation datasets were summarized in this one confusion matrix.

		Estimated yield level			
		Very Low	Low	High	Very High
Observed Yield level	Very Low	9	8	0	0
	Low	4	18	13	2
	High	1	12	73	7
	Very High	0	0	33	12

Table 4.5 Confusion matrices obtained with the second classification strategy (classification into two yield levels). The results obtained with ten different validation datasets were summarized into one confusion matrix.

C5.0 with 71 inputs		Estimated yield level	
		Normal	Low
Observed Yield	Normal	135	10
	Low	17	30

C5.0 with 5 principal components		Estimated yield level	
		Normal	Low
Observed Yield	Normal	130	15
	Low	24	23

ANN reclassification		Estimated yield level	
		Normal	Low
Observed Yield	Normal	130	14
	Low	14	34

Table 4.6 Performance of the ANN models with three different input strategies (1) ANNs with original 71 input variables (ANNs-71inputs), (2) ANNs with 6 spectral bands identified with C5.0 algorithm (C5.0-ANNs), and (3) ANNs with 5 principal components (PCA-ANNs) were compared each other in this study.

	Prediction Accuracies (RMSE)	
	Calibration (%)	Validation (%)
ANNs-71inputs	16.90	19.69
C5.0-ANNs	17.62	19.86
PCA-ANNs	19.16	22.06

Chapter 5 - Summary and Conclusions

5.1 Summary

The potential of two machine learning algorithms, ANNs and DT, for the development of yield mapping and forecasting systems from airborne hyperspectral imagery was explored in this study. The imagery was obtained over experimental plots, cropped with corn at the Emile A. Lods Agronomy Research Center on the Macdonald campus of McGill University, Sainte-Anne-de-Bellevue, Quebec, Canada. The experimental plots were designed to simulate various crop growth scenarios, involving combinations of three different nitrogen application rates (60, 120, and 250 kg N/ha) and four different weed control strategies (broadleaf, grass, broadleaf and grass, and no weed control).

The hyperspectral images were obtained with a compact airborne spectrographic imager (CASI) having a spatial resolution of 2m x 2m, and a spectral range of 408 to 947 nm. The spectral resolutions were approximately 7.5nm. Hyperspectral images were acquired at three times during the year 2000 growing season of year: a) June 30 to represent the early growth stage, b) August 5 at which the corn had reached the tassel stage, and c) August 25 at the fully matured stage. However, the image acquired on the second flight (August 5) was the only one used, since this study was focused on the performance of machine learning algorithms rather than on prediction of the seasonal variations of crop yields.

Although many different algorithms are currently available for ANNs and DT, back-propagation neural network architecture and the C5.0 decision tree estimation algorithms were used in this study, since they are the most commonly used architectures or algorithms. The model developments were all conducted with a data mining software package, *Clementine Data Mining System* (SPSS Inc.). To handle the large amount of redundant information included in hyperspectral imagery, PCA was also used.

In the first set of analyses (Chapter 3), the study evaluated the ANN approach, as compared with conventional modeling methods, such as SMLR models and VI-based modeling methods. Three different VIs (NDIV, SR, and PRI) were tested as performance controls.

In the second set of analyses (Chapter 4), the C5.0 DT estimation algorithms was assessed in terms of (i) performance as a yield classification method, and (ii) performance as a feature band selection tool. For the performance analysis of yield classification, the results were compared with conventional reclassification methods using predicted yield values from the ANNs. Evaluation of the DT algorithm as a feature band selection tool was based on a comparison of its performance with that of PCA.

5.2 Conclusions

This study demonstrated that the potential of machine learning algorithms for the development of in-season yield mapping and forecasting system is generally high. In particular, high prediction accuracies obtained with ANNs demonstrated that ANNs can be an effective alternative to conventional VI-based method. Although further improvement is still required for application of ANNs to precision farming, they have potential as a tool for the development of in-season yield mapping systems from remote sensing imagery.

This study also showed that the performance of the C5.0 algorithm as a yield classification method was comparable with that of conventional reclassification methods. However, the performance still seemed to be unsatisfactory for the practical purposes. Further exploration of the algorithms is necessary if better classification accuracies is to be achieved.

The C5.0 algorithm performed better than PCA as a feature band selection tool. However, certain limitations must be overcome before putting them to practice, because this type of algorithm generally requires a large number of samples for training. It should be noted that data reduction is normally useful in a situation where the numbers of samples is too small for training or model

development. Further analysis is needed to confirm the stability of the model structure as the models were constructed from very limited data.

In general, the data mining approach appeared to be quite effective as a tool for analysis of spectral data. With *Clementine Data Mining System*, various complicated analytical procedures were reduced into relatively simple operations due to the user-friendly interface and highly integrated systems. However, the analytical procedures were not always very clear. This resulted in a limitation to the appropriate understanding and implementation of machine learning algorithms or statistical analysis, especially for scientific studies.

Chapter 6 - Recommendations for further research

Although the potential of machine learning algorithms for yield estimation was demonstrated in this study, further research work is required to develop an in-season yield mapping system, or crop yield forecasting system based on the interpretation of hyperspectral data using machine learning methods.

6.1 Increasing the generality of models

One of the most crucial steps to be taken is to increase the generality of the models. In particular, (i) use of multiple years of yield data and images to predict the seasonal variation of crop yield, (ii) testing various different crop types and cultivars in many locations, and (iii) conducting experiments under various environmental conditions such as water and nutrient deficit conditions, are the essential tasks to develop highly generalized models.

The unavailability of multiple years of yield data and images at a within-field scale is one of the most serious constraints in model development. Indeed, some of the past researches with satellite imagery infer that large numbers of yield data and images (more than nine years) is necessary to develop highly reliable prediction models (Hayes and Decker, 1996 and 1998; Maselli et al., 2000). These researches also infer that the collection of the images often requires multiple observations during the growing season. However, it seems to be unrealistic to obtain these large numbers of yield data and images from an aerial platform.

Some of the latest optical satellite systems, such as IKONOS and Quickbird, which offer high spatial resolutions (1m and 0.61m for panchromatic, and 4m and 2.44m for multi-spectral) and short revisit time (1-5days), can increase the possibility of applying satellite systems to field-scale observations. However, limitations still exist such that (i) the number of the spectral bands is still limited (they only cover four bands in the visible and NIR range), (ii) influence of weather is still high, (iii) it is difficult to obtain many years of yield data at the within-field level in many locations and crops.

Several approaches to overcome these technical and economical limitations have been suggested. Ancillary data, such as meteorological information and field conditions (Simpson, 1994; Hayes and Decker, 1998), may be helpful in reducing the number of images required as well as to increase the prediction accuracies. Integration of multiple sensors, including radar and broadband sensor images (Moran et al., 1997 and 2002) may also help to increase the performance of the models, and also remove the influence of weather conditions. It should be noted that the ability of machine learning algorithms, which can easily incorporate the ancillary information, even if they are non-numerical values, may be helpful in this regard. However, it can not be denied that the development of yield forecasting systems still needs long-term contributions, including the collection of within-field scale yield information over long periods and establishment of basic infrastructure such as spectral and yield databases, which cover various crop types and environmental condition.

6.2 Further exploration of image processing techniques and machine learning algorithms

From the technical aspects of image processing and machine learning algorithms, some further exploration can be suggested. First, use of some pre-processing techniques, typically low-pass filters, may be useful in removing noise, and consequently increase the prediction accuracies. Exploration of the machine learning algorithms, typically further optimization of network architectures and modifying training sample selections, may increase the prediction accuracies to some extent. For C5.0 algorithms, boosting, debugging, and fixing the misclassification cost values can be tested to increase the performance of the models. For PCA, the use of factor rotations and different factor loading methods should be explored. It should also be noted that use of many other machine learning algorithms, which are currently being developed in the AI community, will offer new opportunities for model development in the near future.

One of the constraints associated with ANNs and DT is that they normally require a large number of training samples to obtain reliable results. However, recent proliferation of tractor-mounted yield monitoring system will definitely increase the opportunity to obtain large numbers of yield

samples at the within-field scale. Using this large number of yield samples, more reliable and accurate ANNs and DT models could be developed.

6.3 References

- Hayes, M. J. and W. L. Decker. 1996. Using NOAA AVHRR data to estimate maize production in the United States Corn Belt. *International Journal of Remote Sensing* 17(16): 3189-3200
- Hayes, M. J. and W. L. Decker. 1998. Using satellite and real-time weather data to predict maize production. *International Journal of Biometeorology* 42(1): 10-15
- Maselli, F., S. Romanelli, L. Bottai and G. Maracchi. 2000. Processing of CAC NDVI data for yield forecasting in the Sahelian region. *International Journal of Remote sensing* 21(18): 3509-3523.
- Simpson, G. 1994. Crop yield prediction using a CMAC neural network. In *Proceedings of the Society of Photo-Optical Instrumentation Engineers* 2315, 160-171. Bellingham, WA: The International Society for Optical Engineering.
- Moran, S. M. Vidal A. Troufleau D. Qi J. Clarke T. R. Pinter Jr. P. J. Mitchell T.A. Inoue Y. and C. M. U. Neale 1997. Combining multifrequency microwave and optical data for crop management. *Remote sensing of environment* 61: 96-109
- Moran, S. M. Hymer D. C. Qi J. and Y. Kerr. 2002. Comparison of ESR-2 SAR and Landsat TM imagery for monitoring agricultural crop and soil conditions. *Remote sensing of environment* 79: 243-252

Appendix A: Results of Principal component analysis

Communalities		
	Initial	Extraction
408.73nm_3	1.000	.733
416.13nm_3	1.000	.752
423.53nm_3	1.000	.730
430.95nm_3	1.000	.627
438.36nm_3	1.000	.939
445.79nm_3	1.000	.675
453.21nm_3	1.000	.773
460.65nm_3	1.000	.755
468.09nm_3	1.000	.758
475.53nm_3	1.000	.756
482.98nm_3	1.000	.719
490.44nm_3	1.000	.773
497.90nm_3	1.000	.849
505.37nm_3	1.000	.823
512.84nm_3	1.000	.888
520.32nm_3	1.000	.970
527.80nm_3	1.000	.988
535.29nm_3	1.000	.991
542.79nm_3	1.000	.992
550.29nm_3	1.000	.994
557.79nm_3	1.000	.995
565.35nm_3	1.000	.995
572.82nm_3	1.000	.995
580.34nm_3	1.000	.993
587.86nm_3	1.000	.992
595.39nm_3	1.000	.991
602.93nm_3	1.000	.993
610.47nm_3	1.000	.990
618.02nm_3	1.000	.986
625.57nm_3	1.000	.987
633.13nm_3	1.000	.988
640.69nm_3	1.000	.984
648.26nm_3	1.000	.978

655.83nm_3	1.000	.978
663.41nm_3	1.000	.972
670.99nm_3	1.000	.964
678.57nm_3	1.000	.959
686.17nm_3	1.000	.952
693.76nm_3	1.000	.978
701.36nm_3	1.000	.990
708.97nm_3	1.000	.994
716.58nm_3	1.000	.993
724.20nm_3	1.000	.987
731.82nm_3	1.000	.978
739.45nm_3	1.000	.982
747.08nm_3	1.000	.988
754.71nm_3	1.000	.992
762.35nm_3	1.000	.993
770.00nm_3	1.000	.994
777.65nm_3	1.000	.996
785.30nm_3	1.000	.997
792.96nm_3	1.000	.997
800.62nm_3	1.000	.998
808.29nm_3	1.000	.998
815.96nm_3	1.000	.998
823.64nm_3	1.000	.998
831.32nm_3	1.000	.997
839.01nm_3	1.000	.998
846.70nm_3	1.000	.997
854.39nm_3	1.000	.997
862.09nm_3	1.000	.997
869.80nm_3	1.000	.997
877.51nm_3	1.000	.995
885.22nm_3	1.000	.995
892.93nm_3	1.000	.994
900.66nm_3	1.000	.990
908.38nm_3	1.000	.990
916.11nm_3	1.000	.984
923.84nm_3	1.000	.982
931.58nm_3	1.000	.918
939.33nm_3	1.000	.887

Extraction Method: Principal Component Analysis.

Total Variance Explained						
	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	32.276	45.459	45.459	32.276	45.459	45.459
2	23.978	33.772	79.231	23.978	33.772	79.231
3	7.926	11.164	90.395	7.926	11.164	90.395
4	1.880	2.648	93.043	1.880	2.648	93.043
5	.689	.970	94.013	.689	.970	94.013
6	.500	.704	94.716			
7	.487	.686	95.403			
8	.377	.531	95.934			
9	.353	.497	96.431			
10	.318	.447	96.878			
11	.301	.424	97.302			
12	.279	.392	97.694			
13	.240	.338	98.033			
14	.219	.308	98.341			
15	.196	.276	98.618			
16	.170	.239	98.857			
17	.144	.202	99.059			
18	.101	.142	99.201			
19	9.722E-02	.137	99.338			
20	7.451E-02	.105	99.443			
21	5.613E-02	7.905E-02	99.522			
22	3.610E-02	5.085E-02	99.573			
23	3.117E-02	4.390E-02	99.617			
24	2.594E-02	3.654E-02	99.653			
25	2.456E-02	3.459E-02	99.688			
26	2.199E-02	3.097E-02	99.719			
27	1.972E-02	2.778E-02	99.747			
28	1.524E-02	2.147E-02	99.768			
29	1.414E-02	1.992E-02	99.788			
30	1.202E-02	1.693E-02	99.805			
31	1.126E-02	1.587E-02	99.821			
32	1.057E-02	1.488E-02	99.836			

33	9.806E-03	1.381E-02	99.850			
34	9.524E-03	1.341E-02	99.863			
35	8.159E-03	1.149E-02	99.874			
36	7.416E-03	1.045E-02	99.885			
37	6.727E-03	9.475E-03	99.894			
38	6.142E-03	8.650E-03	99.903			
39	5.428E-03	7.645E-03	99.911			
40	5.338E-03	7.519E-03	99.918			
41	4.940E-03	6.958E-03	99.925			
42	4.430E-03	6.239E-03	99.931			
43	4.212E-03	5.932E-03	99.937			
44	4.013E-03	5.652E-03	99.943			
45	3.776E-03	5.318E-03	99.948			
46	3.373E-03	4.750E-03	99.953			
47	3.167E-03	4.460E-03	99.958			
48	3.035E-03	4.274E-03	99.962			
49	2.774E-03	3.907E-03	99.966			
50	2.278E-03	3.209E-03	99.969			
51	2.156E-03	3.037E-03	99.972			
52	2.066E-03	2.909E-03	99.975			
53	1.830E-03	2.578E-03	99.977			
54	1.704E-03	2.400E-03	99.980			
55	1.572E-03	2.215E-03	99.982			
56	1.448E-03	2.040E-03	99.984			
57	1.305E-03	1.838E-03	99.986			
58	1.247E-03	1.756E-03	99.988			
59	1.135E-03	1.598E-03	99.989			
60	1.055E-03	1.486E-03	99.991			
61	9.862E-04	1.389E-03	99.992			
62	9.416E-04	1.326E-03	99.993			
63	7.858E-04	1.107E-03	99.995			
64	6.779E-04	9.548E-04	99.996			
65	6.103E-04	8.596E-04	99.996			
66	5.472E-04	7.706E-04	99.997			
67	4.906E-04	6.910E-04	99.998			
68	4.400E-04	6.197E-04	99.998			
69	3.839E-04	5.407E-04	99.999			
70	3.688E-04	5.194E-04	100.000			

71	3.281E-04	4.621E-04	100.000			
Extraction Method: Principal Component Analysis.						

Component Matrix(a)					
	Component				
	1	2	3	4	5
408.73nm_3	.161	.185	.700	-.427	-2.509E-02
416.13nm_3	.290	.135	.722	-.353	-5.561E-02
423.53nm_3	.193	.249	.730	-.277	-.143
430.95nm_3	.122	.258	.699	-.235	-5.128E-02
438.36nm_3	.177	.274	.609	-.102	.671
445.79nm_3	.242	.316	.706	-.114	6.451E-02
453.21nm_3	.304	.387	.694	-9.594E-02	-.199
460.65nm_3	.344	.303	.682	-.105	-.263
468.09nm_3	.383	.254	.702	-.113	.203
475.53nm_3	.399	.294	.710	-7.055E-02	-3.167E-02
482.98nm_3	.429	.357	.619	.107	.117
490.44nm_3	.486	.226	.668	7.556E-02	-.183
497.90nm_3	.576	.416	.559	.179	-3.945E-03
505.37nm_3	.667	.475	.353	.168	-4.020E-03
512.84nm_3	.729	.581	9.640E-02	9.727E-02	-1.110E-02
520.32nm_3	.699	.688	-5.781E-02	-5.458E-02	2.960E-02
527.80nm_3	.649	.714	-.214	-.110	-1.467E-03
535.29nm_3	.633	.706	-.263	-.153	-2.758E-03
542.79nm_3	.621	.711	-.279	-.151	2.332E-03
550.29nm_3	.616	.706	-.304	-.153	-5.299E-04
557.79nm_3	.630	.696	-.300	-.153	-5.646E-03
565.35nm_3	.657	.682	-.278	-.142	6.759E-03
572.82nm_3	.693	.660	-.265	-9.546E-02	6.908E-03
580.34nm_3	.725	.640	-.233	-5.804E-02	1.739E-05
587.86nm_3	.749	.620	-.211	-3.296E-02	5.010E-03
595.39nm_3	.749	.621	-.210	-3.665E-03	-1.292E-03
602.93nm_3	.774	.600	-.183	6.295E-03	5.570E-03
610.47nm_3	.791	.580	-.165	3.414E-02	2.798E-03
618.02nm_3	.814	.550	-.123	7.278E-02	-1.392E-02
625.57nm_3	.822	.538	-9.519E-02	.109	7.755E-03
633.13nm_3	.835	.518	-8.160E-02	.127	-1.511E-02
640.69nm_3	.847	.482	-8.579E-02	.162	-5.849E-03

648.26nm_3	.864	.435	2.124E-02	.205	-1.404E-02
655.83nm_3	.874	.369	9.519E-02	.263	-1.110E-03
663.41nm_3	.872	.249	.193	.336	1.659E-02
670.99nm_3	.843	.163	.240	.412	-7.049E-03
678.57nm_3	.811	.119	.255	.471	-4.778E-03
686.17nm_3	.847	.266	.230	.333	-2.451E-02
693.76nm_3	.815	.539	-7.136E-02	.131	9.989E-03
701.36nm_3	.703	.643	-.277	-6.716E-02	2.621E-03
708.97nm_3	.607	.696	-.337	-.164	-3.343E-03
716.58nm_3	.495	.749	-.374	-.218	5.372E-03
724.20nm_3	.292	.830	-.382	-.258	2.357E-03
731.82nm_3	-4.717E-02	.897	-.338	-.239	9.375E-03
739.45nm_3	-.414	.863	-.215	-.141	5.419E-03
747.08nm_3	-.629	.763	-9.741E-02	-4.179E-02	-1.914E-03
754.71nm_3	-.717	.691	-2.063E-02	1.682E-02	-1.864E-03
762.35nm_3	-.746	.659	1.804E-02	4.692E-02	4.403E-03
770.00nm_3	-.757	.644	2.549E-02	7.311E-02	-2.281E-03
777.65nm_3	-.764	.637	3.216E-02	7.316E-02	-7.790E-03
785.30nm_3	-.766	.635	3.586E-02	7.655E-02	-4.211E-03
792.96nm_3	-.766	.635	3.455E-02	7.587E-02	7.313E-05
800.62nm_3	-.768	.633	3.084E-02	8.041E-02	-5.004E-03
808.29nm_3	-.766	.635	3.210E-02	7.917E-02	-3.636E-03
815.96nm_3	-.765	.637	3.719E-02	7.558E-02	-1.526E-03
823.64nm_3	-.763	.639	4.011E-02	7.927E-02	-4.126E-03
831.32nm_3	-.765	.636	3.884E-02	8.176E-02	-7.777E-03
839.01nm_3	-.768	.633	3.887E-02	8.257E-02	-1.427E-03
846.70nm_3	-.763	.638	4.293E-02	7.588E-02	-7.184E-03
854.39nm_3	-.760	.642	3.214E-02	8.383E-02	1.716E-03
862.09nm_3	-.765	.635	4.384E-02	8.298E-02	-7.178E-03
869.80nm_3	-.762	.639	4.189E-02	8.001E-02	-6.460E-03
877.51nm_3	-.757	.643	4.403E-02	8.050E-02	-1.017E-02
885.22nm_3	-.753	.647	5.215E-02	8.168E-02	-4.275E-03
892.93nm_3	-.760	.640	6.174E-02	6.207E-02	-7.447E-03
900.66nm_3	-.749	.648	5.422E-02	7.758E-02	-7.063E-03
908.38nm_3	-.750	.647	6.460E-02	6.304E-02	-1.143E-02
916.11nm_3	-.741	.653	8.003E-02	5.261E-02	1.367E-03
923.84nm_3	-.744	.647	7.763E-02	5.298E-02	-1.465E-02
931.58nm_3	-.665	.676	.138	-9.241E-03	-3.061E-02

939.33nm_3	-.597	.655	.312	-2.564E-02	6.341E-02
Extraction Method: Principal Component Analysis.					
a 5 components extracted.					

Appendix B: Results of SMLR analysis with 71 input variables

Variables Entered/Removed(a)			
Model	Variables Entered	Variables Removed	Method
1	701.36nm_3		Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	900.66nm_3		Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
3	655.83nm_3		Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
4	931.58nm_3		Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
5	482.98nm_3		Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
6	602.93nm_3		Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
7		701.36nm_3	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
a Dependent Variable: Yield in kg/ha			

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.711(a)	.505	.502	1215.5807
2	.759(b)	.576	.570	1129.7521
3	.795(c)	.632	.624	1056.1883
4	.814(d)	.662	.652	1015.2827
5	.820(e)	.672	.660	1004.2076
6	.826(f)	.682	.668	991.6976
7	.825(g)	.681	.669	989.9863
a Predictors: (Constant), 701.36nm_3				
b Predictors: (Constant), 701.36nm_3, 900.66nm_3				
c Predictors: (Constant), 701.36nm_3, 900.66nm_3, 655.83nm_3				
d Predictors: (Constant), 701.36nm_3, 900.66nm_3, 655.83nm_3, 931.58nm_3				
e Predictors: (Constant), 701.36nm_3, 900.66nm_3, 655.83nm_3, 931.58nm_3, 482.98nm_3				
f Predictors: (Constant), 701.36nm_3, 900.66nm_3, 655.83nm_3, 931.58nm_3, 482.98nm_3, 602.93nm_3				
g Predictors: (Constant), 900.66nm_3, 655.83nm_3, 931.58nm_3, 482.98nm_3, 602.93nm_3				

ANOVA(h)						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	214164318.583	1	214164318.583	144.937	.000(a)

	Residual	209824368.350	142	1477636.397		
	Total	423988686.933	143			
2	Regression	244024760.279	2	122012380.139	95.596	.000(b)
	Residual	179963926.654	141	1276339.905		
	Total	423988686.933	143			
3	Regression	267813968.435	3	89271322.812	80.026	.000(c)
	Residual	156174718.498	140	1115533.704		
	Total	423988686.933	143			
4	Regression	280707617.828	4	70176904.457	68.080	.000(d)
	Residual	143281069.105	139	1030799.058		
	Total	423988686.933	143			
5	Regression	284824958.988	5	56964991.798	56.489	.000(e)
	Residual	139163727.945	138	1008432.811		
	Total	423988686.933	143			
6	Regression	289254095.320	6	48209015.887	49.020	.000(f)
	Residual	134734591.613	137	983464.172		
	Total	423988686.933	143			
7	Regression	288738632.435	5	57747726.487	58.922	.000(g)
	Residual	135250054.498	138	980072.859		
	Total	423988686.933	143			
a Predictors: (Constant), 701.36nm_3						
b Predictors: (Constant), 701.36nm_3, 900.66nm_3						
c Predictors: (Constant), 701.36nm_3, 900.66nm_3, 655.83nm_3						
d Predictors: (Constant), 701.36nm_3, 900.66nm_3, 655.83nm_3, 931.58nm_3						
e Predictors: (Constant), 701.36nm_3, 900.66nm_3, 655.83nm_3, 931.58nm_3, 482.98nm_3						
f Predictors: (Constant), 701.36nm_3, 900.66nm_3, 655.83nm_3, 931.58nm_3, 482.98nm_3, 602.93nm_3						
g Predictors: (Constant), 900.66nm_3, 655.83nm_3, 931.58nm_3, 482.98nm_3, 602.93nm_3						
h Dependent Variable: Yield in kg/ha						

Appendix C: Results of SMLR analysis with five principal components

Variables Entered/Removed(a)			
Model	Variables Entered	Variables Removed	Method
1	\$F-PCA-Default-1		Stepwise (Criteria: Probability-of-F-to-enter \leq .050, Probability-of-F-to-remove \geq .100).
2	\$F-PCA-Default-3		Stepwise (Criteria: Probability-of-F-to-enter \leq .050, Probability-of-F-to-remove \geq .100).
3	\$F-PCA-Default-2		Stepwise (Criteria: Probability-of-F-to-enter \leq .050, Probability-of-F-to-remove \geq .100).
4	\$F-PCA-Default-4		Stepwise (Criteria: Probability-of-F-to-enter \leq .050, Probability-of-F-to-remove \geq .100).
a Dependent Variable: Yield in kg/ha			

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.607(a)	.369	.364	1372.7253
2	.703(b)	.494	.487	1233.0257
3	.751(c)	.564	.555	1148.9596
4	.788(d)	.622	.611	1074.4796
a Predictors: (Constant), \$F-PCA-Default-1				
b Predictors: (Constant), \$F-PCA-Default-1, \$F-PCA-Default-3				
c Predictors: (Constant), \$F-PCA-Default-1, \$F-PCA-Default-3, \$F-PCA-Default-2				
d Predictors: (Constant), \$F-PCA-Default-1, \$F-PCA-Default-3, \$F-PCA-Default-2, \$F-PCA-Default-4				

ANOVA(e)						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	156407458.068	1	156407458.068	83.002	.000(a)
	Residual	267581228.865	142	1884374.851		
	Total	423988686.933	143			
2	Regression	209618997.421	2	104809498.711	68.938	.000(b)
	Residual	214369689.512	141	1520352.408		
	Total	423988686.933	143			
3	Regression	239173530.316	3	79724510.105	60.392	.000(c)
	Residual	184815156.617	140	1320108.262		
	Total	423988686.933	143			
4	Regression	263512304.181	4	65878076.045	57.062	.000(d)
	Residual	160476382.752	139	1154506.351		
	Total	423988686.933	143			

a Predictors: (Constant), \$F-PCA-Default-1
b Predictors: (Constant), \$F-PCA-Default-1, \$F-PCA-Default-3
c Predictors: (Constant), \$F-PCA-Default-1, \$F-PCA-Default-3, \$F-PCA-Default-2
d Predictors: (Constant), \$F-PCA-Default-1, \$F-PCA-Default-3, \$F-PCA-Default-2, \$F-PCA-Default-4
e Dependent Variable: Yield in kg/ha