ESTIMATING THE PRODUCTION OF CROPLANDS

Prepared by:

Sico Stevens A. Yao, MIS

Project Report, May 2006

DEPARTMENT OF NATURAL RESSOURCES AND CONSERVATION
UNIVERSITY OF MONTANA

EXECUTIVE SUMMARY

Revenue generated from the utilization of croplands in Montana State is one of the essential measures to ascertain the good operation of the Department of Natural Resources and Conservation (DNRC). The Trust Land Management Division (TMLD) at the DNRC is responsible for leasing lands to interested farmers in return of a production-based allowance. However, there is no established tracking system allowing the TMLD to estimate farmers' productions other than to rely on the producers' reporting.

In order to help the DNRC manage more efficiently revenues from croplands, we developed a statistical model derived from classical linear regression theory. Widely used research method for identifying significant relationships between variables, the regression model incorporates the intrinsic characteristic of the lands in Montana. Models described in the literature usually relate to value of the land instead of the valuation of the agricultural production obtained from the land. While hedonic models – which express the value of the land as a function of its characteristics – have received a large amount of attention, most of the similar research refers to the all country or to a state other than the state of Montana.

This report not only demonstrates how this statistical method can be applied in practice to deliver business solutions but also illustrates how descriptive statistics and multivariate regression analysis are used to determine an adequate model for predicting farmers' production, hence the allowance due to the state. In particular, we illustrate how a set of 304secondary data can be used to estimate the wheat production in Daniels County. We found that a strong relation exists between production levels and dimensions of farm lands and their locations expressed in terms of township and range. Surprisingly, there was no evidence on the type of agricultural practice (continuous cropped vs. summer fallow) having an effect on the production level. Results from the model depicted fourteen cases of production below level of tolerance (α=.05) which need to be placed under investigation and/or require close attention.

TABLE OF CONTENTS

List of Tables

List of Illustrations

1. INTRODUCTION

The Montana Department of Natural Resources and Conservation (DNRC) is a public agency which mission is to provide technical, financial, and administrative assistance to public and private entities to complete projects that put renewable resources to work, increases the efficiency with which natural resources are used, and solve recognized environmental problems. In particular, the Trust Land Management Division (TMLD) of the DNRC aims to manage the State of Montana's resources to produce revenue for the trust beneficiaries. In order to fulfill its mission, the TMLD of the DNRC have shown interest in developing a model that would help estimate the production of dry farmlands in the different county of the state of Montana. Such a model will allow the DNRC to not assess the integrity of the farmers' reporting of their production on which the state revenue is based on, but also eventually help less skilled farmers to optimize their productivity.

Even though the main purpose of our research project will be centered on investigating whether there is a limit under which a level of production can be designated as suspicious, the following hypothesis will be tested as well:

- Does the level of production vary with the type of agricultural practice (continuous cropped vs. summer fallow)?
- Is there a significant difference in the production level depending on farms' locations (township vs. section vs. range)?

We first begin our discussion by providing a description of the data used in the model suggested. Then, we will present different plots used to detect farmers that are underreporting the value of their production and other substantive issues regarding any violations of the model assumptions. Finally, we will discuss the result of our analysis and the implication of the state of Montana.

2. DATA

The data combines 304 observations on the farmland proprieties gathered by the Land Management Division of the Department of Natural Resources and Conservation from secondary sources. This set of cross-sectional data collected in 2004 includes some information describing the geographic location of the farm lands and its method of utilization. The data also contain a few characteristics of the production and incorporate a component for the current leaseholder. The original raw data are composed of 17 variables. Even though some of the variables included in the data set are self-explanatory, other data may require a more comprehensive definition.

2.1 Variables definition

Among the 17 variables, the variables PRODUCT (production), CALYIELD (calculated yield) and CAL$ACRE (calculated dollar per acre) are all candidates for the dependent variables whereas COUNTY (county), LEASE (lease), TYPE (type), COMMOD (commodity), PRACTICE (practice), SEC (sector), TWP (township), TD (township direction), RNG (range), RD (range direction), LEGAL (legal), ACRES (acre) , SHARE (share) and UNIPRICE (unit price) define the potential independent variables for our model. Below, we provide a complete definition of each of those variables in the order they appear in the SPSS variable view.

2.1.1 Leaseholder information

LEASE (lease): LEASE is the only variable that relates the farm land to the leaseholder. It is a five digit number that identifies the current producer. A given producer can manage more than one portion of land.

2.1.2 Farm land geographic location

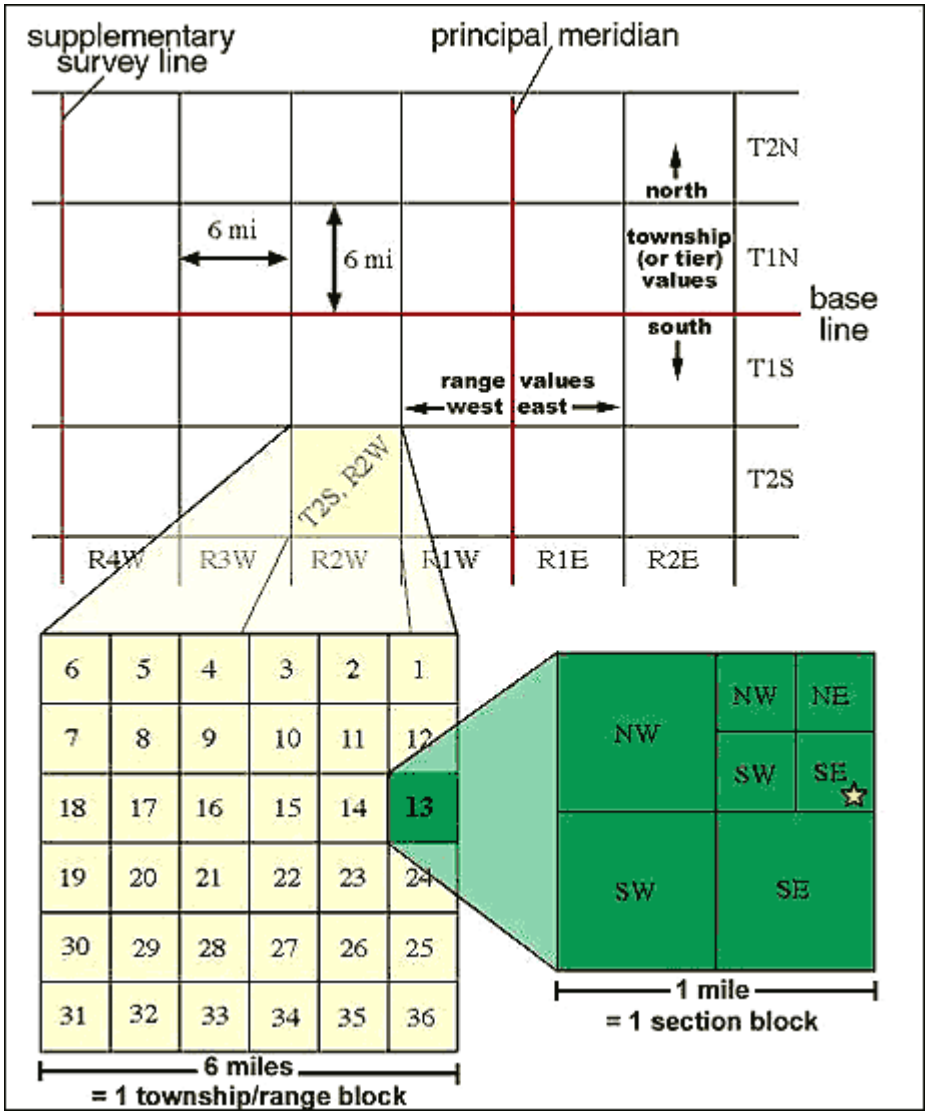
The farm land geographic information is based on the Public Land Survey (PLS) system. It is important to understand PLS system in the context of this research project in order to better understand the data. This understanding will provide us with both a good guidance for making some in depth analysis and a better sense of judgment at the interpretation stage.

The PLS system is a method used in the United States to locate and identify land, particularly for titles and deeds of farm or rural land. States are divided into townships which the primary units of area.

Figure 1: Township Addressing System

6	5	4	3	2	1
7	8	9	10	11	12
18	17	16	15	14	13
19	20	21	22	23	24
30	29	28	27	26	25
31	32	33	34	35	36

Figure 2: PLS System



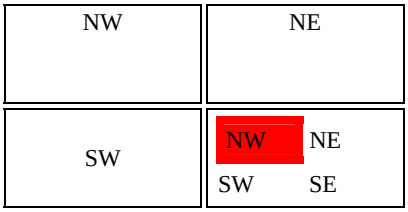
Each township is composed of 36 square miles called sections. Sections constitute the second unit of area in the PLS system and measure one square mile each. They are numbered 1 through 36 in a serpentine fashion from the North East corner of the township and are always laid out like illustrated in figure 1.

Townships are based from a central set of orthogonal axis established by early surveyors, the North-South axis is called the 'Principal Meridian', and the East-West axis is

called the 'Base Line'. Townships are consecutively numbered to the North and South and Ranges are numbered from East to West as shown in the figure 2.

For legal descriptions (and for finding fossil sites) each township can be further subdivided according to quadrant and compass directions. Those can be divided even further, for example the legal description or the fossil site might be listed in literature as the NW 1/4, SE 1/4 corresponding to the section marked in red in figure 3.

Figure 3: Legal Description Notation



COUNTY (county): the alphabetic variable COUNTY indicates the county in which the farm land in located in the state of Montana. The only county in this particular case is 'Daniel'.

SEC (section): the variable SEC represents the section of the land. It is one of the 36 square mile subdivisions of a given township. It is an identification code variable which can take any number between 1 and 36.

TWP (township): TWP is a two-digit number which designates the relative position of the portion of land from the principal meridian of the state.

TD (township direction): it is a variable that provides the orientation of the township which can be N for north or S for south.

RNG (range): RNG is a two-digit number that indicates the position of the farm land relative to the base line of the state.

RD (range direction): it is a variable that provides the orientation of the township which can be E for east or W for west.

LEGAL (legal): LEGAL is an alpha numeric variable that defines a particular farm land referred on the state map by the PLS system.

2.1.3 Farm land utilization method

TYPE (type): TYPE corresponds to the season during which the land is being utilized. It is an alphabetic variable that can take the value spring or winter. However, it only takes the value 'spring' for the given set of data.

COMMOD (commodity): COMMOD is an alphabetic variable that can take one of the following values: durum, wheat, oats, etc. It designates the type of the production which is 'wheat' for all producers included in the data set.

PRACTICE (practice): PRACTICE is a two-character alphabetic variable which characterizes the type of agricultural technique being used on the cropland. PRACTICE can take two values which are CC for continuous cropped or SF for summer fallow. Summer fallowing is a farming practice used in semi-arid regions which purpose is to keep farm land out of production during a cropping season mainly to conserve moisture for the next season. It is common for wheat producers to rotate half their cropland to summer fallow each year.

2.1.4 Production characteristics

PRODUCT (production): product stands for production. It is a numeric variable which represents the total production that was harvested and is measured in bushels.

ACRES (acres): ACRES is a numeric variable expressed in acre. It represents the dimensions of the cropland for a given production.

SHARE (share): share is a numeric variable which correspond to the percentage rate of the total value of the crop that the state receives as return for the lease. Although, the minimum

required by the law is 25%, this rate may be higher due to competitive bidding from other producers.

UNIPRICE (unit price): UNIPRICE is the value of a bushel as determined by the local grain elevator.

CALYIELD (calculated yield): CALYIELD is the production by acre. It is a ratio which is computed as follows: PRODUCT/ACRES.

CAL$ACRE (calculated dollar per acre): CAL$ACRE is another ratio which corresponds to the return on a per acre basis the state receives from the producer. It is obtained as follows: [(PRODUCT * UNIPRICE)/ACRES] * SHARE.

2.2 Data description

Table 1 summarizes the raw data used in our analysis and includes descriptive statistics of the numeric variables. The range, minimum, maximum, mean, standard deviation and variance of the variables are shown in the table.

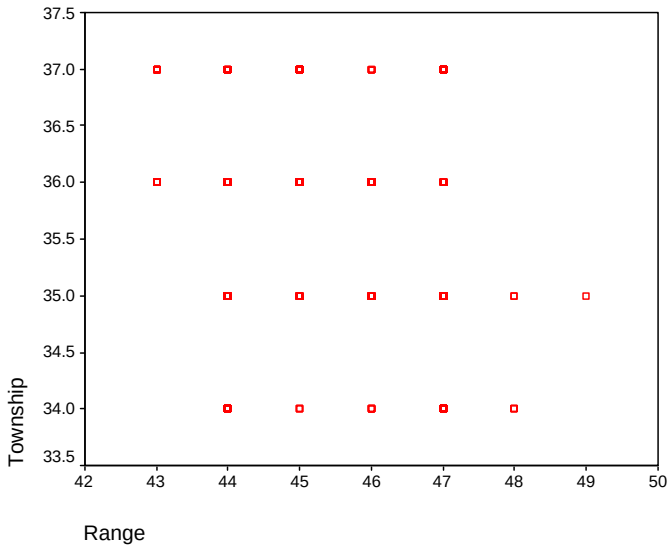
Table 1: Descriptive Statistics

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean		Std.	Variance
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic
Producer Identification	304	10495	34	10529	3319.06	154.20	2688.597	7228551
Section	304	35	1	36	18.32	.64	11.125	123.763
Township	304	3	34	37	35.70	.05	.941	.885
Range	304	6	43	49	45.33	.08	1.375	1.891
Production	304	17425	75	17500	2490.81	152.07	2651.354	7029681
Dimension in Acres	304	619	3	622	98.61	5.55	96.815	9373.140
Share	304	.08	.25	.33	.2503	.0003	.00478	.000
Unit Price per Bushel	304	$4.72	$.06	$4.78	$2.0416	$.0403	$.70238	.493
Calculated Yield	304	46.95	.80	47.76	25.2101	.4996	8.71023	75.868
Calculated $ per Acre	304	$30.54	$.01	$30.55	$13.0271	$.3690	$6.43343	41.389
Valid N (listwise)	304							

A scatter plot allows us to identify 21 different locations of 36 miles square each, in which farm lands are situated. They are shown in the graph below. Interestingly, the contour of the dot on the graph reveals the approximate shape of the county of interest.

Figure 4: Townships Distribution in Daniel County



The next three (3) tables include descriptive statistics summaries of the production per acre (CALYIELD) organized by township (TWP), range (RNG) and township and range. It appears that farmers located in both townships 34, 35 and range 43, 44 tend to be less productive with calculated yields falling below average.

Table 2: Descriptive Statistics (Township vs. Calculated Yield)

Descriptive Statistics

Township		N	Minimum	Maximum	Mean	Std. Deviation
34	Calculated Yield	38	4.71	31.57	20.4474	6.50913
	Valid N (listwise)	38				
35	Calculated Yield	80	.80	42.00	23.7801	9.50577
	Valid N (listwise)	80				
36	Calculated Yield	122	4.12	47.76	27.0810	8.19477
	Valid N (listwise)	122				
37	Calculated Yield	64	6.57	43.80	26.2592	8.57579
	Valid N (listwise)	64				

Table 3: Descriptive Statistics (Range vs. Calculated Yield)

Descriptive Statistics

Range		N	Minimum	Maximum	Mean	Std. Deviation
43	Calculated Yield	28	11.46	34.21	21.9909	5.80358
	Valid N (listwise)	28				
44	Calculated Yield	72	.80	40.00	21.2968	9.50856
	Valid N (listwise)	72				
45	Calculated Yield	65	7.28	47.76	25.8982	8.81832
	Valid N (listwise)	65				
46	Calculated Yield	57	15.99	41.59	29.3899	6.50792
	Valid N (listwise)	57				
47	Calculated Yield	76	6.57	45.00	26.1868	8.76470
	Valid N (listwise)	76				
48	Calculated Yield	5	26.23	31.11	27.9615	2.15683
	Valid N (listwise)	5				
49	Calculated Yield	1	26.16	26.16	26.1558	.
	Valid N (listwise)	1				

Table 4: Descriptive Statistics (Township vs. Range vs. Calculated Yield)

Descriptive Statistics

Township	Range		N	Minimum	Maximum	Mean	Std. Deviation
34	44	Calculated Yield	17	4.71	27.81	17.7221	5.44095
		Valid N (listwise)	17				
	45	Calculated Yield	3	8.19	19.96	14.9772	6.08579
		Valid N (listwise)	3				
	46	Calculated Yield	3	22.58	29.49	26.7984	3.69562
		Valid N (listwise)	3				
	47	Calculated Yield	12	8.66	31.57	22.2356	6.30014
		Valid N (listwise)	12				
	48	Calculated Yield	3	26.23	31.11	27.8576	2.81392
		Valid N (listwise)	3				
35	44	Calculated Yield	23	.80	40.00	19.6939	10.27919
		Valid N (listwise)	23				
	45	Calculated Yield	14	12.00	40.00	23.2812	9.62751
		Valid N (listwise)	14				
	46	Calculated Yield	17	18.00	33.41	24.2914	4.27369
		Valid N (listwise)	17				
	47	Calculated Yield	23	11.28	42.00	27.3115	10.85425
		Valid N (listwise)	23				
	48	Calculated Yield	2	26.96	29.28	28.1174	1.64028
		Valid N (listwise)	2				
	49	Calculated Yield	1	26.16	26.16	26.1558	.
		Valid N (listwise)	1				
36	44	Calculated Yield	25	4.12	38.70	24.0474	10.06769
		Valid N (listwise)	25				
	45	Calculated Yield	30	7.28	47.76	26.6667	7.96540
		Valid N (listwise)	30				
	46	Calculated Yield	33	20.11	41.59	32.6201	5.46071
		Valid N (listwise)	33				
	47	Calculated Yield	14	20.00	45.00	28.8133	6.11235
		Valid N (listwise)	14				
	43	Calculated Yield	20	15.04	30.06	21.1423	4.69093
		Valid N (listwise)	20				
37	44	Calculated Yield	7	13.56	38.66	25.4212	10.03269
		Valid N (listwise)	7				
	45	Calculated Yield	18	15.02	43.80	28.4728	8.63333
		Valid N (listwise)	18				
	46	Calculated Yield	4	15.99	37.61	26.3529	9.21447
		Valid N (listwise)	4				
	47	Calculated Yield	27	6.57	37.94	25.6230	8.60607
		Valid N (listwise)	27				
	43	Calculated Yield	8	11.46	34.21	24.1122	7.93645
		Valid N (listwise)	8				

3. METHODOLOGY

Linear regression is the statistical method that we retained for the purpose of our research project. Popular by its simplicity, it is the technique which will best fit the data at hand by incorporating the relevant variables into the chosen model. In addition, we will make use of ANOVA in appropriate situations.

4. DATA ANALYSIS AND RESULTS

4.1 MODEL AND HYPOTHESIS

It is our goal to capture the variables that are relevant in the value estimation in Daniels County. In the context of our research, we suggest a classical linear regression model. The model classical linear regression model supposes for the stochastic term to be normally distributed and to all have same variance.

We retained PRODUCT (production) as the dependent variable for the model. However, we can see from the histogram in figure X that variable is skewed. Although most of the productions are between 1,000 to 2,000 bushels, a few are as high as 17, 000 bushels. In general, such skewness in the dependent variable can be problematic and may result in problems with violations of the assumptions of the linear model as attested in figure 5 and 6.

Figure 5: Histogram of Production as Dependent Variable

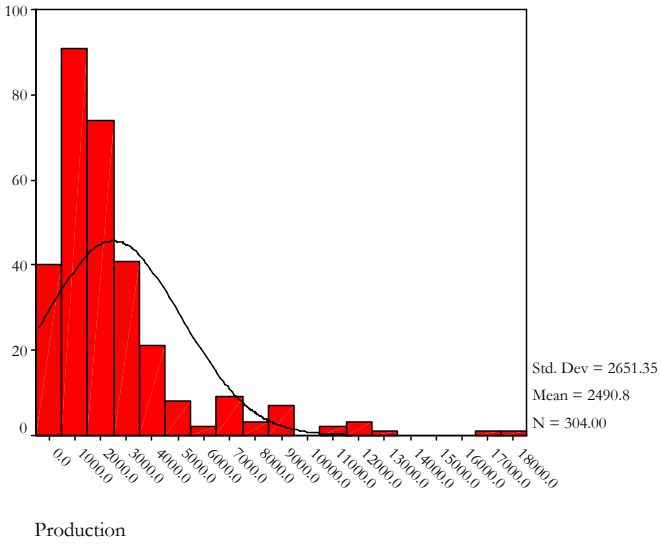
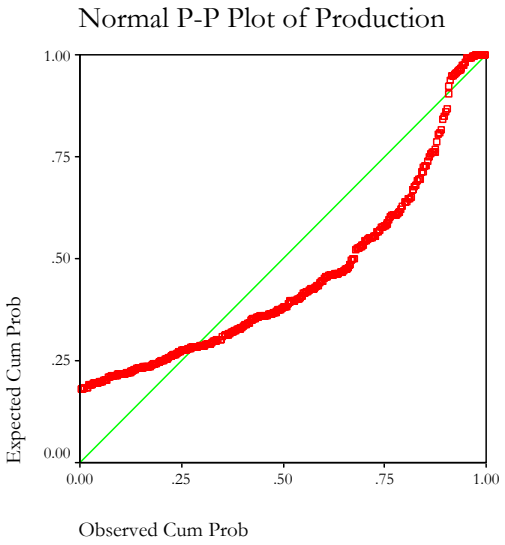


Figure 6: Normal P-P Plot of Production



One way to address this issue is to use a transformation to reduce the skewness of the dependant variable. The distribution and the normal P-P plot of the log transformation are both shown in figure 7 and 8 respectively.

Figure 7: Histogram of Log of Production as Dependent Variable

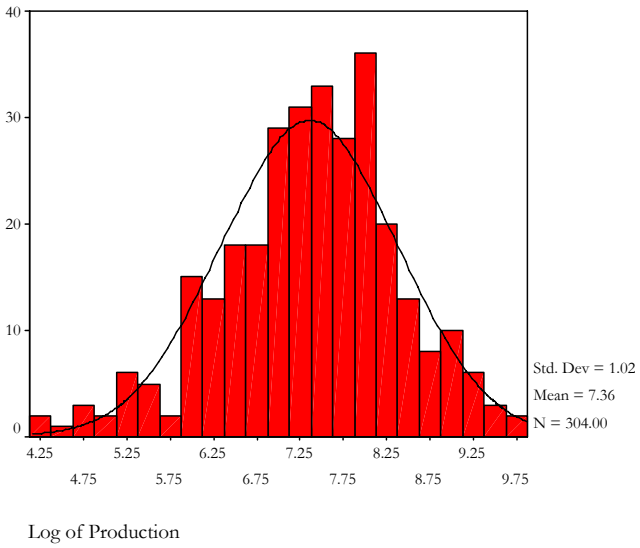
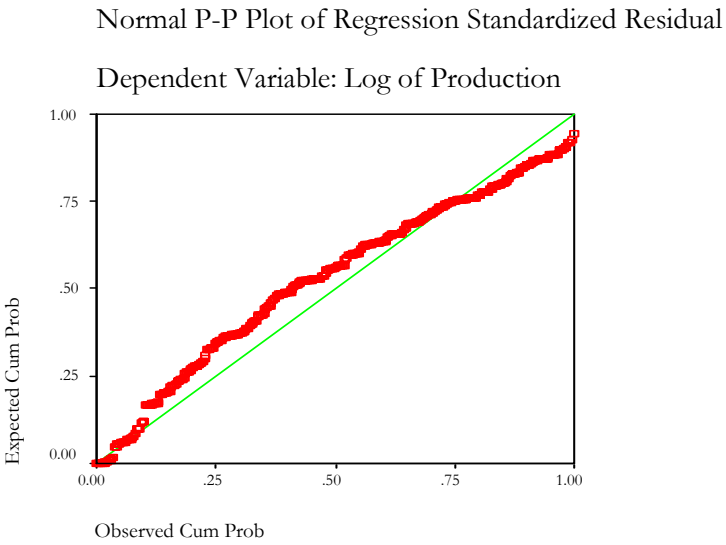


Figure 8: Normal P-P Plot of Log of Production



We have adopted a log-linear model with LOG (ACRES) as independent variable. Not only this functional form has the advantage of offering a clear interpretation for X, but also offers us with a model that fits our data. The graph below in figure 9 provides us with a good illustration.

The scatter plot in figure 9 also shows that the croplands are randomly distributed around the fitted line regardless the type of practice (SF = 1 and CC = 0) they are subject to. The deviation in the production between the types of practice is not perceptible and might not be statistically significant.

Figure 9: Scatter Plot of Log of Production vs. Log of Acre marked by Practice Code (SF=1 and CC=0)

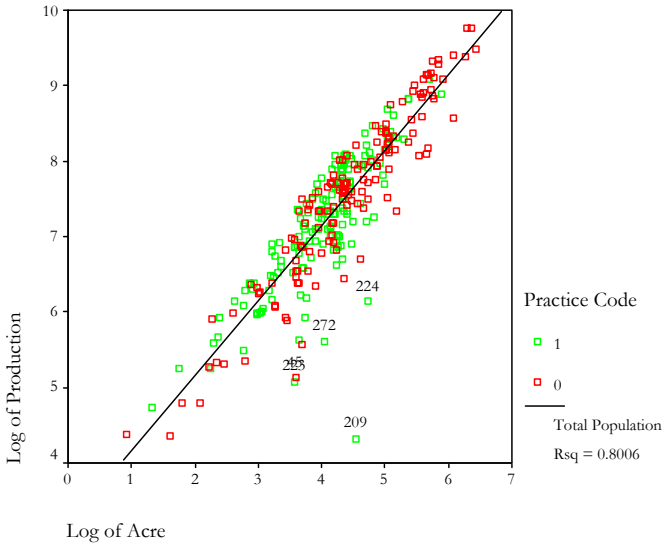


Figure 10: Scatter Plot of Log of Production vs. Log of Acre marked by Range

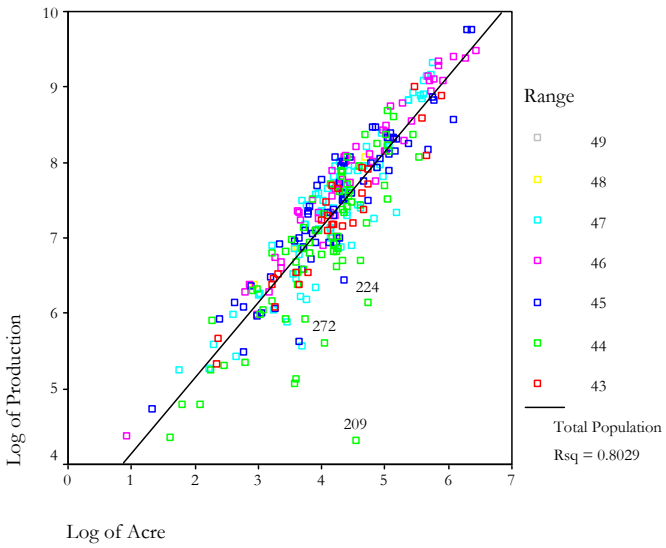
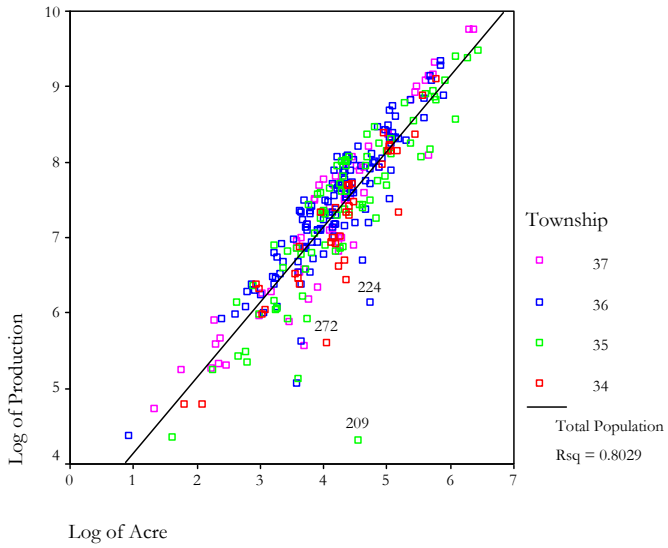


Figure 11: Scatter Plot of Log of Production vs. Log of Acre marked by Township



Furthermore, a detailed analysis of the Scatter Plot of Log of Acre against Log of Production marked by Range and Township respectively in figure 10 and 11 does not capture evidence of some patterns. Though, the farm lands in range 48 appear in average to be above the fitted line while all noticeable outliers seem to belong to the township at range 44. The observations that seem to the most apart from the fitted line correspond to the observations 45, 209, 223, 224 and 272. In order to investigate those observations, we suggest referring to figure 12 and 13.

Figure 12: Scatter Plot of SD residual vs. standardized Predicted

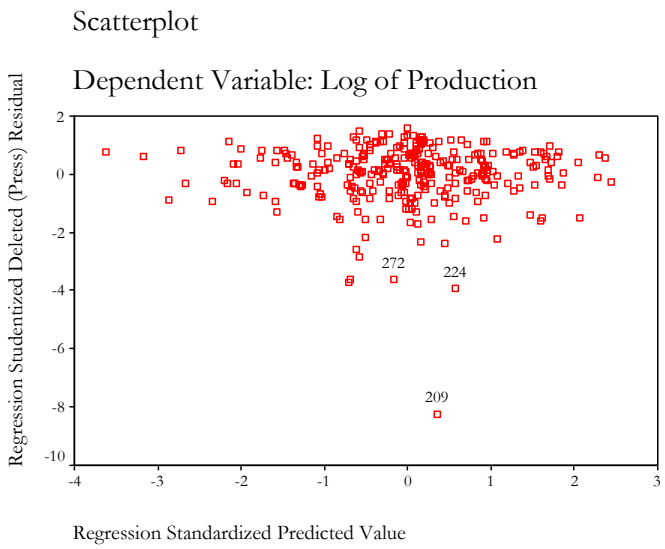
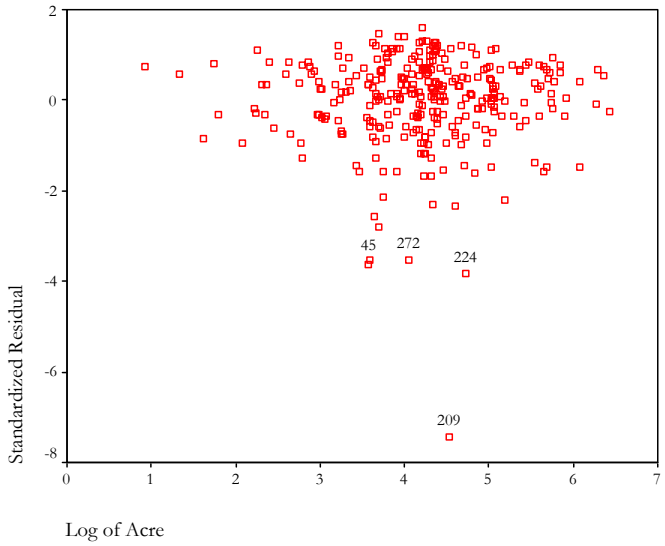


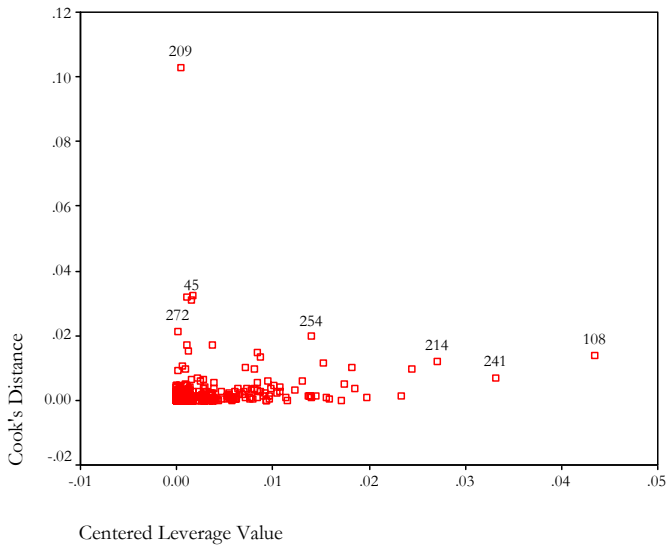
Figure 13: Scatter Plot of SD residual vs. Log of Acre



First, the plot of residuals by the predicted also shows the variance of the errors terms do not increase since we have a good scatter. Therefore, we do not violate the assumption of homoscedasticity. The plot of residuals by LOG (ACRE) also confirms this

result. Second, the plot of residuals by LOG (ACRE) both highlight the observations we were concern with. The point that requires the most of our attention is without any doubt the observation 209. This case has a low leverage and high influence as indicated in Figure 14.

Figure 14: Scatter Plot of Cook's Distance vs. Leverage value



We are not really concern with the effect of the observation 108 on the regression line because of the low centered leverage value (less than .05). To assess the magnitude of the influence of case 209 (D_{209} = .11), we refer to the corresponding distribution, namely, F (209, 95) since in linear models Cook's Distance has, approximately, an F distribution with k and (n-k) degrees of freedom. We find that .11 is the $6.06*10^{-38th}$ percentile of this distribution. Hence, it appears that the observation 209 does not influence the regression fit.

Although Figure 10 and 11 did not reveal any influence of both factors township (TWP) and range (RNG) on the production level, Table 4 suggests there is an apparent difference in productivity based on the location. However, the effect of the variable TWP

and RNG on CALYIELD is not linear which leads us to create two dummy variables defined as follows:

- NS (North-south) = 0, if TWP=34 or 35

  NS (North-south) = 1, otherwise

- WE (West-East) = 0, if RNG=43 or 44

  WE (West-East) = 1, otherwise

Based on the preliminary observations, we propose the following log-linear model:

$$\text{Log}^{1}\ (\text{PRODUCT}_i) = ß_0 + ß_1\text{Log}\ (\text{ACRE}_i) + ß_2\text{WE}_i + ß_3\text{NS}_i + \mu_i \quad (1)$$

We hypothesize a positive linear relationship between Log (PRODUCT) (Log of production) and Log (ACRES) (Log of acres) since farmers with bigger portions of land are more likely to have a larger production, other things the same. We predict $ß_1$ to be around 1 as there is a direct proportionality between PRODUCT and ACRES. Furthermore, we anticipate a positive sign for both the coefficient on West-East (WE) and North-South (NS).

4.2 RESULTS

Equation (1) was estimated using ordinary least squared (OLS) in SPSS version 13. The results of the regression analysis are represented below in Table 5 which is subdivided into four different charts (Model Summary, Excluded Variables, Coefficients and ANOVA). Table 5 includes the results of three different models that were included for comparative purposes. Among the three models taken into consideration, the theory suggests to retain Model 3 for the purpose of our project.

[1] This is a natural logarithm. We also assume $\text{PRODUCT}_i > 0$ and $\text{ACRE}_i > 0$.

Table 5: OLS Estimates Summary

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.896[a]	.803	.802	.45280
2	.906[b]	.821	.820	.43244
3	.911[c]	.830	.828	.42254

a. Predictors: (Constant), Log of Acre

b. Predictors: (Constant), Log of Acre, WestEst

c. Predictors: (Constant), Log of Acre, WestEst, NordSouth

Excluded Variables[c]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance
1	NordSouth	.097[a]	3.867	.000	.218	.996
	WestEst	.135[a]	5.487	.000	.302	.982
2	NordSouth	.093[b]	3.908	.000	.220	.995

a. Predictors in the Model: (Constant), Log of Acre

b. Predictors in the Model: (Constant), Log of Acre, WestEst

c. Dependent Variable: Log of Production

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.128	.123		25.367	.000
	Log of Acre	1.005	.029	.896	35.074	.000
2	(Constant)	3.019	.119		25.271	.000
	Log of Acre	.984	.028	.878	35.640	.000
	WestEst	.292	.053	.135	5.487	.000
3	(Constant)	2.874	.122		23.472	.000
	Log of Acre	.991	.027	.884	36.652	.000
	WestEst	.287	.052	.133	5.513	.000
	NordSouth	.195	.050	.093	3.908	.000

a. Dependent Variable: Log of Production

ANOVA[d]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	252.231	1	252.231	1230.213	.000[a]
	Residual	61.919	302	.205		
	Total	314.150	303			
2	Regression	257.862	2	128.931	689.453	.000[b]
	Residual	56.288	301	.187		
	Total	314.150	303			
3	Regression	260.589	3	86.863	486.527	.000[c]
	Residual	53.561	300	.179		
	Total	314.150	303			

a. Predictors: (Constant), Log of Acre

b. Predictors: (Constant), Log of Acre, WestEst

c. Predictors: (Constant), Log of Acre, WestEst, NordSouth

d. Dependent Variable: Log of Production

In the next section, we will provide an interpretation of our results as well as the limitations of our study and some orientations for future research.

5. DISCUSSIONS

In order to capture the effect of the location on the production, equation (1) can be simplified and rewritten respectively for farmers in the Northwest (WE = 0, NS = 1) and Northeast (WE = 1, NS = 1) of Daniels County as:

$$\text{Log}(\text{PRODUCT}_i) = (\beta_0 + \beta_3) + \beta_1 \text{Log}(\text{ACRE}_i) + \mu_i \quad (2)$$

$$\text{Log}(\text{PRODUCT}_i) = (\beta_0 + \beta_2 + \beta_3) + \beta_1 \text{Log}(\text{ACRE}_i) + \mu_i \quad (3)$$

Likewise, equation (1) can be reduced and re-expressed for farmers situated in the Southeast (WE = 1, NS = 0) and the Southwest (WE = 0, NS = 0) of Daniels County respectively as follows:

$$\text{Log}(\text{PRODUCT}_i) = (\beta_0 + \beta_2) + \beta_1 \text{Log}(\text{ACRE}_i) + \mu_i \quad (4)$$

$$\text{Log (PRODUCT}_i) = \beta_0 + \beta_1 \text{Log (ACRE}_i) + \mu_i \quad (5)$$

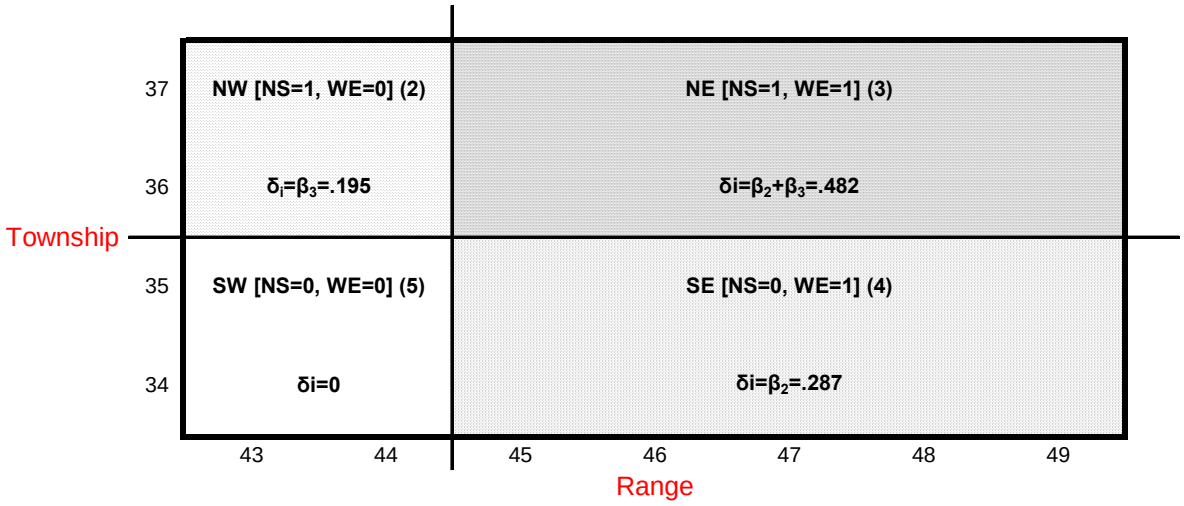
Although all three models are statistically significant, Model 3 has an adjusted R square of .828 which is a little higher than the two other models (.820 and .802). We also found all the coefficients of our model to be significant as reported in the coefficient chart in Table 5. Moreover, our results shows both the coefficient of Log (ACRE) (.991) to be close to 1 and all coefficients to be positive as expected.

According to equation (1), 1% increase in the dimension of the cropland implies a .991% increase in the production at the means, other things the same. In terms of absolute change (or impact at the margin), it means that 1 acre increase in the dimension of the cropland implies a 1.732 bushel[2] increase in production at the means, others things the same.

The impact of the location on our model is only observable at the intercept level. If we treat "being in the Southwest" or β_0=2.874 as the bench mark category, the differential intercepts δ_i equal to β_2, β_3 and (β_2+β_3) tell how much the intercepts of the other three categories differ from the intercept of the base category as illustrated in the perceptual map in figure 15. For instance, farmers located in the Northwest (NW) of Daniels County have their production increased by a factor of $e^{.195}$=1.215 in average compare to the farmers located in the Southwest (SW) Likewise, farmers located in the Southeastern (SE) part of the County are likely to produce $e^{.287}$=1.332 times more crops in average than farmers in the Southwest. But the region the most favorable to growing wheat remains the Northeast (NE) where farmers' production is greater by a factor of $e^{.482}$=1.618 compared to the reference category.

[2] $0.991 * \frac{\sum Log(\text{Pr } oduct)}{\sum Log(Acres)} = 0.991 * \frac{7..357}{4.209}$

Figure 15: Perceptual Map of the Four Distinct Regions in Daniels County



At this time, we are interested in estimating the production of wheat. Determining the expected production will provide us with a base line to detect possible reporting irregularities. We found that the wheat production in Daniels County can be expressed as follows:

$$\text{Log}^{3}(\text{PRODUCT}_i) = 2.874 + .991*\text{Log}(\text{ACRE}_i) + .287*\text{WE}_i + .195*\text{NS}_i + \mu_i \quad (6)$$

The standard error of forecast is equal to .42254, which means that a 95 percent confidence interval would be plus or minus 1.96 standard error around our forecast value for Log (PRODUCT$_i$), or **{2.874 + .991*Log (ACRE$_i$) + .287*WE$_i$ + .195*NS$_i$ ± .828} (7)**. We now need to transform our forecast so that it is interpretable in terms of the original dependent variable. Retransforming the data affects the distribution of the error term μ_i. We have made the assumption when modeling Log (PRODUCT$_i$) that μ_i is normally distributed.

[3] This is a natural logarithm. We also assume PRODUCT$_i$ > 0 and ACRE$_i$ > 0.

Thus, μ_i becomes a log-normally distributed error tem when we transform Log ($PRODUCT_i$) back to $PRODUCT_i$. Assuming that $E(\mu_i) = 0$, equation (7) can be rewritten in terms of $PRODUCT_i$ as described below:

$$PRODUCT_i = {}^4\exp\{E(X) + \frac{\sigma^2}{2}\} = \exp\{\text{Log}(PRODUCT_i) + \frac{\sigma^2}{2}\} = PRODUCT_i * \exp(\frac{\sigma^2}{2}) \quad (8)$$

A simplification of expression (8) gives us the following mean production for wheat in Daniels County:

$$PRODUCT_i = ACRE_i^{\beta 1} * \exp\{\beta_0 + \beta_2 WE_i + \beta_3 NS_i + \frac{\sigma^2}{2}\} \quad (9)$$

Formula (9) can be adapted for each of the four regions within the Daniels County and summarize below:

$$\textbf{NW:}\ PRODUCT_i = ACRE_i^{\beta 1} * \exp\{\beta_0 + \beta_3 + \frac{\sigma^2}{2}\} \quad (10)$$

$$\textbf{NE:}\ PRODUCT_i = ACRE_i^{\beta 1} * \exp\{\beta_0 + \beta_2 + \beta_3 + \frac{\sigma^2}{2}\} \quad (11)$$

$$\textbf{SW:}\ PRODUCT_i = ACRE_i^{\beta 1} * \exp\{\beta_0 + \frac{\sigma^2}{2}\} \quad (12)$$

$$\textbf{SE:}\ PRODUCT_i = ACRE_i^{\beta 1} * \exp\{\beta_0 + \beta_2 + \frac{\sigma^2}{2}\} \quad (13)$$

[4] Exponential function

5.1 RESULTS INTERPRETATION

Using expression (7), it is possible to determine 95% confidence interval for the Log of wheat production in Daniels County. The boundaries of this interval can in turn be used to find the corresponding wheat production at level .05 with formula (8). Any level of production below the lower bound of a given confidence interval will be considered suspicious and placed under investigation.

For example, the farmer on lease 157 has been assigned the land SE4 of 83.6 acres. Given that the land is located in the Northwest, we obtain:

$$6.91 > \text{Log}\,(\text{PRODUCT}_{17}) > 8.57 \text{ or}$$

$$1{,}100.5 > \text{PRODUCT}_{17} > 5{,}764.5$$

Keeping in mind that we are interested in detecting underreported production, we want to compare the lower bound of our confidence interval to the actual production level. For this case, we find that the number 1,100.5 is smaller than the actual of 1,922.8 bushels. Therefore, the farmer on lease 157 will not be subject of any investigation. A list of all suspicious level of production is provided in Table 6.

To fully interpret results from our quantitative analysis, it is important to elaborate a certain number of rules that will serve as qualitative element for better business decision. In fact, a low production can be justified by factors. Those factors can be of technical, natural nature or social such as bad agricultural practices, non-favorable climatic conditions or physical incompetence due to sickness or social troubles.

Table 6: Results Summary Derived from our Statistical Model*

Obsv	Twp	Rng	Production	Acres	Loc	NS	WE	Log (Production)	Lower bound of Log (production) CI	Lower bound of production CI	Underreported Production
8	34	45	631.00	77.00	SE	0	1	7.47	6.64	834.62	TRUE
41	34	47	1,540.00	177.80	SE	0	1	8.30	7.47	1,912.76	TRUE
45	35	44	168.00	36.00	SW	0	0	6.43	5.60	294.87	TRUE
60	36	45	279.00	38.30	NE	1	1	6.97	6.14	507.71	TRUE
128	36	44	815.00	100.40	NW	1	0	7.64	6.81	990.24	TRUE
183	37	47	363.50	31.90	NE	1	1	6.79	5.96	423.57	TRUE
209	35	44	74.60	92.70	SW	0	0	7.36	6.53	752.86	TRUE
223	36	44	160.00	35.48	NW	1	0	6.61	5.78	353.23	TRUE
224	36	44	462.92	112.32	NW	1	0	7.75	6.92	1,106.69	TRUE
272	34	44	272.00	57.70	SW	0	0	6.89	6.06	470.61	TRUE
277	37	47	262.60	40.00	NE	1	1	7.01	6.18	530.04	TRUE
278	37	47	994.30	86.30	NE	1	1	7.77	6.95	1,135.67	TRUE
279	37	47	571.00	50.00	NE	1	1	7.23	6.40	661.22	TRUE
280	37	47	484.90	42.50	NE	1	1	7.07	6.24	562.86	TRUE

* Obsv = Observation | Loc = Location | CI = Confidence Interval

5.2 LIMITATIONS & FUTURE RESEARCH

Model building requires a proper balance of theory, availability of the appropriate data, a good understanding of the statistical proprieties of the various models, and the elusive quality that is called practical judgment[5]. Since we have little information regarding the underlying theory and the data set at hand, we would like in this section to highlight several limitations of our study and provide incentive for further research.

The first restriction of our model stems from its limitation to estimate farmers' production level over time. Time series data would have been useful in this particular case. In addition, very little information has been provided the data set which makes it difficult to assess the validity and the reliability of the results.

The second limitation of our model comes from the fact that climatic conditions at the time the data were collected are not known and have not been taken into account in our research.

For future research, it would be very interesting to generalize our model to the state of Montana to incorporate the intrinsic characteristics of each counties and different type of commodities as well. Moreover, it would be ideal to create an Access database in order to track farmers' level of production over time and manage related information.

[5] Damodar Gujarati (1992), *Essential of Econometrics, second edition* (The McGraw-Hill Companies), pp. 264

BIBLIOGRAPHY

- Cliff T. Ragsdale (2004), *Spreadsheet Modeling & Decision Analysis: A Practical Introduction to Management Science, fourth edition* (Ohio: South Western).
- Damodar Gujarati (1992), *Essential of Econometrics, second edition* (The McGraw-Hill Companies).
- David A. Aaker, V. Kumar, George S. Day (2004), *Marketing Research, eighth edition* (John Wiley & Sons).
- James J. Higgins (2004), *Introduction to Modern Nonparametric Statistics* (Brooks/Cole, a division of Thomson Learning, Inc.)
- James Lattin, J. Douglas Carroll, Paul E. Green (2003), *Analyzing Multivariate Data* (Brooks/Cole, a division of Thomson Learning, Inc.)
- Lawrence C. Hamilton (1992), *Regression with Graphics: A Second Course in Applied Statistics* (California: Duxbury Press)
- Michael H. Kutner, Christopher J. Nachtsheim, John Neter (2004), *Applied Linear Regression Models*, Fourth Edition (McGraw-Hill Irwin)
- Morris H. DeGroot, Mark J. Schervish (2002), *Probability and Statistics, third edition* (Addison-Wesley)
- Peter Kennedy (2003), *A Guide to Econometrics, fifth edition* (The MIT Press)
- Richard D. De Veaux, Paul F. Velleman (2004), *Intro Stats* (Pearson Education, Inc.)
- Royce A. Singleton, Bruce C. Straits, *Approaches to Social Research, third edition* (New York: Oxford).

Figure 16: Montana State Map

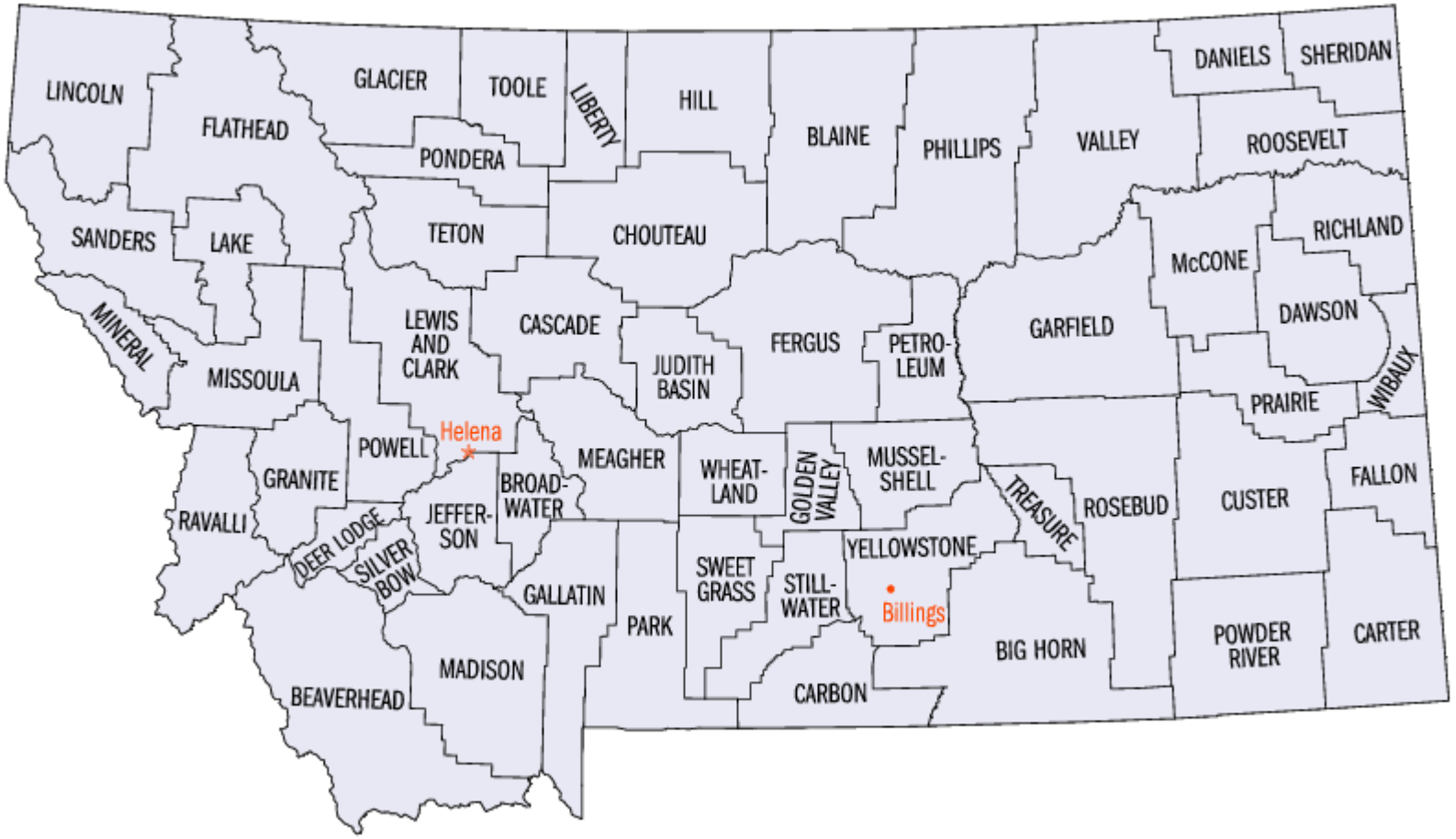
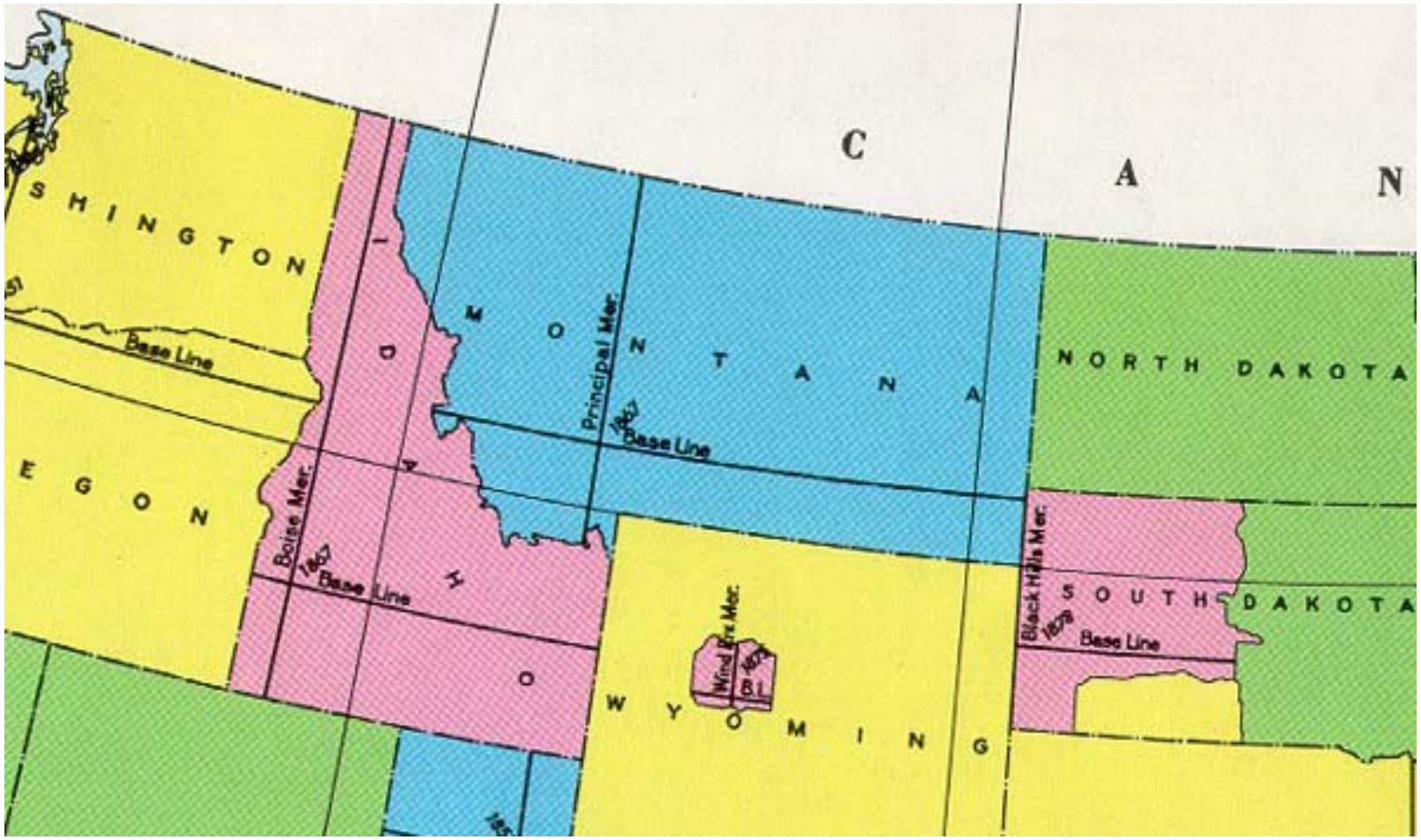


Figure 17: Meridians and baselines locations in the United States[6]



[6] Source: http://upload.wikimedia.org/wikipedia/en/f/ff/Meridians-baselines.jpg