TIME SERIES AND SPATIAL ANALYSIS OF CROP YIELD

by

YARED ASSEFA

BSc., Alemaya University, Ethiopia, 2000
MSc., Wageningen University, The Netherlands, 2006
PhD., Kansas State University, USA, 2010

A THESIS

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2013

Approved by:

Major Professor:
Dr. Juan Du

Abstract

Space and time are often vital components of research data sets. Accounting for and utilizing the space and time information in statistical models become beneficial when the response variable in question is proved to have a space and time dependence. This work focuses on the modeling and analysis of crop yield over space and time. Specifically, two different yield data sets were used. The first yield and environmental data set was collected across selected counties in Kansas from yield performance tests conducted for multiple years. The second yield data set was a survey data set collected by USDA across the US from 1900-2009. The objectives of our study were to investigate crop yield trends in space and time, quantify the variability in yield explained by genetics and space-time (environment) factors, and study how spatio-temporal information could be incorporated and also utilized in modeling and forecasting yield. Based on the format of these data sets, trend of irrigated and dryland crops was analyzed by employing time series statistical techniques. Some traditional linear regressions and smoothing techniques are first used to obtain the yield function. These models were then improved by incorporating time and space information either as explanatory variables or as auto- or cross-correlations adjusted in the residual covariance structures. In addition, a multivariate time series modeling approach was conducted to demonstrate how the space and time correlation information can be utilized to model and forecast yield and related variables. The conclusion from this research clearly emphasizes the importance of space and time components of data sets in research analysis. That is partly because they can often adjust (make up) for those underlying variables and factor effects that are not measured or not well understood.

Table of Contents

List of Figures

List of Tables

Acknowledgments

My first thanks goes to God, the reason being that this (completing research or getting a degree) is possible only when circumstances beyond human control line up with the human's plan. All those circumstances such as being alive, being healthy, having peace…, I believe, are under his control. Therefore, I really want to thank God, the uncaused causer, the unmoved mover, and the ultimate reality.

Second, I want to thank Dr. Juan Du for accepting to be my Major Professor. Her expertise on spatio-temporal analysis, her trust, kindness, and her comments shaped my project a lot. Similarly, I would like to thank Dr. Weixing Song and Dr. Scott Staggenborg for serving as my committee members and for the valid comments and support.

While pursuing my masters in statistics, I was fortunate to find teaching, research, and consulting positions in the Department of Agronomy. I want to thank the department as a whole and specially Dr. Scott Staggenborg, Dr. Kevin Donnelly, Dr. Nathan Nelson, Dr. Tessfaye Tesso, Dr. Chuck Rice, and Dr. Kraig Roozeboom for the opportunities they have given me to work under their supervision.

Last but the most important credit is to my wife. As a father of two girls, this work was only possible because my wife was covering for me at home. I have no words for that. Thanks and lots of love to my wife Betelihem Lakew.

Dedication

To my Wife
AND
My Daughters:
AXUM and MARY

Chapter I

GENERAL INTRODUCTION AND BACKGROUND OF THE STUDY

Data observed over time result in time series (temporal) data (Anderson, 1971; Shumway and Stoffer, 2010) and data observed across space result in a spatial data (Haining, 2003). In many aspects, time and space have similar characteristics and are closely related (Schlesinger, 1975; Peuquet, 2001). However, space and time are not the same in absolute sense, i.e., unlike time which is uni-directional, space is multi-directional and unlike space which is relatively static and given, time is dynamic and irreversible (Johnsons, 1913).

Time series and spatial data analysis enables one to understand the behavior of data across time and space. The goal of both time series and spatial analysis procedures is to identify patterns, investigate possible causes of these patterns and often forecast for the unknown (space or time). The time and space information of a data plays an important role at pattern identification step. If temporal or spatial pattern is identified, the possible reasons that cause the pattern that are behind space and time will be investigated. Then the response variable can be modeled as function of time and space or as a function of the direct cause for forecasting.

The importance of space and time in a data set is not restricted to cases where we intend to do time series and spatial analysis. Space and time are often vital components of research data set. That is partially because they can often adjust for those underlying

variables and factor effects that are not measured or not well understood. In the present paper, we will study whether and how those natural components adjust or make up for unmeasured variables or less understood relations. For example, in crop and soil science studies, yield is proved to be affected by genetics and multi-environmental factors (Lobell et al., 2009; Machado et al., 2002; Perez-Quezada et al., 2003). Due to the vast number of factors involved in determining crop yield, it is impossible to measure all possible factors in every experiment. In addition, it is possible that there are some factors that are not discovered. Moreover, how these many factors combined to influence yield was not adequately understood.

In such cases, time and space information usually can make up for those unmeasured or less understood variables in the sense of explaining the variability of response. In experiments that involve measuring variable like crop yield, which are influenced by many factors, the selection of space and time of experiment should, therefore, be done as carefully as possible. The analysis of such a data set should also take into account the space and time information as importantly as targeted treatments.

RESEARCH GAPS

The research gaps that this report intends to address is as follows. Different forms of spatial and temporal data sets are available for crop yield. One of these data sets is annual hybrid performance test in which yield and management data are collected as part of yearly comparisons of different hybrids of crops. The comparisons of hybrids have been done in different locations and for long time. However, more than the specific purpose of

comparing yields at a time and space; the story that might come out from time and spatial analysis of these data sets was less exploited.

A number of things can be accomplished using a time series and spatial analysis of the hybrid yield trial data. First of all, the trend of yield over time and space can be determined. In the hybrid yield trial data sets we have multiple hybrids, i.e., we have a multiple yield data annually for a crop, at a time and at same space. Therefore, the variation in yield due to genetics and environment can be studied. Hybrid trials have separate components for different management factors such as irrigated and dryland that can go long years back than survey data. Using these data, the relationships between yield and a number of management factors can be investigated.

The other spatial and temporal crop data set is a survey data set collected by USDA National Statistical Service. This data set is rich in its spatial and temporal dimension and it has been used by many for trend analysis. However, not many studies utilized the potential of the data set for auto-and cross correlation analysis between different crop parameters for modeling and forecasting purposes. The data provides the opportunity to look at relationships between past crop area, total production, and prices to forecast future.

Moreover, the importance of time and space components of experimental design and research data analysis can be well illustrated by studying these data sets. In crop and soil science studies, it is common to conduct a research in multiple years (time) and locations

(space). However, the following two points are observed on the selection and use of space and time components of a research data set.

1. **Selection of time and space at the design phase of an experiment is often considered secondary.**

Researchers usually initiate a research with an interest in studying the impact of certain independent variables (treatments) on a dependant variable, say yield. They often carefully design the study in regard to how treatments should be arranged and replicated. The selection of where and when this research should be conducted did not seem to be given enough attention, i.e., it is usually based on where there is enough space and personnel than demand of the actual research. Considering the factors that might affect the dependant variable (crop yield), however, the treatments (independent variables) in this experiments are usually one or up to three factors with all the spatio-temporal variability largely neglected.

2. **In the analysis step, the use of time and space information is inconsistent across research.**

In spite of an experiment being conducted across space and time, the use of the space and time information from a data varies from one research to another.

a. In most cases, researchers tend to analyze the effect of treatments just location by location and year by year without checking whether effect is environment specific.

b. Some don't use space and location information and just analyze effect of treatment
c. In some cases, space and year information become random effects and in other cases space and time information (both or one) becomes fixed effect for unjustified reasons.

We felt these inconsistencies can be addressed by showing how important the space and time information are in a data analysis and by providing a logical approach on how time and space information should be incorporated in a model.

GENERAL OBJECTIVES OF THE RESEARCH

Based on research gaps identified above, the present study was initiated with the following general objectives:

1. to investigate and determine the magnitude of crop yield trends in space and time,
2. to quantify the variability in yield explained by genetics and space-time (environment) factors,
3. to build optimal crop models for early forecasting using the temporal and spatial characteristics of the data, and
4. to study how spatio-temporal information could be incorporated and also utilized in modeling and forecasting yield.

GENERAL OVERVIEW AND METHODOLOGIES

In the next chapter of this thesis, analysis result from a data assembled from irrigated and dryland corn performance trials conducted in Kansas for the years 1939 through 2009

will be presented. The data contains only the average yield of hybrids tested for each trial site for these years. A simple regression analysis of data over the time period and a comparison of means of a data at different time and space intervals were conducted and the trend analysis result was presented. The detail of this chapter was published in Agronomy journal 104:473-482 (2012) in a title "Dryland and irrigated corn yields with climate, management, and hybrid changes from 1939 through 2009. Therefore, we presented only part of the report with a little modification, on the software used for analysis. The units in this chapter are international (S.I) units.

In the third chapter, a report based on a data assembled from Kansas Corn Performance Trials (KCPT) and Kansas Grain Sorghum Performance Trials (KGSPT) conducted at 11 counties of Kansas within the years between 1992 and 2009 will be presented. Traditional regression and smoothing techniques were used to develop varieties of yield functions with and without assumption of independence. Possible improvement of these models with time and space information will be demonstrated. The units in chapter three are U.S. units.

In the fourth chapter, we will present a multivariate analysis result using the annual corn yield, harvest area, and price survey data available in USDA National Statistics Service website for the years 1900-2011. Auto-and cross correlation, spatial autocorrelation using Morons I, and semivariogram analysis were carried out. Based on the analysis, a vector autoregressive (VAR) models that can predict price, area, and total production was developed. Another vector autoregressive model that is capable of using past yield

information from state and its neighbors was also developed. This detail will be presented in fourth chapter. The units in this chapter are S.I units.

REFERENCES

Anderson T.W. 1971. The statistical analysis of time series. pp 1. John Willey and Sons INC, NY

Haining, R.P. 2003. Spatial Data Analysis: Theory and Practice. pp 22. Cambridge: Cambridge University Press. Johnson JT. 1913. Bergson's space and time: a criticism. University of Wisconsin (thesis).

Lobell D.B., Cassman K.G., and Field C.B. 2009. Crop yield gabs: their importance, magnitude, and causes. Annu. Rev. Environ. Resour. 2009. 34:179–204.

Machado S., Bynum E. D., Archer T. L., Lascano R. J., Wilson L. T, Bordovsky J., Segarra E., Bronson K., Nesmith D. M., and Xu W. 2002. Spatial and Temporal Variability of Corn Growth and Grain Yield: Implications for Site-Specific Farming. Crop Sci. 42:1564–1576 (2002).

Perez-Quezada J.F., Pettygrove G. S. and Plant R. E. 2003. Spatial–Temporal Analysis of Yield and Soil Factors in Two Four-Crop–Rotation Fields in the Sacramento Valley, California. Agron. J. 95:676–687.

Peuquet D.J, 2001. Making Space for Time: Issues in Space-Time Data Representation. GeoInformatica 5(1): 11-32.

Shumway and Stoffer. 2010. Time Series Analysis and Its Applications: With R Examples. 3rd edition. pp 1. Springer.

Chapter II

Trend Analysis of Dryland and Irrigated Corn

(The detail of this chapter is published in Agronomy journal 104:473-482 (2012) in a title "Dryland and irrigated corn yields with climate, management, and hybrid changes from 1939 through 2009)

ABSTRACT

A non-seasonal long term pattern of a time series data is called a trend. A simple linear regression model was fitted for crop yield with an explanatory variable, time (year) to investigate trend in irrigated and dryland corn. The objective of the study was to determine the magnitude of yield changes in irrigated and dryland corn for the years 1939 through 2009 at different districts of Kansas. The data for this analysis was assembled from irrigated and dryland corn performance trials conducted in Kansas for the years 1939 through 2009. Results of our analysis suggested a significant trend for both dryland and irrigated corn yields over time and space but the interaction of the two factors was not significant. Spatially, average dryland yields in Kansas decreases significantly from east to west and slightly from north to south. Temporally, corn yields have increased at an average rate of 94 kg ha^{-1} yr^{-1} in dryland and 127 kg ha^{-1} yr^{-1} in irrigated trials. The rate of corn yield changes over time, however, was not regular for the seven decades considered. Both irrigated and dryland yields increased significantly at least every two decades until the last three, during which dryland yields stagnated.

INTRODUCTION

A data that is observed sequentially in time is called a time series data. A time series data can be decomposed into trend, seasonality, cycle, and random fluctuation components. The trend component of a time series data is the long term pattern or change in the mean of the data. Existence of a trend in a time series data can be identified by simply comparing the means of a time series data at different intervals or by a simple regression analysis of data over the time period. In the present research, we used these simple techniques to identify trend in dryland and irrigated corn yields in Kansas.

For the detail of the result in this chapter:

https://www.agronomy.org/publications/aj/abstracts/104/2/473

Title: Dryland and Irrigated Corn Yield with Climate, Management, and Hybrid Changes from 1939 through 2009

Yared Assefa, Kraig L. Roozeboom, Scott A. Staggenborg, and Juan Du

Corresponding author (yareda@ksu.edu)

Published in Agron J 104:473-482 (2012)
DOI: 10.2134/agronj2011.0242

5585 Guilford Rd., Madison, WI 53711 USA

Chapter III

Space and Time Adjustment in Dryland Crop Yield Models

ABSTRACT

A number of statistical models have been developed to describe crop yield, yet, there is no one conclusive crop yield model that works everywhere. The objectives of the present study were to study the complex relationship between yield and factors responsible for its variability, to quantify the variability that is explained by genetics and time-space (environment) factors, to develop and compare statistical yield functions, and to study how time and spatial information could be incorporated and utilized in modeling yield. The data for the study was assembled from Kansas Corn Performance Trials (KCPT) and Kansas Grain Sorghum Performance Trials (KGSPT) conducted at 11 counties of Kansas within the years between 1992 and 2009. Climatic data was assembled from Kansas State Weather Data Library. First using space, time and genetics information and then with other environmental variables, varieties of yield functions were developed. Models with only environmental variables were then improved by incorporating time and space information as an explanatory variable. Consequently, spatial and temporal dependence remaining on the error was brought as a case. In this instance a method of generalized least square estimation with correlated error variance was suggested as opposed to ordinary least square regression.

INTRODUCTION

A number of statistical models have been developed to describe crop yield in relation to environmental and technological factors. The majority of statistical crop models developed before the late 1950s relate crop yield with climatic conditions alone (Compton, 1943; Mathews and Brown, 1938; Sanderson, 1954). Models released after the 1950s started to incorporated impact of technological factors on yield models (Shaw, 1964; Oury, 1965; Kaylen et al., 1991; Nelson and Dale., 1978; Thompson, 1975; Sclenker et al., 2004; Lobell and Asner, 2003).

Different statistical techniques have also been suggested and employed in modeling yield, i.e., simple regression models, multiple regression models, nonlinear models, time series models, spatial models, spatio-temporal models, and others (Stone, 2006; Sclenker and Roberts, 2006; Gumpertz and Rawlings,1992; Lobell, 2010; Ozaki et al., 2008). Not only different approaches are used but also mixed messages are forwarded about the impact of different variables on crop yield. For example, the impact of increasing temperature on crop yield is reported to be negative (Lobell and Asner, 2003; Deschenes and Greenstone, 2007) and also reported positive (McCarl et al., 2008; crops. Lobell et al. 2008). Similar mixed conclusions were also forwarded for the relationship between yield and rainfall, tillage, fertilizer, and other variables.

Considering the age of agricultural research, the statement that yield is one of the longest and most frequently studied variable in history of human kind research can not be an over statement. Yet, there is no one conclusive crop yield model that works everywhere. That

is perhaps due to number of factors involved in determining yield. Crop yield is a function of crops genetics and multi-environment factors. For the major cereal crops of the world like corn, rice, wheat, barley, and sorghum, the number of days that the crop is out in field (length of growing season) is often more than 120 days. Theoretically, what happens in terms of environmental conditions and what is available in terms of resources, every single day the crop is out in the field affects final yield. However, one can imagine that a statistical model that accounts for the relationship between yield and every environmental factor for the length of the growing season are complex, if conceived. Mechanistic crop models attempt to capture this complex relationship between yield, genetics, and multi-environmental factors.

Our objectives were to study the complex relationship between yield and factors responsible for its variability, to quantify the variability in yield that is explained by genetics and time-space (environment) factors, to develop and compare statistical yield functions, and to study how time and spatial information could be incorporated and adjust statistical crop models.

The data for the present study was assembled from Kansas Corn Performance Trials (KCPT) and Kansas Grain Sorghum Performance Trials (KGSPT) conducted at 11 counties of Kansas within the years between 1992 and 2009. Occasionally, a survey data of crop yield collected by USDA across Kansas was used to explain certain facts. The detail of how data is assembled and a brief on analysis steps will be presented in next section, and results and detail of analysis will follow.

DATA ASSEMBLY AND ANALTICAL STEPS

The data for this research was assembled from Kansas Corn Performance Trials (KCPT) and Kansas Grain Sorghum Performance Trials (KGSPT) conducted in 11 counties of Kansas between 1992 and 2009 (Fig. 1). In these trials, different corn and sorghum hybrids were planted every year at each location and were evaluated for yield and other traits. In addition to yield data, information on the names of hybrids, the amount of nitrogen (N) fertilizer applied, planting and harvesting date, cropping system (irrigated or dryland), average daily maximum and minimum temperature, and average daily rainfall data were also assembled.

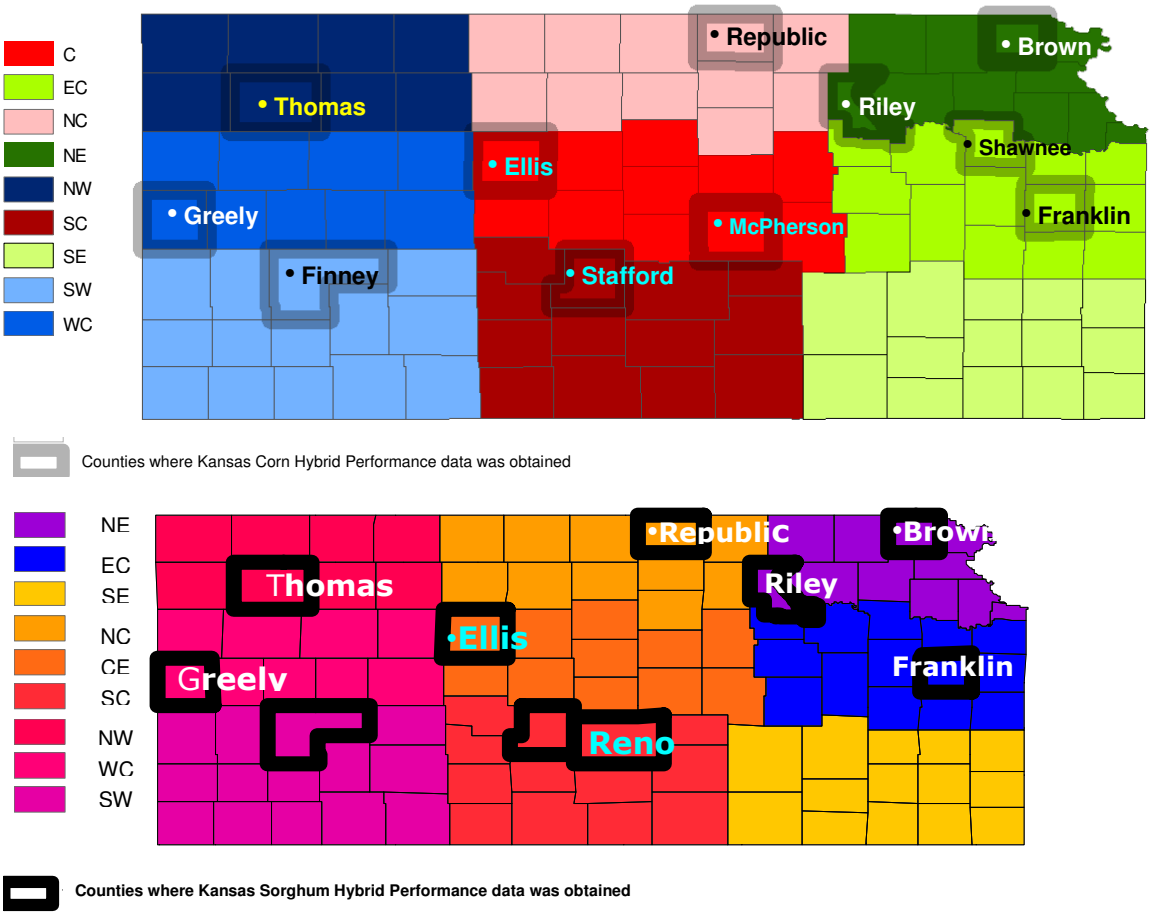

Figure 1 Map of Kansas depicting counties and their district where the Kansas Corn Hybrid Performance Trials data for the years 1992-2009 was assembled from.

The data set which constitute name of each county (space), years the data was collected (time), name of hybrid, yield of each hybrid, amount of nitrogen fertilizer applied, length of growing season, cropping system (irrigated or dryland), rainfall, minimum and maximum temperature of months from April to September, were organized for corn and sorghum separately in long data format. With county and year, a column called environment was created.

Analysis was performed in R (R 2.15 statistical program) and whenever necessary in SAS. First the distribution of irrigated and dryland yield of each crop was determined by plotting yield values in x-axis and their frequency in y-axis. The mean and standard deviation of yield for dryland and irrigated corn and sorghum was also analyzed.

Second, the variability in yield that is explained by genetics and environment was partitioned using a mixed model with random environment and hybrid variables. By doing so, the variability that is explained by environment, genetics, and genetics by environment interaction was determined.

Environment explained the largest variation in crop yield. In the next step of our analysis, we split environment into space and time, because it is first defined as a combination of these two factors. Within space and time, environmental and management factors that might directly affect crop yield were listed based on pure prior knowledge. Since we had a data on few of the important weather and management factors, yield was modeled against these variables using multiple linear regression and robust locally weighted

regression and smoothing techniques. The significance of incorporating time and space information as an explanatory variable to improve model fit was demonstrated.

The dependence of yield over time and space was demonstrated on a survey data set that has wide spatial coverage and long time components. Semivarograms and autocorrelation functions were used to make this point. Literature is provided on the importance and how adjustments should be done on the residual covariance structure when models suffer with correlated spatial, temporal, and spatial temporal errors.

RESULT

Yield Distribution

The distribution of crop yield and the variability associated with it was analyzed and plotted for irrigated and dryland corn and sorghum (Fig. 2). From the result, we can conclude that yields of irrigated and dryland corn and sorghum were approximately normally distributed with means 204,126,143, 104 bushel/acre, respectively.

The mean, standard deviation, coefficients of variation, and a 95% confidence intervals for yield in Kansas are presented in Table 1. The deviations from the mean were higher for corn than for grain sorghum. The deviation from the mean in irrigated and dryland yields appear to be similar for both crops. However, the coefficients of variations were higher for dryland yields than they are for irrigated.

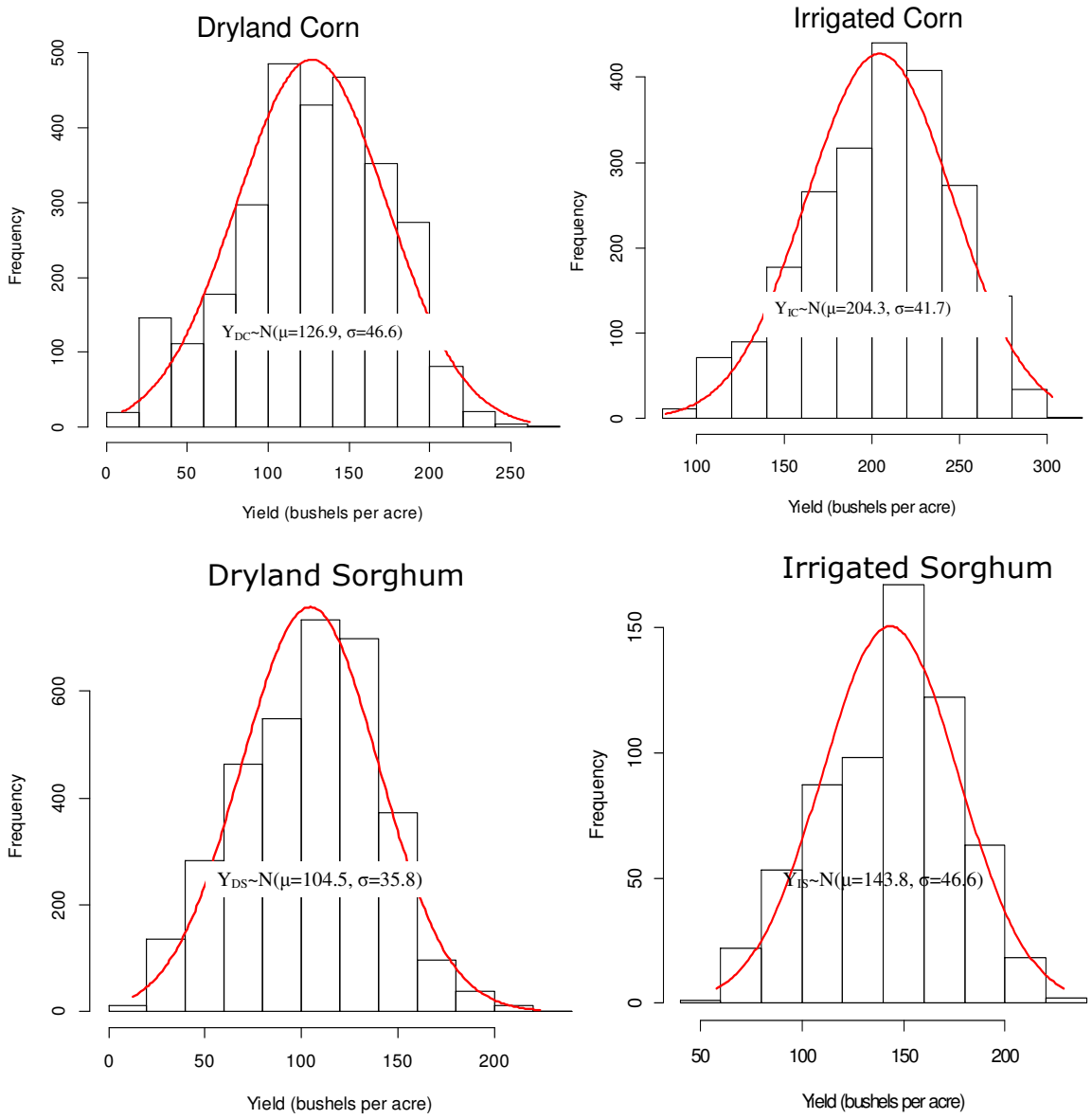

Figure 2. Distribution of dryland and irrigated sorghum yields from Kansas Hybrid Performance Trial Data.

Table 1. The mean, the standard deviation, coefficient of variation, and 95% confidence intervals for irrigated and dryland corn yields in Kansas

Crop	Cropping System	Mean	Standard Deviation	Coefficient of Variation	Upper Limit	Lower Limit
					95% Confidence Interval	
Corn	Irrigated	204.4	41.7	0.20	286	123
	Dryland	126.9	46.6	0.37	218	36
Sorghum	Irrigated	143.5	33.5	0.23	210	78
	Dryland	104.5	35.8	0.34	175	34

Partitioning the variability

How much of the variability in yield of corn and sorghum can be explained by environment and genetics was another important question. Since our data for each of dryland and irrigated corn is from multiple hybrids from one location and a year, hybrid (genetics) is one source of variation. In a given year, we had multiple locations; it makes space another source of variation. We collected the data in multiple years; therefore, time is third source of variation. To make it simpler, space and time were combined and are called environment. Therefore, at this stage of the analysis, we have two sources of variation, i.e., environment (space and time) and hybrid (genetics). How much of the variability come from genetics and environment can be estimated in two different ways. One way is by fitting a random yield model, with both environment and hybrids as random effects, and estimating the percentage of variance explained by environment, hybrid, and residual (g x e).

σ^2 $\left\{ \begin{array}{ll} \sigma_e^2 & \text{environmental variance} \\ \sigma_g^2 & \text{genetic variance} \\ \sigma_{eg}^2 & \text{residual (g x e)} \end{array} \right.$

Table 2. The mean, the standard deviation, coefficient of variation, % of the variation explained by environment, genetics, and genetics by environment interaction for irrigated and dryland corn and sorghum yields in Kansas

Statistics		Dryland		Irrigated	
		Corn	Sorghum	Corn	Sorghum
mean		126.9	104.5	204.4	143.5
Stand Deviation		46.6	35.8	41.7	33.5
% of the deviation explained by	Environment	92.0	87.3	86.2	77.5
	Genetics	1.5	3.3	3.1	10.5
	GbyE	6.5	9.3	10.7	12.0
Confidence interval upper limit		218.2	174.7	286.0	209.5
Confidence interval lower limit		35.6	34.3	122.6	78.1

Table 2 show the variability explained by environment and genetics. The variability in yield that is explained by environment is much higher than that of the variability in yield that is explained by either genetics or genetics by environment interaction. The percentage of variation due to environment (sensitivity to environmental changes) is higher in corn than in sorghum.

The second way to achieve the same result is by fitting yield as function of fixed environment and random hybrid effects. We can call this environment specific model.

$$Y_{ij} = \mu_i + H_j + \varepsilon_{ij} \qquad (1)$$

Y_{ij} is the yield at environment i from hybrid j; μ_i mean for the i^{th} environment ; H_j is differential random effect of j^{th} hybrid; and εij is the residual or interaction effect of j^{th} hybrid and i^{th} environment

Figure 3 shows the relationship between observed yield and fitted values of model (1). This environment specific model explained the variability in yield with coefficient of determination (R-square) value of 0.92 for dryland corn and R-square value of 0.87 for dryland sorghum. This is value is equivalent to the variability in yield explained by environment for dryland corn and dryland sorghum, respectively, presented in by fitting random environment and hybrid model (Table 2). In the next section of this paper, we will try to model dryland yield with few of the environmental factors that, perhaps, are underlying reasons for yield variation in space and time. This model, environment specific model (1), sets the maximum variability that we can explain with environmental factors if we were able to measure all of them and if we understood the way they interact to affect yield. But let us start by explaining the underlying factors in space and time that affect yield.

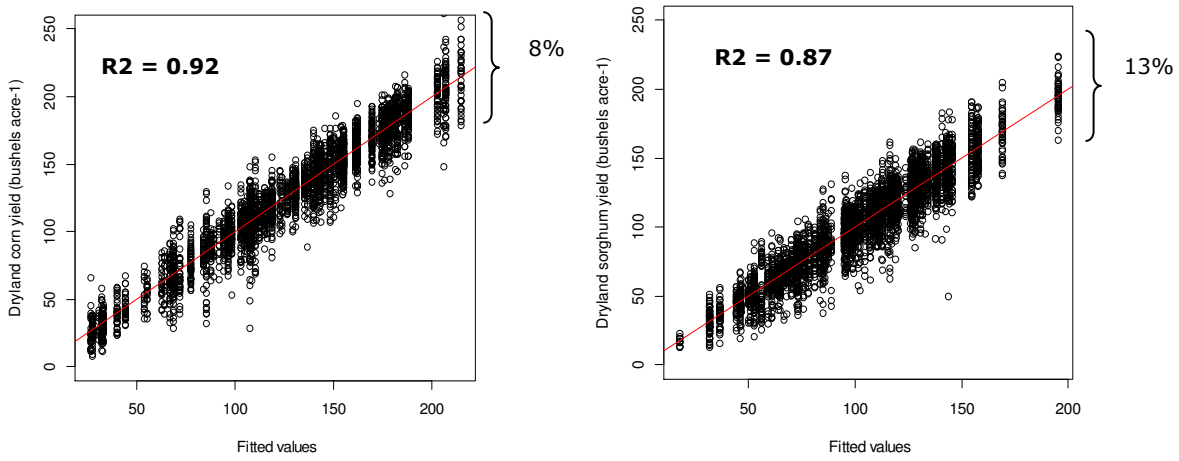

Figure 3. The relationship between observed dryland crop yield and fitted values of a model with fixed environment and random hybrid effects (environment specific model).

Environment is a combination of space and time. Space and time can also be further divided into factors that are the direct reasons for variability in yield but are embedded in space and time. Figure 4 shows a space-time diagram with these major factors responsible to yield variability. When we analyze both space and time, they come with variation in climate, resource, management, and genetics. The extent of how these variables change in time and space depends on how far or how long we travel in space and time, respectively. Some factors such as elevation, slope, soil type and depth might not change in time, at least in short period. Also, it should be noted that this is not a complete list of factors that vary in space and time and affect yield. It is an example and each factor by itself is broader than they look in the diagram. In the following section of this paper we will model yield with one factor from few of the environmental factors. Finally we will try to combine all measured environmental factors to model yield and compare how much of the variability we can explain.

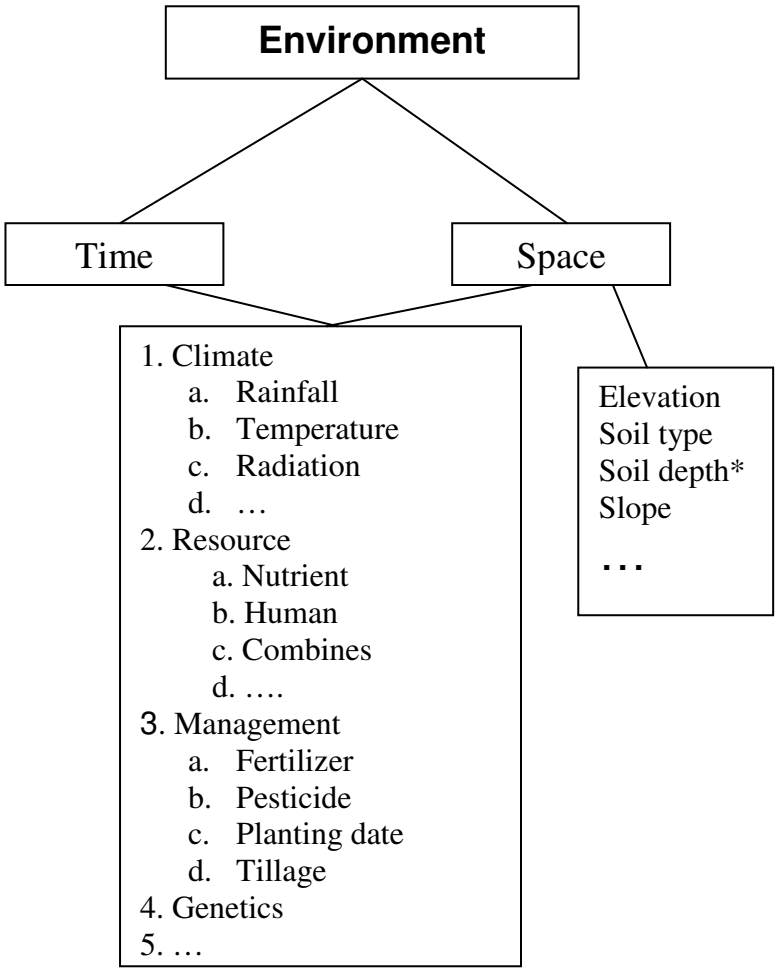

Figure 4. A conceptual diagram of environmental factors that vary in space and time and considered the underlining causes of yield variability.

Yield as function of seasonal rainfall

Crop yield, particularly dryland yield, is highly dependant on the amount of rainfall in the growing season. The relationship between dryland yields of corn and sorghum and total April to September rainfall is presented in figure 5. Here we have used a robust local weighted scatter smoothing regression technique to fit the data. Both dryland corn and sorghum yields seem to have a non linear relationship with total rainfall. Both crops yield increase with different slopes from approximately 5 to 27 inches of rainfall and no changes or decrease afterwards.

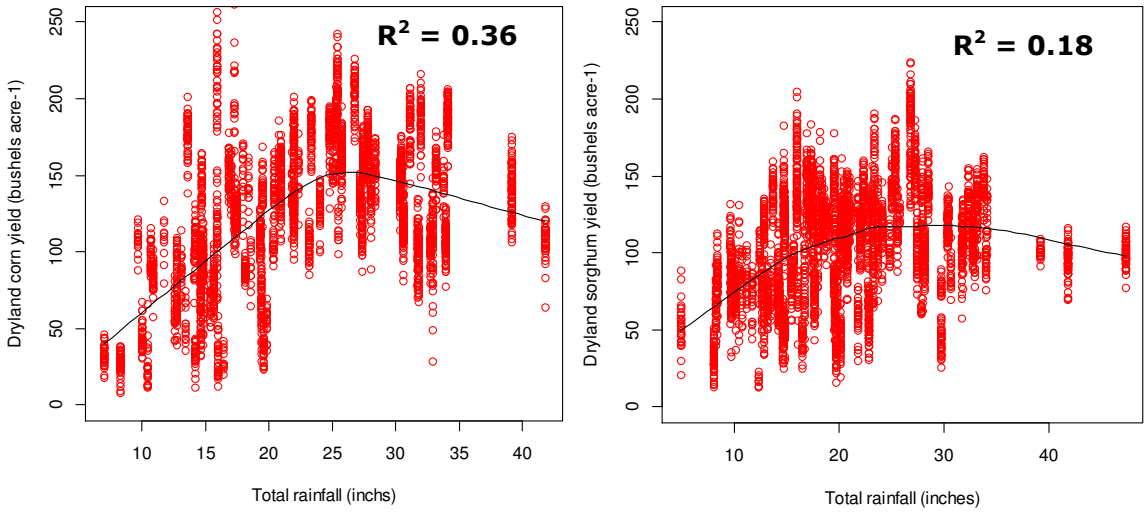

Figure 5. Dryland corn and dryland sorghum as a function of total seasonal (April to September) rainfall.

Intuitively, it is correct to assume total rainfall is not the best variable to model yield. For example, if a crop stands without a rain for the first one month in the season and gets plenty of water after that, the damage in the rainless month might be not recoverable. Therefore, modeling yield using average monthly rainfall may better describe the relationship between rain and yield than the total rainfall of season (Figure 6).

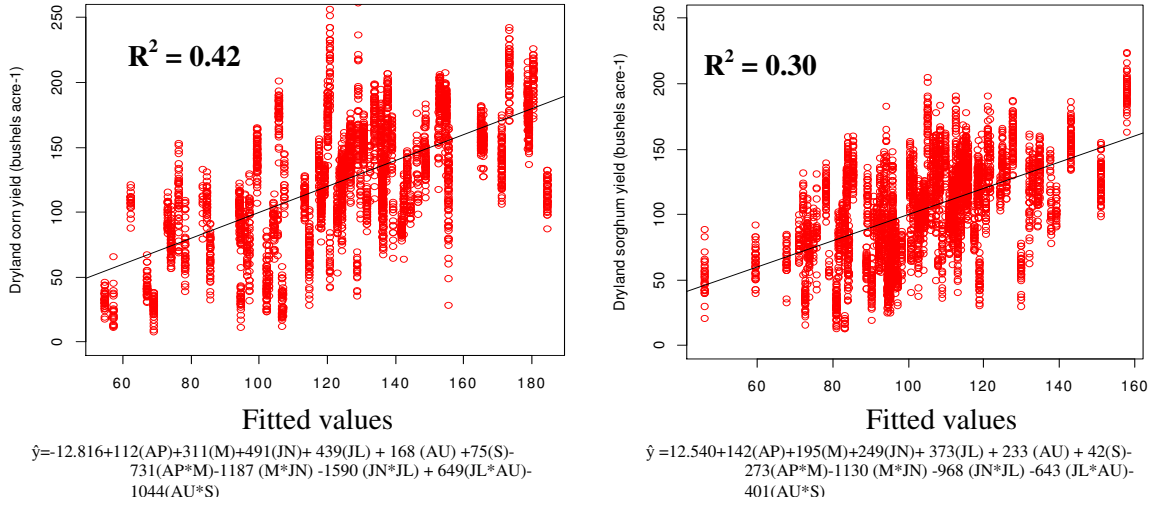

Figure 6. Observed dryland corn and sorghum yield against fitted values of a model that linearly predicts yield with daily average monthly rainfall and interaction of two consecutive months.

The regression equation in Figure 6 presented the coefficients for daily average monthly rainfall for the months April to September and interaction effect of two consecutive months. From the regression equation we can infer that the magnitude of impact of rain of each month on crop yield varies. However, we are still far from explaining the total variability in yield that is due to environment in either of the rainfall models above.

Yield as function of seasonal temperature

Seasonal temperature is also a very important factor in determining crop yield. In fact, crops are grouped into summer and winter crops based on their requirement of seasonal temperature. One of the reasons that corn is planted in early April and sorghum is planted late in May is due to requirement in minimum temperature for germination and growth. Figure 7 illustrates how a linear model that relates yield with the effect of average temperature of each month explains dryland corn and dryland sorghum yields. From the coefficients of the model we can infer that the impact of increasing average temperature on yield might be positive or negative depending on when it happens.

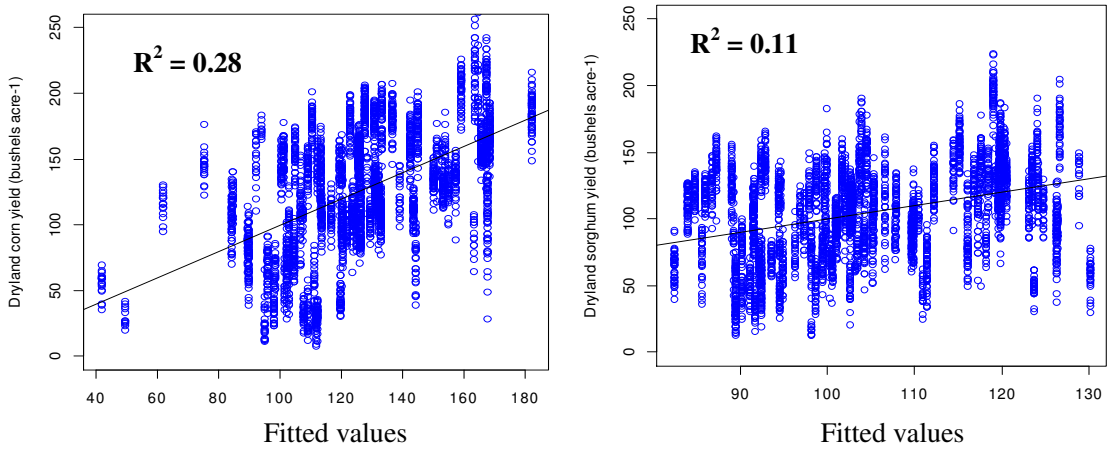

Figure 7. Observed dryland corn and sorghum yields against fitted values of a model that linearly predicts yield from average temperature of the months April through September.

The relationship between yield and temperature can also be further improved by splitting temperature into its minimum and maximum temperature components.

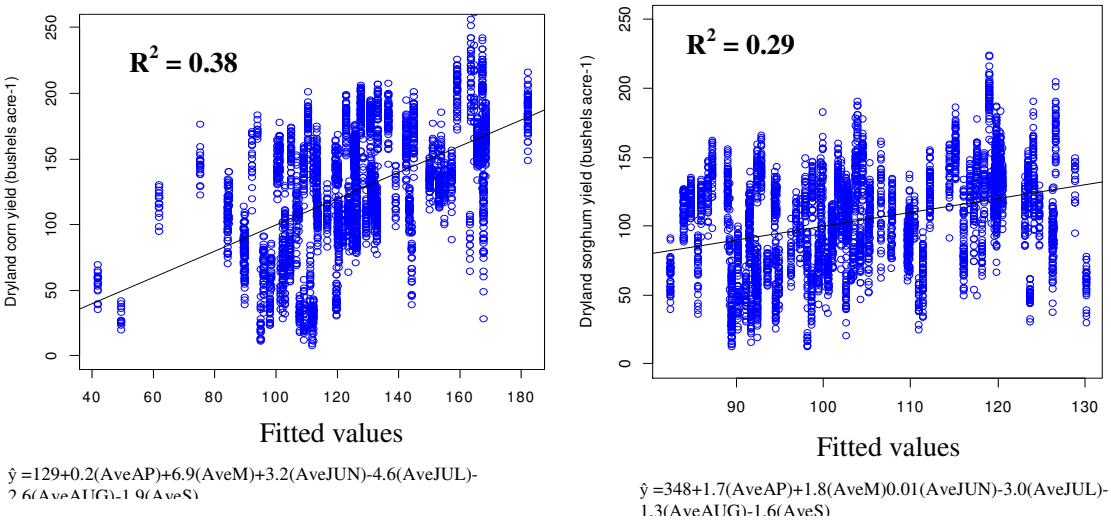

Figure 8. Observed dryland corn and sorghum yields against fitted values of a model that linearly predicts yield from average monthly maximum and minimum temperature of the months April through September.

The logical reasoning behind splitting temperature into minimum and maximum is that both how low the temperature drops and how high it might get that matter more important than the average for growth and development (Fig. 8). This model improved the relationship between yield and temperature but its model fit was much lower than total variability that should be explained by environmental factors.

Yield as function of management factors

Crop management factors such as growing season length and amount of fertilizer applied relate to crop yield. Figure 9 presents the relationship between dryland corn and sorghum yields as function of length of the growing season. Length of growing season was calculated by counting the number of days between date of planting and date of harvest.

As can be seen in Figure 9, dryland corn was more responsive to growing season length than dryland sorghum.

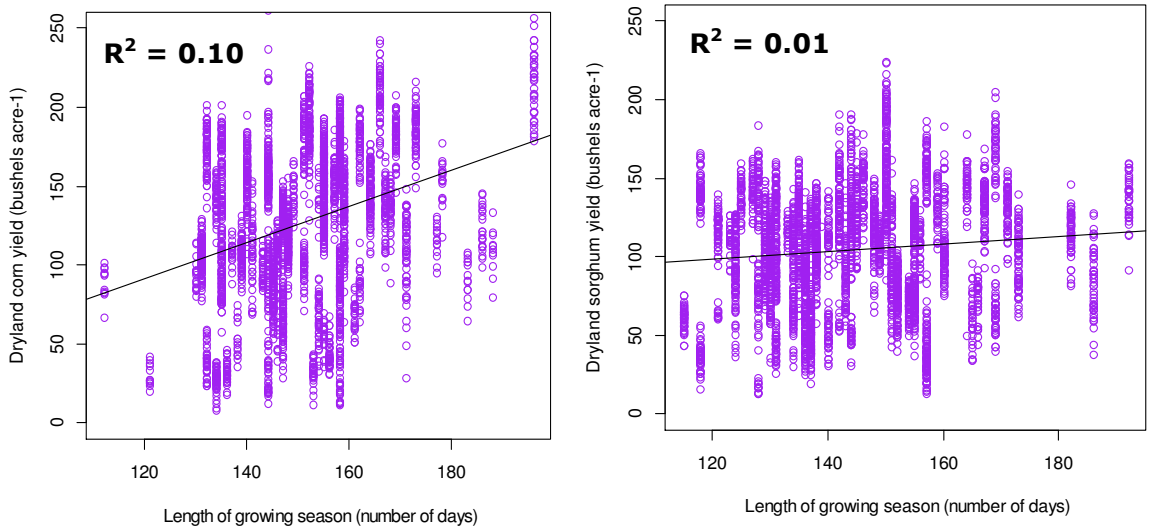

Figure 9. Dryland corn and sorghum yields as a function of the length of growing season.

Yield as function of rainfall, temperature, and management factors

We might further improve the yield functions by assuming that the relationship between yield and environmental factors is linear with a combination of these environmental factors. However, when we combine all of these variables together, multicolliniarity might be a problem. For that reason, the correlation between these factors was investigated before modeling (Table 3). Here, we defined a strong correlation between two variables when the Pearson Correlation Coefficient is above 60% and then we modeled yield with factors whose correlation is less than 60%. This can be achieved with different combination of the factors in Table 3. The Variance Inflation Factor (VIF) is also calculated for each model to check whether multicolliniarity is problem or not. Figure 10 and 11 depicts the model fit for the different combinations of these factors for

dryland corn and sorghum, respectively. Table 4 presents VIF values for explanatory variables for the fitted models.

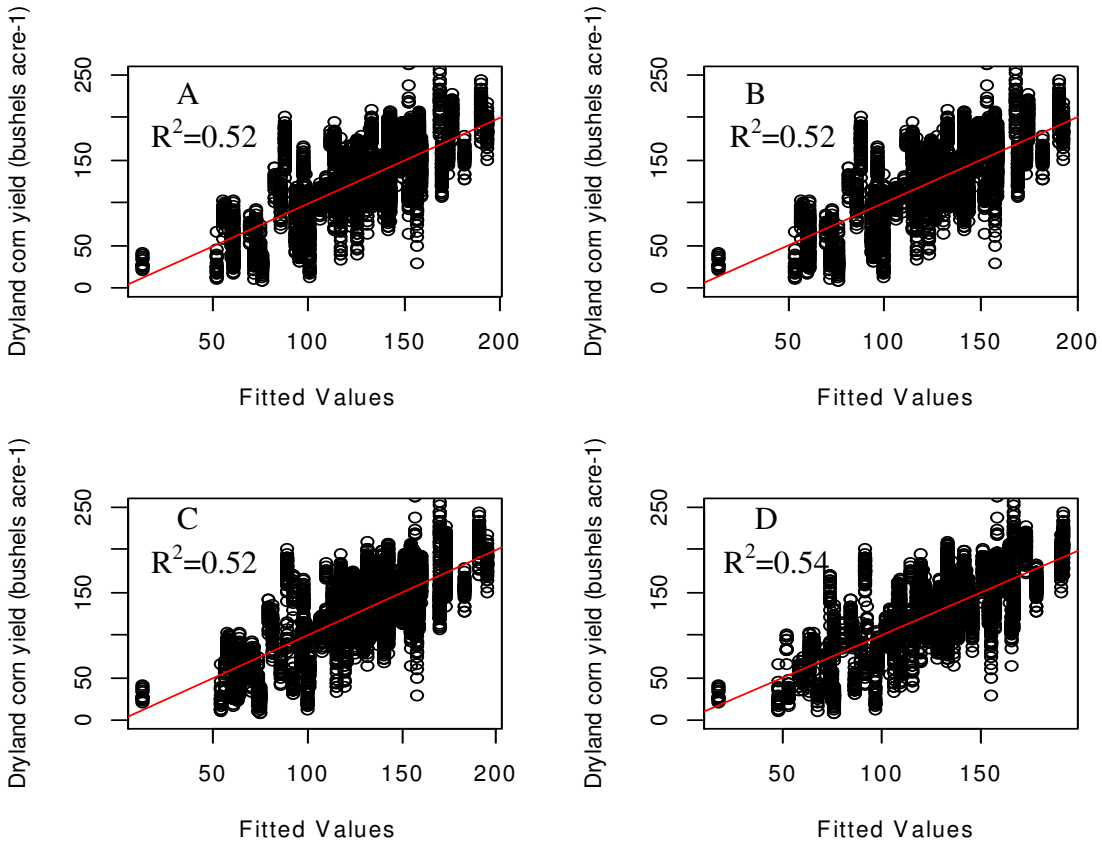

Figure 10. Observed dryland corn yield against fitted values of a model that linearly predicts yield from climatic and environmental factors in table 3 but to avoid multicolliniarity in (A) all minimum temperatures are not include except September minimum temperature, in (B) all minimum temperatures are not include except May and September minimum temperatures, in (C) all minimum temperatures are not included except June and September minimum temperatures, (D) contains all factors and it suffers multicolliniarity.

Table 3. Pearson correlation coefficient between the continuous explanatory variables of dryland yield in Kansas

Variables	LGS	N	Maximum temperature						Minimum temperature						Rainfall					
			APR	MAY	JUN	JUL	AUG	SEP	APR	MAY	JUN	JUL	AUG	SEP	APR	MAY	JUN	JUL	AUG	SEP
LGS	1.00																			
N	0.28	1.00																		
Aprtmax	-0.10	-0.07	1.00																	
Maytmax	-0.13	0.04	0.43	1.00																
Juntmx	0.03	-0.04	0.39	0.18	1.00															
Jultmax	-0.22	0.00	0.34	-0.02	0.49	1.00														
Augtmax	-0.47	-0.11	0.20	0.12	0.05	0.51	1.00													
Septmax	-0.31	0.06	0.01	0.45	-0.18	0.05	0.38	1.00												
Aprtmin	-0.18	-0.20	**0.69**	0.27	0.02	0.01	0.15	0.02	1.00											
Maytmin	-0.19	-0.16	0.30	0.40	-0.23	-0.35	0.24	0.18	**0.71**	1.00										
Juntmin	-0.09	-0.26	0.26	0.05	0.14	-0.25	0.03	-0.09	**0.68**	**0.68**	1.00									
Jultmin	-0.32	-0.31	0.37	0.13	0.00	0.24	0.39	0.03	**0.74**	**0.65**	**0.64**	1.00								
Augtmiin	-0.37	-0.19	0.17	0.06	0.00	0.19	**0.69**	0.18	0.49	**0.64**	0.58	**0.75**	1.00							
Septmin	-0.15	-0.15	-0.02	0.23	-0.19	-0.29	0.08	0.57	0.43	0.58	0.58	0.44	0.47	1.00						
AprilRF	-0.17	-0.21	-0.05	-0.16	-0.05	-0.20	-0.02	-0.29	0.28	0.17	0.37	0.18	0.11	0.02	1.00					
MayRF	-0.18	-0.21	-0.13	-0.28	-0.30	-0.06	0.30	0.07	0.22	0.44	0.37	0.48	0.55	0.33	0.07	1.00				
JunRF	-0.11	-0.25	-0.06	0.06	-0.23	-0.45	-0.13	0.02	0.23	0.33	0.41	0.17	0.14	0.41	0.17	0.24	1.00			
JulRF	0.12	-0.18	-0.06	0.06	-0.20	-0.58	-0.37	-0.02	0.11	0.29	0.21	-0.01	-0.14	0.19	0.14	0.12	0.35	1.00		
AugRF	0.07	-0.16	0.36	-0.04	0.25	0.03	-0.20	-0.24	0.44	0.07	0.27	0.16	0.03	0.06	0.19	-0.10	0.17	0.13	1.00	
SepRF	-0.03	-0.32	-0.10	-0.09	-0.02	-0.20	-0.14	-0.23	0.22	0.22	0.48	0.40	0.19	0.34	0.25	0.22	0.50	0.22	0.20	1.00

Table 4. **Variance Inflation Factor** for explanatory variables in models A, B, C, and D in figure 11 and 12. As rule of thumb models with VIF> 10 could be considered to seriously suffer from multicollinarity problems. NI in the table refers to parameter not included.

Variables	LGS	N	Maximum temperature						Minimum temperature						Rainfall					
			APR	MAY	JUN	JUL	AUG	SEP	APR	MAY	JUN	JUL	AUG	SEP	APR	MAY	JUN	JUL	AUG	SEP
Model A	1.7	1.1	2.1	2.3	1.6	3.4	2.5	4.5	NI	NI	NI	NI	NI	3.4	1.3	1.8	2.1	1.7	2.0	2.2
Model B	1.7	1.1	2.6	5.8	1.7	4.8	3.8	7.6	NI	**10.2**	NI	NI	NI	6.2	1.3	2.8	2.3	1.8	2.0	2.1
Model C	1.7	1.2	2.4	2.3	2.1	4.0	2.6	5.5	NI	NI	6.3	NI	NI	6.5	1.6	2.0	2.1	1.8	2.0	2.2
Model D	1.9	1.3	7	7	3.7	9.1	**11.2**	**10.5**	**17.2**	**14.7**	9.4	**10.7**	**12.7**	**11.7**	2.1	3.0	2.7	2.1	2.7	2.7

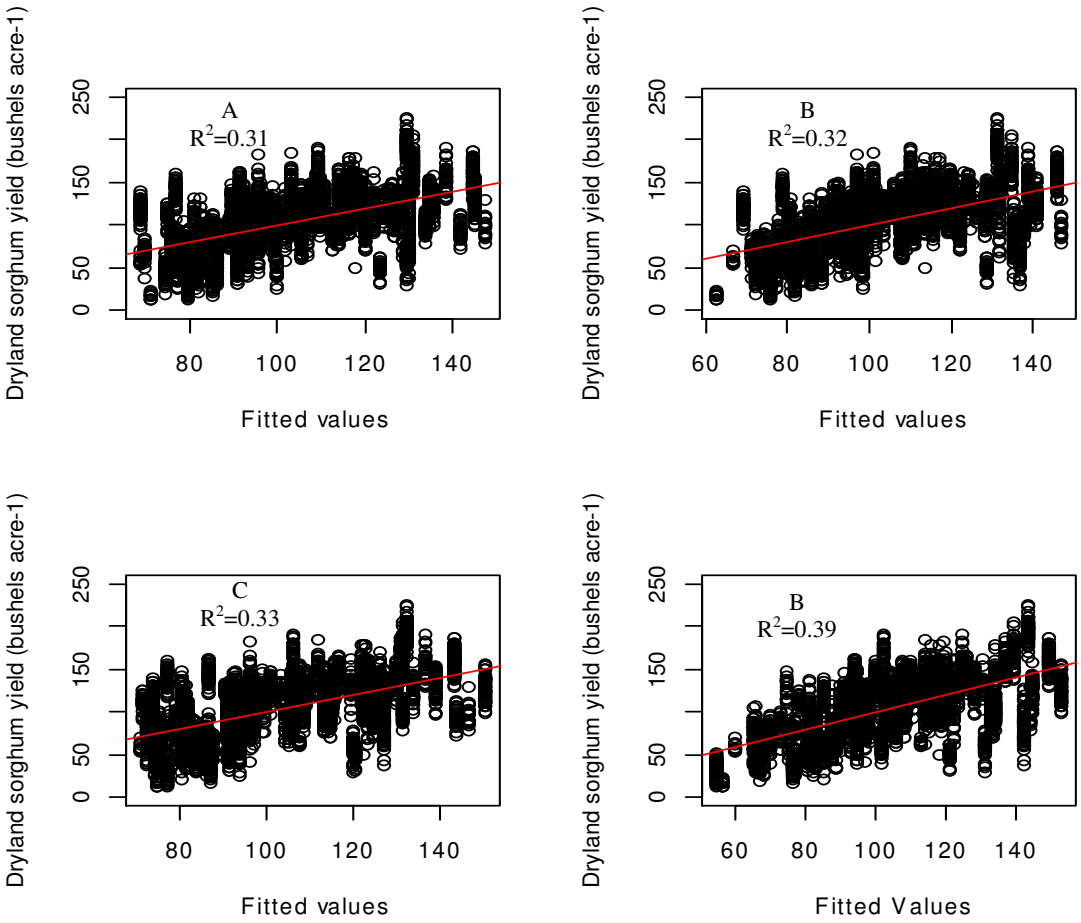

Figure 11. Observed dryland sorghum yield against fitted values of a model that linearly predicts yield from climatic and management factors in table 3 but to avoid multicolliniarity in (A) all minimum temperatures are not included except September minimum temperature, in (B) all minimum temperatures are not included except May and September minimum temperatures, in (C) all minimum temperatures are not included except June and September minimum temperatures, (D) contains all factors and it suffers multicolliniarity.

From the results in the two figures (10 and 11), we noted a better model fit when all factors come together. However, even these models that contain the most relevant environmental factors did not fit to the level where the first environmental specific model did. The most probable reason (claim) being that the information contained in space and time is richer than what five or six environmental factors can explain. We can strengthen this claim by including a space time adjustment on the best of the models above.

Space-Time Adjustment: As Explanatory Variable

Among the reasons why our models above did not have the best fit could be lack of additional information contained in space and time but not measured. Therefore, we might benefit by a systematic space and time adjustment. This adjustment should be conducted in such a way that we do not over fit the model. To do so, we first grouped every five years of our data set, the years 1992 to 2009, into about 4 time lines and called them lustrum, i.e. lustrum 1(1992-1996), lustrum 2 (1997-2001), lustrum 3(2002-2006), and one partial 4th lustrum (2007-2009). Then we included a categorical space information, county, and continuous time information, lustrum, on the best of the models above. Figure 12 depicts how this model fits our data. Using a space and time adjustment, we were be able to explain about 70% of the variability out of 92% possible for corn and 54% of the variability out of 87% possible for dryland sorghum.

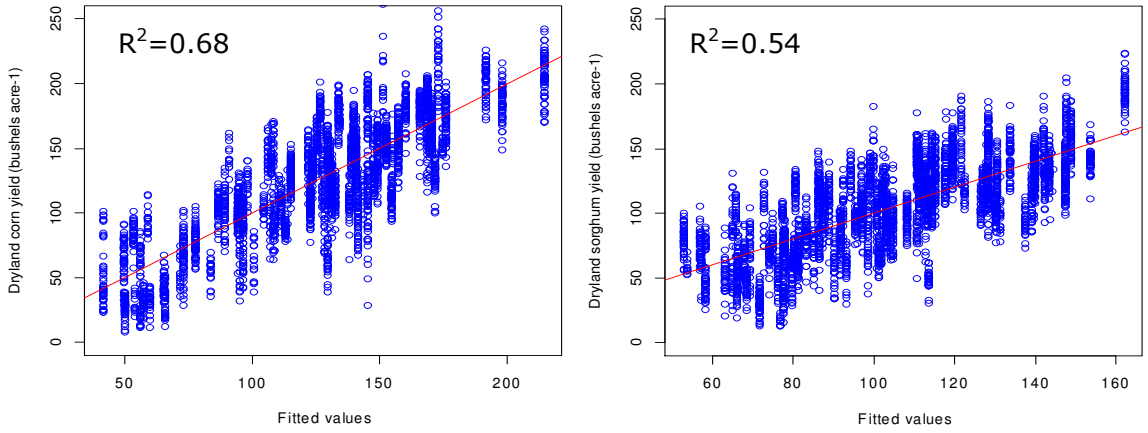

Figure 12. Observed dryland sorghum yield against fitted values of a model that linearly predicts yield from climatic and management factors (model A in figure 11 and 12) plus space and time adjustment.

The model fit for this model with combination of management and climatic factors and space time adjustment is depicted in Figure 13. The residuals were normally distributed with mean zero (Fig. 13). There was no significant trend or correlation noted on the residuals over space or time.

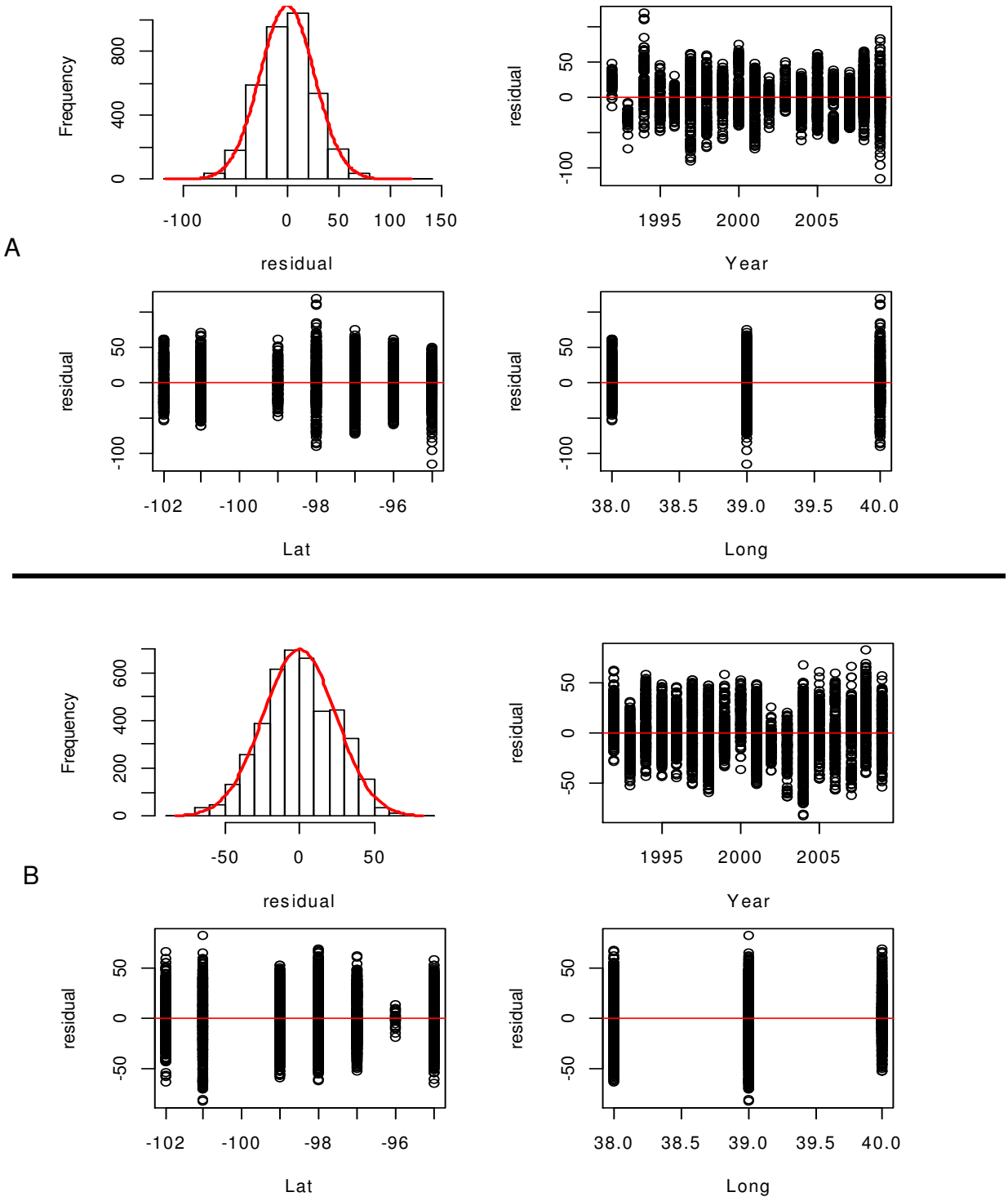

Figure 13. Residual plots of dryland corn (A) and dryland sorghum (B) models that linearly predict yield from climatic and management factors (model A in figure 10 and 11) plus space and time adjustment. Lat and Long in the graphs refer to latitude and longitude.

Space-Time Adjustment: On the Residual Covariance Structures

The assumption of normality, independence, and constant variance seem to hold in our models (Fig. 13). This was possible perhaps because we captured, not all, but the most important variables responsible for variability in yield. However, our response variable yield is spatially and temporally dependent. Here we demonstrated the spatial and temporal dependence of yield using USDA dryland corn data. In order to study the spatial dependency of yield, the average dryland corn yield for all 105 counties of Kansas for the time period 1992-2009 was used. Figure 14 depicts dryland corn yield map of Kansas, semivariogram cloud, and semivariogram curve of dryland corn. From the semivariogram curve in Figure 14, it is clear that dryland corn yield of counties that are closer in distance are more similar (dependant). This is a clear indicator of spatial dependence in crop yield.

In order to study the temporal dependency of yield, the average dryland corn yield reported for Kansas for the years 1972 to 2011 by USDA was studied. Figure 15 depicts the time series plot for yield, annual yield growth, and the autocorrelation and partial autocorrelation for the de-trended (annual) yield. The autocorrelation and partial autocorrelation graphs indicate a significant temporal correlation with time lag 1.

Therefore, this spatial and temporal dependence in yield might be reflected in error terms in situations where the explanatory variables available in model did not capture the variability. If this is realized, modeling of crop yield should be done by spatial, temporal, or spatio-temporal adjustments made mainly on the residual covariance. When there is only a spatial or temporal dependence, they can be accounted in model with spatial or temporal correlated error variance

through generalized least square regression approach as oppose to ordinary regression (Direnzo et al., 2000; Anselin, 1988; Anselin, 2007). Similarly, when there is spatio-temporal dependence, model should be adjusted for both space and time (Anselin et. al., 2008; Demel and Du, 2012; Gneiting, 2002; Millo and Piras, 2012).

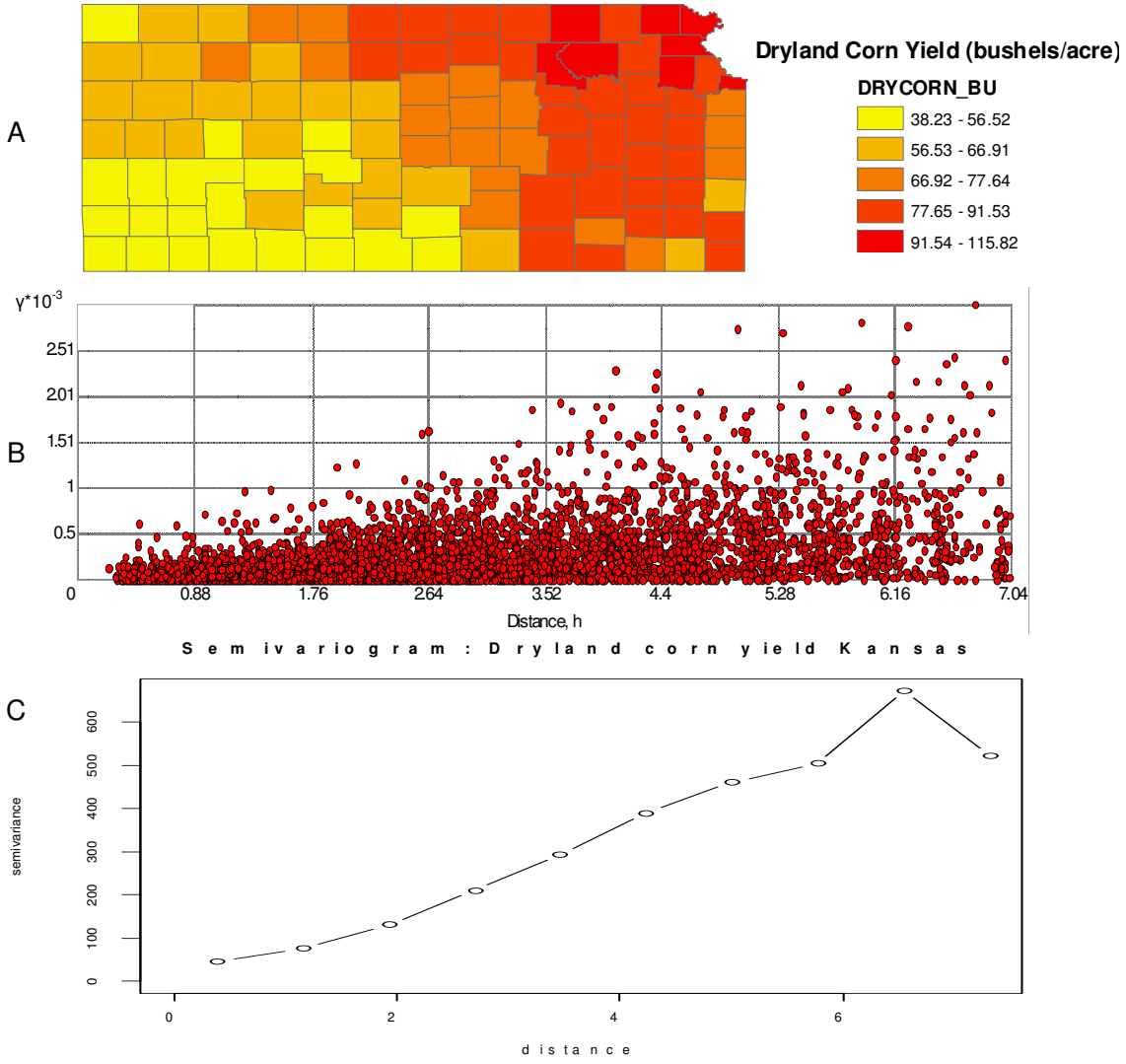

Figure 14. (A) Dryland corn yield in Kansas, (B) semivariogram cloud for dryland corn yield in Kansas, and (C) a semivariogram curve.

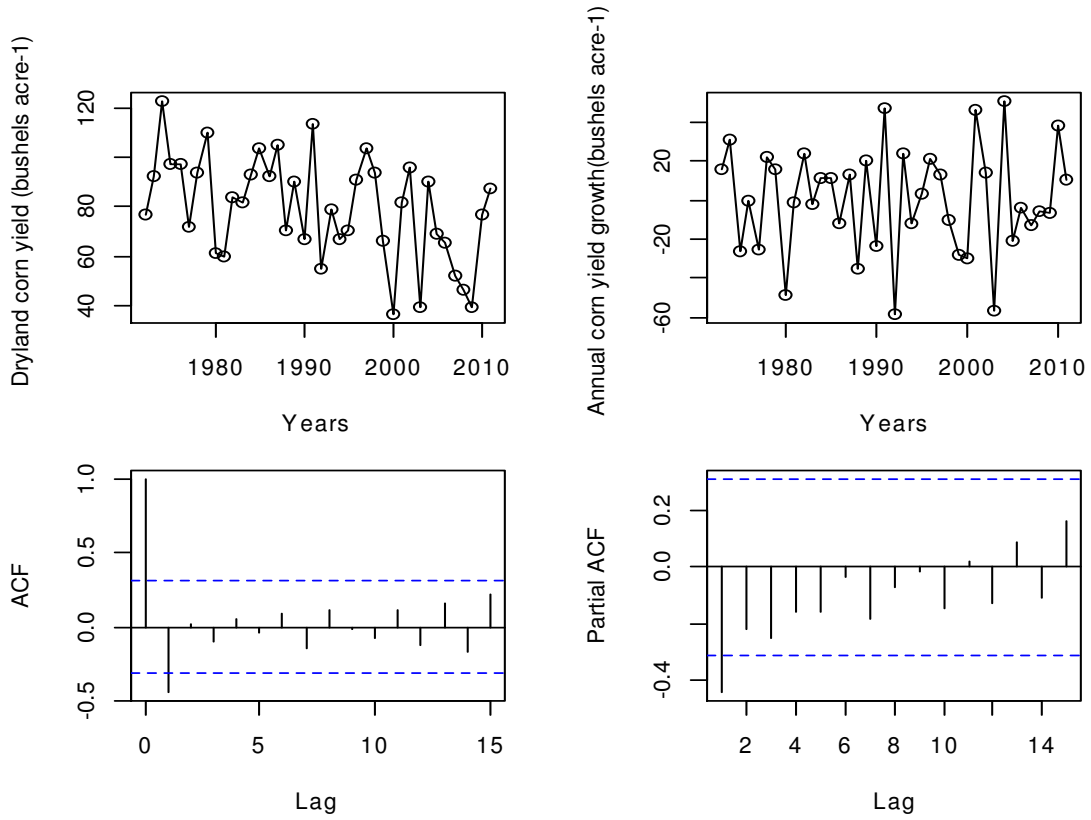

Figure 15. Time series plot of dryland corn yield in Kansas, de-trended yield (annual yield growth), autocorrelation (ACF), and partial autocorrelation (PACF) of annual yield for the years 1972 to 2011.

SUMMARY AND CONCLUSION

Determining the distribution of crop yield has been a research topic by many economists and statisticians in the past and has resulted in mixed conclusions of normality versus non-normal distribution (Just and Weninger, 1999; Nelson and Preckel, 1989; Ramírez, 1997). Determining the distribution of crop yield was not main objective of our research but was an important assumption in least square estimation. Our analysis of the frequency distribution of yield in the

present research data set show that the assumption that dryland and irrigated corn and sorghum yields are approximately normally distributed seem to hold.

The variability in yield that is explained by environment is much higher than the variability that is explained by genetics within each cropping system. This result signifies how environment plays a significant role in determining crop yield. This further emphasizes the notion that crop recommendation should be environmental specific. Corn yield variation explained by environment is much higher than sorghum. A relative stability of sorghum yields across environmental variations compared to corn yields in this analysis support previous findings (Boyer 1970; Beadle, 1973; Stone et al., 1996; Fischer et al., 1982). Obviously, dryland yields are relatively more environmental dependant than irrigated yields just because one environmental factor, water, is less of a limitation in irrigated systems.

Table 5. The AIC, BIC, and R^2 values for seven models and relationships fitted to dryland corn and sorghum yield

Model	AIC(x10000)	BIC(x10000)	R^2	AIC(x10000)	BIC(x10000)	R^2
		Dryland corn			**Dryland sorghum**	
Space-time	2.84	2.89	0.92	3.35	3.42	0.87
Total Rainfall	3.61	3.61	0.36	4.16	4.16	0.18
Monthly RF	3.57	3.57	0.42	4.16	4.17	0.30
Average Temp.	3.65	3.65	0.28	4.18	4.19	0.11
Min and Max Temp.	3.59	3.60	0.38	4.09	4.10	0.29
Length of GS	3.72	3.73	0.10	4.23	4.23	0.01
All factors	3.50	3.52	0.52	4.07	4.08	0.36
All factors and space-time adj.	3.36	3.37	0.68	3.91	3.93	0.54

Here we have done an exploratory analysis of different yield functions. We started with an environmental specific model. Then we break environment into its components and looked at relationships between yield and different environmental factors. The model goodness of fit for the seven models is given in Table 5. For example, we have used rainfall, which is the major environmental factor that is known to affects crop yield. However, total seasonal rainfall

explains only a portion of the variability in dryland corn and dryland sorghum yields. The fitted values from the robust locally weighted scatter smoothing regression that describe the relationship between rainfall and dryland yields are presented in figure 16. As can be seen in the equation, the conclusion on relationship between yield and rainfall varies depending on where we are in range of rainfall value in our data set and on the type of crop we have. That is why there is a mixed conclusion in the literature about the effect of increasing rainfall, temperature, or other factors on crop yield.

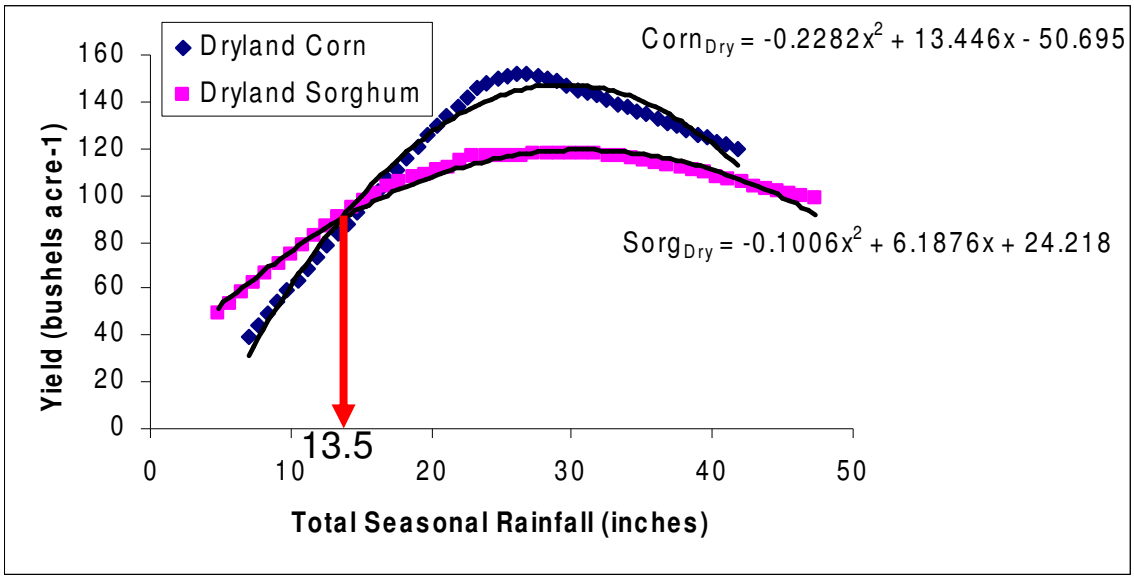

Figure 16. Weighted means and fitted equations for the relationship between total rainfall and corn and sorghum yields.

In the present study we also demonstrated the different ways that we might model yield with one factor , i.e., total rainfall versus monthly rainfall with interaction terms; using monthly average temperature versus monthly minimum and maximum temperature. The effects of management factors such as length of growing season on yield were also explored. The combined effect of these factors explained yield much better than each of the individual factors. However, the variation that was explained by environment specific model was out of reach of the individual or

combined factors model. From this we can deduce that either the number of variables in the model or the way they interact in these models was not enough to explain the possible yield variability explained by environment specific model. This is always the case on variables like yield that can be impacted with multi factors.

Time and space adjustments are made as an explanatory variable in best of our models and witnessed a significant improvement in model fit. In our case, the assumptions on error seem to hold, however, we commented on cases when the assumption of independence on error did not hold. In such an occurrence, modeling corn yield using spatial, temporal, and spatiotemporal adjustments on the covariance of residuals was suggested. In conclusion, our analysis indicated the importance of space and time components of data sets because they can often adjust (make up) for those underlying variables and factor effects that are not measured or not well understood.

REFERENCE

Anselin, L. 1988. Spatial econometrics: methods and models. Kluwer Academic Publishers, London.

Anselin L. 2007. Spatial regression analysis in R: A work book. Available at https://geodacenter.asu.edu/system/files/rex1.pdf.

Anselin L, Le Gallo J, Jayet H. 2008. Spatial Panel Econometrics. In L Matyas, P Sevestre (eds.), The Econometrics of Panel Data : Fundamentals and Recent Developments in Theory and Practice, pp. 624{660. Springer-Verlag.

Beadle, C.L., K.R. Stevensor, H.H. Neumann, G.W. Thurtell and K.W. King. 1973. Diffuse resistance, transpiration, and photosynthesis in single leaves of corn and

sorghum in relation to leaf water potential. Can. J. Plant Sci. 53:537-544.

Boyer, J.S. 1970a. Differing sensitivity of photosynthesis to low leaf water potentials in corn and soybeans. IRRI, Philippines. Plant Physiol 46:819-820.

Compton, L L. 1943. Relationship of moisture to whaet yields on western Kansas farms. Kansas Agr. Exp. Sta. Circ. 168.

Demel, S. S. and Du, J. 2011. Spatio-temporal covariance modeling with some ARMA temporal margins, Conference for Applied Statistics in Agriculture Proceedings, vol. 23.

Deschenes, O., and M. Greenstone. 2007. "The Economic Impacts of Climate Change: Evidence from Agricultural Output and Random Fluctuations inWeather." American Economic Review 97(1):354–85.

Dirienzo C, Fackler P., Goodwin B.K. 2000. Modeling spatial dependence and spatial heterogeneity in county yield forecasting models. AAEA conference, Tampa, FL.

Fisher, K.S., E.C. Johnson, and G.O. Edmeades. 1982. Breeding and selection for drought resistance in tropical maize. In Drought resistance in crops with emphasis on rice. Intern. Rice Res. Inst. Laguna, Philippines.

Gneiting, T. 2002. Nonsepable, stationary covariance functions for space-time data. J. Amer. Stati. Asso. 97: 590-600.

Gumpertz, ML, Rawlings, JO . 1992. Nonlinear regression with variance components: Modeling effects of ozone on crop yield. Crop Sci. 32:219-224.

Just RE, Weninger Q. Are Crop Yields Normally Distributed? American Journal of Agricultural Economics. 1999;81:287–304.

Kaylen, Michael S., and Suffyanu S. Koroma. "Trend, Weather Variables, and the Distribution of U.S. Corn Yields." Review of Agricultural Economics, Vol. 13,

No. 2, 1991: 249-258.

Lobell D.B., Cassman K.G., and Field C.B. 2009. Crop yield gabs: their importance, magnitude, and causes. Annu. Rev. Environ. Resour. 2009. 34:179–204.

Lobell, David B., and Gregory P. Asner. "Climate and Management Contributions to Recent Trends in U.S. Agricultural Yields." Science, New Series Vol. 299, No. 5609, 2003: 1032.

Lobell, D.B., M.B. Burke, C. Tebaldi, M.D. Mastrandrea, W.P. Falcon, and R.L. Naylor. 2008. "Prioritizing Climate Change Adaptation Needs for Food Security in 2030." Science 319: 607- 610.

Lobell D. 2010. Crop Responses to Climate: Time-Series Models . In Climate Change and Food Security. Lobell and Burke (eds.). Springer: New York, 123–140.

Machado S., Bynum E. D., Archer T. L., Lascano R. J., Wilson L. T, Bordovsky J., Segarra E., Bronson K., Nesmith D. M., and Xu W. 2002. Spatial and Temporal Variability of Corn Growth and Grain Yield: Implications for Site-Specific Farming. Crop Sci. 42:1564–1576 (2002).

Mathews, O.R., and Brown, L.A. 1938. Winter whaet and sorghum production in the southern Great Plains under limited rainfall. USDA Circ. 477.

McCarl, B.A., X. Villavicencio, and X. Wu. 2008. "Climate Change and Future Analysis: Is Stationary Dying?" American Journal of Agricultural Economics 90(5): 1241 1247.

Millo G. and Piras G. 2012. splm: Spatial Panel Data Models in R. J. of Statistical Software 47(1).

Nelson CH, Preckel PV. The Conditional Beta Distribution as a Stochastic Production

Function. American Journal of Agricultural Economics. 1989;71:370–378. Nelson, William L., and Robert F. Dale. "Effect of Trend on Technology Variables and Record Period Prediction of Corn Yields with Weather Variables." Journal of Applied Meteorology, Vol. 17, 1978: 926-933.

Oury, Bernard. "Allowing for Weather in Crop Production Model Building." Journal of Farm Economics, Vol. 47, No. 2, 1965: 270-283.

Ozaki, V., Ghosh, S., Goodwin, B. & Shirota, R. 2008. Spatio-temporal modeling of agricultural yield data with an application to pricing crop insurance contracts. American Journal of Agricultural Economics, 90, 951–961.

Perez-Quezada J.F., Pettygrove G. S. and Plant R. E. 2003. Spatial–Temporal Analysis of Yield and Soil Factors in Two Four-Crop–Rotation Fields in the Sacramento Valley, California. Agron. J. 95:676–687.

Ramírez OA. Estimation and Use of a Multivariate Parametric Model for Simulating Heteroskedastic, Correlated, Nonnormal Random Variables: The Case of Corn Belt Corn, Soybean and Wheat Yields. American Journal of Agricultural Economics. 1997;79:191–205.

Schlenker W. and Roberts M J. 2006. Nonlinear Effects of Weather on Corn Yields . Rev. Agric. Econ. 28 (3):391-398.

Schlenker, Wolfram, Michael Hanemann, and Fisher Anthony. The Impact of Global Warming on U.S. Agriculture: An Econometric Analysis of Optimal Growing Conditions. Berkeley: Department of Agricultural And Resource Economics, UCB, UC Berkeley, 2004.

Schlenker, Wolfram, and Michael Roberts. "Nonlinear Temperature Effects Indicate

Severe Damages to U.S. Crop Yields Under Climate Change." Proceedings of the National Academy of Sciences (PNAS), 2009: 15594-15598.

Shaw, Lawrence H. "The Effect of Weather on Agricultural Output: A Look at Methodology." Journal of Farm Economics, Vol. 46, No. 1, 1964: 218-230.

Stone, L. R. A.J. Schlegel, R.E. Gwin, A. H. Khan. 1996. Response of corn, grain sorghum, and sunflower to irrigation in the High Plains of Kansas. Agricultural Water Management 30(3):251-259

Stone L.R. and Schlegel A.J. 2006. Yield–Water Supply Relationships of Grain Sorghum and Winter Wheat. Agron. J. 98:1359–1366

Thompson, Louis M. "Weather Variability, Climatic Change, and Grain Production." Science, New Series, Vol. 188, No. 4188, 1975: 535-541.

Chapter IV

Multivariate Time Series Analysis of Corn Production in the USA

ABSTRACT

A multivariate time series analysis presents an opportunity to analyze the characteristics of variables in time, space, and in relation to other variables. Our objectives were to examine the temporal and spatial characteristics of corn harvest area, price, yield and total production in the US and to build an optimal model for early forecasting using the temporal and spatial characteristics of the data. The annual corn yield, harvest area, and price survey data available in USDA National Statistics Service database collected for the years 1900-2011 was used for the study. Multivariate time series plots, temporal auto-and cross correlation, spatial autocorrelation analysis were completed. Based on the analysis, models were developed and compared. Results suggested that corn harvested area, price, yield, and total production trends in the US varied for the number of years and states considered. All of these variables demonstrated a significant autocorrelation with their value at time t and their value at t-1. A significant cross correlation was also found between price, area, and total production. Base on this auto-and cross correlation analysis result, a vector autoregressive, VAR(1), model was developed for area, price and production. This model proved better in model fit and forecasting qualities than an ad-hoc model. A significant spatial dependence was found for these variables by Moran's I spatial autocorrelation and semivariogram analysis. This spatial dependence information was then used to develop a state based yield forecasting model. The VAR model was capable of using past (t-1) yield values of the state and its neighbors to predict yield at time t. This study demonstrates how data rich in time and space can be used for modeling and early forecasting.

INTRODUCTION

Early crop production forecasting help producers, consumers, researchers, policy makers, and grain marketing agencies in decision making. A timely and accurate crop production forecast help these parties make better decision on crop selection, soil and crop management, marketing, storage, transport, and assessing risk associated with these activities (Hammer et al., 2001; Kantanatha, 2010; Potgieter et al. 2005; Stone and Meinke, 2005; Rasmussen et al., 1998).

Crop production estimates could be pre-harvest estimates that can be forecasted as weather and plant condition data becomes available (as the growing season unfolds) or they can be made even before planting by surveys and other methods. United States Department of Agriculture National Agricultural Statistics Service (USDA NASS), for example, forecasts crop yield based on surveys and crop conditions on a monthly basis (USDA, 2006). Many available crop forecasting methods are dependent on seasonal weather information (Dudley and Hearn, 1993; Hammer et al., 1996; Meinke and Hammer, 1997; Singels and Potgieter, 1997). These seasonal weather dependant forecasting methods are usually relatively accurate but their contribution in terms of making better crop choice and management decision is, obviously, limited. For a better decision making at all stages of crop production process, i.e., starting from selecting a crop for the season, early forecasting methods independent of actual seasonal information are crucial. However, finding a reliable method for early crop production estimate is a challenge.

A chronological sequence of observations on a particular variable results in time series data (Chatfield, 2000). Time series data collected for a reasonably long period of time may reflect an internal structure that contains information on how the variable relates to its past

(autocorrelation) or history of other variables (factors) that have an influence (cross correlation). For example, a corn yield data set that is collected for over a hundred years might reflect information on both the relationship between yield at time (t) and its past (t-k) and information on yield potential of corn, change of yield of corn due to changes over factors like year to year weather variation, genetics (hybrid or variety), nutrient rates and application methods, and management methods that varied over this time period. Therefore, a time series analysis presents a potential to early forecasting (Boken, 2000; Chatfield, 2000; Kumar et al., 2010).

Corn ranks as the number one crop in both area of production and in quantities produced in the U.S. (Elbehri and Paarlberg, 2003; O'Brien, 2010). Corn has an innumerable uses in food, feed, alcohol, and biofuel industries. A time series analysis of corn production and early forecasting, therefore, presents a great opportunity for better decision for corn producers, consumers, researchers, policy makers, and grain marketing agencies in the US.

The main objective of this research was to understand the temporal and spatial characteristics of corn production area, price of corn, corn yield and total corn production in the U.S. and develop a model to forecast these variables as early as possible. We hypothesized that models developed based on multivariate time series analysis of corn from historic data will result in a more effective prediction of the future than ad-hoc approach, which is assuming next year will be almost same as this year.

Here we have presented a general exploratory analysis of corn harvest area, yield, and price in the USA from 1900-2011. A temporal analysis of these variables, i.e., trend analysis, auto-and

cross correlation among area, yield, and price, followed the exploratory analysis. A temporal model was developed to predict area, price, and expected total production a year ahead for the USA. The model was then used to predict observed values, which were not included at model development step, and the prediction was compared with the ad-hoc approach.

As a second approach the spatial characteristics of area, yield, and price was analyzed. The intent of this analysis was to show how information from spatial analysis can be integrated into a multivariate time series approach to modeling and early prediction. A spatial autocorrelation analysis was conducted for the yield and how this information can be used in modeling and prediction at state level was elaborated on using Kansas as an example. We have presented the detail analysis methodology first, then results, and a conclusion at the end.

DATA ASSEBLY AND ANALYTICAL STEPS

The annual yield, harvested area, and price of corn reported by USDA for 41 states (Fig. 1) from 1900 to 2011 was used to explore time and space dependence of corn yield in the USA and to eventually develop models for forecasting (USDA, 2011).

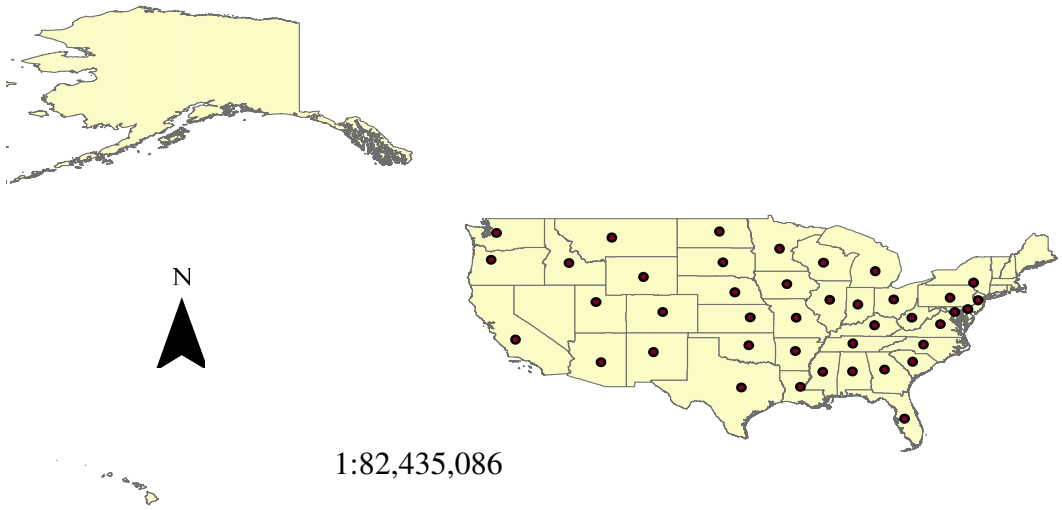

Figure 1. Map of the USA with points indicating the 41 states that corn data for the years 1900-2011 was collected.

An exploratory analysis was conducted by plotting corn yield, harvested area, and price of the 41 states and 112 years, 1900-2011, data in a multivariate time series plot using Mvtsplot (Peng 2008) function in R version 2.15.0 (R Development Core Team, 2012). The multivariate time series plot is an image plot of time series matrix. The value of each time series data is divided into three categories as “low”, “medium”, and “high”. In the plot, these categories are colored purple, grey, and green, respectively. These categories are based on quantiles of the time series value. In our case, the whole data was divided into tertiles with about equal number of values in each tertile so that values can be compared globally (within state or among states at any year). The average (for yield per hectar and price) and the total (for area and total yield) values of the variables are graphed under each color plot to illustrate trends over time.

The changes (trend) in yield, harvest area, and price of corn over time and space were discussed based on the colored plot and graph. The spatial and temporal dependence of each of these variables (corn yield, area, and price) and relationship among them were further studied in ArcGIS 10 (ESRI, 2011), R version 2.15.0, and SAS software as indicated below.

Temporal Analysis for Modeling at National level

To prepare the data from the 41 states for temporal analysis, the annual average yield (Mg ha^{-1}), total harvested area (ha), total corn production (Mg), and average price of corn ($) were calculated for the USA using equations (1) to (4), respectively. A time series graph of these variables (Y_t, A_t, TY_t and P_t) in the USA were then plotted and discussed.

$$Y_t = \frac{1}{n_t}\sum_x Y_{x,t} \qquad (1) \qquad A_t = \sum_x A_{x,t} \qquad (2)$$

$$TY_t = \sum_x Y_{x,t} A_{x,t} \qquad (3) \qquad P_t = \frac{1}{n_t}\sum_x P_{x,t} \qquad (4)$$

Where Y_t, A_t, TY_t and P_t are average yield (Mg ha^{-1}), total area (ha), total yield (Mg), and average price at time t ($); n_t is number of states that report these variables; $Y_{x,t}$, $A_{x,t}$, and $P_{x,t}$ are yield, area, and price of corn at state x and time t, respectively.

Following the discussion on the time series plots, an autocorrelation function (ACF) was calculated and plotted for average yield (Mg ha^{-1}) and annual yield growth of corn (5) for up to 40 year lags. The autocorrelation function shows the relationship between values of a variable at time t with its value at different time lag. The annual yield growth is the difference between yield at time t and yield at time lag 1 (t-1). Calculating annual yield growth was essential to detrend (remove trend) from annual yield and make it a stationary process.

$$R_{s,t} = \frac{E[(X_s - \mu_s)(X_t - \mu_t)]}{\sigma_s \sigma_t} \qquad (5)$$

$R_{s,t}$ = autocorrelation between value of yield (Mg ha^{-1}) at time s and t ; X_s and X_t are yield measured at time s and time t ; μ_s and μ_t are mean yields for time s and t, respectively ; σ_s and σ_t are standard deviation for yield at time s and time t

A cross correlation analysis among the variables harvested area, price, and total corn yield were conducted and plotted for up to 40 years lag (6). Cross correlation analysis describes the relationship between a variables at time t with value of another variable at different time lags.

$$R_{XY(s,t)} = \frac{E[(X_s - \mu_{Xs})(Y_t - \mu_{Yt})]}{\sigma_{Xs}\sigma_{Yt}} \quad (6)$$

$R_{xy(s,t)}$= cross correlation between value of X and Y variables (total yield, area, price) at time s and t, respectively; X_s is the value of X variable measured at time s ; Y_t is the value of Y variable measured at time t ; μ_X is means for the X variable at time s; μ_{Yt} is means for the Y variable at time t; σ_{Xs} and σ_{Yt} are standard deviation for variable X and Y variables at tim s and time t, respectively.

Based on the information on autocorrelation and cross correlations between these variables, obtained from the analysis described above, outcomes of three variables (harvested area, price, and total production) at a time t were modeled using vector autoregressive models. A vector autoregressive model (VAR) is multivariate time series model that utilizes information from variable's own history and history of related variables to predict outcomes in the future (Pfaff, 2008). Since, yield (Mg ha^{-1}) is not a function of area and price (it is function of genetics, weather, inputs, and management) we did not attempt to model and predict yield using information from these variables.

$$\begin{bmatrix} A_t \\ P_t \\ TY_t \end{bmatrix} = \begin{bmatrix} \Phi_{111} & \Phi_{121} & \Phi_{131} \\ \Phi_{112} & \Phi_{122} & \Phi_{132} \\ \Phi_{113} & \Phi_{123} & \Phi_{133} \end{bmatrix} \begin{bmatrix} A_{t-1} \\ P_{t-1} \\ TY_{t-1} \end{bmatrix} + \ldots + \begin{bmatrix} \Phi_{p11} & \Phi_{p21} & \Phi_{p31} \\ \Phi_{p12} & \Phi_{p22} & \Phi_{p32} \\ \Phi_{p13} & \Phi_{p23} & \Phi_{p33} \end{bmatrix} \begin{bmatrix} A_{t-p} \\ P_{t-p} \\ TY_{t-p} \end{bmatrix} + \begin{bmatrix} C_A \\ C_P \\ C_{TY} \end{bmatrix} + \begin{bmatrix} T_A \\ T_P \\ T_{TY} \end{bmatrix} + \begin{bmatrix} e_{A,t} \\ e_{P,t} \\ e_{TY,t} \end{bmatrix} \quad (7)$$

A_t, P_t, and TY_t are area, price, and total yield at time t; Φ_{ijk} are coefficients for the i^{th} time lag, j^{th} predictor (area, price, total yield), k^{th} variable; C_k is constant for k^{th} variable; T_k is a trend for k^{th} variable; $e_{k,t}$ is residual for the k^{th} variable at time t

Since our analysis on the auto and cross correlation between the three variables is different in the years between 1900 and1960 and from 1960-2011, two models were developed at this step. The first VAR model was developed using data from 1900-1999. The data from 2000 to 2011 was left for model testing. A second autoregressive model was developed using information from years 1960-1999. The appropriate time lag that should be included in the models using either of

the data sets (1900-1999 or 1960-1999) was determined automatically in R based on the four model selection criterions, AIC, HQ, SC, FPE. Using the selected time lag, appropriate models were developed. Consequently, the value of each of the variables for the years 2001-2011 was predicted using the models developed. The VAR models were then evaluated based on their prediction of these observed values, i.e., there (R^2).

Spatial Analysis for Modeling at a State Level

To prepare the data for spatial analysis, the data set (41 states by 112 years) was divided into three periods, i.e., period 1 (1900-1939), period 2 (1940-1979), and period 3 (1980-2011). Average total corn yield, average yield per area, average harvested area, and average price for each state at these periods were calculated with equations (8)-(11).The reason for dividing the data set into three periods, rather than averaging the entire period for each state, was chosen because we had information that the variables changed over time and the average of all the 112 years may not have the entire story.

$$Y_{x,p} = \frac{1}{n_y}\sum_{y} Y_{x,y} \qquad (8) \qquad A_{x,p} = \frac{1}{n_y}\sum_{y} A_{x,y} \qquad (9)$$

$$TY_{x,p} = \frac{1}{n_y}\sum_{y} Y_{x,y}A_{x,y} \qquad (10) \qquad P_{x,p} = \frac{1}{n_y}\sum_{y} P_{x,y} \qquad (11)$$

Where $Y_{x,p}$, $A_{x,p}$, $TY_{x,p}$ and $P_{x,p}$ are average yield, total area, total yield, and average price at state s and period p; n_y is number of years at a period; $Y_{x,y}$, $A_{x,y}$, $TY_{x,y}$ and $P_{x,y}$ are yield, area, total yield, and price of corn at state x and year y, respectively.

After calculating the average yield, total area, total yield, and average price at each state for the three periods, using equations (8) to (11), maps of the USA states with value of this variable by period was produced in ArcMap. In the map the value of the variables at a state was categorized into five groups, rather than the exact value for each state, to help visualize spatial characteristics (clustering or dispersion). For a formal confirmation of a spatial characteristics of these

variables, the existence of spatial autocorrelation was analyzed using Moran's I statistics in ArcGIS 10. The equation that was used in ArcGIS to calculate Moran's I statistics for these variables is given in equation (13).

$$I = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (12) \qquad Z_I = \frac{I - E(I)}{\sqrt{V[I]}} \qquad (13)$$

I = Moran's I index; n= total number of states = 41; $w_{i,j}$ = is the spatial weight between state i, and j based on their distance from one another (inverse distance in this case); Z_I = is score of the test statistics; E[I] = is the expected value of Moran's index = $\frac{-1}{n-1} = \frac{-1}{41-1} = -0.025$; V[I] = is the variance of index = $E[I^2] - E[I]^2$

Based on spatial correlation study above and semivariogram analysis, we demonstrated that we can use past information from the state and its neighborhood to model and forecast yield. We used Kansas data as an example. We defined neighbors as those states that share borders. For state of Kansas, for example, Colorado, Nebraska, Missouri, and Oklahoma, fall into this definition of neighbor.

A VAR model was developed for yield. The appropriate time lag that should be included in the models was determined automatically in R, based on the four model selection criterions, AIC, HQ, SC, FPE. Based on the selected time lags, the appropriate models were developed. The VAR model were then evaluated and compared with ad-hoc model.

RESULTS AND DISCUSSION

Corn Yield, Harvested Area, and Price in the USA from 1900 to 2010

The annual yield (Mg ha^{-1}) of corn for 41 states in the years 1900 to 2011 is depicted in Figure 2. Yield was variable among states and there was an obvious increasing trend for yield from 1900 through 2011. A significant yield increase across states seems to have started in the early 1940's but most of states realized it in 1960's.

In the first four decades, corn yield (Mg ha^{-1}) was relatively higher in states of Iowa, Ohio, Illinois, Indiana, Pennsylvania, and New York. In the last four decades, however, corn yield was the highest in west and south west states (Washington, Oregon, California, New Mexico and Arizona). The main reason for higher yield (Mg ha^{-1}) in western states could be relatively high percentage of irrigated corn area compared to dryland corn in the states than is the case in the Corn Belt region.

The harvested area of corn in the 41 states for the years 1900 to 2000 is depicted in Figure 3. As can be seen from the graph, area allocated to corn varied by state and by year. Over all the harvested area of corn was high in the early 1900's for many states and in total for the U.S. Corn harvested area in the U.S. seems to have declined between years 1919-1960, was variable in the years between 1960 and 1980's, and it is on the increase from 1980's onwards. Area allocated for corn seems to have highly varied from state to state over the 112 years considered.

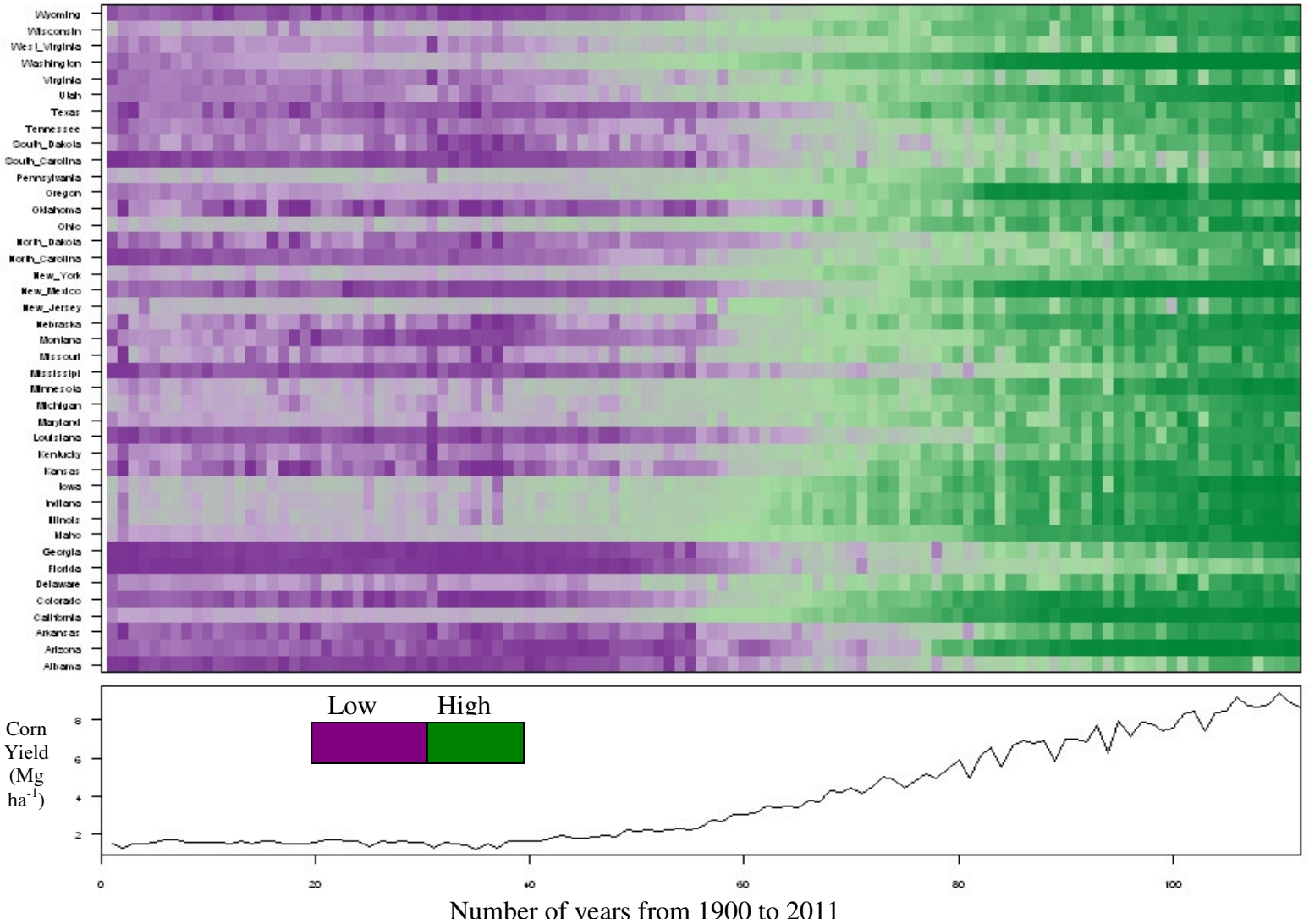

Figure 2. Multivariate time series plot of corn yield per hectare for 41 states in the USA from 1900 to 2011. The row data is grouped into three tertiles, i.e., purple representing low, gray representing medium, and green representing relatively high yields.

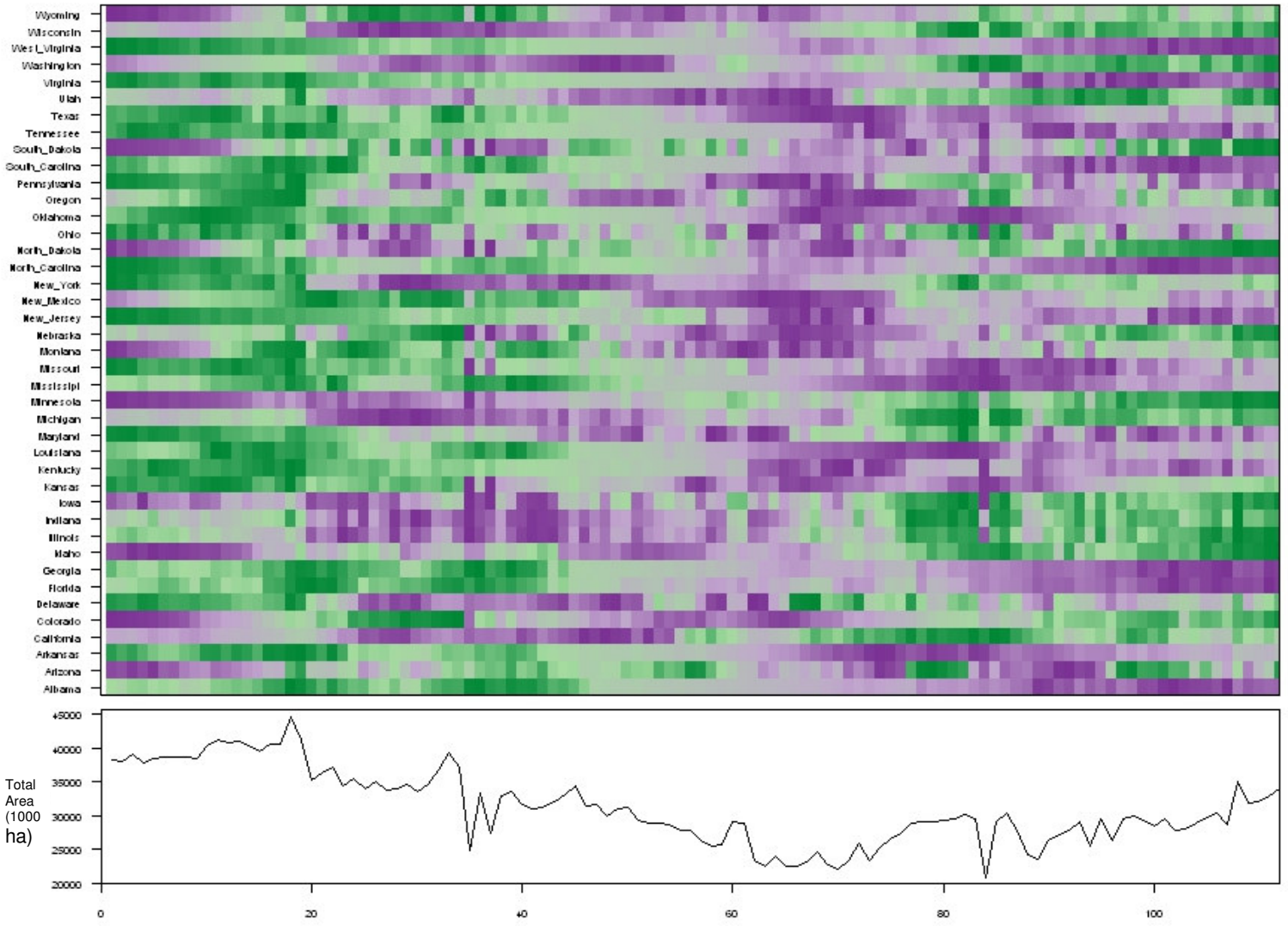

Figure 3. Multivariate time series plot of corn harvested area for 41 states in the USA from 1900 to 2011. The row data is grouped into three tertiles, i.e., purple representing low, gray representing medium, and green representing relatively high harvesting areas.

The total corn production (product of yield (Mg ha-1) and area (ha)) for the 41 states in the USA is depicted in figure 4. Similar to corn harvested area and yield per hectare, total corn production in the USA varied in time and space. Total production for the years 1900 to 1940 was about 50 million tons per year. However, it significantly increased from 1940 onwards and currently it is about 300 million tons per year. In the first few decades (1900-1930), states like West Virginia, Virginia, Tennessee, Oklahoma, New Jersey, and Kansas contributes the highest corn production. In recent decades, total corn production is high in most of the 41 states considered.

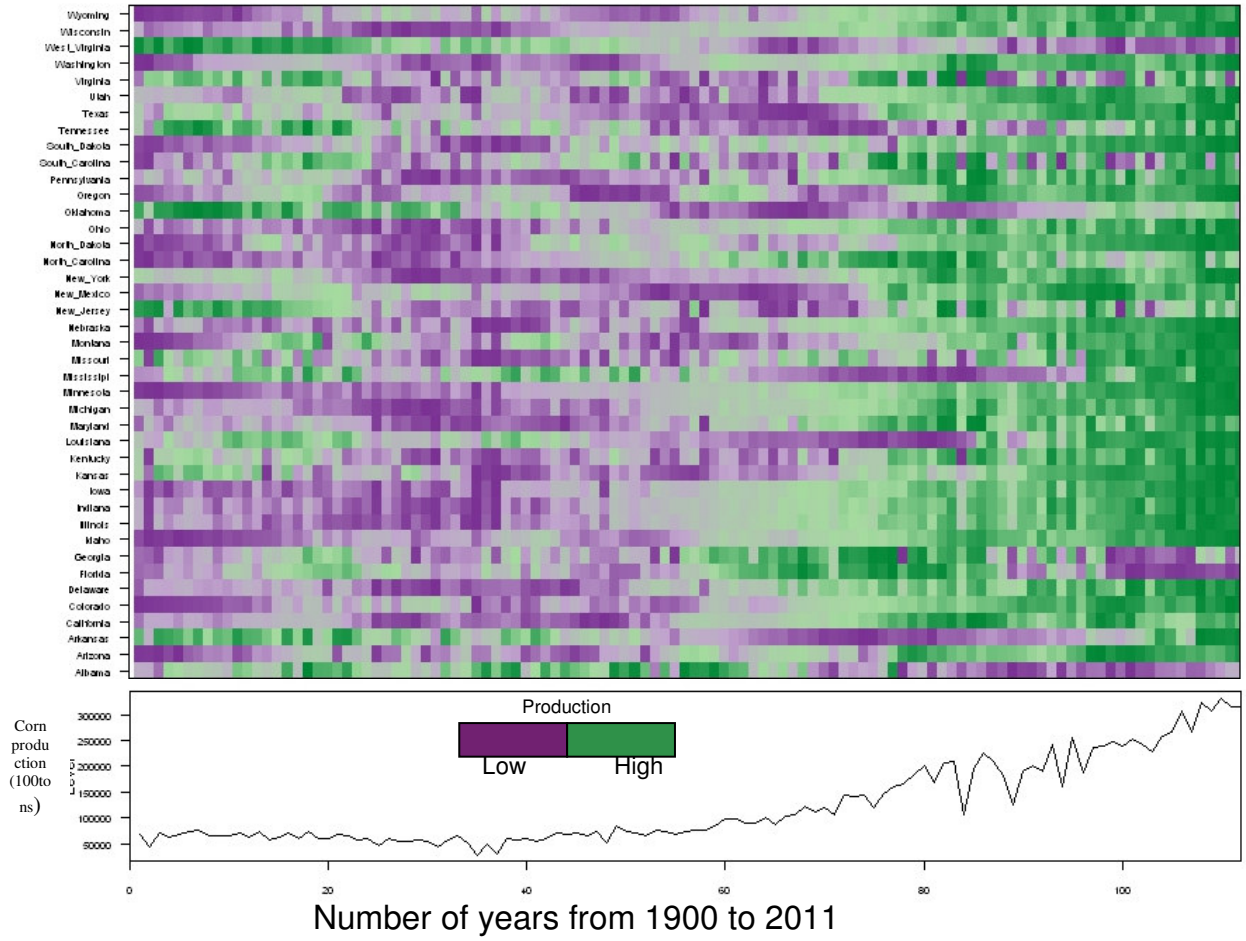

Figure 4. Multivariate time series plot of total corn production for 41 states in the USA from 1900 to 2011. The row data is grouped into three tertiles, i.e., purple representing low, gray representing medium, and green representing relatively high total production.

The price of corn for the 41 states for the years 1900 to 2000 is depicted in Figure 6. For most of states, price data is only available from 1907. For 6 out of the 41 sates in our study, price data

was only available from 1949. As can be seen from the graph, price of corn varied very little between states. However, it varied significantly over time. The recent decade seem to have the record high price for corn (at about \$0.25 kg^{-1} of corn).

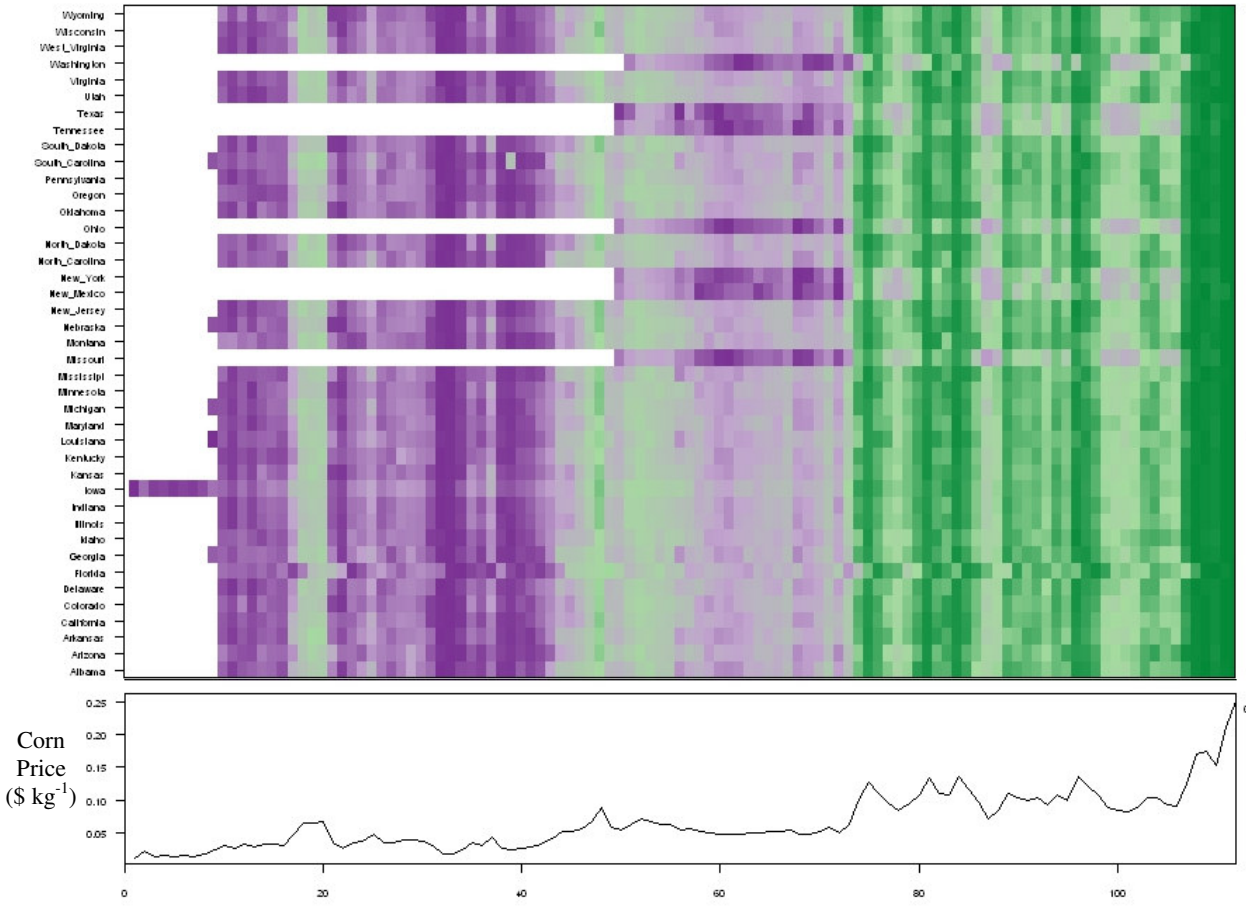

Figure 5. Multivariate time series plot of price of corn for 41 states in the USA from 1900 to 2011. The row data is grouped into three tertiles, i.e., purple representing low, gray representing medium, and green representing relatively high prices.

Autocorrelation in Yield and Annual Yield Growth

The average corn yield per hectare and its autocorrelation and partial autocorrelation function are depicted in figure 6. The autocorrelation in yield slowly decays from the first time lag to the 28 years time lag and becomes almost zero afterwards. This type of autocorrelation function (ACF)

indicates that there is a trend which renders the non-stationarity of the variable yield over time. That also means yield was not a random process and it looked like a strong correlation between yields in adjacent and near adjacent years. However, the partial correlation proves the significant autocorrelation that we see in the ACF plot for time lag 2 and above is propagation from autocorrelation at time lag 1, which was the only significant time lag in the PACF plot.

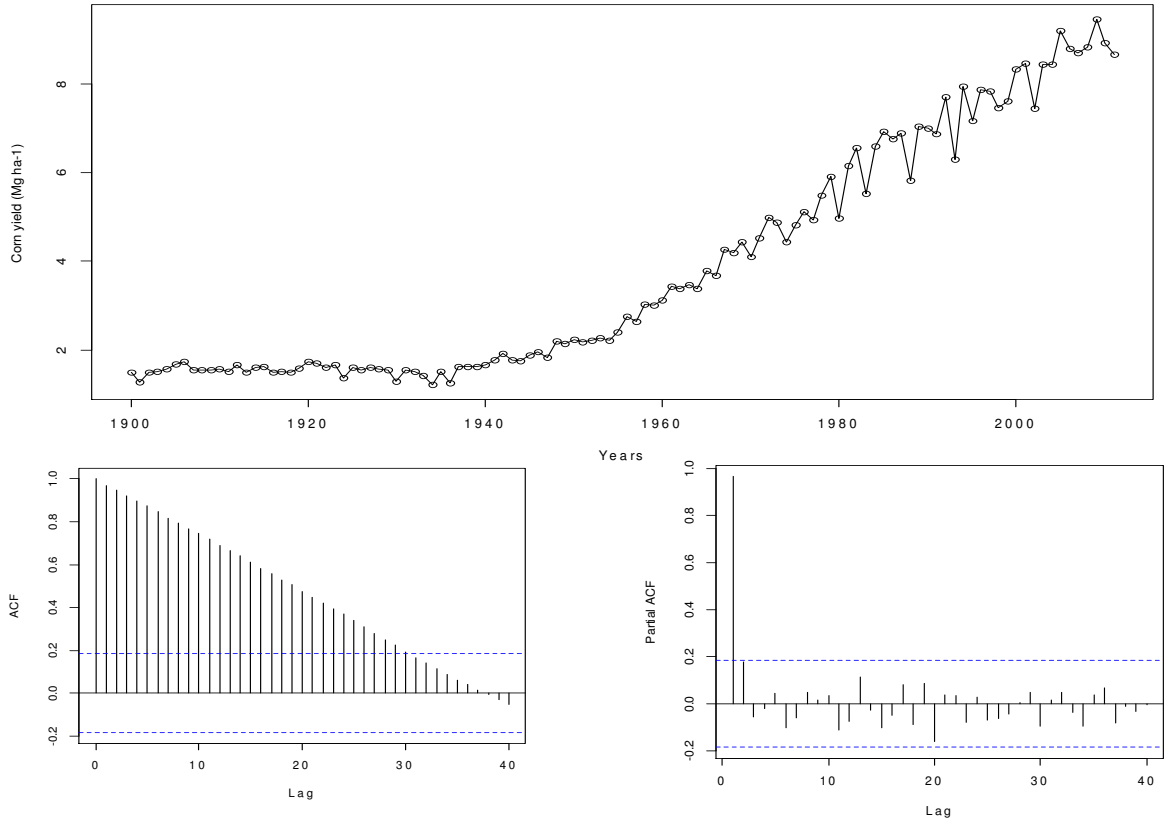

Figure 6. Average corn yield in the USA (1900-2011) and its autocorrelation and partial autocorrelation functions for up to 40 years lag.

To make the yield data stationary, detrending the data (removing trend) was necessary. The trend was removed by taking a first difference, i.e., difference between yield at time t and yield at t-1. This differencing resulted in time series data of annual yield growth. The annual yield growth, its ACF, and PACF are depicted in figure 7. The ACF of annual yield growth was not a slow decay but rather a sharp "cut off" after time lag 1 indicates that there is no trend in this variable.

A significant spike at time lag 1 indicates a negative and strong significant correlation between annual yield growth at time t and annual yield growth at t-1.

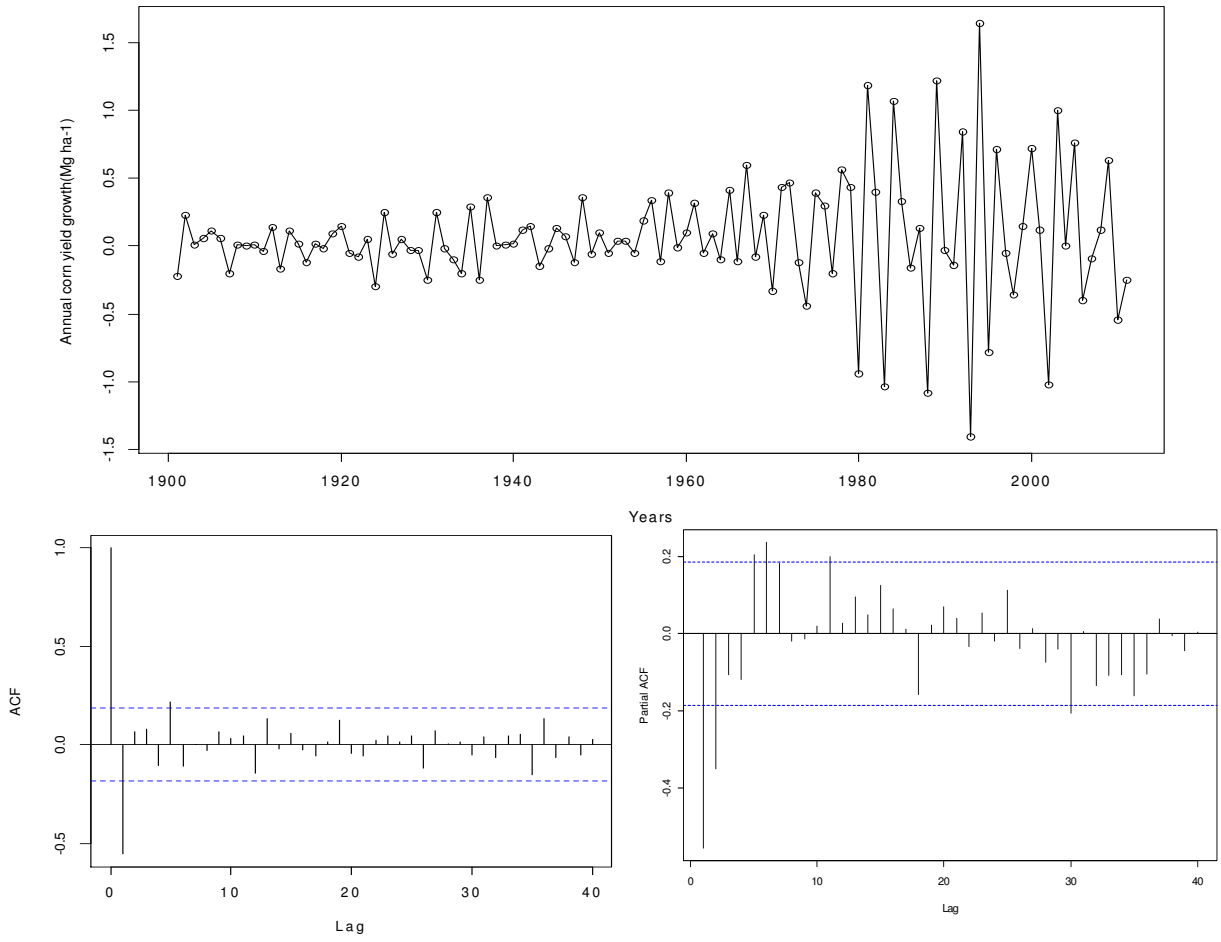

Figure 7. Average annual corn yield growth in the USA (1900-2011), its autocorrelation and partial autocorrelation functions for up to 40 years lag.

Corn yield (Mg ha^{-1}) is a function of genetics, weather, and other management factors (Assefa et al., 2012; Cardwell, 1982; Duvick, 2005). There is no reason to anticipate a cause and effect type of relationship between corn harvested area, price of corn, and/or total corn production with yield per hectare. Therefore, we did not perform a cross correlation between yield and the other three variables.

Auto- and Cross Correlation within and among Harvested Area, Price, and Total Corn Yield

Figure 8 depicted the trend of harvested area, price, and total corn production in the USA. We can deduce that the trend in these variables before and after 1960s was different. Corn harvested area seems to have been declining from early 1920s to the 1960s and has been increasing from the 1960s onward. Corn price before the 1960s was relatively cheaper and showed little variation over time. Starting from around the end of the 1960s, however, corn price have slowly increasing. Similarly, total corn production in the USA did not show significant change between the years 1900 to1960s. Since the 1960s, however, total corn production is on the rise. For this reason, the auto- and cross correlations within and between these variables was studied in two categories, i.e., for the period 1900-2011 and for the time period 1960-2011.

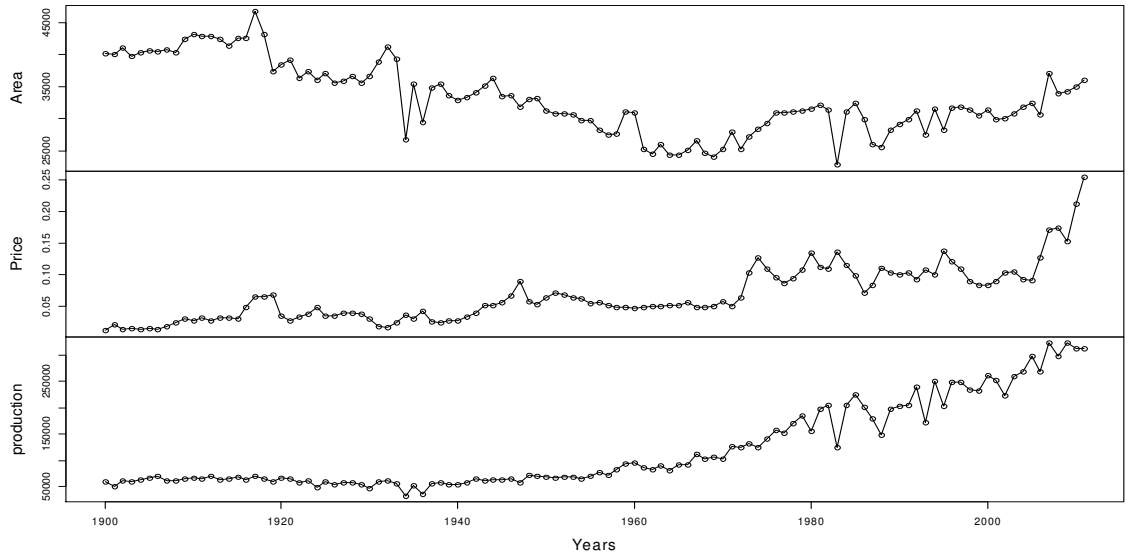

Figure 8. Total harvested area, annual average price, and total corn production in the USA from 1900 through 2011.

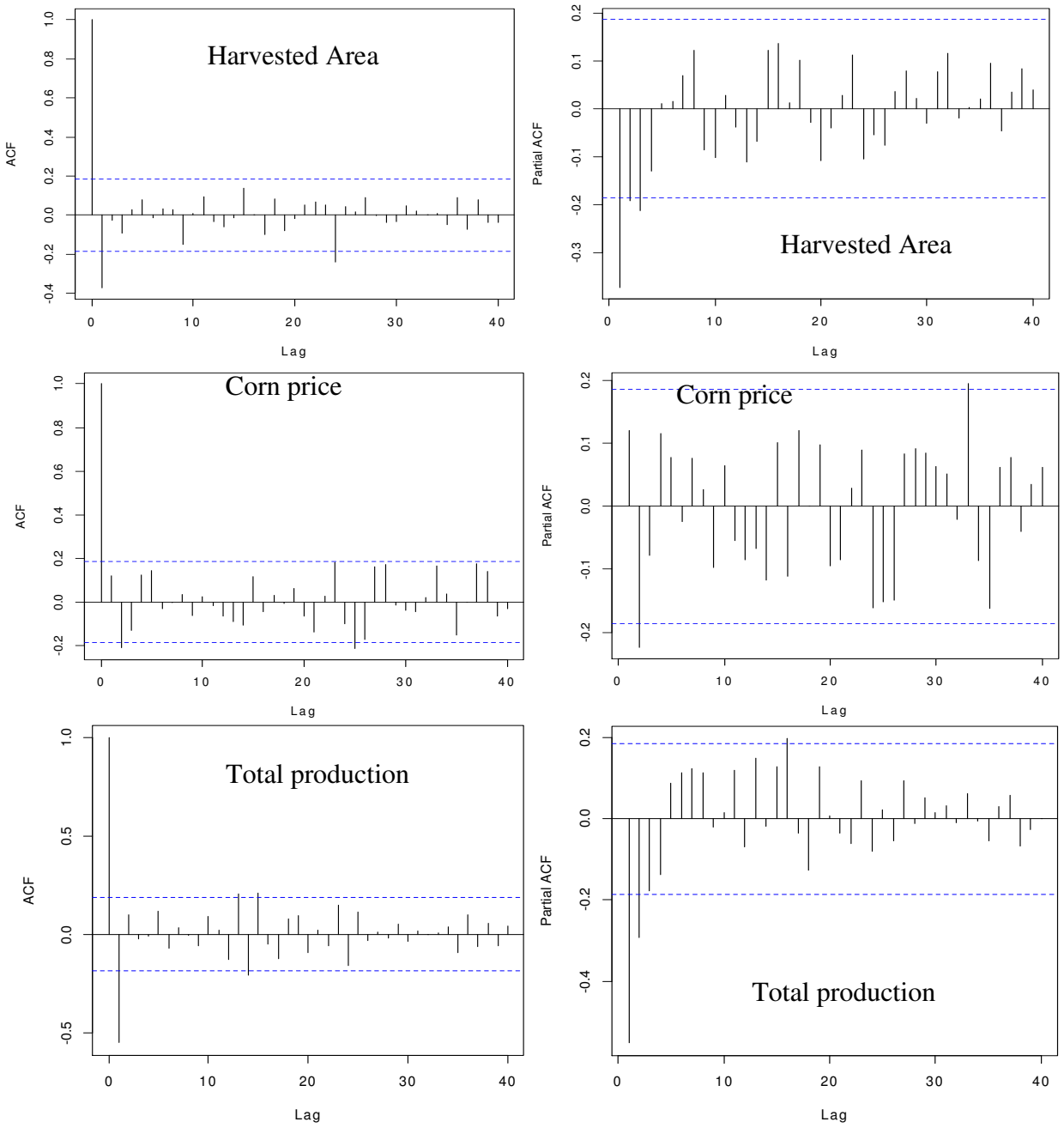

Figure 9. Auto- and partial correlation within annual growth (detrended) values of harvested area, annual average price, and total corn yield in the USA.

The autocorrelation indicated a slow decay due to trends that made the variables a non-stationary process. Therefore, the variables were detrended to make them stationary. Auto and partial autocorrelations for detrended corn harvested area; price and total corn production is depicted in figure 9. Similar results are obtained when the entire data, 1900-2011, or part of data,1960-2011, is used for the auto and partial correlation analysis.

The cross correlation between the detrended harvested area, price, and total corn yield is presented in figure 10. For the two data sets (1900-2011 and 1960-2011), the cross correlation results were slightly different for the relationship between price with total yield, otherwise, they were same. In both data sets, a negative significant cross correlation between price at time t with total yield and harvested area at time t was evidenced. Which means price was high when harvest area or total production is low and price was low when harvest area or total production were high. This is simple demand and supply relationship. Annual price increase at time t significantly increased annual harvest area increase at t+1 but did not have a significant impact for remaining time periods. This is an indicator that price motivates people to produce more.

Annual harvest area increase at time t, obviously, positively impacted annual total yield at time t but negatively impacted total yield change at t+1 and vice versa. In the 1900-2011 data, annual total yield increase seem to positively related to price at time t+3. However, in the 1960-2011 data, no significant relationship between total yield increase at time t and price at t+h ($h \neq 0$) was observed. On the other hand, price increase at time t positively related to total yield at t+1 (Fig 10).

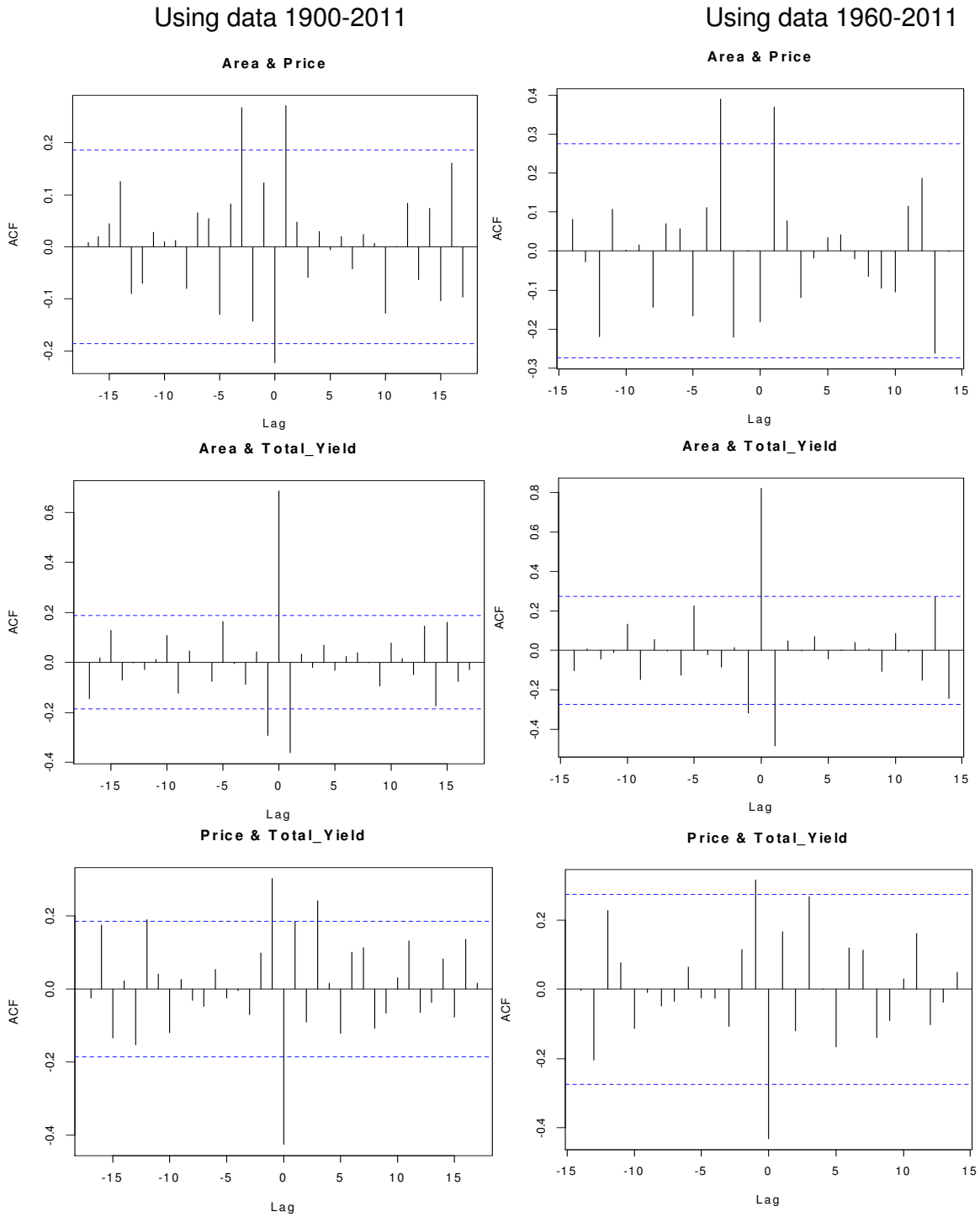

Figure 10. cross correlation between detrended values of harvested area, annual average price, and total corn yield in the USA using data from 1900-2011 (left) and data from 1960-2011 (right). In the positive side of this graphs, the variable whose name comes second leads (comes first in time) before the variable which's name comes first. For example on the top, we have area and price. The positive side of that graph shows the correlation between price at time t and area at t+1, t+2…. On the negative side the variable whose name appears first in the title leads in time.

Temporal Modeling of Price, Total Area, and Total Corn Yield in the USA

The preliminary analysis described above indicated corn production variables might best be modeled in an autoregressive process. The dependence between harvested area, price, and total production suggest that we could utilize information from the other two variables and own history when modeling one. For these reasons, a vector autoregressive (VAR) model was selected because it has the potential to utilize information from variables own history and history of related variables.

Initially two competitive VAR models were developed, using the whole data set (1900-1999) and with data set from 1960-1999. The reason why the whole data set (1900-1999) and the data set 1960-1999 was used to develop models was, as it is presented in previous section, the relationships between harvested area, price, and total yield, differ before and after 1960's.

When the data from 1900-1999 was used and the number of lags that should be included in the VAR model were searched (with maximum lag of 10), a VAR model with three years lag was found better in terms of its goodness of fit values, i.e., AIC, HQ, and FPE. On the other hand, when the data from 1960-1999 is used and best time lag was searched, a one year lag VAR model comes out to be better in terms of the entire model selection criterion used, i.e., AIC, HQ, SC, and FPE. Therefore, a VAR model with three years lag and a VAR model with one year lag were developed and compared. Figure 11 depicts the observed and fitted values of area, price, and production from these two models. Figure 12 depicts the observed and forecasted values of area, price, and total production for the years 2000-2011 based on same VAR models. In table 1 we presented a comparison of

these two models and additional ad-hoc model that assumes this years area, price, and production is almost same as next year. As can be seen in the table, the VAR model with one year lag predicts better than the two.

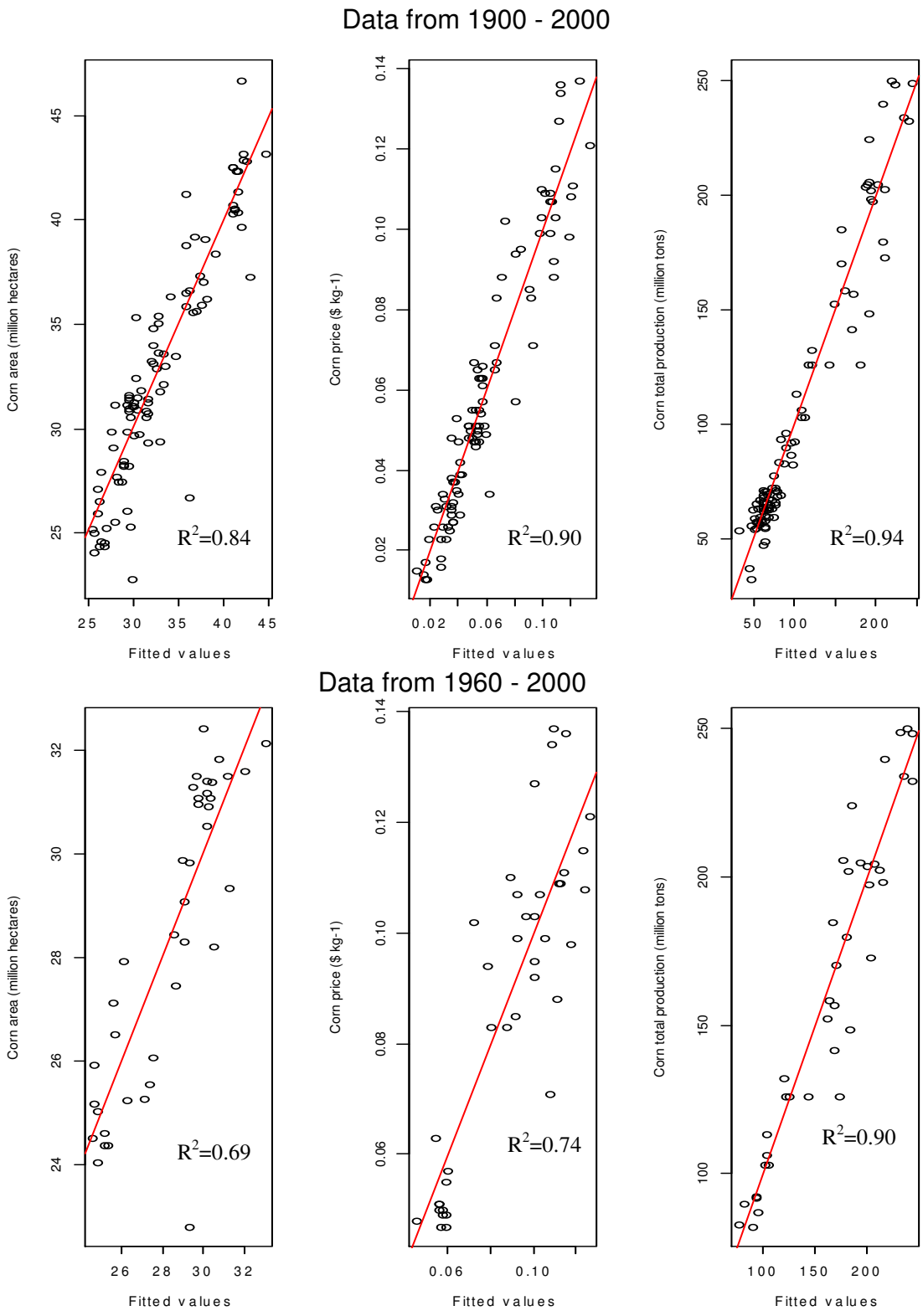

Figure 11. Observed and fitted values of area, price, and production using VAR model with 3 lag time (top) and VAR model with 1 lag time (bottom).

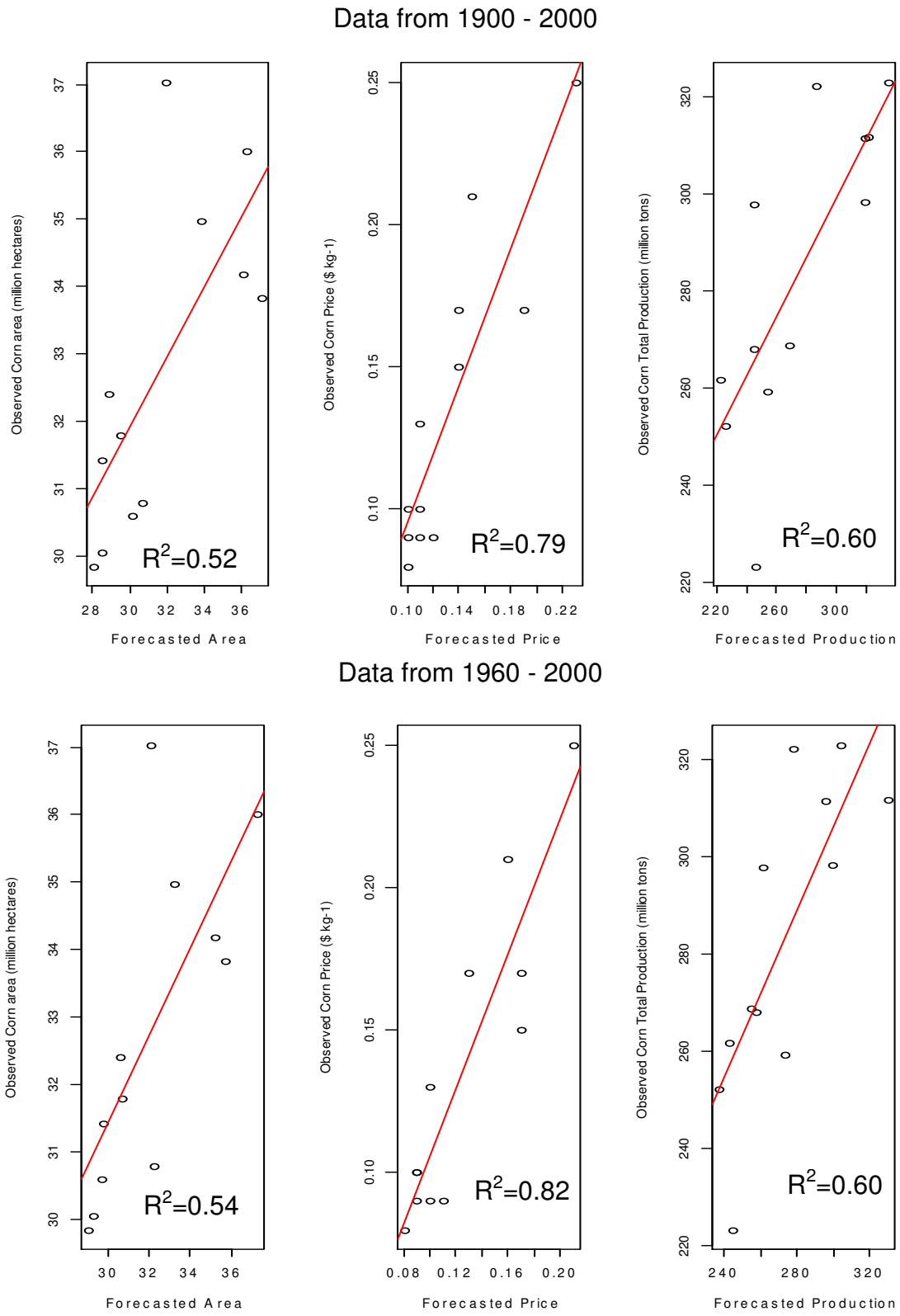

Figure 12. Observed and forcasted area, price, and production using VAR model with 3 lag time (top) and VAR model with 1 lag time (bottom).

Table 1. A comparison of the VAR models with three lag, VAR model with one lag, and Ad-Hoc model with R2, AIC, BIC, and Mean Square Error of Prediction

Comparison Criterion	VAR model with three lag years			VAR model with one lag year			Ad-hoc model with years value predicted same as previous year		
	Area	Price	T.Prod.	Area	Price	T.Prod.	Area	Price	T.Prod.
R^2	0.52	0.79	0.60	**0.54**	**0.82**	**0.60**	0.25	0.80	0.41
AIC	51.2	-49.5	110.9	**50.7**	**-51.9**	**110.9**	56.6	-50.5	115.8
BIC	52.7	-48.1	112.4	**52.1**	**-50.4**	**112.4**	58.0	-49.1	117.3
MSEP	3.0	0.0006	**440.3**	**2.9**	**0.0006**	442.1	4.8	0.0006	661.4

The coefficients for predicting the average area, price, and total production at time t based on the selected VAR model with one year lag is presented in (15) below.

$$\begin{bmatrix} \hat{A}_t \\ \hat{P}_t \\ \hat{TY}_t \end{bmatrix} = \begin{bmatrix} 0.52 & 42.7 & -0.05 \\ -0.004 & 0.865 & 0.007 \\ 4.6 & 276.51 & -0.52 \end{bmatrix} \begin{bmatrix} A_{t-1} \\ P_{t-1} \\ TY_{t-1} \end{bmatrix} + \begin{bmatrix} 12.77 \\ 0.074 \\ -30.00 \end{bmatrix} + \begin{bmatrix} 0.25 \\ -0.002 \\ 5.82 \end{bmatrix} [t] \qquad (15)$$

To predict area at time t, for example, this model suggest multiplying the previous years area by 0.52, previous years price by 42.7, multiplying the total production by -0.05, and adding a constant 12.77, and trend 0.25 values. From this we can infer a positive impact of price on producers' motivation to produce more. On the other hand, we can infer a smaller but negative impact of previous years area of production on price at time t.

Spatial dependence (Spatial Autocorrelation)

The spatial characteristics of harvest area of corn, corn yield per hectare, total corn production and corn price in the USA can also be used as a modeling tool. In order to demonstrate that here we started by analyzing whether area, yield, total production, and

price have a spatial dependence. Moran's I analysis was conducted and a significant spatial clustering was proved for all of these variables (table 2). A significant clustering of all of this variables implies that harvested area, yield per hectare, total production, and price of corn tend to be similar in places that are closer than distant.

Table 2. Moran's I analysis result for test for randomness of corn production variables

Moran's I Summary	Harvested area	Yield per acre	Total corn yield	Price
Moran's Index:	0.43	0.42	0.41	0.36
Expected Index:	-0.03	-0.03	-0.03	-0.03
Variance:	0.01	0.01	0.01	0.01
z-score:	5.98	5.41	5.71	4.71
p-value:	<0.001	<0.001	<0.001	<0.001

The average (average of 1900-1939, 1940-1979, and 1980-2011) spatial characteristics of yield, harvested area, and price are depicted in figure 13. Corn area was relatively high in the Corn Belt region (Illinois, Iowa, Indiana, Missouri, Kansas, and Nebraska) almost in all the time periods studied. West and South West regions had lowest corn harvested area.

Average yield per hectare for the 1900-1939 and 1940-1979 periods was higher and equivalent in the Corn Belt region and West Coast. In the recent years, 1980-2011, the West Coast and South West regions seem to have the highest yield per hectare followed by the Corn Belt region.

Total corn production is by far higher in the Corn Belt region (particularly in Iowa and Indiana) than any place in the USA. Relative price of corn follows almost the reverse of

total production. Corn prices are higher in places where total production is lower and vice versa.

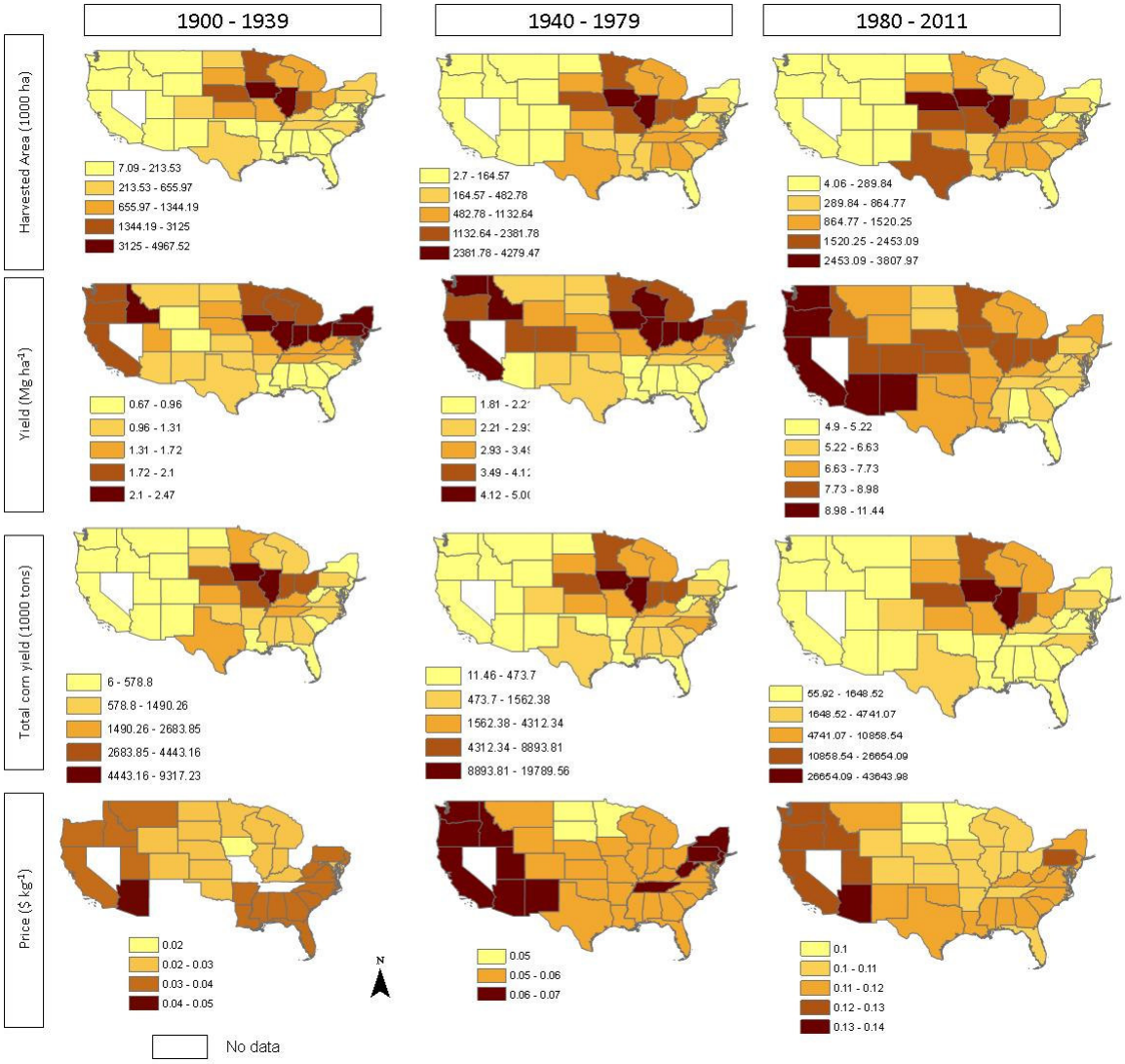

Figure 13. The spatial characteristics of corn harvest area, corn yield per hectare, total corn production, and price of corn in the USA at different periods from 1900-2011.

Time series modeling using spatial information

The analysis above and particularly the semivariogram in figure 14 below prove a spatial association in corn yield across the USA. Since states that are closer to each other tend to have a similar yield, we can use a states own yield history plus the history of yield in

neighboring state for modeling and forecasting yield. Here, we used a multivariate time series approach in modeling and forecasting yield for state of Kansas using the time series corn yield data information for state and its neighboring states.

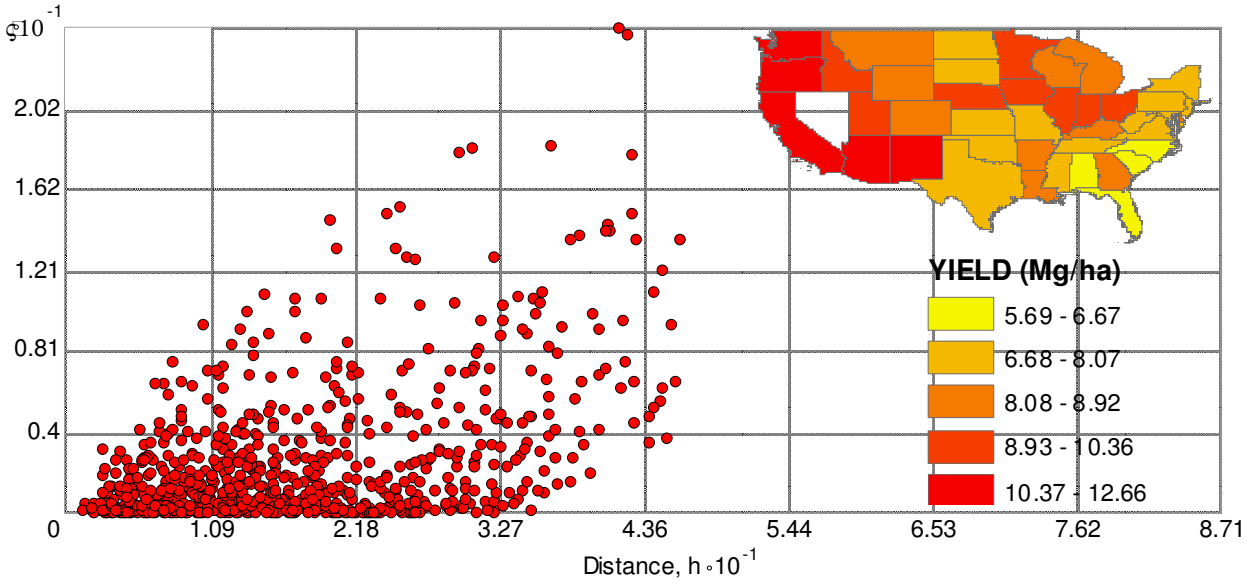

Figure 14. A semivariogram cloud and yield map of average corn yield in the US for the years 2000-2011.

Kansas shares border in the east with Colorado, in west with Missouri, in north with Nebraska, and in the south with Oklahoma. Based on the semivariogram analysis above, it is a safe assumption to say corn yield in Kansas is much more related to these neighbors. Therefore, here we hypothesized a yield model that can utilize information from Kansas own yield history and history of yield across neighboring state might do well than an ahoc yield model. To test this hypothesis we assembled yield information from Kansas and neigboring states for the years 1900-2000 (Fig. 15). Then the time lag that had the highest correlation to predict yield at time t was auto-searched to model yield using VAR model. The yield data for Kansas and its neighboring states for the years 1960-2011 were used to develop the model.

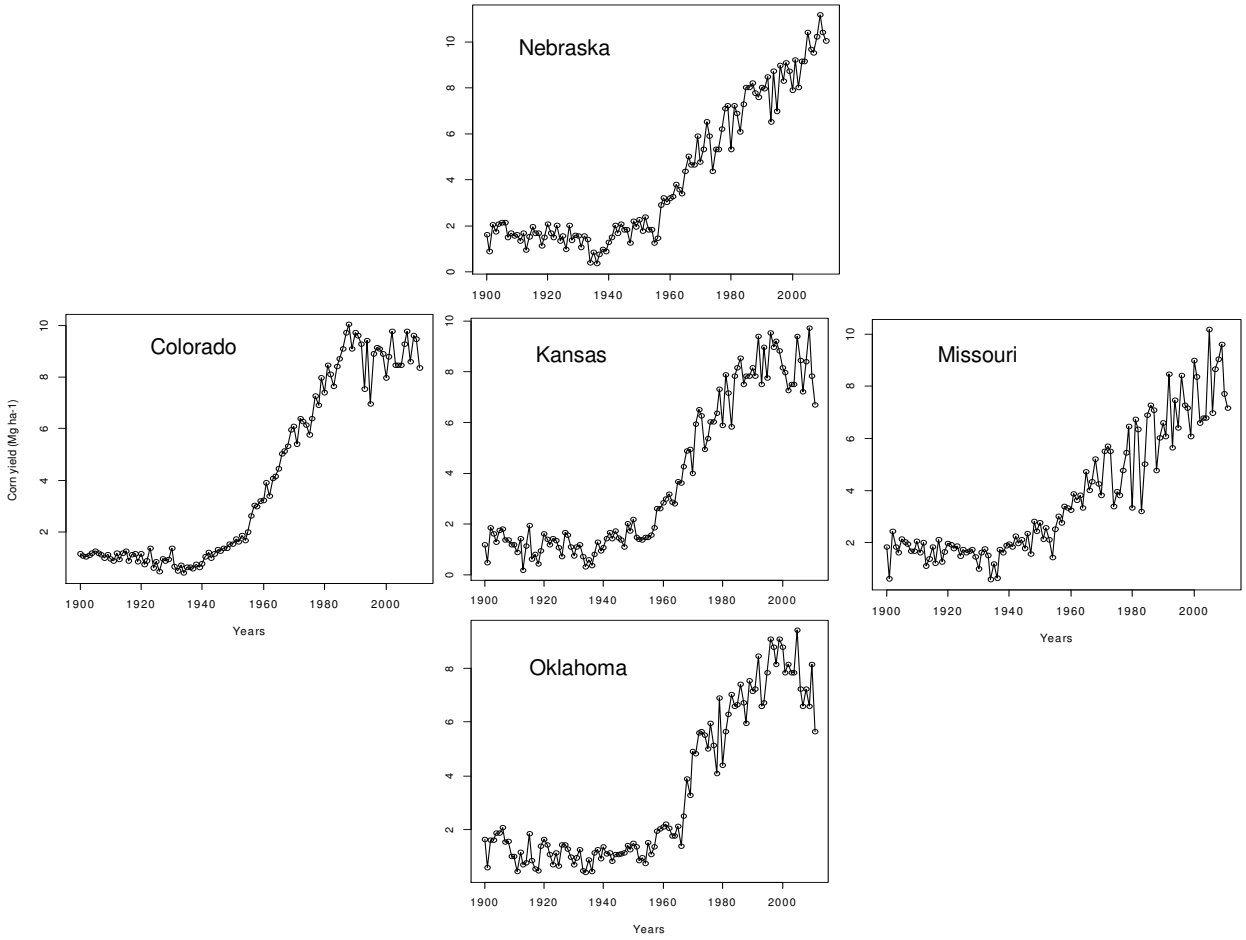

Figure 15. Time series plots of corn yield in states of Colorado, Kansas, Missouri, Nebraska, and Oklahoma for the years 1900-2011.

When the yield data for the five states for the years 1960-2011 were assembled and the time lag that best predicts time t is searched, a model with time lag 1 was selected with all of model selection criterions employed. This selected model and the ad-hoc model that assumes time t is almost same as time t-1 were compared. In Figure 16 we presented the comparison of the VAR model and the ad-hoc model in terms of some model evaluation graphs. As can be seen in the graphs, both models did well but the VAR(1) model was better in terms of explaining the variability and meeting model assumptions.

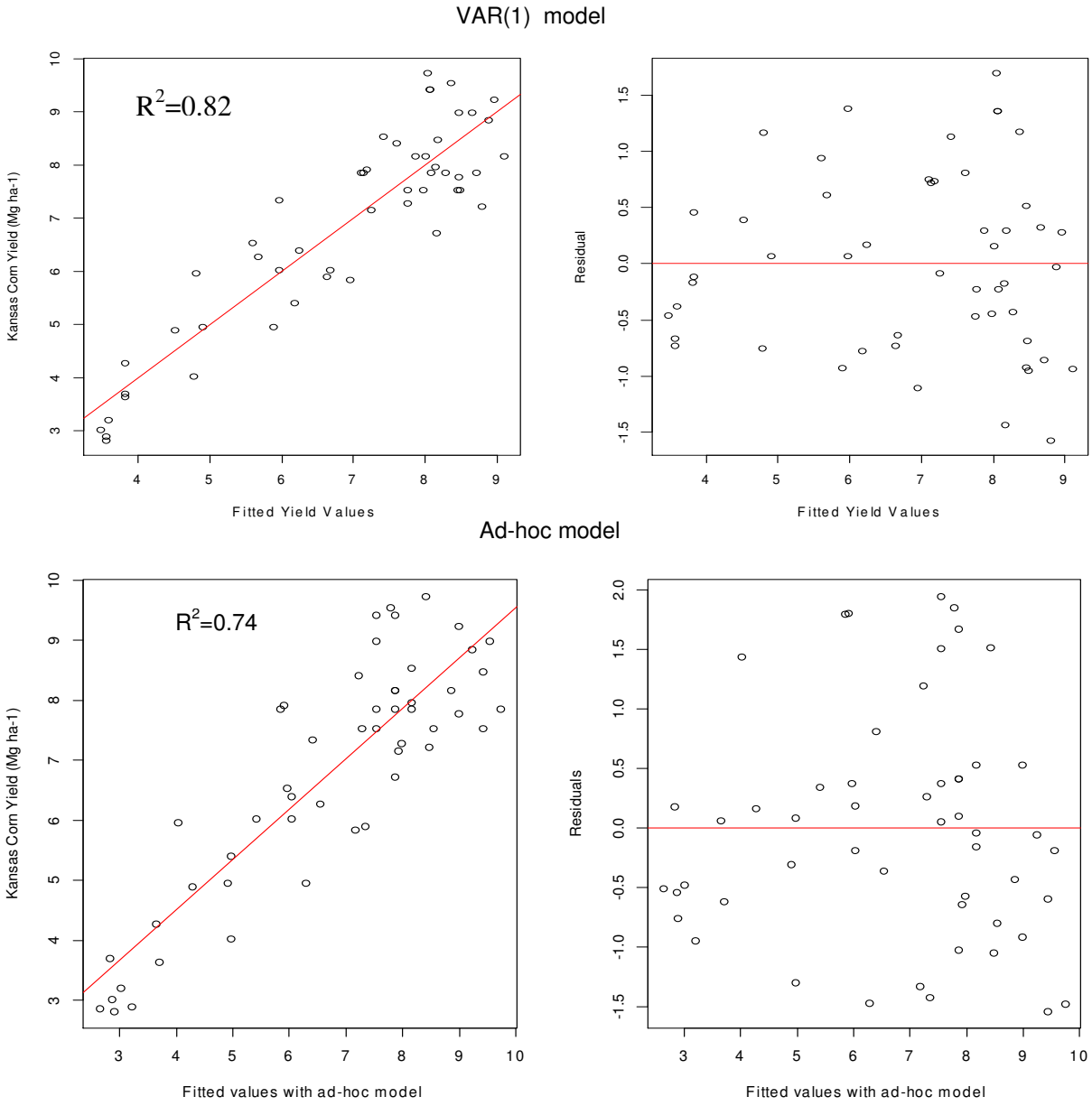

Figure 16. Observed and fitted model values (on the left) and residual verses fitted values (on the right) for VAR(1) model (on top) and ad-hoc model(on bottom).

$$\begin{bmatrix} \hat{Y}_{KS_t} \\ \hat{Y}_{CO_t} \\ \hat{Y}_{MO_t} \\ \hat{Y}_{NE_t} \\ \hat{Y}_{OK_t} \end{bmatrix} = \begin{bmatrix} 0.78 & 0.14 & -0.25 & -0.46 & 0.01 \\ 0.47 & 0.43 & -0.15 & -0.16 & -0.06 \\ 0.49 & -0.21 & -0.17 & -0.34 & -0.33 \\ 0.16 & 0.04 & -0.08 & -0.08 & -0.14 \\ 0.78 & -0.04 & -0.14 & -0.11 & 0.22 \end{bmatrix} \begin{bmatrix} Y_{KS_{t-1}} \\ Y_{CO_{t-1}} \\ Y_{MO_{t-1}} \\ Y_{NE_{t-1}} \\ Y_{OK_{t-1}} \end{bmatrix} + \begin{bmatrix} 2.91 \\ 2.34 \\ 5.09 \\ 3.76 \\ 1.29 \end{bmatrix} + \begin{bmatrix} 0.09 \\ 0.05 \\ 0.17 \\ 0.14 \\ 0.06 \end{bmatrix} [t] \qquad (16)$$

The VAR(1) model in (16) is valid for Kansas only because the neighbors are selected because they share border with state of Kansas. Based on this model, yield was forecasted for Kansas for the years 2012-2021. The forecasted yield and the 95% confidence intervals are presented in figure 17.

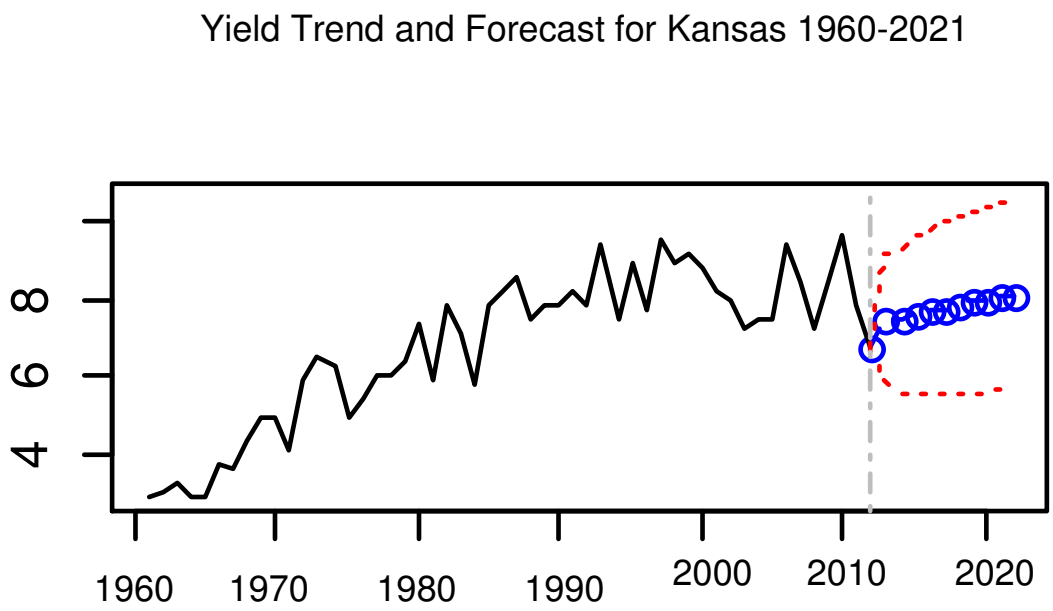

Figure 17. Yield trend from 1960-2011 for Kansas and forecasted yield and 95% confidence band for 2012 to 2021based on VAR(1) model presented in equation 16.

SUMMARY AND CONCLUSION

Here we started by conducting an exploratory multivariate time series analysis of corn in the U.S. Our analysis found that corn yield, total corn production, corn area, and price trends varied in time and in space in the U.S. for the years from 1900-2011. Reasons such introduction of hybrid technology, changes in management, irrigation, subsidies, rising in soybean crop acreage, weather, emerging new uses of crop and other similar reasons are

indicated in the literature for variation in corn trends (Duvick 2005; Runge, 2002; Kucharik and Ramankuty, 2005).

Second we performed an autocorrelations analysis within yield, total production, area, and price of corn. Almost all of these variables demonstrated a significant autocorrelation at time lag 1, i.e., the value of these variables is highly correlated with their value at time t-1. Then we performed and proved that there is a cross correlation between price, area, and total production. That means the value of each of these variables is not only highly correlated with their own past but also the value of the other two variables in previous years.

Third we developed a vector autoregressive model for area, price, and total production. This model predicts the value of area, price, and total production at time t using not only the information from each of the variables past but also the past information from the other two variables. We developed two of these models and proved that VAR(1) model performed much better than the ad-hoc model.

Fourth we proved spatial dependence of corn yield in addition to the time dependence that we have proved above. Since, yields of states that are closer in distance tend to be similar; we demonstrated how we can use this information in early modeling and forecasting yield. We took Kansas as an example and developed a model to predict Kansas yield at time t using yield information from its past and past yield information of

neighboring states. Then we compared this model and showed a better model fit than the ad-hoc approach.

In conclusion, this analysis demonstrates how data that is rich in its temporal and spatial dimension can be used in modeling. Multivariate time series approach presents the capability to study how variables behave in time, in space, and in relation to other variables. The models that we have developed herein are examples of how this capability can be translate into systematic use of data.

REFERENCES

Assefa Y., K. L. Roozeboom, S. A. Staggenborg, and J. DU. 2012. Dryland and irrigated corn yield with climate, management, and hybrid changes from 1939 through 2009. Agron. J. 104:473-482.

Boken V. K. 2000. Forecasting spring wheat yield using time series analysis: A case study for the canadian prairies. Agron. J. 92(6):1047-1053.

Cardwell, V.B. 1982. Fifty years of Minnesota corn production: sources of yield increase. Agron. J. 74:984-990.

Chatfield C.. 2000. Time series forecasting. Chapman & Hall/CRC, New York.

Dudley N.J., A.B. Hearn. 1993. El Niño effects hurt manoi irrigated cotton growers, but they can do little to ease the pain. Agricultural Systems, 42: 103–126.

Duvick, D.N. 2005. The contribution of breeding to yield advances in maize. Adv. Agron. 86:83 - 145.

ESRI. 2011. ARCGIS DESKTOP. Release 10. Environmental Systems Research Institute. Redlands, CA.

Elbehri A. and P. Paarlberg. 2003. Price Behavior in Corn Market with Identity Preserved Types. American Agri. Econ. Assoc. Annual Meeting, Montreal, Canada, July 27 30, 2003.

Hammer G.L., J.W. Hansen, J.G. Phillips, J.W. Mjelde, H. Hill, A. Love, and A. Potgieter. 2001. Advances in application of climate prediction in Agriculture. Agricultural Systems. 70 (2–3): 515–553

Hammer G.L., D.P. Holzworth, R. Stone. 1996. The value of skill in seasonal climate forecasting to wheat crop management in a region with high climatic variability.

Australian J. of Agric. Res. 47: 717–737.

Kantanantha N., N. Serban, and P Griffin. 2010. Yield and price forecasting for stochastic crop decision planning. J. of Agric., Biol., and Enviro. Stat. 15 (3): 362–380.

Kucharik C. J. and N. Ramankutty. 2005. Trends and variability in U.S. corn yields over the twentieth century. Earth Interactions. 9(1).

Kumar N., S. Ahuja, V. Kumar, and A. Kumar. 2010. Fuzzy time series forecasting of wheat production. IJCSE. 02, (3): 635-640.

Meinke H., G.L. Hammer. 1997. Forecasting regional crop production using SOI Phases an example for the Australian peanut industry. Australian J. of Agric. Res. 48:789–793

O'Brien D. 2010. World corn market supply-demand trends. K-State Research and Extension.

Peng R. D. 2008. A Method for Visualizing Multivariate Time Series Data. J. of Statistical Software. 25, Code Snippet 1.

Pfaff B. and K. Taunus. 2008. VAR, SVAR and SVEC Models: implementation Within R Package vars. J. of Statistical Software. 27 (4).

R Development Core Team. 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Available at http://www.R project.org.

Runge C. F. .2002. King corn: the history, trade and environmental consequences of corn (Maize) production in the United States. World Wildlife Fund Inc. Washington DC.

Singels A., A.B. Potgieter. 1997. A technique to evaluate ENSO-based maize production strategies. African J. of Plant and Soil, 14 (3):93–97.

Stone R.C.and H. Meinke. 2005. Operational seasonal forecasting of crop performance. Phil. Trans. R. Soc. B 2005 360, 2109-2124.

USDA National Agricultural Statistics Service. 2006. Yield Forecasting program of NASS. U.S. Gov. Print. Offi ce, Washington, DC.

Chapter V

A GENERAL SUMMARY AND CONCLUSION

In this study we presented analysis of a time series and spatial data collected across different counties in Kansas and a survey data set collected by USDA across the US. The objectives of the study were to investigate crop yield trends in space and time, quantify the variability in yield explained by genetics and space-time (environment) factors, and study how spatio-temporal information could be incorporated and also utilized in modeling and forecasting yield.

In our trend analysis (Chapter II), we performed a simple regression analysis of data over the time period and compared the means of the data at different time and space intervals. From this analysis we found spatial and temporal trends in corn yield in Kansas. Spatially, average dryland yields in Kansas decreases significantly from east to west and slightly from north to south. Temporally, corn yields have increased at an average rate of 94 kg ha^{-1} yr^{-1} in dryland and 127 kg ha^{-1} yr^{-1} in irrigated trials. The rate of corn yield changes over time, however, was not regular for the seven decades considered. Both irrigated and dryland yields increased significantly at least every two decades until the last three, during which dryland yields stagnated.

Then we modeled yield using genetic and environmental information (Chapter III). In the initial step of our modeling, we checked distribution of data and declared yield can be assumed approximately normally distributed. We partition the variability and found from about 77 to 93% of the variability in yield was due to environmental factors. With the

assumption of spatial and temporal independence, varieties of yield functions were developed using traditional regression and smoothing techniques. These models are then improved by incorporating time and space information as an explanatory variable. Consequently, spatial and temporal dependence of yield was proved using spatially and temporally rich data set. In this case a method of generalized least square estimation with correlated error variance was suggested as oppose to ordinary least square regression.

We conducted a multivariate time series analysis on corn production in the US (Chapter IV). Multivariate time series plots, temporal auto-and cross correlation, spatial autocorrelation analysis were carried out. Based on the analysis, models were developed and compared. Corn area, price, yield, and total production demonstrated a significant autocorrelation with their value at time t and their value at t-1, proving they are best modeled as an autoregressive process. A significant cross correlation was also found between price, area, and total production. Base on this auto-and cross correlation analysis result, a vector autoregressive, VAR(1), model was developed proved better in model fit and forecasting qualities than an ad-hoc model which assumes next year is same as this year. A significant spatial dependence was found for these variables by Moran's I spatial autocorrelation and semivariogram analysis. This spatial dependence information was then used to develop a state based yield forecasting model. The VAR model was capable of using past (t-1) yield values of the state and its neighbors to predict yield at time t. This study demonstrates how data rich in time and space can be used for modeling and early forecasting.

Over all, we demonstrated how time and space information are useful (i) in analyzing trends and possible reasons for trend in data set, (ii) in adjusting a model for unknown or unmeasured variables, and (iii) in developing models that are purely based on temporal and spatial history of a variable. Assuming our results emphasized enough on the importance of time and space in a research, we want to comment on how time and space (location) should be treated in a model. This is especially important for researchers and consultants who want to have a logical way on how to determine the time and space information as random or fixed effect.

From many aspects, the information contained within space and time can be considered equivalent because time and space have similar characteristics, i.e., in both time and space; climate, technology, management, and resources vary. However, space and time are not same in absolute sense, i.e., the extent in which climate, technology, management, and resource variation in time and space is not likely same. The way we select space and time and the way we treat them in the analysis, therefore, might be different.

Space is static and given, time is dynamic and irreversible. Since space is given and it is in our hand, we can carefully choose it depending on what we wanted to conclude at the end of our experiment. That means, if we are going to model yield and conclude yield relationship with our independent variables for a targeted space (location), then we can do so by conducting our experiment in that specific location. However, if our intention is to make conclusion on similar locations after experimenting in representative locations, then the selection of this representative locations should be unbiased (random).

Therefore, it is our selection of the space at the beginning of the experiment that determines whether our space information should be treated random or fixed effect. Time, however, is not in our hand and it is usually impossible to randomly select time of our experiment. The reason for randomization is to avoid bias and time is out of this bias because it is unknown. Therefore, time could be considered a random effect in our model unless one justifies that the environmental conditions that occurred in time of his experiment are extreme and want to infer the result of conclusion for times like that.

Time and space information in a data set contain both known and unknown or measured and unmeasured factors that might have an influence on a response variable. If the numbers of factors that influence a response factor are few and the relationship between them is well known, time and space could possibly be replaced by measuring all those factors. This is possible for research in controlled environment or industrial researches. However, for response variables like yield that is proved to be influenced by number of factors and with the present knowledge that how these factors combined to have an effect is unknown, space and time information are irreplaceably important. For the above reason, time and space selection at the beginning of a study and utilization of time and space information at the analysis stage of a study that wishes to model yield should be carefully done.