NCCC-134
APPLIED COMMODITY PRICE ANALYSIS, FORECASTING AND MARKET RISK MANAGEMENT

Forecasting Crop Yields and Condition Indices

by

Paul L. Fackler and Bailey Norwood

Suggested citation format:

Fackler, P. L., and B. Norwood. 1999. "Forecasting Crop Yields and Condition Indices." Proceedings of the NCR-134 Conference on Applied Commodity Price Analysis, Forecasting, and Market Risk Management. Chicago, IL. [http://www.farmdoc.uiuc.edu/nccc134].

FORECASTING CROP YIELDS WITH CONDITION INDICES

Paul L. Fackler and Bailey Norwood[1]

A model relating crop condition indices to average yields is developed. The model is used to motivate a crop yield forecasting model, which in turn yields estimates of the time path of information flows into the commodity market. An empirical assessment of the forecasting model is undertaken.

Introduction

Crop yields have a significant impact on both commodity prices and farmer income. Growing season forecasts of crop yields are therefore of considerable interest to commodity market participants and price analysts. For example, grain futures prices tend to be quite volatile during crop growing seasons, with the markets being quite sensitive to weather information that impacts the yield potential of the growing crop.

Yield forecasting can be a very sophisticated enterprise utilizing crop growth models together with weather and geographical data. As such yield forecasts can be very expensive to develop. It would be useful to have a simpler alternative, if it could be shown to provide reasonably accurate forecasts. The National Agricultural Statistics Service (NASS) of the USDA publishes weekly crop condition data for various crops and regions. The crop condition indices estimate the percent of crop acres in each of five categories: very poor, poor, fair, good, and excellent. Crop condition data has the potential to provide a simple, regular source of information about the eventual realized yield and has been used for forecasting at the Food and Agricultural Policy Resource Institute (FAPRI) using a methodology developed in Kruse and Darnell.

This paper attempts to develop a explicit model of the relationship between the condition indices and the average yield and to use that model to guide statistical estimation of model parameters. The model provides not only a yield forecasting model but estimates of the seasonality of forecast volatility, which is related to the volatility of harvest contract futures prices. In the process, we also address a empirical puzzle that yield forecasts may increase when the fraction of the crop in the worst condition category increases.

The paper is organized as follows. The next section discusses the nature of the crop condition indices and develops a model that relates these to average yields. In the third section, we discuss the implications of the model for estimation and develop an econometric model. The fourth section discusses empirical results for winter and spring wheat, corn, soybeans and cotton. The last section presents a summary and conclusions.

Crop Condition Reports

Each week during growing seasons the *Weekly Weather and Crop Report* provides estimates of the fraction of acreage for selected crops and states in each of five condition classes. The USDA defines crop condition classes as follows:

[1] Associate Professor and Graduate Assistant, North Carolina State University. Address correspondence to paul_fackler@ncsu.edu

- Very Poor: Extreme degree of loss to yield potential, complete or near crop failure.
- Poor: Heavy degree of loss to yield potential which can be caused by excess soil moisture, drought, disease, etc.
- Fair: Less than normal crop condition. Yield loss is a possibility but the extent is unknown.
- Good: Yield prospects are normal. Moisture levels are adequate and disease, insect damage, and weed pressures are minor.
- Excellent: Yield prospects are above normal. Crops are experiencing little or no stress. Disease, insect damage, and weed pressures are insignificant.

County level estimates are reported by extension agents based on a subjective assessment. It is not clear that agents use a consistent set of criteria to make their assessments. Nevertheless, changes in the indices should give some indication of changes in yield expectations at the state and country level.

For each crop/region, we assume that the condition classes represents a yield interval, and that each interval has an average yield, y_i, i=1,...,5. Let c_1, c_2,..., c_5 represent the fraction of crop in the five condition classes, with the sum of the c_i identically equal to one. Given these assumptions, the average yield can calculated as

$$\text{average yield on all acres} = \sum_{i=1}^{5} y_i c_i .$$

This suggests a simple forecasting rule can be obtained by determining the applicable yield weights and using a simple weighted sum of the five condition numbers. It is clear that weights constructed in this fashion should be increasing in i.

The situation is made more complicated by the fact that the realized yield is measured in terms of the harvested production. If some acreage is abandoned then the average should be taken with only with respect to the harvested acreage. We will assume that some fixed fraction of the acres in each class, λ_i, are abandoned. The total production is then the total number of planted acres times $\sum_{i=1}^{5} y_i \lambda_i c_i$, the harvested acreage is $\sum_{i=1}^{5} \lambda_i c_i$ and the average yield with abandonment is

$$\text{average yield} = \frac{\sum_{i=1}^{5} y_i \lambda_i c_i}{\sum_{i=1}^{5} \lambda_i c_i}.$$

The possibility of abandonment and hence truncation of the lower tail of the yield distribution leads to the curious phenomenon that movement of acres from poor to very poor condition can lead to an increase in the average yield. The intuition is that increasing the acreage in very poor condition increases the fraction being abandoned and higher yielding acres make up a larger percent of total acres harvested.

To illustrate, suppose that the yield levels defining the classes are 10, 20, 30, 40, 50 and initially the fraction of acreage in each class is 0.20. With no abandonment, the average yield is 30. If all acres in the lowest condition class are abandoned ($\lambda_1 = 0$), the average yield on harvested acres increases to 35. Compare this to a situation in which 40% of the acreage is in the

worst class and no acres are in the second class. The average yield on harvested acres actually increases to 40 even though the average yield on all acres decreases to 28.

Although we have assumed that abandonment percentages are fixed, it is likely that the acres abandoned responds to price and hence is related to the total production, which in turn is related to the crop condition. Given the complexity and circularity of this relationship and the need for both a harvesting cost and a demand relationship, it was deemed expedient to use the simpler assumption.

Forecast Error Covariance Structure

Although there are reports issued every week, the forecast errors from one week to the next should be highly correlated. Useful estimates of the forecasting model should incorporate information about the nature of the forecast error covariance structure. Let time t represent the current week and $t+h$ represent h weeks hence. Let e_t be the forecast error at time t, i.e. $y=f_t+e_t$. Identically, $e_t=e_{t+h}+f_{t+h}-f_t$ and hence the variance of e_t is equal to the variance of e_{t+h} plus the variance of the change in the forecast between t and $t+h$ (the covariance of e_{t+h} and $f_{t+h}-f_t$ is zero). This has two consequences. First, the forecast error variance is declining in time. Second, the error covariance is equal to the variance of the error in the later period:

$$\mathrm{cov}(e_t, e_{t+h}) = \mathrm{var}(e_{t+h}).$$

Let Σ be the error covariance matrix. It will be block diagonal with each block representing the observations in a single year. Within a year it will exhibit a structure such that all elements above and to the left of a diagonal element will equal that diagonal element:

$$\begin{bmatrix} v_1 & v_2 & v_3 & \cdots & v_n \\ v_2 & v_2 & v_3 & & v_n \\ v_3 & v_3 & v_3 & & v_n \\ & & & & v_n \\ v_n & v_n & v_n & v_n & v_n \end{bmatrix},$$

where v_i are the error variances, which must be decreasing over time.

It is straightforward to verify that $R' D^{-1} R = \Sigma^{-1}$, where

$$R = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ & & & & \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

and D is a diagonal matrix with element (i,i)

$$d_i = \begin{cases} v_i - v_{i+1} & i < n \\ v_n & i = n \end{cases}$$

or, equivalently, $D = diag(Rv)$.

Estimation

Condition data is reported for selected locations on a weekly basis throughout the growing season. Let c_{iSYt} be the value of the *i*th condition index in location *S*, year *Y* and time of year *t*. Kruse and Darnell assume that the condition weights differ across states only by a multiplicative constant and that yields should be adjusted by a state specific time trend and that deviations from expectations exhibit state specific heterskedasticity. Furthermore, they assume that the weights on the condition indices change over the growing season. These assumptions lead to the following estimation model:

$$y_{SY} = \beta_s \sum_{i=1}^{5} w_{it} c_{iSYt} + A_s Y + \sigma_s e_{SYt}$$

where y_{SY} is the yield in location *S* in year *Y* and β, w, A and σ are parameters to be estimated. They estimated separate models for each *t*, where t is measured as the week of the year in which the condition data was reported. They also imposed the restriction that $w_i \leq w_{i+1}$.

The model developed in the previous section suggests that the weights on the condition indices can be interpreted as average yields in each condition class. If this is a reasonable description of the meaning of the classes, the weights should not change over the season. We also choose to estimate the model in terms of ratios of yield to trend yield. This makes it easier to interpret the forecasts and weights. Given the assumption that all of the very poor class is abandoned and none of the crop in other classes is, the model can be expressed

$$\frac{y_Y}{\alpha + \beta Y} = \sum_{i=2}^{5} x_{iYt} w_i + e_{Yt},$$

where

$$x_{iYt} = \frac{c_{iYt}}{\sum_{i=2}^{5} c_{iYt}}$$

(each location is treated separately rather than pooled and hence the location subscript is dropped). Pooling the data over time allows estimation of the error variance, here using a polynomial approximation constrained to decrease with *t*. In matrix notation the forecasting equation has the form

$$\tilde{y} = Xw + e.$$

The short period over which condition indices have been reported (1986-present) led us to use a two stage estimation strategy, first estimating the yield trend using data for crop years 1960-1998 and then use the ratios of yield to estimated trend as the dependent variable in the second stage.

The model can be estimated by first applying the operator *R* to the forecasting equation:

1) $R\tilde{y} = RXw + u$

where $u=Re$, the error covariance of which is the diagonal matrix *D* discussed above. Unfortunately, feasible GLS and maximum likelihood procedures are not useful in estimating the elements of *D*, because of an identification problem that arises because $x_{Yt} - x_{Yt+h}$ identically sum to one. In our reported estimates, therefore we apply OLS to (1) and utilize a heteroskedasticity-consistent covariance matrix estimator (Davidson and McKinnon, p. 552) to compute standard errors.

The conventional estimation techniques (GLS and maximum likelihood) are not applicable to this estimation problem, however. To see why, let E represent the set of observations at the end of each year and I be the set of all other observations. Assuming Gaussian forecast errors, the likelihood can be written (up to an affine transformation) as

$$-\sum_{i\in E}\left[\ln v_i - \frac{(\tilde{y}_i - f_i)^2}{v_i}\right] - \sum_{i\in I}\left[\ln(v_i - v_{i+1}) - \frac{(f_i - f_{i+1})^2}{v_i - v_{i+1}}\right]$$

where f_i is the forecast for the ith observation and v_i is the variance of the associated forecast error. In as much as the forecast function can always be set equal to a constant, f, the restricted likelihood

$$-\sum_{i\in E}\left[\ln v_i - \frac{(\tilde{y}_i - f)^2}{v_i}\right] - \sum_{i\in I}[\ln(v_i - v_{i+1})]$$

is feasible. This function, however, can be made arbitrarily large by making the change in the variance arbitrarily small. Thus, paradoxically, the likelihood is maximized by using an unconditional forecast (one that ignores conditioning information). Clearly this cannot be used as a basis for estimating a conditional expectation.

As an alternative, consider how weighted least squares procedures selectively give different weights to different observations. GLS procedures make these weights inversely proportional to the variance of the observation. This prevents observations with high variance, and hence that are expected to have large errors, from unduly influencing estimates of the conditional mean. Given that feasible GLS methods fail in the current situation, a feasible weighting scheme is used based on the following reasoning. The forecast error variance should decline over the growing season and hence is taken to be a quadratic function of the time of year, $v(t)$. This function is estimated from the forecasts errors and the condition coefficients are estimated using weights that are inversely proportional to v (GLS would use $Rv(t)$ as a weighting function). This generally will put the highest weight on the final observations of each year, but it also uses the information in earlier periods. The next section will demonstrate that this approach yields reasonable estimates of both w and $v(t)$.

Empirical Results

Crop condition reports have been issued since 1986 for five major crops (corn, cotton, soybeans, spring wheat and winter wheat) at the national level and for selected states. Only the national level is considered in this study; also winter wheat is not considered due to the difficulties in handling the winter months when the crop is dormant. The reports are issued weekly throughout the growing season; Table 1 summarizes the timing of the releases.

Although a large number of reports have been issued, they cover a period of only 13 crop years. In order to reduce the number of parameters estimated, it is assumed only crops in very poor condition are abandoned. On examination it was determined that similarly sized yield forecast errors results for a wide range of restrictions on the harvest fraction for the very poor class and that it could be set to zero without loss of precision. In this case the yield weight on the very poor class plays no role in the forecast and only four parameters need be estimated.

Estimates of the four parameters using the method discussed in the previous section are provided in Table 2. Also provided are estimates based on other estimation criteria: OLS of $R\tilde{y}$ on RX and OLS of $\tilde{y}$ on X, using both the whole sample and only the observations in E (the

last observations of each year). Heteroskedasticity-consistent covariance estimators were used to compute coefficient standard errors in the former two cases (shown in parentheses).

It is obvious from Table 2 that the use of R to diagonalize the error structure has a profound effect on the estimated parameters. Although OLS on the untransformed data results in small sample errors, it does so by overfitting the data. In the case of cotton and spring wheat, OLS resulted in weights that fell as the condition improved. Furthermore, the estimates are on the poor condition for these crops is too low to be believable as an estimate of the average yield in this class. Similarly, the weights estimated using the last observations of each year do not conform to theoretical expectations.

The regressions that use R to transform the data, on the other hand, not only increase as crop category improves, but also have reasonable magnitudes. The estimates with the time varying weights, in particular, are quite reasonable. In addition to satisfying minimal consistency requirements, the values range from poor condition weights of about 50% (spring wheat) to 80% (cotton) to excellent condition weights of 130% (corn) to 145% (soybeans). Furthermore the weights on intermediate condition classes do not cluster but are spread over the poor to excellent range.

The usefulness of the forecast model, of course, depends on how well it forecasts. To provide a benchmark, the forecasts are compared to forecast yields issued by the USDA in the monthly publication *Crop Production*. The comparison is imperfect as the USDA forecasts are true forecasts, whereas the forecasts from the crop condition model are in-sample. Although this caveat should be borne in mind, there is really no other reasonable alternative given the short time period over which condition reports have been issued. For the purposes of the comparison, forecasts were selected for the last date before the USDA forecast was made. Plots of the forecasts are provided in Figures 5-8, with the condition-based forecasts represented by small, connected dots, the USDA forecasts by circles and the final yield estimates by flat lines. Root mean square error comparisons are provided in Table 3.

In general, the condition-based forecasts are about as good as the USDA forecasts early in the season but are not competitive towards the end of the season. This suggests that the condition reports may be most useful in providing an early signal about upcoming yields, but that better forecasts are available latter in the growing season.

Conclusions

This study develops simple methods for generating yield forecasts from readily available crop condition information. The approach differs from a previous effort in several ways. First, estimates are based on data for individual regions rather than pooled across locations and the forecast function does not change over time. The issue of abandonment is addressed, resolving the puzzle that yield forecasts can increase when crop condition worsens.

The study raises a puzzle concerning how to estimate conditional forecasts from information that is revised over time. It is shown that traditional GLS and maximum-likelihood methods are inherently incapable of providing estimates of both conditional forecasts and forecast error variances. Although the puzzle is not resolved, a workable estimation strategy is developed that results in improved estimates relative to those obtained using ordinary least squares. Crop condition-based forecasts compare favorably to USDA estimates early in the crop year and may be useful as an early warning signal.

References

R. Davidson and J.G. MacKinnon, *Estimation and Inference in Econometrics*. Oxford University Press, Oxford. 1993.

Kruse, John R. and Darnell Smith. "Yield Estimation Throughout the Growing Season." In *Applied Price Analysis, Forecasting, and Market Risk Management*, Proceedings of the NCR-134 Conference, Chicago, IL; April 18-19, 1994. B. Wade Brorsen, ed. Department of Agricultural Economics, Oklahoma State University, Stillwater, OK. 1994.

USDA. "Crop Progress/Crop Weather: Terms and Definitions." National Agricultural Statistics Service (NASS), Washington. Web site: http://www.usda.gov/nass/pubs/cwterms.htm. 1998.

USDA. *Crop Production*. National Agricultural Statistics Service (NASS), Washington. Monthly.

Table 1. Crop Condition Reports: Summary Information 1986-1998

	Corn	Cotton	Soybeans	Spring Wheat
Usual planting date	3/25-6/15	3/15-6/30	4/15-7/15	4/1-5/31
Average date of first report	5/30	5/28	6/16	5/21
Average date of last report	10/6	10/14	10/5	8/21
Usual harvest date	8/15-11/30	9/1-12/15	9/15-12/15	7/15-9/30
Average Number of Reports	19.3	20.8	16.8	13.9
Total Number of Reports	251	270	219	181

Usual planting and harvesting dates are estimated from crop progress data.

Table 2. Condition Weight Estimates

	Poor	Fair	Good	Excellent
Corn				
OLS	0.5695	0.6569	1.1681	1.2099
OLS with last obs only	0.7101	0.5569	1.1542	1.3218
Transformed Data	0.6175	0.9011	1.0247	1.2437
	(0.1762)	(0.0684)	(0.0303)	(0.0646)
Transformed Data with weighting	0.6030	0.7501	1.0863	1.2938
	(0.0867)	(0.0639)	(0.0239)	(0.0495)
Cotton				
OLS	0.3148	1.1304	0.9637	1.8141
OLS with last obs only	0.3001	1.0021	1.0812	1.7944
Transformed Data	0.8442	0.9666	1.0791	1.2855
	(0.0762)	(0.0332)	(0.0244)	(0.1117)
Transformed Data with weighting	0.8076	0.9573	1.0865	1.3310
	(0.0808)	(0.0353)	(0.0249)	(0.1201)
Soybeans				
OLS	0.5457	0.8069	1.1540	1.4085
OLS with last obs only	0.7104	0.7160	1.1710	1.4433
Transformed Data	0.7976	0.8878	1.0553	1.3535
	(0.1147)	(0.0492)	(0.0333)	(0.0942)
Transformed Data with weighting	0.7542	0.8475	1.0703	1.4388
	(0.1128)	(0.0426)	(0.0359)	(0.1006)
Spring Wheat				
OLS	0.3311	0.9837	1.0703	1.0521
OLS with last obs only	0.4514	0.7170	1.0555	2.6034
Transformed Data	0.5768	0.9562	1.0727	1.1516
	(0.2158)	(0.0484)	(0.0373)	(0.1550)
Transformed Data with weighting	0.4738	0.8914	1.1200	1.3371
	(0.1671)	(0.0647)	(0.0331)	(0.2420)

Table 3. Root Mean Squared Forecast Errors

	July 1	Aug. 1	Sept.	Oct. 1
Corn				
Condition Based	NA	6.3640	7.3218	6.1145
USDA	NA	7.5180	6.3870	4.2406
Cotton				
Condition Based	NA	51.5720	51.2649	44.5990
USDA	NA	49.5526	38.9753	27.3397
Soybeans				
Condition Based	NA	1.6407	1.8777	1.5753
USDA	NA	1.5916	1.7265	0.8494
Spring Wheat				
Condition Based	3.9454	3.3538	3.0361	NA
USDA	4.2980	2.9626	1.0763	NA

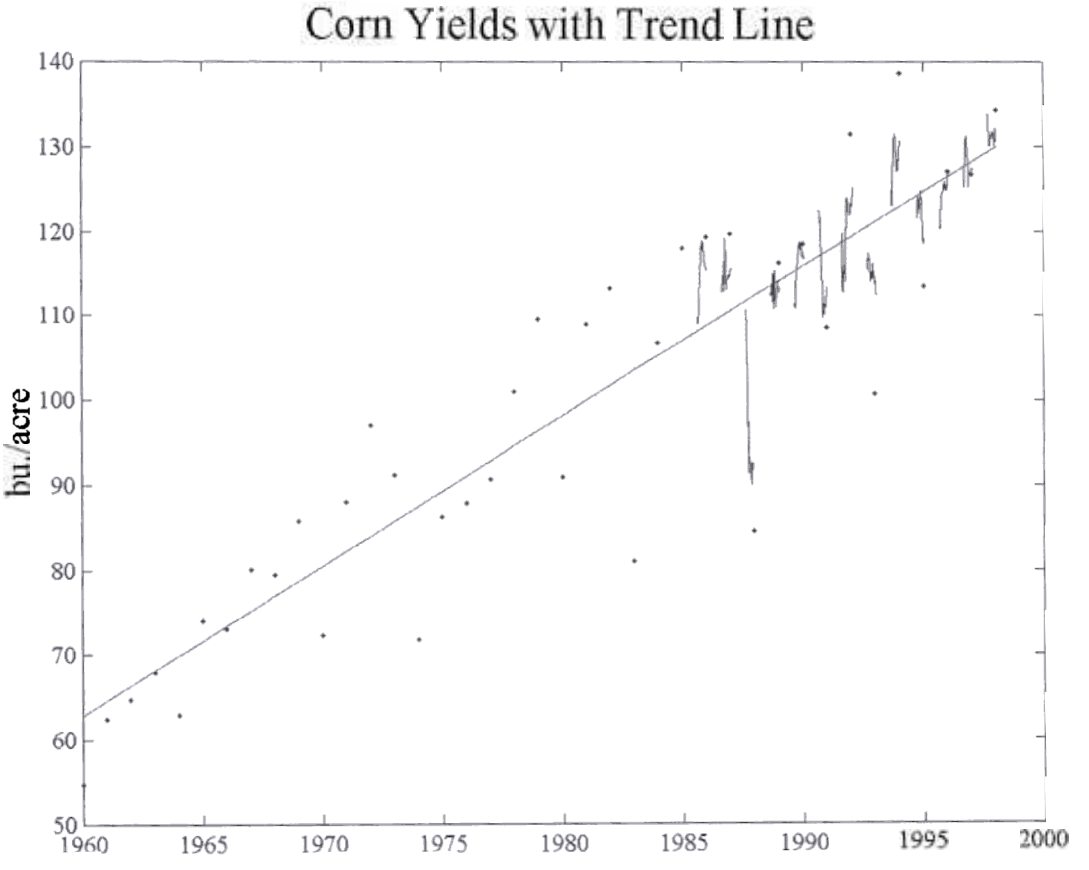

Corn Yields with Trend Line
bu./acre
140
130
120
110
100
90
80
70
60
50
1960
1965
1970
1975
1980
1985
1990
1995
2000

Figure 1

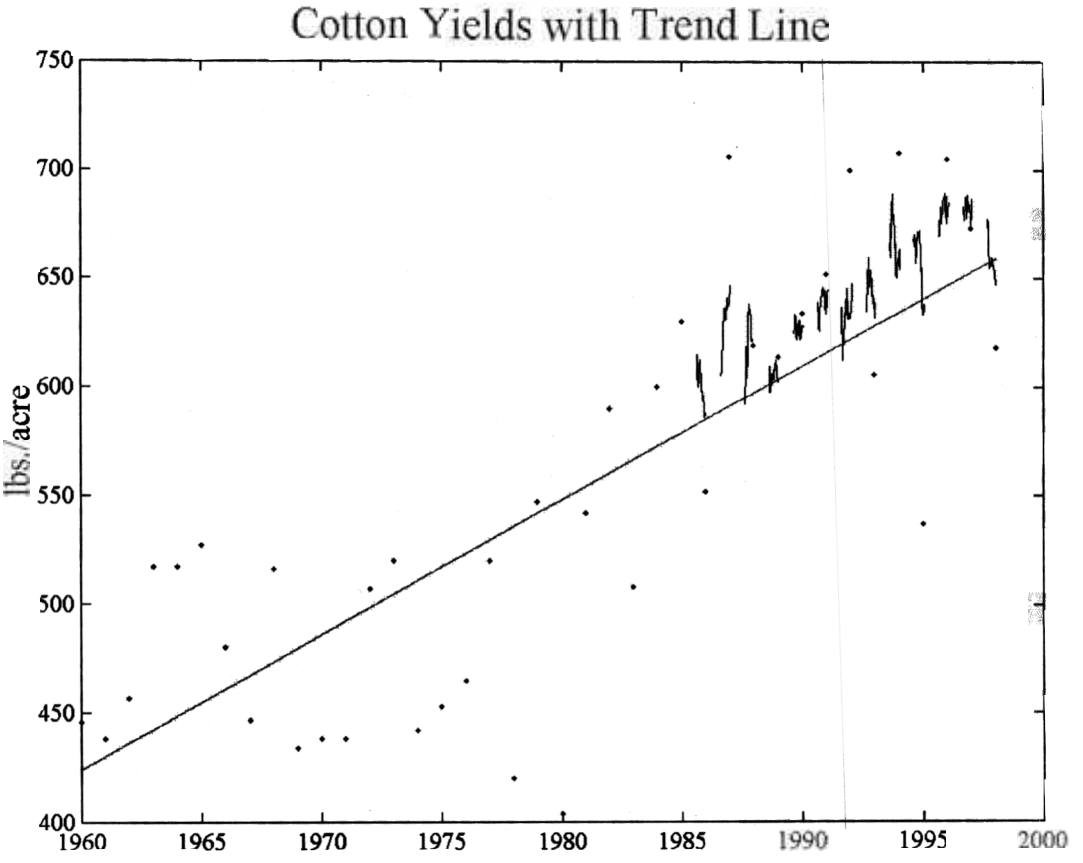

Cotton Yields with Trend Line
lbs./acre
750
700
650
600
550
500
450
400
1960
1965
1970
1975
1980
1985
1990
1995
2000

Figure 2

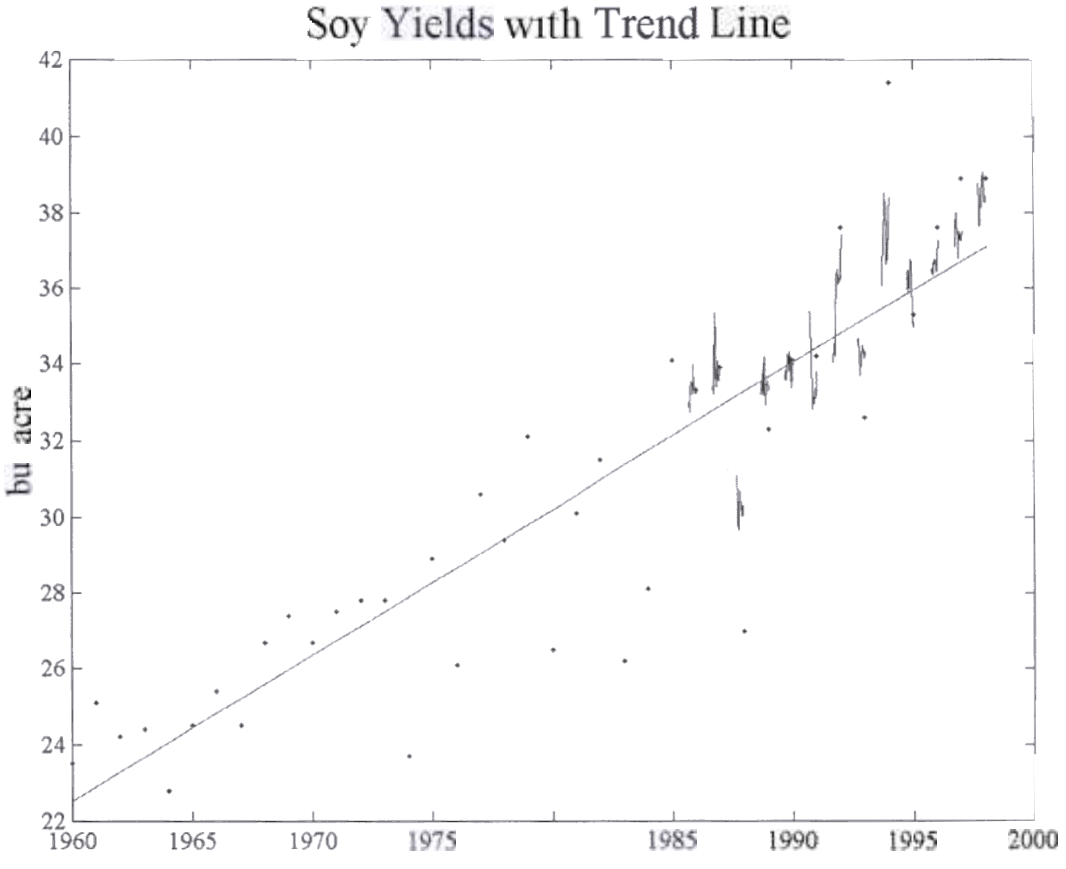

Soy Yields with Trend Line
bu acre
42
40
38
36
34
32
30
28
26
24
22
1960
1965
1970
1975
1985
1990
1995
2000

Figure 3

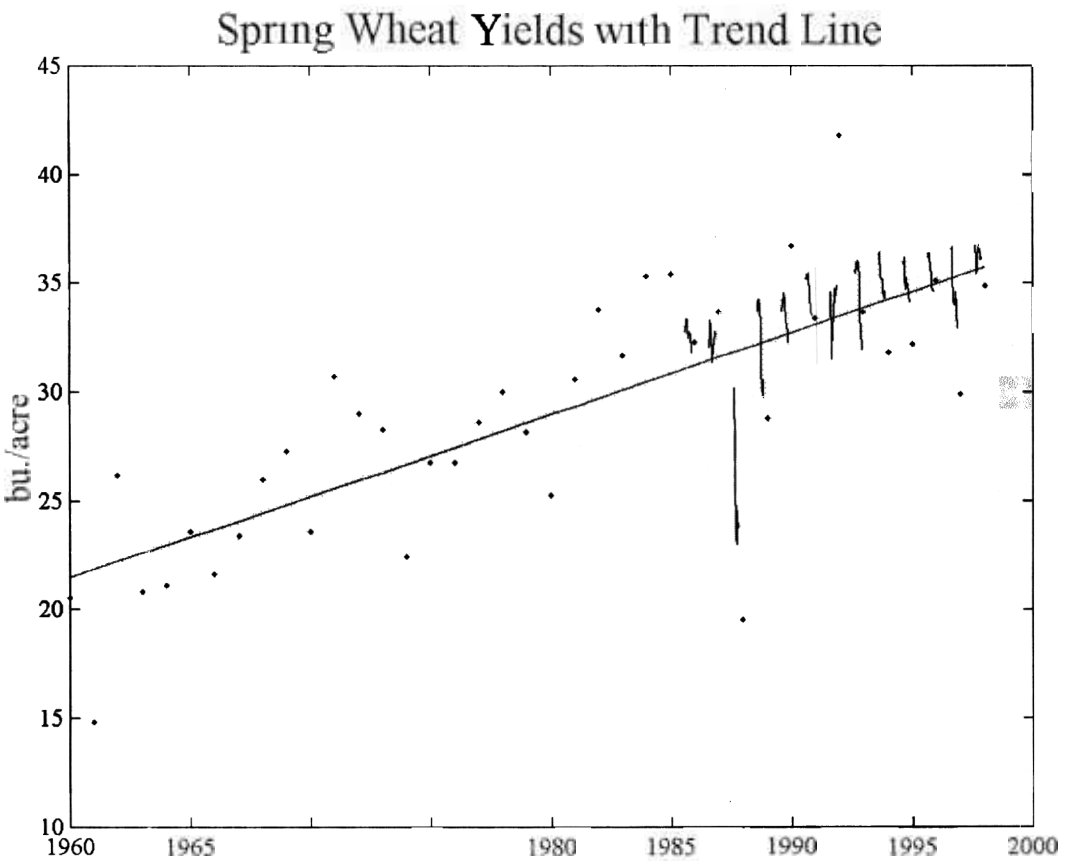
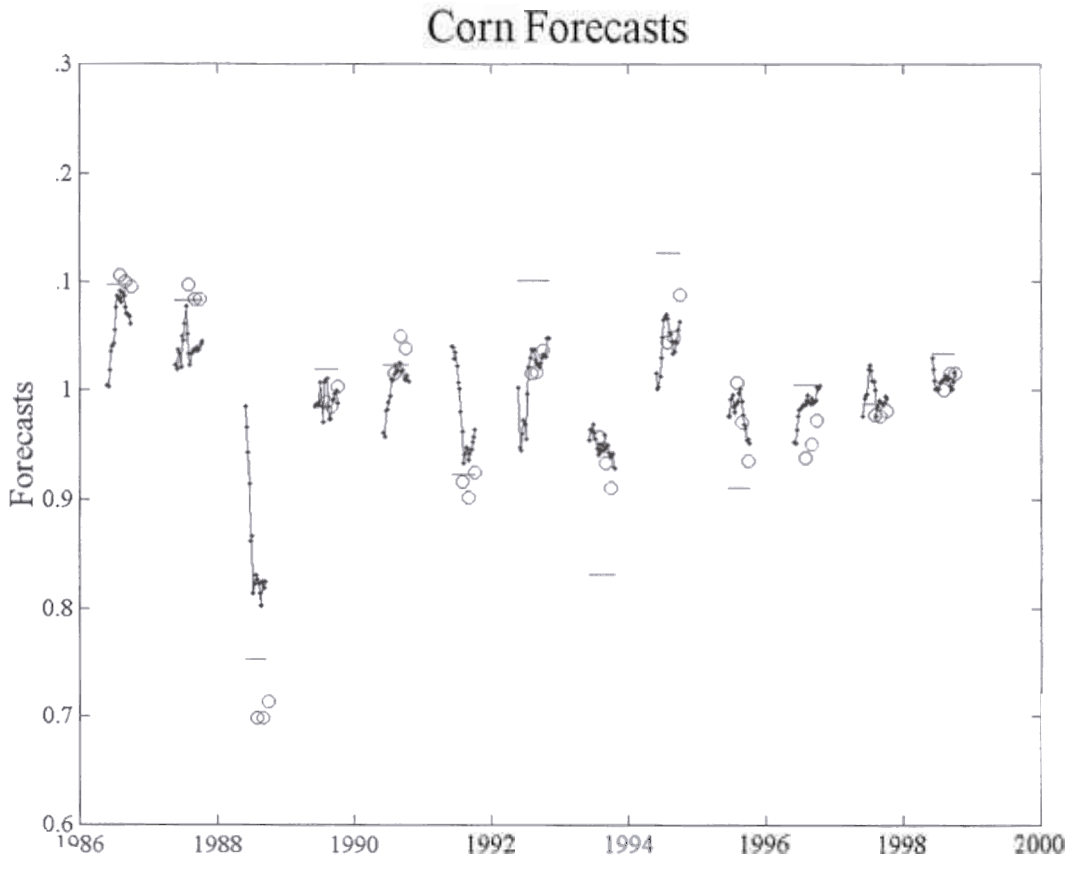

Spring Wheat Yields with Trend Line
bu./acre
45
40
35
30
25
20
15
10
1960
1965
1980
1985
1990
1995
2000

Figure 4


Corn Forecasts
Forecasts
.3
.2
.1
1
0.9
0.8
0.7
0.6
1986
1988
1990
1992
1994
1996
1998
2000

Figure

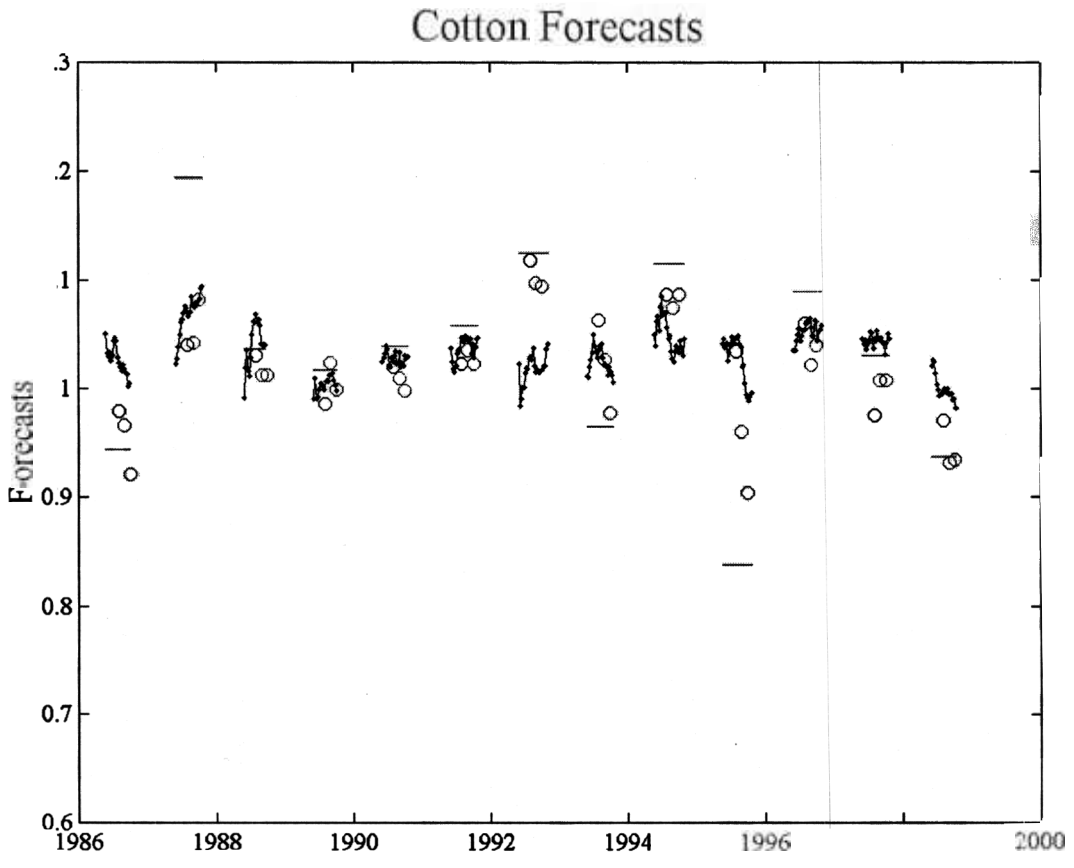
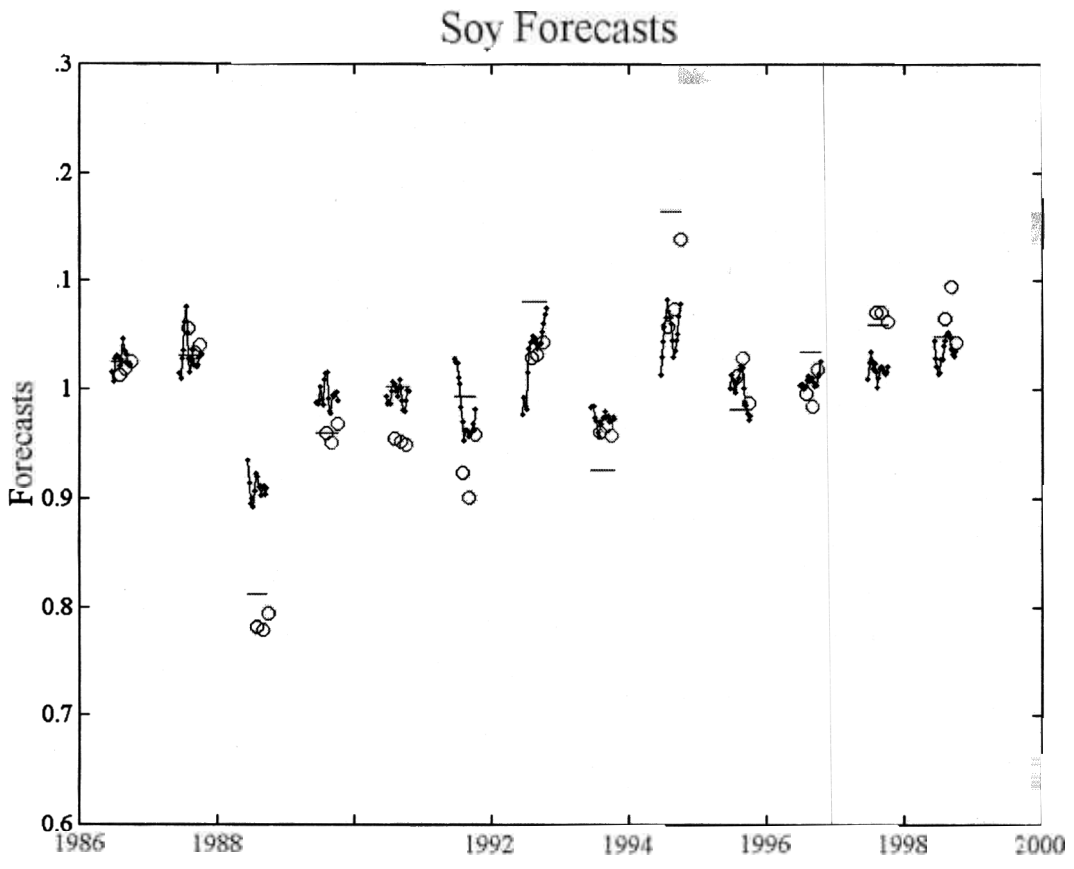
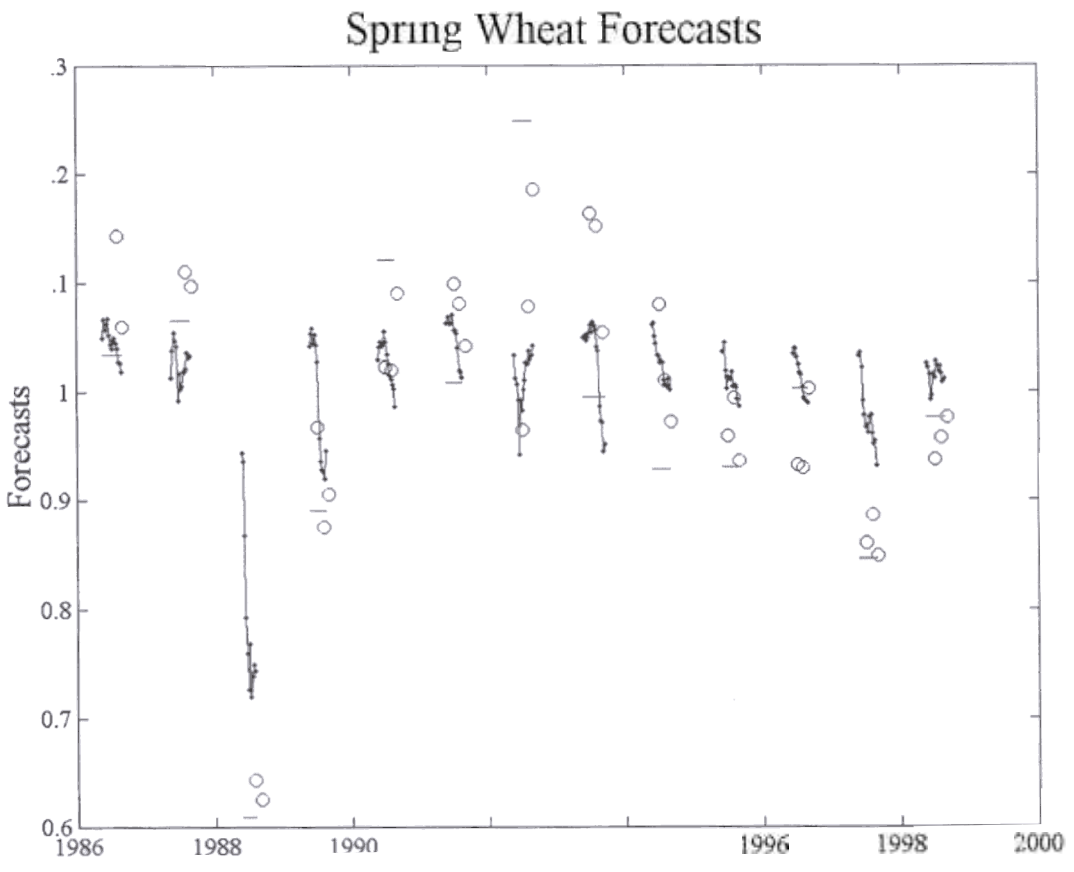

Cotton Forecasts
Forecasts
.3
.2
.1
1
0.9
0.8
0.7
0.6
1986
1988
1990
1992
1994
1996
2000

Figure 6


Soy Forecasts
Forecasts
.3
.2
.1
1
0.9
0.8
0.7
0.6
1986
1988
1992
1994
1996
1998
2000

Figure


Spring Wheat Forecasts
Forecasts
.3
.2
.1
1
0.9
0.8
0.7
0.6
1986
1988
1990
1996
1998
2000

Figure