INTEGRATING REMOTE SENSING, GEOGRAPHIC INFORMATION SYSTEM AND MODELING FOR ESTIMATING CROP YIELD

by

LUIS ALONSO SALAZAR

A dissertation submitted to the Graduate Faculty in Electrical Engineering in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2007

UMI Number: 3283213

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

This manuscript has been read and accepted for the Graduate Faculty in Electrical Engineering in satisfaction of the dissertation requirements for the degree of Doctor of Philosophy.

_____	_____
Date	Leonid Roytman, Ph. D. Chair of Examining Committee
_____	_____
Date	Mumtaz K. Kassir, Ph. D. Executive Officer

Supervisory Committee:

Reza Khanbilvardi, Ph. D.
Irina Gladkova, Ph. D.
Michael Grossberg, Ph. D.

Outside readers:

Felix Kogan, Ph. D.

THE CITY UNIVERSITY OF NEW YORK

Abstract

INTEGRATING REMOTE SENSING, GEOGRAPHIC INFORMATION SYSTEM AND MODELING FOR ESTIMATING CROP YIELD

by

LUIS ALONSO SALAZAR

Advisor: ***Professor Leonid Roytman***

This thesis explores various aspects of the use of remote sensing, geographic information system and digital signal processing technologies for broad-scale estimation of crop yield in Kansas.

Recent dry and drought years in the Great Plains have emphasized the need for new sources of timely, objective and quantitative information on crop conditions. Crop growth monitoring and yield estimation can provide important information for government agencies, commodity traders and producers in planning harvest, storage, transportation and marketing activities. The sooner this information is available the lower the economic risk translating into greater efficiency and increased return on investments.

Weather data is normally used when crop yield is forecasted. Such information, to provide adequate detail for effective predictions, is typically feasible only on small research sites due to expensive and time-consuming collections. In

order for crop assessment systems to be economical, more efficient methods for data collection and analysis are necessary.

The purpose of this research is to use satellite data which provides 50 times more spatial information about the environment than the weather station network in a short amount of time at a relatively low cost. Specifically, we are going to use Advanced Very High Resolution Radiometer (AVHRR) based vegetation health (VH) indices as proxies for characterization of weather conditions.

Acknowledgments

The road from personal curiosity to academic understanding can never be traveled alone. There are countless people over the last years who have helped me along the way. Some have been family members, some have been teachers, others have been friends and yet others have been strangers whose comments in an informal meeting might have made something open up in my mind.

All the people I need to thank are too numerous to mention but I would like to start by thanking Dr Felix Kogan who introduced me to remote sensing. In all likelihood, had he not done that I would have traveled an entirely different path and one that, without a doubt, would not have been as rich or rewarding. In addition, I want to thank Dr. Kogan for giving me a conceptual understanding of satellite remote sensing technology and for his well-structured talks in which I grasped many complex concepts. Rochelle Lapidus, my friend and teacher, for having boundless patience when helping me with my own projects. The faculty of City College of New York whose diverse set of expertise have helped me to expand my understanding of engineering. Leonid Roytman and Reza Kabilbardy who opened doors for me that would have remained closed without their help. The faculty of Columbia University for providing me with statistical skills that have proved to be indispensable in every aspect of my research. The faculty of Hunter College for providing me with the geographic information system knowledge fundamental for this project especially Dr. Jack Eichenbaum. I give many thanks to my dissertation committee, Prof. Leonid

Roytman, Mentor, Dept. of Electrical Engineering, The City College, Prof. Reza M. Khanbilvardi, Dept. of Civil Engineering, The City College, Prof. Irina Gladkova, Dept. of Computer Science, The City College, Prof. Michael Grossberg, Dept. of Computer Science, The City College, for making numerous helpful comments.

Very special thanks to all my friends who have been there for me over the years. Of course, I cannot mention them all by name; however, the acknowledgements would not be complete without thanking Michael Lapidus who taught me what it means to give a professional presentation.

Table of Contents

List of Tables

List of Figures

1. Introduction

In six of the last seven years world grain production has fallen short of consumption. In 2006, harvest of 1,967 million tons fell short of the estimated consumption of 2,040 million tons by some 73 million tons (USDA 2006) (Figure 1.1).

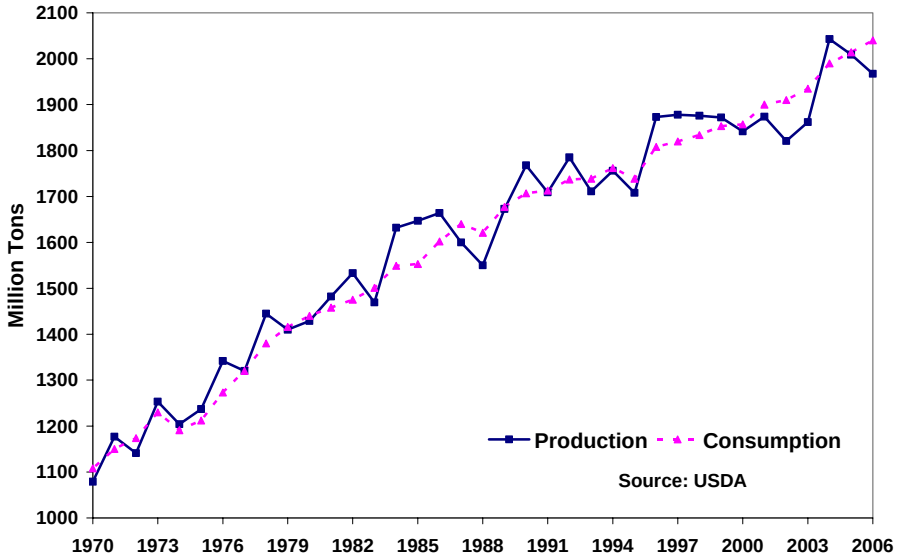

Figure 1.1: World crop production and consumption, 1970-2006

As a result, world carryover stocks of grain have been drawn down to 57 days of consumption, the lowest level in 34 years (Figure 1.2). World carryover stocks of grain, the amount stored in the bin when the next harvest begins, are the most basic measure of food security. Whenever crop stocks drop below 60 days of consumption, prices begin to rise. The last time they were this low wheat and rice prices doubled (Figure 1.3).

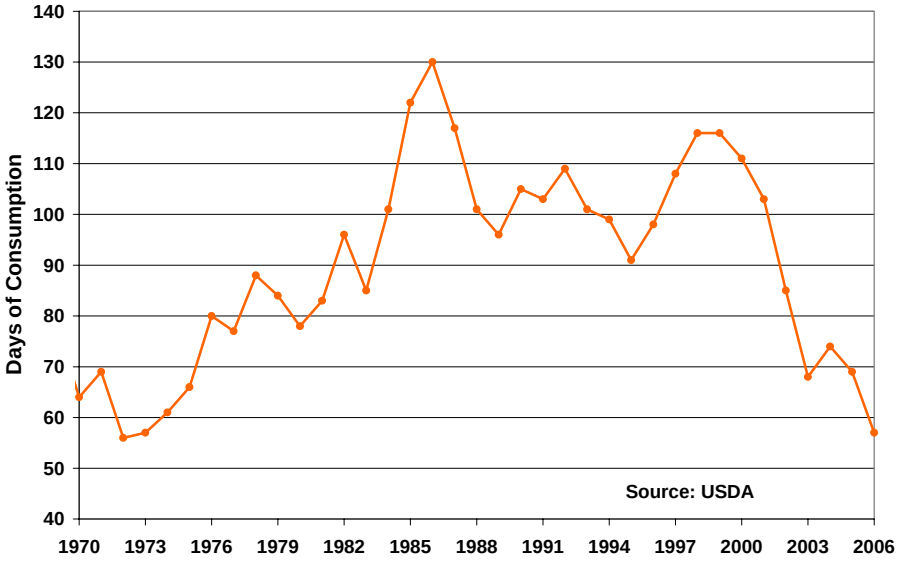

Figure 1.2: World crop stocks as days of consumption, 1970-2006

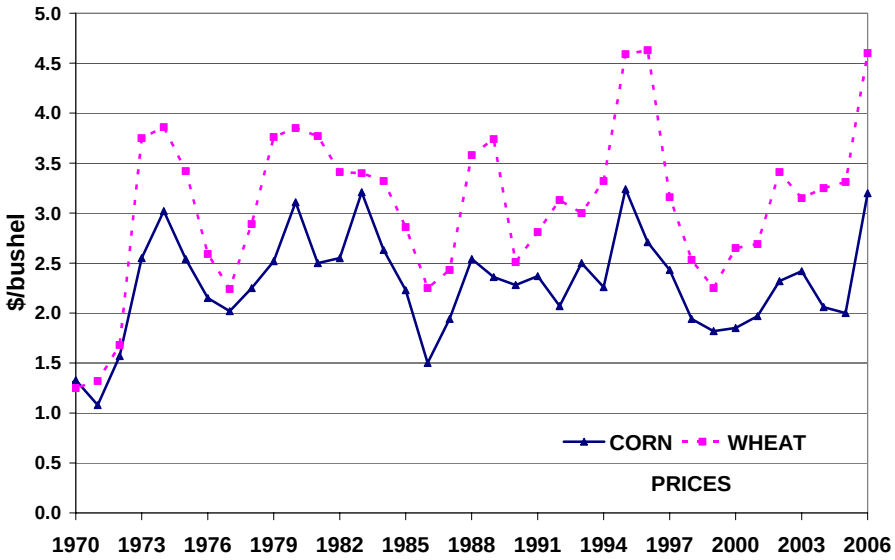

Figure 1.3: Corn and wheat prices, 1970-2006

In addition, if harvest is sharply reduced by heat or drought, prices could rise far away the projected values. Each year the world's farmers must try to feed an additional 70 million people. Farmers are facing a record growth in the demand for crops at a time when agricultural technology used to raise crop yields is shrinking, when underground water reserves are being depleted and when rising temperatures threaten to reduce future harvests. Perhaps the most dangerous threat to future world food security is the rise in temperature. Among crop ecologists there is now a general consensus that for each temperature rise of 1 degree Celsius above the historical average, during the growing season, we can expect a 10% reduction in crop yields.

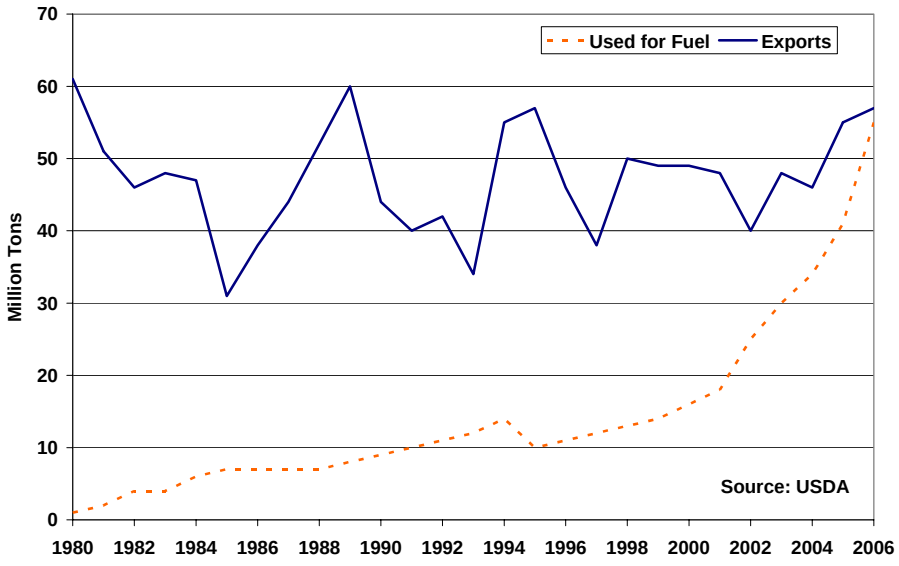

Figure 1.4: U.S. corn used for fuel ethanol and for export, 1980-2006

Approximately 60% of the world crop production is consumed as food, 36% as feed and 3% as fuel. While the use of crops for food and feed grows by about 1% per

year, the use for fuel is growing by over 20% per year. With the current price of ethanol double its cost of production, the conversion of agricultural commodities into fuel for cars has become hugely profitable (Figure 1.4). Increasing competition for the U.S. corn crop is already driving up prices. In some corn-growing states such as Iowa, Indiana and South Dakota, completion of the ethanol plants under construction and those projected to be built means distillery requirements would take virtually all the states' corn production.

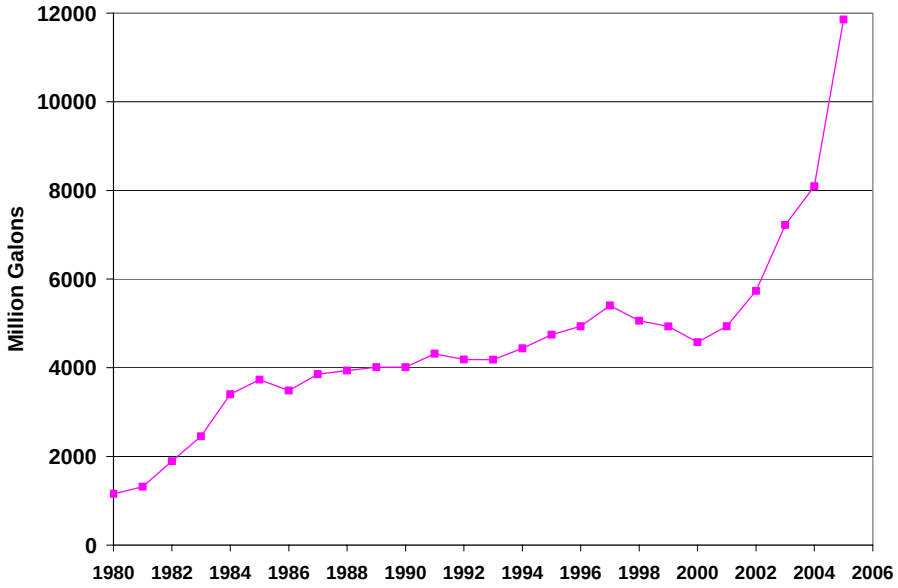

Figure 1.5: World ethanol production, 1980-2005

Corn importers like Japan, Egypt and Mexico are also worried that the likely reduction in U.S. corn exports, which are approximately 70% of the world total, will disrupt their livestock and poultry industries. In the United States, most corn is

consumed indirectly. The milk, eggs, cheese, chicken, ham, ground beef, ice cream and yogurt are all produced with corn. The price of every of these products is affected by the price of corn. Wheat and corn prices have climbed by a third or more over the past several months (USDA 2007). Corn and wheat futures are both trading at 10-year highs. With corn stocks at the lowest level on record and demand soaring, corn prices appear headed for historic highs. Wheat and rice prices will likely follow corn prices upward. If crop prices do climb to all-time highs, food riots and political instability in lower-income countries that import grain could disrupt global economic progress.

Therefore, timely knowledge concerning the condition, production and availability of our agricultural resources is critical to the decision making process at all levels of the food distribution chain. Accurate methods of acquiring crop condition information in a timely manner are critical in the global grain–market. Any technique that helps reduce uncertainties with respect to marketing decisions is therefore important

Crop yield assessment models are valuable in aiding the management process during a growing season. A farmer is prepared in advance when to expect certain physiological occurrences, reducing the amount of guesswork in managing a crop. Most often, the required information includes weather data, soil physical properties and soil fertility data, each requiring multiple and complex collection methods. This information is typically possible only on small research sites due to expensive and

time-consuming collections. In order for crop yield assessment models to be economical, more efficient methods for data collection and analysis are necessary.

Since satellites cover large areas in a short amount of time at a relatively low cost, remote sensing serves as a potential alternative to common data collection methods. In the past, remote sensing has been used to identify correlation between crops and their spectral reflectance. Data required by crop models to perform predictions can be obtained through remote sensing if the correlation between spectral reflectance and crop parameters is present. Timely knowledge concerning the condition, production and availability of our agricultural resources is critical to the decision making process at all levels of the food distribution chain.

1.1 Importance of Crop Yield Forecasts

Accurate methods of acquiring crop condition information in a timely manner are critical in the global grain–market. Any technique that helps reduce uncertainties with respect to marketing decisions is therefore important. Statistical information on acreage, production, stocks and prices is essential for managing programs in such areas as consumer protection, conservation and environmental quality, trade and education. Moreover, the regular updating of information helps to ensure a systematic flow of goods and services among agriculture's producing, processing and marketing sectors. Reliable, timely and detailed crop statistics help to maintain a stable

economic climate and minimize the uncertainties and risks associated with the production, marketing and distribution of commodities.

Farmers and ranchers rely on statistical reports to decide on specific production plans, such as how much of a crop to plant, how many cattle to raise and when to sell. In addition, these reports are used by the transportation sector, warehouse and storage companies, banks, commodity traders and food processors. Those who provide farmers with seeds, equipment, chemicals and other goods and services use these reports when planning their marketing strategies. The degree of accuracy of the statistical reports affects productivity. The more accurate the crop yield statistics, the wiser the economic decisions and the higher may be the profits.

1.2 Common Forecasting Techniques

Traditionally, yield estimations are made through agro-meteorological, meteorological modeling or by compiling survey information provided throughout the growing season. Yield estimates derived from agro-meteorological models use soil properties and daily weather data as inputs to simulate various plant processes at a field level (Wiegand et al. 1979, Wiegand et al. 1986, Wiegand and Richardson 1990). At this scale, agro-meteorological crop yield modeling provides useful results. However, at regional scales these models are of limited practical use because of spatial differences in soil characteristics and crop growth determining factors such as

nutrition levels, plant disease, herbicide and insecticide use, crop type, and crop variety, which would make informational and analytical costs excessive.

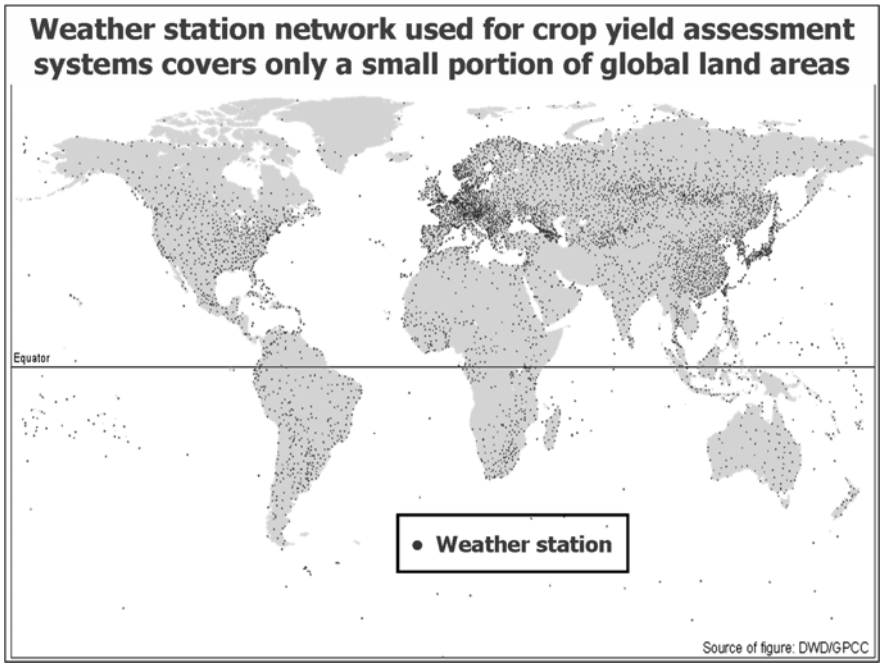

Figure 1.6: The weather station network used for crop yield assessment systems covers only a small portion of global land areas

Additionally, Rudorff and Batista (1991) indicated that, at a regional level, agro-meteorological models are unable to completely simulate the different crop growing conditions that result from differences in climate, local weather conditions, and land management practices. The scale of applicability of agro-meteorological models is getting larger through the integration of remote sensing data. For instance, Doraiswamy et al. (2003) developed a method using AVHRR NDVI data as proxy

inputs to an agro-meteorological model in estimating spring wheat yields at county and sub-county scales in the U.S. state of North Dakota.

Another approach that is very popular is to use statistical models that analyze the relationship between one or more independent climatological variables such as precipitation, temperature, moisture, most often precipitation and a dependent crop response variable, which is normally crop yield. The problem with this methodology is that we do not have a reliable source of climatological information. Reliable, readily available gauge measurements cover only a small portion of global land areas (Figure 1.6).

In Kansas, where we developed our modes, the weather station network used in these assessment systems covers only a very small portion of land surface and some counties do not have weather stations. In addition, some of the weather data sets are incomplete. In addition, weather stations tend to be located in urban settings. Measures of climate variables across the landscape, especially where farms are located, are difficult to acquire (Figure 1.7).

In the U.S. yield estimates are derived by compiling survey information provided throughout the growing season. Each month, the U.S. Department of Agriculture (USDA) publishes information about crop production. The National Agricultural Statistics Service (NASS) agency is responsible for preparing these statistics.

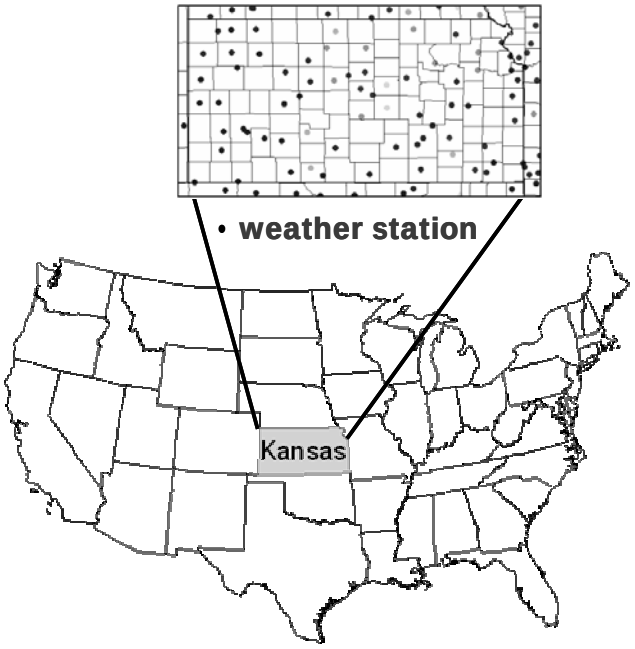

Figure 1.7: The weather station networks used for crop yield assessment systems in Kansas, U.S.

NASS uses two major survey techniques for crop yield estimation and forecasting:

1. Monthly Yield Survey
2. Objective Yield Survey

The first is the Monthly Yield Survey of farm operators. A stratified sample of farm operators is selected from a list frame, and then multiple interviews are conducted to acquire the farmer's projection of crop yield and the final yield after harvest as well. The second is the Objective Yield Survey. A self weighting sample of small plots are laid out in a statistical sample of fields and several types of

observations, counts, measurements and weighing are done throughout the crop season.

For example, for corn in the first month, plots are laid out and plant counts are collected. As the season progresses, other measurements are taken such as number of ears, the length and diameter of the ears, weights of ears and several laboratory measurements such as corn moisture content. After harvest, sample plots are gleaned for any harvest lost grain as well. Yield forecasting models are used during the season utilizing the above information and a final yield is derived using the grain weight and laboratory data and ear population. All this information of crop condition and yield information is quantified. An expert panel of NASS statisticians meets monthly to set a yield forecast or final estimate, using all of the above sources and data from multiple years as inputs. Because of the high cost of in-field observations, objective yield surveys are normally used only in major producing states.

The Agricultural Statistics Board composed of statisticians permanently assigned and rotating State office statisticians review indications from both surveys and the State office recommendations to set official forecasts and estimates at the State, regional and national levels. There currently is no direct quantitative use of weather data, soils data or satellite data in the operational program to forecast or estimate crop yields. In some cases weather variables, such as early season rainfall at the State level, were tried in the operating program, but later dropped. The objective yield data and the farmer reported data indirectly reflect, to some degree, weather and

soil effects as well as plant diseases, plant pests and excessive weed growth (Vogel and Bange 1999). A detailed document of USDA/NASS crop yield estimation methods using agricultural statistics is available on the Internet (http://www.usda.gov/nass/).

This methodology is time consuming and very costly. Crop yield assessment systems should have the qualities of objectivity, reliability, timeliness, adequacy of coverage, efficiency and effectiveness. Crop yield assessment systems in many important agricultural countries do not meet these standards. In the U.S. the complete procedure used by the USDA to forecast crop yield is secret and official estimates are released several weeks after harvest has been completed.

In order for crop yield assessment systems to be economical and timely, more efficient methods for data collection and analysis are necessary.

1.3 Remote Sensing as an Alternative to Common Data Collection Methods

The role of satellite data as part of a crop yield estimation system is a natural alternative because of the ability of satellites to provide relatively economical, consistent and repeated coverage over large areas. These characteristics of satellites allow collecting data useful for timely estimation of crop conditions throughout an entire growing season covering either important agricultural production regions or remote regions where accurate information is normally unavailable. Even though we have little control over the impacts of weather on crops, with remote sensing

technology we have the ability to monitor and assess the impacts that weather is having on crops. This information is critical to reducing economic risk. The sooner this information is available, the lower the economic risk translating into greater efficiency and increased return on investments.

Tremendous advances in remote sensing technology and computing power over the last two decades are now providing scientists with the opportunity to investigate, measure and model environmental patterns and processes with increasing confidence. Remote sensing of the Earth is playing an increasing role in understanding the natural environment and its inherent physical, biological and chemical processes.

The uses of remote sensing already represent a very active field of research and application in agriculture. In Europe the MARS (Monitoring of Agriculture by Remote Sensing) Project of the Joint Research Centre has taken a leading role in such development (Csornai et al. 2002, ITA 2002). In the U.S. the United States Department of Agriculture (USDA) National Agricultural Statistics Service (NASS) uses satellite data to enhance its program of crop acreage estimates. This program is used for: construction of the nation's area sampling frame for agricultural statistics, improvement of the statistical precision of crop acreage estimate indicators, especially at the county level and application of GIS based Cropland Data Layer used for watershed monitoring, soil utilization analysis, agribusiness planning, crop rotation

practice analysis, animal habitat monitoring and prairie water pothole monitoring (Craig 2001, Mueller et al. 2003).

1.4 Related Research

In the past 25 years, many scientists have utilized remote sensing techniques to assess agricultural yield, production and crop conditions. Wiegand et al. (1979) and Tucker et al. (1980) first identified a relationship between the NDVI and crop yield using experimental fields and ground-based spectral radiometer measurements. Final grain yields were found to be highly correlated with accumulated NDVI (a summation of NDVI between two dates) around the time of maximum greenness (Tucker et al. 1980). In another experimental study, Das et al. (1993) used remotely sensed data to predict wheat yield 85-110 days before harvest in India. These early experiments identified relationships between NDVI and crop response, paving the way for crop yield estimation using satellite imagery.

Rudorff and Batista (1991) used NDVI values as inputs into an agro-meteorological model to explain nearly 70% of the variation in 1986 wheat yields in Brazil. Rasmussen (1992) used 34 AVHRR images of Burkina Faso, Africa, for a single growing season to estimate millet yield. Using accumulated NDVI and statistical regression techniques, he found strong correlations between accumulated NDVI and yield. Potdar (1993) estimated sorghum yield in India using 14 AVHRR images from the same growing season. He was able to forecast actual yield at an

accuracy of ±15% up to 45 days before harvest. Hayes and Decker (Hayas and Decker 1996) used AVHRR NDVI data to explain more than 50% of the variation in corn yields in the United States Corn Belt. Each of these studies found positive relationships between crop yield and NDVI, but the strength of the relationships depended upon the amount and quality of the remote sensing data used.

Some studies have used large, multi-year AVHRR NDVI data sets. Maselli et al. (1992) found strong correlations between NDVI and final crop yields in the Sahel region of Niger using 3 years of AVHRR imagery (60 images). In India, Gupta et al. (1993) used 3 years of AVHRR data to estimate wheat yields within ±5% up to 75 days before harvest. The success of this study was dependent on the fact that over 80% of the study area was covered with wheat. In Greece, actual harvested rice yields were predicted with an accuracy of ±10%, and wheat yields were predicted with an accuracy of ±12% at the time of maximum greenness, using two years of AVHRR imagery (Quarmby et al. 1993). Groten (Groten 1993) predicted crop yield with a ±15% estimation error 60 days before harvest in Burkina Faso using regression techniques and 5 years of AVHRR NDVI data (41 images). Doraiswamy and Cook (1995) used 3 years of AVHRR NDVI imagery to assess spring wheat yields in North and South Dakota in the U.S. Maselli and Rembold (2001) used time series of annual crop yields and monthly NDVI to develop cropland masks for four Mediterranean African countries. They found that application of the derived cropland masks improved relationships between NDVI and final yields. Ferencz et al. (2004) found

yields of eight different crops in Hungary to be highly correlated with optimized, weighted seasonal NDVI sums using 1-km AVHRR NDVI from 1996 to 2000. They used non-forest vegetation masks and a novel time series interpolation approach and actually obtained their best results when using a greenness index equivalent to the numerator of the NDVI formula (NIR-RED).

Furthermore, many researchers have found that crop conditions and yield estimation are improved through the inclusion of metrics that characterize crop development stage (Badhwar and Henderson 1981, Rasmussen 1992, Groten 1993, Quarmby et al. 1993, Kastens 1998, Lee et al. 1999). Ancillary data have been found useful as well. For example, Rasmussen (1997) found that soil type information improved the explanation of millet and ground nut yield variation using 3 years of AVHRR NDVI from the Peanut Basin in Senegal.

In addition, AVHRR–based vegetation health indices were found to be very useful for early drought detection and monitoring drought impacts on crop and pasture production around the world, including such major agricultural producers as China, Russia, Brazil, Argentina and Kazakhstan (Dabrowska-Zielinska et al. 2002, Kogan 2002, Liu and Kogan 2002, Kogan et al. 2003, Domenikiotis et al. 2004, Kogan et al. 2005).

1.5 Goals and Objectives

In this research, we incorporate satellite data into a statistical model to make timely estimation of crop yield at State, Crop Reporting District (CRD) and County levels. We use AVHRR–based Vegetation Health (VH) indices (VCI, TCI) as proxy for modeling crop yield and for early warning of drought related losses of agricultural production in the U.S.

Furthermore, we investigate the entire procedure associated with including satellite data in a crop yield estimation model. First, techniques for the development of timely and easily applicable methods that include satellite data within a model are established. After these methods are developed, it is determined that the resulting models provide enough beneficial information to assist with the estimation of crop yield.

Since few image masking techniques have been used, likely due to the inherent complexities underlying this phase in any remote sensing-based yield forecasting methodology, one important objective of this research is to use historical yield information and historical time series AVHRR-based VH indices imagery to devise a robust statistical procedure for obtaining early crop yield forecasts, with particular emphasis on image masking or classification. The techniques described in this thesis can be applied to any region and crop pair that possesses sufficient historical yield information and corresponding time series VH indices imagery. Since few meaningful crop phenology metrics can be accurately derived at early points in the growing

season, our research does not attempt to use this information. In addition, no ancillary information is used, to prevent dependence on the availability of such data.

2. Background Information

This chapter introduces the background information and basic concepts of optical remote sensing. Optical remote sensing makes use of visible, near infrared and short–wave–infrared sensors (0.4–1.8 μm) to form images of the earth's surface by detecting the solar radiation reflected from targets on the ground. Different materials reflect and absorb differently at different wavelengths. Thus, the targets can be differentiated by their spectral reflectance signatures in the remote sensing images. This chapter consists of three parts. The first part (Sections 2.1) discuses the physical principles of remote sensing. The second part (Section 2.2) describes NOAA satellites characteristics and part three (Section 2.3) introduces the basic concepts of satellite data processing and presents calibration results for AVHRR data.

Section 2.1 defines some basic physical concepts and introduces some background on remote sensing. Section 2.1.1 deals with issues related to the remote sensing of vegetation. The section starts with a discussion on how leaf structure and chlorophyll affect absorption and reflection of electromagnetic radiation. Then, the reflective properties of canopies are described. The section finishes with an overview of the most commonly used remote sensing technique for vegetation monitoring–vegetation indices (VI). We finish the discussion about vegetation indices by listing those most frequently encountered in the literature

In section 2.2, we give a description of NOAA satellites. NOAA operational weather satellite system is composed of two types of satellites: geostationary operational environmental satellites (GOES) for short–range warning and "now-casting" and polar-orbiting satellites for longer-term forecasting. Both types of satellite are necessary for providing a complete global weather monitoring system. Section 2.3 introduces some background information, including the importance of sensor calibration, basic concepts and principles. Section 2.3.1 reviews the process of passing from the digital number (DN) provided by the satellite instrument to a reflectance value. Two problems must be solved: (1) the calibration factors of the DN into a physical value such as radiance at satellite level, and (2) the computation of a reflectance at ground level. Section 2.3.1.2 presents various post-launch calibration methods that are currently being used. Section 2.3.1.3 shows calibration results for NOAA-AVHRR, which is essential for use of these datasets in quantitative applications.

2.1 Physical Principles of Remote Sensing

Remote sensing is the science concerned with acquiring information about an object without physically interfering with that object. Any remote sensing system requires a source of energy, a target and a sensor for recording the interactions of electromagnetic radiation with that object. On striking an object, energy may be reflected, absorbed or transmitted. Typically, the reflected energy is measured using a

specially designed sensor system. Measurement of the amount of energy reflected allows some inferences to be made about the nature of the target.

The typical reflectance characteristics of four major earth surface targets are shown in Figure 2.1. The difference between living and dead vegetation is clear in both the red and near infrared parts of the spectrum. Absorption of red light by chlorophyll pigments causes a lower reflectance for healthy vegetation in that part of the spectrum. Due to scattering of light by leaf internal tissues and the low absorption by pigments, healthy vegetation reflects more near infrared light than dead or stressed vegetation (Curran 1984, Jensen 2000, Liang 2004). Water has a typically low reflectance in all wavelengths, and the reflectance is inversely related to wavelength. Soil reflectance shows a clear linear relationship with wavelength (Curran 1984, Jensen 2000, Liang 2004).

Reflectance measurements can be used to differentiate between various earth surface targets, and in some cases infer differences in vegetation condition from the spectral measurements (Tucker et al. 1979, Tucker et al. 1983, Tucker et al. 1985, Choudhury and Tucker 1987). Reflectance measurements were found to be very useful for early drought detection and monitoring drought impacts on crop and pasture production around the world (Hayas and Decker 1996, Dabrowska-Zielinska et al. 2002, Kogan 2002, Liu and Kogan 2002, Kogan et al. 2003, Domenikiotis et al. 2004, Kogan et al. 2005).

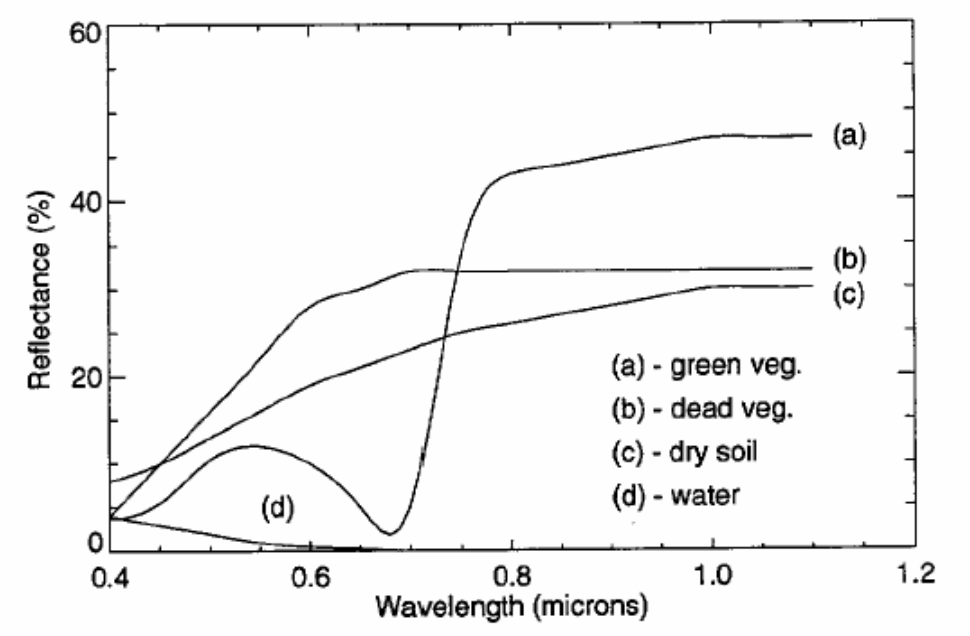

Figure 2.1: Reflectance characteristics of four major earth surface targets

The measurements shown in Figure 2.1 are derived from laboratory and field based studies. For remote sensing from satellite platforms, the atmosphere modifies the signal recorded by the satellite as depicted in Figure 2.2. This shows that the total radiance recorded at the satellite (L) is given by:

$$L = P + D + I \qquad \textbf{(2.1)}$$

The direct radiance component D is often described as "top of canopy radiance", and the total radiance received at the sensor L is often called "top of atmosphere radiance" (TOA). Field based measurements which are unaffected by the atmosphere, are equivalent to D.

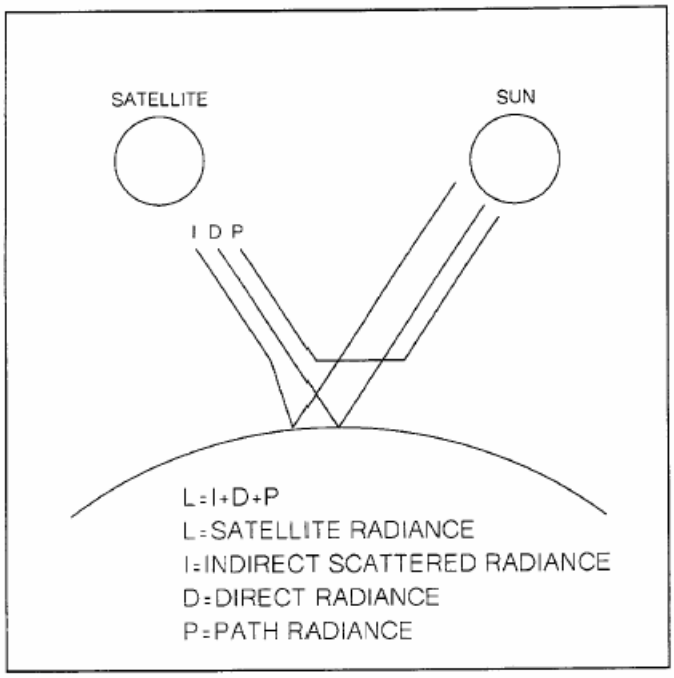

Figure 2.2: Atmospheric component in remote sensing

2.1.1 Principles of Remote Sensing of Vegetation

Approximately 70% of the land surface is covered with vegetation. In addition, vegetation is one of the most important components of ecosystems. Knowledge about variations in vegetation, alteration in periodic biological phenomena that are correlated with climatic conditions (vegetation growth) cycles, and modifications in the plant physiology and morphology provide valuable information into the climatic, geologic and physiographic characteristics of an area (Jones et al. 1996). Scientists have devoted a significant amount of effort to develop sensors and digital image processing algorithms to extract vegetation information from remote sensing data (Frohn 1998, Didan and Huete 2004, Unsalan and Boyer 2004). Many of

the remote sensing techniques may be applied to a variety of vegetated landscapes, including (Fensholt et al. 2002, Lunetta and Lyon 2004, Feng et al. 2006, Lambin and Linderman 2006, Morisette et al. 2006):

1. crop growing,
2. forests,
3. rangeland,
4. wetland, and
5. urban vegetation

2.1.1.1 Structure of the Leaf

Remote sensing techniques for vegetation monitoring depend on the knowledge of the spectral properties of individual leaves and plants. These properties are understood by studying the leaf structure.

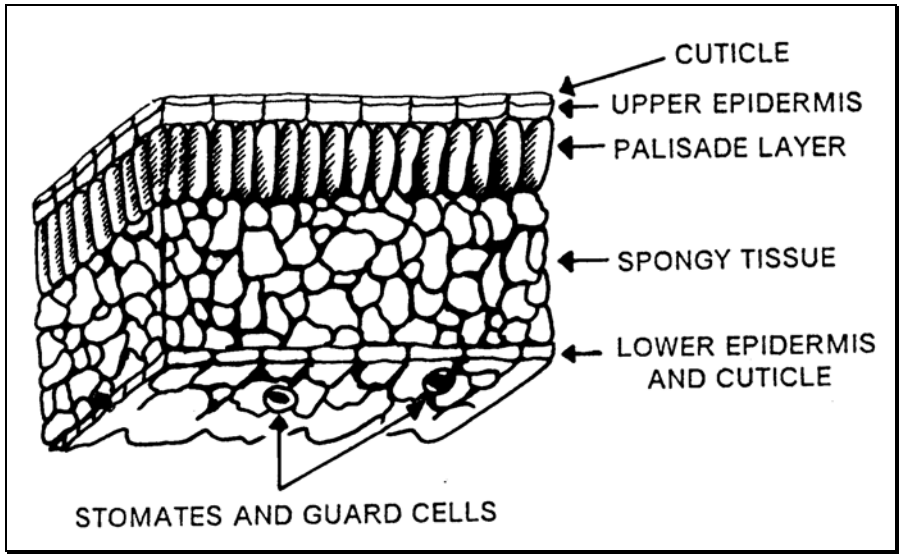

Figure 2.3: Diagram cross section of a typical leaf

The cross section of a typical leaf is shown in Figure 2.3. The upper surface of the leaf is covered with a translucent layer known as the cuticle, which serves to prevent uncontrolled moisture loss from the leaf interior. Below the cuticle lies the upper epidermis. The leaf underside has a similar structure, consisting of a layer of cuticle and the lower epidermis. However, it also includes very tiny openings called stomates. Each stomate is guarded by a pair of cells that open and close to control the movement of air through stomates into the leaf interior. The stomates also play a critical role in maintaining the moisture balance of the leaf (Campbell et al. 2004).

Below the upper epidermis is the palisade tissue. The vertically elongated cells of the palisade include chloroplasts, cells composed of chlorophyll and other pigments active in the process of photosynthesis. Below the palisade, tissue is the spongy, mesophyll tissue. It consists of irregularly shaped cells separated by interconnected openings. The mesophyll tissue utilizing its large surface area serves as the exchange medium for the oxygen and carbon dioxide required for photosynthesis and respiration (Campbell and Reece 2002, Campbell et al. 2004).

2.1.1.2 Photosynthesis

Chlorophyll is one of the most important biological compounds for life on Earth as a whole. It acts as a photoreceptor and catalyst for the conversion of sunlight into chemical energy, in the form of carbohydrates stored by plants. Light energy

entering the plant splits the water into oxygen and hydrogen. The photosynthetic process is described by the equation:

$$6CO_2 + 6H_2O + \text{light energy} \rightarrow C_6H_{12}O_6 + 6O_2 \qquad \textbf{(2.2)}$$

The role of the chlorophyll in the process of photosynthesis is to capture light (energy) which is needed for the above chemical reaction. Photosynthesis is an energy–storing process that takes place in leaves and other green parts of vegetation in the presence of light. The light energy is stored in a simple sugar molecule (glucose) that is produced from carbon dioxide (CO_2) present in the air and water (H_2O) absorbed by the plant through the root system. When the carbon dioxide and the water are combined and form a sugar molecule ($C_6H_{12}O_6$) in a chloroplast, oxygen gas (O_2) is released as a by–product. The oxygen diffuses out into the atmosphere. The photosynthetic process begins when sunlight strikes chloroplasts, small bodies in the leaf that contain a green substance called chlorophyll (Jensen 2000, Campbell et al. 2004).

Plants have adapted their internal and external structure to perform photosynthesis. This structure and its interaction with electromagnetic energy has a direct impact on how leaves and canopies appear spectrally when recorded using remote sensing technology.

2.1.1.3 Spectral Characteristics of Vegetation

A healthy green leaf intercepts incident radiant flux (Φ_i) directly from the sun or from diffuse skylight scattered onto the leaf. This incident electromagnetic energy interacts with the pigments, water and intercellular air spaces within the plant leaf. The amount of radiant flux reflected from the leaf (Φ_r), the amount of radiant flux absorbed by the leaf (Φ_α), and the amount of radiant flux transmitted through the leaf (Φ_τ) can be carefully measured as we apply the energy balance equation and attempt to keep track of what happens to all the incident energy. The general equation for the interaction of spectral (λ) radiant flux on and within the leaf is (Jensen 2000, Campbell et al. 2004):

$$\Phi_i = \Phi_r + \Phi_\alpha + \Phi_\tau \qquad \textbf{(2.3)}$$

Dividing each of the variables by the original incident radiant flux, Φ_i:

$$\frac{\Phi_i}{\Phi_i} = \frac{\Phi_r}{\Phi_i} + \frac{\Phi_\alpha}{\Phi_i} + \frac{\Phi_\tau}{\Phi_i} \qquad \textbf{(2.4)}$$

yields

$$i = r + \alpha + \tau \qquad \textbf{(2.5)}$$

where r is spectral hemispherical reflectance of the leaf, α is spectral hemispherical absorptance and τ is spectral hemispherical transmittance by the leaf. Most remote

sensing systems function in the 0.35-3.0 μm region measuring primarily reflected energy. Therefore, it is useful to think of this relationship as:

$$r = i - (\alpha + \tau) \qquad \textbf{(2.6)}$$

where the energy reflected from the plant leaf surface is equal to the incident energy minus the energy absorbed directly by the plant for photosynthetic or other purposes and the amount of energy transmitted directly through the leaf onto other leaves or to the ground beneath the canopy.

2.1.1.4 Dominant Factors Controlling Leaf Reflectance

Chlorophyll controls much of the spectral response of the living leaf in the visible part of the spectrum (0.4-0.75 μm) by absorbing strongly in the blue and red parts (Figure 2.4) by as much as 70-90%. This is the reason for the green appearance of live vegetation.

In the near infrared (NIR) part of the spectrum from 0.75 μm to about 1.35 μm, spectral characteristics of the leaf are primarily controlled by the inner structure of the leaf and not its pigments. The cuticle and epidermis are practically transparent to NIR radiation. The majority of NIR radiation passing through the upper epidermis layer is scattered by the mesophyll tissue. Very little of the scattered NIR radiation is absorbed internally, and up to about 60% is scattered outside the leaf (Jensen 2000, Campbell and Reece 2002).

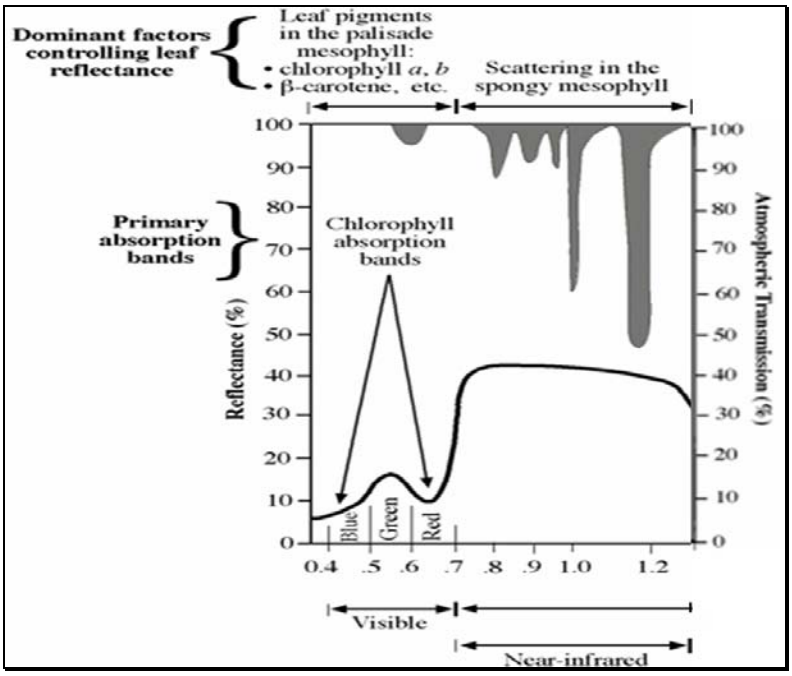

Figure 2.4: Dominant Factors Controlling Leaf Reflectance

At the edge of the visible spectrum, roughly between 0.68 μm and 0.75 μm, lies a noticeably sharp increase in reflectance caused by the decline of chlorophyll pigment absorption. This steep part of the leaf reflectance curve is known as the red edge. It is a dynamic feature of the leaf spectrum, as its location in the spectrum tends to shift as the plant matures (Campbell et al. 2004). This red-shift of the chlorophyll absorption edge was first reported by Gates et al. (1965) and later by Collins (1978).

Any plant during its lifetime is subject to some kind of stress by disease, pest attack, lack of moisture, etc. Stresses in general have an impact on the leaf spectrum altering both its visible and NIR parts. The changes in the NIR spectrum are better documented due to the high reflectivity than in the visible spectrum (Myneni et al.

1998). Since leaf reflectance in the NIR depends strongly on the structure of the complex cavities of the mesophyll tissue and internal leaf reflection, any effect causing destruction of this tissue or the internal structure of the leaf is detected as a decrease in NIR reflectivity. There is still much uncertainty about how different stresses to the plant affect the change in NIR reflectivity (Campbell and Reece 2002).

Some scientists maintain that moisture stress and natural maturing causes these cavities to collapse; others are of the opinion that NIR reflectance decreases mostly due to the changes in the cell walls of the mesophyll and palisade tissues rather than physical changes in the cavities themselves. In conclusion, any decrease of NIR reflectivity in subsequent measurements of the same plant is a clear indication of some kind of stress.

2.1.1.5 Vegetation Indices

Vegetation indices (VI) are dimensionless numbers that are generated by some combination of remote sensing bands and have some relationship to the amount of vegetation in a given image pixel. A vegetation index should (Xiang et al. 2003, Unsalan and Boyer 2004, Fensholt and Sandholt 2005, Jiang et al. 2006):

1. maximize sensitivity to plant biophysical parameters, preferably with a linear response in order that sensitivity is available for a wide range of vegetation conditions, and to facilitate validation and calibration of the index;
2. normalize or model external effects such as Sun angle, viewing angle, and the atmosphere for consistent spatial and temporal comparisons;

3. normalize internal effects such as canopy background variations, including topography (slope and aspect), soil variations, and differences in senesced or woody vegetation (non-photosynthetic canopy components); and
4. be coupled to some specific measurable biophysical parameter such as biomass, Leaf Area Index (LAI), or Absorbed Photo–synthetically Active Radiation (APAR) as part of the validation effort and quality control.

Vegetation indices are quantitative measures that attempt to relate remote sensing data to different biophysical properties of vegetation. These indices are calculated from either raw (as measured) or post-processed (reflectance, BRDF and/or atmospheric corrected) data. Vegetation indices commonly include observations of several spectral values that are manipulated to yield a single value that predicts vegetation status within a pixel. Usually vegetation indices attempt to predict biomass or leaf area index, but they have also been used for leaf water content, chlorophyll, percentage cover and other biophysical characteristics of vegetation (Purevdorj et al. 1996, Purevdorj et al. 1998).

The applications of vegetation indices can be loosely grouped in two categories: quantitative and qualitative. Quantitative applications attempt to use vegetation indices to predict biophysical properties of vegetation as accurately as possible. Such studies typically examine test plots during an entire growing season and compare the derived vegetation indices measured throughout the season with samples of vegetation taken at the same time. Qualitative applications use vegetation indices as “mapping” tools, to assist in image classification, separation of vegetated

from non–vegetated areas, distinguishing between different types and densities of vegetation, monitoring seasonal variations in vegetation abundance, etc.

Over the years, many types of vegetation indices have been developed. A major part of their derivation is dedicated to the treatment of the so-called soil line. The soil line is a hypothetical line in spectral space that describes the variation in the spectrum of bare soil in the image; this line is considered to be the line of zero vegetation. The spectral space which is commonly used to observe this line is the plot of NIR vs. RED channel reflectance because these channels have the strongest contrast in spectral signature of vegetation (Campbell and Reece 2002). The assumption inherent in all vegetation indices is that bare soil pixels in an image lie on or very near the soil line in the NIR-RED reflectance graph. At this point, the vegetation indices diverge depending on the orientation of lines of equal vegetation (isovegetation lines) (Asrar et al. 1988, Asrar and Murphy 1989, Myneni et al. 1992, Mynenil and Asrar 1992):

1. All isovegetation lines converge at a single point. The indices that use this assumption are the ratio-based indices, which measure the slope of the line between the point of convergence and the NIR-RED point of the pixel.

2. All isovegetation lines remain parallel to the soil line. This family of indices is known as perpendicular indices since they measure the perpendicular distance from the soil line to the NIR-RED point of the pixel.

We finish the discussion about vegetation indices by listing those most frequently encountered in the literature (in the following equations R and NIR refer to the red and near-infrared channel reflectance, respectively):

- Ratio Vegetation Index (RVI) (Jordan 1969)

$$RVI = NIR/R \quad \textbf{(2.7)}$$

- Normalized Difference Vegetation Index (NDVI) (Kriegler et al. 1969)

$$NDVI = NIR - R/NIR + R \quad \textbf{(2.8)}$$

- Infrared Percentage Vegetation Index (IPVI) (Crippen 1990)

$$IPVI = NDVI + 1/2 = NIR/NIR + R \quad \textbf{(2.9)}$$

- Perpendicular Vegetation Index (PVI) (Richardson and Wiegand 1977)

$$PVI = \sin(\alpha)\, NIR - \cos(\alpha)\, R \quad \textbf{(2.10)}$$

where α is the angle between the soil line and the NIR axis.

- Weighted Difference Vegetation Index (WDVI) (Clevers 1988)

$$WDVI = NIR - aR \quad \textbf{(2.11)}$$

where a is the slope of the soil line in the NIR-R plot.

- Soil Adjusted Vegetation Index (SAVI) (Huete 1988)

$$SAVI = (NIR - R/NIR + R + L)(1 + L) \quad \textbf{(2.12)}$$

where L is an adjustment factor, ranging from 0 (very high vegetation density) to 1 (very low dense vegetation). The standard value typically used in applications is 0.5 (Lillesand and Kiefer 1987).

- Transformed Soil Adjusted Vegetation Index (TSAVI) (Baret et al. 1989, Baret and Guyot 1991)

$$TSAVI = a(NIR - aR - b)/aNIR + R - a\,b + X(1 + a2) \quad \textbf{(2.13)}$$

where a and b are the soil line slope and intercept, respectively, and X is an adjustment factor which is set to minimize soil noise (0.08 in the original papers).

- Modified Soil Adjusted Vegetation Index (MSAVI) (Qi et al. 1994)

$$MSAVI = (NIR - R/NIR + R + L)\,(1 + L) \quad \textbf{(2.14)}$$

$$L = 1 - a*NDVI*WDVI$$

where *a* is the slope of the soil line

- Second Modified Soil Adjusted Vegetation Index (MSAVI2) (Qi et al. 1994)

$$MSAVI2 = 1/2(2NIR + 1 - [(2NIR + 1)^2 - 8(NIR - R)]^{1/2} \quad \textbf{(2.15)}$$

- Global Environmental Monitoring Index (GEMI) (Pinty and Verstraete 1991)

$$GEMI = \eta(1 - \eta/4) - (R - 0.125)/(1 - R) \quad \textbf{(2.16)}$$

$$\eta = [2(NIR2 - R^2) + 1.5NIR + 0.5R]/(NIR + R + 0.5) \quad \textbf{(2.17)}$$

The utility of the normalized difference vegetation index (NDVI) and related indices for satellite assessment of vegetation cover has been demonstrated for almost three decades. The time series analysis of seasonal NDVI data have provided a method of estimating net primary production, of monitoring phenological patterns of vegetated surface and of assessing the length of the growing season and dry-down periods (Huete and Liu 1994).

Global vegetation analysis was initially based on linearly regressing NDVI values (derived from AVHRR, Landsat MSS, Landsat TM and SPOT HRV data) with in situ measurements of leaf–area–index (LAI), absorbed photosynthetically active radiation (APAR), percent cover, and/or biomass. This empirical approach revolutionized global science land–cover biophysical analysis in just one decade (Justice et al. 1998).

2.2 NOAA Satellite Details

The National Oceanic and Atmospheric Administration (NOAA), National Environmental Satellite Data and Information Service (NESDIS) operates the system of environmental (weather) satellites and manages the processing and distribution of

data and images theses satellites produce every day. NOAA National Weather Service uses satellite data to create forecasts for the public, television, radio and weather advisory services. Satellite information is also shared with various Federal agencies, such as the Departments of Agriculture, Interior, Defense and Transportation; with other countries, such as Japan, India, Russia, members of the European Space Agency (ESA) and the United Kingdom Meteorological Office; and with the private sector.

NOAA operational weather satellite system is composed of two types of satellites: geostationary operational environmental satellites (GOES) for short-range warning and "now-casting" and polar-orbiting satellites for longer-term forecasting. Both types of satellite are necessary for providing a complete global weather monitoring system.

The new GOES-I through M series provide higher spatial and temporal resolution images and full-time operational soundings (vertical temperature and moisture profiles of the atmosphere). The newest polar-orbiting meteorological satellites (that began with NOAA-K in 1998) provide improved atmospheric temperature and moisture data in all weather situations. This new technology will help provide the National Weather Service the most advanced weather forecast system in the world.

GOES satellites provide the kind of continuous monitoring necessary for intensive data analysis. They circle the Earth in a geosynchronous orbit, which means they orbit the equatorial plane of the Earth at a speed matching the rotation of the

Earth. This allows them to fly continuously over one position on the surface. The geosynchronous plane is about 35,800 km (22,300 miles) above the Earth, high enough to allow the satellites a full–disc view of the Earth. Because they stay above a fixed spot on the surface, they provide a constant vigil for the atmospheric "triggers" for severe weather conditions such as tornadoes, flash floods, hailstorms and hurricanes. When these conditions develop, the GOES satellites are able to monitor storm development and track their movements.

GOES satellite imagery is also used to estimate rainfall during thunderstorms and hurricanes for flash flood warnings, as well as to estimate snowfall accumulations and overall extent of snow cover. Such data help meteorologists issue winter storm warnings and spring snowmelt advisories. Satellite sensors also detect ice fields and map the movements of sea and lake ice.

NASA launched the first GOES for NOAA in 1975 and followed it with another in 1977. Currently, the United States is operating GOES-10 and GOES-12. (GOES-9, which is partially operational, is being provided to the Japanese Meteorological Agency to replace their failing geostationary satellite). GOES-11 is being kept in orbit as a replacement for GOES-12 or GOES-10 in the event of failure.

Complementing the geostationary satellites are two polar-orbiting satellites known as Advanced Television Infrared Observation Satellite (TIROS-N or ATN), constantly circling the Earth in an almost north–south orbit, passing close to both poles. The orbits are circular, with an altitude between 830 (morning orbit) and 870

(afternoon orbit) km, and are sun synchronous. One satellite crosses the equator at 7:30 a.m. local time, the other at 1:40 p.m. local time. The circular orbit permits uniform data acquisition by the satellite and efficient control of the satellite by the NOAA Command and Data Acquisition (CDA) stations located near Fairbanks, Alaska and Wallops Island, Virginia. Operating as pair, these satellites ensure that data for any region of the Earth are no more than six hours old.

A suite of instruments is able to measure many parameters of the Earth's atmosphere, its surface, cloud cover, incoming solar protons, positive ions, electron-flux density and the energy spectrum at the satellite altitude. The primary instrument aboard the satellite is the Advanced Very High Resolution Radiometer or AVHRR.

The polar orbiters are able to monitor the entire Earth, tracking atmospheric variables and providing atmospheric data and cloud images. They track weather conditions that eventually affect the weather and climate of the United States. The satellites provide visible and infrared radiometer data that are used for imaging purposes, radiation measurements and temperature profiles. The polar orbiters' ultraviolet sensors also provide ozone levels in the atmosphere and are able to detect the "ozone hole" over Antarctica during mid-September to mid-November. These satellites send more than 16,000 global measurements daily via NOAA-CDA station to NOAA computers, adding valuable information for forecasting models, especially for remote ocean areas, where conventional data are lacking.

Table 2.1: Essential characteristics of NOAA satellites and AVHRR sensor

Spatial Resolution	1.1 km at NADIR ~ 8 km at 55 degree view angle	
Temporal Resolution	9 day repeat cycle for orbit Giving 2 images per day (one in day, one at night) Sun synchronous orbit at 830 km altitude	
Spectral Resolution	Name	Bandwidth (microns)
	Channel 1:	0.580 – 0.68 (RED)
	Channel 2:	0.725 – 1.00 (NIR)
	Channel 3:	3.550 – 3.93 (MIR)
	Channel 4:	10.300 – 11.30 (TIR)
	Channel 5:	11.500 – 12.50 (TIR)
Radiometric Resolution	10 bit quantization	
Calibration Details	Bands 3, 4, 5: on board calibration Bands 1, 2: deep space counts only	
Archive Commenced	1982 (NOAA)	

Currently, NOAA is operating five polar orbiters. A new series of polar orbiters, with improved sensors, began with the launch of NOAA-15 in May 1998 and NOAA-16 on September 21, 2000. The newest, NOAA-17, was launched June 24, 2002. NOAA-12, NOAA-14 and NOAA-15 all continue transmitting data as stand-by satellites. NOAA-16 and NOAA-17 are classified as the "operational" satellites.

2.2.1 Characteristics

The essential characteristics of the NOAA satellites and AVHRR sensor are shown in Table 2.1

Table 2.2: Ascending and descending node times in LST

Satellite	***Ascending Node (Launch)***	***Descending Node (Launch)***	***Ascending Node (3/95)***	***Descending Node (3/95)***
TIROS-N	1500	0300	n/a	n/a
NOAA-6	1930	0730	n/a	n/a
NOAA-7	1430	0230	n/a	n/a
NOAA-8	1930	0730	n/a	n/a
NOAA-9	1420	0220	2116	0916
NOAA-10	1930	0730	1753	0553
NOAA-11	1330	0130	1723	0523
NOAA-12	1930	0730	1915	0715
NOAA-13	1340	0140	n/a	n/a
NOAA-14	1330	0130	1330	0130

2.2.2 Orbital information

The nominal equatorial overpass times for each satellite are shown in Table 2.2. For vegetation studies, the early afternoon overpass satellites have been used almost exclusively. For NOAA-16 (2001-2005), NOAA-18 (May 2005 to present), these satellites were/are in the afternoon orbits. From September 1994 through February 1995, morning satellite orbits were used in the dataset, since afternoon satellite was not available during this period. Therefore, data for this period is not reliable.

The launch dates of each satellite, and dates from which the data were archived by NOAA, are shown in Table 2.3. The small gap between launch and operational recording of the data (by NOAA) is used for testing of the satellite systems.

Table 2.3: Launch and data available dates for the TIROS-N series satellites

Satellite	***Launch Date***	***Date Range***
TIROS–N	October 13, 1978	October 19, 1978 – January 30, 1980
NOAA–6	June 27, 1979	June 27, 1979–March 5, 1983 July 3, 1984–November 16, 1986
NOAA–B	May 29, 1980	Failed to achieve orbit
NOAA–7	June 23, 1981	August 19, 1981–June 7, 1986
NOAA–8	March 28, 1983	June 20, 1983–June 12, 1984
NOAA–9	December 12, 1984	February 25, 1985–November 7, 1988
NOAA–10	September 17, 1986	November 17, 1986–September 16, 1991
NOAA–11	September 24, 1988	November 8, 1988 – April 11, 1995
NOAA–12	May 14, 1991	May 14, 1991 – December 14, 1998
NOAA–13	August 9, 1983	August 9, 1993 – August 21, 1993
NOAA–14	December 30, 1994	April 11, 1995 – present

2.2.3 Satellite Data Acquisition

Two modes of data acquisition are available for AVHRR data on each of the NOAA satellites. During orbit the data are broadcast continuously (called high-resolution picture transmission (HRPT) direct readout) and can be recorded by independent ground stations. In addition, each NOAA satellite has on board recorders,

which can be programmed from earth control stations for data acquisition. When the satellite comes within view of a NOAA receiving station, the recorded data can be transmitted to earth and subsequently archived. This on board recording facility can only hold ten minutes of full resolution data (Kidwell 1997). Thus, global coverage of full resolution (1 km) data is currently not possible.

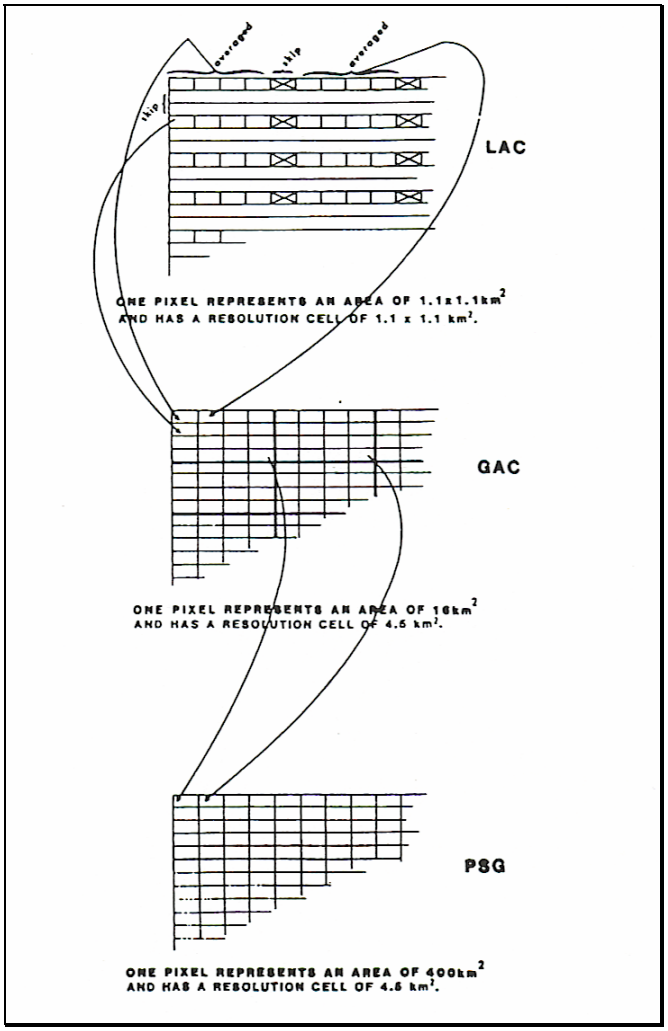

Figure 2.5: Global Area Coverage dataset generation

To achieve global coverage using the on board storage facility, NOAA samples the HRPT data to (approximately) 5 km pixels. This data format is called global area coverage (GAC). GAC data are generated by averaging the first four pixels out of every five along a scan line of image data, and retaining these averages from every third scan line. The spatial resolution at nadir is thus 1.1 x 4 km (Kidwell 1997). The averaged pixel is then assumed to represent a 4 x 4 km area on the ground. This scheme reduces the data handling (processing and storage) required on board the satellite (Figure 2.5) (Kidwell 1997, http://history.nasa.gov).

2.3 Satellite Data Processing

The extraction of useful information from satellite data is a complex process involving many individual steps, and is essentially a serial process. The acquisition of the raw data is the beginning of the process of information extraction. This process starts with the acquisition of raw data (digital numbers DNs). DNs are the scaled integers from quantization, which is not a physical quantity. Therefore, DNs should be converted to physical quantities for estimating land surface variables such as radiance. Normally, DNs are linearly related to radiance, and most remote sensing data providers produce the conversion coefficients for the users. The procedure that determines these conversion coefficients is called sensor calibration. This is an important procedure in remote sensing since many sensors deteriorate in space after satellite launch and the preflight conversion coefficients seldom remain valid.

2.3.1 Satellite Data Calibration

An essential component of any monitoring program is instrument calibration. In satellite remote sensing, calibration involves the conversion of the observed digital numbers (DN) to a physical quantity like radiance. The need for calibration has been encouraged by an increasing interest in the use of long-term data sets for monitoring biophysical phenomena (Gutman 1999b, Kogan and Zhu 2001, Simoniello et al. 2004).

The AVHRR instruments carried on each successive NOAA mission were calibrated against a standard to derive a linear relationship (i.e. gain and offset) between digital numbers and spectral radiance (Kidwell 1997). This calibration was adopted by NOAA and is known as the pre-launch calibration. Once in space, there is no on-board facility to determine the gain of the solar (visible) channels (one and two). The offset for these channels is routinely determined by observations of deep space.

Pre-launch calibration assumes that both the gain and offset of the instrument were stable over time. However, several studies demonstrated that the actual calibration values were not stable over time and differed from the published pre-flight values (Price 1987a, Price 1987b, Price 1988, Che and Price 1992, Kaufman et al. 1997, Kaufmann and Zhou 2000). The launch trauma, and subsequent degassing of the instrument (when in space), both contribute to the calibration differences reported (Che and Price 1992). In addition to these difficulties, the calibration coefficients for

NOAA-11 were changed by NOAA after a re-examination of the original instrument data and calibration methodology (Rao and Chen 1999).

With the lack of on board calibration facilities for measuring the instrument gain, and the obvious importance to multi-temporal studies, many different techniques of post flight calibration for the AVHRR have been proposed. Unfortunately, the results of several post-flight calibration studies using different methods vary substantially. In a review by Che and Price (1992), the results of several calibration studies were compiled and compared. Significant differences were noted, between both the techniques used and results from individual studies. However, in all cases, the post-flight calibration values were significantly different from the pre-flight values recommended by NOAA (Kidwell 1997).

2.3.1.1 Radiometric Calibration of the Satellite Data

Most applications of remote sensing require image data to be expressed in the form of reflectance values in order to allow comparisons with field or laboratory measurements. Two problems must be solved in order to pass from the digital number (DN) provided by the satellite instrument to a reflectance value: (1) the calibration factors of the DN into a physical value as luminance/radiance at satellite level, and (2) the computation of a reflectance at ground level.

The Working Group on Calibration and Validation (WGCV) of the international Committee on Earth Observation Satellites (CEOS) defines remote

sensing calibration as the process of quantitatively defining the system response to known controlled signal inputs (Defries and Belward 2000, Justice et al. 2000, Miura et al. 2000). The main fundamental aspects that need to be calibrated are the sensor system's response to electromagnetic radiation as a function of:

- Wavelength and/or spectral band (spectral response)
- The intensity of the input signals (radiometric response)
- Different locations across the instantaneous field of view and/or overall scene (spatial response or uniformity)
- Different integration times and lens or aperture settings wanted signals such as stray light and leakage from other spectral bands

Estimating land surface bio/geophysical variables accurately from remotely sensed data relies largely on the quality of the data and, in particular, on the accuracy of radiometric calibration. Those who build remote sensing devices must have accurate measurements of a sensor's radiometric properties before that sensor is sent into space-this is usually called preflight calibration. This calibration may change in space in response to variations in the environment surrounding the sensor in a space borne environment such as:

2.3.1.1.1 Satellite Orbit Drift

The NOAA satellites are intended to operate in nearly Sun-synchronous orbits. However, they show a non-negligible shift in local observation time (about 20-40 min per year) due to an orbit drift effect. This leads to an increase in the sun zenith angle

as well as to a change in the relative azimuth between sun and satellite direction with respect to the observed target during the satellite's lifetime. The change in the Sun-satellite-target geometry increases the optical path lengths of the incoming solar radiation, which increase the BRDF effects. Such effects are different for the two short-wave channels (Channel 1 and 2); therefore, the impact of solar zenith angle on NDVI changes depends on land covers (Price 1991, Privette et al. 1995, Gutman 1999).

2.3.1.1.2 Calibration Residuals

The AVHRR was designed for meteorological applications, like the other instruments on board of NOAA satellites. Particularly, the red and NIR channels were devoted to getting images of cloud distribution in weather analysis and prediction. Short-term meteorological analyses may be satisfied by relative coarse calibration since day-to-day spurious changes in sensor response are negligible. Therefore, the two short-wave channels are not calibrated onboard. Because of increasing interest in global coverage of albedo and vegetation information from AVHRR for long-term climatological and environmental applications, more stringent calibration procedures are needed in order to detect changes over a long period. The agency responsible for the satellite (NOAA-NESDIS) continues to provide the most updated calibration information concerning the AVHRR (Rao and Chen 1995a, Rao and Chen 1996, Rao and Chen 1999, Heidinger et al. 2003) following the guidelines given by the CEOS–WGCV (Working Group on Calibration and Validation).

The calibration of the AVHRR on board NOAA-9 satellite is the most reliable as confirmed by several studies (Rao and Chen 1995b, Gutman 1999), whereas the NOAA-11 short-wave channels seem to still be affected by calibration residuals (Koslowsky 1997, Gutman 1999). Observations (Koslowsky 1997, Gutman 1999, Rao and Chen 1999) performed on NOAA-14 data showed that the Rao and Chen (1996) calibration provides unreal upward trends in albedo and an illogical greening of deserts in NDVI. Therefore, the post-launch calibration coefficients of NOAA-14 were revised by NOAA-NESDIS (Rao and Chen 1999) and they have been available since the end of 1999.

However, the positive impact obtained from the new calibration was not yet widely explored and analyzed. This is also true for the AVHRR/3 onboard NOAA-16 operative since March 2001 of which the post-launch calibration coefficients were reviewed in April 2002 (NOAA-NESDIS, Updated Calibration Coefficients for NOAA-16 AVHRR, http://noaasis.noaa.gov/NOAASIS/ ml/nl6calup,html7). The update showed that there seems to be little loss in sensitivity of the first two channels and the coefficients proposed by Heidinger et al. (2003) overestimate their signal degradation. Such coefficients are based on a cross calibration with the onboard-calibrated MODIS data, the methods were devised in order to take into account the dual-gain approach specifically adopted on the AVHRR/3 to increase the sensitivity for dark targets.

2.3.1.1.3 Inter–satellite Calibration

The inter-calibration procedure is a mandatory step for integrating data from different satellite platforms. NOAA-NESDIS has provided inter-calibration coefficients only for NOAA-7, -9 and -11 (Rao and Chen 1995). Although such an inter-calibration minimizes the difference between the responses from different AVHRR, this problem is retained as shown by (Kogan et al. 1996, Koslowsky 1997, Gutman 1999, Kogan and Zhu 2001). Moreover, an official inter-calibration between NOAA-11 and NOAA-14 data is not yet available. Evident discontinuities in the mean values of the different AVHRR data sets are observed in correspondence to this satellite switch (Cuomo et al. 2001). Such discontinuities are able to alter the actual magnitude of natural changes or induce nonexistent long-term trends that can be erroneously ascribed to anthropic/climatic forcing (Teillet et al. 2006).

In-flight absolute calibration is usually performed on a routine basis for the thermal infrared channels to allow for precise temperature information, but the solar channels used for imaging on most operational satellites do not have onboard calibration capabilities mainly because of the limitations in satellite power, weight and space. Some orbiting satellites even have simple onboard calibration systems, but they change in sensitivity with time. Post-launch calibration data have to be obtained from vicarious calibration techniques. Vicarious calibration usually refers to techniques that make use of natural or artificial sites on the surface of the Earth for the post-launch calibration of sensors (Teillet et al. 2001).

The usual approach to sensor calibration starts with the formulation of a sensor calibration model. The simplest calibration model is the linear formula that links the sensor output (DN) to the radiance L at the entrance pupil of the sensor

$$Y = AL \qquad \textbf{(2.18)}$$

where Y represents the DN values and A is the matrix of the absolute calibration coefficients. Matrix A can be determined from the accurate preflight measurements, then monitored on orbit by onboard calibration devices using secondary or tertiary standard light sources (lamps or the Sun), and finally vicarious methods using images of specific well-known ground targets or the Moon (Dinguirard and Slater 1999). Since the sensor builders provide preflight calibration coefficients and thermal infrared sensors can be accurately calibrated using onboard devices, we focus on the vicarious calibration techniques and results for solar channels (visible and near-infrared spectra) in the following sections (Chen et al. 2005).

2.3.1.2 Post–launch Calibration Methods

There are roughly two typical post-launch calibration techniques. The first is to fly an aircraft with a calibrated radiometer that measures the spectral radiance of the target observed by the satellite in the same illumination and observing directions (Smith et al. 1997). It is often involved in simultaneous radiometric measurements of spatially and spectrally homogeneous Earth targets. This method is usually called the radiance-based calibration method. The reflectance-based method requires an

accurate measurement of the spectral reflectance of the ground target and measurement of spectral extinction depths and other meteorological variables (Teillet et al. 1990).

Although these may be the most direct methods, they are relatively expensive and complex and cannot be used to calibrate historical data (Kaufmann and Zhou 2000). Another technique is to compare the observe radiance with radiative transfer calculations using well-known physical characteristics of the atmosphere and surface targets. Most of the methods discussed below largely belong to this category.

The surface targets used for post–launch calibration include ocean, cloud, snow, dry lake, ice sheet and the Moon (Miura et al. 2000, Kogan and Zhu 2001, Heidinger et al. 2003).

2.3.1.2.1 Ocean Calibration

Ocean calibration relies on either molecular scattering in the atmosphere over the ocean or the Sun glint. For a cloudless air mass with a small amount of haze that is far away from the ocean glint, the major contribution (~ 70-80%) to upward radiance over deep oceans in the visible part of the spectrum is from molecular scattering in the atmosphere. To reduce the influence of other variables (e.g., aerosol scattering, foam and glint from the ocean water), particular viewing conditions need to be chosen. These conditions include deep oceans to get clear water, large viewing and solar zenith angles to increase the photon travel path length, and viewing the western

direction to avoid specular reflection. The non-Rayleigh component of the scattering is deduced from the signals in the near-IR spectrum where Rayleigh scattering is negligible. This method has been used to calibrate SPOT (Brown et al. 2006), MODIS (Kaufman et al. 1997) and AVHRR (Kaufmann and Zhou 2000).

Approximately 87% of glint radiance is due to specular reflectance (Kaufman and Zhou 2000). This reflectance cannot be theoretically established with the same accuracy as molecular scattering because of its dependence on speed and wave structure, but it is independent of the radiation wavelength and therefore can be used to determine the relative calibration of the near-IR bands to the visible bands. Good conditions usually correspond to wind speeds between two and five m/s. This approach has to be over as many glitter images as possible.

2.3.1.2.2 Desert Calibration

Desert sites have been widely used for sensor calibration since they have a stable spectral response over time. Because of their high reflectances, the atmospheric effect on the upward radiance is relatively minimal. They are also spatially uniform. Their temporal instability without atmospheric correction has been determined to be less than 1–2% over a year. Several sites have been used for sensor calibration, including the Libyan desert (Rao and Chen 1996, Kaufmann and Zhou 2000, Kogan and Zhu 2001) for calibrating AVHRR data, the North Africa desert for calibrating SPOT imagery (Brown et al. 2006), and the Egyptian desert, which was identified by

the international remote sensing community for sensor inter-calibration (Dinguirard and Slater 1999).

For calibration of high-resolution imagery, the White Sands Missile Range test site in New Mexico has been extensively used since the 1980s. It is located in the desert southwest of the U.S. in a region of low aerosol loading and an elevation of 1.2 km.

2.3.1.2.3 Clouds

Very-high-altitude (10-km) bright clouds are good validation targets in the visible and near-IR spectra because of their high spectrally consistent reflectance (Masek et al. 2006). If the clouds are very high, we do not need to correct aerosol scattering and water vapor absorption as both aerosol and water vapor are distributed near the surface. Only Rayleigh scattering and ozone absorption need to be considered. This method has been found (Csiszar et al. 2001) to give a 4% uncertainty for the intercalibration of the POLDER spectral bands.

2.3.1.2.4 Others Targets Used for Post-Launch Calibration

Dry lakes and other large homogeneous areas are also used for calibration, for example, Railroad Valley Playa, a dry lakebed in Nevada (U.S.) with a composition dominated by clay and Rogers dry lake at Edwards Air Force Base in California.

Tahnk and Coakley Jr (2001a) determined the sensor degradation coefficients for NOAA-14 AVHRR first two channels using the permanent ice sheets of the central Antarctica.

Cloud shadows over water have been used for calibrating high-resolution sensors (Carder et al. 1993). This cloud shadow method uses the differences between the total radiance values observed at the sensor for these two regions of sunlit and shadowed, thus removing the nearly identical atmospheric radiance contributions to the two signals (e.g., path radiance and Fresnel reflected skylight). What remains is due largely to solar photons backscattered from beneath the sea to dominate the residual signal. Normalization by the direct solar irradiance reaching the sea surface and correction for some second-order effects provide the remote sensing reflectance of the ocean at the location of the neighbor region, providing a known ground target spectrum for use in testing the calibration of the sensor. A similar approach may be useful for land targets if horizontal homogeneity of scene reflectance exists about the shadow.

The Moon is another calibration target. The stability of its reflectance is extremely high, but its radiance is not. The Moon can be used to (1) check the in-flight stability of a solar diffuser and (2) provide a direct calibration of the sensor (Dinguirard and Slater 1999). From 1995 to 1998, the USGS and the Northern Arizona University Department of Physics and Astronomy constructed an observatory in Flagstaff, Arizona dedicated to making long–term radiometric measurements of the

Moon. The purpose of this ongoing program with respect to EOS calibration is to utilize the radiometric stability of the lunar surface to provide long-term, on-orbit calibration and cross-calibration of EOS and non-EOS sensors flown on similar and different platforms. Currently, accurate measurements of the radiance and irradiance of the Moon are made at a number of wavelengths in the 348-2385 nm wavelength region using two telescopic imaging systems. The observatory measurements are used to produce exoatmospheric radiance images of the Moon that can be compared with orbiting spacecraft lunar observations. The lunar radiometric data are being archived in the NASA Goddard Space Flight Center (GSFC) Distributed Active Archive Center (DAAC).

2.3.1.3 Calibration Coefficients for AVHRR Reflective Bands

Archived AVHRR data span the operational lifetime of several satellites; however, sensor degradation has been identified as a major factor affecting the stability of the data quality. The degradations of NOAA-9, -11 and -14 crossing times are shown in Figure 2.6.

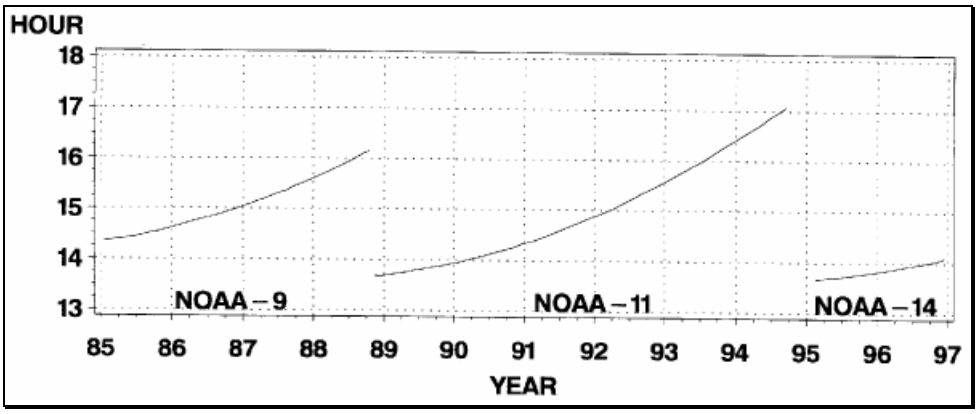

Figure 2.6: Equator crossing time of NOAA satellites afternoon passing

There have been many efforts to calibrate the two AVHRR solar bands (e.g., (Gutman and Ignatov 1995, Kogan et al. 1996, Rao and Chen 1996, Gitelson et al. 1998, Los 1998, Rao and Chen 1999, Kaufmann and Zhou 2000, Csiszar et al. 2001, Cuomo et al. 2001, Heidinger et al. 2003, Cihlar et al. 2004, Simoniello et al. 2004, van Leeuwen et al. 2006).

Table 2.4: NOAA Satellite Lifetimes

Spacecraft	***Launch Date***	***Operational Dates***
NOAA–7	June 23, 1981	Aug. 24, 1981 – Feb. 1, 1985
NOAA–9	Dec. 12, 1984	Feb. 25, 1985 – Nov. 7, 1988
NOAA–11	Sept. 24, 1988	Nov. 8, 1988 – April 11, 1995
NOAA–12	May 14, 1991	Sep. 16, 1991 – Dec. 14, 1998
NOAA–14	Dec. 30, 1994	Current
NOAA–15	May 13, 1998	Current
NOAA–16	Sept. 21, 2000	Current

The satellite lifetimes are listed in Table 2.4. In the following sections, we discuss the calibration coefficients for each NOAA satellite (-7, -9, -11, -14 and -15), **d** is the day since satellite launch, and $\mathbf{C_{10}}$ represents a DN value with quantization level 10.

2.3.1.3.1 NOAA-7

The following formulas involve converting DN values to radiance (W m^{-2} sr^{-1} μm^{-1}).

$$L_1 = 0.5753e^{0.000101d}(C_{10} - 36) \quad \textbf{(2.19)}$$

$$L_2 = 0.3914e^{0.00012d}(C_{10} - 37) \quad \textbf{(2.20)}$$

2.3.1.3.2 NOAA-9

Set 1:

$$L_1 = 0.5415e^{0.000166(d - 65)}(C_{10} - 37) \quad \textbf{(2.21)}$$

$$L_2 = 0.3832e^{0.000098(d - 65)}(C_{10} - 39.6) \quad \textbf{(2.22)}$$

Set 2:

$$L_1 = 0.5406e^{0.000166d}(C_{10} - 37) \quad \textbf{(2.23)}$$

$$L_2 = 0.3808e^{0.000098d}(C_{10} - 39.6) \quad \textbf{(2.24)}$$

2.3.1.3.3 NOAA–11

$$L_1 = 0.5496e^{0.000033d}(C_{10} - 40) \quad \textbf{(2.25)}$$

$$L_2 = 0.3680e^{0.000055d}(C_{10} - 40) \quad \textbf{(2.26)}$$

2.3.1.3.4 NOAA-14

The NOAA-14 spacecraft was launched into a nominal Sun-synchronous orbit on December 30, 1994. Because the two solar AVHRR bands do not have on board calibration devices, Rao and Chen (1996) developed calibration coefficients using the vicarious calibration method based on the data of about one year from the Libyan Desert.

The calibration coefficients are given for converting DN to TOA reflectance:

$$\rho = S_i (C_{10} - C_0)D^2 \quad \textbf{(2.27)}$$

where:

i = 1, 2 for two bands,

D^2 characterizes the distance between the Sun and the Earth that is given by

$$E_0 = \frac{\overline{E}_0}{D^2} = \overline{E}_0[1 + 0.033\cos(\frac{2\pi d_n}{365})], \text{ and}$$

$$E_0 = \overline{E}_0[1.00011 + 0.034221\cos x + 0.00128\sin x + 0.000719\cos 2x + 0.000077\sin 2x]$$

where $x = 2\pi(d_n - 1)/365$, and

$$S_1 = 0.0000135d + 0.111 \quad \textbf{(2.28)}$$

$$S_2 = 0.0000133d + 0.134 \quad \textbf{(2.29)}$$

C_0 is set equal to 41 counts in both channels.

Canada Centre for Remote Sensing (CCRS) recommendations available at: http://www.ccrs.nrcan.gc.ca/ccrs/rd/ana/calval/noaag14_e.html are given below.

The conversion is for radiance L (W m^{-2} sr^{-1} μm^{-1}) at the sensor (top of the atmosphere):

$$L = (DN - 41)/(gain) \quad \textbf{(2.30)}$$

where gain = A × d + B where d is the days since launch (d = 0 for Dec. 30, 1994); A and B are coefficients given in Table 2.5.

Table 2.5: Coefficients for Calculating Gain Value for NOAA-14 AVHRR Sensor

Time	***Band 1***		***Band 2***	
	A	***B***	***A***	***B***
1995	-3.532×10^{-4}	1.796	-6.161×10^{-4}	2.364
1996	-3.055×10^{-4}	1.778	-5.090×10^{-4}	2.325
1997	-2.671×10^{-4}	1.750	-4.275×10^{-4}	2.265
1998	-2.356×10^{-4}	1.715	-3.644×10^{-4}	2.196
1999	-1.209×10^{-4}	1.587	-3.714×10^{-4}	1.883
2000 / 2001	-1.249×10^{-4}	1.639	-3.837×10^{-4}	1.946

Tahnk and Coakley Jr (2001b) suggested a set of different calibration coefficients for TOA reflectance:

$$\rho_1 = (-5.35829\ 10^{-9}\ d^2 + 1.70469\ 10^{-5}\ d + 0.11414)\ (C_{10} - 41) \quad \textbf{(2.31)}$$

From launch to January 1, 2000

$$\rho_2 = (-1.46883 \times 10^{-9}\ d^2 + 5.59073 \times 10^{-6}\ d + 0.14302) \times (C_{10} - 41) \quad \textbf{(2.32)}$$

After January 2, 2000

$$\rho_2 = (4.38569 \times 10^{-5}\ d + 0.06829)\ (C_{10} - 41) \quad \textbf{(2.33)}$$

2.3.1.3.5 NOAA-15

Some piecewise linear (PWL) calibration coefficients for NOAA-15 AVHRR spectral data in channels 1, 2, and 3A acquired from 1998 and 2002 are presented here.

The NOAA-15 AVHRR/3 radiometer has a dual-gain response; thus, there are dual calibration gains for each channel. The recommended PWL coefficients for channels 1 and 2 are based on time-dependent calibration equations provided by Tahnk and Coakley Jr (2001b). Their calibration dataset was derived from the analysis of ice sheets in the Antarctic and Greenland (Appendix D of the NOAA KLM User's Guide on the web page from NOAA-NESDIS):

For the low-radiance range:

$$L = (DN - 38.9)/47.420 \quad \textbf{(2.34)}$$

For the high-radiance range:

$$L = (DN - 423.7)/7.064 \quad \textbf{(2.35)}$$

where L is the TOA radiance (W m^{-2} sr^{-1} μm^{-1}) at the sensor.

For bands 1 and 2:

$$L = (DN - DN_0)/(A \times d + B) \quad \textbf{(2.36)}$$

where d = days since launch (d = 0 for 5/13/1998). Since the DN_0 and B are consistent for a given radiance range (for low radiance range, DN_0 = 38.5 and B = 3.269 for band 1, DN_0 = 40.4 and B = 3.564 for band 2; for the high radiance range, DN_0 = 336.9 and B = 1.137 for band 1, DN_0 = 338.8 and B = 1.696 for band 2), Table 2.6 gives only the coefficient A.

Table 2.6: Coefficient A for Equation 2.36

Year	***Band 1***		***Band 2***	
	Low	***High***	***Low***	***High***
1998	1.389×10^{-6}	4.832×10^{-7}	3.564×10^{-5}	1.304×10^{-5}
1999	1.390×10^{-6}	4.834×10^{-7}	3.580×10^{-5}	1.310×10^{-5}
2000	1.390×10^{-6}	4.835×10^{-7}	3.600×10^{-5}	1.317×10^{-5}
2001	1.391×10^{-6}	4.837×10^{-7}	3.621×10^{-5}	1.325×10^{-5}
2002	1.391×10^{-6}	4.838×10^{-7}	3.641×10^{-5}	1.332×10^{-5}

Table 2.6 indicates that the AVHRR sensors on NOAA-15 are quite stable since the calibration coefficients do not change much with time.

2.3.1.3.6 NOAA-16

The preflight calibrations are provided below. The NOAA-16 AVHRR/3 radiometer also has a dual-gain response; thus, there are dual calibration coefficients for each channel:

Low radiance range:

$$L_1 = (DN - 38.5)/3.653 \quad \textbf{(2.37)}$$

$$L_2 = (DN - 37.9)/5.920 \quad \textbf{(2.38)}$$

$$L_{3A} = (DN - 71.25)/44.944 \quad \textbf{(2.39)}$$

High radiance range:

$$L_1 = (DN - 339.7)/1.250 \quad \textbf{(2.40)}$$

$$L_2 = (DN - 342.8)/2.011 \quad \textbf{(2.41)}$$

$$L_{3A} = (DN - 432.1)/7.142 \quad \textbf{(2.42)}$$

2.4 Summary

This is an introduction chapter that lays the foundations for this thesis. One of the major motivation in quantitative remote sensing is to estimate land surface biophysical variables. This chapter presents the statistical method based on different

vegetation indices. A comprehensive overview of the most commonly used vegetation indices was provided. The NDVI, VCI and TCI are only few of a number of vegetation indices, which have been used for the extraction of vegetation information from spectral observations in the red and near infrared regions of the spectrum. The NDVI values reported by different studies should theoretically be comparable. Unfortunately, this is not the case. Variations exist depending on whether the NDVI was computed with digital counts, radiances or reflectances. Thus, when using the results from other studies, care needs to be taken to ascertain how the NDVI was computed in that study.

The NDVI is computed from measurements of red and near infrared reflectance. Numerous studies have quoted the effect on the NDVI due to errors in these measurements caused by atmospheric contamination of the signals. In addition, quantization of the signal can also cause large errors in the NDVI. Quantization is a fundamental process behind all digital systems, and sets a finite limit on the precision with which samples can be represented. It would be a waste of time to make corrections, which were significantly smaller than the finite limit of precision for these data.

Conversion of the digital numbers recorded by the AVHRR instrument to radiance is a linear calibration problem requiring a gain and offset. These values were originally determined at the time of construction of each instrument, resulting in a pre-flight calibration. A number of studies have established that the pre-flight calibration

parameters have changed following launch on all NOAA satellites. Two distinct changes have been recognized:

1. Large change in calibration following launch,
2. Slow drift in calibration over time.

The changes in calibration for the individual channels will affect the NDVI if those changes are not identical.

The maximum value compositing (MVC) technique has been used extensively to remove cloud from time series images. This technique selects the maximum NDVI from a number of images and writes that value to the output composite image.

Sensor radiometric calibration is a very important process in quantitative remote sensing that converts the digital numbers to TOA radiance. There are three stages of calibration activities: pre-launch, in-flight, and post-launch. This chapter focuses on the post-launch vicarious calibration methods on various targets, such as ocean, desert, cloud and moon.

In Section 2.3.1.3, we provided the calibration coefficients for NOAA-AVHRR sensors. These coefficients indicate that sensor degradation has been a serious problem.

3. Study Area and Data

The U.S. is the largest exporter of agricultural products in the world. The forecasted 2006 U.S. agricultural export value is $68 billion (Figure 3.1) (USDA 2006).

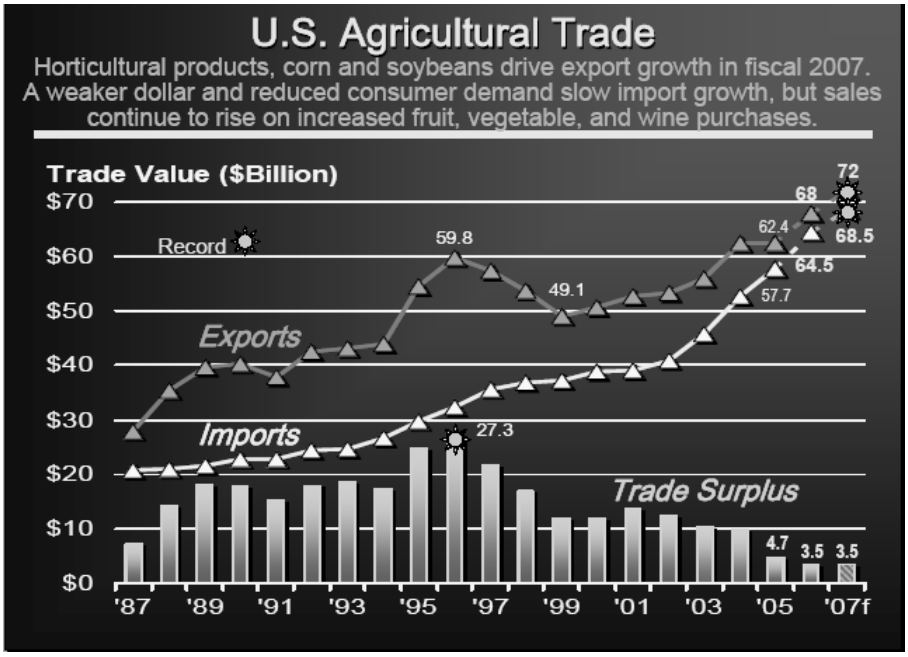

Figure 3.1: Fiscal year trade forecasts for agricultural, fishery and solid wood products

The economic well-being of many U.S. States is strongly linked to agricultural supplies and demands. Perhaps no region of the U.S. is tied more closely to agriculture and its economic impacts than Kansas. Agriculture in Kansas plays a leading role in the economy of the State. Every county in one form or another contributes to the agricultural output of the State. Kansas is proud to be called the granary of the country (Shroyer et al. 2004). Kansas constantly figures among the top

states in the production of wheat, sorghum, soybeans, and beef. The crop industry in Kansas provides income not only for the farmers and their families but also for other agriculturally-related occupations (USCRB 2005).

The study area, Kansas, lies between longitude: 94° 38' W to 102° 1' 34" W, and latitude: 37°N to 40°N (Figure 3.2).

Figure 3.2: Map of U.S. showing the study area Kansas

Kansas is divided into nine Crop Reporting Districts (CRD) and 105 counties (Figure 3.3). The districts are designated as follows: Northwest (NW-10), West Central (WC-20), Southwest (SW-30), North Central (NC-40), Central (C-50), South Central (SC-60), North East (NE-70), East Central (EC-80) and Southeast (SE-90) (Figure 3.3).

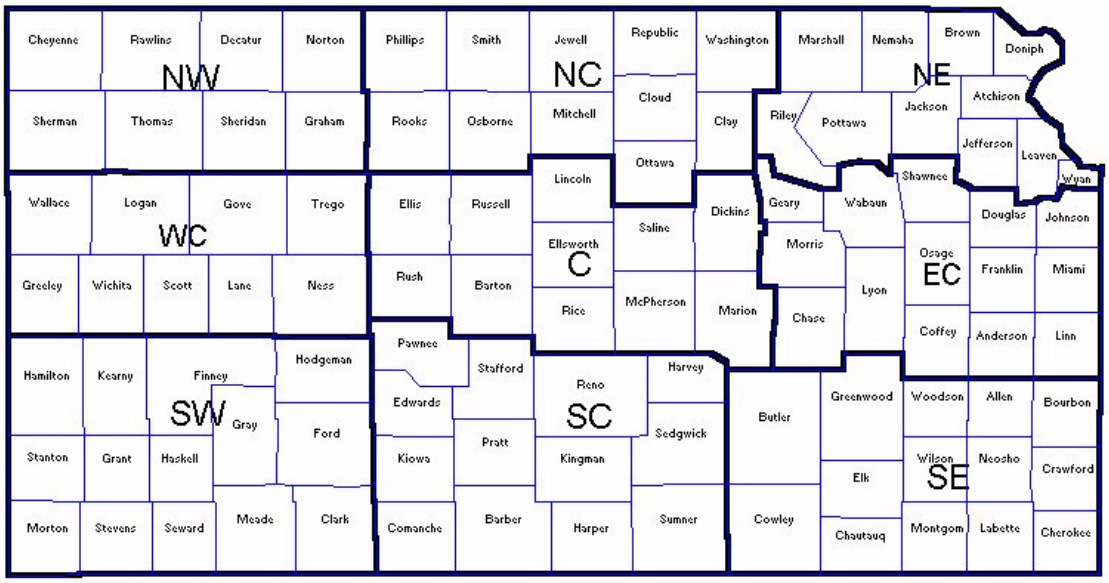

Figure 3.3: Kansas, Kansas Crop Reporting Districts (CRDs) and Kansas Counties

In Kansas, a CRD is made up of approximately 10 counties and covers an area as large as 10,000 square miles. The area within each CRD is considered to be relatively homogeneous with respect to vegetation, climate, topography and soil. Using CRD and county levels for regional assessments makes a Geographic Information System (GIS) a practical tool.

3.1 Kansas Cropland Utilization

There are about 48 million acres of farmland in Kansas, of which 65%, or about 31.1 million acres, is available for crops. Pasture and grazing land make up the remaining acres. During the twenty-one year period (1985 to 2005), crops were planted on an estimated 22.0 million acres and harvested from an average of 20.6

million acres (including 2.5 million acres utilized for hay). Wheat, the number one crop in the State, was planted on 38.7% of the total cropland. Sorghum, a distant second to wheat, was planted on 10.3% of the total cropland; corn 6.2% and soybeans 6.5%. Hay was harvested from 7.9% of all cropland and all other crops accounted for 1.3%. The residual 29.1% was idle cropland (Figure 3.4) (USCRB 2005).

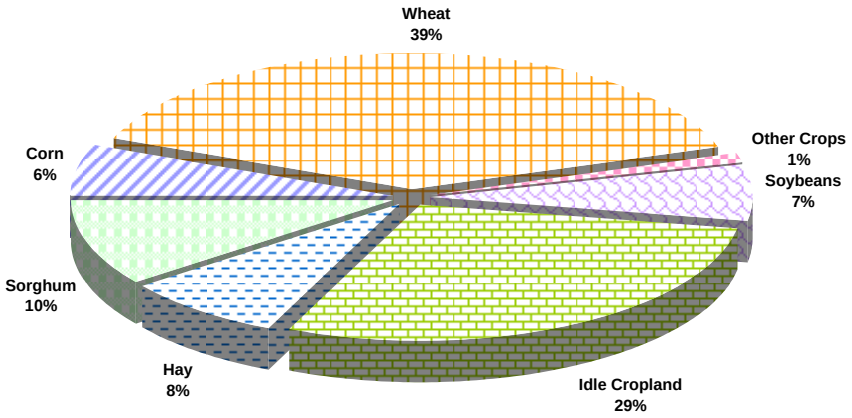

Figure 3.4: Kansas Crop Land Utilization

3.2 Climate of Kansas

Kansas has what is typically described as a continental climate, without the influence of any major bodies of water. Summers are warm, with the majority of the annual precipitation occurring during this period. Winters tend to be cold with an occasional mild spell and moderate snowfall amounts. Annual average precipitation ranges between approximately 40 inches in the southeastern part of the state to less

than 20 inches in the western part of the state (Figure 3.5). Annual average snowfall ranges from 40 inches in the northwest to less than 15 inches in the southeast (NOAA National Weather Service [NWS] 2004).

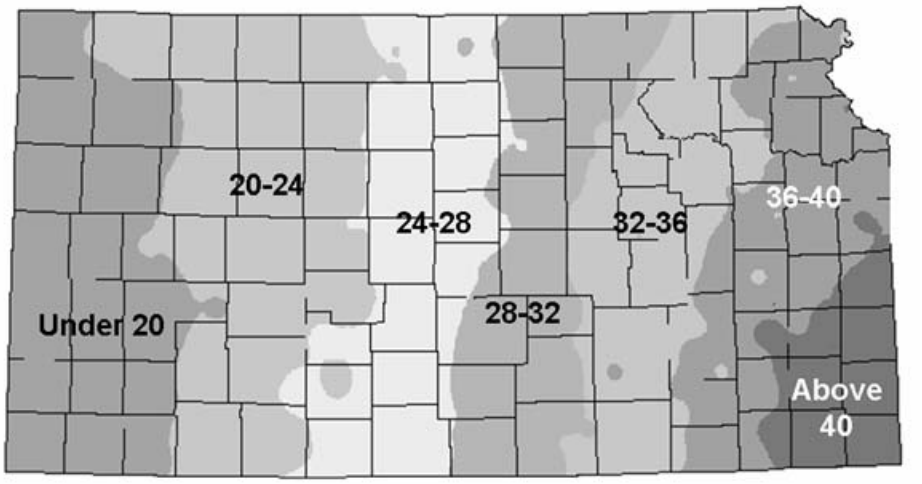

Figure 3.5: Average (1985 to 2005) annual precipitation in Kansas (inches)

3.3 Crops and Weather

In Kansas resides one of the most vulnerable agro ecosystems in North America. Annual and inter-annual variations in weather strongly affect crop production, and shifts in the temporality and quantity of precipitation have major impacts. High temperatures and strong winds during crop growing season stimulate high evapotranspiration rates, while weather vagaries during ripening and harvesting add further unpredictability in yield output.

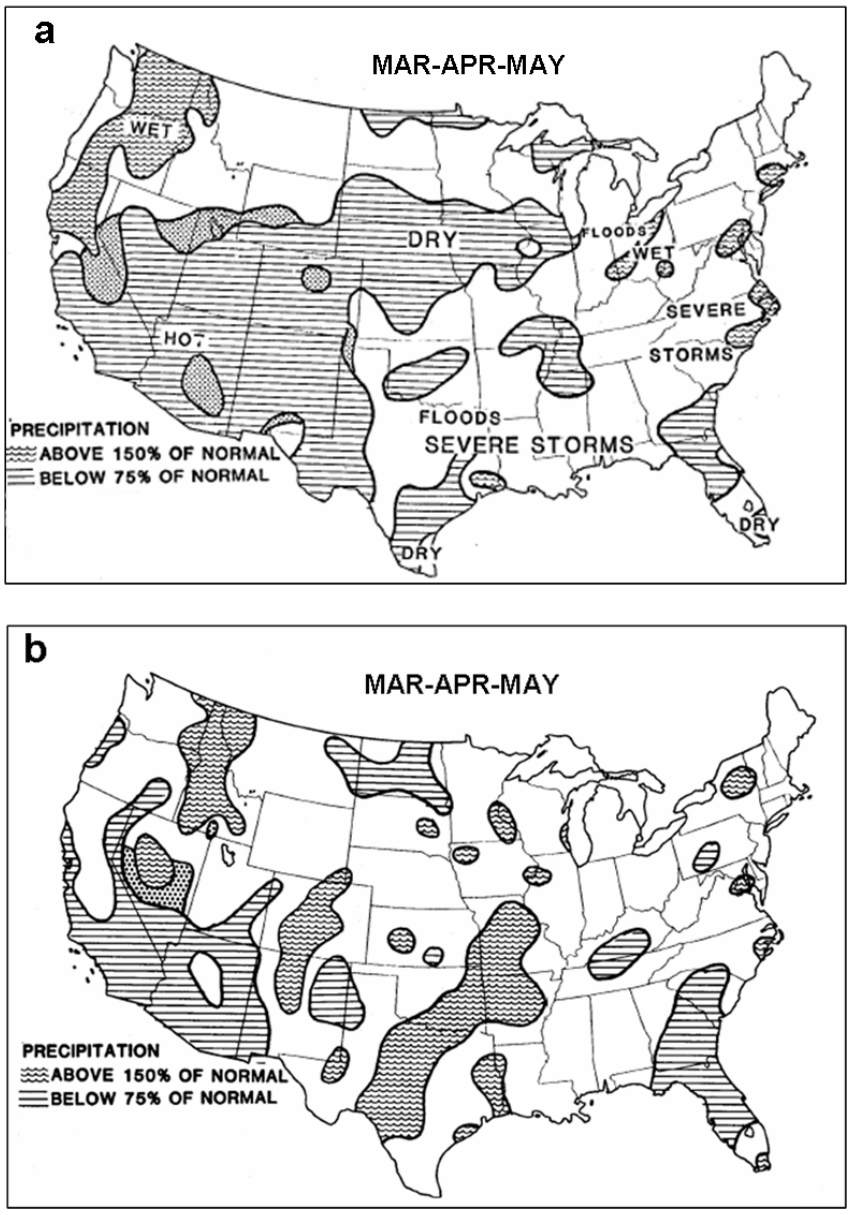

Figure 3.6: Percentage of normal precipitation in: (a) spring 1989, and (b) spring 1990 (WWCB 1989, 1990)

A clear example of the effects of weather on crop production can be seen in the following Kansas statistics: In 1989, winter wheat acres seeded were 12,400,000;

harvested acres were 8,900,000; abandonment 3,500,000 acres or 28.2% of the planted acres, the largest percent abandonment since 1951. Most of the state had through the entire winter wheat growing season less than normal rainfall. Moisture was below average in early spring. April and May were the driest on record for many counties in Kansas (Figure 3.6a), which contributed to a drought-stressed crop. The final yield was 24 bushels per acre (35% below the average). Total winter wheat production was 214,000,000 bushels, the lowest production since 1963.

In 1990, winter wheat acres seeded were 12,400,000; harvested acres were 11,800,000; abandonment 600,000 acres or 4.8% of the planted acres, the lowest abandonment since 1974. Weather conditions were near ideal through late winter and especially spring (Figure 3.6b). The final yield was 40 bushels per acre. Total winter wheat production was 472,000,000 bushels, more than double the drought stricken 1989 crop and exceeded the previous record of 458,900,000 bushels in 1982.

Such weather extremes demonstrate the sensitivity of crop production to weather variation. The impacts of wide swings in crop production affect not just primary producers and end users, but a wide range of enterprises and a variety of people dependent on a successful agricultural economy. Therefore, it is vital early estimation of crop production, which is important for both domestic and world production and consumption.

3.4 Yield

Yield is the measurement of the effects of all other crop growth factors. Nutrient management, irrigation management, residue management and weed management all affect yield. All management practices, as well as the external environment (heat, cold, rain, snow, pollution, etc.) can affect yield. In other words, yield is the integrator of crop health.

The desire for higher yields and higher profits causes most growers to manage their fields to get the maximum possible yield. This may, or may not, mean maximum possible profit. Often the point of maximum profit is not at the maximum yield. Hence, the term 'maximum economic yield' has been coined to indicate the point of maximum profit. Farmers are in a unique position in the business community. They operate on a narrow profit margin frequently too narrow for sustained operation.

Four factors are important in determining profitability: crop yield, selling price, fixed cost and production cost. Of these factors, farmers have relatively little control over the latter three. The opportunity for greater profits is most often contained in the optimal crop yield. However, reducing production costs, such as precision application, can also increase profits. Higher yields often increase production efficiency (yield divided by total production costs). Most farmers are operating with fixed costs approaching 80% of total production costs. More production per acre means less cost per unit of output, that is, lower cost to produce each pound of grain.

Yield is the single most important factor in the economic sustainability of any crop farm. Maximum yields can also mean maximum on-and off-site environmental damage (costs to society as a whole, as well as the farming operation). These costs can come from soil erosion, non-point source water pollution or groundwater damage.

The management decisions of growers, applying crop enhancements, nutrients, or pesticides at the right time, place, and amount, have the most significant effect on reducing environmental damage and costs. This is often called the Bill of Rights of Yield Management. Hence, the informed grower, aided by remote sensing information (regarding amounts and time of application of crop ameliorants) and guided by precision GPS systems (the right place) can have a significant impact on both yield and the wellbeing of the surrounding environment.

3.5 Data

The research presented in this thesis relies on two data sets.

3.5.1 Ground Data

The fist data set is historical, final, crop yield data obtained electronically from USDA National Agricultural Statistics Service (NASS) database site (http://www.usda.gov/nass/) for the entire state of Kansas, for each CRD and for each County from 1970 through to 2005. Initial release dates for USDA state level estimates are approximately May 11 for winter wheat and August 11 for corn,

soybeans, and grain sorghum. After this, winter wheat (WW) will be classified as an early-season crop and corn and grain sorghum as late-season crops. This database is updated annually for all crops, with each particular crop's final regional yield estimates released well after harvest completion. Updates to the final regional yield estimates can occur up to 3 years after their initial release, but generally these changes are not large. No historical or expected error statistics for these estimates are published below the national spatial scale, but they are nonetheless accepted by the industry as the best widely available record for average regional crop yield in the U.S. Yield data were used to investigate the nature of the relationship with satellite-based VH indices.

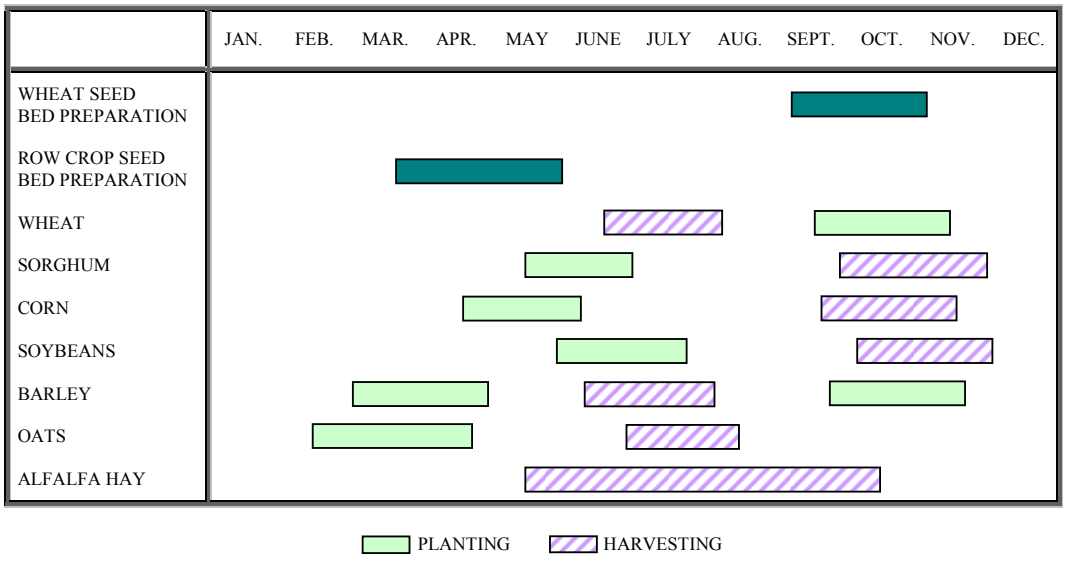

Figure 3.7: Kansas Crop Calendar (USCRB 2006)

3.5.1.1 Description of Crops and Time Periods under Investigation

Figure 3.7 shows the planting date patterns and harvesting dates for the most important crops grown in Kansas. The crops under investigation in this research are winter wheat, grain sorghum and corn.

3.5.1.1.1 Winter Wheat

In the U.S., wheat is the fourth leading field crop and the leading export crop. Total U.S. wheat production in 2003 reached 64 million tones, while in 2005 wheat production was only 57 million tones (FAO 2006). This reduction was due to unfavorable weather. Two types of wheat are grown in the U.S.: winter wheat sown in fall and harvested in early summer (Figure 3.8) and spring wheat planted in spring and harvested in late summer/early fall (Figure 3.9). Winter wheat provides 70 to 80% of the total wheat production (USCRB 2006).

In Kansas, WW planting usually takes place in late August. The peak of the planting period is from the middle of September to the middle of October. WW begins to head usually during the last week of April. Turning color closely follows heading. Ripening of wheat starts around the first week of June in southern Kansas and is virtually ripe Statewide by the first part of July. Ripening can vary from year to year. WW harvesting usually begins in mid June and lasts until mid July depending on precipitation amounts during these months (Figure 3.7).

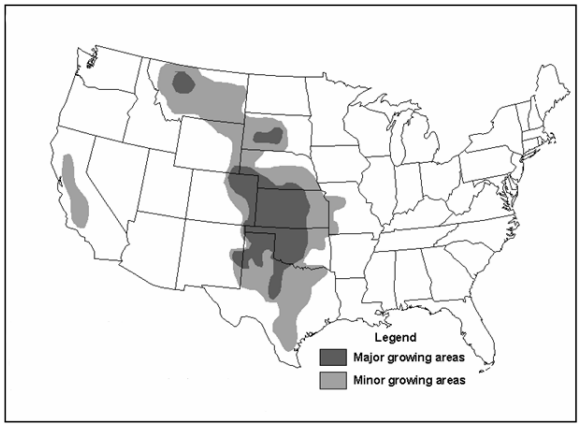

Figure 3.8: United States winter wheat production region

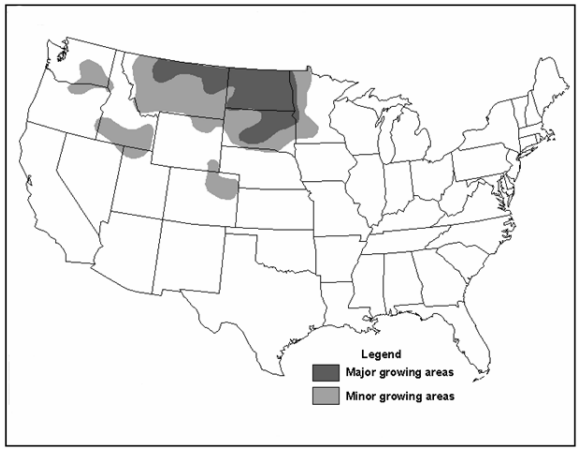

Figure 3.9: United States spring wheat production region

Kansas is the largest WW producing state. Nearly one-fifth of all U.S. WW volume is produced in Kansas (USCRB 2006). Annual average wheat production in Kansas for the past twenty-one years has been about 371 million bushels harvested from an average ten million acres. State WW yields over this period averaged 37

bushels per acre. Kansas ranks number one in wheat and wheat products exported (USCRB 2006).

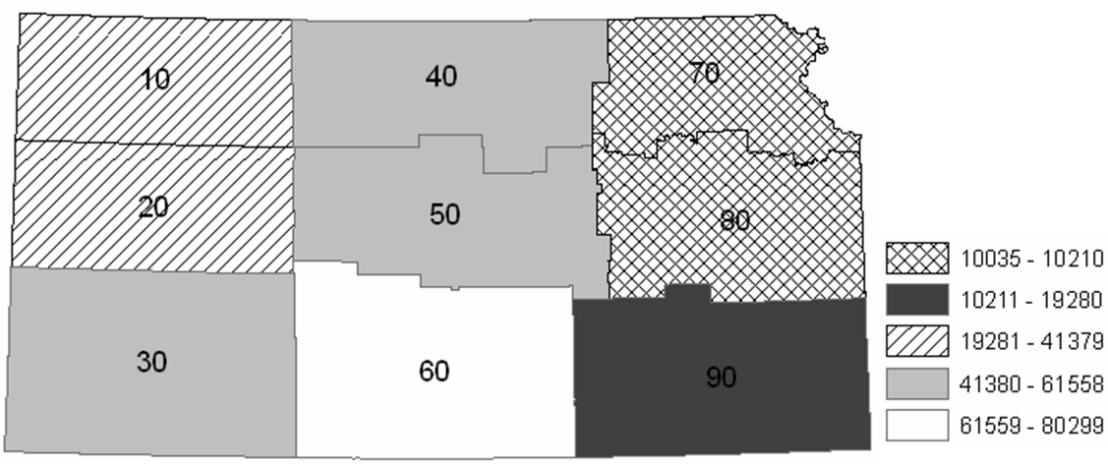

Figure 3.10: Kansas CRDs average winter wheat production thousand bushels (1985-2005) (USCRB 2006)

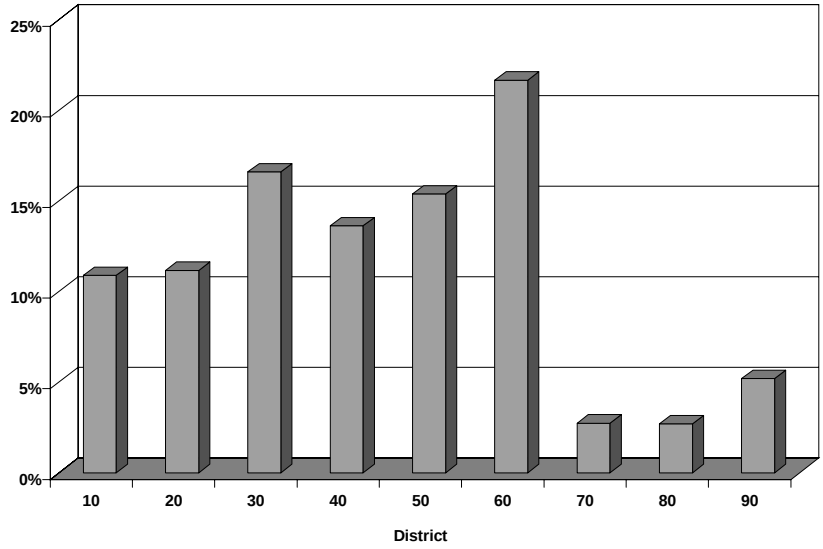

Figure 3.11: Percent CRD winter wheat production from total Kansas

Western and central CRD are the major producers of WW (Figure 3.10). CRD 60 is the major producer. Annual average WW production in CRD 60 for the past twenty-one years has been about 80 million bushels harvested from an average 2.2 million acres. CRD 60 WW yields over this period averaged 35.49 bushels per acre. CRD 60 is followed by CRD 30 and CRD 50 (Figure 3.11).

Sumner County is the major producer of WW in Kansas, during the 1985 to 2005 period, WW production has been about 13.5 million bushels harvested from an average 397 thousand acres. Summer WW yields over this period averaged 34.11 bushels per acre. Summer is followed by Reno and Finney. Two out of three of these counties are located in the south central CRD. WW county production representing the average for 1985 to 2005 is shown in Figure 3.12.

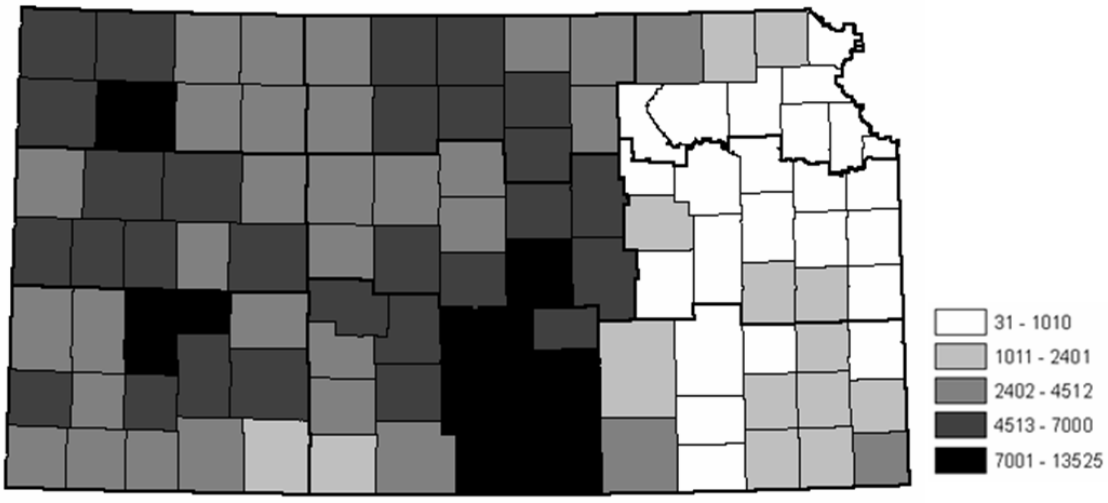

Figure 3.12: Kansas counties average winter wheat production thousand bushels (1985-2005) (USCRB 2006)

3.5.1.1.2 Sorghum

In Kansas, sorghum planting starts in early May. Mid July is about the time sorghum begins to head, and by the end of September, all sorghum has usually headed. Harvesting of sorghum usually begins in September and is well underway everywhere by mid October (Figure 3.7).

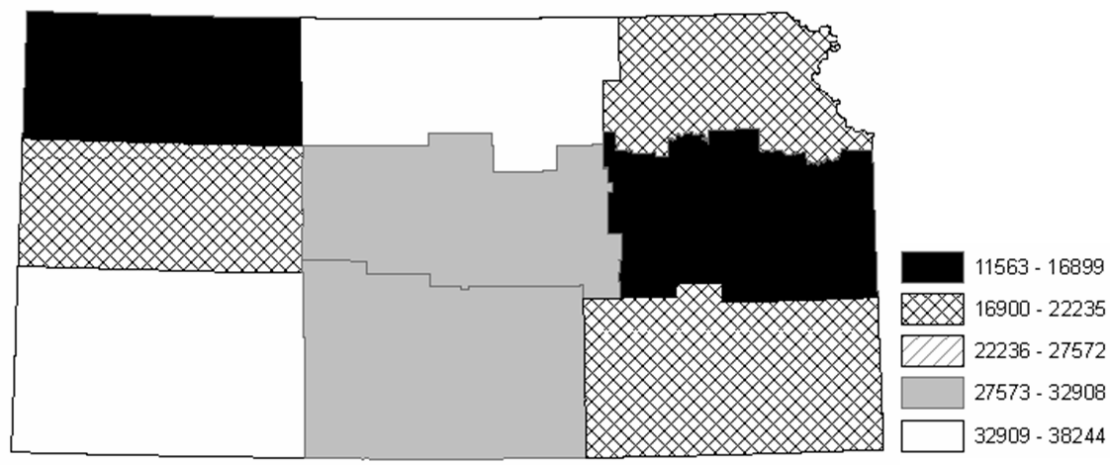

Figure 3.13: Kansas CRDs average sorghum production, thousand bushels (1985-2005) (USCRB 2006)

Annual average sorghum production in Kansas for the past twenty-one years has been about 224.5 million bushels harvested from an average 3.3 million acres. State sorghum yields over this period averaged 66.81 bushels per acre (USCRB 2006).

CRD 40 and CRD 30 are the major producers of sorghum (Figure 3.13). CRD 40 is the major producer. Annual average sorghum production in CRD 40 for the past twenty-one years has been about 38.2 million bushels harvested from an average 541

thousand acres. CRD 40 sorghum yields over this period averaged 70.38 bushels per acre. CRD 40 is followed by CRD 30 and CRD 50 (Figure 3.14).

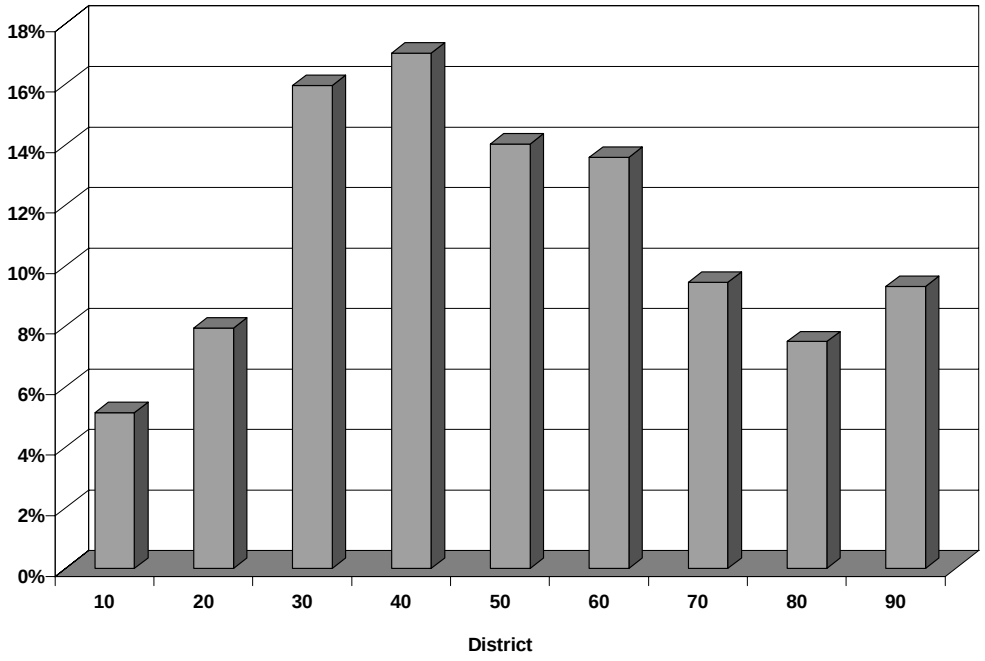

Figure 3.14: Percent CRD sorghum production from total Kansas

Marshall County, located in CRD 70, led all counties in sorghum production over the years 1985 to 2005, averaging 6.4 million bushels harvested from an average 81 thousand acres. Marshall sorghum yields over this period averaged 79.71 bushels per acre. Sorghum County production representing the average for 1985 to 2005 is shown in Figure 3.15.

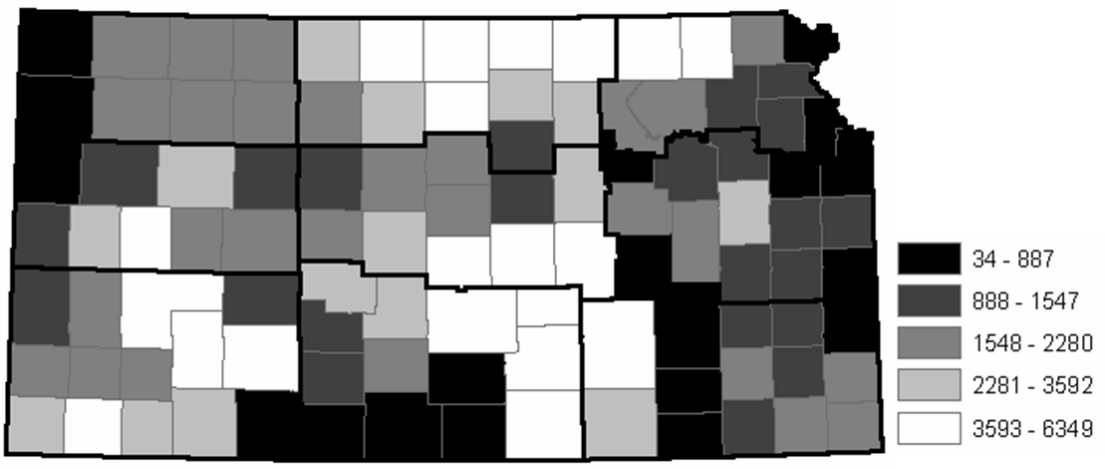

Figure 3.15: Kansas counties average sorghum production thousand bushels (1985-2005) (USCRB 2006)

3.5.1.1.3 Corn

In Kansas, corn planting starts in early April and is generally completed by the second week of June. Planting is at its peak during the first week of May. By the first week of July, corn starts to silk. The silking stage lasts four to five weeks, with most of the corn having silked by mid August. Corn reaches maturity between August and September. Corn harvest covers a three-month period of September through November (Figure 3.7).

During the 1985-2005 period average corn production in Kansas was 288.7 million bushels harvested from an average 2.2 million acres. State corn yields over this period averaged 132.81 bushels per acre (USCRB 2006).

Figure 3.16: Kansas CRD's average corn production (1985-2005) (USCRB 2006)

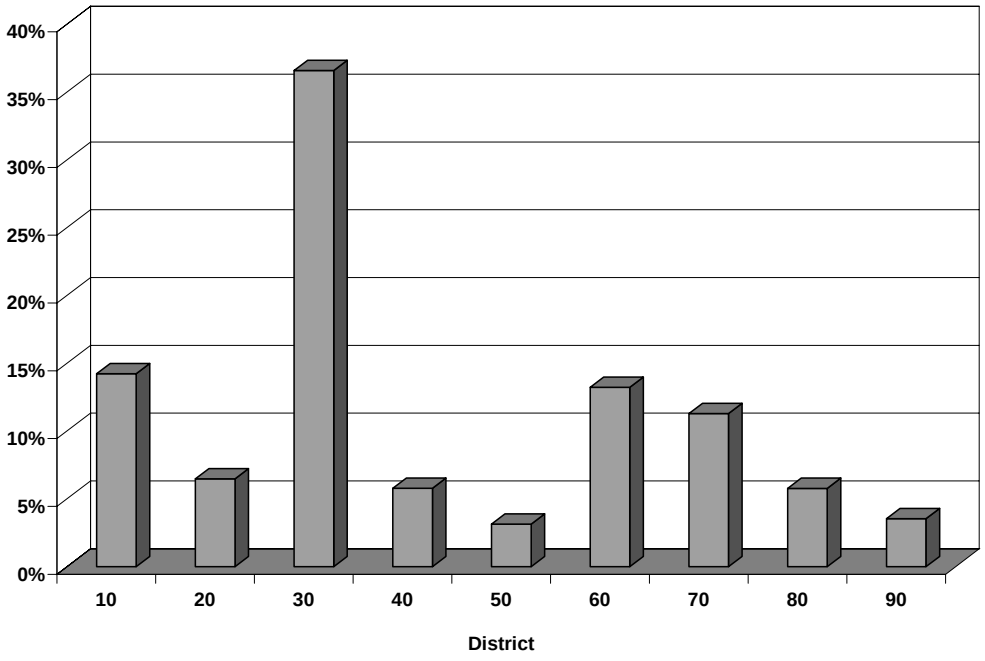

Figure 3.17: Percent corn CRD production from total Kansas

CRD 30 is the major producer of corn in Kansas (Figure 3.17). Annual average corn production in CRD 30 for the past twenty-one years has been about 105.6 million bushels harvested from an average 615 thousand acres. CRD 30 corn yields over this period averaged 169 bushels per acre. CRD 30 is followed by CRD 10 and CRD 60 (Figure 3.17).

Haskell County, located in CRD 30, led all counties in corn production over the years 1985-2005, averaging 17.5 million bushels harvested from an average 97 thousand acres. Haskell corn yields over this period averaged 177.31 bushels per acre (USCRB 2006). Corn County production representing the average for 1985-2005 is shown in Figure 3.18.

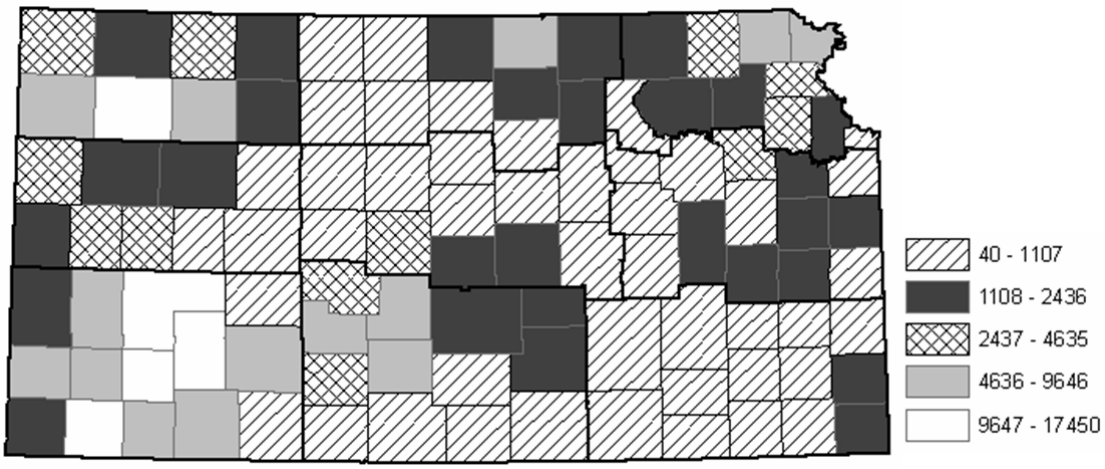

Figure 3.18: Kansas counties average corn production thousand bushels (1985-2005) (USCRB 2006)

3.5.2 Satellite Data

The second data set is a time series of weekly AVHRR-based VH indices (VCI and TCI) imagery obtained from NOAA Global Vegetation Index (GVI) data set from 1985 through to 2005. This data set was chosen because it is relatively inexpensive, reliable, and is updated in near real-time. Satellite data included AVHRR-measured solar energy reflected/emitted from the land surface (in 8-bit counts). Spatial data resolution was 4 km^2, sampled to 16 km^2 and temporal one day sampled to seven-day composite (Kidwell 1997). The GVI counts in the visible (VIS), near infrared (NIR) and infrared (IR, 10.3-11.3 μm, Channel 4 (Ch4)) spectral regions were used in this research. Post-launch-calibrated VIS and NIR counts were converted to reflectances following (Rao and Chen 1995b, Rao and Chen 1996, Kidwell 1997, Rao and Chen 1999, Heidinger et al. 2003) and used to calculate the Normalized Difference Vegetation Index (NDVI). Channel 4 counts were converted to brightness (radiative) temperature (BT) following the method in Kidwell (1997).

The AVHRR data are available in several forms. The highest spatial resolution recorded by the AVHRR is the 1.1 km Local Area Coverage (LAC) data. This spatial resolution has made LAC data attractive for some studies (Maselli et al. 1998, Rasmussen 1998). Unfortunately, for the same reason, LAC data have some disadvantages. Unless a user has access to the direct transmission of data from the satellite, LAC data can be stored only for limited regions and storage must be

requested in advance (Kidwell 1997). In addition, the large amount of data necessary for a regional study makes the use of LAC data difficult.

The second AVHRR data form available is the Global Area Coverage (GAC) data. In contrast to LAC data, GAC data are collected and archived daily for nearly the entire Earth and are available for research. The GAC data are produced from LAC data by averaging the first four pixels in a 3x5 array of LAC observations, and have a spatial resolution of approximately 4 km^2 (Kidwell 1997).

The third form of AVHRR data are the Global Vegetation Index (GVI) data. These GVI data are produced by selecting one GAC pixel from a 4×4 array of GAC pixels to represent the entire array. Thus, the GVI resolution is approximately 16 km^2 at the equator (Kidwell 1997). The data for each GVI pixel contain values for four of the five AVHRR channels, as well as the scan angle and solar zenith angle (Kidwell 1997).

The GVI product from the AVHRR was selected for this study. One advantage of the GVI data is that weekly GVI data currently exist for the world between 70° N and 55° S beginning in 1982. These data continue to be collected and are available in real-time from NOAA-NESDIS. The VH indices and data are delivered in real time (every Monday) to http://orbit.nesdis.noaa.gov/smcd/emcb/vci. They show global and regional vegetation health, moisture and thermal conditions and fire risk potential. They also discuss climate issues and VH utility in global observing system. Despite the spatial sampling reducing the resolution to 16 km^2, the GVI data

have been found to contain valuable information about vegetation conditions. In addition, for a regional study, GVI data are manageable and economical to use. This makes GVI data practical for any organization needing to anticipate and assist governmental or private agencies designing farm policies or marketing strategies.

Unfortunately, studies have found that the empirically derived NDVI products can be unstable, varying with soil color and moisture conditions, bidirectional reflectance distribution function (BRDF) and atmospheric conditions. Clouds and other atmospheric constituents obscure the land surface, reducing NDVI considerably. Changes in viewing geometry can lead to both an increase and decrease in NDVI, depending on location, type of vegetation and illumination. Satellite orbit drift, sensor degradation and satellite change create long-term noise in NDVI data, especially after the satellite has been in service for more than 3 years (Los 1998, Csiszar et al. 2001, Cihlar et al. 2004, Los et al. 2005).

Many techniques have been developed to reduce noise in AVHRR data. Several cloud-screening algorithms are presently available. Post launch calibration correction reduces the generally upward trend in NDVI time series. Alternative techniques for high frequency noise reduction in NDVI and brightness temperature (BT) data have been developed and widely used in the past fifteen years (Kogan and Sullivan 1993).

In order to reduce long-term systematic errors in GVI time series (Gutman 1999, Kogan and Zhu 2001, Simoniello et al. 2004) the following techniques were

used in this research. To minimize the effects of changes in solar zenith angle, satellite scan angle and earth-sun distance, satellite images were aggregated over a seven-day period by saving those values that have the largest NDVI. In addition, to correct instability in the VIS and NIR channels due to AVHRR sensor degradation, sensor change and to degradation in satellite orbit over time, the standard data preparation procedure included a correction of VIS and NIR values following the most updated calibration information concerning the AVHRRs (Rao and Chen 1995b, Rao and Chen 1996, Rao and Chen 1999, Heidinger et al. 2003). In order to reduce cloud and bidirectional effects and also some effects of sun-sensor geometry NDVI and BT were smoothed over time (Kogan et al. 1996).

The post-launch corrections and time series smoothing considerably improved the stability of the NDVI and BT over time. However, we should admit that there is some reduction in NDVI and BT values by the end of AVHRR sensor life for each satellite. Investigation of this problem by Kogan and Zhu (2001), Simoniello et al. (2004) showed that in most crop-related vegetative areas, maximum NDVI and BT change by the end of the satellite life is less than 10%. Moreover, this 10% reduction is much smaller than variation in NDVI and BT values related to inter and intra annual weather change. This accuracy is appropriate for monitoring vegetation health in vegetative areas (Kogan et al. 2003, Simoniello et al. 2004).

In addition, vegetation oriented techniques were used in this study to reduce the remaining noise in AVHRR data. These techniques considerably reduce high

frequency noise and enhance low-frequency variations related to weather fluctuations. These techniques smooth the weekly time series with a combination of a compound median filter and the least square technique (Kidwell 1997). This smoothing completely eliminates high frequency outliers, including random effects and pulled out low-frequency weather related fluctuations (valleys and hills in the NDVI and BT time series) during the annual cycle (Kogan 1997). After smoothing, inter-annual differences in NDVI and BT became more apparent. These differences are due to weather variations. For example, in dry years, the NDVI curves are lower and the BT curves are higher than in normal and wet years (Kogan et al. 2003).

For heterogeneous land cover, the NDVI, which reflects vegetation greenness and vigor, are normally higher in the areas with more favorable climate, soil, cultivation practices (irrigation) and more productive ecosystems (forest) compared to the areas with less favorable environmental conditions (dry steppe). These differences were taken into consideration. In addition, weather-induced NDVI variations are quite different for various climatic conditions, and, what is even more important, that weather signal is much weaker and not easily detectable compared to the ecological one. Therefore, when weather impacts on vegetation were estimated from NDVI, the weather-related portion was enhanced by separating it from the ecological one following Kogan (1997).

Therefore, for each pixel and week, the NDVI values were stratified based on the 1985-2005 absolute minimum and maximum of NDVI. The assumption was that

maximum amount of vegetation is developed in years with optimal weather because such weather stimulates efficient use of ecosystem resources (for example, increase in the rate of soil nutrition uptake). Conversely, minimum vegetation amount develops in years with extremely unfavorable weather (mostly dry and hot), which suppresses vegetation growth directly and through a reduction in the rate of ecosystem resources use (for example, lack of water in drought years reduces considerably the amount of soil nutrient uptake). Therefore, the absolute maximum and minimum of NDVI calculated from several years of data that contain the extreme weather events (drought and no drought years) can be used as criteria for quantifying the potential of geographic areas (Kogan 1997).

The required adjustment of the NDVI values for a given set of geographically determined criteria provides an additional filtering over and above median filtering, thereby enhancing the weather-related signal in NDVI values. The next stage of the algorithm development consisted of comparison of any year's NDVI in the archive with the absolute maximum and minimum NDVI values, which define ecosystem resources (Kogan 1997). The entire procedure is given in the following expression of the vegetation condition index (VCI):

$$VCI = 100\times(NDVI - NDVI_{min})/(NDVI_{max} - NDVI_{min}), \qquad \textbf{(3.1)}$$

where NDVI is the smoothed weekly NDVI, $NDVI_{max}$, and $NDVI_{min}$ are absolute maximum and minimum NDVI, respectively, calculated for each pixel and week from multiyear smoothed NDVI data.

The VCI captures rainfall dynamics better than the NDVI particularly in geographically no homogeneous areas. The VCI not only permits the description of land cover and spatial and temporal vegetation change but also allows quantifying the impact of weather on vegetation. It is also important to note that the VCI makes it possible for one to compare the weather impact in areas with different ecological and economic resources. VCI values indicate easily how much the vegetation has advanced or deteriorated in response to weather and how far vegetation development is from the potential maximum and minimum defined by the ecological limits. The selection of the absolute maximum and minimum NDVI as a criterion for index modification is justified by the fact that the patterns of this criterion match the established agro ecological zoning. The technique minimizes noise in the AVHRR data by means of elimination of no uniformity and increases the vegetation signal related to the impact of weather alone on vegetation.

In addition, during the rainy season, it is not uncommon for cloudy conditions to prevail for long periods. If these periods last for more than 3 weeks, the weekly NDVI values tend to be depressed giving the false impression of water stress or drought conditions. To remove the effects of cloud contamination in the satellite assessment of vegetation condition, Kogan (1997) proposed the temperature condition index (TCI). The TCI is derived from BT, and its algorithm is similar to VCI except that the formula was modified to reflect the opposite to the NDVI vegetation's

response to temperature (high temperature is less favorable for vegetation). The resultant expression for TCI is:

$$TCI = 100 \times (BT_{max} - BT)/(BT_{max} - BT_{min}), \quad \textbf{(3.2)}$$

where BT, BT_{max}, and BT_{min} are smoothed brightness temperature, its maximum and minimum, respectively, calculated for each pixel and week from multiyear data.

It is important to mention here that BT was calculated from channel 4 (10.3-11.3 μm), because the channel 4 measured radiance is more representative of drought conditions being much less responsive to the amount of water vapor in the atmosphere than Channel 5. Although the BT only partially represents land surface conditions, analysis showed that during drought years the BT is much higher than normal and wet years (Kogan 1997). Therefore, similar to the NDVI, the same principle of comparison of dry year with other years in the GVI archive was used for TCI approximation.

High surface temperature or outgoing long wave radiation tends to be associated with clear skies, and the converse is also true. Therefore, the TCI is an appropriate complementary tool to VCI. When used together, the VCI and TCI provide a reliable crop condition assessment scheme.

4. Geographic Information Systems

Geographic Information System technology can be used as a data analysis and dissemination tool on a crop yield assessment system. A geographical information system (GIS) consists of a system of hardware and software used for storage, retrieval, mapping and analysis of geographic data. Spatial features are stored in a coordinate system (latitude/longitude, state plane, UTM, etc.), which references a particular place on the earth. Descriptive attributes in tabular form are associated with spatial features. Spatial data and associated attributes in the same coordinate system can then be layered together for mapping and analysis (Figure 4.1). GIS can be used for scientific investigations, resource management and development planning.

GIS differs from CAD and other graphical computer applications in that all spatial data is geographically referenced to a map projection in an earth coordinate system. Spatial data can be re-projected from one coordinate system into another, thus data from various sources can be brought together into a common database and integrated using GIS software. Boundaries of spatial features should register or align properly when re-projected into the same coordinate system. Another property of a GIS database is that it has topology, which defines the spatial relationships between features. The fundamental components of spatial data in a GIS are points, lines (arcs) and polygons. When topological relationships exist, we can perform analyses, such as

modeling the flow through connecting lines in a network, combining adjacent polygons that have similar characteristics and overlaying geographic features.

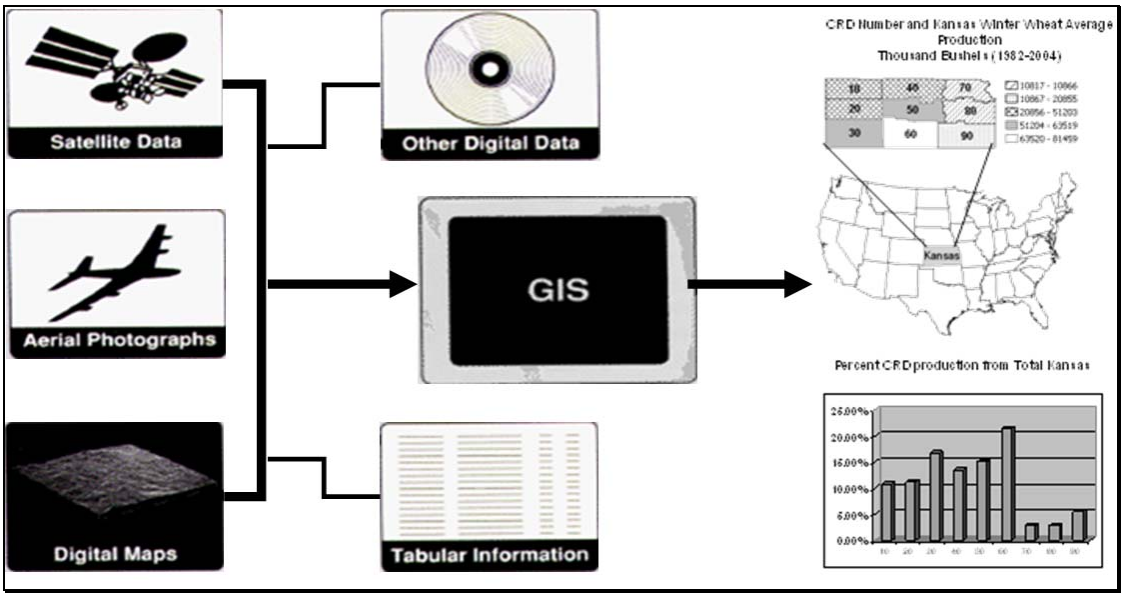

Figure 4.1: Spatial data and associated attributes in the same coordinate system can be layered together for mapping and analysis

The combination of remote sensing and GIS technologies facilitates the acquisition, interpretation and dissemination of information on crop conditions and yield estimation during the growing season.

Using remote sensing technology, data are collected and the vegetation indices calculated. Then, the vegetation indices are summarized in a GIS to the State, Crop Reporting Districts (CRD) and county levels. Then, theme maps of ranges of the vegetation indices can be produced. We can produce a ratio image comparing the

closest annual calendar weekly period from the previous year, as well as show the two images side by side. For events, such as droughts, floods, crop diseases or pest infestations that reduce plant chlorophyll the vegetation indices images are of considerable value in an early warning system.

The major strength of using remote sensing and GIS technologies is that we have complete spatial coverage and we can display the results graphically on a weekly basis.

4.1 GIS in the Analysis of Remote Sensing Data

Figure 4.2 shows the data flow between the GIS (ARCINFO), the image analysis software (ENVI/IDL) and the statistical analysis software (SAS) that we used for processing, analyzing and modeling the data in this research.

4.1.1 Geographic Base

The data flow begins in the GIS environment with basic map layers for geographic control. Thematic map layers derived from a variety of sources are added to the GIS database. Of particular interest in this research are State, CRD and county levels (Figure 4.2a).

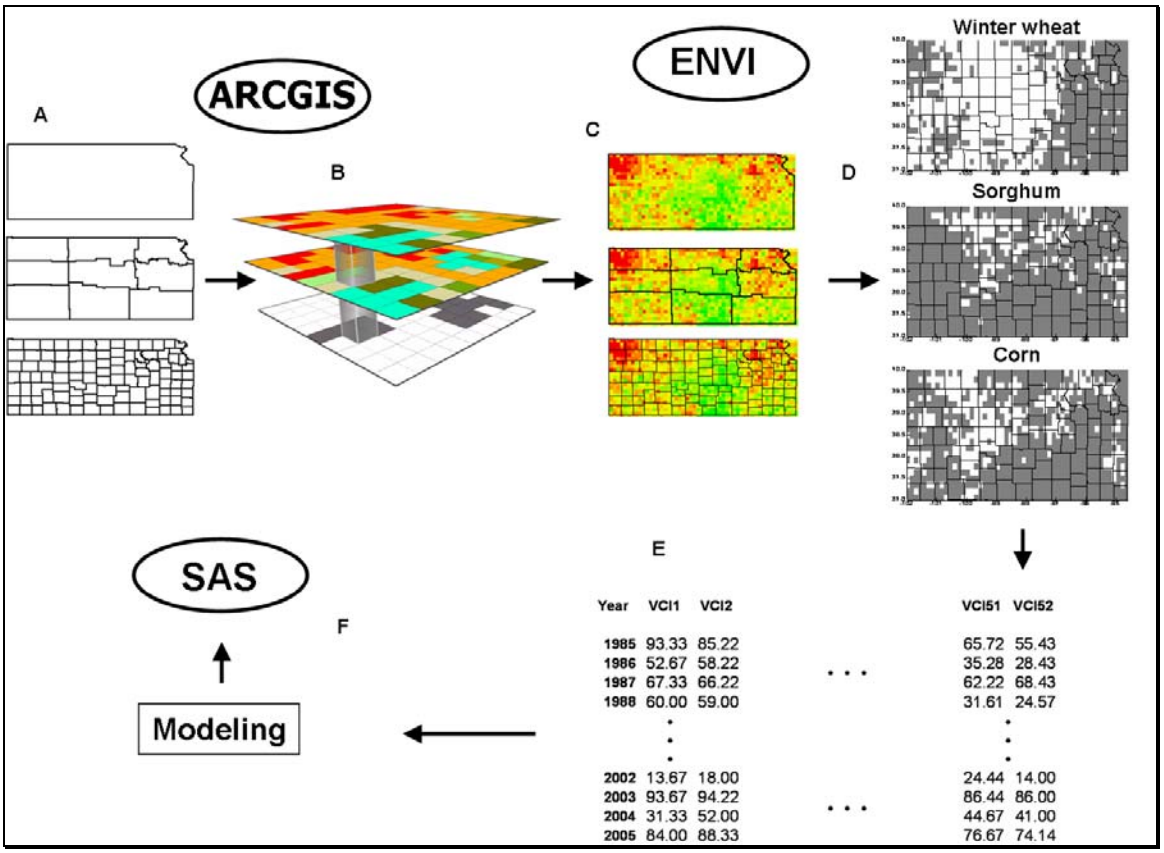

Figure 4.2: Data flow between the GIS, the image analysis and the statistical analysis software used for processing, analyzing and modeling the data

4.1.2 Image registration

The first stage of processing of the satellite imagery is geometric correction to a map projection the same map projection used in the GIS environment (Figure 4.2c). Since, this study includes several spatial data sets, such as AVHRR-based Vegetation Health Indices (VHI) imagery and ground crop yield data; data integration is an important step. Data integration can increase the effectiveness of the data

dramatically. Data layers that are un-rectified with respect to each other, such as different maps at different projections, create different problems and thus limit their usefulness.

In this research, we integrate GIS and remote sensing data sets to estimate crop yields. Correctly rectified satellite imagery had a benefit for their later correlation with detailed ground crop yield data. Rectification to single-pixel accuracy is required for matching crop yield data to AVHRR-based VH indices data.

A frequent problem with integrating spatial data involves uncertainty about map projections and datums. Often the datum and projection information are not known. However, if the spatial data in question is to be integrated with other data sources, the projection information must be known; otherwise, the absolute accuracy of spatial locations is uncertain. Even if the projections are known, problems may occur with transforming the spatial data. Certainly all projection changes should involve transformation to geographic coordinates (latitude/longitude) as an intermediate step to ensure stable spherical trigonometry (Steinwand et al. 1995).

It is also instructive to know the magnitude of the errors involved in projection conversion. For example, to determine the magnitude of the errors resulting from a conversion from geographic coordinates (assuming a spherical earth) to geographic coordinates using the WGS84 datum, the distance difference for points on a graticule were plotted and contoured, as shown in Figure 4.3. Errors up to 100m occur in the

western U.S., or even greater in the NW of Nevada, much of California, and all of Washington and Oregon. Errors greater than 200m occur for much of South America.

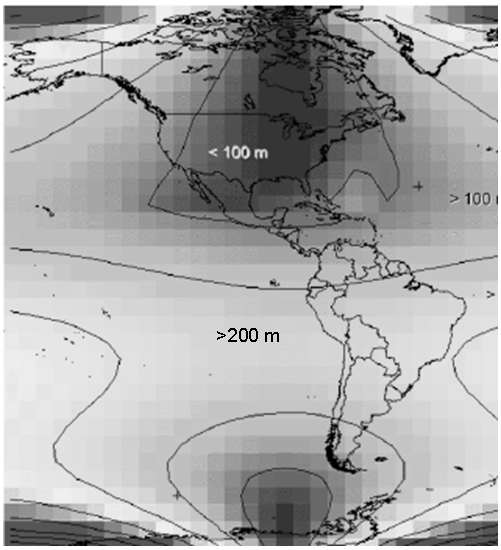

Figure 4.3: Errors between latitude/longitude graticule for perfect sphere and for WGS84. Dark areas are below 100 m, but lighter grey areas are over 300 m

4.1.3 Image rectification

Since all data have errors, these errors are compounded when multiple data sets are compared. Rubber sheeting algorithms are used to help minimize errors on individual data sets as well as the relationship among data sets. In this research, layers in step a (Figure 4.2a) are chosen as the base and the VH indices images rectified to it. The data set chosen as a base has errors associated with it, but it is the relative spatial error among the data sets that is important rather than their absolute position in the world.

Remote sensing imagery is a valuable source of GIS data and has somewhat predictable distortion. However, most data still require non-parametric methods of rectification. The non-parametric method requires taking control points from an image or map you are warping from and another set of like control points from the data you are warping to. The algorithms employed are polynomial equations such as those that are used in curve fitting.

There are several problems in applying these algorithms:

1. The algorithms do not remove topographic distortion. Topography is random and a Digital Terrain Model (DTM) must be used to correct for elevation. The random topographic effects are minimized in flat terrain and data collected from high orbits.

2. The residual errors are only statistical. This does not guarantee levels of accuracy everywhere; however, you can derive probabilities by contouring residuals and kriging.

3. Data outside of the control points have no guarantee as to their spatial accuracy. Typically, the higher the order of the polynomial the worse the integrity of the data outside of the controlled area.

4. Collect more control points than are needed by the coefficients to the polynomial. Doubling the number of control points needed by the coefficients in the equation is mathematically preferred, but this can be difficult.

4.1.4 Pixel Classification

Automatic grouping of pixels having a similar characteristic in a multivariate image is an important problem. One obstacle to successful modeling and prediction of crop yields using remote sensing imagery is the identification of image masks. Image masking involves restricting an analysis to a subset of a region's pixels rather than using all of the pixels in the scene. Cropland masking, where all sufficiently cropped pixels are included in the mask regardless of crop type, has been shown to generally improve crop yield forecasting ability, but it requires the availability of a land cover map depicting the location of cropland. In this research, we use an alternative image masking technique called statistical masking, which can be used for the development and implementation of regional crop yield forecasting models and eliminates the need for land cover maps. This is a pixel based approach, employing classical inference, and requires time series of satellite images and corresponding time series of the region's crop yields, and involves correlating historical, pixel-level VH indices values with historical regional yield values.

4.1.4.1 Approaches to image classification in crop yield forecasting

The purpose of image classification in the context of crop yield forecasting is to identify subsets of a region's pixels that lead to VH indices variable values that are optimal indicators of a particular crop's final yield.

4.1.4.1.1 Cropland masking

Cropland masking refers to using pixels dominated by crop production. Kastens (1998 and 2000) and Lee et al. (1999) obtained some of their best yield modeling results using this approach. Rasmussen (1998) used a percent-cropland map to improve his yield modeling by splitting the data into two categories based on cropland density and building different models for the two classes. Cropland masks are derived from existing land use/land cover maps. If relatively small amounts of land in a study area have been taken out of or put into agricultural crop production during a study period, a single mask can be obtained and applied to all years of data. Considering that all traditional agricultural crops are now grouped in the general class of ‘‘cropland’’, heavily cropped pixels are more prevalent in heavily cropped regions, which allows for the construction of well-populated masks dominated by cropland. However, the generation of such masks becomes difficult when low-producing regions are encountered, as well as in regions where cropland is widely interspersed with non-cropland.

4.1.4.1.2 Statistical Masking of Satellite Images

We used a different approach called statistical masking. All vegetation in a region integrates the season’s cumulative growing conditions in some fashion and may be more indicative of a crop’s potential than the crop itself. Thus, all pixels are considered for use in crop yield prediction. Each VH-based variable captures a

different aspect of the current growing season. This aspect manifests itself in different ways within the region's vegetation, suggesting that optimal masks for the different VH-based variables are not identical. Thus, for each crop, statistical masking generates a unique mask for each VH variable (Figure 4.4).

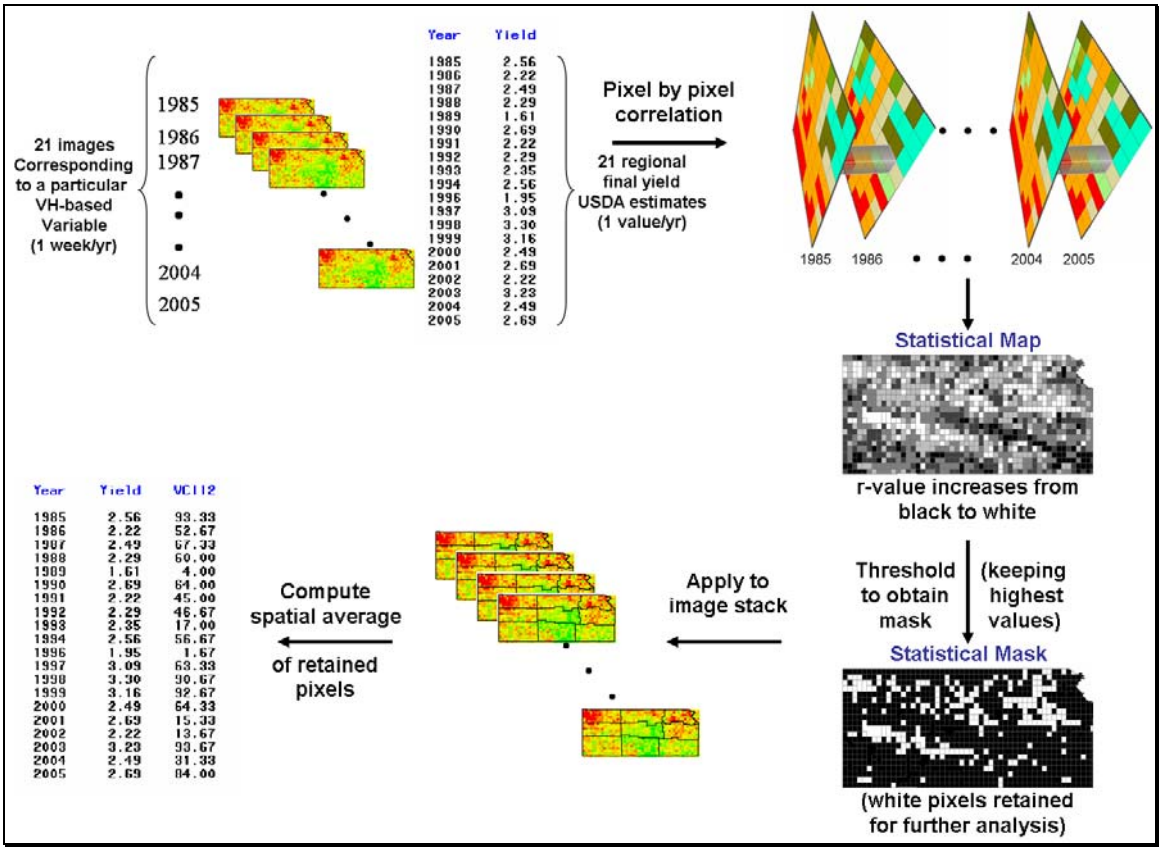

Figure 4.4: Flowchart for a single-variable application of the statistical masking technique. Example shown is for Kansas WW using VCI16 during the 21-year span 1985-2005

The technique is initiated by correlating each of the historical, pixel-level VH variable values with the region's final yield history (Figure 4.4). Once statistical maps

are available, the highest correlating pixels, thresholded so that some pre-specified number of pixels is included in the mask, are retained for further processing and evaluation of the variable at hand. Figure 4.4 shows a diagram outlining this process for a single variable. Unlike cropland masking, statistical masking can be applied to low-producing regions and regions possessing sparse crop distribution.

4.1.4.2 Mask application and evaluation

The generated masks were used to reduce the spatial stacks of VH indices variables to single time series (Figure 4.4). To do this, geo-political overlays boundaries in the GIS were used to define training areas in each image (Figure 4.4d). Then, the masks were applied to each variable stacks, and one-dimensional annual time series were subsequently produced for all the variables by averaging the values of the retained pixels. Each of the variable arrays that comprised the VH variable pool thus consisted of 21 points (21 years) (Figure 4.4e). Then, these time series were displayed and manipulated like any other data base values employing the full analytical and modeling power of ARCINFO and SAS systems (Figure 4.2f).

4.2 Summary

Knowledge of the history and origin of a data set is critical to successful modeling and predicting crop yields using remote sensing imagery. A review of map accuracy, map projections and datums can help to find where errors might be

introduced and to what magnitude. These errors can be minimized by a suite of readily available computer algorithms employing polynomials for rubber sheeting.

Automatic grouping of pixels having a similar characteristic in a multivariate image is an important problem. One obstacle to successful modeling and prediction of crop yields using remote sensing imagery is the identification of image masks. Image masking involves restricting an analysis to a subset of a region's pixels rather than using all of the pixels in the scene.

5. Methodology

The research strategy of this thesis was to extract the weather component from crop yield, NDVI and BT time series, and to correlate the weather related component of the crop yield with the corresponding NDVI and BT components. The latter two were expressed in the form of VH indices (Kogan 1997). The goal was to investigate the strength of the relationship and determine if the strongest correlation coincides with crop's critical period, which is the period when crop production is highly sensitive to weather conditions. In order to explain the methodology, results for winter wheat (WW), for total Kansas, are presented.

5.1 Crop Yield Time Series

Figure 5.1 shows the WW yield time series for total Kansas. Following Brockwell and Davis (2000) the WW yield time series was approximated by the following equation:

$$Y_t = T_t + dY_t,\ t = 1,\ldots,21, \qquad \textbf{(5.1)}$$

where T_t is a slowly changing function representing the deterministic component or trend that is regulated by agricultural technology, and dY_t is a random component regulated by weather fluctuations.

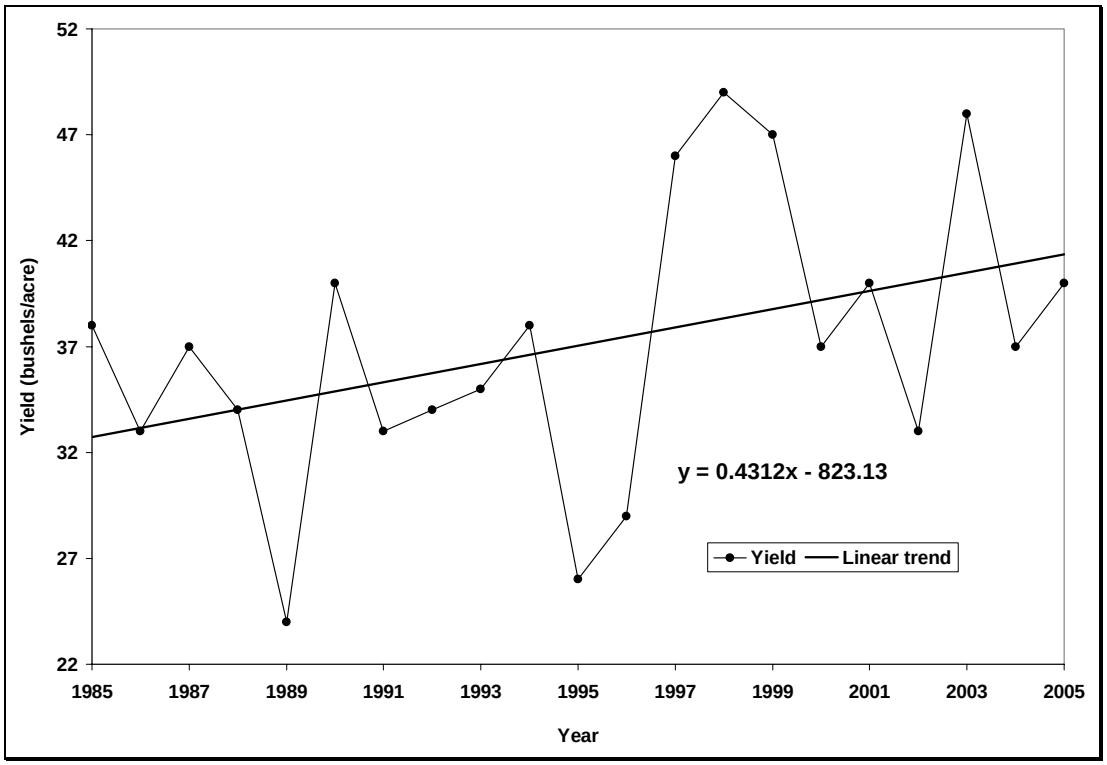

Figure 5.1: Winter wheat yield time series Kansas, U.S.

The deterministic component (T_t) was estimated using the least squares method. If the yield time series are longer than 30 to 35 years, they might be approximated by a second-degree polynomial of the form:

$$T_t = a_0 + a_1 t + a_2 t^2, \qquad \textbf{(5.2)}$$

by choosing the parameters a_0, a_1 and a_2 to minimize $\sum_{t=1}^{n} (Y_t - T_t)^2$. For shorter time series, as in our case, linear approximation is sufficient to satisfy the minimum criteria. The parameters of linear equations are shown in Table 5.1.

Table 5.1: Intercept and slope for winter wheat linear trend yield estimates

Crop	*Intercept*	*Slope*
WW	-823.13	0.4312

The random component (dY_t) was expressed as:

$$dY_t = Y_t - T_t. \qquad \textbf{(5.3)}$$

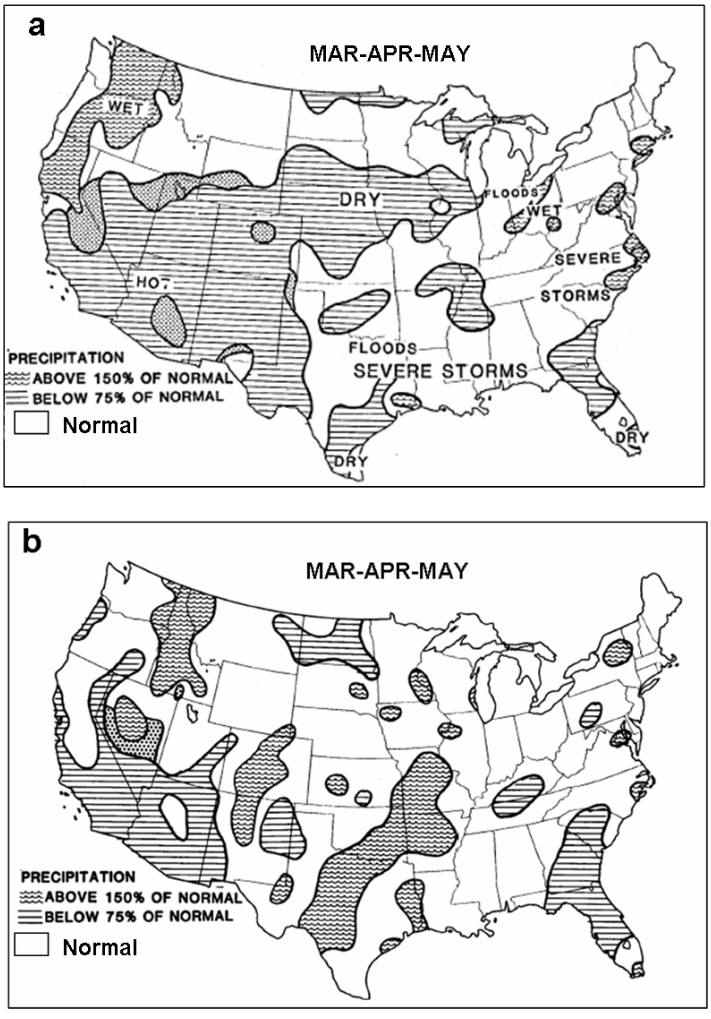

Figure 5.2: Percentage of normal precipitation in: (a) spring 1989, and (b) spring 1998 (WWCB 1989, 1998)

Figure 5.1 shows that in Kansas, WW yield increases due to technology improvement, management and cultivar changes. In addition, from Figure 5.1 we can see that WW yield variations from trend (dY) in Kansas are large. For example, in Kansas dY in 1989 and 1998 were estimated at -10.46 and 10.66, respectively, indicating a 30% yield reduction in 1989 due to unfavorable and a 28% increase due to favorable weather in 1998. In 1989, April and May were the driest on record in many counties in Kansas (Figure 5.2(a)). This contributed to a drought stressed crop. On the other hand, in 1998 spring rainfall was near and above normal (Figure 5.2(b)), which resulted in above trend WW yield.

Since dY and VH indices were similarly expressed as a deviation from climatology (from trend for crop yield and from maximum to minimum range for NDVI and BT time series), further examination included correlation and regression analysis of these deviations to investigate the association among them for the area of WW growth.

5.2 Regression Analysis

A useful starting point in any multiple regression analysis is to compute correlations among all variables. This provides a first look at the simple linear relationships among them.

5.2.1 The Correlation Matrix

For WW, VCI (weeks 12 to 21) variables have significant correlations (at the α= 0.01 level of significance) with dY (0.82-0.88). VCI15 has the highest correlation with dY. It would account for 77% (0 $.88^2$) of the variation in dY if used separately as the only independent variable in the regression model. TCI (weeks 16 to 18) variables also have significant correlation (p-value <0.001) with dY (0.69-0.73) (Table 5.2).

Table 5.2: Correlation Matrix dY (WW) with VCI (weeks 12 to 21) and TCI (weeks 16 to18) Kansas

Pearson Correlation Coefficients	
	dY
VCI12	0.8222
VCI13	0.8645
VCI14	0.8641
VCI15	0.8772
VCI16	0.8724
VCI17	0.8686
VCI18	0.8749
VCI19	0.8722
VCI20	0.8753
VCI21	0.8433
TCI16	0.6930
TCI17	0.7318
TCI18	0.7205

As seen in Table 5.2, dY is highly correlated with VCI (weeks 12 to 21) and TCI (weeks 16 to 18) April through to May. This period is known as critical for WW

yield in Kansas because WW goes through the reproductive period from the end of biomass development to the beginning of maturation. The actual number of kernels that will form in the spike is determined at this stage (Shroyer et al. 2004). Positive correlation of dY with VCI and TCI indicates that above trend WW yield is associated with favorable moisture and thermal conditions (VCI and TCI above 60) and below trend WW yield is associated with moisture and thermal stress (VCI and TCI below 40).

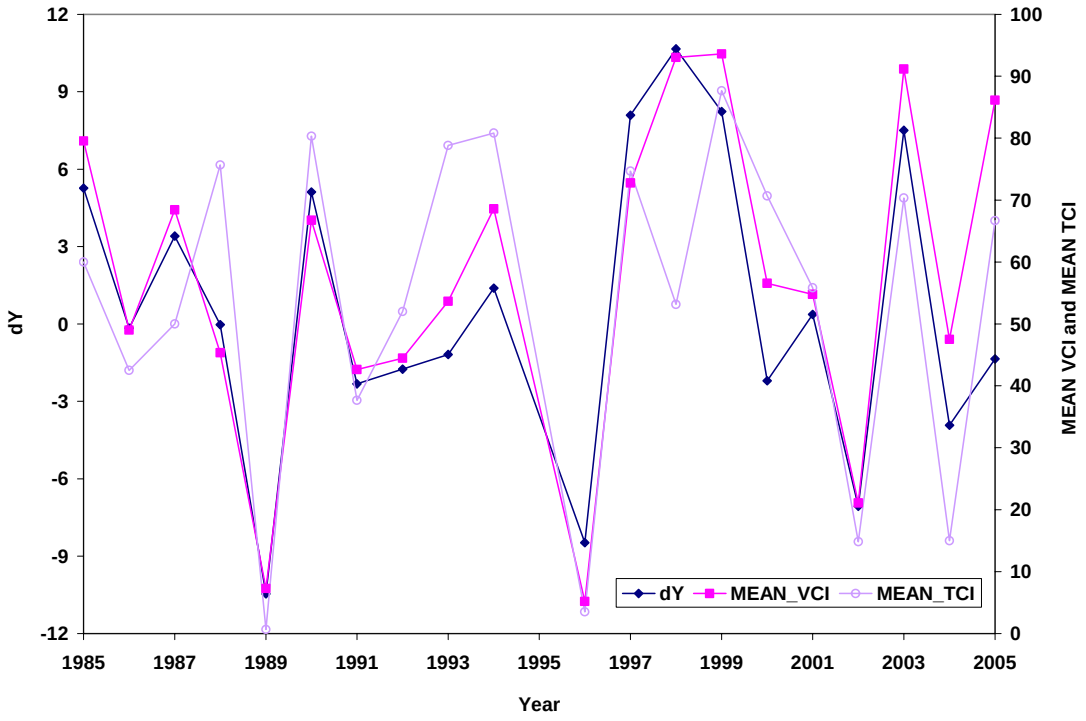

Figure 5.3: Dynamics of dY (WW), MEAN VCI (weeks 12 to 21) and MEAN TCI (weeks 16 to 18) Kansas, U.S.

Average VCI and TCI ($\overline{vci}, \overline{tci}$), for weeks with significant Pearson's correlation coefficients (at the α= 0.01 level of significance) of dY with VH indices, were used as a predictors of dY. Figure 5.3 shows dynamics of dY (WW), MEAN VCI (weeks 12 to 21) and MEAN TCI (weeks 16 to 18) for total Kansas. This figure shows how closely VH dynamics follows dY dynamics. This is a strong indication that VH indices can be used as predictors in a forecast model for WW yield in Kansas.

Table 5.3 shows correlation coefficients of dY (WW) with $\overline{vci}$ (weeks 12 to 21) and $\overline{tci}$ (weeks 16 to 18). These correlation coefficients are significantly different from zero with p-value ($p<0.001$). For WW, $\overline{vci}$ would account for 81% (0.90^2) of the variation in dY if used separately as the only independent variable in the regression model. Besides, $\overline{tci}$ would account for 53% (0.73^2) of the variation in dY (WW) if used separately as the only independent variable in the regression model.

Table 5.3: Correlation matrix of dY (WW) with MEAN VCI ($\overline{vci}$) and MEAN TCI ($\overline{tci}$), Kansas, U.S.

Pearson Correlation Coefficients	
	dY
MEAN VCI	0.9004
MEAN TCI	0.7267

5.2.2 Multiple Regression Results

The results of fitting the ordinary least squares (OLS) regression model approximated by Equation 5.4 to Kansas are shown in Table 5.4:

$$dY = a_0 + b_1 \cdot \overline{vci} + b_2 \cdot \overline{tci} + \varepsilon \,. \qquad \textbf{(5.4)}$$

Table 5.4: Results of the regression of dY (WW) on the two independent variables MEAN VCI and MEAN TCI, Kansas

Source	***DF***	***Sum of Squares***	***Mean Square***	***F Value***	***Pr > F***
Model	2	510.8820	255.4410	37.1885	<.0001
Error	17	116.7699	6.8688	–	–
Corrected Total	19	627.6519	–	–	–

R-Square	***Root MSE***
0.8140	2.6208

Parameter	***Estimate***	***Standard Error***	***t Value***	***Pr > |t|***
Intercept	-11.0062	1.4769	-7.4524	0.0000
MEAN VCI	0.1836	0.0359	5.1110	0.0001
MEAN TCI	0.0192	0.0349	0.5498	0.5896

Table 5.4 shows that there is a strong relationship between dY and the independent variables. The coefficient of determination, R-Square, is 0.81 or 81% of the sum of squares in dY (WW) can be associated with the variation in these independent variables. The test of the composite hypothesis that all regression

coefficients are zero is highly significant with F-value of 37.19 compared to F (0.01, 2, 17) = 6.11. The residual standard error (Root MSE) is 2.6208 with 17 degrees of freedom is an unbiased estimate of σ if this model is the correct model.

This may not be the correct model because (1) important variables may have been excluded, or (2) the mathematical form of the model may not be correct. Including unimportant variables generally will not bias the estimate of σ^2. Therefore, s^2 must be regarded as the tentative best estimate of σ^2 and will be used for tests of significance and for computing the standard errors of the estimates.

5.2.3 Model Validation

Validation of a fitted regression equation is the confirmation that the model is effective for the purpose for which it was intended. This is not equivalent to demonstrating that the fitted equation agrees well with the data from which it was computed. Validation of the model requires assessing the effectiveness of the fitted equation against an independent set of data, and is essential if confidence in the model is to be expected.

Results from the regression analysis R-Square, MSE do not necessarily reflect the degree of agreement one might obtain from future applications of the equation. In addition, least squares estimation has given the best possible agreement of the model Equation 5.4 with the observed data. As a result, the fitted equation is expected to fit the data from which it was computed better than it will fit an independent set of data.

The fitted model should be validated for the specific objective for which it was planned. An equation that is good for predicting Y_i in a given region of the X-space might be a poor predictor in another region of the X-space. Two criteria are of interest:

1. Does the fitted regression equation provide unbiased predictions of the quantities of interest?
2. Is the precision of the prediction good enough (the variance small enough) to accomplish the objective of the study?

When fitting a model to noisy data, we make the assumption that the data have been generated from some model (the 'truth') by making predictions at given values of the inputs, then adding some amount of noise to each point, where the noise is drawn from a normal distribution with an unknown variance. Our task is to discover both this model and the width of the noise distribution. In doing so, we look for a compromise between bias, where our model does not follow the right trend in the data (and so does not match well with the underlying truth), and variance, where our model fits the data points too closely, and so 'chases' the noise rather than trying to capture the true trend. These two extremes are known as under-fitting and over-fitting.

An important concept in this context is the number of parameters in a model. As this number increases, the model can bend in more complicated ways. If the number of parameters in our model is larger than that in the truth, then we risk over-fitting, and if our model contains fewer parameters than the truth, we could under-fit.

5.2.3.1 Mean Square Error of Prediction

Bias and variance are incorporated into a single measure called the mean squared error of prediction (MSEP). Mean squared error of prediction is defined as the average squared difference between independent observations and predictions from the fitted equation for the corresponding values of the independent variables. The mean squared error of prediction incorporates both the variance of prediction and the square of the bias of the prediction:

$$MSEP = \frac{(n-1)s^2(\delta)}{n} + (\bar{\delta})^2 . \qquad \textbf{(5.5)}$$

In regression models, prediction error refers to the expected squared difference between the future response and its prediction from the model:

$$PE = E(y - \hat{y})^2 . \qquad \textbf{(5.6)}$$

The expectation refers to repeated sampling from the true population. Here we consider how well will our model predict dY (crop yield anomalies)? To answer this question, we could look at the average residual squared error for all n = 20 responses:

$$SSE/n = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / n = 116.77/20 = 5.84 . \qquad \textbf{(5.7)}$$

but this will tend to be too optimistic; it will probably underestimate the true prediction error. The reason is that we are using the same data to assess the model as

were used to fit it, using parameters estimates that are fine-tuned to our particular data set.

A method for improving on Equation 5.7 is to divide by $n - p$ instead of n, where p is the number of parameters to be estimated. This gives the usual unbiased estimate of residual variance (MSE) $\hat{\sigma}^2 = \sum (y_i - \hat{y}_i)^2 / (n - p) =$ 6.86. However, bigger corrections are necessary for the prediction problem.

5.2.3.2 Cross Validation

In order to get more realistic estimates of prediction error, we would like to have a test sample that is separated from our training sample. Ideally this would come in the form of some new data from the same population that produced our original sample. Usually, additional data are not often available, for reasons of logistic or cost. Methods have been devised for estimating the mean squared error of prediction, MSEP, when it is not practical to obtain new independent data. One approach is cross-validation that uses part of the available data to fit the model, and a different part to test it. With large amounts of data, an alternative is to use it for both estimation and validation. One approach is to divide the data set into two representative halves; one-half is then used to develop the regression model and the other half is used for validation of the model. Snee (1977) suggests that the total sample size should be greater than $2p + 25$ before splitting the sample is considered. With smaller data sets like our case, *K*-fold cross-validation makes more efficient use of the available data.

In *K*-fold cross-validation we split the data into *K* parts. Let *k (i)* be the part containing observation *i*. Denote by $\hat{y}_i^{-k(i)}$ the fitted value for observation *i*, computed with the $k\,(i)_{th}$ part of the data removed. Then the cross-validation estimate of prediction error is:

$$CV = 1/n\sum_{i=1}^{n}(y_i - \hat{y}_i^{-k(i)})^2 . \qquad \textbf{(5.8)}$$

We chose *k* = n, resulting in leave-one-out cross-validation. For each observation *i*, we refit the model leaving that observation out of the data, and then compute the predicted value for the i_{th} observation, denoted by $\hat{y}_i^{-i}$. We do this for each observation and then compute the average cross-validation sum of squares $CV = \sum(y_i - \hat{y}_i^{-i})^2 / n$.

5.2.3.3 Advantages of Using Cross-Validation

Why use cross-validation when simpler alternatives such as MSE, C_p, AIC, SBC, etc. are available? The main reason is that for fitting problems more complicated than least squares, the number of parameters *p* is not known. The MSE, C_p, AIC and SBC statistics require knowledge of *p*, while cross-validation does not. Cross-validation tends to give similar answers as standard methods in simple problems and its real power stems from its applicability in more complex situations. A second advantage of cross-validation is its robustness. The C_p and SBC statistics require a

roughly correct working model to obtain the estimate $\hat{\sigma}^2$. Cross-validation does not require this and will work well even if the models being assessed are far from correct.

5.2.3.4 Predicted vs. Observed

For model validation we regressed observed yield versus independently simulated yield and the corresponding statistics were generated (Table 5.5). The overall correlation coefficient 0.8933 is good. An R-Square value of 0.7980 shows that in most years, WW yield in Kansas can be modeled by variables considered in model Equation 5.4. For WW the model forecast captured 80% of the variability in yield anomalies.

The average prediction bias is 0.0276 (systematic error). The variance of the prediction error is 8.6023 or the standard error of prediction (SEP) is 2.9330 (non-systematic error). The standard error of the estimated mean bias is 0.6558 (Table 5.5). T-tests of the hypothesis that the bias is zero gives t = 0.0421 which, with 19 degrees of freedom and $\alpha = 0.05$, is not significant.

The mean square error of prediction (MSEP) is $= \frac{19(8.6023)}{20} + (0.0276)^2 = 8.6031$. The simplest and most efficient measure of the uncertainty on future predictions is the RMSEP (Willmott 1982, Westad and Martens 2000). The root mean square error of prediction (RMSEP) is 2.9331 an approximate

7.8% error in prediction. This is an acceptable error in this kind of application, which has uncertainty in measurements.

Table 5.5: Statistics of an independent testing for model Equation (5.4) WW, Kansas

Crop	*Correlation*	*R-Square*	*Bias*	*Variance*	*SEP*	*Std Error*	*RMSEP*
WW	0.8933	0.7980	0.0276	8.6023	2.9330	0.6558	2.9331

Figure 5.4 displays observed versus independently simulated crop yield time series of WW 1985 to 2005 for total Kansas. The results of cross-validation analysis suggest that winter wheat yield can be forecast with fairly high accuracy based on remote sensing data particularly VH indices (Figure 5.4 and Table 5.5). For WW total Kansas the model forecasts captured 82% of the variability in yield anomalies.

Interestingly, the models did very well at forecasting extremely low yields such as in 1989, 1996 and 2002. These years were characterized by dry conditions. For example, in 1989, most of the State struggled through nearly a year of less than normal rainfall. Temperatures were generally mild but precipitation was on the short side through the fall and early winter. Moisture was on the short side in March. April was the driest on record for many counties in the State. In 1996, by late November, emergence and growth was severely stunted by dry conditions. Dry soil and high winds, over winter, reduced crop conditions. Several hard freezes in mid-March caused severe damage in the western third of the State. Large acreages were abandoned as conditions declined in April.

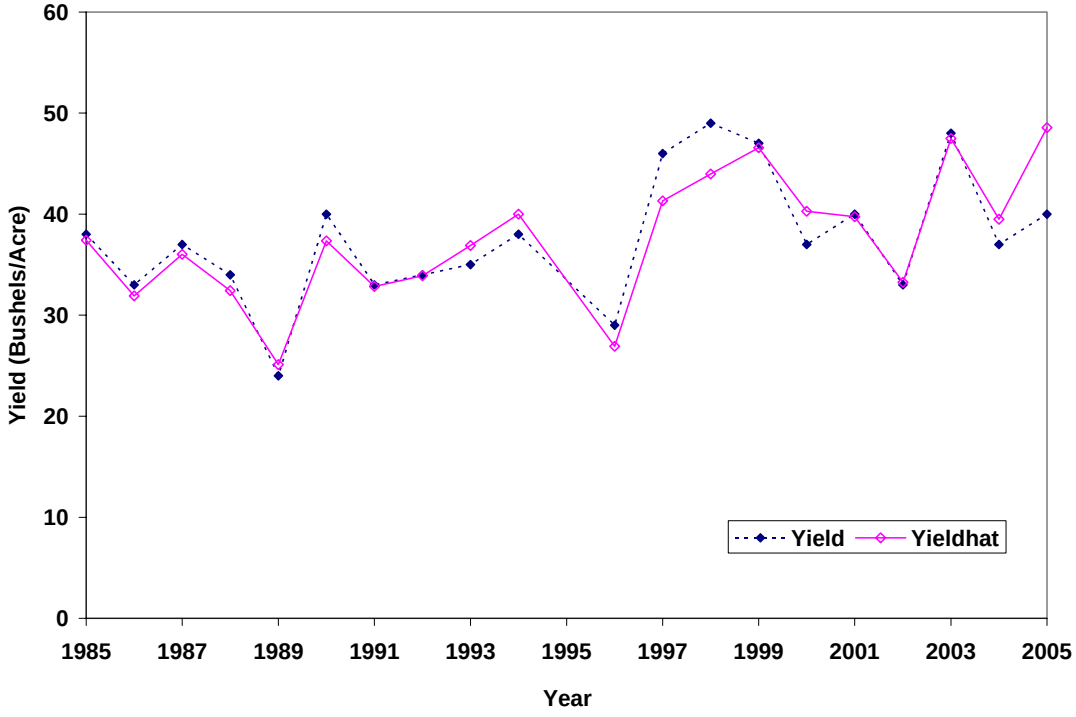

Figure 5.4: Observed yield (WW) versus independently simulated yield (yieldhat), Kansas

In general, the model correctly predicted the direction of yield anomalies for most of the years. That is, the models correctly predicted whether the yield would be above or below the trend.

5.3 Alternative Statistical Approaches

Alternative statistical approaches to the multiple linear regression that we used may improve accuracies. In addition, using VCI and TCI values independently without taking the average allows us to see what variables are more important in

predicting crop yield. However, VH indices of neighboring weeks are highly correlated as seen in Table 5.6.

Table 5.6: Correlation Matrix among dY (WW) and VCI (weeks 12 to 21), Kansas

	dY	***VCI12***	***VCI13***	***VCI14***	***VCI15***	***VCI16***	***VCI17***	***VCI18***	***VCI19***	***VCI20***	***VCI21***
dY	1.0000	0.8222	0.8645	0.8641	0.8772	0.8724	0.8686	0.8749	0.8722	0.8753	0.8433
VCI12	0.8222	1.0000	0.9699	0.9407	0.9125	0.8986	0.8595	0.8139	0.7675	0.7228	0.6580
VCI13	0.8645	0.9699	1.0000	0.9905	0.9728	0.9589	0.9313	0.8941	0.8553	0.8135	0.7486
VCI14	0.8641	0.9407	0.9905	1.0000	0.9913	0.9821	0.9577	0.9284	0.8912	0.8498	0.7827
VCI15	0.8772	0.9125	0.9728	0.9913	1.0000	0.9952	0.9795	0.9575	0.9267	0.8893	0.8281
VCI16	0.8724	0.8986	0.9589	0.9821	0.9952	1.0000	0.9918	0.9741	0.9471	0.9128	0.8562
VCI17	0.8686	0.8595	0.9313	0.9577	0.9795	0.9918	1.0000	0.9908	0.9735	0.9465	0.9027
VCI18	0.8749	0.8139	0.8941	0.9284	0.9575	0.9741	0.9908	1.0000	0.9928	0.9754	0.9438
VCI19	0.8722	0.7675	0.8553	0.8912	0.9267	0.9471	0.9735	0.9928	1.0000	0.9900	0.9704
VCI20	0.8753	0.7228	0.8135	0.8498	0.8893	0.9128	0.9465	0.9754	0.9900	1.0000	0.9809
VCI21	0.8433	0.6580	0.7486	0.7827	0.8281	0.8562	0.9027	0.9438	0.9704	0.9809	1.0000
TCI16	0.6930	0.6514	0.7032	0.7337	0.7411	0.7486	0.7685	0.7875	0.7678	0.7766	0.7318
TCI17	0.7318	0.6617	0.6866	0.7035	0.7072	0.7126	0.7249	0.7344	0.7062	0.7179	0.6778
TCI18	0.7205	0.6316	0.6789	0.6986	0.7127	0.7227	0.7493	0.7731	0.7602	0.7690	0.7366

For example correlation coefficient between VCI17 with VCI18 (weeks 17 and 18, April) is 0.99. This high correlation among the independent variables is called collinearity. In situations where collinearity is severe, ordinary least squares (OLS) produces unbiased estimators at a potentially high cost in estimator variability (Gunst and Mason 1980, Goutis 1996, Chatterjee et al. 2000). Therefore, we considered an alternative estimation procedure.

5.3.1 Principal Component Analysis

We find the main variations in the satellite measurements by decomposing the variables with Principal Component Analysis (PCA). The data matrix is decomposed into eigenvalues, eigenvectors (loadings), principal components (PCs, scores) and residuals. We will interpret these results to see whether we can say something about the satellite data.

5.3.1.1 Eigenvalues and Eigenvectors of the Correlation Matrix

The first part of Table 5.7 shows the eigenvalues of the correlation matrix for the thirteen independent variables in Table 5.2. From the 'Eigenvalue' column, it is clear that the first principal component has a very large variance (11.1228), the second has much smaller variance (1.0309), and the others have negligible variances. The 'Difference' column gives the differences between adjacent eigenvalues. This statistic shows the rate of decrease in variances of the PCs.

The proportion of total variation accounted for by each of the components is obtained by dividing each of the eigenvalues by the total variation. These quantities are given in the 'Proportion' column (Table 5.7). The first component accounts for 86% of the total variation, a result that is typical when a single factor, in this case moisture (VCI), is a common factor in the variability among the original variables.

The cumulative proportions printed in the 'Cumulative' column indicate that 99.40% of the total variation in the thirteen variables is explained by only four components.

Table 5.7: Eigenvalues of the correlation matrix, Kansas

Eigenvalues of the Correlation Matrix				
	Eigenvalue	***Difference***	***Proportion***	***Cumulative***
1	11.1228	10.0919	0.8556	0.8556
2	1.0309	0.3963	0.0793	0.9349
3	0.6346	0.5498	0.0488	0.9837
4	0.0848	0.0365	0.0065	0.9902
5	0.0483	0.0099	0.0037	0.9940
6	0.0384	0.0192	0.0030	0.9969
7	0.0192	0.0091	0.0015	0.9984
8	0.0102	0.0048	0.0008	0.9992
9	0.0054	0.0024	0.0004	0.9996
10	0.0030	0.0016	0.0002	0.9998
11	0.0014	0.0005	0.0001	0.9999
12	0.0009	0.0007	0.0001	1.0000
13	0.0002	–	0.0000	1.0000

Table 5.8 shows the eigenvectors for each of the PCs. These coefficients, which relate the components to the original variables listed on the first column, are scaled so that their sum of squares is unity. This allows for finding which of the original variables dominate a component. The coefficients of the first PC show a positive relationship with all variables, with somewhat larger contributions from VCI15 (0.2910), VCI16 (0.2932) VCI17 (0.2949) and VCI18 (0.2946). As expected,

these components are in the middle of the critical period of WW. The second component is dominated by TCI (thermal conditions).

Table 5.8: Eigenvectors of the correlation matrix, Kansas

	Prin1	***Prin2***	***Prin3***	***Prin4***	***Prin5***	***Prin6***	***Prin7***	***Prin8***	***Prin9***	***Prin10***
VCI12	0.2624	-0.2455	0.4617	0.5946	0.2270	-0.0644	-0.4444	-0.0771	0.1650	-0.0075
VCI13	0.2806	-0.2361	0.3043	0.1138	0.0582	-0.1586	0.4974	0.0647	-0.4576	0.1049
VCI14	0.2868	-0.2117	0.2194	-0.2154	0.0100	-0.0212	0.4241	0.0215	0.1123	-0.5234
VCI15	0.2910	-0.1969	0.1044	-0.2913	-0.1466	0.0673	0.1435	-0.0234	0.5041	0.6155
VCI16	0.2932	-0.1808	0.0349	-0.2626	-0.1661	0.1292	-0.2612	-0.0266	0.1200	-0.0028
VCI17	0.2949	-0.1302	-0.0833	-0.2059	-0.1635	0.0796	-0.3657	0.1939	-0.6488	0.2118
VCI18	0.2946	-0.0721	-0.2003	-0.1313	-0.0805	-0.0006	-0.2399	0.1238	0.0929	-0.3654
VCI19	0.2892	-0.0605	-0.3176	0.0012	-0.0146	-0.1059	-0.0936	-0.0037	0.1348	-0.3194
VCI20	0.2839	0.0010	-0.3855	0.1302	0.1252	0.0307	0.1128	-0.8147	-0.1345	0.0944
VCI21	0.2707	0.0196	-0.5064	0.4344	0.1656	0.0953	0.2384	0.5148	0.1159	0.1750
TCI16	0.2552	0.4718	0.1187	-0.3564	0.7214	-0.1696	-0.0946	0.0750	-0.0011	0.0761
TCI17	0.2455	0.5097	0.2437	0.1422	-0.1936	0.7249	0.0885	-0.0294	-0.0324	-0.0924
TCI18	0.2508	0.5091	0.0866	0.1439	-0.5113	-0.6060	0.0066	-0.0083	0.0441	0.0439

5.3.1.2 The Score Plot

The score plot (PCs plot), also called a map of samples, displays information about the samples in the PCA model (Figure 5.5). This graph shows the projected locations of the samples onto eigenvectors, and by studying patterns we may find the meaning of the PCs.

Figure 5.5 shows the score plot for total Kansas; you will notice that the 20 samples are not arranged in a random way on the plot. When you move from the left

to the right part of the plot, you first encounter samples 1989, 1996, 2002 and finally 2003 and 1999.

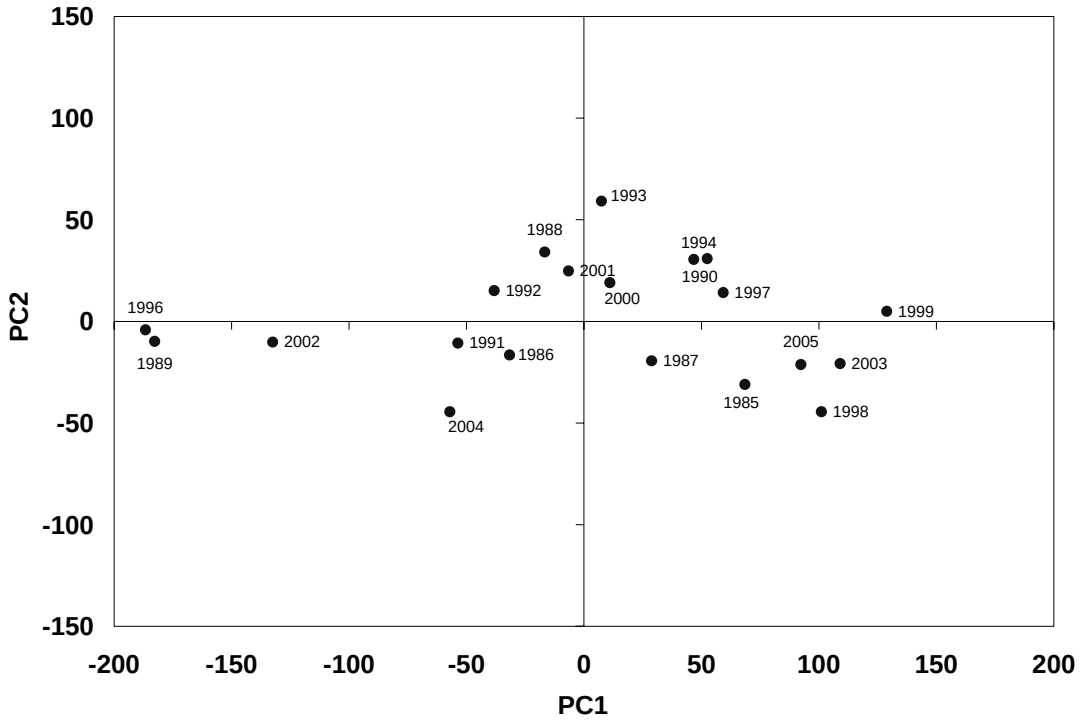

Figure 5.5: Score plot, Kansas

Reading the literature, we found the following facts: in 1989, most of the State struggled through nearly a year of less than normal rainfall. Precipitation was on the short side through the fall and early winter. Moisture continued on the short side in March. April was the driest on record for many counties in the State. The May 1 yield forecast was 21 bushels per acre; the final was 24 bushels. Total wheat production was 213,600,000 bushels, the lowest production since 1963.

In 1996, by late November, emergence and growth were severely stunted by dry conditions. Dry soil and high winds, over winter, reduced crop conditions. Large acreages were abandoned as conditions continued to decline in April. Total production was 255,200,000 with a yield of 29 bushels.

On the other hand, 1999, 2003 were years with average weather conditions. These years had large yields.

5.3.2 Principal Component Regression

Principal component regression has been suggested as a means of obtaining estimates with smaller mean squared errors in the presence of collinearity. This alternative has the potential to produce more precision in the estimated coefficients and smaller prediction errors when the predictions are generated using data other than those used for estimation (Draper and Smith 1981, Myers 1986).

Using PCR methodology, the variables in model Equation (5.9) were transformed into a new set of orthogonal or uncorrelated variables called principal components (PCs) of the correlation matrix. This transformation ranks the new orthogonal variables in order of their importance and the procedure then involves eliminating some of the PCs to get a reduction in variance. After elimination of the least important PCs, a multiple regression analysis of the response variable dY against the reduced set of PCs was performed using OLS estimation. Since the PCs are orthogonal, they are pair-wise independent and hence OLS is appropriate. Once the

regression coefficients for the reduced set of orthogonal variables were calculated, they were mathematically transformed into a new set of coefficients that correspond to the original or initial correlated set of variables in model Equation (5.9). These new coefficients are principal component estimators (Gunst and Mason 1980).

$$dY = c_1 + a_1\ VCI_{12} + a_2\ VCI_{13} + a_3\ VCI_{14} + a_4\ VCI_{15} + a_5\ VCI_{16} + a_6\ VCI_{17}$$

$$a_7\ VCI_{18} + a_8\ VCI_{19} + a_9\ VCI_{20} + a_{10}\ VCI_{21} + b_1\ TCI_{16} + b_2\ TCI_{17} + b_3\ TCI_{18}. \quad \textbf{(5.9)}$$

The model Equation (5.9) can be expressed as:

$$Y = \beta_0 + \sum_{i=1}^{13} \beta_i \cdot X_i + \varepsilon \qquad \textbf{(5.10)}$$

where dY = Y , VCI_j and $TCI_j = X_j$ and n = 13. Let $\bar{y}$ and $\bar{x}_j$ be the means of Y and X_j, respectively. Also, let $s_y = (\sum_{i=1}^{n} (y_i - \bar{y})^2 /(n-1))^{1/2}$ and $s_j = (\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 /(n-1))^{1/2}$ be the standard deviations of the response and j_{th} predictor variable, respectively. Equation (5.10) can be written in terms of standardized variables as:

$$\tilde{Y} = \theta_1 \cdot \tilde{X}_1 + \theta_2 \cdot \tilde{X}_2 + ... + \theta_7 \cdot \tilde{X}_{13} + \varepsilon' \qquad \textbf{(5.11)}$$

where $\widetilde{Y} = (y_i - \bar{y})/s_y$ is the standardized version of the response variable (dY) and $\widetilde{X} = (x_{ij} - \bar{x}_j)/s_j$ is the standardized version of the j_{th} predictor variable.

The estimated coefficients satisfy:

$$\beta_j = (s_y / s_j) \cdot \theta_j, \quad \text{j} = 1,2,...,13 \qquad \textbf{(5.12)}$$

$$\beta_0 = \bar{y} - \beta_1 \cdot \bar{x}_1 - \beta_2 \cdot \bar{x}_2 - ... - \beta_{13} \cdot \bar{x}_{13}. \qquad \textbf{(5.13)}$$

The principal components of the standardized predictor variables are given by

$$Z_j = \sum_{i=1}^{13} c_{ij} \cdot X_i, \quad \text{j} = 1,...,13 \qquad \textbf{(5.14)}$$

where c_{ij} are elements of the eigenvectors of the matrix of bivariate correlation between pairs of the explanatory variables. The model in Equation (5.11) may be written in terms of the principal components as:

$$\tilde{Y} = \alpha_1 \cdot Z_1 + \alpha_2 \cdot Z_2 + ... + \alpha_7 \cdot Z_{13} + \varepsilon', \qquad \textbf{(5.15)}$$

where the α's and θ's are related as

$$\alpha_j = \sum_{i=1}^{13} c_{ij} \cdot \theta_i, \text{ j=1, 2,…, 13} \qquad \textbf{(5.16)}$$

or conversely

$$\theta_j = \sum_{i=1}^{13} c_{ij} \cdot \alpha_i, \text{ j=1, 2,…, 13.} \qquad \textbf{(5.17)}$$

5.3.2.1 Residual Validation Variance

This plot illustrates how much of the variation in the response (dY) is accounted for by each different component. Total residual variance is computed as the sum of squares of the residuals, divided by the number of degrees of freedom. This variance can be computed after 0, 1, 2… components have been extracted from the data.

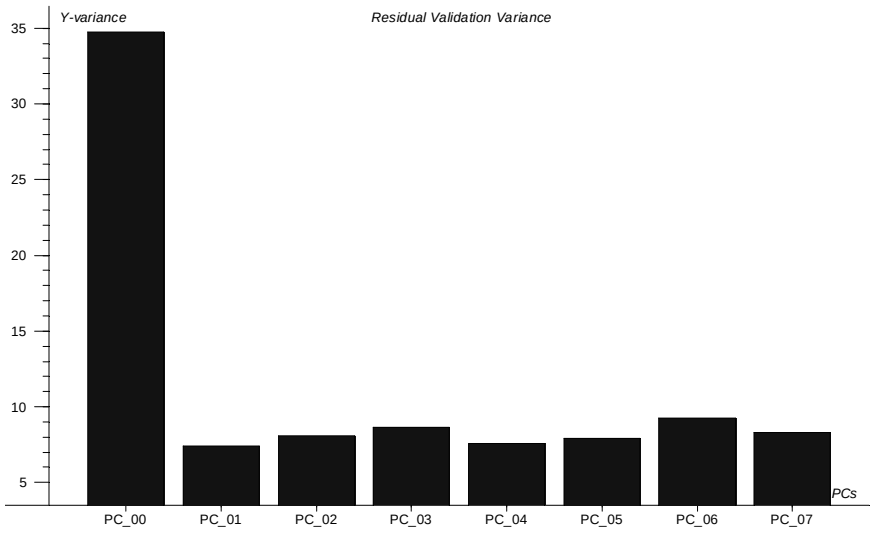

Figure 5.6: Residual validation variance (PCR) WW, Kansas

Models with small (close to zero) total residual variance explain most of the variation in dY. Ideally, one would like to have simple models, where the residual variance goes to zero with as few components as possible.

The residual variance plot is excellent for selection of the optimal number of components in the model. From Figure 5.6 we see that the residual variance decreases

until PC 01 is reached (7.4427). Then, the residual variance increases again due to over-fitting. Therefore, we will use a model with one PC.

5.3.2.2 Interpretation of the Regression Coefficients

The regression coefficients are used to calculate the response variable (dY) from the satellite measurements. The size of the coefficients gives an indication of which variables have an important impact on the response variable. Regression coefficients summarize the relationship between all predictors and the response variable. For PCR, the regression coefficients can be computed for any number of components. For total Kansas data, the regression coefficients were calculated for one PC. They summarize the relationship between the predictors and the response, as it is approximated by a model with one component.

In order to simplify the final model and to make it more reliable and stable, non-significant predictors were eliminated following Westad and Martens (2000). We used cross-validation above to determine the number of factors which should be retained. When we did cross-validation, we retained all the estimates of the regression coefficients for each candidate model to use them to estimate the variance of the coefficients and test if they were significantly different from zero. Once we had the test, we proceeded to use it to decide which variables should be dropped from the original data set. For each variable, we calculated the difference between the estimated coefficient B_i in a sub-model and the B_{tot} for the total model. We took the

sum of the squares of the differences (SSD) in all sub-models to get an expression of the variance of the estimated coefficient for a variable in the model. $SSD = \sum_{i=1}^{N} (B_{tot} - B_i)^2$. B_{tot} is the regression coefficient for the model using all the n=20 observations, B_i is the regression coefficient for the model using all observations except the observation left out in cross validation.

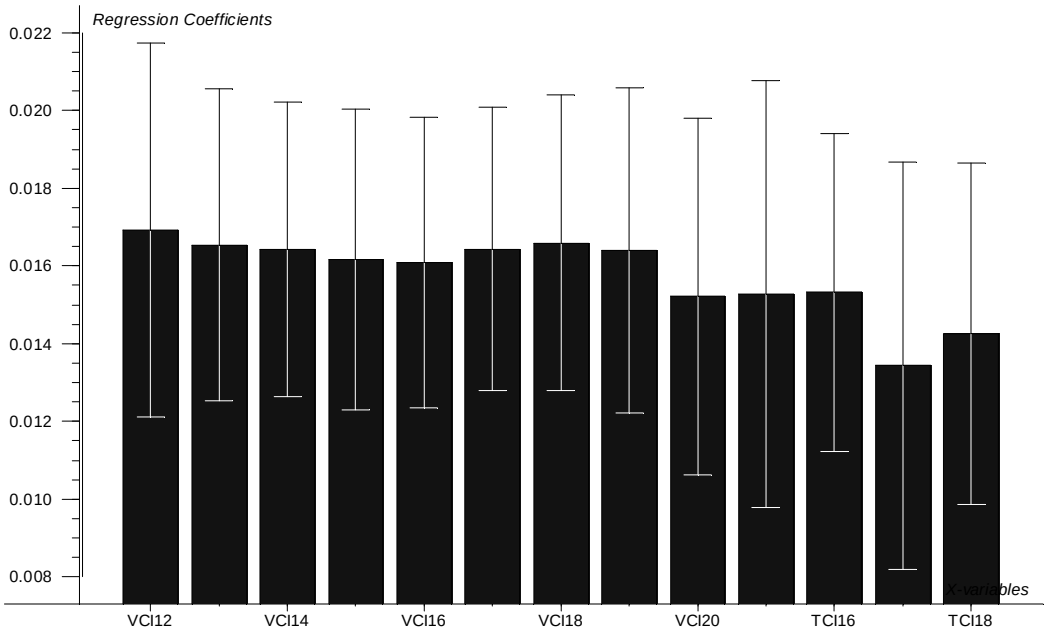

Figure 5.7: Regression coefficients for model Equation (5.9) calculated using PCR, Kansas

When these variances were estimated, they were used to find significant parameters. A Student's t-test was performed for each estimated coefficient relative to the square root of its estimated variance, giving the significance level for each parameter. Figure 5.7 shows the resulting significant predictors with uncertainty

limits that correspond to two standard deviations. Figure 5.7 shows that uncertainty limits for significant predictors do not cross the zero line.

Figure 5.7 shows the regression coefficients for model Equation (5.9) calculated using PCR methodology (Table 5.9). Each predictor variable defines one bar of the plot. All predictors VCI (weeks 12 to 21) and TCI (weeks 16 to 18) have large regression coefficients; thus, they play an important role in the regression model. All are positive showing a positive link with the response dY.

Table 5.9: Coefficients for model Equation (5.9) calculated following principal components regression methodology total Kansas

VCI17	***VCI18***	***VCI19***	***VCI20***	***VCI21***	***VCI22***	***VCI23***	***Intercept***
0.085884	0.08963	0.094772	0.089554	0.090925	0.090361	0.08911	64.15381

5.3.2.3 Predicted vs. Observed

For model validation we regressed observed yield versus independently simulated yield and the corresponding statistics were generated (Table 5.10). The overall correlation coefficient 0.9019 is good. An R-Square value of 0.8134 shows that in most years, WW yield in Kansas can be modeled by variables considered in model Equation (5.9) with regression coefficients estimated using PCR methodology. For WW the model forecast captured 81% of the variability in yield anomalies.

The average prediction bias is 0.0279 (systematic error). The variance of the prediction error is 7.8333 or the standard error of prediction (SEP) is 2.7988 (non-systematic error). The standard error of the estimated mean bias is 0.6258 (Table 5.10). T-tests of the hypothesis that the bias is zero gives t = 0.0446 which, with 19 degrees of freedom and $\alpha = 0.05$, is not significant.

Table 5.10: Statistics of an independent testing for model Equation (5.9) with coefficients calculated following PCR methodology, WW, Kansas, U.S.

Crop	*Correlation*	*R-Square*	*Bias*	*Variance*	*SEP*	*Std Error*	*RMSEP*
WW	0.9019	0.8134	0.0279	7.8333	2.7988	0.6258	2.7281

The mean square error of prediction (MSEP) is $= \frac{19(7.8333)}{20} + (0.0279)^2 = 7.4424$. The simplest and most efficient measure of the uncertainty on future predictions is the root mean square error of prediction (RMSEP) (Willmott 1982, Westad and Martens 2000, Dingstad et al. 2004, Anderssen et al. 2006). The RMSEP is 2.7281 an approximate 7% error in prediction for total Kansas using PCR methodology. This is an acceptable error in this kind of application, which has uncertainty in measurements. Therefore, we can conclude that models based on Equation (5.9) with the coefficients estimated using PCR methodology perform well.

Figure 5.8 displays observed versus independently simulated crop yield time series of WW 1985 through to 2005 for total Kansas. The results of cross-validation analysis suggest that WW yield can be forecast with fairly high accuracy based on

remote sensing data particularly VH indices (Figure 5.8). For WW total Kansas the model forecasts captured 82% of the variability in yield anomalies.

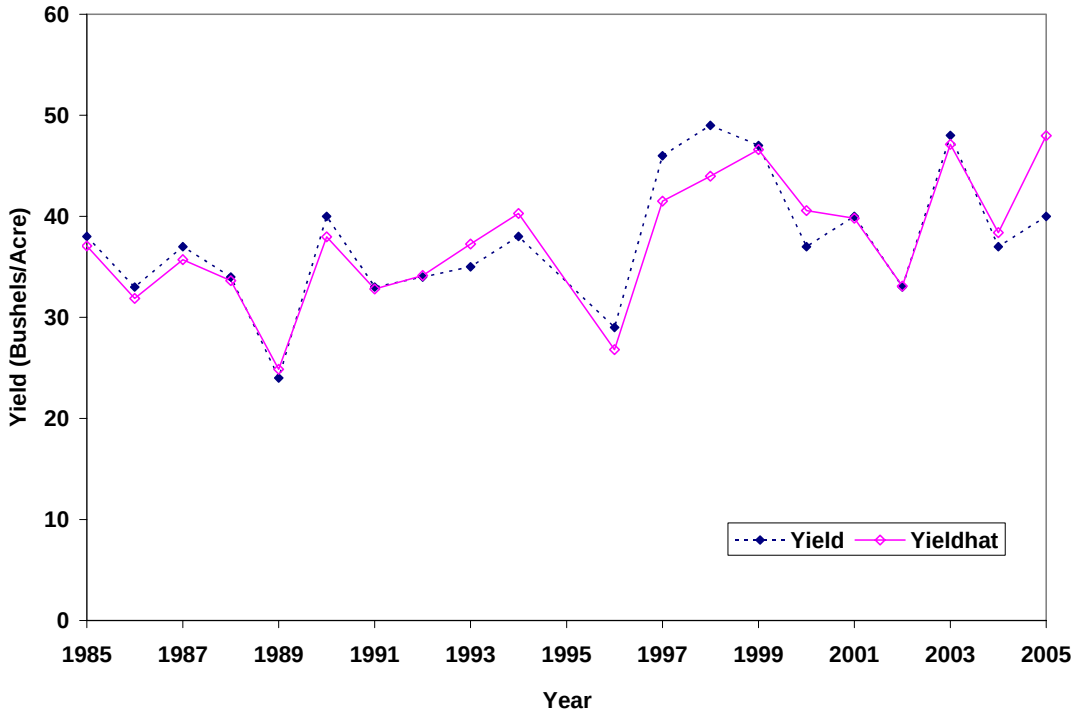

Figure 5.8: Observed yield versus independently simulated yield (yieldhat), Kansas

5.3.3 Partial Least Squares

As was mentioned in section 1.3.4, one approach to deal with collinearity is principal component regression (PCR) (Draper and Smith 1981). Even though the collinearity problem is efficiently avoided using PCR, it is not entirely clear how to select the principal components to use in the linear regression model. It has been found that the natural ordering in which the principal components are included in the

analysis may not be the best way to select the predictors from the list of *p* principal components (Hadi and Ling 1998, Jolliffe and ebrary Inc. 2002, Jackson 2003).

PCR method uses the linear combinations or PCs of **X** (predictors), derived using principal component analysis (PCA) to model the relationship between **X** and **y** (response). The derivation of the PCs is independent of **y** (the dependent variable) and therefore, this procedure neglects low variance components that may have predictive value. In an improved approach, only those PCs are used that show a good correlation with the **y-**variable of interest but often this is not sufficient. In contrast to PCR, the subspace retained in partial least squares (PLS) is constructed with reference to the vector of observations, **y**. Therefore, this alternative estimation procedure to cope with the collinearity problem is more efficient than PCR. In PLS, the data is projected so that correlation between input and output variables is maximized. PLS methodology works by extracting successive linear combinations of the predictors called factors, which explain both variation of the dependent and independent variables (Martens 1985, Garthwaite 1994, Goutis 1996, Butler and Denham 2000, Westad and Martens 2000).

5.3.3.1 Residual Validation Variance

When building the PLS model we investigated dY versus VCI (weeks 12 to 21) and TCI (weeks 16 to 18) for WW, which are weeks with correlation coefficients significantly different from zero (p-value<0.001) (Table 5.2). Full cross-validation

was used to determine the appropriate number of factors in the model. We set aside the first observation from the sample, and we used the remaining (n – 1) observations to estimate the coefficients for a particular candidate model. Then, the first observation was replaced and the second observation withheld with coefficients estimated again.

We removed each observation one at a time, and thus the candidate model was fit n = 20 times. The deleted response was estimated each time, resulting in n independent prediction residuals $y_i - \hat{y}_{i,-i} = e_{i,-i}$ (i = 1, 2, 3, 20). These residuals are true prediction errors with $\hat{y}_{i,-i}$ being independent of y_i. Thus, in this way, the observation y_i was not simultaneously used for fit and model assessment, this being the true test of validation. For choice of the best model, the RMSEP was calculated for each candidate model. It is a measurement of the average difference between predicted and measured response values and is interpreted as the average prediction error. Figure 5.9 shows that the model with one factor (PC_01) has the smallest RMSEP. Thus, for total Kansas one factor is the optimum number for modeling purposes (Westad and Martens 2000).

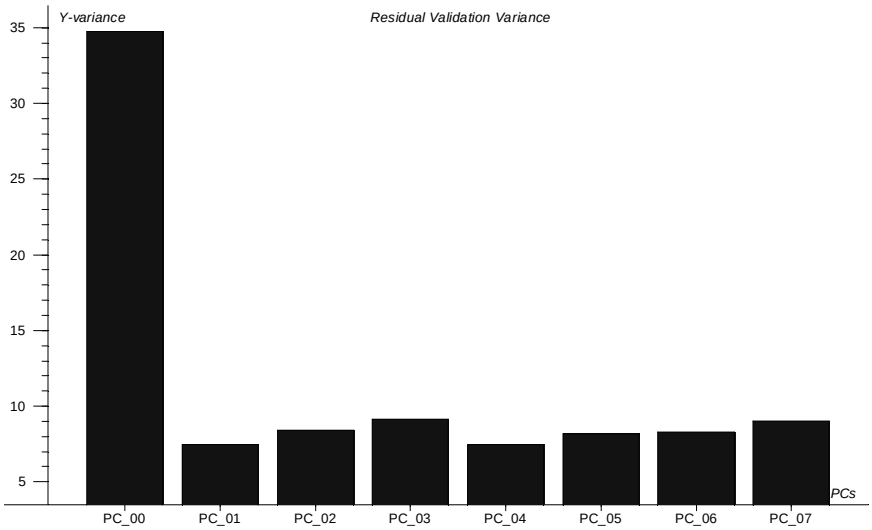

Figure 5.9: Residual validation variance (PLS), WW, Kansas

From Figure 5.9 we see that the residual variance decreases until PC 01 is reached (7.496). Then, the residual variance increases again due to over-fitting. Therefore, we will use a model with one PC.

5.3.3.2 Interpretation of the Regression Coefficients

One problem with PLS is that there are no simple expressions for the variances of the regression coefficients. This is because the use of the **y**-values in constructing the factors makes the formulae for these coefficients non-linear in **y**. Many methods of obtaining either analytical or numerical approximations to those variances have been suggested over the years. A technique known as the Jackknife can be used for variance estimation. The basic idea is to use the variation in a calculated statistic when observations are left out one at a time to get some handle on the uncertainty of

the statistic. Suppose we want to estimate the variance of a PLS regression coefficient *b* calculated from a set of *n* observations. Let b_i be the value of *b* obtained when observation *i* is omitted from the data set, and let $\bar{b}$ be the mean of the *n* values b_i. Then the jackknife estimate of the variance of *b* is:

$$v(B) = (\frac{n-1}{n})\sum_{i=1}^{n}(b_i - \bar{b})^2 \ . \qquad \textbf{(5.18)}$$

To see how it works it is easier to look at the estimation of the variance of a mean. Suppose the raw data are $x_1, x_2, \ldots, x_n$ and the statistic we want to calculate is the sample mean $\bar{x} = (1/n)\sum x_i$. Then, we do not need to use the jackknife, because there is a formula for the variance of the mean. This is s^2/n, where $s^2 = (1/n-1)\sum_{i=1}^{n}(x_i - \bar{x})^2$ is the sample variance of the x_i.

Suppose however, we did compute the jackknife estimate, if $\bar{x}_i$ is the mean of the $n - 1$ observations that remain when x_i is omitted, algebra produces the relationship:

$$(n-1)\,\bar{x}_i = n\,\bar{x} - x_i, \qquad \textbf{(5.19)}$$

both sides being the sum of all the observations except x_i. Summing both sides of this from i = 1 to *n* results in:

$$(n-1)n\,\bar{\bar{x}} = n^2\,\bar{x} - n\,\bar{x}\ , \qquad \textbf{(5.20)}$$

which shows that $\bar{\bar{x}} = \bar{x}$, i.e. that the average of all the leave-one-out means is the overall mean. Now subtract $(n-1)\ \bar{\bar{x}}$ from the left hand side of Equation (5.20) and $(n-1)\,\bar{x}$, which we have just shown is the same, from the right hand side to give:

$$(n-1)(\ \bar{x}_i - \bar{\bar{x}}\ ) = \bar{x} - x_i. \qquad \textbf{(5.21)}$$

Finally, summing the squares of both sides gives:

$$(n-1)^2 \sum_{i=1}^{n} (\bar{x}_i - \bar{\bar{x}})^2 = \sum_{i=1}^{n} (\bar{x} - x_i)^2 , \qquad \textbf{(5.22)}$$

which, if we divide both sides by $n(n-1)$ reduces to:

$$\frac{(n-1)}{n} \sum_{i=1}^{n} (\bar{x}_i - \bar{\bar{x}})^2 = s^2 / n , \qquad \textbf{(5.23)}$$

showing that in this case the jackknife estimate of the variance of $\bar{x}$ is exactly the same as that given by the standard formula and explaining where the fabricate factor of $(n - 1)/n$ comes from.

We are not interested in using the jackknife for estimating the mean variance because we have a formula. The idea is to use it when we do not have a formula, for example in the case of PLS regression coefficients. We saved the intermediate results in the cross-validation calculations to get the jackknife variances. Then a plot of *b*-coefficient against the independent variables with error bars of $\pm 2\sqrt{v(b)}$ will give a good indication where the coefficients are stable and significantly different from zero. These procedure can be used to select variables for inclusion in the PLS regression model (Westad and Martens 2000).

In order to simplify the final model and to make it more reliable and stable, non-significant predictors were eliminated following Westad and Martens (2000). Figure 5.10 shows the resulting significant predictors with uncertainty limits that correspond to two standard deviations ($\pm 2\sqrt{v(B)}$). Figure 5.10 shows that uncertainty limits for significant predictors do not cross the zero line.

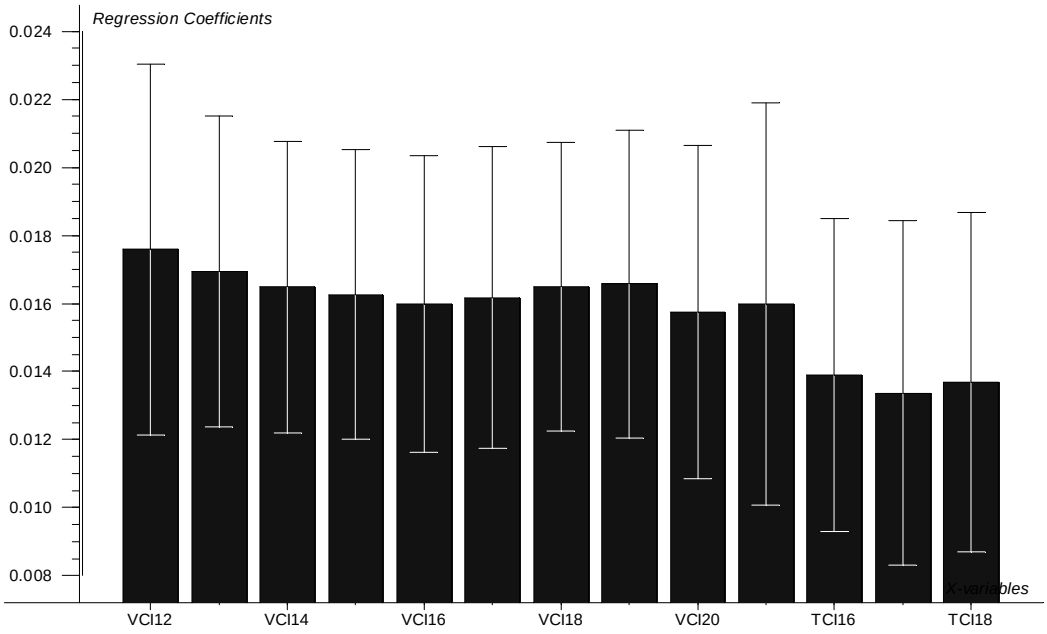

Figure 5.10: Regression coefficients for model Equation (5.9) calculated using PLS methodology, Kansas, U.S.

Figure 5.10 shows the regression coefficients for model Equation (5.9) calculated using PLS methodology. Each predictor variable defines one bar of the plot. All predictors VCI (weeks 12 to 21) and TCI (weeks 16 to 18) have large regression coefficients; thus, they play an important role in the regression model. All are positive showing a positive link with the response dY.

5.3.3.3 Predicted vs. Observed

For model validation we regressed observed yield versus independently simulated yield and the corresponding statistics were generated (Table 5.11). The overall correlation coefficient 0.9013 is good. An R-Square value of 0.8123 shows that in most years, WW yield in Kansas can be modeled by variables considered in model Equation (5.9) with regression coefficients estimated using PLS methodology. For WW the model forecast captured 81% of the variability in yield anomalies.

Table 5.11: Statistics of an independent testing for model Equation (5.9) with coefficients estimated following PLS methodology, WW, Kansas, U.S.

Crop	***Correlation***	***R-Square***	***Bias***	***Variance***	***SEP***	***Std Error***	***RMSEP***
WW	0.9013	0.8123	0.0247	7.8899	2.8089	0.6281	2.7379

The average prediction bias is 0.0247 (systematic error). The variance of the prediction error is 7.8899 or the standard error of prediction (SEP) is 2.8089 (non-systematic error). The standard error of the estimated mean bias is 0.6281 (Table 5.11). T-tests of the hypothesis that the bias is zero gives t = 0.0394 which, with 19 degrees of freedom and $\alpha = 0.05$, is not significant.

The mean square error of prediction (MSEP) is $= \frac{19(7.8899)}{20} + (0.0247)^2 = 7.4960$. The RMSEP is 2.7379 an approximate 7% error in prediction for total Kansas. This is an acceptable error in this kind of application, which has uncertainty in measurements. Therefore, we can conclude that

models based on Equation (5.9) with the coefficients estimated using PLS methodology perform well.

6. Results and Discussion

As mentioned in Chapter 5, the strategy of this research was to: (1) extract the weather component from ground data (crop yield), and satellite data (NDVI and BT) values and (2) correlate the weather related component of ground data with the corresponding weather related component of satellite data. The goal was to investigate the strength of the relationship and to determine if the strongest correlation coincides with crop's critical period, which is the period when crop production is highly sensitive to weather conditions.

Since ground data and satellite data were similarly expressed as a deviation from climatology, further examination included correlation and regression analysis of these deviations to investigate the association among them for each CRD and for total Kansas for the area of crop growth.

6.1 Winter Wheat

6.1.1 Crop Reporting District 10

6.1.1.1 Crop Yield Time Series

Figure 6.1 shows slight decrease in the long-term yield trend for CRD 10. Although agriculture technology is improving here as well, analysis of the literature indicates that this reduction is related to low precipitation rates in western Kansas (as shown in Figure 3.5) and intensive irrigation practices. Irrigation has stimulated an

increase in soil salinity, which has become a severe environmental hazard in this region. Farmers are facing decreasing crop yields due, in part, to high levels of salinity. In some areas in western Kansas, land is being taken out of production due to unsustainable crop yields (Miles et al. 1977, Hillel 2000, Eldeiry and Garcia 2004).

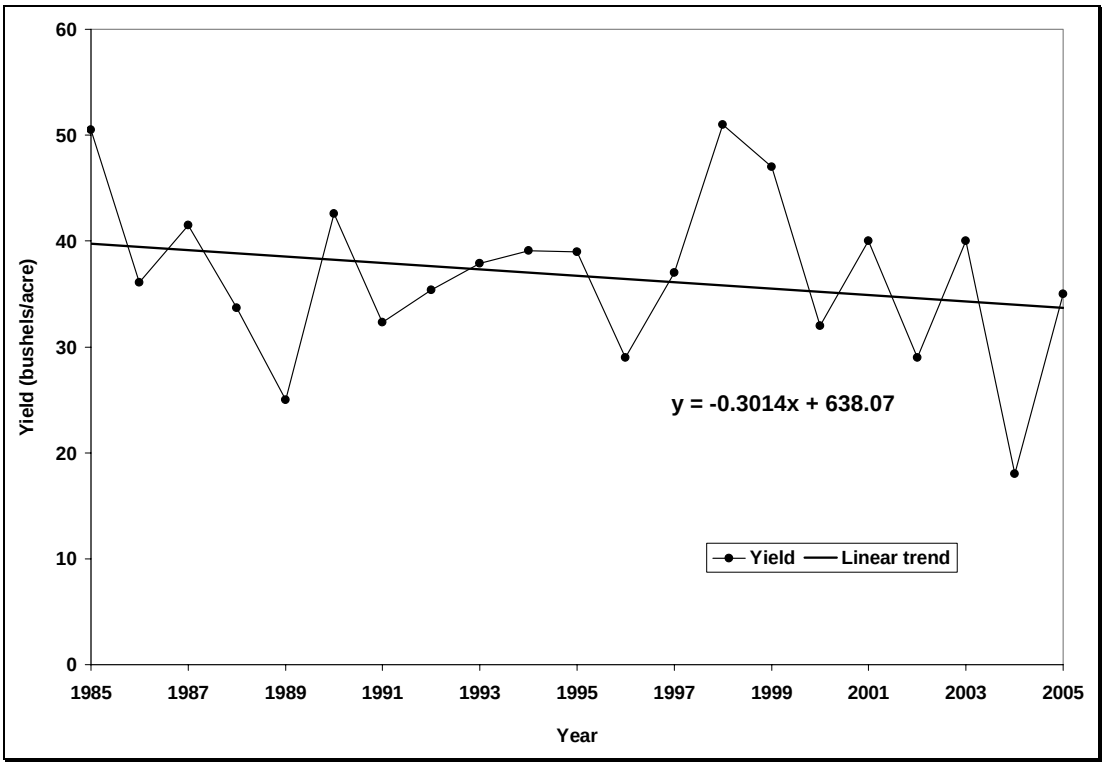

Figure 6.1: Winter wheat yield time series CRD 10, Kansas

6.1.1.2 Regression Analysis

A useful starting point in any multiple regression analysis is to compute correlations among all variables. This provides a first look at the simple linear relationships among them.

6.1.1.2.1 The Correlation Matrix

For CRD 10, VCI (weeks 15 to 23) variables have high correlations with dY (0.81-0.86). VCI19 has the highest correlation with dY. It would account for 74 % (0 $.86^2$) of the variation in dY if used separately as the only independent variable in the regression model.

Table 6.1: Correlation Matrix of dY with VCI (weeks 15 to 23) and TCI (weeks 22 to 23) CRD 10, Kansas

Pearson Correlation Coefficients	
	dY
VCI15	0.8308
VCI16	0.8452
VCI17	0.8520
VCI18	0.8511
VCI19	0.8635
VCI20	0.8594
VCI21	0.8531
VCI22	0.8374
VCI23	0.8127
TCI22	0.6935
TCI23	0.7130

TCI (weeks 22 to 23) variables also have significant correlation (p-value <0.001) with dY (0.69-0.71) (Table 6.1). Average VCI and TCI ($\overline{vci}, \overline{tci}$) for weeks with significant Pearson's correlation coefficients at p-level (p<0.001) among dY and VH indices were used as a predictors of dY. Figure 6.2 shows dynamics of dY (WW),

MEAN VCI (weeks 15 to 23) and MEAN TCI (weeks 22 to 23) for CRD 10. This figure shows how closely VH indices dynamics follow dY dynamics. This is a strong indication that VH indices can be used as a predictor in a forecast model for WW yield in CRD 10, Kansas.

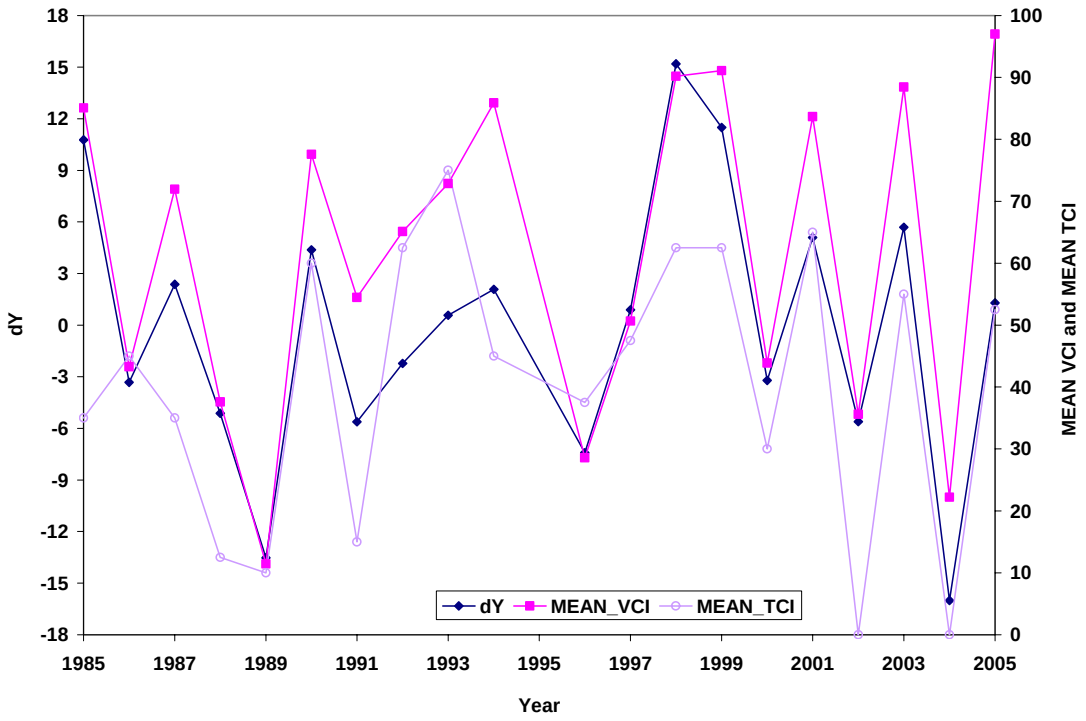

Figure 6.2: Dynamics of dY (WW), MEAN VCI (weeks 15 to 23) and MEAN TCI (weeks 22 to 23) CRD 10, Kansas

Table 6.2 shows correlation coefficients between dY (WW) with $\overline{vci}$ (weeks 15 to 23) and $\overline{tci}$ (weeks 22 to 23). These correlation coefficients are significantly different from 0 with p-value ($p<0.001$). For WW, $\overline{vci}$ would account for 77% (0.88^2) of the variation in dY if used separately as the only independent variable in the

regression model. Besides, $\overline{tci}$ would account for 50% (0.71^2) of the variation in dY (WW) if used separately as the only independent variable in the regression model.

Table 6.2: Correlation matrix dY, MEAN VCI ($\overline{vci}$) and MEAN TCI ($\overline{tci}$) CRD 10, Kansas

Pearson Correlation Coefficients	
	dY
MEAN VCI	0.8813
MEAN TCI	0.7103

6.1.1.2.2 Multiple Regression Results

The results of fitting the ordinary least squares (OLS) regression model approximated by Equation (6.1) to CRD 10 are shown in Table 6.3.

$$dY = a_0 + b_1 \cdot \overline{vci} + b_2 \cdot \overline{tci} + \varepsilon \,. \qquad \textbf{(6.1)}$$

Table 6.3 shows that there is a strong relationship between dY and the independent variables. The coefficient of determination, R-Square, is 0.78 or 78% of the sum of squares in dY (WW) can be associated with the variation in these independent variables. The test of the composite hypothesis that all regression coefficients are 0 is highly significant with F-value of 30.60 compared to F (0.01, 2, 17) = 6.11. The residual standard error (Root MSE) is 3.8819 with 17 degrees of freedom is an unbiased estimate of σ.

Table 6.3: Results of the regression of dY on the two independent variables MEAN VCI and MEAN TCI CRD10, Kansas

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	2	922.1702	461.0851	30.5977	<.0001
Error	17	256.1778	15.0693	–	–
Corrected Total	19	1178.348	–	–	–

R-Square	*Root MSE*
0.7826	3.8819

Parameter	*Estimate*	*Standard Error*	*t Value*	*Pr > |t|*
Intercept	-16.5074	2.2753	-7.2550	0.0000
MEAN VCI	0.2391	0.0513	4.6630	0.0002
MEAN TCI	0.0399	0.0588	0.6784	0.5067

6.1.1.2.3 Model Validation

For model validation we regressed observed yield versus independently simulated yield and the corresponding statistics were generated (Table 6.4). The overall correlation coefficient 0.8427 is good. An R-Square value of 0.7101 shows that in most years, WW yield in CRD 10, Kansas can be modeled by variables considered in model Equation (6.1). For WW the model forecast captured 71% of the variability in yield anomalies.

Table 6.4: Statistics of an independent testing for model equation (6.1), WW, CRD 10, Kansas

Crop	*Correlation*	*R-Square*	*Bias*	*Variance*	*SEP*	*Std Error*	*RMSEP*
WW	0.8427	0.7101	–0.0585	19.0609	4.3659	0.9767	4.2557

The average prediction bias is -0.0585 (systematic error). The variance of the prediction error is 19.0609 or the standard error of prediction (SEP) is 4.3659 (non-systematic error). The standard error of the estimated mean bias is 0.9767 (Table 6.4). T-tests of the hypothesis that the bias is zero gives t = -0.0599 which, with 19 degrees of freedom and $\alpha = 0.05$, is not significant. The mean square error of prediction (MSEP) is $= \frac{19(19.0609)}{20} + (-0.0585)^2 = 18.1113$. The root mean square error of prediction (RMSEP) is 4.2557 an approximate 12% error in prediction.

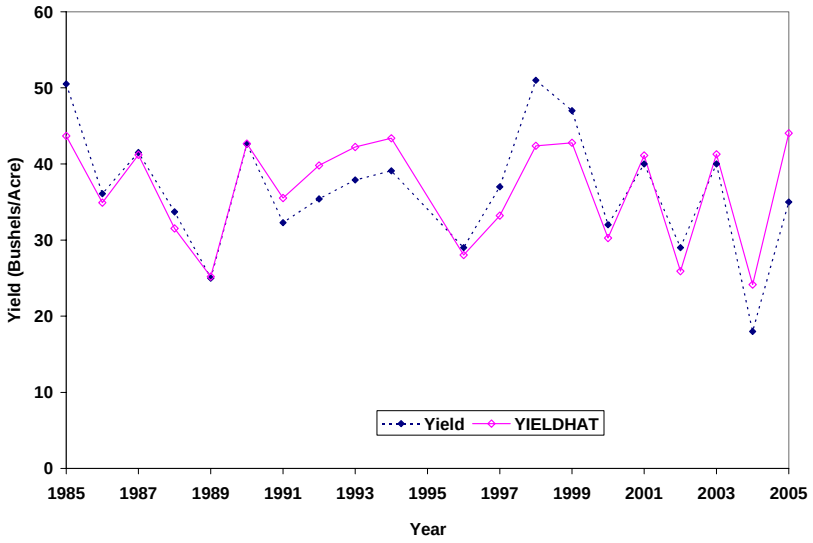

Figure 6.3: Observed yield (WW) versus independently simulated yield (yieldhat) CRD 10, Kansas

Figure 6.3 displays observed versus independently simulated crop yield time series of WW 1985 through to 2005 for CRD 10, Kansas. For WW CRD 10 the model forecasts captured 71% of the variability in yield anomalies.

6.1.2 Crop Reporting District 20

6.1.2.1 Crop Yield Time Series

Figure 6.4 shows slight decrease in the long-term yield trend for CRD 20, Kansas. Although agriculture technology is improving here as well, reduction is related to low precipitation rates in western Kansas (Figure 3.5) and intensive irrigation practices.

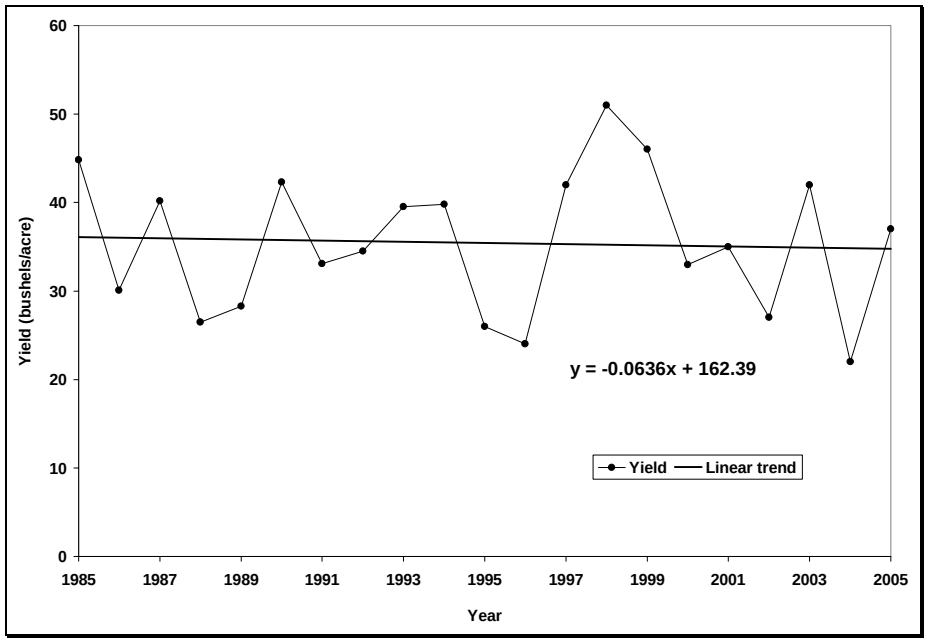

Figure 6.4: Winter wheat yield time series CRD 20, Kansas

6.1.2.2 Regression Analysis

A useful starting point in any multiple regression analysis is to compute correlations among all variables. This provides a first look at the simple linear relationships among them.

6.1.2.2.1 The Correlation Matrix

For CRD 20, VCI (weeks 15 to 24) variables have high correlations with dY (0.80-0.90). VCI20 has the highest correlation with dY. It would account for 81% (0 $.90^2$) of the variation in dY if used separately as the only independent variable in the regression model. TCI (weeks 22 to 23) variables also have significant correlation (p-value <0.001) with dY (0.69-0.72) (Table 6.5).

Average VCI and TCI ($\overline{vci}, \overline{tci}$) for weeks with significant Pearson's correlation coefficients at p-level (p<0.001) among dY and VH indices were used as a predictors of dY.

Figure 6.5 shows dynamics of dY (WW), MEAN VCI (weeks 15 to 24) and MEAN TCI (weeks 22 to 23) for CRD 20, Kansas. This figure shows how closely VH indices dynamics follows dY dynamics. This is a strong indication that VH indices can be used as a predictor in a forecast model for WW yield in CRD 20, Kansas.

Table 6.5: Correlation Matrix dY with VCI (weeks 15 to 24) and TCI (weeks 22 to 23) CRD 20, Kansas

Pearson Correlation Coefficients	
	dY
VCI15	0.8255
VCI16	0.8569
VCI17	0.8623
VCI18	0.8791
VCI19	0.8935
VCI20	0.9026
VCI21	0.9022
VCI22	0.8921
VCI23	0.8638
VCI24	0.7993
TCI22	0.6923
TCI23	0.7236

Figure 6.5: Dynamics of dY (WW), MEAN VCI (weeks 15 to 24) and MEAN TCI (weeks 22 to 23) CRD 20, Kansas

Table 6.6 shows correlation coefficients between dY (WW) with $\overline{vci}$ (weeks 15 to 24) and $\overline{tci}$ (weeks 22 to 23). These correlation coefficients are significantly different from 0 with p-value ($p<0.001$). For WW, $\overline{vci}$ would account for 85% (0.92^2) of the variation in dY if used separately as the only independent variable in the regression model. Besides, $\overline{tci}$ would account for 52% (0.72^2) of the variation in dY (WW) if used separately as the only independent variable in the regression model.

Table 6.6: Correlation matrix dY, MEAN VCI ($\overline{vci}$) and MEAN TCI ($\overline{tci}$) CRD 20, Kansas

Pearson Correlation Coefficients	
	dY
MEAN VCI	0.9249
MEAN TCI	0.7175

6.1.2.2.2 Multiple Regression Results

The results of fitting the ordinary least squares (OLS) regression model approximated by Equation (6.1) to CRD 20 are shown in Table 6.7.

Table 6.7 shows that there is a strong relationship between dY and the independent variables. The coefficient of determination, R-Square, is 0.88 or 88% of the sum of squares in dY (WW) can be associated with the variation in these independent variables. The test of the composite hypothesis that all regression coefficients are 0 is highly significant with F-value of 64.06 compared to F (0.01, 2,

17) = 6.11. The residual standard error (Root MSE) is 2.8541 with 17 degrees of freedom is an unbiased estimate of σ.

Table 6.7: Results of the regression of dY on the two independent variables MEAN VCI and MEAN TCI CRD 20, Kansas

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	2	1043.691	521.8457	64.0625	<.0001
Error	17	138.4800	8.1459	–	–
Corrected Total	19	1182.171	–	–	–

R-Square	*Root MSE*
0.8829	2.8541

Parameter	*Estimate*	*Standard Error*	*t Value*	*Pr > |t|*
Intercept	-17.1776	1.7035	-10.0839	0.0000
MEAN VCI	0.2441	0.0334	7.3089	0.0000
MEAN TCI	0.0736	0.0368	1.9974	0.0620

6.1.2.2.3 Model Validation

For model validation we regressed observed yield versus independently simulated yield and the corresponding statistics were generated (Table 6.8). The overall correlation coefficient 0.9102 is good. An R-Square value of 0.8285 shows that in most years, WW yield in CRD 20 can be modeled by variables considered in

model Equation (6.1). For WW the model forecast captured 83% of the variability in yield anomalies.

The average prediction bias is -0.0398 (systematic error). The variance of the prediction error is 10.7813 or the standard error of prediction (SEP) is 3.2835 (non-systematic error). The standard error of the estimated mean bias is 0.7342 (Table 6.8). T-tests of the hypothesis that the bias is zero gives t = -0.0542 which, with 19 degrees of freedom and $\alpha = 0.05$, is not significant.

The mean square error of prediction (MSEP) is $= \frac{19(10.7813)}{20} + (-0.0398)^2 = 10.2438$. The RMSEP is 3.2006 an approximate 9% error in prediction.

Table 6.8: Statistics of an independent testing for model equation (6.1), WW, CRD 20, Kansas

Crop	***Correlation***	***R-Square***	***Bias***	***Variance***	***SEP***	***Std Error***	***RMSEP***
WW	0.9102	0.8285	-0.0398	10.7813	3.2835	0.7342	3.2006

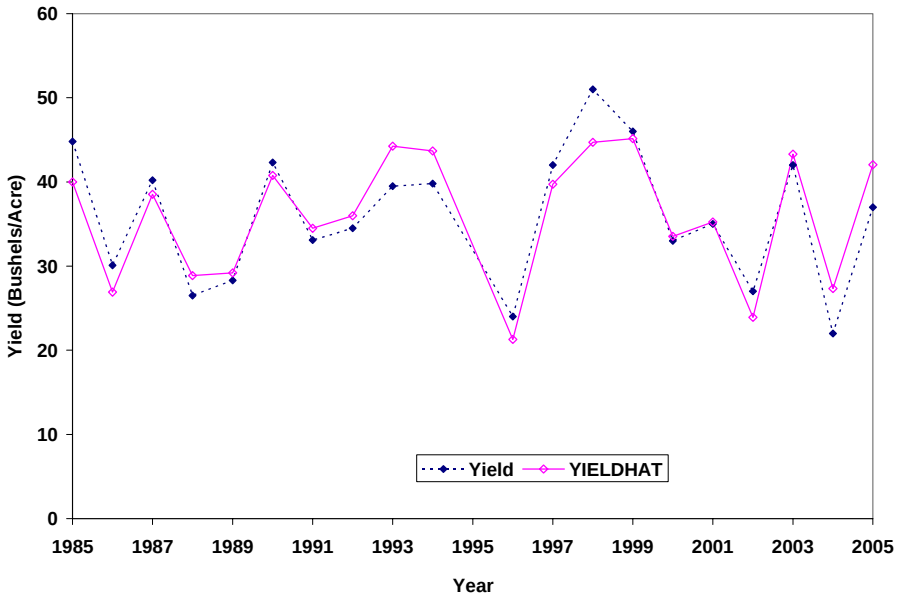

Figure 6.6: Observed yield (WW) versus independently simulated yield (yieldhat) CRD 20, Kansas

Figure 6.6 displays observed versus independently simulated crop yield time series of WW 1985 through to 2005 for CRD 20, Kansas. For WW CRD 20 the model forecasts captured 83% of the variability in yield anomalies.

6.1.3 Crop Reporting District 30

6.1.3.1 Crop Yield Time Series

Figure 6.7 shows slight decrease in the long-term yield trend for CRD 30. Although agriculture technology is improving here as well, reduction is related to low precipitation rates in western Kansas (Figure 3.5) and intensive irrigation practices.

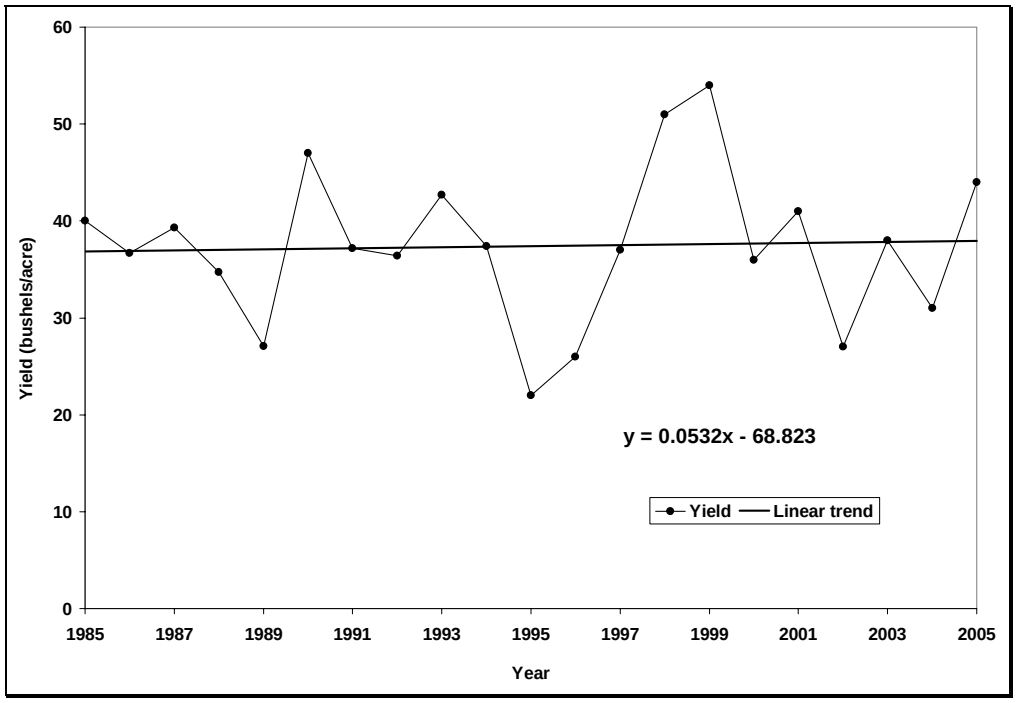

Figure 6.7: Winter wheat yield time series CRD 30, Kansas

6.1.3.2 Regression Analysis

A useful starting point in any multiple regression analysis is to compute correlations among all variables. This provides a first look at the simple linear relationships among them.

For CRD 30, VCI (weeks 14 to 22) variables have high correlations with dY (0.82-0.89). VCI19 has the highest correlation with dY. It would account for 79% (0 $.89^2$) of the variation in dY if used separately as the only independent variable in the

regression model. TCI (weeks 15 to 19) variables also have significant correlation (p-value <0.001) with dY (0.72-0.77) (Table 6.9).

Table 6.9: Correlation Matrix dY with VCI (weeks 14 to 22) and TCI (weeks 15 to 19) CRD 30, Kansas

Pearson Correlation Coefficients	
	dY
VCI14	0.8179
VCI15	0.8575
VCI16	0.8668
VCI17	0.8771
VCI18	0.8905
VCI19	0.8916
VCI20	0.8836
VCI21	0.8740
VCI22	0.8535
TCI15	0.7577
TCI16	0.7549
TCI17	0.7698
TCI18	0.7693
TCI19	0.7227

6.1.3.2.1 The Correlation Matrix

Average VCI and TCI ($\overline{vci}, \overline{tci}$) for weeks with significant Pearson's correlation coefficients at p-level (p<0.001) among dY and VH indices were used as a predictors of dY.

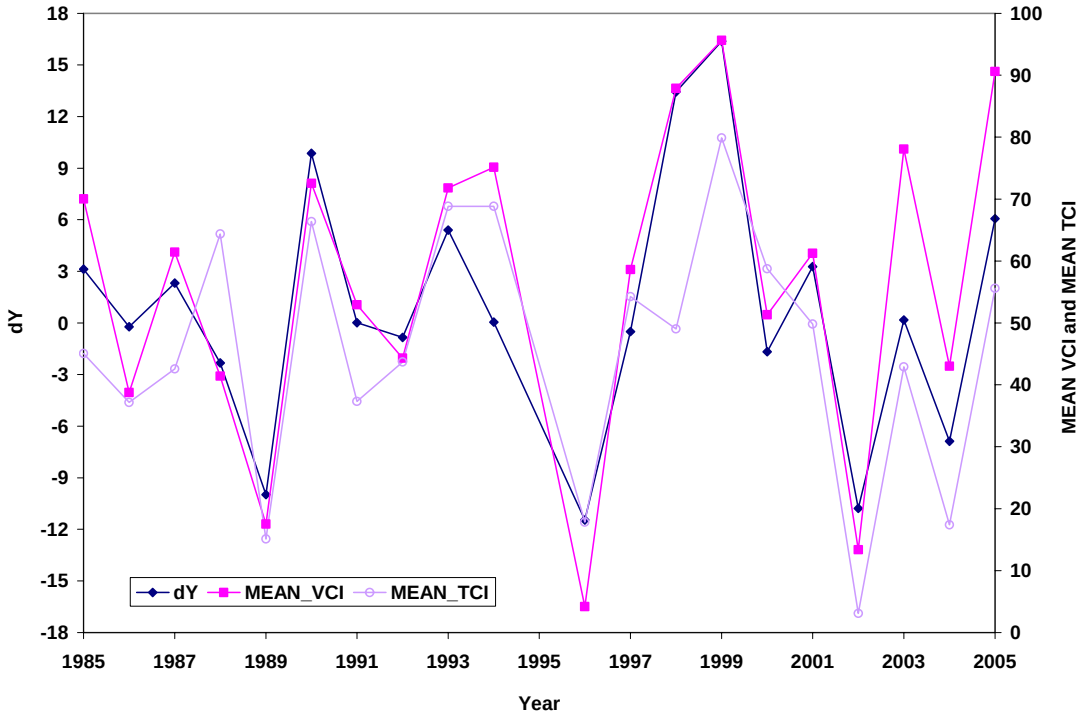

Figure 6.8: Dynamics of dY (WW), MEAN VCI (weeks 14 to 22) and MEAN TCI (weeks 15 to 19) CRD 30, Kansas

Figure 6.8 shows dynamics of dY (WW), MEAN VCI (weeks 14 to 22) and MEAN TCI (weeks 15 to 19) for CRD 30. This figure shows how closely VH indices dynamics follows dY dynamics. This is a strong indication that VH indices can be used as a predictor in a forecast model for WW yield in CRD 30, Kansas.

Table 6.10 shows correlation coefficients between dY (WW) with $\overline{vci}$ (weeks 14 to 22) and $\overline{tci}$ (weeks 15 to 19). These correlation coefficients are significantly different from 0 with p-value ($p<0.001$). For WW, $\overline{vci}$ would account for 81%

(0.90^2) of the variation in dY if used separately as the only independent variable in the regression model. Besides, $\overline{tci}$ would account for 61% (0.78^2) of the variation in dY (WW) if used separately as the only independent variable in the regression model.

Table 6.10: Correlation matrix dY, MEAN VCI ($\overline{vci}$) and MEAN TCI ($\overline{tci}$) CRD 30, Kansas

Pearson Correlation Coefficients	
	dY
MEAN VCI	0.8976
MEAN TCI	0.7754

6.1.3.2.2 Multiple Regression Results

The results of fitting the ordinary least squares (OLS) regression model approximated by Equation (6.1) to CRD 30 are shown in Table 6.11.

Table 6.11 shows that there is a strong relationship between dY and the independent variables. The coefficient of determination, R-Square, is 0.82 or 82% of the sum of squares in dY (WW) can be associated with the variation in these independent variables. The test of the composite hypothesis that all regression coefficients are zero is highly significant with F-value of 39.43 compared to F (0.01, 2, 17) = 6.11. The residual standard error (Root MSE) is 3.2778 with 17 degrees of freedom is an unbiased estimate of σ.

Table 6.11: Results of the regression of dY on the two independent variables MEAN VCI and MEAN TCI, CRD 30, Kansas

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	2	847.2883	423.6441	39.4312	<.0001
Error	17	182.6461	10.7439	–	–
Corrected Total	19	1029.934	–	–	–

R-Square	*Root MSE*
0.8227	3.2778

Parameter	*Estimate*	*Standard Error*	*t Value*	*Pr > |t|*
Intercept	-14.7425	1.9339	-7.6233	0.0000
MEAN VCI	0.2145	0.0465	4.6079	0.0003
MEAN TCI	0.0740	0.0581	1.2749	0.2195

6.1.3.2.3 Model Validation

For model validation we regressed observed yield versus independently simulated yield and the corresponding statistics were generated (Table 6.12). The overall correlation coefficient 0.8665 is good. An R-Square value of 0.7508 shows that in most years, WW yield in CRD 30 can be modeled by variables considered in model Equation (6.1). For WW, CRD 30, the model forecast captured 75% of the variability in yield anomalies.

The average prediction bias is 0.0090 (systematic error). The variance of the prediction error is 13.6551 or the standard error of prediction (SEP) is 3.6953 (non-

systematic error). The standard error of the estimated mean bias is 0.8263 (Table 6.12). T-tests of the hypothesis that the bias is zero gives t = 0.0109 which, with 19 degrees of freedom and $\alpha = 0.05$, is not significant.

Table 6.12: Statistics of an independent testing for model equation (6.1), WW, CRD 30, Kansas

Crop	*Correlation*	*R-Square*	*Bias*	*Variance*	*SEP*	*Std Error*	*RMSEP*
WW	0.8665	0.7508	0.0090	13.6551	3.6953	0.8263	3.6017

The mean square error of prediction (MSEP) is $= \frac{19(13.6551)}{20} + (0.0090)^2 = 12.9724$. The RMSEP is 3.6017 an approximate 9% error in prediction.

Figure 6.9 displays observed versus independently simulated crop yield time series of WW 1985 through to 2005 for CRD 30, Kansas. For WW CRD 30 the model forecasts captured 75% of the variability in yield anomalies.

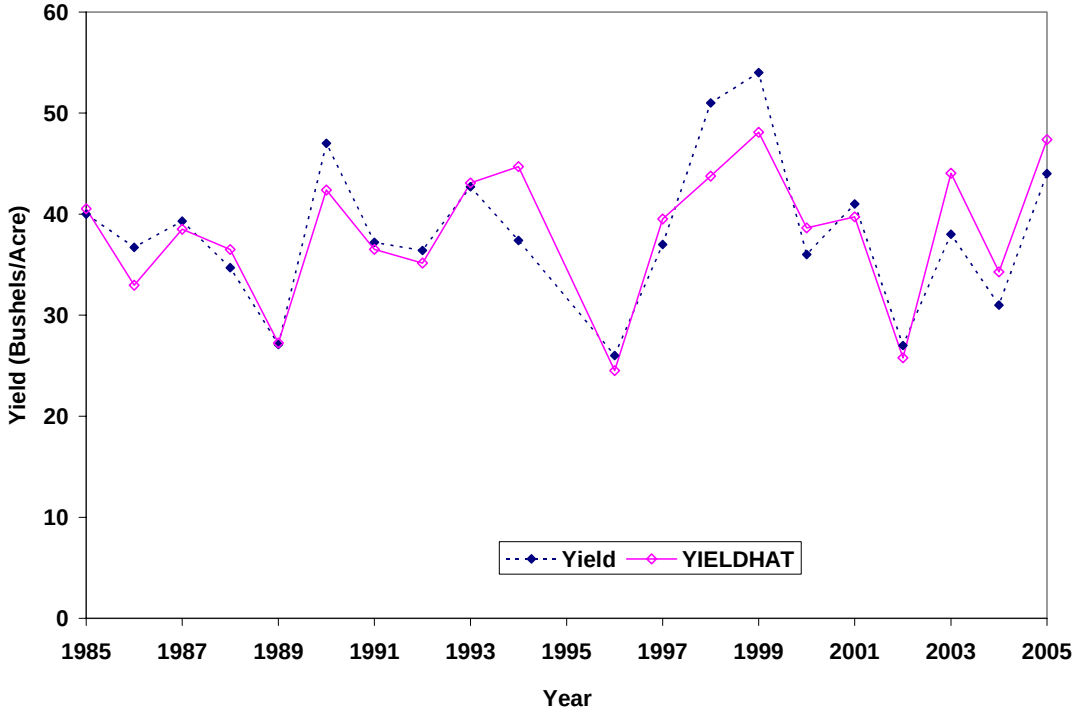

Figure 6.9: Observed yield (WW) versus independently simulated yield (yieldhat) CRD 30, Kansas

6.1.4 Crop Reporting District 40

6.1.4.1 Crop yield time series

Figure 6.10 shows that in CRD 40 WW yield increases due to technology improvement.

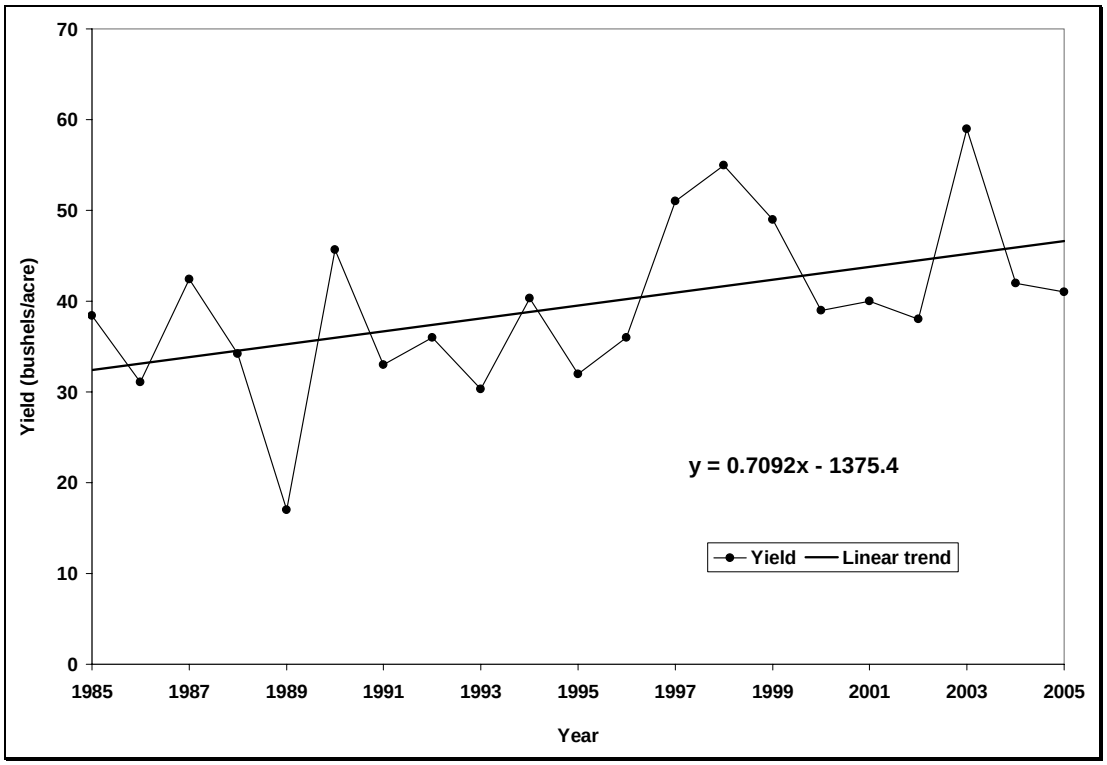

Figure 6.10: Winter wheat yield time series CRD 40, Kansas

6.1.4.2 Regression Analysis

A useful starting point in any multiple regression analysis is to compute correlations among all variables. This provides a first look at the simple linear relationships among them.

6.1.4.2.1 The Correlation Matrix

For CRD 40, VCI (weeks 14 to 22) variables have high correlations with dY (0.69-0.76). VCI20 has the highest correlation with dY. It would account for 58% (0 $.76^2$) of the variation in dY if used separately as the only independent variable in the regression model. TCI (weeks 19 to 20) variables also have significant correlation (p-value <0.001) with dY (0.55-0.56) (Table 6.13).

Table 6.13: Correlation matrix dY with VCI (weeks 14 to 22) and TCI (weeks 19 to 20) CRD 40, Kansas

Pearson Correlation Coefficients	
	dY
VCI14	0.7108
VCI15	0.7236
VCI16	0.7265
VCI17	0.7110
VCI18	0.7366
VCI19	0.7387
VCI20	0.7633
VCI21	0.7610
VCI22	0.6907
TCI19	0.5485
TCI20	0.5595

Average VCI and TCI ($\overline{vci}, \overline{tci}$) for weeks with significant Pearson's correlation coefficients at p-level (p<0.001) among dY and VH indices were used as a predictors of dY. Figure 6.11 shows dynamics of dY (WW), MEAN VCI (weeks 14

to 22) and MEAN TCI (weeks 19 to 20) for CRD 40. This figure shows how closely VH indices dynamics follows dY dynamics. This is a strong indication that VH indices can be used as predictors in a forecast model for WW yield in CRD 40, Kansas.

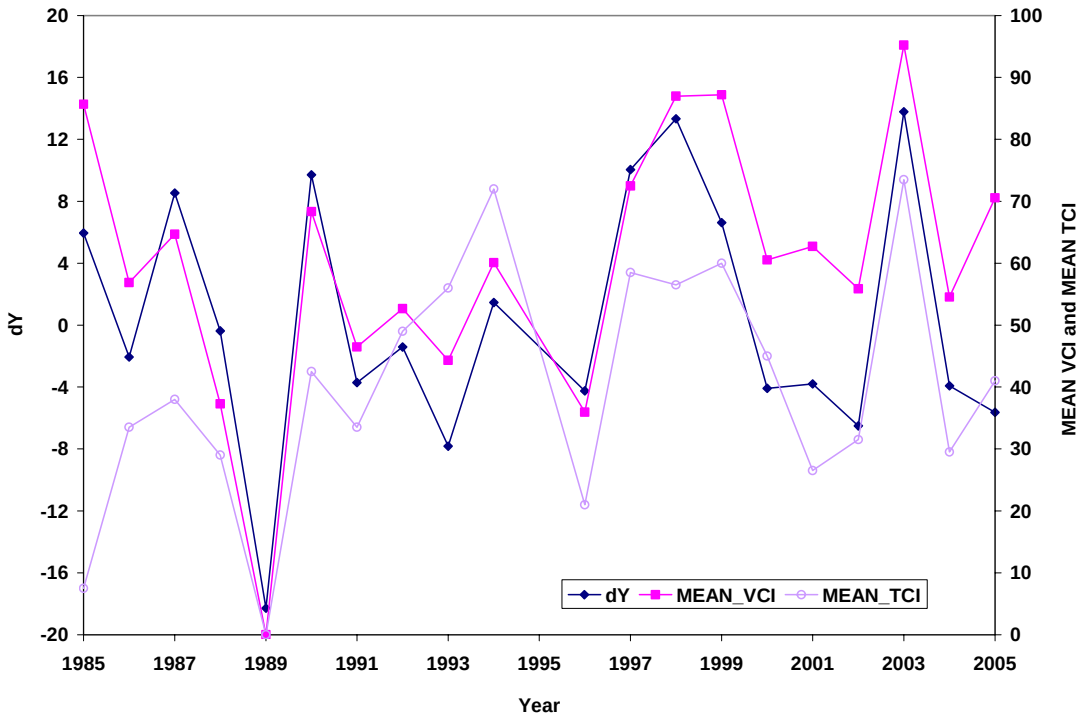

Figure 6.11: Dynamics of dY (WW), MEAN VCI (weeks 14 to 22) and MEAN TCI (weeks 19 to 20) CRD 40, Kansas

Table 6.14 shows correlation coefficients between dY (WW) with $\overline{vci}$ (weeks 14 to 22) and $\overline{tci}$ (weeks 19 to 20) CRD 40. These correlation coefficients are significantly different from zero with p-value ($p<0.001$). For WW, $\overline{vci}$ would account for 69% (0.83^2) of the variation in dY if used separately as the only

independent variable in the regression model. Besides, $\overline{tci}$ would account for 31% (0.56^2) of the variation in dY (WW) if used separately as the only independent variable in the regression model.

Table 6.14: Correlation matrix dY, MEAN VCI ($\overline{vci}$) and MEAN TCI ($\overline{tci}$) CRD 40, Kansas

Pearson Correlation Coefficients	
	dY
MEAN VCI	0.8300
MEAN TCI	0.5646

6.1.4.2.2 Multiple Regression Results

The results of fitting the ordinary least squares (OLS) regression model approximated by Equation (6.1) to CRD 40 are shown in Table 6.15.

Table 6.15 shows that there is a strong relationship between dY and the independent variables. The coefficient of determination, R-Square, is 0.70 or 70% of the sum of squares in dY (WW) can be associated with the variation in these independent variables. The test of the composite hypothesis that all regression coefficients are 0 is highly significant with F-value of 19.97 compared to F (0.01, 2, 17) = 6.11. The residual standard error (Root MSE) is 4.7195 with 17 degrees of freedom is an unbiased estimate of σ.

Table 6.15: Results of the regression of dY on the two independent variables MEAN VCI and MEAN TCI, CRD 40, Kansas

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	2	889.7589	444.8794	19.9736	<.0001
Error	17	378.6476	22.2734	–	–
Corrected Total	19	1268.407	–	–	–

R-Square	*Root MSE*
0.7015	4.7195

Parameter	*Estimate*	*Standard Error*	*t Value*	*Pr > |t|*
Intercept	-18.8915	3.2367	-5.8366	0.0000
MEAN VCI	0.2830	0.0606	4.6684	0.0002
MEAN TCI	0.0575	0.0680	0.8458	0.4094

6.1.4.2.3 Model Validation

For model validation we regressed observed yield versus independently simulated yield and the corresponding statistics were generated (Table 6.16). The overall correlation coefficient 0.8651 is good. An R-Square value of 0.7483 shows that in most years, WW yield in CRD 40, Kansas can be modeled by variables considered in model Equation (6.1). For WW CRD 40, the model forecast captured 75% of the variability in yield anomalies.

The average prediction bias is 0.0347 (systematic error). The variance of the prediction error is 24.0566 or the standard error of prediction (SEP) is 4.9048 (non-systematic error). The standard error of the estimated mean bias is 1.0967 (Table 6.16). T-tests of the hypothesis that the bias is 0 gives t = 0.0317 which, with 19 degrees of freedom and $\alpha = 0.05$, is not significant.

The MSEP is $= \frac{19(24.0566)}{20} + (0.0347)^2 = 22.8550$. The RMSEP is 4.7807 an approximate 12% error in prediction.

Table 6.16: Statistics of an independent testing for model Equation (6.1), WW, CRD 40, Kansas

Crop	*Correlation*	*R-Square*	*Bias*	*Variance*	*SEP*	*Std Error*	*RMSEP*
WW	0.8651	0.7483	0.0347	24.0566	4.9048	1.0967	4.7807

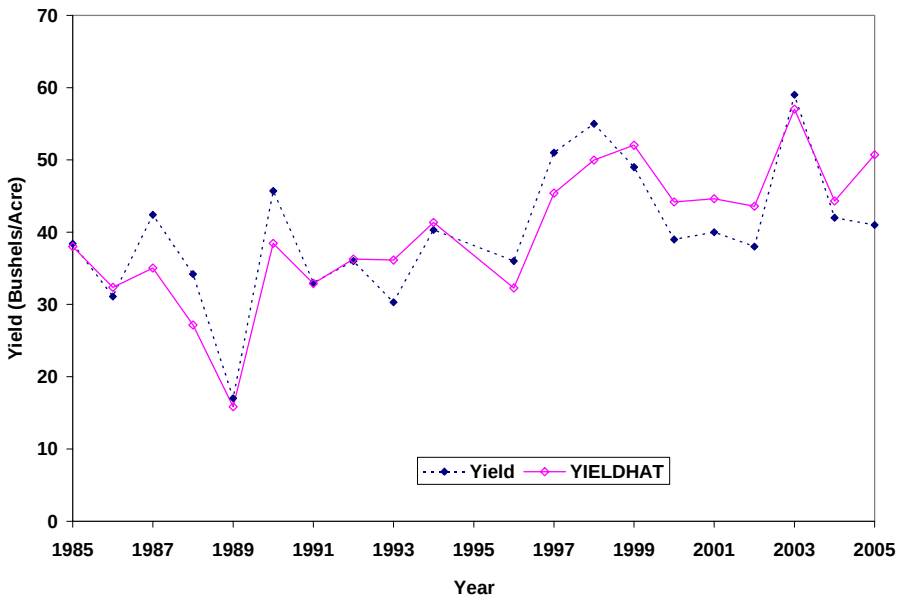

Figure 6.12: Observed yield (WW) versus independently simulated yield (yieldhat) CRD 40, Kansas

Figure 6.12 displays observed versus independently simulated crop yield time series of WW 1985 through to 2005 for CRD 40, Kansas. For WW CRD 40 the model forecasts captured 75% of the variability in yield anomalies.

6.1.5 Crop Reporting District 50

6.1.5.1 Crop Yield Time Series

Figure 6.13 shows that in CRD 50 WW yield increases due to technology improvement.

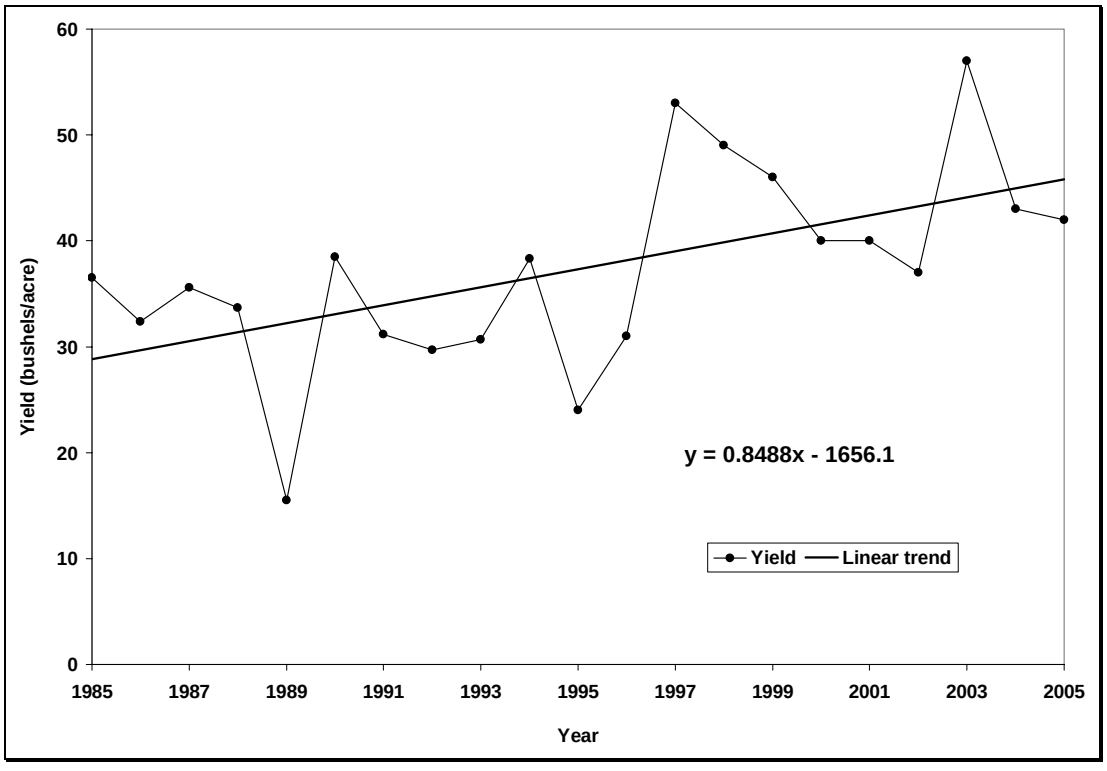

Figure 6.13: Winter wheat yield time series CRD 50, Kansas

6.1.5.2 Regression Analysis

A useful starting point in any multiple regression analysis is to compute correlations among all variables. This provides a first look at the simple linear relationships among them.

6.1.5.2.1 The Correlation Matrix

For CRD 50, VCI (weeks 12 to 21) variables have high correlations with dY (0.70-0.78). VCI15 has the highest correlation with dY. It would account for 61% (0 $.78^2$) of the variation in dY if used separately as the only independent variable in the regression model. TCI (week 16) variables also have significant correlation (p-value <0.001) with dY (0.59) (Table 6.17).

Average VCI and TCI ($\overline{vci}, \overline{tci}$) for weeks with significant Pearson's correlation coefficients at p-level (p<0.001) among dY and VH indices were used as a predictors of dY. Figure 6.14 shows dynamics of dY (WW), MEAN VCI (weeks 12 to 21) and MEAN TCI (weeks 16) for CRD 50. This figure shows how closely VH indices dynamics follows dY dynamics. This is a strong indication that VH indices can be used as a predictor in a forecast model for winter wheat yield in CRD 50, Kansas.

Table 6.17: Correlation matrix dY with VCI (weeks 12 to 21) and TCI (week 16) CRD 50, Kansas

Pearson Correlation Coefficients	
	dY
VCI12	0.7338
VCI13	0.7529
VCI14	0.7610
VCI15	0.7752
VCI16	0.7504
VCI17	0.7495
VCI18	0.7325
VCI19	0.7005
VCI20	0.7535
VCI21	0.7233
TCI16	0.5859

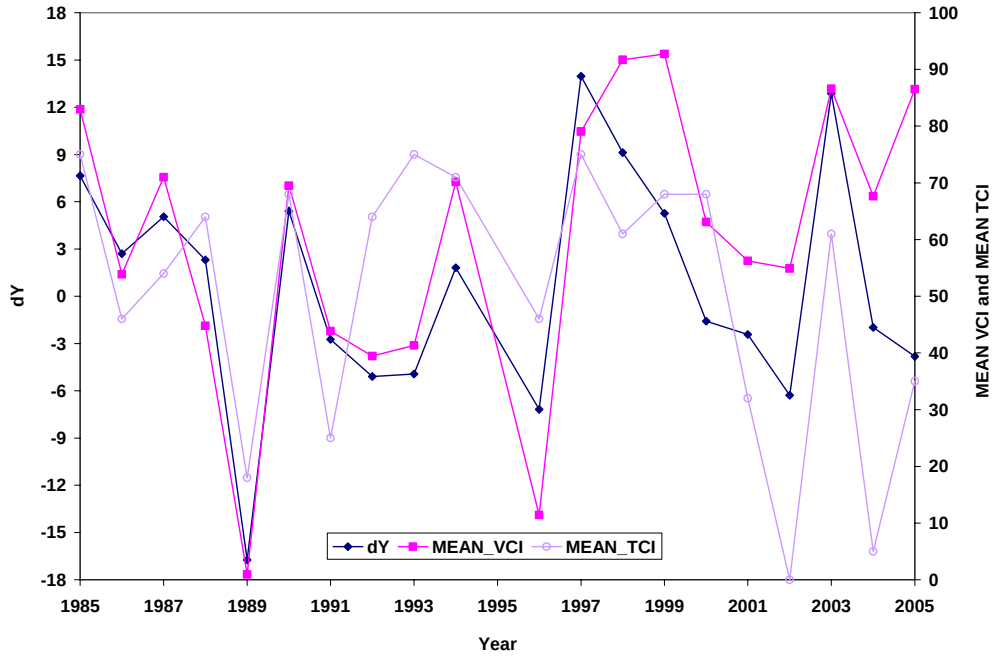

Figure 6.14: Dynamics of dY (WW), MEAN VCI (weeks 12 to 21) and MEAN TCI (week 16) CRD 50, Kansas

Table 6.18 shows correlation coefficients between dY (WW) with $\overline{vci}$ (weeks 12 to 21) and $\overline{tci}$ (weeks 16). These correlation coefficients are significantly different from zero with p-value ($p<0.001$). For WW, $\overline{vci}$ would account for 64% (0.80^2) of the variation in dY if used separately as the only independent variable in the regression model. Besides, $\overline{tci}$ would account for 35% (0.59^2) of the variation in dY (WW) if used separately as the only independent variable in the regression model.

Table 6.18: Correlation matrix dY, MEAN VCI ($\overline{vci}$) and MEAN TCI ($\overline{tci}$), CRD 50, Kansas

Pearson Correlation Coefficients	
	dY
MEAN VCI	0.7950
MEAN TCI	0.5859

6.1.5.2.2 Multiple Regression Results

The results of fitting the ordinary least squares (OLS) regression model approximated by Equation (6.1) to CRD 50 are shown in Table 6.19.

Table 6.19 shows that there is a strong relationship between dY and the independent variables. The coefficient of determination, R-Square, is 0.74 or 74% of the sum of squares in dY (WW) can be associated with the variation in these independent variables. The test of the composite hypothesis that all regression coefficients are 0 is highly significant with F-value of 24.54 compared to F (0.01, 2,

17) = 6.11. The Root MSE is 3.9824 with 17 degrees of freedom is an unbiased estimate of σ.

Table 6.19: Results of the regression of dY on the two independent variables MEAN VCI and MEAN TCI, CRD 50, Kansas

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	2	778.4590	389.2295	24.5421	<.0001
Error	17	269.6145	15.8597	–	–
Corrected Total	19	1048.073	–	–	–

R-Square	*Root MSE*
0.7428	3.9824

Parameter	*Estimate*	*Standard Error*	*t Value*	*Pr > |t|*
Intercept	-16.9840	2.6912	-6.3109	0.0000
MEAN VCI	0.1994	0.0388	5.1383	0.0001
MEAN TCI	0.1110	0.0410	2.7044	0.0150

6.1.5.2.3 Model Validation

For model validation we regressed observed yield versus independently simulated yield and the corresponding statistics were generated (Table 6.20). The overall correlation coefficient 0.8839 is good. An R-Square value of 0.7812 shows that in most years, WW yield in CRD 50, Kansas can be modeled by variables

considered in model Equation (6.1). For WW CRD 50, the model forecast captured 78% of the variability in yield anomalies.

The average prediction bias is -0.0743 (systematic error). The variance of the prediction error is 19.0963 or the standard error of prediction (SEP) is 4.3699 (non-systematic error). The standard error of the estimated mean bias is 0.9771 (Table 6.20). T-tests of the hypothesis that the bias is zero gives t = -0.0761 which, with 19 degrees of freedom and $\alpha = 0.05$, is not significant.

The MSEP is $= \frac{19(19.0963)}{20} + (-0.0743)^2 = 18.1470$. The RMSEP is 4.2599 an approximate 11% error in prediction.

Table 6.20: Statistics of an independent testing for model Equation (6.1), WW, CRD 50, Kansas

Crop	***Correlation***	***R-Square***	***Bias***	***Variance***	***SEP***	***Std Error***	***RMSEP***
WW	0.8839	0.7812	-0.0743	19.0963	4.3699	0.9771	4.2599

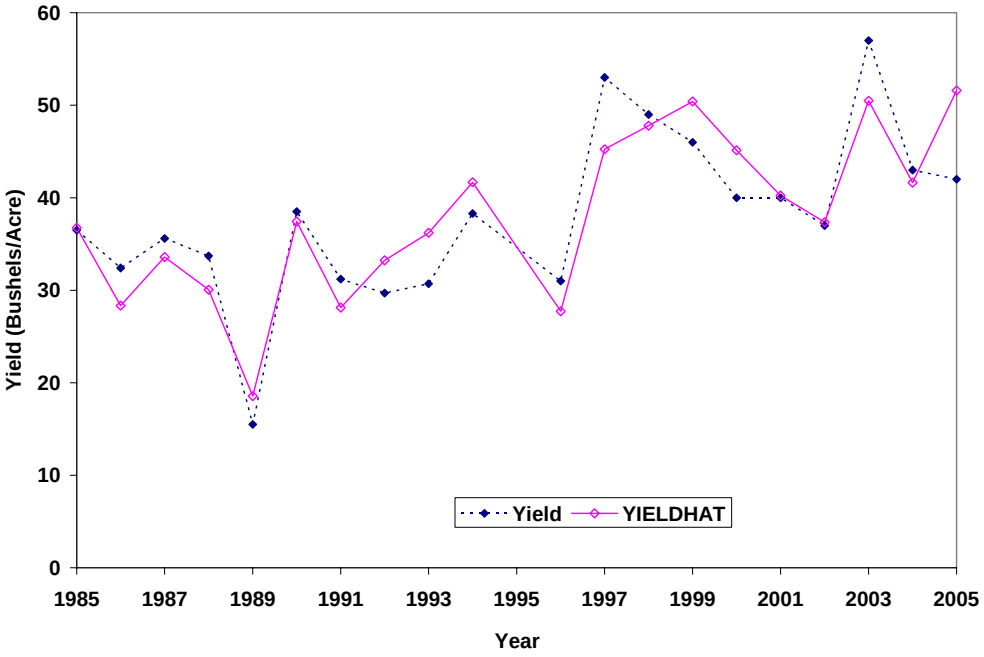

Figure 6.15: Observed yield (WW) versus independently simulated yield (yieldhat), CRD 50, Kansas

Figure 6.15 displays observed versus independently simulated crop yield time series of WW 1985 through to 2005 for CRD 50, Kansas. For WW CRD 50 the model forecasts captured 78% of the variability in yield anomalies.

6.1.6 Crop Reporting District 60

6.1.6.1 Crop Yield Time Series

Figure 6.16 shows that in CRD 60 WW yield increases due to technology improvement.

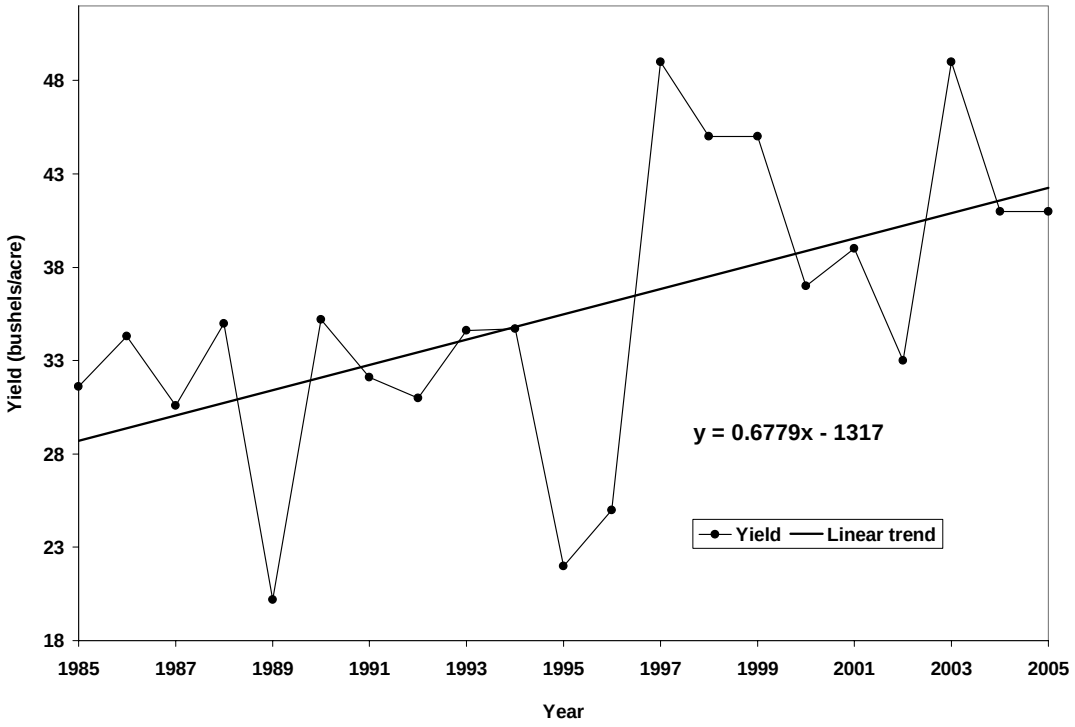

Figure 6.16: Winter wheat yield time series CRD 60, Kansas

6.1.6.2 Regression Analysis

A useful starting point in any multiple regression analysis is to compute correlations among all variables. This provides a first look at the simple linear relationships among them.

6.1.6.2.1 The Correlation Matrix

For CRD 60, VCI (weeks 13 to 19) variables have high correlations with dY (0.77-0.88). VCI18 has the highest correlation with dY. It would account for 77% (0 $.88^2$) of the variation in dY if used separately as the only independent variable in the

regression model. TCI (week 17 to 19) variables also have significant correlation (p-value <0.001) with dY (0.68-0.69) (Table 6.21).

Table 6.21: Correlation matrix dY with VCI (weeks 13 to 19) and TCI (week 17 to 19), CRD 60, Kansas

Pearson Correlation Coefficients	
	dY
VCI13	0.8587
VCI14	0.8728
VCI15	0.8664
VCI16	0.8736
VCI17	0.8615
VCI18	0.8845
VCI19	0.7659
TCI17	0.6791
TCI18	0.6789
TCI19	0.6854

Average VCI and TCI ($\overline{vci}, \overline{tci}$) for weeks with significant Pearson's correlation coefficients at p-level (p<0.001) among dY and VH indices were used as a predictors of dY. Figure 6.17 shows dynamics of dY (WW), MEAN VCI (weeks 13 to 19) and MEAN TCI (weeks 17 to 19) for CRD 60. This figure shows how closely VH indices dynamics follows dY dynamics. This is a strong indication that VH indices can be used as a predictor in a forecast model for winter wheat yield in CRD 60, Kansas.

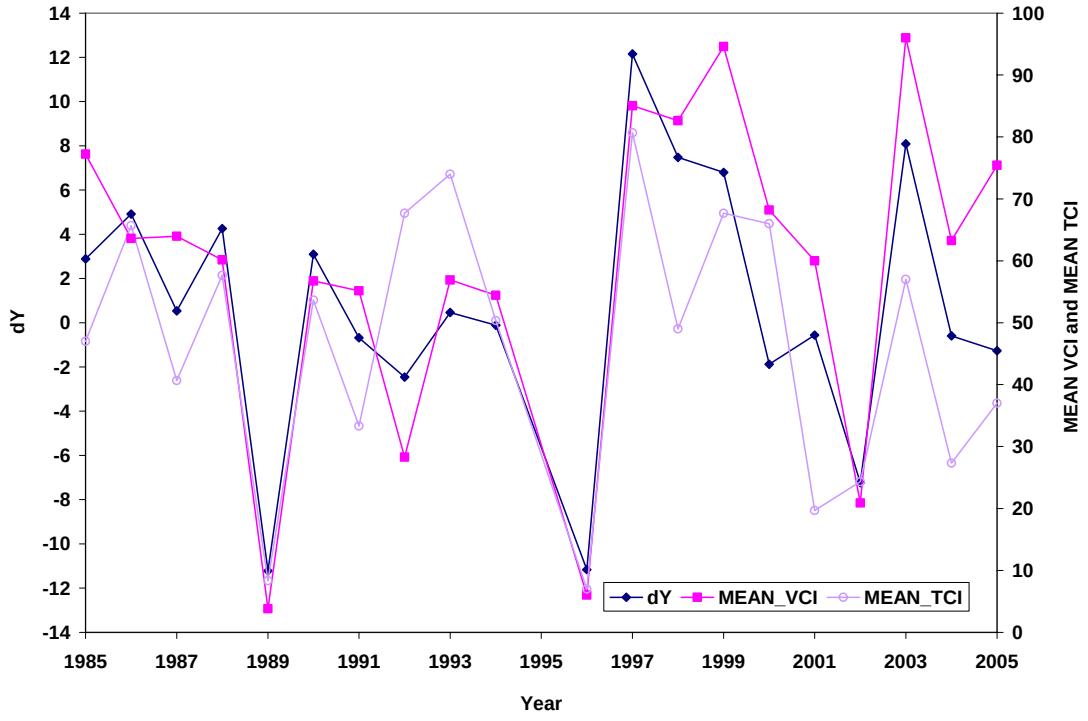

Figure 6.17: Dynamics of dY (WW), MEAN VCI (weeks 13 to 19) and MEAN TCI (weeks 17 to 19), CRD 60, Kansas

Table 6.22: Correlation matrix dY, MEAN VCI ($\overline{vci}$) and MEAN TCI ($\overline{tci}$), CRD 60, Kansas

Pearson Correlation Coefficients	
	dY
MEAN VCI	0.8889
MEAN TCI	0.7455

Table 6.22 shows correlation coefficients between dY (WW) with $\overline{vci}$ (weeks 13 to 19) and $\overline{tci}$ (weeks 17 to 19). These correlation coefficients are significantly different from 0 with p-value ($p<0.001$). For WW, $\overline{vci}$ would account for 79%

(0.89^2) of the variation in dY if used separately as the only independent variable in the regression model. Besides, $\overline{tci}$ would account for 56% (0.75^2) of the variation in dY (WW) if used separately as the only independent variable in the regression model.

6.1.6.2.2 Multiple Regression Results

The results of fitting the ordinary least squares (OLS) regression model approximated by Equation (6.1) to CRD 60 are shown in Table 6.23.

Table 6.23: Results of the regression of dY on the two independent variables MEAN VCI and MEAN TCI, CRD 60, Kansas

Source	***DF***	***Sum of Squares***	***Mean Square***	***F Value***	***Pr > F***
Model	2	585.3155	292.6577	51.3887	<.0001
Error	17	96.8147	5.6950	–	–
Corrected Total	19	682.1302	–	–	–

R-Square	***Root MSE***
0.8581	2.3864

Parameter	***Estimate***	***Standard Error***	***t Value***	***Pr > |t|***
Intercept	-12.8962	1.4536	-8.8722	0.0000
MEAN VCI	0.1588	0.0264	6.0173	0.0000
MEAN TCI	0.0913	0.0320	2.8506	0.0111

Table 6.23 shows that there is a strong relationship among dY and the independent variables. The coefficient of determination, R-Square, is 0.86 or 86% of

the sum of squares in dY (WW) can be associated with the variation in these independent variables. The test of the composite hypothesis that all regression coefficients are 0 is highly significant with F-value of 51.39 compared to F (0.01, 2, 17) = 6.11. The Root MSE is 2.3864 with 17 degrees of freedom is an unbiased estimate of σ.

6.1.6.2.3 Model Validation

For model validation we regressed observed yield versus independently simulated yield and the corresponding statistics were generated (Table 6.24). The overall correlation coefficient 0.9393 is good. An R-Square value of 0.8822 shows that in most years, WW yield in CRD 60 can be modeled by variables considered in model Equation (6.1). For WW CRD 60, the model forecast captured 88% of the variability in yield anomalies.

The average prediction bias is 0.0024 (systematic error). The variance of the prediction error is 6.6741 or the standard error of prediction (SEP) is 2.5834 (non-systematic error). The standard error of the estimated mean bias is 0.5777 (Table 6.24). T-tests of the hypothesis that the bias is zero gives t = 0.0042 which, with 19 degrees of freedom and $\alpha = 0.05$, is not significant.

The MSEP is $= \frac{19(6.6741)}{20} + (0.0024)^2 = 6.3404$. The RMSEP is 2.5180 an approximate 7% error in prediction.

Table 6.24: Statistics of an independent testing for model equation (6.1), WW, CRD 60, Kansas

Crop	*Correlation*	*R-Square*	*Bias*	*Variance*	*SEP*	*Std Error*	*RMSEP*
WW	0.9393	0.8822	0.0024	6.6741	2.5834	0.5777	6.3404

Figure 6.18 displays observed versus independently simulated crop yield time series of WW 1985 through to 2005 for CRD 60, Kansas. For WW CRD 60 the model forecasts captured 88% of the variability in yield anomalies.

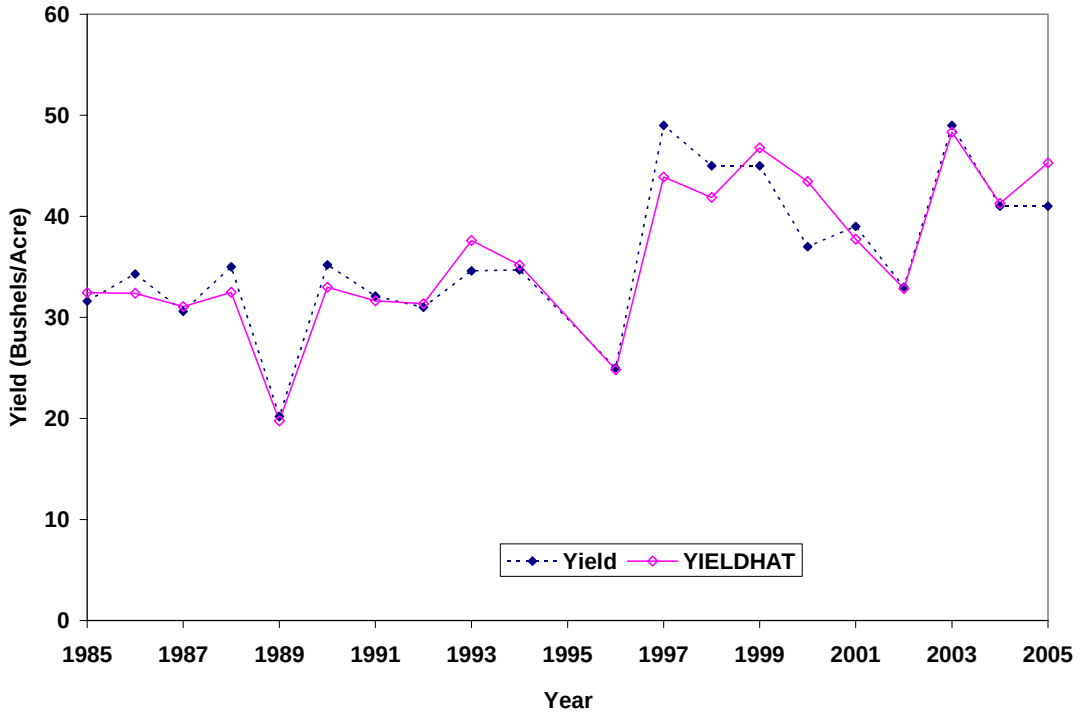

Figure 6.18: Observed yield (WW) versus independently simulated yield (yieldhat), CRD 60, Kansas

6.2 Sorghum

6.2.1 Total Kansas

6.2.1.1 Crop yield time series

Figure 6.19 shows slight decrease in the long-term sorghum yield trend for Kansas.

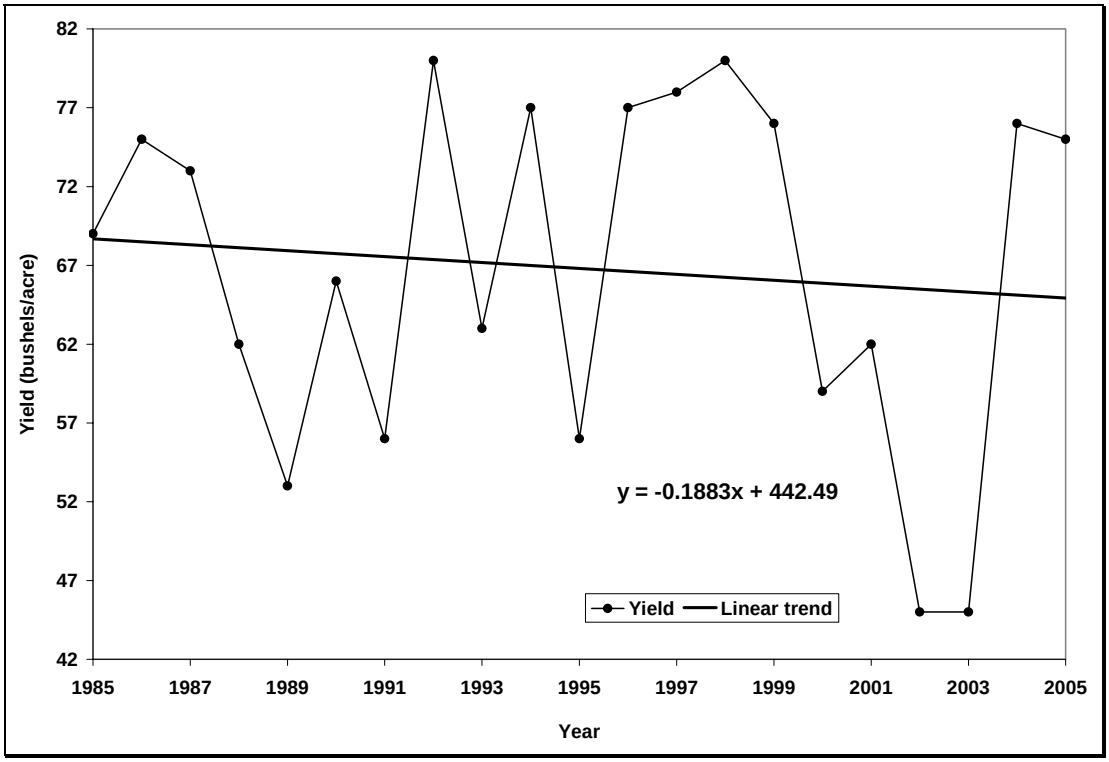

Figure 6.19: Sorghum yield time series, Kansas

6.2.1.2 Regression Analysis

A useful starting point in any multiple regression analysis is to compute correlations among all variables. This provides a first look at the simple linear relationships among them.

6.2.1.2.1 The Correlation Matrix

For total Kansas, VCI (weeks 28 to 37) variables have high correlations with dY (0.76-0.88). VCI33 has the highest correlation with dY. It would account for 77% (0.88^2) of the variation in dY if used separately as the only independent variable in the regression model. TCI (week 27 to 33) variables also have significant correlation (p-value <0.001) with dY (0.77-0.87) (Table 6.25).

Average VCI and TCI ($\overline{vci}, \overline{tci}$) for weeks with significant Pearson's correlation coefficients at p-level (p<0.001) among dY and VH indices were used as a predictors of dY. Figure 6.20 shows dynamics of dY (sorghum), MEAN VCI (weeks 28 to 37) and MEAN TCI (weeks 27 to 33) for Kansas. This figure shows how closely VH indices dynamics follows dY dynamics. This is a strong indication that VH indices can be used as a predictor in a forecast model for sorghum yield in Kansas.

Table 6.25: Correlation matrix dY (sorghum) with VCI (weeks 28 to 37) and TCI (week 27 to 33) Kansas

Pearson Correlation Coefficients	
	dY
VCI28	0.7873
VCI29	0.8088
VCI30	0.8182
VCI31	0.8682
VCI32	0.8761
VCI33	0.8848
VCI34	0.8792
VCI35	0.8751
VCI36	0.8506
VCI37	0.7649
TCI27	0.7825
TCI28	0.8416
TCI29	0.8602
TCI30	0.8667
TCI31	0.8283
TCI32	0.8023
TCI33	0.7661

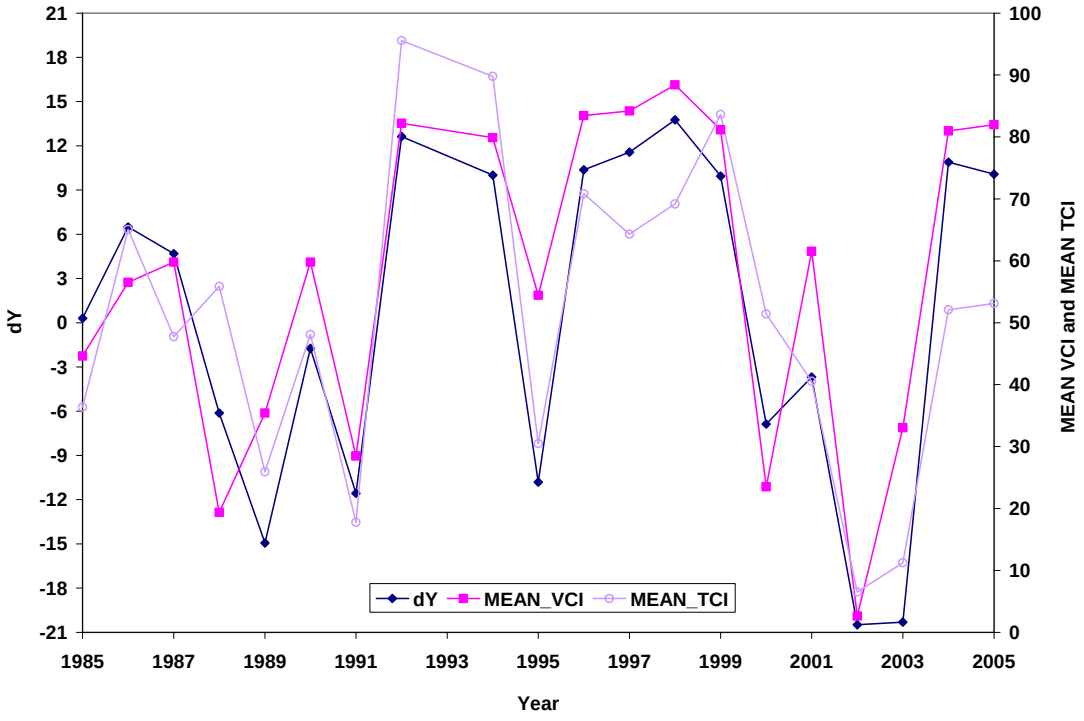

Figure 6.20: Dynamics of dY (sorghum), MEAN VCI (weeks 28 to 37) and MEAN TCI (weeks 27 to 33), Kansas

Table 6.26 shows correlation coefficients between dY (sorghum) with $\overline{vci}$ (weeks 28 to 37) and $\overline{tci}$ (weeks 27 to 33). These correlation coefficients are significantly different from 0 with p-value ($p<0.001$). For sorghum, $\overline{vci}$ would account for 79% (0.89^2) of the variation in dY if used separately as the only independent variable in the regression model. Besides, $\overline{tci}$ would account for 76% (0.87^2) of the variation in dY (sorghum) if used separately as the only independent variable in the regression model.

Table 6.26: Correlation matrix dY (sorghum), MEAN VCI ($\overline{vci}$) and MEAN TCI ($\overline{tci}$), Kansas

Pearson Correlation Coefficients	
	dY
MEAN VCI	0.8920
MEAN TCI	0.8714

6.2.1.2.2 Multiple Regression Results

The results of fitting the ordinary least squares (OLS) regression model approximated by Equation 6.1 to Kansas are shown in Table 6.27.

Table 6.27: Results of the regression of dY (sorghum) on the two independent variables MEAN VCI and MEAN TCI, Kansas

Source	***DF***	***Sum of Squares***	***Mean Square***	***F Value***	***Pr > F***
Model	2	2212.209	1106.105	70.3408	<.0001
Error	17	267.3237	15.7249	–	–
Corrected Total	19	2479.533	–	–	–

R-Square	***Root MSE***
0.8922	3.9655

Parameter	***Estimate***	***Standard Error***	***t Value***	***Pr > |t|***
Intercept	-24.3011	2.2521	-10.7905	0.0000
MEAN VCI	0.2392	0.0523	4.5779	0.0003
MEAN TCI	0.2138	0.0548	3.9025	0.0011

Table 6.27 shows that there is a strong relationship between dY and the independent variables. The coefficient of determination, R-Square, is 0.89 or 89% of the sum of squares in dY (sorghum) can be associated with the variation in these independent variables. The test of the composite hypothesis that all regression coefficients are 0 is highly significant with F-value of 70.34 compared to F (0.01, 2, 17) = 6.11. The Root MSE is 3.9655 with 17 degrees of freedom is an unbiased estimate of σ.

6.2.1.2.3 Model Validation

For model validation we regressed observed yield versus independently simulated yield and the corresponding statistics were generated (Table 6.28). The overall correlation coefficient 0.9261 is good. An R-Square value of 0.8577 shows that in most years, sorghum yield in Kansas can be modeled by variables considered in model Equation (6.1). For sorghum Kansas, the model forecast captured 86% of the variability in yield anomalies.

The average prediction bias is -0.0260 (systematic error). The variance of the prediction error is 18.8722 or the SEP is 4.3442 (non-systematic error). The standard error of the estimated mean bias is 0.9714 (Table 6.28). T-tests of the hypothesis that the bias is zero gives t = -0.0268 which, with 19 degrees of freedom and $\alpha = 0.05$, is not significant.

The MSEP is $= \frac{19(18.8722)}{20} + (-0.0260)^2 = 17.9293$. The RMSEP is 4.2343

an approximate 6% error in prediction.

Table 6.28: Statistics of an independent testing for model Equation (6.1), sorghum, Kansas

Crop	*Correlation*	*R-Square*	*Bias*	*Variance*	*SEP*	*Std Error*	*RMSEP*
Sorghum	0.9261	0.8577	-0.0260	18.8722	4.3442	0.9714	4.2343

Figure 6.21 displays observed versus independently simulated crop yield time series of sorghum 1985 through to 2005 for Kansas. For sorghum, Kansas, the model forecasts captured 86% of the variability in yield anomalies.

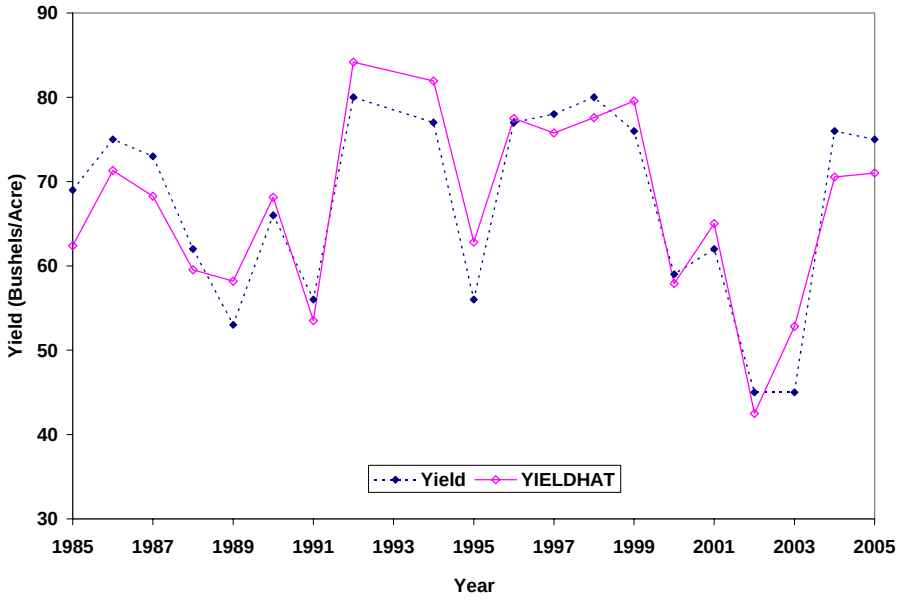

Figure 6.21: Observed yield (sorghum) versus independently simulated yield (yieldhat), Kansas

6.2.2 Crop Reporting District 40

6.2.2.1 Crop yield time series

Figure 6.22 shows that in CRD 40 sorghum yield increases due to technology improvement.

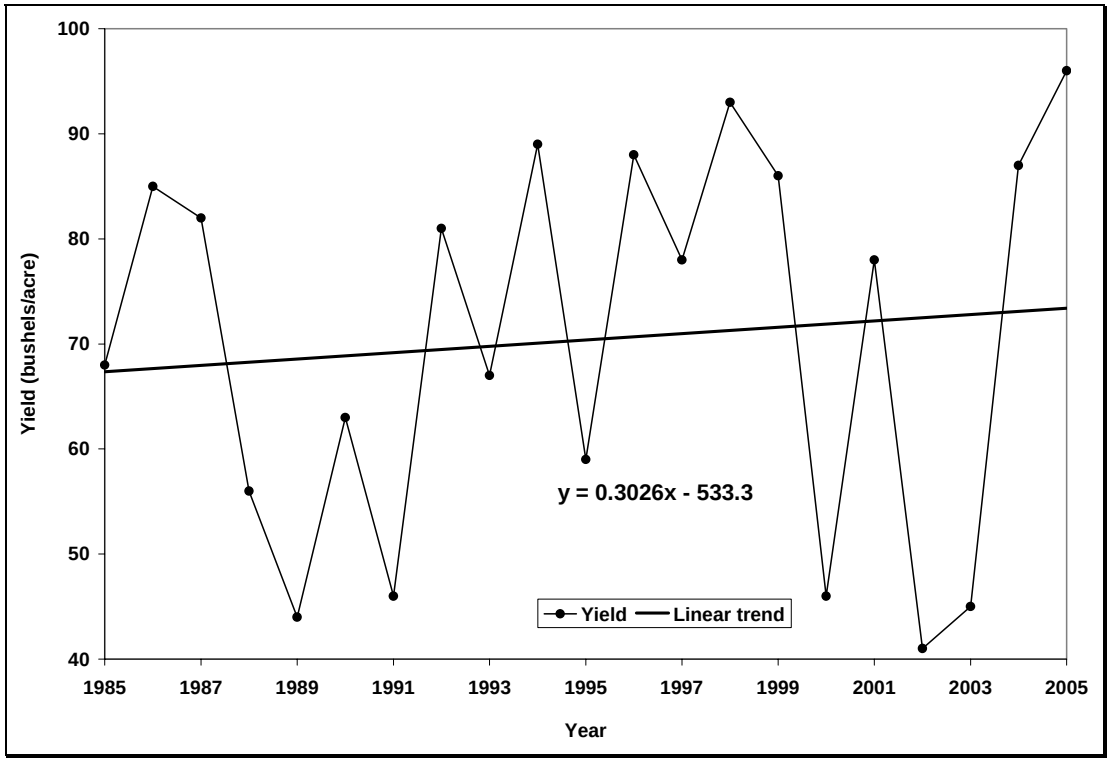

Figure 6.22: Sorghum yield time series CRD 40, Kansas

6.2.2.2 Regression Analysis

A useful starting point in any multiple regression analysis is to compute correlations among all variables. This provides a first look at the simple linear relationships among them.

6.2.2.2.1 The Correlation Matrix

For CRD 40, VCI (weeks 27 to 35) variables have high correlations with dY (0.76-0.90). VCI31 has the highest correlation with dY. It would account for 81% (0 $.90^2$) of the variation in dY if used separately as the only independent variable in the regression model. TCI (week 29 to 30) variables also have significant correlation (p-value <0.001) with dY (0.77-0.80) (Table 6.29).

Average VCI and TCI ($\overline{vci}, \overline{tci}$) for weeks with significant Pearson's correlation coefficients at p-level (p<0.001) among dY and VH indices were used as a predictors of dY. Figure 6.23 shows dynamics of dY (sorghum), MEAN VCI (weeks 27 to 35) and MEAN TCI (weeks 29 to 30) for CRD 40, Kansas. This figure shows how closely VH indices dynamics follows dY dynamics. This is a strong indication that VH indices can be used as a predictor in a forecast model for sorghum yield in CRD 40, Kansas.

Table 6.29: Correlation matrix dY (sorghum) with VCI (weeks 27 to 35) and TCI (week 29 to 30), CRD 40, Kansas

Pearson Correlation Coefficients	
	dY
VCI27	0.7630
VCI28	0.8098
VCI29	0.8272
VCI30	0.8685
VCI31	0.9024
VCI32	0.8990
VCI33	0.8862
VCI34	0.8297
VCI35	0.7681
TCI29	0.7713
TCI30	0.7997

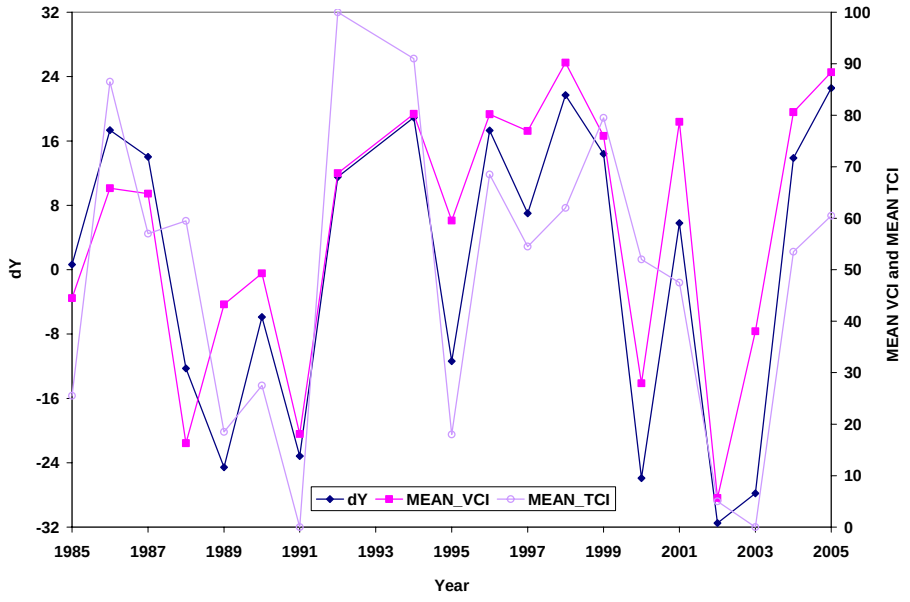

Figure 6.23: Dynamics of dY (sorghum), MEAN VCI (weeks 27 to 35) and MEAN TCI (weeks 29 to 30), CRD 40, Kansas

Table 6.30 shows correlation coefficients between dY (sorghum) with $\overline{vci}$ (weeks 27 to 35) and $\overline{tci}$ (weeks 29 to 30). These correlation coefficients are significantly different from 0 with p-value ($p<0.001$). For sorghum, $\overline{vci}$ would account for 79% (0.89^2) of the variation in dY if used separately as the only independent variable in the regression model. Besides, $\overline{tci}$ would account for 62% (0.79^2) of the variation in dY (sorghum) if used separately as the only independent variable in the regression model.

Table 6.30: Correlation matrix dY (sorghum), MEAN VCI ($\overline{vci}$) and MEAN TCI ($\overline{tci}$), CRD 40, Kansas

Pearson Correlation Coefficients	***dY***
MEAN VCI	0.8922
MEAN TCI	0.7891

6.2.2.2.2 Multiple Regression Results

The results of fitting the ordinary least squares (OLS) regression model approximated by Equation (6.1) to CRD 40 are shown in Table 6.31.

Table 6.31 shows that there is a strong relationship among dY and the independent variables. The coefficient of determination, R-Square, is 0.88 or 88% of the sum of squares in dY (sorghum) can be associated with the variation in these independent variables. The test of the composite hypothesis that all regression

coefficients are 0 is highly significant with F-value of 62.04 compared to F (0.01, 2, 17) = 6.11. The Root MSE is 6.8624 with 17 degrees of freedom is an unbiased estimate of σ.

Table 6.31: Results of the regression of dY (sorghum) on the two independent variables MEAN VCI and MEAN TCI, CRD 40, Kansas

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	2	5843.781	2921.890	62.0455	<.0001
Error	17	800.5765	47.0927	–	–
Corrected Total	19	6644.357	–	–	–

R-Square	*Root MSE*
0.8795	6.8624

Parameter	*Estimate*	*Standard Error*	*t Value*	*Pr > |t|*
Intercept	-38.4553	3.8575	-9.9690	0.0000
MEAN VCI	0.4751	0.0789	6.0196	0.0000
MEAN TCI	0.2318	0.0676	3.4305	0.0032

6.2.2.2.3 Model Validation

For model validation we regressed observed yield versus independently simulated yield and the corresponding statistics were generated (Table 6.32). The overall correlation coefficient 0.9151 is good. An R-Square value of 0.8374 shows that in most years, sorghum yield in CRD 40 can be modeled by variables considered

in model Equation (6.1). For sorghum CRD 40, the model forecast captured 84% of the variability in yield anomalies.

The average prediction bias is -0.0185 (systematic error). The variance of the prediction error is 58.3723 or the SEP is 7.6402 (non-systematic error). The standard error of the estimated mean bias is 1.7084 (Table 6.32). T-tests of the hypothesis that the bias is zero gives t = -0.0108 which, with 19 degrees of freedom and $\alpha = 0.05$, is not significant.

The MSEP is $= \frac{19(58.3723)}{20} + (-0.0185)^2 = 55.4540$. The RMSEP is 7.4467 an approximate 11% error in prediction.

Table 6.32: Statistics of an independent testing for model Equation (6.1), sorghum, CRD 40, Kansas

Crop	*Correlation*	*R-Square*	*Bias*	*Variance*	*SEP*	*Std Error*	*RMSEP*
Sorghum	0.9151	0.8374	-0.0185	58.3723	7.6402	1.7084	7.4467

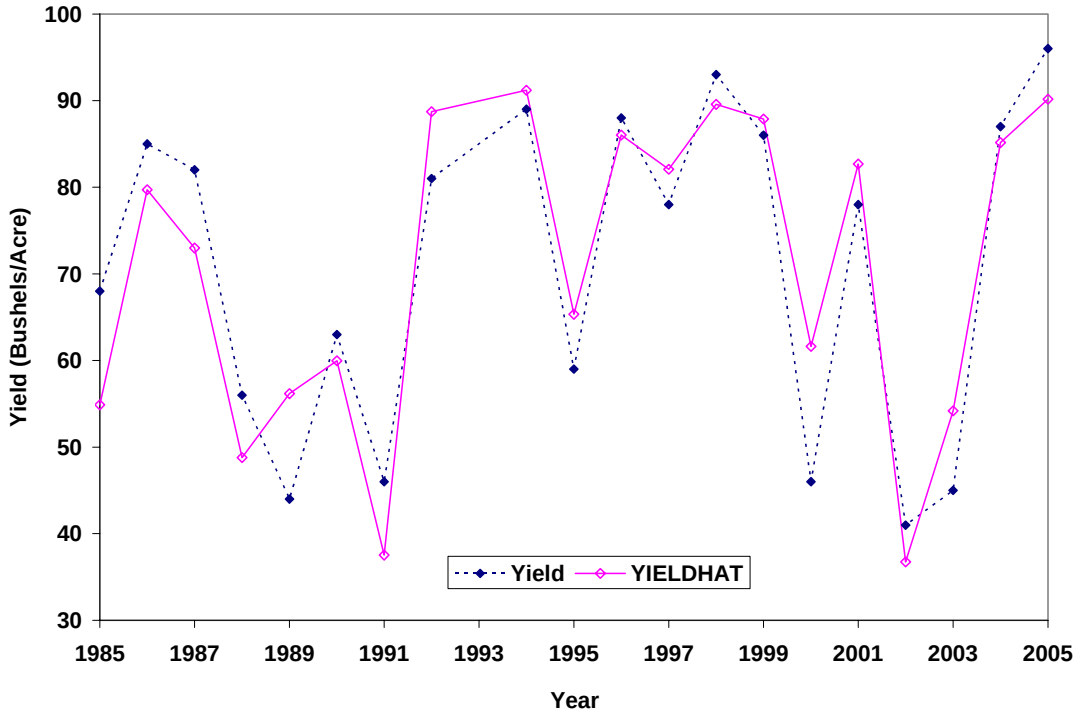

Figure 6.24: Observed yield (sorghum) versus independently simulated yield (yieldhat), CRD 40, Kansas

Figure 6.24 displays observed versus independently simulated crop yield time series of sorghum 1985 through to 2005 for CRD 40, Kansas. For sorghum CRD 40 the model forecasts captured 84% of the variability in yield anomalies.

6.2.3 Marshall County

6.2.3.1 Crop Yield Time Series

Figure 6.25 shows that in Marshall County sorghum yield increases due to technology improvement.

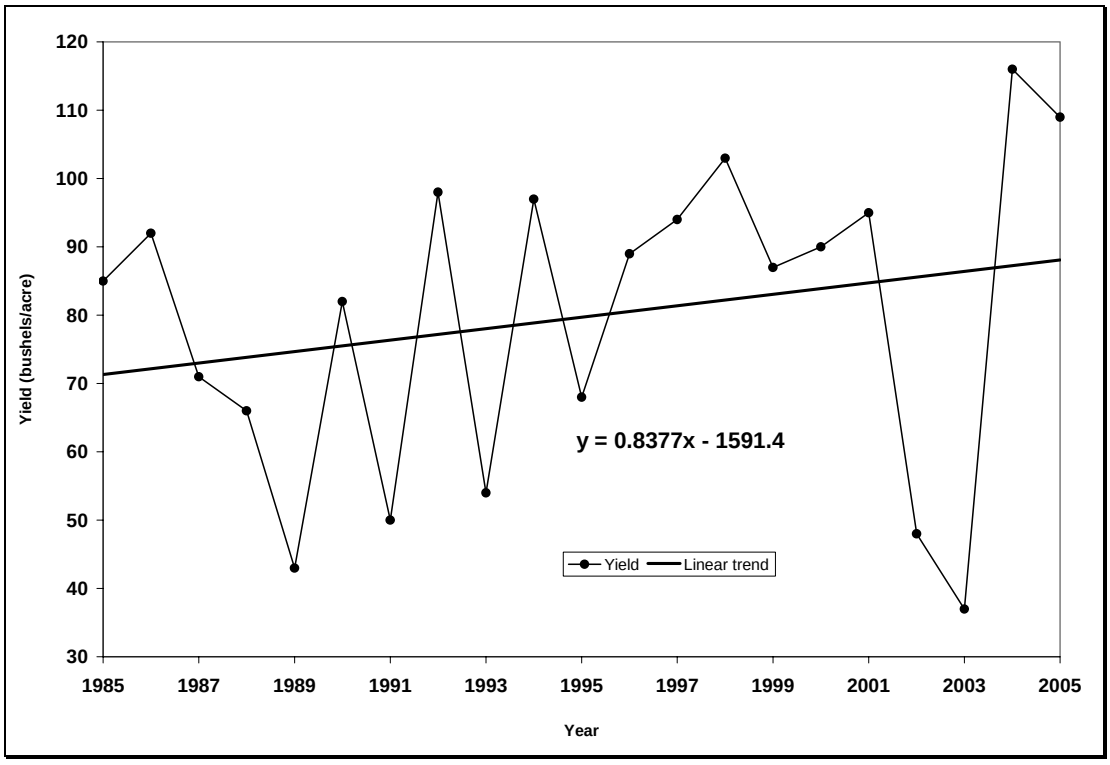

Figure 6.25: Sorghum yield time series Marshall County, Kansas

6.2.3.2 Regression Analysis

A useful starting point in any multiple regression analysis is to compute correlations among all variables. This provides a first look at the simple linear relationships among them.

6.2.3.2.1 The Correlation Matrix

For Marshall County, VCI (weeks 31 to 34) variables have high correlations with dY (0.69-0.74). VCI32 has the highest correlation with dY. It would account for

55% (0 .74^2) of the variation in dY if used separately as the only independent variable in the regression model. TCI (week 32) variable also has significant correlation (p-value <0.001) with dY (0.79) (Table 6.33).

Table 6.33: Correlation matrix dY (sorghum), VCI (weeks 31 to 34) and TCI (week 32) Marshall County, Kansas

Pearson Correlation Coefficients	
	dY
VCI31	0.7369
VCI32	0.7414
VCI33	0.7085
VCI34	0.6921
TCI32	0.7906

Average VCI and TCI ($\overline{vci}, \overline{tci}$) for weeks with significant Pearson's correlation coefficients at p-level (p<0.001) among dY and VH indices were used as a predictors of dY. Figure 6.26 shows dynamics of dY (sorghum), MEAN VCI (weeks 31 to 34) and MEAN TCI (week 32) for Marshall County. This figure shows how closely VH indices dynamics follows dY dynamics. This is a strong indication that VH indices can be used as a predictor in a forecast model for sorghum yield in Marshall County, Kansas.

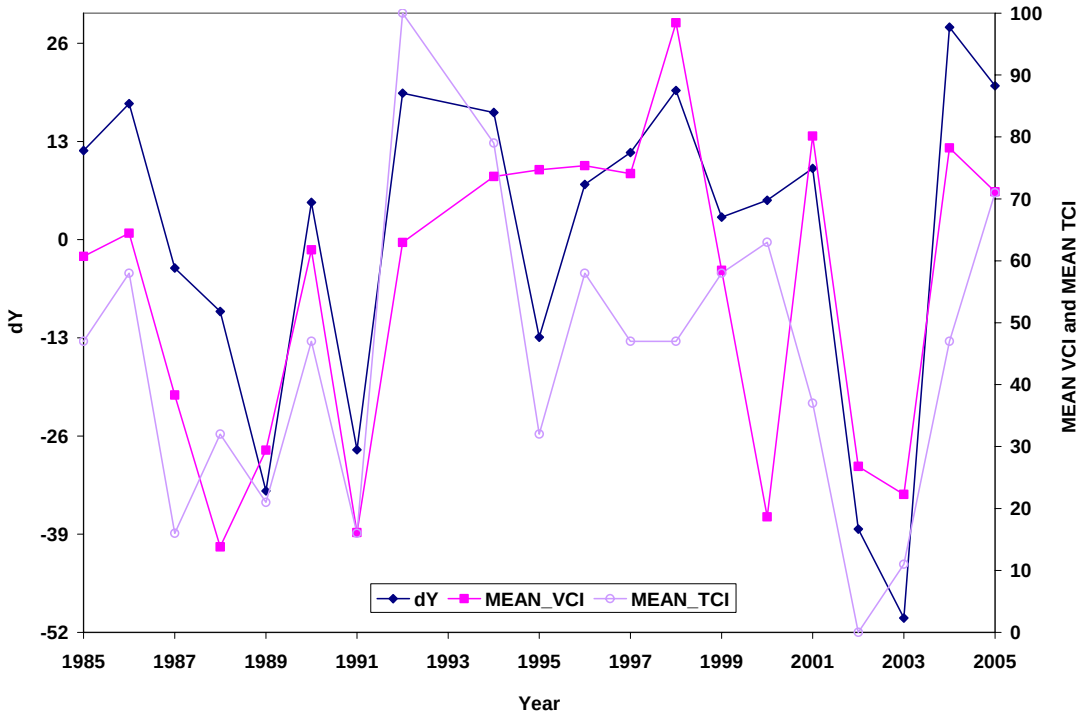

Figure 6.26: Dynamics of dY (sorghum), MEAN VCI (weeks 31 to 34) and MEAN TCI (week 32) for Marshall County, Kansas

Table 6.34 shows correlation coefficients among dY (sorghum) with MEAN VCI and MEAN TCI for Marshal County. These correlation coefficients are significant different from zero ($p<0.0001$). MEAN VCI would account for 55% ($r^2 = 0.74$) of the variation in dY if used separately as the only independent variable in the regression model. In addition, the correlation coefficient between dY and MEAN TCI would account for 62% ($r^2 = 0.79$) of the variation in dY if used separately as the only independent variable in the regression model.

Table 6.34: Correlation matrix dY (sorghum), MEAN VCI (weeks 31 to 34) and MEAN TCI (week 32) for Marshall County, Kansas

Pearson Correlation Coefficients	
	dY
MEAN VCI	0.7363
MEAN TCI	0.7906

6.2.3.2.2 Multiple Regression Results

The results of fitting the ordinary least squares (OLS) regression model approximated by Equation (6.1) to Marshal County are shown in Table 6.35.

Table 6.35: Results of the regression of dY (sorghum) on the two independent variables MEAN VCI and MEAN TCI, Marshal County, Kansas

Source	***DF***	***Sum of Squares***	***Mean Square***	***F Value***	***Pr > F***
Model	2	7103.430	3551.715	28.9384	<.0001
Error	17	2086.471	122.7336	–	–
Corrected Total	19	9189.900	–	–	–

R-Square	***Root MSE***
0.7730	11.0785

Parameter	***Estimate***	***Standard Error***	***t Value***	***Pr > |t|***
Intercept	-43.3805	6.3449	-6.8371	0.0000
MEAN VCI	0.3858	0.1160	3.3273	0.0040
MEAN TCI	0.4999	0.1202	4.1578	0.0007

Table 6.35 shows that there is a strong relationship among dY and the independent variables. The coefficient of determination, R-Square, is 0.77 or 77% of the sum of squares in dY (sorghum) can be associated with the variation in these independent variables. The test of the composite hypothesis that all regression coefficients are 0 is highly significant with F-value of 28.94 compared to F (0.01, 2, 17) = 6.11. The Root MSE is 11.0785 with 17 degrees of freedom is an unbiased estimate of σ.

6.2.3.2.3 Model Validation

For model validation we regressed observed yield versus independently simulated yield and the corresponding statistics were generated (Table 6.36). The overall correlation coefficient 0.8340 is good. An R-Square value of 0.6955 shows that in most years, sorghum yield in Marshal County can be modeled by variables considered in model Equation (6.1). For sorghum Marshal County, the model forecast captured 70% of the variability in yield anomalies.

The average prediction bias is -0.2061 (systematic error). The variance of the prediction error is 159.5506 or the SEP is 12.6313 (non-systematic error). The standard error of the estimated mean bias is 2.8245 (Table 6.36). T-tests of the hypothesis that the bias is 0 gives t = -0.0730 which, with 19 degrees of freedom and α = 0.05, is not significant. The MSEP is $= \frac{19(159.5506)}{20} + (-0.2061)^2 = 151.6155$. The RMSEP is 12.3132 an approximate 15% error in prediction.

Table 6.36: Statistics of an independent testing for model Equation (6.1), sorghum, Marshal County, Kansas

Crop	*Correlation*	*R-Square*	*Bias*	*Variance*	*SEP*	*Std Error*	*RMSEP*
Sorghum	0.8340	0.6955	-0.2061	159.5506	12.6313	2.8245	12.3132

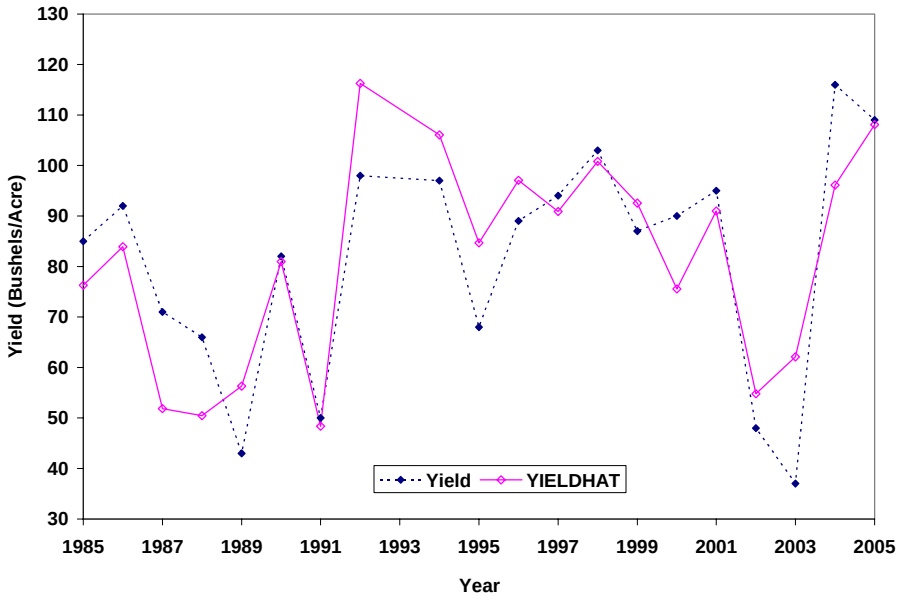

Figure 6.27: Observed yield (sorghum) versus independently simulated yield (yieldhat) Marshal County, Kansas

Figure 6.27 displays observed versus independently simulated crop yield time series of sorghum 1985 through to 2005 for Marshal County, Kansas. For sorghum Marshal County the model forecasts captured 70% of the variability in yield anomalies.

6.3 Corn

6.3.1 Total Kansas

6.3.1.1 Crop Yield Time Series

Figure 6.28 shows that in total Kansas, corn yield increases due to technology improvement.

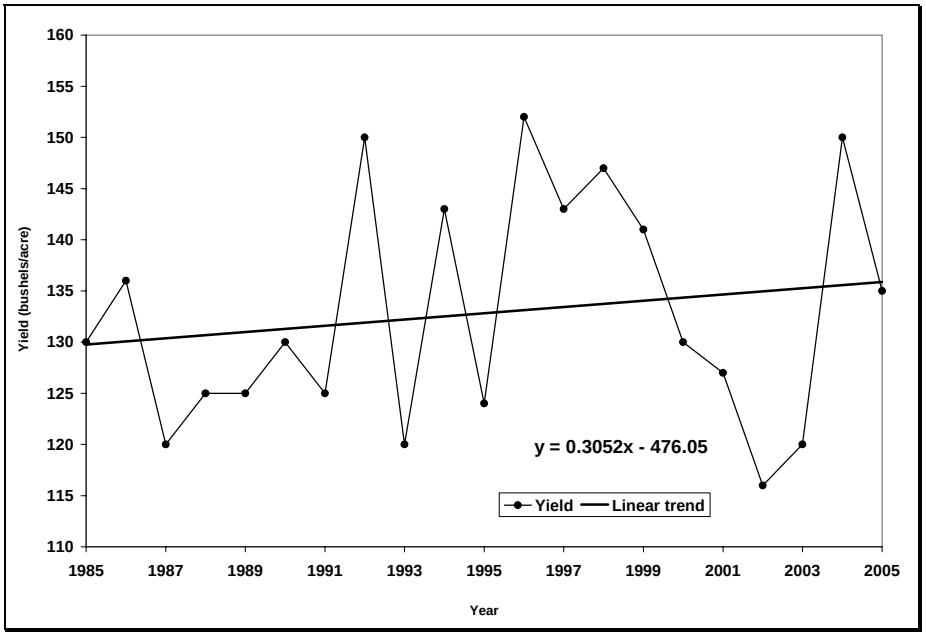

Figure 6.28: Corn yield time series, Kansas

6.3.1.2 Regression Analysis

A useful starting point in any multiple regression analysis is to compute correlations among all variables. This provides a first look at the simple linear relationships among them.

6.3.1.2.1 The Correlation Matrix

Table 6.37 contains the Pearson correlation coefficients among the dependent variable dY (corn) with VCI (weeks 31 to 37) and TCI (weeks 28 to 31) for Kansas. All VCI variables have high correlations with dY (0.78-0.87). VCI33 has the highest correlation with dY. It would account for 77% (r^2 =0 .88) of the variation in dY if used separately as the only independent variable in the regression model. In addition, TCI variables have significant correlation with dY (0.77-0.90).

Table 6.37: Correlation matrix of dY (corn) with VCI (week 31 to 37) and TCI (week 28 to 31), Kansas

Pearson Correlation Coefficients	
	dY
VCI31	0.8539
VCI32	0.8638
VCI33	0.8798
VCI34	0.8528
VCI35	0.8629
VCI36	0.8194
VCI37	0.7771
TCI28	0.7692
TCI29	0.8497
TCI30	0.8746
TCI31	0.8966
TCI32	0.8834
TCI33	0.8905
TCI34	0.8210

Average VCI and TCI ($\overline{vci}, \overline{tci}$) for weeks with significant Pearson's correlation coefficients at p-level ($p<0.001$) among dY and VH indices were used as predictors of dY. Figure 6.29 shows dynamics of dY (corn), MEAN VCI (weeks 31 to 37) and MEAN TCI (weeks 30 to 32) for Kansas. This figure shows how closely VH indices dynamics follows dY dynamics. This is a strong indication that VH indices can be used as predictors in a forecast model for corn yield in Kansas.

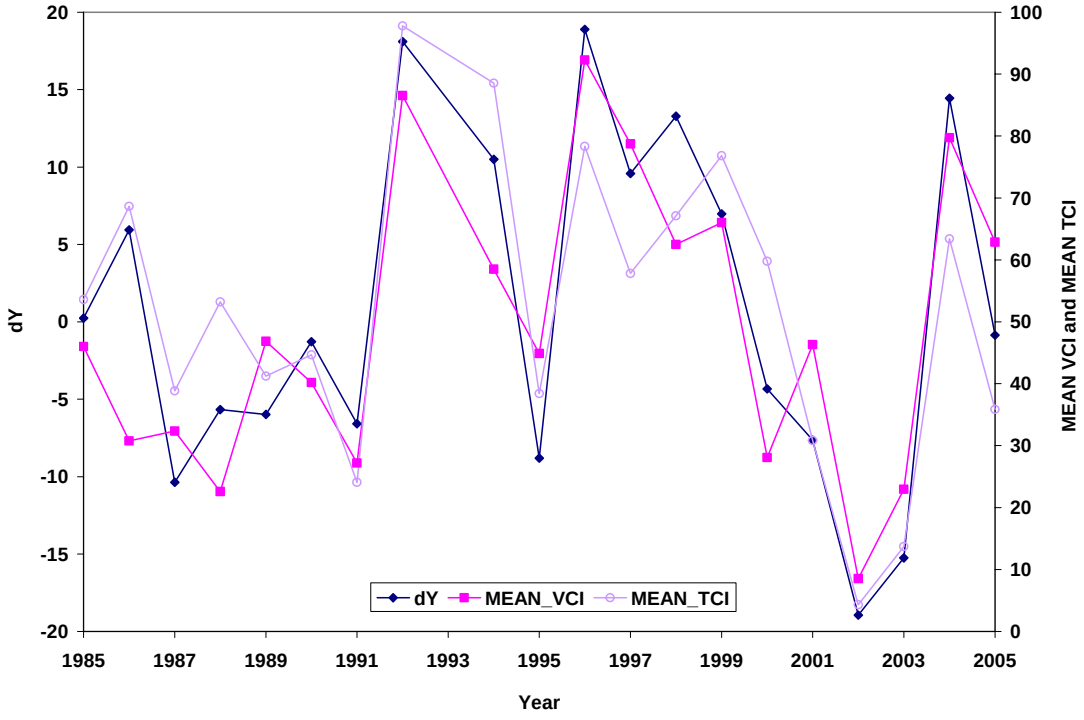

Figure 6.29: Dynamics of dY (corn), MEAN VCI (weeks 31 to 37) and MEAN TCI (weeks 28 to 31), Kansas

Table 6.38 shows correlation coefficients between dY (corn) with $\overline{vci}$ (weeks 31 to 37) and $\overline{tci}$ (weeks 28 to 31). These correlation coefficients are significantly

different from 0 with p-value ($p<0.001$). For corn, $\overline{vci}$ would account for 76% (0.87^2) of the variation in dY if used separately as the only independent variable in the regression model. Besides, $\overline{tci}$ would account for 79% (0.89^2) of the variation in dY (corn) if used separately as the only independent variable in the regression model.

Table 6.38: Correlation matrix of dY (corn) with MEAN VCI ($\overline{vci}$) and MEAN TCI ($\overline{tci}$), Kansas

Pearson Correlation Coefficients	
	dY
MEAN VCI	0.8718
MEAN TCI	0.8906

6.3.1.2.2 Multiple Regression Results

The results of fitting the ordinary least squares (OLS) regression model approximated by Equation (6.1) to Kansas are shown in Table 6.39.

Table 6.39 shows that there is a strong relationship among dY and the independent variables. The coefficient of determination, R-Square, is 0.92 or 92% of the sum of squares in dY (corn) can be associated with the variation in these independent variables. The test of the composite hypothesis that all regression coefficients are 0 is highly significant with F-value of 91.77 compared to F (0.01, 2, 17) = 6.11. The Root MSE is 3.3977 with 17 degrees of freedom is an unbiased estimate of σ.

Table 6.39: Results of the regression of dY (corn) on the two independent variables MEAN VCI and MEAN TCI, Kansas

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	2	2119.002	1059.501	91.7757	<.0001
Error	17	196.2557	11.5445	–	–
Corrected Total	19	2315.257	–	–	–

R-Square	*Root MSE*
0.9152	3.3977

Parameter	*Estimate*	*Standard Error*	*t Value*	*Pr > |t|*
Intercept	-23.6035	1.9420	-12.1542	0.0000
MEAN VCI	0.2289	0.0463	4.9481	0.0001
MEAN TCI	0.2498	0.0448	5.5801	0.0000

6.3.1.2.3 Model Validation

For model validation we regressed observed yield versus independently simulated yield and the corresponding statistics were generated (Table 6.40). The overall correlation coefficient 0.9432 is good. An R-Square value of 0.8897 shows that in most years, corn yield in Kansas can be modeled by variables considered in model Equation (6.1). For corn, in Kansas, the model forecast captured 89% of the variability in yield anomalies.

The average prediction bias is 0.0613 (systematic error). The variance of the prediction error is 13.8262 or the SEP is 3.7184 (non-systematic error). The standard

error of the estimated mean bias is 0.8314 (Table 6.40). T-tests of the hypothesis that the bias is zero gives t = 0.0738 which, with 19 degrees of freedom and $\alpha = 0.05$, is not significant. The MSEP is $= \frac{19(13.8262)}{20} + (0.0613)^2 = 13.1386$. The RMSEP is 3.6247 an approximate 3% error in prediction.

Table 6.40: Statistics of an independent testing for model Equation (6.1), corn, Kansas

Crop	***Correlation***	***R-Square***	***Bias***	***Variance***	***SEP***	***Std Error***	***RMSEP***
Corn	0.9432	0.8897	0.0613	13.8262	3.7184	0.8314	3.6247

Figure 6.30 displays observed versus independently simulated crop yield time series of corn 1985 through to 2005 for Kansas. For corn, in Kansas, the model forecasts captured 89% of the variability in yield anomalies.

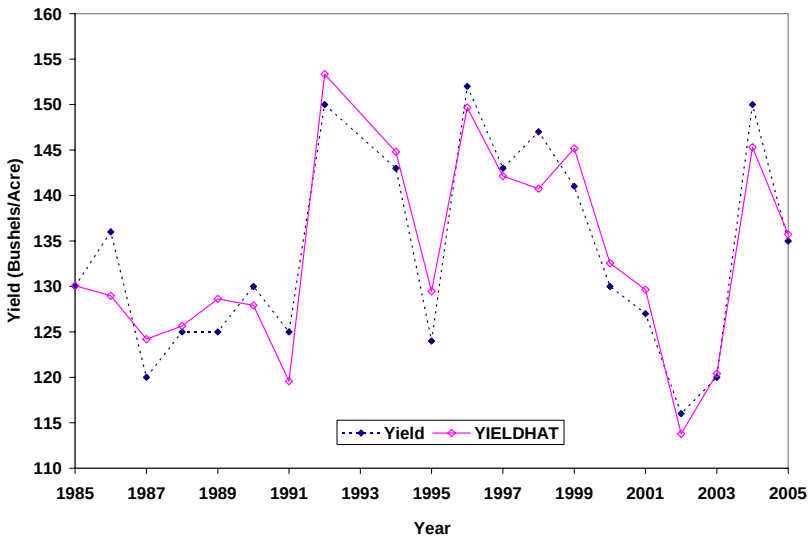

Figure 6.30: Observed yield (corn) versus independently simulated yield (yieldhat), Kansas

6.3.2 Crop Reporting District 30

6.3.2.1 Crop yield time series

Figure 6.31 shows that in CRD 30, corn yield increases due to technology improvement.

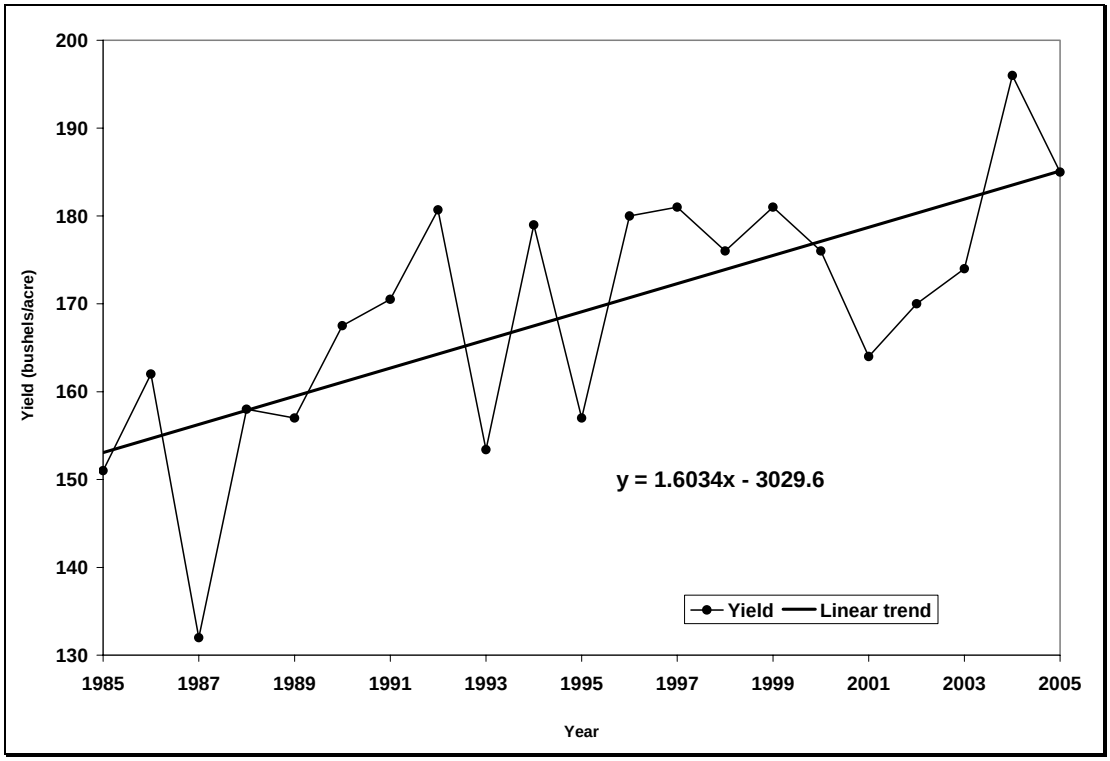

Figure 6.31: Corn yield time series CRD 30, Kansas

6.3.2.2 Regression Analysis

6.3.2.2.1 The Correlation Matrix

Table 6.41 contains Pearson's correlation coefficients among the dependent variable dY (corn) with VCI (weeks 28 to 36) and TCI (weeks 30 to 33) for CRD 30. All VCI variables have high correlations with dY (0.62-0.70). VCI33 has the highest correlation with dY. It would account for 49 % (r^2 =0 .70) of the variation in dY if used separately as the only independent variable in the regression model. Besides, TCI variables have significant correlation with dY (0.79-0.84).

Table 6.41: Correlation matrix dY (corn), VCI (weeks 28 to 36) and TCI (weeks 30 to 33) for CRD 30, Kansas

Pearson Correlation Coefficients	
	dY
VCI28	0.6400
VCI29	0.6600
VCI30	0.6558
VCI31	0.6496
VCI32	0.6401
VCI33	0.7015
VCI34	0.7026
VCI35	0.6848
VCI36	0.6170
TCI30	0.8446
TCI31	0.8099
TCI32	0.8042
TCI33	0.7915

Average VCI and TCI ($\overline{vci}, \overline{tci}$) for weeks with significant Pearson's correlation coefficients at p-level ($p<0.001$) among dY and VH indices were used as a predictors of dY. Figure 6.32 shows dynamics of dY (corn), MEAN VCI (weeks 28 to 36) and MEAN TCI (weeks 30 to 33) for CRD 30, Kansas. This figure shows how closely VH indices dynamics follows dY dynamics. This is a strong indication that VH indices can be used as predictors in a forecast model for corn yield in CRD 30, Kansas.

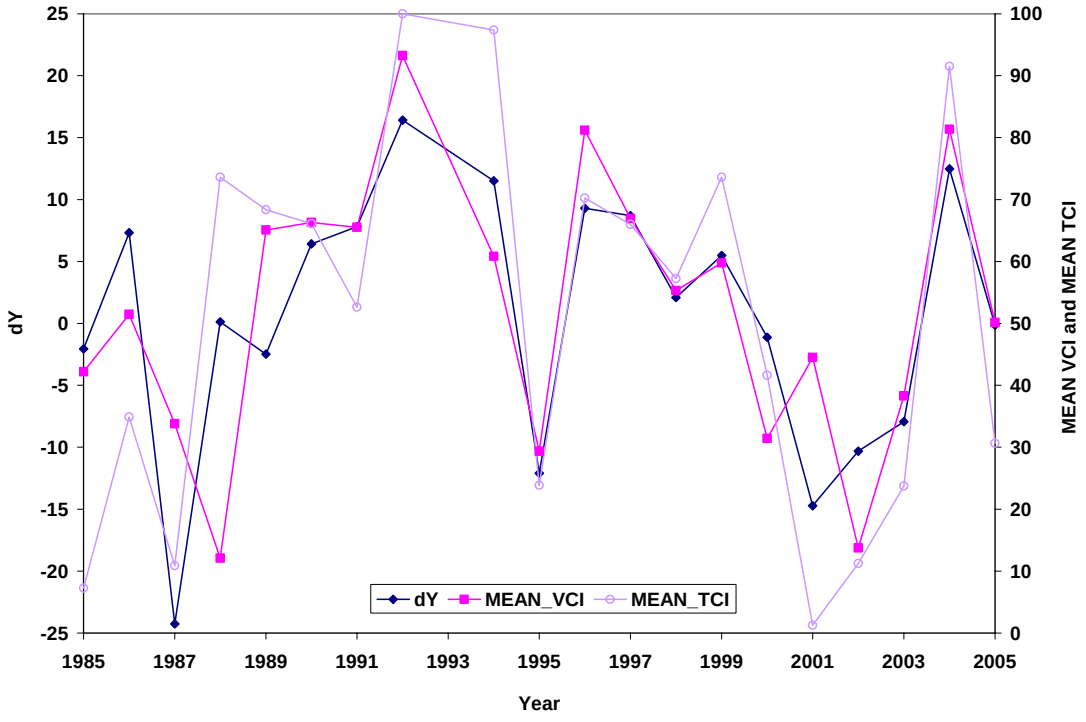

Figure 6.32: Dynamics of dY (corn), MEAN VCI (weeks 28 to 36) and MEAN TCI (weeks 30 to 33), CRD 30, Kansas

Table 6.42 shows correlation coefficients among dY (corn) with $\overline{vci}$ (weeks 28 to 36) and $\overline{tci}$ (weeks 30 to 33). These correlation coefficients are significantly different from 0 with p-value ($p<0.001$). For corn, $\overline{vci}$ would account for 53% (0.73^2) of the variation in dY if used separately as the only independent variable in the regression model. Besides, $\overline{tci}$ would account for 67% (0.82^2) of the variation in dY (corn) if used separately as the only independent variable in the regression model.

Table 6.42: Correlation matrix dY (corn), MEAN VCI ($\overline{vci}$) and MEAN TCI ($\overline{tci}$), CRD 30, Kansas

Pearson Correlation Coefficients	
	dY
MEAN VCI	0.7292
MEAN TCI	0.8245

6.3.2.2.2 Multiple Regression Results

The results of fitting the ordinary least squares (OLS) regression model approximated by Equation (6.1) to CRD 30 are shown in Table 6.43.

Table 6.43 shows that there is a strong relationship between dY and the independent variables. The coefficient of determination, R-Square, is 0.74 or 74 % of the sum of squares in dY (corn) can be associated with the variation in these independent variables. The test of the composite hypothesis that all regression coefficients are 0 is highly significant with F-value of 24.26 compared to F (0.01, 2,

17) = 6.11. The Root MSE is 5.5733 with 17 degrees of freedom is an unbiased estimate of σ.

Table 6.43: Results of the regression of dY (corn) on the two independent variables MEAN VCI and MEAN TCI, CRD 30, Kansas

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	2	1507.086	753.5429	24.2600	<.0001
Error	17	528.0401	31.0612	–	–
Corrected Total	19	2035.126	–	–	–

R-Square	*Root MSE*
0.7405	5.5733

Parameter	*Estimate*	*Standard Error*	*t Value*	*Pr > |t|*
Intercept	-17.6911	3.3042	-5.3542	0.0001
MEAN VCI	0.1559	0.0781	1.9959	0.0622
MEAN TCI	0.2034	0.0550	3.6988	0.0018

6.3.2.2.3 Model Validation

For model validation we regressed observed yield versus independently simulated yield and the corresponding statistics were generated (Table 6.44). The overall correlation coefficient 0.9088 is good. An R-Square value of 0.8258 shows that in most years, corn yield in CRD 30 can be modeled by variables considered in

model Equation (6.1). For corn CRD 30, the model forecast captured 83% of the variability in yield anomalies.

The average prediction bias is 0.0260 (systematic error). The variance of the prediction error is 36.0806 or the SEP is 6.0067 (non-systematic error). The standard error of the estimated mean bias is 1.3431 (Table 6.44). T-tests of the hypothesis that the bias is 0 gives t = 0.0193 which, with 19 degrees of freedom and $\alpha = 0.05$, is not significant. The MSEP is $= \frac{19(36.0806)}{20} + (0.0260)^2 = 34.2772$. The RMSEP is 5.8547 an approximate 3% error in prediction.

Table 6.44: Statistics of an independent testing for model Equation (6.1), corn, CRD 30, Kansas

Crop	***Correlation***	***R-Square***	***Bias***	***Variance***	***SEP***	***Std Error***	***RMSEP***
Corn	0.9088	0.8258	0.0260	36.0806	6.0067	1.3431	5.8547

Figure 6.33 displays observed versus independently simulated crop yield time series of corn 1985 through to 2005 for CRD 30, Kansas. For corn CRD 30 the model forecasts captured 83% of the variability in yield anomalies.

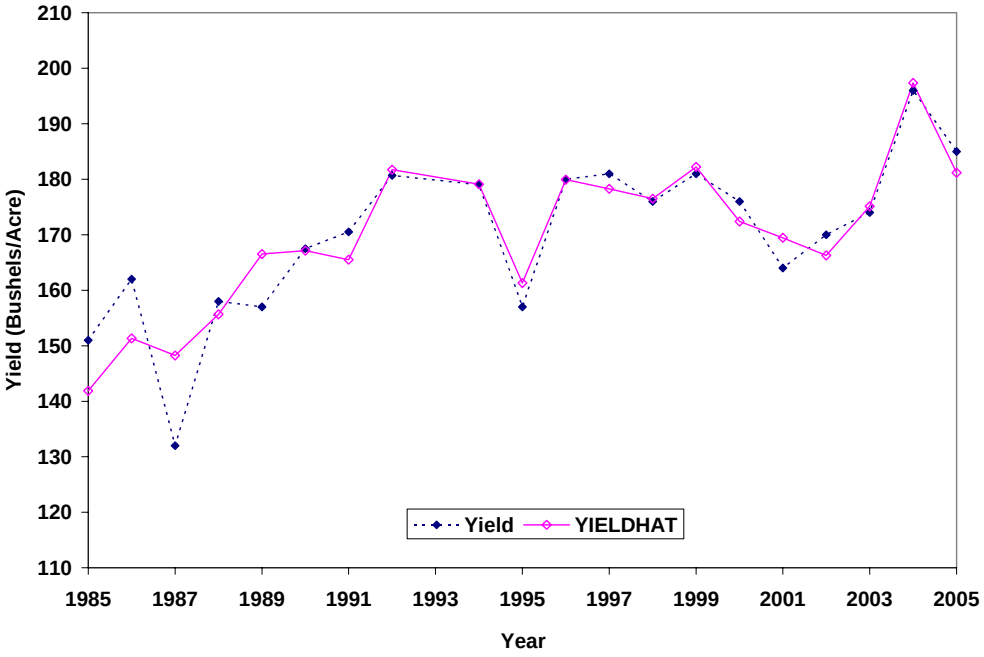

Figure 6.33: Observed yield (corn) versus independently simulated yield (yieldhat), CRD 30, Kansas

6.3.3 Haskell County

6.3.3.1 Crop yield time series

Figure 6.34 shows that in Haskell County corn yield increases due to technology improvement.

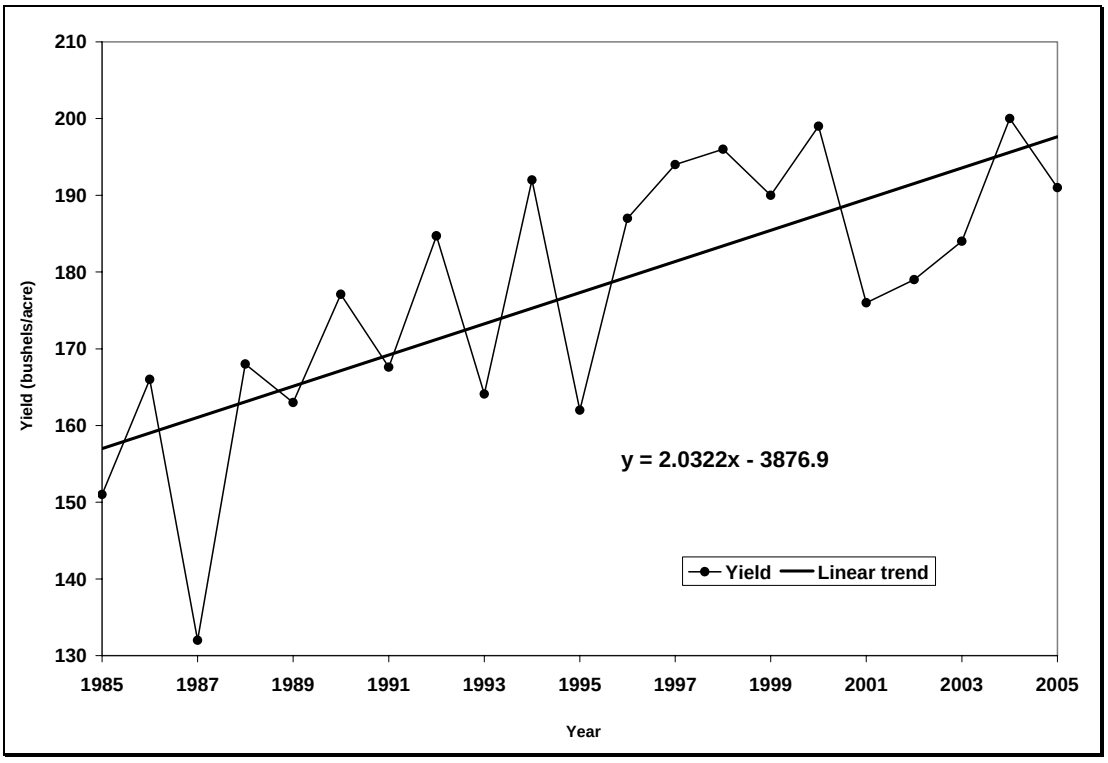

Figure 6.34: Corn yield time series, Haskell County, Kansas

6.3.3.2 Regression Analysis

A useful starting point in any multiple regression analysis is to compute correlations among all variables. This provides a first look at the simple linear relationships among them.

6.3.3.2.1 The Correlation Matrix

Table 6.45 contains Pearson's correlation coefficients among the dependent variable dY (corn) with VCI (weeks 32 to 34) and TCI (weeks 29 to 33) for Haskell County, Kansas. All VCI variables have high correlations with dY (0.72-0.74).

VCI33 has the highest correlation with dY. It would account for 55% (r^2 =0 .74) of the variation in dY if used separately as the only independent variable in the regression model. Besides, TCI variables have significant correlation with dY (0.45-0.54).

Table 6.45: Correlation matrix dY (corn), VCI (weeks 25 to 35) and TCI (weeks 31 to 32) Haskell County, Kansas

Pearson Correlation Coefficients	
	dY
VCI32	0.7313
VCI33	0.7398
VCI34	0.7228
TCI29	0.4771
TCI30	0.4507
TCI31	0.5280
TCI32	0.5450
TCI33	0.5143

Table 6.45 contains Pearson's correlation coefficients among the dependent variable dY (corn) with VCI (weeks 32 to 34) and TCI (weeks 29 to 33) for Haskell County, Kansas. VCI variables have high correlations with dY (0.72-0.74). VCI33 has the highest correlation with dY (0.78). TCI variables have significant correlation with dY (0.45-0.54) at $p<0.05$ level.

Average VCI and TCI ($\overline{vci}, \overline{tci}$) for weeks with significant Pearson's correlation coefficients at p-level ($p<0.001$) among dY and VH indices were used as a predictors of dY. Figure 6.35 shows dynamics of dY (corn), MEAN VCI (weeks 32 to

34) and MEAN TCI (weeks 29 to 33) for Haskell County, Kansas. This figure shows how closely VH indices dynamics follows dY dynamics. This is a strong indication that VH indices can be used as predictors in a forecast model for corn yield in Haskell County, Kansas.

Figure 6.35: Dynamics of dY (corn), MEAN VCI (weeks 32 to 34) and MEAN TCI (weeks 29 to 33) for Haskell county, Kansas

Table 6.46 shows correlation coefficients among dY (corn) with $\overline{vci}$ (weeks 32 to 34) and $\overline{tci}$ (weeks 29 to 33). These correlation coefficients are significantly different from 0 with p-value ($p<0.001$). For corn, $\overline{vci}$ would account for 55% (0.74^2) of the variation in dY if used separately as the only independent variable in the

regression model. Besides, $\overline{tci}$ would account for 28% (0.53^2) of the variation in dY (corn) if used separately as the only independent variable in the regression model.

Table 6.46: Correlation matrix dY (corn), MEAN VCI ($\overline{vci}$) and MEAN TCI ($\overline{tci}$) Haskell County, Kansas

Pearson Correlation Coefficients	
	dY
MEAN VCI	0.7372
MEAN TCI	0.5319

6.3.3.2.2 Multiple Regression Results

The results of fitting the ordinary least squares (OLS) regression model approximated by Equation (6.1) to Haskell County are shown in Table 6.47.

Table 6.47 shows that there is a strong relationship among dY and the independent variables. The coefficient of determination, R-Square, is 0.67 or 67% of the sum of squares in dY (corn) can be associated with the variation in these independent variables. The test of the composite hypothesis that all regression coefficients are 0 is highly significant with F-value of 17.23 compared to F (0.01, 2, 17) = 6.11. The Root MSE is 5.9965 with 17 degrees of freedom is an unbiased estimate of σ.

Table 6.47: Results of the regression of dY (corn) on the two independent variables MEAN VCI and MEAN TCI, Haskell County, Kansas

Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	2	1238.792	619.3959	17.2253	<.0001
Error	17	611.2958	35.9586	–	–
Corrected Total	19	1850.088	–	–	–

R-Square	*Root MSE*
0.6696	5.9965

Parameter	*Estimate*	*Standard Error*	*t Value*	*Pr > |t|*
Intercept	77.3883	4.1336	18.7216	0.0000
MEAN VCI	0.2451	0.0549	4.4601	0.0003
MEAN TCI	0.1397	0.0549	2.5466	0.0209

6.3.3.2.3 Model Validation

For model validation we regressed observed yield versus independently simulated yield and the corresponding statistics were generated (Table 6.48). The overall correlation coefficient 0.7219 is good. An R-Square value of 0.5211 shows that in most years, corn yield in Haskell County can be modeled by variables considered in model Equation (6.1). For corn Haskell County, the model forecast captured 52% of the variability in yield anomalies.

The average prediction bias is 0.1363 (systematic error). The variance of the prediction error is 150.3631 or the SEP is 12.2623 (non-systematic error). The

standard error of the estimated mean bias is 2.7413 (Table 6.48). T-tests of the hypothesis that the bias is 0 gives t = 0.0497 which, with 19 degrees of freedom and α = 0.05, is not significant. The MSEP is $= \frac{19(150.3631)}{20} + (0.1363)^2 = 142.8635$. The RMSEP is 11.9526 an approximate 7% error in prediction.

Table 6.48: Statistics of an independent testing for model Equation (6.1), corn, Haskell County, Kansas

Crop	***Correlation***	***R-Square***	***Bias***	***Variance***	***SEP***	***Std Error***	***RMSEP***
Corn	0.7219	0.5211	0.1363	150.3631	12.2623	2.7413	11.9526

Figure 6.36 displays observed versus independently simulated crop yield time series of corn 1985 through to 2005 for Haskell County, Kansas. For corn Haskell County the model forecasts captured 52% of the variability in yield anomalies.

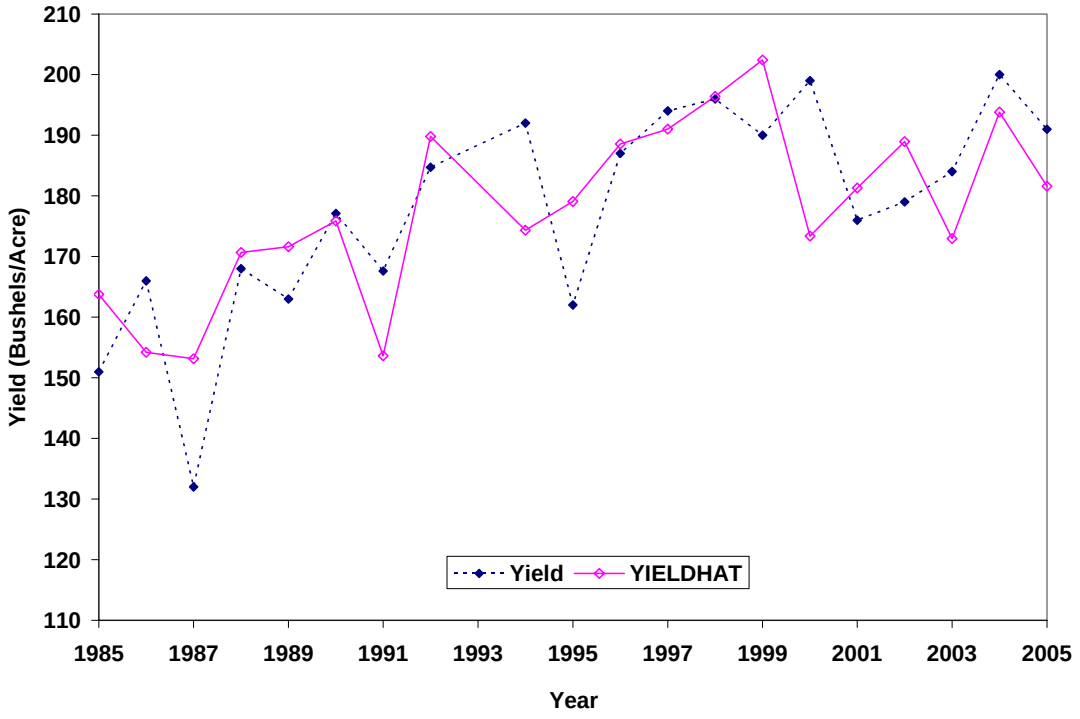

Figure 6.36: Observed yield (corn) versus independently simulated yield (yieldhat), Haskell County, Kansas

7. Concluding Remarks

7.1 Thesis Conclusions

Crop yield forecasts provide important information for government agencies, commodity traders, producers and other users in planning harvest, storage, transportation and marketing activities. Yield estimates of several crops in Kansas are currently based on field surveys and farmer interviews although official forecasts do not exist for many crops. This methodology is typically expensive and time-consuming.

In addition, over the years, large-area assessment of crop production has been based on a complex of in situ observations such as weather and climate data, soil physical properties, technological and other factors using statistical methods. This approach provided quite reasonable large-area predictions. However, one of the principal shortcomings of this approach is that in situ observations are not dense enough to adequately represent the volatile nature of spatial precipitation patterns. This is especially important in areas with limited water supplies where the majority of staple food crops are grown. An additional problem with the availability of in situ observations often appeared due to economic, political and social situations.

In order for crop assessment systems to be economical, more efficient methods for data collection and analysis are necessary. Since satellites cover large

areas in a short amount of time at a relatively low cost, remote sensing has the potential to provide reliable, timely and cost-effective predictions.

We developed remote sensing data-based models of statewide, CRDs and county yields for three major Kansas crops and tested their accuracy using cross-validation from 1985 to 2005. The forecasts were highly accurate, as judged by the percentage of yield variation explained by the forecast, the number of yields with correctly predicted direction of yield change, the number of yields with correctly predicted extreme values and the error on prediction. Predictions for most crops relied on satellite data measurements well before harvest time, allowing longer lead times than existing procedures.

Remote sensing is a promising alternative approach. The satellite data as part of a crop yield estimation system can be considered a natural component because of the ability of satellites to provide relatively economical, consistent and repeated coverage over large areas. These characteristics of satellites allow collecting data useful for timely estimation of crop conditions throughout an entire growing season covering either important agricultural production regions or remote regions where accurate information is normally unavailable. Thus, remote sensing is a potential alternative to previous data collection methods. This approach would be attractive because yields could potentially be forecast at lower cost, with greater accuracy and longer lead times.

To test the ability of remote sensing data to forecast crop yields prior to harvest, we studied the statistical relationships among Vegetation Health (VH) Indices (Vegetation Condition Index (VCI) and Temperature Condition Index (TCI)) computed for each week during a period of 21 years (1985-2005) from Advance Very High Resolution Radiometer (AVHRR) data and crop–yield records. We selected three crops (winter wheat, sorghum and corn) that are among the most valuable in Kansas, and obtained state crop production (bushels, b), area (acres, a) and yield (b a^{-1}) electronically from USDA/NASS database site (http://www.usda.gov/nass/) for the entire state of Kansas, each CRD and county, from 1980 through to 2005.

Several crops have exhibited significant positive yield trends since 1985 due to improvement in technology, management and cultivar changes, so we removed a linear trend from each crop to produce a time series of yield anomalies, or departures from expected yields. A positive anomaly indicates yields higher than expected based on time trends, and a negative anomaly indicates yields lower than expected.

The satellite data and yield data were then combined in linear regression models to test how well yield anomalies could be predicted before harvest based on weekly satellite measurements. An important step in building statistical models is to independently test model predictions, because tests using the same data that were used to calibrate the model tend to be very optimistic. The straightforward approach of reserving part of the data during model calibration, however, is problematic when the quantity of data is limited as in our case. An alternative approach, which we

employed in this research, is 'leave-one-out' cross-validation. In this approach, a single year was left out of the calibration step and subsequently compared to model predictions in that year. This comparison was done for each year, in this case resulting in 21 comparisons between model predictions and observations.

The results of cross–validation analysis suggest that winter wheat, sorghum and corn yields can be forecast with fairly high accuracy based on remote sensing data particularly VH indices. In general, the models correctly predicted the direction of yield anomalies for most of the years. That is, the models correctly predicted whether the yield would be above or below the trend.

This study shows that crop yield can be estimated from remote sensing data (VH indices) at approximately two months before the end of harvest, giving farmers and others in the industry the opportunity to use this information to make decisions. We have used the two AVHRR–based VH indices characterizing moisture (VCI) and thermal (TCI) conditions in Kansas as predictors of crop yield.

The models developed in this study are promising for forecasting winter wheat, sorghum and corn yields based on satellite measurements. Such forecasts could be of great relevance to commodities trade and management decisions. For example, farmers and other people related to the agribusiness industry could have used satellite data to correctly predict the low yield in 1989, 1996 and 2002 and adjust marketing practices for that reason, well before the forecast from USDA became available.

7.2 Future Research

The models developed in this study are promising for forecasting statewide, Crop Reporting District and county levels crop yields based on remote sensing data specifically Vegetation Health Indices (VCI and TCI) values. The potential value of such forecasts will depend on the acceptable types and magnitude of errors for particular applications. The potential to forecast crop yields also depends on crop type.

Even though field-based surveys are more accurate than other forecasts, it is important to consider the tradeoff between forecast accuracy, cost and timing. The low cost and long lead times that are possible with remote sensing data-based models would likely provide a useful complement to existing approaches for crops that are currently surveyed. For crops that are not currently forecast by USDA, these models present an opportunity to develop forecasts with low cost.

The current research was limited to few crops of the many grown in Kansas. Open questions are how well other crops can be modeled and whether other variables such as weather would improve forecast accuracies. These models can be further improved with growing historical and higher spatial resolution data. In addition, these models can be properly modified to add or drop variables and can be easily extended for other crops and regions.

7.3 Scholarly Publications

7.3.1 Peer Reviewed Journal Articles

SALAZAR, L., KOGAN, F. & ROYTMAN, L., 2007, Use of remote sensing data for estimation of winter wheat yield in the United States. *International Journal of Remote Sensing*, **In press**, pp. doi: 10.1080/01431160601050395.

SALAZAR, L., KOGAN, F. & ROYTMAN, L., 2007, Using vegetation health indices and partial least squares method for estimation of corn yield. *International Journal of Remote Sensing*, **In press**, pp. doi: 10.1080/01431160701271974.

7.3.2 Unpublished Work (submitted/in revision)

SALAZAR, L., KOGAN, F. & ROYTMAN, L., 2007, Early warning of drought related losses of agricultural production in U.S., submitted to *International Journal of Remote Sensing*.

7.3.3 Conferences Proceedings and Oral Presentations

2006 Salazar, L., F. Kogan and L. Roytman, Estimation of crop losses using remote sensing data and partial least squares, SPIE Europe Remote Sensing Conference September 17 to 20, 2007 that is going to be held in Florence, Italy.

2007 Salazar, L., Agricultural crops yield prediction using NOAA environmental satellite data, First NOAA and Kellogg's corporation Symposium, CCNY, New York, NY, March 27, 2007, Goal of the meeting discussions on possible areas of cooperation (oral presentation).

2006 Salazar, L., F. Kogan and L. Roytman, Estimation of crop losses using remote sensing data, Fourth Education and Science Forum of the National Oceanic and Atmospheric Administration – Educational Partnership Program (NOAA-EPP), Florida A & M University, Tallahassee, Florida, October 30 to November 1, 2006 (poster).

2006 Salazar, L., Using remote sensing data and partial least squares method for estimation of corn yield, Fourth Annual NOAA-CREST Symposium, Mayaguez, Puerto Rico, February 23 to 25, 2006 (oral presentation).

2005 Salazar, L., F. Kogan and L. Roytman, A comparison of support vector machines and PLSR on spectral data, The 2006 Joint Assembly, Baltimore, Maryland, May 23 to 26, 2006 (Poster).

2004 Salazar, L., F. Kogan and L. Roytman, Analysis of NOAA–AVHRR NDVI images for crop monitoring, Third Education and Science Forum of the National Oceanic and Atmospheric Administration – Educational Partnership Program (NOAA-EPP), City College of New York, NY, October 21 to 23, 2004 (poster).

8. Appendices

8.1 Appendix 1: The Collinearity Problem

The partial regression coefficient and partial sum of squares for any independent variable are, in general, dependent on which other independent variables are in the model. This dependence of the regression results for each variable on what other variables are in the model derives from the independent variables not being mutually orthogonal. Lack of orthogonality of the independent variables is to be expected in observational studies, those in which the researcher is restricted to making observations on nature as it exists. In such studies, the researcher cannot impose on a subject, or withhold from the subject, a procedure or treatment whose effects he desires to discover, we cannot assign subjects at random to different procedures. (Cochran 1983). On the other hand, controlled experiments are usually designed to avoid collinearity problems. The extreme case of nonorthogonality, where two or more independent variables are very nearly linearly dependent, creates severe problems in least squares regression. This is referred to as the collinearity problem.

Multiple linear regression with correlated explanatory variables is relevant to a broad range of problems in the physical, environmental, chemical, and engineering sciences. In this annex, we review a general theory for selecting principal components that yield estimates of regression coefficients with low mean squared error. The

results can be applied to any problem in which estimation of regression coefficients for correlated explanatory variables is of interest.

8.1.1 Overview

It is well known of the inability of classical least squares to provide reasonable point estimates when the matrix of regressor variables **X** is ill conditioned. Despite possessing a very desirable property of being minimum variance in the class of linear unbiased estimators, the ordinary least squares regression coefficients may have extremely large variances when there are near multicollinearities in **X** the matrix of explanatory variables, which is one form of ill conditioning. The variance of the ordinary least squares estimates becomes inflated when one or more eigenvalues of the correlation matrix of explanatory variables are close to zero. This results in an estimator that may have low probability of being close to the true value of the vector of regression coefficients β (Draper and Smith 1981, Myers 1986, Chatterjee et al. 2000).

There are many methodologies for combating this problem. Principal components regression (PCR) and partial least squares regression (PLSR) (Wold et al. 1987, Geladi et al. 1989, Kettaneh-Wold 1992, Trygg and Wold 1998, Wold 2001, Wold et al. 2001a, Wold et al. 2001b, Li et al. 2002) are two related methodologies that are often used to cope with the collinearity problem in multiple linear regression (MLR). Both involve selecting a subspace of the column space of **X** on which to

project the response vector **Y**. The two methodologies differ in the subspaces that they consider. PCR considers subspaces spanned by subsets of the principal components of **X**. PLSR considers subspaces spanned by subsets of the partial least squares (PLS) components which depend on both **X** and **Y**.

In general, users of PCR and PLSR regress **Y** against the first *k* components where *k* is determined by leave-one-out cross-validation. For PLSR, the first *k* PLS components are by design the ones most relevant to **Y**. The first *k* principal components (PCs), however, correspond to the largest *k* eigenvalues of the correlation matrix of **X** and are constructed independently of **Y**. Limiting the analysis to PCs with the largest eigenvalues helps to control variance inflation but can introduce high bias by discarding PCs with small eigenvalues that may be most closely associated with **Y** (Jolliffe 1972, Jolliffe 1973, Jolliffe 1982, Hadi and Ling 1998, Jolliffe and ebrary Inc. 2002).

In this thesis, we present methods for choosing a subset of components that attempt to minimize the mean squared error (MSE) of the estimator of β. Given a particular set of basis vectors for the column space of **X,** we derive the subset of the basis vectors that leads to an estimator of β with lowest MSE. This optimal subspace depends on the unknown error variance σ^2 and the unknown regression coefficients. We use cross-validation to select the optimal subspace and thus approximate the optimal vector of regression coefficients.

Here we analyze the collinearity problem and a description is given of how principal components regression (PCR) works in practice.

8.1.2 Notation and Conventions

Throughout the remainder of this annex, we will follow the notational conventions listed below. Greek letters are used to denote the parameters of a model while Roman letters are used to denote the estimates of the corresponding parameter.

Scalars are denoted by lower case, italic Roman and Greek letters.

a, b, c ….and α, β, γ

Vectors are always defined as column vectors and are denoted by lower case, bold Roman and Greek letters.

a, b, c… and α, β, γ…

A vector with p elements will be of dimension ($p \times 1$) with its i_{th} element given by the corresponding lower-case italic letter with an appropriate subscript, e.g. the i_{th} element of vector **v** is $\mathbf{v}_i$. The vector **1** denotes a vector whose elements are all ones.

Matrices are represented by upper case, bold Roman and Greek type.

A, B, C… and Ξ, Γ…

Their elements are represented by the corresponding lower-case, italic letter with subscripts identifying the row and column, respectively. For example, the elements of matrix **X** are given by x_{ij}. Sometimes, matrices will be written as **A** ($n \times$

p) to emphasize that the matrix **A** has n rows and p columns. It is also convenient to define matrices as a series of column vectors. **A** ($n \times p$) is also defined as $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p]$ where the $\mathbf{a}_i$, i = 1, 2, …, p are the ($n \times 1$) columns of **A**. If we want to consider the matrix formed by the first $m < p$ columns of **A**, we will write it as $\mathbf{A}_m$, where $\mathbf{A}_m$ ($n \times m$) = $[\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m]$. The identity matrix of order p is written as $\mathbf{I}_p$ and is defined as a square ($p \times p$) matrix that has 1's on the main diagonal and 0's elsewhere.

The columns of the ($n \times p$) matrix **X** might be measurements of p variables taken at n different samples. These variables will be denoted by the corresponding lower-case italic letter with a subscript indicating the column number, e.g., x_1, x_2, …, x_p.

8.1.3 Review of Methodologies

In this section, we will not give a detailed mathematical, statistical and geometrical interpretation of each methodology since the interested reader can find it in the references. Instead, we will start giving a review of ordinary least squares regression method. The shortcomings of this method are outlined. Then PCR is reviewed and its shortcomings analyzed. Finally, PLS method will be introduced as a practical solution to the collinearity problem.

8.1.4 The Multiple Regression Model

The least squares regression model is given by

$$\mathbf{y} = \beta_0 \mathbf{1} + \mathbf{X}\beta + \boldsymbol{\varepsilon}$$

where **y** is the ($n \times 1$) vector containing ground data (crop yield), **1** is an ($n \times 1$) vector of ones; $\mathbf{X} = [\mathbf{X}_1\ \mathbf{X}_2\ \ldots\ \mathbf{X}_p]$ is a ($n \times p$) full–rank matrix of non-stochastic regressor variables (VH indices) which is standardized so that $\mathbf{X}_j^{'}\mathbf{1} = 0$ and $\mathbf{X}_j^{'}\ \mathbf{X}_j = 1$ for j = 1,2,…,*p*, β_0 is an unknown constant, β is the ($p \times 1$) vector containing the regression coefficients or parameters to be estimated and $\boldsymbol{\varepsilon}$ is the ($n \times 1$) vector of unobservable random-error variables which are independent and identically distributed with mean zero and variance σ^2_i. In addition, for inference, testing hypothesis and calculating confidence intervals, it is assumed that $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\ \mathbf{I})$. The least squares solution for β is (assuming **X** is of full column range) given by $\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'y}$ and the variance-covariance matrix of the estimated regression coefficients is $\text{Var}(\mathbf{b}) = \sigma^2\ (\mathbf{X}^{'}\mathbf{X})^{-1}$. If collinearity exits it can degrade the precision of the results and make the **b** coefficients unstable (Draper and Smith 1981, Myers 1986).

8.1.5 The Problem

Singularity of **X** results when some linear function of the columns of **X** is exactly equal to the zero vector. Such cases become obvious when the least squares analysis is attempted because the unique $(\mathbf{X'X})^{-1}$ does not exist. A more critical

situation arises when the matrix is only close to being singular, meaning that a linear function of the vectors is nearly zero. It is caused because of redundant independent variables-the same information expressed in different forms. Interdependent variables that are closely linked in the system being studied can cause near-singularities in **X**. A unique solution to the normal equations exists in these nearly singular cases but the solution is very unstable (Draper and Smith 1981, Jolliffe 1986, Myers 1986, Chatterjee et al. 2000). Small changes (random noise) in the variables **y** or **X** can cause drastic changes in the estimates of the regression coefficients. The variances of the regression coefficients, for the independent variables involved in the near–singularity, become very large. Variables involved in the near-singularity can serve as surrogates for each other so that different combinations of the independent variables can be used to give nearly the same value of **y** (Draper and Smith 1981). Therefore, using ordinary regression procedures under high levels of correlation among predictor variables affects several characteristics of the model such as magnitude, sign, standard error of the estimated coefficients and the coefficient of determination. The sign indicates the direction of the relationship, the magnitude indicates the degree of influence, and the significance indicates whether the influence was due to chance or not. These difficulties that result from **X** being nearly singular are referred as the collinearity problem.

The impact of collinearity on least squares is very serious if primary interest is in the regression coefficients or if the purpose is to identify important variables in the

process. The estimates of the regression coefficients can differ greatly from the parameters they are estimating, even to the point of having incorrect sign. The collinearity will allow important variables to be replaced in the model with minor variables that are involved in the near-singularity. Hence, the regression analysis provides little indication of the relative importance of the independent variables.

The use of the regression equation for prediction is not seriously affected by collinearity if the correlational structure observed in the sample persists in the prediction population and prediction is carefully restricted to the sample **X** space. However, prediction to a system, where the observed correlation structure is not maintained or for points outside the sample space, can be misleading. The sample **X**-space in the presence of near-collinearities becomes very narrow in certain dimensions so that it is easy to choose prediction points that fall outside the sample space and, at the same time, difficult to detect when this has been done. Points well within the limits of each independent variable may be far outside the sample space. To detect near collinearities in the data, certain clues are present: unreasonable values for regression coefficients, large standard errors, nonsignificant partial regression coefficients when the model provides a reasonable fit, and known important variables appearing as unimportant (or with an opposite sign from what the theory would suggest) in the regression results. High correlations between independent variables will identify near-collinearities involving two variables but may miss those involving

more than two variables. A more direct approach to detecting the presence of collinearity is with a singular value decomposition of **X** or an eigenanalysis of **X'X**.

The remedies for the collinearity problem depend on the objective of the model. If the objective is prediction, collinearity causes no serious problem within the sample **X**-space. When primary interest is in estimation of the regression coefficients, one of the biased regression methods may be useful. A better solution, when possible, is to obtain new data or additional data such that the sample **X**-space is expanded to remove the near-singularity. It is not likely that this will be possible when the near-singularity is the result of internal constraints of the system being studied. When the primary interest of the research is to identify the important variables in a system or to model the system, the regression results in the presence of severe collinearity will not be very helpful and can be misleading. It is more productive for this purpose to concentrate on understanding the correlational structure of the variables and how the dependent variable fits into this structure. Principal component analysis and principal component regression can be helpful in understanding this structure.

*8.1.6 Understanding the Structure of the **X**-space*

The correlation matrix of the standardized independent variables $\mathbf{X}'\mathbf{X}$ is used for understanding the structure of the **X**-space. The off-diagonal elements of this matrix are the cosines of the angles between the corresponding standardized vectors in

X-space. Values near 1.0 or -1.0 indicate nearly collinear vectors; values near zero indicate nearly orthogonal vectors.

In our data, there are high correlations among the independent variables. For example, the correlation coefficient between VCI for week 20 and VCI for week 21 is 0.9839 as seen in Table 8.1.

Table 8.1: Correlation matrix among dY (WW) and VCI (week 17-23) for CRD 50, Kansas, U.S.

	dY	***VCI17***	***VCI18***	***VCI19***	***VCI20***	***VCI21***	***VCI22***	***VCI23***
dY	1.00000	0.83322	0.84091	0.84398	0.84858	0.84569	0.83353	0.82327
VCI17	0.83322	1.00000	0.98924	0.95386	0.89391	0.83018	0.76140	0.70926
VCI18	0.84091	0.98924	1.00000	0.98667	0.93071	0.88345	0.82662	0.77647
VCI19	0.84398	0.95386	0.98667	1.00000	0.95199	0.92808	0.88988	0.84335
VCI20	0.84858	0.89391	0.93071	0.95199	1.00000	0.98391	0.93709	0.87701
VCI21	0.84569	0.83018	0.88345	0.92808	0.98391	1.00000	0.98289	0.93699
VCI22	0.83353	0.76140	0.82662	0.88988	0.93709	0.98289	1.00000	0.97781
VCI23	0.82327	0.70926	0.77647	0.84335	0.87701	0.93699	0.97781	1.00000

The difficulties encountered when using remote sensing data are not only confined to the mere handling of the large amounts of data, but also, and perhaps more important, to the analysis and interpretation of the data. Multispectral data possess a high degree of correlation due to natural spectral correlation, topographic slope and aspect, and overlap of spectral sensitivities between adjacent spectral bands (Schowengerdt 1983). Hence, it is necessary to utilize techniques that can help in extracting significant information and eliminate redundancy.

Correlations will reveal linear dependencies involving two variables, but they will not reveal linear dependencies involving several variables. Near linear dependencies, involving any number of variables, are detected with a singular value decomposition of the matrix of independent variables, or with an eigenanalysis of the correlation matrix. The independent variables are scaled so that the vectors are of equal length and centered to remove collinearities with the intercept.

Table 8.2: Eigenvalues of the correlation matrix for the seven independent variables (VCI, weeks 17-23), CRD 50, Kansas

	Eigenvalue	***Difference***	***Proportion***	***Cumulative***
1	6.39092765	5.88683691	0.9130	0.9130
2	0.50409074	0.42771912	0.0720	0.9850
3	0.07637162	0.05502020	0.0109	0.9959
4	0.02135142	0.01471412	0.0031	0.9990
5	0.00663730	0.00628283	0.0009	0.9999
6	0.00035446	0.00008764	0.0001	1.0000
7	0.00026682	–	0.0000	1.0000

The eigenvectors of $\mathbf{X}'\mathbf{X}$ that correspond to the smaller eigenvalues identify the linear functions of the **X**s that show least dispersion and that cause the collinearity problem. The result of the eigenanalysis of the correlation matrix for CRD 50 data are shown in Table 8.2 and 8.3. The first part of Table 8.2 shows the eigenvalues of the correlation matrix of the seven independent variables VCI (Vegetation Condition Index, weeks 17-23) for CRD 50. From the "Eigenvalue" column, it is clear that the first principal component has a very large variance (6.39), the second has much

smaller variance (0.51), and the others have negligible variances. The first component accounts for 91% of the total variation. The cumulative proportions printed in the "Cumulative" column indicate that 98.50% of the total variation in the seven variables is explained by only two components. The eigenvalues reflect an extreme collinearity problem, with the condition number being $(6.390927/0.0002668)^{1/2}=154.77$

Table 8.3: Eigenvectors for each of the principal components for the seven independent variables (VCI, weeks 17-23), CRD 50, Kansas

	Prin1	***Prin2***	***Prin3***	***Prin4***	***Prin5***	***Prin6***	***Prin7***
VCI17	0.363085	-.542199	0.166207	0.548184	0.329027	0.253174	0.271485
VCI18	0.378227	-.406123	0.165804	-.104324	-.096032	-.606898	-.525439
VCI19	0.387758	-.221802	0.184769	-.731118	-.173639	0.331356	0.303030
VCI20	0.389129	0.008743	-.626240	0.195810	-.459916	0.352703	-.286466
VCI21	0.387356	0.232046	-.422137	-.015523	0.142980	-.533348	0.559255
VCI22	0.377201	0.415745	0.003543	-.150112	0.675088	0.220784	-.397290
VCI23	0.361954	0.514901	0.583401	0.304895	-.405782	-.013249	0.075591

Table 8.3 shows the eigenvectors for each of the PCs. The coefficients of the first PC show a positive relationship with all variables, with somewhat larger contributions from VCI20 (0.3891), VCI19 (0.33878) and VCI21 (0.3874). As expected, these components are in the middle of the critical period of winter wheat. The second component is dominated by VCI23 (0.5149). The final yield component, kernel weight, is determined during maturation that for WW occurs in week 23.

8.1.7 Principal Components

8.1.7.1 Principal Component Analysis

Consider a full-rank matrix of nonstochastic variables $\mathbf{X} = [\mathbf{X}_1\ \mathbf{X}_2\ \ldots\ \mathbf{X}_p]$ which is standardized so that $\mathbf{X}_j^{'}\ \mathbf{1} = 0$ and $\mathbf{X}_j^{'}\ \mathbf{X}_j = 1$ for j = 1,2,…,p, in which many of the columns are linear or near linear dependent. The total information or variation contained in **X** is proportional to tr (**X'X**), the trace of the matrix **X'X.** The idea in principal component analysis (PCA) is to construct successive linear combinations of the independent variables so that each one accounts for as much of the total variation as possible subject to the constraint that it be orthogonal to the linear combinations already extracted (Jolliffe and ebrary Inc. 2002).

Let us write any linear combination of the remote sensing data as **z** = **Xv**, where **v** is the *p*-element vector of weights. Then the variance of **z** can be written as **var(z) = var(Xv) = v'X'Xv** and we would like to find the vector of weights, **v**, that will maximize this variance. Since it is possible to achieve a maximum for infinite **v**, a constraint is imposed on the norm of the vector of weights. The simplest procedure usually applied is to specify a unit-norm constraint on **v**, hence, **v'v** = 1. It can be shown that this problem reduces to an eigenvalue problem yielding $\mathbf{X'X}\ \mathbf{v} = \lambda\ \mathbf{v}$ for which it is clear that the solution, which we will denote by the vector $\mathbf{v_i}$, is given by an eigenvector of **X'X**, that is, $\mathbf{v_1}$ satisfies the eigenvalue/eigenvector equation $\mathbf{X'X}\ \mathbf{v_1} = \lambda_1\mathbf{v}_1$ in which λ_1 is the eigenvalue corresponding to the eigenvector $\mathbf{v_1}$ **(**Jolliffe and

ebrary Inc. 2002). The question now is which eigenvector accounts for the maximal variation. Pre-multiplying the above equation by $\mathbf{v'_1}$ and application of the normalization constraint yields $\mathbf{v'_1X'Xv_1} = \lambda_1\mathbf{v'_1v_1} = \lambda_1$.

Since the variance is set to be as large as possible, λ_1 must be the largest eigenvalue of $\mathbf{X'X}$ and $\mathbf{v_1}$ is the associated eigenvector. Thus, the linear combination of $\mathbf{X}$ which has the largest variance is $\mathbf{z_1} = \mathbf{Xv_1}$ and is the first or largest principal component of $\mathbf{X}$, and its variance is λ_1. Successive principal components (PCs) are derived in a similar manner with the additional constraint that they be orthogonal to all previous PCs (Jolliffe and ebrary Inc. 2002).

The i_{th} PC derived in this way is $\mathbf{z}_i = \mathbf{Xv}_i$, and its variance, $\mathbf{v'}_i\mathbf{X'Xv}_i$ or $\mathbf{z'}_i\mathbf{z}_i$ equal to λ_i, the i_{th} largest eigenvalue of $\mathbf{X'X}$. The weighting vector $\mathbf{v}_i$, is the unit-norm eigenvector associated with λ_i. In the literature on PCA, $\mathbf{v}_i$ is known as a vector of weights or loadings. The loadings yield information about the interrelationships among the variables. The PCs, on the other hand, carry information about the interrelationships among the *n* observations. The elements of the i_{th} linear combination $\mathbf{z}_i$ are known as scores and are grouped in a matrix. If the first $m < p$ PCs are extracted, then the $(n \times m)$ score matrix, $\mathbf{Z}_m$, whose columns are the first *m* PCs, will be given by $\mathbf{Z}_m = \mathbf{X}\,\mathbf{V}_m$ or $\mathbf{[z_1, z_2, \ldots, z_m]} = \mathbf{X[v_1, v_2, \ldots, v_m]}$ where the columns of the $(p \times m)$ matrix $\mathbf{V}_m$ are the first *m* eigenvectors of $\mathbf{X'X}$. Because principal components are orthogonal, we can write that $\mathbf{Z'Z} = \mathbf{V'}_m\mathbf{X'XV}_m = \mathbf{L}_m$ where $\mathbf{L}_m$ is an $(m \times m)$ diagonal matrix whose elements, λ_i, are the first *m* eigenvalues of $\mathbf{X'X}$. As

many PCs can be extracted as the rank, p, of **X**. The total variation of **X**, tr(**X'X**), can be written as tr(**X'X**) = $\sum_{i=1}^{p} \lambda_i$

If only $m < p$ PCs are extracted, then the total variance that they account for is simply $\sum_{i=1}^{m} \lambda_i$, the sum of the first m eigenvalues of **X'X**. The fraction of the total variance accounted for by the first m PCs is $\alpha_m = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$.

If the last eigenvalues are very small, α_m will be very close to one when only m PCs are retained. This makes it possible to summarize the information in **X** using fewer PCs than the original number of **X** variables.

When all PCs have been extracted we can write for $m = p$ **Z = X V.** Post multiplying by **V'** and noting that **V** is an orthonormal matrix yields **X = ZV'.** In other words, the data matrix can be reconstructed exactly if all p PCs are retained. If only the first m PCs are retained, however, **X** cannot be reproduced exactly. Instead, the product $\mathbf{Z_m V'_m} = \mathbf{z_1 v'_1} + \mathbf{z_2 v'_2} + \ldots + \mathbf{z_m v'_m} \equiv \hat{X}^{m}_{PCA}$ provides the best rank m approximation of **X** in a least squares sense.

In general, the rank m approximation of **X** is an m–dimensional hyper-plane in p-dimensions. For example, consider a standardized data matrix **X**, consisting of only two ($n \times 2$) but highly collinear variables. Each row of this matrix represents a point

in two-dimensional space defined by the orthogonal axes corresponding to the variables $\mathbf{x}_1$ and $\mathbf{x}_2$ and the *n* points plotted together form a two-dimensional cloud. If the elements of $\hat{X}^{1}_{PCA}$ (a rank one approximation) are plotted, they are all positioned along a straight line. The first eigenvector, $\mathbf{v_1}$, lies in the direction of the greatest spread or variation, while the second eigenvector, $\mathbf{v_2}$, is orthogonal to the first and lies in the next greatest direction of spread. The straight line formed by the first eigenvector is the best fitting one in the sense of minimizing the sum of the squares of the perpendicular distances from the data points to any line. The data used in this example has most of the variation in one direction only, that of $\mathbf{v_1}$. The variation in the other direction, described by $\mathbf{v_2}$ is minor compared to the first one and the according eigenvalue, λ_2, is much smaller than λ_1. By retaining only the first PC (rank 1 approximation), most of the variation in the data is accounted for and the direction defined by the second PC is regarded redundant.

In the case of three **x**-variables, the first PC also defines a straight line while the first two PC's define a two-dimensional plane in the three-dimensional space spanned by the original variables $\mathbf{x_1}$, $\mathbf{x_2}$ and $\mathbf{x_3}$. Higher dimensions cannot be illustrated by a graphical example but the methodology remains the same. The basic idea in PCA is that the directions associated with small eigenvalues are unimportant and that the relevant information (most of the variation) is concentrated in the first few PCs.

8.1.7.2 Principal Components Regression

The least squares estimators of the regression coefficients are the best linear unbiased estimators. That is, of all possible estimators that are both linear functions of the data and unbiased for the parameters being estimated, the least squares estimators have the smallest variance. In the presence of collinearity, however, this minimum variance may be unacceptably large. Relaxing the least squares condition that estimators be unbiased opens for consideration a much larger set of possible estimators from which one with better properties in the presence of collinearity might be found. Biased regression refers to this class of regression methods in which unbiasedness is no longer required. Such methods have been suggested as possible solutions for the collinearity problem (Draper and Smith 1981, Myers 1986, Jong and Kotz 1999, Chatterjee et al. 2000, Kim et al. 2005). Biased regression has been motivated for the possibility of obtaining estimators that are closer, on average, to the parameter being estimated than are the least squares estimators.

Principal components regression is a technique to handle the problem of collinearity and produces stable and meaningful estimates for regression coefficients. Principal component regression approaches the collinearity problem by eliminating those dimensions of the **X**-space that are causing the collinearity problem. This is similar to dropping an independent variable from the model when there is insufficient dispersion in that variable to contribute meaningful information on **y**. However, in

principal component regression the dimension dropped is defined by a linear combination of the independent variables rather than by a single independent variable.

As described mathematically in the previous sections, PCA is in fact a method for transforming the coordinate system. After PCA is carried out, the original data contained in $\mathbf{X}$ and described by p variables is represented by $\mathbf{Z}$, or by $\mathbf{Z_m}$ if one is satisfied with the rank m approximation. By doing so, the number of variables is also reduced from p to m without a significant loss of information. The matrix $\mathbf{V_m}$ is a transformation matrix needed to toggle between the two coordinate systems. In addition, the m new variables are orthogonal meaning that the columns of $\mathbf{Z_m}$ are not collinear - not (near) linear dependent.

Principal component regression (PCR) consists of replacing the original regressors, or independent variables by a subset of PCs, usually those that account for most of the variation. Consider the conventional multiple linear regression model which is expressed as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, in which $\mathbf{y}$ is an $(n \times 1)$ vector of observations on the response variable, $\mathbf{X}$ is the $(n \times p)$ matrix of observations on the predictor variables, $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of parameters, $\boldsymbol{\varepsilon}$ an $(n \times 1)$ vector of errors independent and identically distributed with mean zero and variance $\sigma^2\mathbf{I}$. $\mathbf{X}$ and $\mathbf{y}$ are mean-centered to avoid collinearity with the intercept. In the case of PCR the above equation is rewritten as $\mathbf{y} = \mathbf{Z_m}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$.

Here, $\mathbf{Z}_\mathrm{m}$ is an $(n \times m)$ matrix whose columns are the first m PCs ($\mathbf{z}_\mathrm{i}$) and $\boldsymbol{\alpha}$ is an $(m \times 1)$ vector of unknown parameters. Using least squares to estimate the

elements of α leads to $\mathbf{a} = (\mathbf{Z'_m Z_m})^{-1}\mathbf{Z'_m y} = \mathbf{L^{-1}Z'_m y}$ in which $\mathbf{L}_m$ is a diagonal matrix consisting of the first *m* eigenvalues of **X'X**.

The vector of fitted values, $\hat{y}$, is given by $\mathbf{Z_m a}$. We can also express $\hat{y}$ as a linear combination of **X** since $\mathbf{Z_m = XV_m}$, where the columns of the matrix $\mathbf{V_m}(p \times m)$ consists of the first *m* eigenvectors of **X'X**. Thus, we can write

$\hat{Y}^{m}_{PCR} = \mathbf{XV_m L_m^{-1} V_m' X' y}$. If we compare this with the multiple regression equation, the PCR estimate of β is $b^{m}_{PCR} = \mathbf{V_m L_m^{-1} V_m' X' y}$.

However, the PCR method has a significant drawback. The PCR as described above is the standard formulation for which the PCs defining $\mathbf{Z_m}$ are in accordance with the size of their corresponding eigenvalues alone and not on their predictive value. This procedure neglects low variance components that may have predictive value on **y**. The PCs solely describe the largest variance components of the satellite data, **X**, and it is not necessarily true that this is the most informative variance to model the crop yield data, **y** (Hadi and Ling 1998, Chatterjee et al. 2000, Jolliffe and ebrary Inc. 2002). Therefore, several authors suggest retaining, in the final PCR model, only those PCs that show a good correlation with the **y**-variable (Jolliffe 1972, Jolliffe 1973, Jolliffe 1982, Jolliffe 1986, Preisendorfer and Mobley 1988, Jackson 1991, Hadi and Ling 1998).

9. References

ANDERSSEN, E., DYRSTAD, K., WESTAD, F. & MARTENS, H., 2006, Reducing over-optimism in variable selection by cross-model validation. ***Chemometrics and Intelligent Laboratory Systems***, **84**(1-2), pp. 69-74.

ASRAR, G. & MURPHY, R. E., 1989, Remote Sensing Science Program: A research program Within Land Processes Branch of NASA.

ASRAR, G., WICKLAND, D. E., BALTUCK, M., RUZEK, M. J. & MURPHY, R. E., 1988, Remote Sensing Of Land Processes: Programs Of Study At NASA Headquarters.

BADHWAR, G. D. & HENDERSON, K. E., 1981, Estimating development stages of corn from spectral data—an initial model. Agronomy Journal, **73**, pp. 748–755.

BARET, F. & GUYOT, G., 1991, Potentials and limits of vegetation indices for LAI and APAR assessment. Remote Sensing of Environment, **35**, pp. 161-173.

BARET, F., GUYOT, G. & MAJOR, D. J., 1989, TSAVI: A Vegetation Index Which Minimizes Soil Brightness Effects On LAI And APAR Estimation.

BROCKWELL, P. J. & DAVIS, R. A., 2000, Introduction to time series and forecasting (New York: Springer).

BROWN, M. E., PINZON, J. E., DIDAN, K., MORISETTE, J. T. & TUCKER, C. J., 2006, Evaluation of the consistency of long-term NDVI time series derived from AVHRR,SPOT-vegetation, SeaWiFS, MODIS, and Landsat ETM+ sensors. IEEE Transactions on Geoscience and Remote Sensing, **44**(7), pp. 1787-1793.

BUTLER, N. A. & DENHAM, M. C., 2000, The Peculiar shrinkage properties of partial least squares regression. Journal of the Royal Statistical Society: Series B, **62**(3), pp. 585-593.

CAMPBELL, N. A. & REECE, J. B., 2002, Biology (San Francisco: Benjamin Cummings).

CAMPBELL, N. A., SIMON, E. J. & REECE, J. B., 2004, Essential biology with physiology (San Francisco, CA: Pearson/Benjamin Cummings).

CARDER, K. L., REINERSMAN, P., CHEN, R. F., MULLER-KARGER, F., DAVIS, C. O. & HAMILTON, M., 1993, AVIRIS calibration and application in coastal oceanic environments. Remote Sensing of Environment, **44**(2-3), pp. 205-216.

CHATTERJEE, S., HADI, A. S. & PRICE, B., 2000, Regression analysis by example (New York: Wiley).

CHE, N. & PRICE, J. C., 1992, Survey of radiometric calibration results and methods for visible and near infrared channels of NOAA-7, -9, and -11 AVHRRs. Remote Sensing of Environment, **41**(1), pp. 19-27.

CHEN, X., VIERLING, L. & DEERING, D., 2005, A simple and effective radiometric correction method to improve landscape change detection across sensors and across time. Remote Sensing of Environment, **98**(1), pp. 63-79.

CHOUDHURY, B. J. & TUCKER, C. J., 1987, Satellite observed seasonal and inter-annual variation of vegetation over the Kalahari, The Great Victoria Desert, and The Great Sandy Desert: 1979-1984. Remote Sensing of Environment, **23**(2), pp. 233-241.

CIHLAR, J., LATIFOVIC, R., CHEN, J., TRISHCHENKO, A., DU, Y., FEDOSEJEVS, G. & GUINDON, B., 2004, Systematic corrections of AVHRR image composites for temporal studies. Remote Sensing of Environment, **89**(2), pp. 217-233.

CLEVERS, J. G. P. W., 1988, The derivation of a simplified reflectance model for the estimation of leaf area index. Remote Sensing of Environment, **35**, pp. 53-70.

COCHRAN, W. G., 1983, Planning and analysis of observational studies (New York: Wiley).

COLLINS, W., 1978, Remote sensing of crop type and maturity. Photogrammetric Engineering and Remote Sensing, **44**, pp. 43-55.

CRAIG, M., 2001, A resource sharing approach to crop identification and estimation, 2001 ASPRS/ACSM Annual Convention Technical Papers. Proceedings of the ASPRS 2001 Conference Bethesda, MD – USA.

CRIPPEN, R. E., 1990, Calculating the Vegetation Index Faster. Remote Sensing of Environment, **34**, pp. 71-73.

CSISZAR, I., GUTMAN, G., ROMANOV, P., LEROY, M. & HAUTECOEUR, O., 2001, Using ADEOS/POLDER data to reduce angular variability of NOAA/AVHRR reflectances. Remote Sensing of Environment, **76**(3), pp. 399-409.

CSORNAI, G., WIRNHARDT, C., SUBA, Z., NADOR, G., MARTINOVITCH, L., TIKASZ, L., LELKES, M., KOCSIS, A. & ZELEI, G., 2002, The operational crop monitoring and production forecast program (CROPMON) and others RS based applications. Geoinformation for European wide Integration. Proceeding of the 22nd Symposium of the European Association of Remote Sensing Laboratories. Prague, Czech Republic, Millpress.

CUOMO, V., LANFREDI, M., LASAPONARA, R., MACCHIATO, M. F. & SIMONIELLO, T., 2001, Detection of interannual variation of vegetation in middle and southern Italy during 1985-1999 with 1 km NOAA AVHRR. Journal of Geophysical Research, **106**(D16), pp. 17863-17876.

CURRAN, P. J., 1984, Principles of remote sensing (London ; New York: Longman).

DABROWSKA-ZIELINSKA, K., KOGAN, F., CIOLKOSZ, A., GRUSZCZYNSKA, M. & KOWALIK, W., 2002, Modelling of crop growth conditions and crop yield in Poland using AVHRR-based indices. International Journal of Remote Sensing, **23**(6), pp. 1109-1123.

DAS, D. K., MISHRA, K. K. & KALRA, N., 1993, Assessing growth and yield of wheat using remotely sensed canopy temperature and spectral indices. International Journal of Remote Sensing, **14**, pp. 3081-3092.

DEFRIES, R. S. & BELWARD, A. S., 2000, Global and regional land cover characterization from satellite data: an introduction to the Special Issue. International Journal of Remote Sensing, **21**(6), pp. 1083-1092.

DIDAN, K. & HUETE, A., 2004, Analysis of the global vegetation dynamic metrics using MODIS Vegetation Index and land cover products. Geoscience and Remote Sensing Symposium, 2004. IGARSS '04. Proceedings. IEEE International.

DINGSTAD, G. I., WESTAD, F. & NAES, T., 2004, Three case studies illustrating the properties of ordinary and partial least squares regression in different mixture models. Chemometrics and Intelligent Laboratory Systems, **71**(1), pp. 33-45.

DINGUIRARD, M. & SLATER, P. N., 1999, Calibration of Space-Multispectral Imaging Sensors: A Review. Remote Sensing of Environment, **68**(3), pp. 194-205.

DOMENIKIOTIS, C., SPILIOTOPOULOS, M., TSIROS, V. & DALEZIOS, N. R., 2004, Early cotton yield assessment by the use of the NOAA/AVHRR derived Vegetation Condition Index (VCI) in Greece. International Journal of Remote Sensing, **25**(14), pp. 2807-2819.

DORAISWAMY, P. D. & COOK, P. W., 1995, Spring wheat yield assessment using NOAA AVHRR data Canadian Journal of Remote Sensing, **21**, pp. 43– 51.

DORAISWAMY, P. D., MOULIN, S., COOK, P. W. & STERN, A., 2003, Crop yield assessment from remote sensing. Photogrammetric Engineering and Remote Sensing, **69**(6), pp. 665– 674.

DRAPER, N. R. & SMITH, H., 1981, Applied regression analysis (New York: Wiley).

ELDEIRY, A. & GARCIA, L., 2004, Spatial modeling using remote sensing, GIS, and field data to assess crop yield and soil salinity. Hydrology Days, **7**, pp. 55-66.

FAO, 2006, Crop production. Available online at: http://fao.org (Rome Italy: FAO).

FENG, D., CHEN, J. M., PLUMMER, S., MINGZHEN, C. & PISEK, J., 2006, Algorithm for global leaf area index retrieval using satellite imagery. IEEE Transactions on Geoscience and Remote Sensing, **44**(8), pp. 2219-2229.

FENSHOLT, R. & SANDHOLT, I., 2005, Evaluation of MODIS and NOAA AVHRR vegetation indices with in situ measurements in a semi-arid environment. International Journal of Remote Sensing, **26**(12), pp. 2561-2594.

FENSHOLT, R., SANDHOLT, I. & RASMUSSEN, M. S., 2002, Earth observation of vegetation status in a semi-arid environment: comparison of TERRA MODIS and NOAA AVHRR satellite data.

FERENCZ, C., BOGNAR, P., LICHTENBERGER, J., HAMAR, D., TARCSAI, G. & TIMAR, G., 2004, Crop yield estimation by satellite remote sensing. International Journal of Remote Sensing, **20**, pp. 4113–4149.

FROHN, R. C., 1998, Remote sensing for landscape ecology : new metric indicators for monitoring, modeling, and assessment of ecosystems (Boca Raton ; London: Lewis Publishers).

GARTHWAITE, P. H., 1994, An Interpretation of Partial Least Squares. Journal of the American Statistical Association, **89**(425), pp. 122-127.

GATES, D. M., KEEGAN, H. J., SCHLETER, J. C. & WEIDNER, V. R., 1965, Spectral properties of plants. ***Applied Optics***, **4**, pp. 11-20.

GELADI, P., ISAKSSON, H., LINDQVIST, L., WOLD, S. & ESBENSEN, K., 1989, Principal component analysis of multivariate images. Chemometrics and Intelligent Laboratory Systems, **5**(3), pp. 209-220.

GITELSON, A. A., KOGAN, F., ZAKARIN, E., SPIVAK, L. & LEBED, L., 1998, Using AVHRR data for quantitive estimation of vegetation conditions: Calibration and validation. Advances in Space Research, **22**(5), pp. 673-676.

GOUTIS, C., 1996, Partial least squares algorithm yields shrinkage estimators. The Annals of Statistics, **24**(2), pp. 816-824.

GROTEN, S. M. E., 1993, NDVI-crop monitoring and early yield assessment of Burkina Faso. International Journal of Remote Sensing, **14**, pp. 1495– 1515.

GUNST, R. F. & MASON, R. L., 1980, Regression analysis and its application: a data-oriented approach (New York, USA: Marcel Dekker, Inc.).

GUPTA, R. K., PRASAD, S., RAO, G. H. & NADHAM, T. S. V., 1993, District level wheat yield estimation using NOAA/AVHRR NDVI temporal profile. Advanced Space Research, **13**, pp. 253–256.

GUTMAN, G., 1999a, On the use of long-term global data of land reflectances and vegetation indices derived from the Advanced Very High Resolution Radiometer. Journal of Geophysical Research – Atmospheres, **104**, pp. 6241-6255.

GUTMAN, G. & IGNATOV, A., 1995, Global land monitoring from AVHRR: Potential and limitations. International Journal of Remote Sensing, **16**, pp. 2301-2309.

HADI, A. S. & LING, R. F., 1998, Some cautionary notes on the use of principal components regression. The American Statistician, **52**(1), pp. 15-19.

HAYAS, M. J. & DECKER, W. L., 1996, Using NOAA AVHRR data to estimate maize production in the United States Corn Belt. International Journal of Remote Sensing, **17**, pp. 3189-3200.

HEIDINGER, A. K., SULLIVAN, J. T. & RAO, C. R. N., 2003, Calibration of visible and near-infrared channels of the NOAA-12 AVHRR using time series of observations over deserts. International Journal of Remote Sensing, **24**(18), pp. 3635.

HELLAND, K., BERNTSEN, H. E., BORGEN, O. S. & MARTENS, H., 1992, Recursive algorithm for partial least squares regression. Chemometrics and Intelligent Laboratory Systems, **14**(1-3), pp. 129-137.

HILLEL, D., 2000, Salinity management for sustainable irrigation : integrating science, environment, and economics (Washington, DC: World Bank).

HOY, M., STEEN, K. & MARTENS, H., 1998, Review of partial least squares regression prediction error in Unscrambler. Chemometrics and Intelligent Laboratory Systems, **44**(1-2), pp. 123-133.

HUETE, A. R., 1988, A soil-adjusted vegetation index (SAVI). Remote Sensing of Environment, **25**(3), pp. 295-309.

HUETE, A. R. & LIU, H. Q., 1994, An error and sensitivity analysis of the atmospheric- and soil-correcting variants of the NDVI for the MODIS-EOS. IEEE Transactions on Geoscience and Remote Sensing, **32**(4), pp. 897-905.

ITA, 2002, Integrated crop area estimates. Final MARS/JRC Report.

JACKSON, J. E., 1991, A user's guide to principal components (New York: Wiley).

JACKSON, J. E., 2003, A user's guide to principal components (Hoboken, N.J.: Wiley).

JENSEN, J. R., 2000, Remote sensing of the environment: an earth resource perspective (New Jersey: Prentice Hall).

JIANG, Z., HUETE, A. R., CHEN, J., CHEN, Y., LI, J., YAN, G. & ZHANG, X., 2006, Analysis of NDVI and scaled difference vegetation index retrievals of vegetation fraction. Remote Sensing of Environment, **101**(3), pp. 366-378.

JOLLIFFE, I. T., 1972, Discarding Variables in a Principal Component Analysis. I: Artificial Data. Applied Statistics, **21**(2), pp. 160-173.

JOLLIFFE, I. T., 1973, Discarding Variables in a Principal Component Analysis. II: Real Data. Applied Statistics, **22**(1), pp. 21-31.

JOLLIFFE, I. T., 1982, A Note on the Use of Principal Components in Regression. Applied Statistics, **31**(3), pp. 300-303.

JOLLIFFE, I. T., 1986, Principal component analysis (New York: Springer-Verlag).

JOLLIFFE, I. T. & EBRARY INC., 2002, Principal component analysis (New York: Springer).

JONES, P. D., JOUZEL, J. & BRADLEY, R. S., 1996, Climatic variations and forcing mechanisms of the last 2000 years (Berlin ; London: Springer).

JONG, J.-C. & KOTZ, S., 1999, On a relation between principal components and regression analysis. The American Statistician, **53**(4), pp. 349-351.

JORDAN, C. F., 1969, Derivation of leaf area index from quality of light on the forest floor. Ecology, **50**, pp. 663-666.

JUSTICE, C., BELWARD, A., MORISETTE, J., LEWIS, P., PRIVETTE, J. & BARET, F., 2000, Developments in the 'validation' of satellite sensor products for the study of the land surface. International Journal of Remote Sensing, **21**(17), pp. 3383-3390.

JUSTICE, C. O., VERMOTE, E., TOWNSHEND, J. R. G., DEFRIES, R., ROY, D. P., HALL, D. K., SALOMONSON, V. V., PRIVETTE, J. L., RIGGS, G., STRAHLER, A., LUCHT, W., MYNENI, R. B., KNYAZIKHIN, Y., RUNNING, S. W., NEMANI, R. R., ZHENGMING, W., HUETE, A. R., VAN LEEUWEN, W., WOLFE, R. E., GIGLIO, L., MULLER, J., LEWIS, P. & BARNSLEY, M. J., 1998, The Moderate Resolution Imaging Spectroradiometer (MODIS): land remote sensing for global change research. IEEE Transactions on Geoscience and Remote Sensing, **36**(4), pp. 1228-1249.

KASTENS, D. L., 1998, Estimating wheat yields from time series analysis of remotely sensed data. United States Department of Agriculture small business innovative research phase I grant (Grant agreement number 99-33610-7495).

KASTENS, D. L., 2000, Forecasting pre-harvest winter wheat yields in the Great Plains using remotely sensed data. United States Department of Agriculture small business innovative research phase II grant (Grant agreement number 00-33610-9453).

KAUFMAN, Y. J., WALD, A. E., REMER, L. A., BO-CAI, G., RONG-RONG, L. & FLYNN, L., 1997, The MODIS 2.1μm channel-correlation with visible reflectance for use in remote sensing of aerosol. IEEE Transactions on Geoscience and Remote Sensing, **35**(5), pp. 1286-1298.

KAUFMANN, R. K. & ZHOU, L., 2000, Effect of orbital drift and sensor changes on the time series of AVHRR vegetation index data. IEEE Transactions on Geoscience and Remote Sensing, **38**(6), pp. 2584.

KETTANEH-WOLD, N., 1992, Analysis of mixture data with partial least squares. Chemometrics and Intelligent Laboratory Systems, **14**(1-3), pp. 57-69.

KIDWELL, K. B., 1997, Global Vegetation Index user's guide (Camp Springs MD: US Department of Commerce, NOAA, National Environmental Satellite Data

and Information Service, National Climatic Data Center, Satellite Data Services Division).

KIM, K., LEE, J.-M. & LEE, I.-B., 2005, A novel multivariate regression approach based on kernel partial least squares with orthogonal signal correction. Chemometrics and Intelligent Laboratory Systems, **79**(1-2), pp. 22-30.

KOGAN, F., 1997, Global drought watch from space. Bulletin American Meteorological Society, **78**(4), pp. 621-636.

KOGAN, F., 2002, World droughts in the new millennium from AVHRR-based Vegetation Health Indices. EOS, TRANSACTIONS AMERICAN GEOPHYSICAL UNION, **83**(48), pp. 557-564.

KOGAN, F., BANGJIE, Y., GUO, W., PEI, Z. & JIAO, X., 2005, Modelling corn production in China using AVHRR-based vegetation health indices. International Journal of Remote Sensing, **26**(11), pp. 2325-2336.

KOGAN, F., GITELSON, A., ZAKARIN, E., SPIVAK, L. & LEBED, V., 2003, AVHRR-based spectral vegetation indices for quantitative assessment of vegetation state and productivity: calibration and validation. Photogrammetry Engineering and Remote Sensing, **69**(8), pp. 899-906.

KOGAN, F. & SULLIVAN, J., 1993, Development of global drought-watch system using NOAA/AVHRR data. Advances in Space Research, **13**(5), pp. 219-222.

KOGAN, F. N., SULLIVAN, J. T. & BU CIREN, P., 1996, Testing post-launch calibration for the AVHRR sensor on world desert targets during 1985-1993. Advances in Space Research, **17**(1), pp. 47-50.

KOGAN, F. N. & ZHU, X., 2001, Evolution of long-term errors in NDVI time series: 1985-1999. Advances in Space Research, **28**(1), pp. 149-153.

KOSLOWSKY, D., 1997, Signal degradation of the AVHRR shortwave channels of NOAA-11 and NOAA-14 by daily monitoring of desert targets. Advances in Space Physics, **19**, pp. 355-1358.

KRIEGLER, F. J., MALILA, W. A., NALEPKA, R. F. & RICHARDSON, W., 1969, Preprocessing transformations and their effects on multispectral recognition.

Sixth International Symposium on Remote Sensing of Environment. University of Michigan, Ann Arbor, MI.

LAMBIN, E. F. & LINDERMAN, M., 2006, Time series of remote sensing data for land change science. IEEE Transactions on Geoscience and Remote Sensing, **44**(7), pp. 1926-1928.

LEE, R., KASTENS, D. L., PRICE, K. P. & MARTINKO, E. A., 1999 Forecasting corn yield in Iowa using remotely sensed data and vegetation phenology information. Proceedings, PECORA 14 land satellite information regional conference Denver, CO December 6 –10, American society of photogrammetric Engineering and remote sensing.

LI, B., MORRIS, J. & MARTIN, E. B., 2002, Model selection for partial least squares regression. Chemometrics and Intelligent Laboratory Systems, **64**(1), pp. 79-89.

LIANG, S., 2004, Quantitative remote sensing of land surfaces (Hoboken, NJ: Wiley).

LILLESAND, T. M. & KIEFER, R. W., 1987, Remote Sensing and Image Interpretation (New York, Chichester, Brisbane, Toronto, Singapore: John Wiley and Sons).

LIU, W. T. & KOGAN, F., 2002, Monitoring Brazilian soybean production using NOAA/AVHRR based vegetation condition indices. International Journal of Remote Sensing, **23**(6), pp. 1161-1179.

LOS, S. O., 1998, Estimation of the ratio of sensor degradation between NOAA-AVHRR channels 1 and 2 from monthly NDVI composites. IEEE Transactions on Geoscience and Remote Sensing, **36**(1), pp. 206-213.

LOS, S. O., NORTH, P. R. J., GREY, W. M. F. & BARNSLEY, M. J., 2005, A method to convert AVHRR Normalized Difference Vegetation Index time series to a standard viewing and illumination geometry. Remote Sensing of Environment, **99**(4), pp. 400-411.

LUNETTA, R. S. & LYON, J. G., 2004, Remote sensing and GIS accuracy assessment (Boca Raton, Fla.: CRC Press).

MARTENS, H., 2001, Reliable and relevant modelling of real world data: a personal account of the development of PLS Regression. Chemometrics and Intelligent Laboratory Systems, **58**(2), pp. 85-95.

MARTENS, H., HOY, M., WESTAD, F., FOLKENBERG, D. & MARTENS, M., 2001, Analysis of designed experiments by stabilised PLS Regression and jack-knifing. Chemometrics and Intelligent Laboratory Systems, **58**(2), pp. 151-170.

MARTENS, H. A., 1985, Doctoral thesis, Multivariable Calibration-Quantitative interpretation of non-selective chemical data. R. Trondheim, Technical University of Norway.

MASEK, J. G., VERMOTE, E. F., SALEOUS, N. E., WOLFE, R., HALL, F. G., HUEMMRICH, K. F., FENG, G., KUTLER, J. & TENG-KUI, L., 2006, A Landsat surface reflectance dataset for North America, 1990-2000. IEEE Geoscience and Remote Sensing Letters, **3**(1), pp. 68-72.

MASELLI, F., CONESE, C., PETKOV, L. & GILABERT, M. A., 1992, Use of NOAAAVHRR NDVI data for environmental monitoring and crop forecasting in the Sahel. Preliminary results. International Journal of Remote Sensing, **13**, pp. 2743–2749.

MASELLI, F., GILABERT, M. A. & CONESE, C., 1998, Integration of High and Low Resolution NDVI Data for Monitoring Vegetation in Mediterranean Environments. Remote Sensing of Environment, **63**(3), pp. 208-218.

MASELLI, F. & REMBOLD, F., 2001, Analysis of GAC NDVI data for cropland identification and yield forecasting in Mediterranean African countries. Photogrammetric Engineering and Remote Sensing, **67**(5), pp. 593– 602.

MILES, D. L., COLORADO STATE UNIVERSITY. COOPERATIVE EXTENSION SERVICE. & UNITED STATES. ENVIRONMENTAL PROTECTION AGENCY. REGION VIII., 1977, Salinity in the Arkansas Valley of Colorado (Denver: Region VIII Environmental Protection Agency).

MIURA, T., HUETE, A. R. & YOSHIOKA, H., 2000, Evaluation of sensor calibration uncertainties on vegetation indices for MODIS. IEEE Transactions on Geoscience and Remote Sensing, **38**(3), pp. 1399-1409.

MORISETTE, J. T., BARET, F. & LIANG, S., 2006, Special Issue on Global Land Product Validation. IEEE Transactions on Geoscience and Remote Sensing, **44**(7), pp. 1695-1697.

MUELLER, R., BORYAN, C., CRAIG, M., FLEMING, M. & HANUSCHAK, G., 2003, Pilot Research Project: Investigation of Very High Resolution Spaceborne Imagery for Citrus Tree Counting. Report for FDOC Contract No. 02-17.

MYERS, R. H., 1986, Classical and modern regression with applications (Boston, Mass.: Duxbury Press).

MYNENI, R. B., ASRAR, G., TANRE, D. & CHOUDHURY, B. J., 1992, Remote sensing of solar radiation absorbed and reflected by vegetated land surfaces. IEEE Transactions on Geoscience and Remote Sensing, **30**(2), pp. 302-314.

MYNENI, R. B., TUCKER, C. J., ASRAR, G. & KEELING, C. D., 1998, Interannual variations in satellite-sensed vegetation index data from 1981 to 1991. Journal of Geophysical Research, **103**, pp. 6145-6160.

MYNENIL, R. B. & ASRAR, G., 1992, Simulation Of Space Measurements Of Vegetation Canopy Bidirectional Reflectance Factors. Geoscience and Remote Sensing Symposium, 1992. IGARSS '92. International.

NAES, T., IRGENS, C. & MARTENS, H., 1986, Comparison of Linear Statistical Methods for Calibration of NIR Instruments. Applied Statistics, **35**(2), pp. 195-206.

PINTY, B. & VERSTRAETE, M. M., 1991, GEMI: A Non-Linear Index to Monitor Global Vegetation from Satellites. Plant Ecology, **101**(1), pp. 15-20.

POTDAR, M. B., 1993, Sorghum yield modeling based on crop growth parameters determined from visible and near-IR channel NOAA AVHRR data. International Journal of Remote Sensing, **14**, pp. 895– 905.

PREISENDORFER, R. W. & MOBLEY, C. D., 1988, ***Principal component analysis in meteorology and oceanography*** (Amsterdam ; New York New York, NY: Elsevier ; Distributors for the U.S. and Canada Elsevier Science Pub. Co.).

PRICE, J. C., 1987a, Calibration of satellite radiometers and the comparison of vegetation indices. Remote Sensing of Environment, **21**(1), pp. 15-27.

PRICE, J. C., 1987b, Radiometric calibration of satellite sensors in the visible and near infrared: History and outlook. Remote Sensing of Environment, **22**(1), pp. 3-9.

PRICE, J. C., 1988, An update on visible and near infrared calibration of satellite instruments. Remote Sensing of Environment, **24**(3), pp. 419-422.

PRICE, J. C., 1991, Beltsville symposium XV remote sensing for agriculture. Remote Sensing of Environment, **35**(2-3), pp. 77-77.

PRIVETTE, J. L., FOWLER, C., WICK, G. A., BALDWIN, D. & EMERY, W. J., 1995, Effects of orbital drift on advanced very high resolution radiometer products: Normalized difference vegetation index and sea surface temperature. Remote Sensing of Environment, **53**(3), pp. 164-171.

PUREVDORJ, T., ISHIYAMA, T., TATEISHI, R. & FURUYA, T., 1996, Estimating of percent vegetation cover using vegetation indices. Conference Proceedings of the Japan Society for Photogrammetry and Remote Sensing. Yamagata, Japan.

PUREVDORJ, T., TATEISHI, R., ISHIYAMA, T. & HONDA, Y., 1998, Relationship between percent vegetation cover and vegetation indices. International Journal of Remote Sensing, **19**, pp. 3519-3535.

QI, J., CHEHBOUNI, A., HUETE, A. R. & KERR, Y. H., 1994, Modified Soil Adjusted Vegetation Index (MSAVI). Remote Sensing of Environment, **48**, pp. 119-126.

QUARMBY, N. A., MILNES, M., HINDLE, T. L. & SILLEOS, N., 1993, The use of multi-temporal NDVI measurements from AVHRR data for crop yield estimation and prediction. International Journal of Remote Sensing, **14**, pp. 199– 210.

RAO, C. R. N. & CHEN, J., 1995a, Inter-satellite calibration linkages for the visible and near-infrared channels of the Advanced Very High Resolution Radiometer on the NOAA-7, -9 and -11 spacecraft. International Journal of Remote Sensing **16**, pp. 1931-1942.

RAO, C. R. N. & CHEN, J., 1995b, Inter-satellite calibration linkages for the visible and near-infrared channels of the Advanced Very High Resolution Radiometer on the NOAA-7, -9, and -11 spacecraft. International Journal of Remote Sensing, **16**, pp. 1931-1942.

RAO, C. R. N. & CHEN, J., 1996, Post-launch calibration of the visible and near-infrared channels of the Advanced Very High Resolution Radiometer on the NOAA-14 spacecraft. International Journal of Remote Sensing, **17**, pp. 2743-2747.

RAO, C. R. N. & CHEN, J., 1999, Revised post-launch calibration of the visible and near-infrared channels of the Advanced Very High Resolution Radiometer on the NOAA-14 spacecraft. International Journal of Remote Sensing, **20**(18), pp. 3485-3491.

RASMUSSEN, M. S., 1992, Assessment of millet yields and production in northern Burkina Faso using integrated NDVI from the AVHRR. International Journal of Remote Sensing, **13**, pp. 3431– 3442.

RASMUSSEN, M. S., 1997, Operational yield forecast using AVHRR NDVI data: Reduction of environmental and inter-annual variability. International Journal of Remote Sensing, **18**, pp. 1059– 1077.

RASMUSSEN, M. S., 1998, Developing simple, operational, consistent NDVI-vegetation models by applying environmental and climatic information. Part II: Crop yield assessment. International Journal of Remote Sensing, **19**(1), pp. 119-139.

RICHARDSON, A. J. & WIEGAND, C. L., 1977, Distinguishing vegetation from soil background information. Photogrammetric Engineering and Remote Sensing, **43**, pp. 1541-1552.

RUDORFF, B. F. T. & BATISTA, G. T., 1991 Wheat yield estimation at the farm level using TM Landsat and agro meteorological data. International Journal of Remote Sensing, **12**, pp. 2477– 2484.

SCHANDA, E., 1986, Physical fundamentals of remote sensing : with 102 figures and 14 tables (Berlin ; New York: Springer-Verlag).

SCHOWENGERDT, R. A., 1983, Techniques for image processing and classification in remote sensing (New York, N.Y. : Academic Press, Inc.).

SHROYER, J. P., WHITNEY, D. & PATERSON, D., 2004, Wheat production handbook (Manhattan, Kansas: K-State research & Extension).

SIMONIELLO, T., CUOMO, V., LANFREDI, M., LASAPONARA, R. & MACCHIATO, M., 2004, On the relevance of accurate correction and validation procedures in the analysis of AVHRR-NDVI time series for long-term monitoring. Journal of Geophysical Research – Atmospheres, **109**(D20), pp. 20101-20113.

SMITH, D. L., READ, P. D. & MUTLOW, C. T., 1997, The calibration of the visible/near infra-red channels of the Along-Track-Scanning-Radiometer-2 (ATSR-2). Proceedings of the SPIE Conference on Sensors, Systems, and Next Generation Satellites. Bellingham, Washington, USA, SPIE.

SNEE, R. D., 1977, Validation of Regression Models: Methods and Examples. Technometrics, **19**(4), pp. 415-428.

STEINWAND, D. R., HUTCHINSON, J. A. & SNYDER, J. P., 1995, Map projections for global and continental data sets and an analysis of pixel distortion caused by reprojection. Photogrammetric Engineering and Remote Sensing, **61**(1487-1497).

TAHNK, W. R. & COAKLEY JR, J. A., 2001a, Improved calibration coefficients for NOAA-14 AVHRR visible and near-infrared channels. International Journal of Remote Sensing, 22(7), pp. 1269-1283.

TAHNK, W. R. & COAKLEY JR, J. A., 2001b, Updated calibration coefficients for NOAA-14 AVHRR Channels 1 and 2. International Journal of Remote Sensing, 22(15), pp. 3053-3057.

TEILLET, P. M., FEDOSEJEVS, G., GAUTHIER, R. P., O'NEILL, N. T., THOME, K. J., BIGGAR, S. F., RIPLEY, H. & MEYGRET, A., 2001, A generalized approach to the vicarious calibration of multiple Earth observation sensors using hyperspectral data. Remote Sensing of Environment, 77(3), pp. 304-327.

TEILLET, P. M., MARKHAM, B. L. & IRISH, R. R., 2006, Landsat cross-calibration based on near simultaneous imaging of common ground targets. Remote Sensing of Environment, 102(3-4), pp. 264-270.

TEILLET, P. M., SLATER, P. N., DING, Y., SANTER, R. P., JACKSON, R. D. & MORAN, M. S., 1990, Three methods for the absolute calibration of the NOAA AVHRR sensors in-flight. Remote Sensing of Environment, **31**(2), pp. 105-120.

TRYGG, J. & WOLD, S., 1998, PLS regression on wavelet compressed NIR spectra. Chemometrics and Intelligent Laboratory Systems, **42**(1-2), pp. 209-220.

TUCKER, C. J., ELGIN, J. J. H., MCMURTREY, I. J. E. & FAN, C. J., 1979, Monitoring corn and soybean crop development with hand-held radiometer spectral data. Remote Sensing of Environment, **8**(3), pp. 237-248.

TUCKER, C. J., HOLBEN, B. N., ELGIN, J. H., JR. & MCMURTREY, J. E., 1980, Relationship of spectral data to grain yield variation. Photogrammetric Engineering and Remote Sensing, **45** pp. 657–666.

TUCKER, C. J., VANPRAET, C., BOERWINKEL, E. & GASTON, A., 1983, Satellite remote sensing of total dry matter production in the Senegalese Sahel. Remote Sensing of Environment, **13**(6), pp. 461-474.

TUCKER, C. J., VANPRAET, C. L., SHARMAN, M. J. & VAN ITTERSUM, G., 1985, Satellite remote sensing of total herbaceous biomass production in the senegalese sahel: 1980-1984. Remote Sensing of Environment, **17**(3), pp. 233-249.

UNITED STATES CROP REPORTING BOARD [USCRB], 2005, Crop production (Washington, D.C.: Crop Reporting Board Statistical Reporting Service U.S. Dept. of Agriculture).

UNITED STATES CROP REPORTING BOARD [USCRB], 2006, Crop production (Washington, D.C.: Crop Reporting Board Statistical Reporting Service U.S. Dept. of Agriculture).

UNSALAN, C. & BOYER, K. L., 2004, Linearized vegetation indices based on a formal statistical framework. IEEE Transactions on Geoscience and Remote Sensing, **42**(7), pp. 1575-1585.

US CROP REPORTING BOARD (USCRB), 2006, Crop production (Washington DC Crop Reporting Board Statistical Reporting Service, US Department of Agriculture).

US DEPARTMENT OF AGRICULTURE (USDA), 2006, Outlook for US Agricultural Trade. Available online at: http://www.ers.usda.gov/publications (Washington DC, USDA).

VAN LEEUWEN, W. J. D., ORR, B. J., MARSH, S. E. & HERRMANN, S. M., 2006, Multi-sensor NDVI data continuity: Uncertainties and implications for vegetation monitoring applications. Remote Sensing of Environment, **100**(1), pp. 67-81.

VOGEL, F. A. & BANGE, G. A., 1999, Understanding Crop Statistics. IN SERVICE, N. A. S. & BOARD, W. A. O. (Eds.) Office of the Chief Economist. U.S. Department of Agriculture.

WESTAD, F. & MARTENS, H., 2000, Variable selection in near infrared spectroscopy based on significance testing in partial least squares regression. Journal of Near Infrared Spectroscopy, **8**(2), pp. 117-124.

WIEGAND, C. L. & RICHARDSON, A. J., 1990, Use of spectral vegetation indices to infer leaf area, evapotranspiration and yield. Agronomy Journal, **82**, pp. 623– 636.

WIEGAND, C. L., RICHARDSON, A. J., JACKSON, R. D., PINTER JR., P. J., AASE, J. K. & SMIKA, D. E., 1986, Development of agrometeorological crop model inputs from remotely sensed information. IEEE Transactions on Geoscience and Remote Sensing, **24**(1), pp. 83– 89.

WIEGAND, C. L., RICHARDSON, A. J. & KANEMASU, E. T., 1979, Leaf area index estimates for wheat from Landsat and their implications for evapotranspiration and crop modeling. Agronomy Journal, **71**, pp. 336-342.

WILLMOTT, C. J., 1982, Some comments on the evaluation of model performance. Bulletin American Meteorological Society, **63**(11), pp. 1309-1313.

WOLD, S., 2001, Personal memories of the early PLS development. Chemometrics and Intelligent Laboratory Systems, **58**(2), pp. 83-84.

WOLD, S., ESBENSEN, K. & GELADI, P., 1987, Principal component analysis. Chemometrics and Intelligent Laboratory Systems, **2**(1-3), pp. 37-52.

WOLD, S., SJOSTROM, M. & ERIKSSON, L., 2001a, PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems, **58**(2), pp. 109-130.

WOLD, S., TRYGG, J., BERGLUND, A. & ANTTI, H., 2001b, Some recent developments in PLS modeling. Chemometrics and Intelligent Laboratory Systems, **58**(2), pp. 131-150.

XIANG, G., HUETE, A. R. & DIDAN, K., 2003, Multisensor comparisons and validation of MODIS vegetation indices at the semiarid Jornada experimental range. ***IEEE Transactions on Geoscience and Remote Sensing***, 41(10), pp. 2368-2381.