

## Topic 6

# Developing Predictive Models

### 6.1 Predicting Yield Maps

Talk about the future of Precision Ag—how about maps of things yet to come? Sounds a bit far fetched but *Spatial Data Mining* is taking us in that direction. For years non-spatial statistics has been predicting things by analyzing a sample set of data for a numerical relationship (equation) then applying the relationship to another set of data. The drawbacks are that the non-approach doesn't account for geographic relationships and the result is just a table of numbers addressing an entire field.

Extending predictive analysis to mapped data seems logical. After all, maps are just organized sets of numbers.

And the GIS toolbox enables us to link the numerical and geographic distributions of the data. The past several topics have discussed how the computer can “see” spatial data relationships including “descriptive techniques” for assessing *map similarity*, and *data zones*. The next logical step is to apply “predictive techniques” to generate extrapolative maps that forecast future conditions.

A geo-business application to extend a test market project for a phone company might serve to introduce the basic approach. The example doesn't relate crop yield and farm inputs, but it does relate the sales (think crop yield) to demographics (think farm inputs).

The new product that was test marketed in 1991 enabled two phone numbers with distinctly different rings to be assigned to a single home phone—one for the kids and one for the parents. When customers purchased the new product their addresses were used to geo-code the sales. Like pushpins on a map, the pattern of sales throughout the city emerged with some areas doing very well (high Sales yield), while in other areas sales were few and far between (low sales yield).

The assumption was that a pattern existed between conditions throughout the city, such as income level, education, number in household, etc. (analogous to farm inputs throughout a field) that determine sales

yield. The demographic data for the city was analyzed to calculate a prediction equation between product sales and census block data.

The prediction equation derived from the test market sales in one city was applied to another city by evaluating exiting demographics to “solve the equation” for a predicted sales

map. In turn the predicted map was combined with a wire-exchange map to identify switching facilities that required upgrading before release of the product in the new city.

The ability to model a spatial relationship then apply it to another area or time period fuels the multi-billion dollar industry in retail sales forecasting.

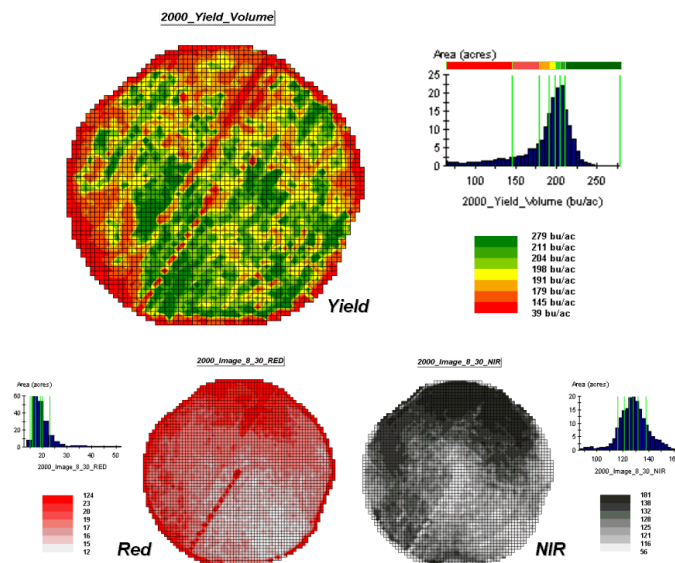


Figure 6-1. The corn yield map (top) identifies the pattern to predict; the red and near-infrared maps (bottom) are used to build the spatial relationship.

Discovery of spatial and other relationships in product sales directly translates into key business decisions. Spatial data mining in agriculture holds a similar opportunity.

To illustrate spatial data mining for precision agriculture data, the approach can be applied to the case study cornfield data. The top portion of figure 6-1 shows the yield pattern for the field varying from a low of 39 bushels per acre (red) to a high of 279 (green). Corn yield, like “sales yield,” is termed the *dependent map variable* and identifies the phenomena one wants to predict.

The *independent map variables* depicted in the bottom portion of the figure are used to uncover the spatial relationship—

**prediction equation.**

In this instance, digital aerial imagery will be used to

explain the corn yield patterns, instead of demographics to explain sales. The map on the left indicates the relative reflectance of red light off the plant canopy while the map on the right shows the near-infrared response (a form of light just beyond what we can see).

While it is difficult for you to assess the subtle relationships between corn yield and the red and near-infrared images, the computer “sees” the relationship quantitatively. Each grid location in the analysis frame has a value for each of the map layers— 3,289 values defining each geo-registered map covering the 189-acre field.

For example, top portion of figure 6-2 identifies that the example location has a “joint” condition of red equals 14.7 counts and yield equals 218 bu/ac. The red lines in the scatter plot on the right show the precise position of the pair of map values—X= 14.7 and Y= 218. Similarly, the near-infrared and yield values for the same

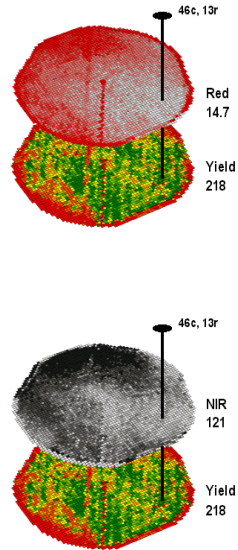
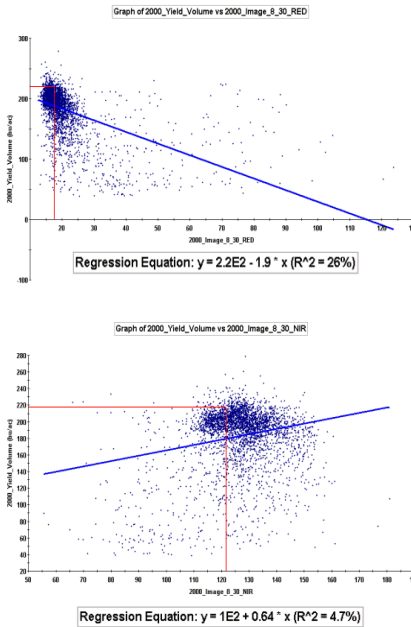


Figure 6-2. The joint conditions for the spectral response and corn yield maps are summarized in the scatter plots.



location are shown in the bottom portion of the figure.

In fact the set of “blue dots” in both of the scatter plots represents data pairs for each grid location. The blue lines in the plots represent the prediction equations derived through regression analysis. While

the mathematics is a bit complex, the effect is to identify a line that “best fits the data”— just as many data points above as below the line.

In a sense, the line sort of identifies the average yield for each step along the X-axis (red and near-infrared responses respectively). Come to think of it, wouldn’t that make a reasonable guess of the yield for each level of spectral response? That’s how a regression prediction is

used—a value for red (or near-infrared) in another field is entered and the equation for the line is used to predict corn yield. Repeat for all of the locations in the field and you have a prediction map of yield from an aerial image... but alas, if it were only that simple and exacting.

**6.2 Assessing Prediction Model Results**

A major problem is that the “r-squared” statistic for both of the prediction equations is fairly small ( $R^2= 26\%$  and  $4.7\%$ ) which suggests that the prediction lines do not fit the data very well. One way to improve the predictive model might be to combine the information in both of the images. The “Normalized Density Vegetation Index (NDVI)” does just that by calculating a new value that indicates relative plant vigor—  $NDVI= ((NIR - Red) / (NIR + Red))$ .

Figure 6-3 shows the process for calculating NDVI for the sample grid location—  $((121-14.7) / (121 + 14.7))= 106.3 / 135.7= .783$ . The scatter plot on the

right shows the yield versus NDVI plot and regression line for all of the field locations. Note that the R<sup>2</sup> value is a higher at 30% indicating that the combined index is a better predictor of yield than red or NIR alone.

The bottom portion of the figure evaluates the NDVI prediction equation's performance over the field. The two smaller maps show the actual yield (left) and predicted yield (right). As you would expect the prediction map doesn't contain the extreme high and low values actually measured.

The larger map on the right calculates the error of the estimates by simply subtracting the actual measurement from the predicted value at each map location. The error map suggests that overall the yield "guesses" aren't too bad— average error is a 2.62 bu/ac over-guess; 67% of the field is within +/- 20 bu/ac. Also note that most of the over estimating occurs along the edge of the field while most of the under estimating is scattered along curious NE-SW bands.

While evaluating a prediction equation on the data that generated it isn't validation, the procedure provides at least some empirical verification of the technique. It suggests a glimmer of hope that with some refinement the prediction model might be useful in predicting yield before harvest. In the next section we'll investigate some of these

refinement techniques and see what information can be gleaned by analyzing the error surface.

### 6.3 Stratifying Maps for Better Predictions

The last section described procedures for predictive analysis of mapped data. While the underlying

theory, concerns and considerations can easily consume a graduate class for a semester, the procedure is quite simple. The grid-based processing preconditions the maps so each location (grid cell) contains the appropriate data. The "shishkebab" of numbers for each location within a stack of maps are analyzed for a prediction equation that summarizes the relationships.

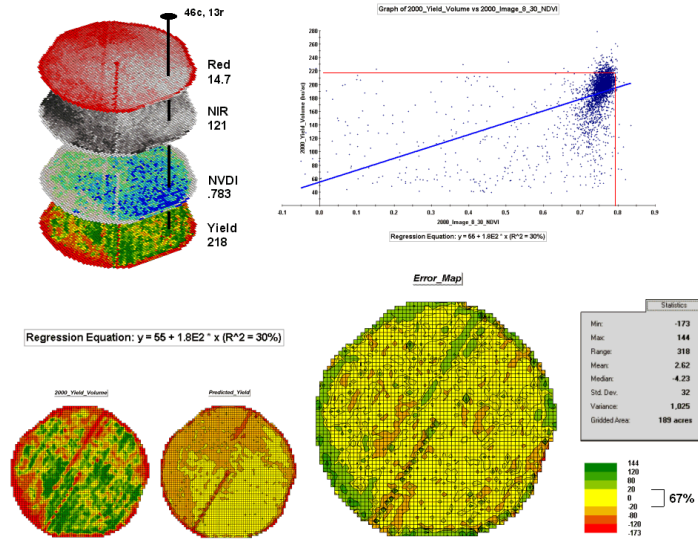


Figure 6-3. The red and NIR maps are combined for NDVI value that is a better predictor of yield.

In the example discussed in the last section, regression analysis was used to relate a map of NDVI ("normalized density vegetation index" derived from remote sensing imagery) to a map of corn yield for a farmer's field. Then the equation was used to derive a map of predicted yield based on the NDVI values and the results evaluated for how well the prediction equation performed.

The left side of figure 6-4 shows the evaluation procedure. Subtracting the actual yield values from the predicted ones for each map location derives an **Error Map**. The

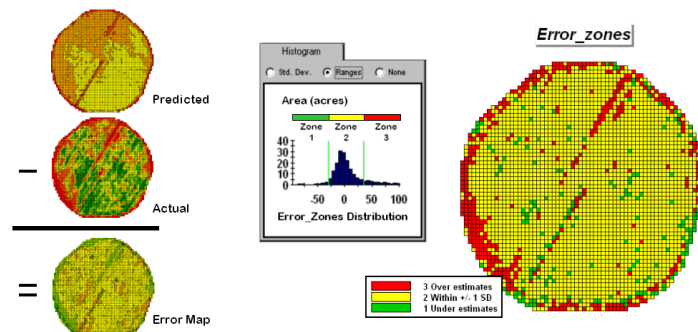


Figure 6-4. Using prediction errors to stratify.

previous discussions noted that the yield "guesses" weren't too bad—average error of 2.62 bu/ac with 67% of the estimates within 20 bu/ac of the actual

yield. However, some locations were as far off as 144 bu/ac (over-guess) and -173 bu/ac (under-guess).

One way to improve the predictions is to **stratify** the data set by breaking it into groups of similar characteristics. The idea is that set of prediction equations tailored to each stratum will result in better predictions than a single equation for an entire area. The technique is commonly used in non-spatial statistics where a data set might be grouped by age, income, and/or education prior to analysis. In spatial statistics additional factors for stratifying, such as neighboring conditions and/or proximity, can be used.

While there are several alternatives for stratifying, subdividing the error map will serve to illustrate the conceptual approach. The histogram in the center of figure 6-5 shows the distribution of values on the *Error Map*. The vertical bars identify the breakpoints at plus/minus one standard deviation and divide the map values into three strata—zone 1 of unusually high under-guesses (red), zone 2 of typical error (yellow) and zone 3 of unusually high over-guesses (green). The map on the right of the figure locates the three strata throughout the field.

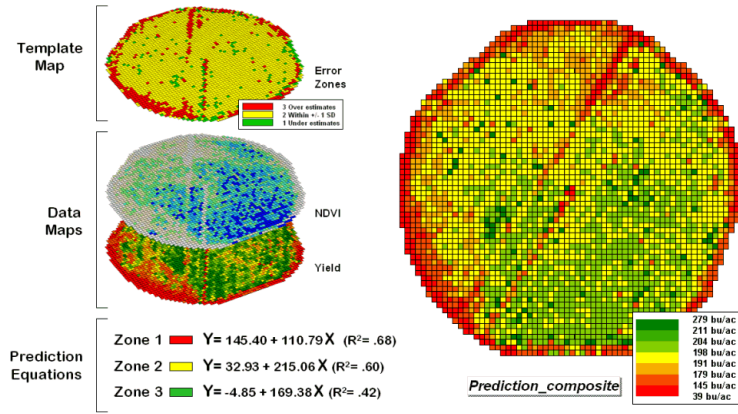


Figure 6-5. After stratification, prediction equations can be derived for each element.

The rationale behind the stratification is that the whole-field prediction equation works fairly well for zone 2 but not so well for zones 1 and 3. The

assumption is that conditions within zone 1 makes the equation under estimate while conditions within zone 3 cause it to over estimate. If the assumption holds one would expect a tailored equation for each zone would be better at predicting than an overall

equation. Figure 6-6 summarizes the results of deriving and applying a set of three prediction equations.

The left side of the figure illustrates the procedure. The *Error Zones* map is used as a template to identify the NDVI and Yield values used to calculate three separate prediction equations. For each map location, the algorithm first checks the value on the *Error*

*Zones* map then sends the data to the appropriate group for analysis. Once the data has been grouped, a regression equation is generated for each zone. The “r-squared” statistic for all three equations (.68, .60, and .42 respectively) suggests that the equations fit the data fairly well and ought to be good predictors. The right side of

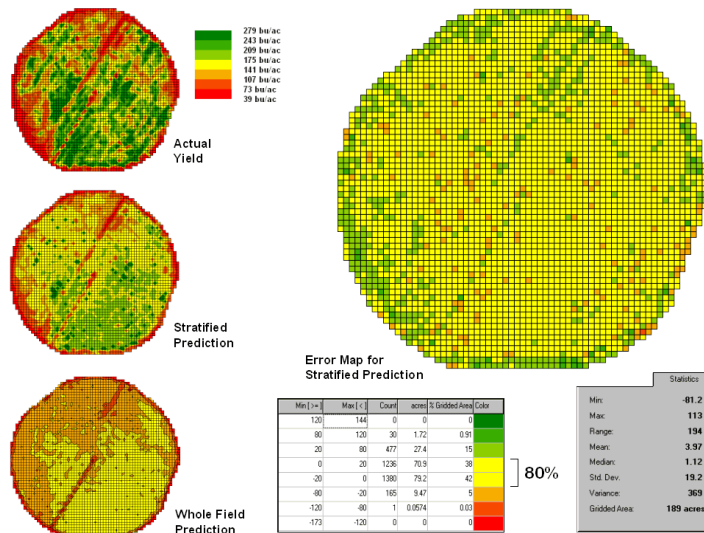


Figure 6-6. Stratified and whole-field predictions can be compared using statistical techniques.

figure 2 shows the composite prediction map generated by applying the equations to the NDVI data respecting the zones identified on the template map.



The left side of figure 6-6 provides a visual comparison between the actual yield and predicted maps. The “stratified prediction” shows detailed estimates that more closely align with the actual yield pattern than the “whole-field” derived prediction map. The error map for the stratified prediction shows that eighty percent of the estimates are within +/- 20 bushels per acre. The average error is only 4bu/ac and a maximum of under and over-estimate of -81.2 and 113, respectively. All in all, not bad guessing of yield based on a remote sensing shot of the field nearly a month before the field was harvested.


A couple of things should be noted from this example of spatial data mining. First, that there is a myriad of other ways to stratify mapped data—1) *Geographic Zones*, such as proximity to the field edge; 2) *Dependent Map Zones*, such as areas of low, medium and high yield; 3) *Data Zones*, such as areas of similar soil nutrient levels; and 4) *Correlated Map Zones*, such as micro terrain features identifying small ridges and depressions. The process of identifying useful and consistent stratification schemes is an emerging research frontier in the spatial sciences.

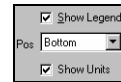
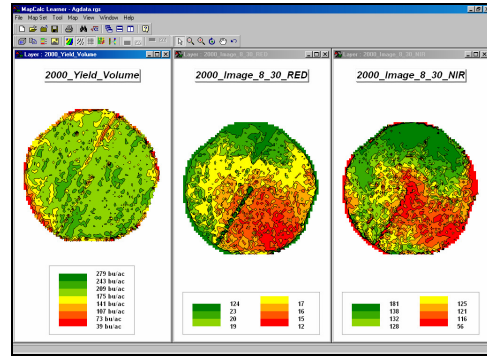
Second, the error map is important in evaluating and refining the prediction equations. This point is particularly important if the equations are to be extended in space and time. The technique of using the same data set to develop and evaluate the prediction equations isn’t always adequate. The results need to be tried at other locations and dates to verify performance. While spatial data mining methodology might be at hand, good science is imperative.

The bottom line is that maps are increasingly seen as organized sets of data that can be quantitatively analyzed for spatial relationships— we have only scratched the surface. The applications of spatial statistics and data mining in production agriculture are in their infancy. As the agricultural sciences embrace spatial technology, research will tailor the procedures for the unique data and situations on individual farms.

Access MapCalc using the *Agdata.rgs* data set by selecting **Start → Programs → MapCalc Learner → MapCalc Learner → Open existing map set → PA\_AgData.rgs**. The following set of exercises utilizes this database.

**6.4.1 Predictive Modeling**

 Using the *View* and *Tile* buttons create side-by-side displays of the **2000\_Yield\_Volume**, **2000\_Image\_8\_30\_RED** and **2000\_Image\_8\_30\_NIR** maps.



To place the legend at the bottom of a map display, right-click on the map, select **Properties, Legend** tab and set the **Pos** (Position) to **Bottom**. Press the **Apply to Open Maps** button to place the legend at the bottom in all of the open display windows. Also note that when you press the *Tile* button the active map window (blue strip at the top) is set to the left-most position.

Can you detect any pattern in the Red and NIR images that relates to the pattern in the Yield image? How about the patterns around the access road and edges?

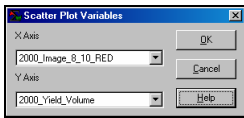
Double-click on the Yield map to pop-up the data inspection tool. Record the Yield, Red and NIR values for map locations (42,44) and (45,54).

[42,44]= 193 Yield, 23.3 Red, 128 NIR				[45,54]= 232 Yield, 22.1 Red, 150 NIR			
Feature Name	Value	Units		Feature Name	Value	Units	
2000_Yield_Volume	193	bu/ac		2000_Yield_Volume	232	bu/ac	
2000_Yield_Mass	10,780	lb/ac		2000_Yield_Mass	12,590	lb/ac	
2000_Yield_Moisture	15	%		2000_Yield_Moisture	15.4	%	
2000_Image_8_30_NIR	128			2000_Image_8_30_NIR	150		
2000_Image_8_30_RED	23.3			2000_Image_8_30_RED	22.1		

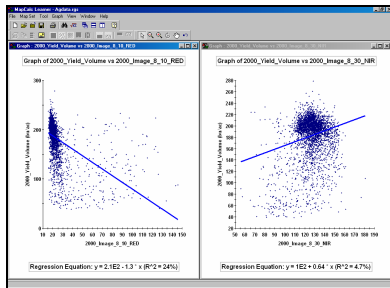
In analyzing data patterns, the computer mathematically investigates these sequences of values. Note that yield increases from 193 to 232 while red decreases from 23.3 to 22.1 and NIR

**6.4 Exercises**

increases from 128 to 150. A way to visualize the relationships for all of the field locations is to generate a scatter plot of the paired data.



Select **Map Set** → **New Graph** → **Scatter Plot** from the main menu and specify the appropriate maps for the X-axis (independent variable; Red and NIR) and Y-axis (dependent variable; Yield). Create side-by-side displays of the plots as shown below.



Each point in a scatter plot shows the paired values for a grid location in the field. Notice that there is a dense cluster of points around paired values that typically occur—*joint response* for most of the field.

The pattern of points outside of the cluster shows how one variable changes as the other changes. The line in the plot summarizes the relationship by effectively balancing the points above and below the *regression line*. Note that the relationship is negative for the Yield vs. Red with increasing red reflectance values associated with decreasing yield values. The opposite is true for Yield vs. NIR with increasing NIR values related to increasing yield values.

The regression equations at the bottom of the plots mathematically depict the relationships:

$$\text{Yield} = 210 \text{bu/ac} - 1.3 * \text{Red\_value}$$

$$\text{Yield} = 100 \text{bu/ac} + 0.64 * \text{NIR\_value}$$

Use the equations to predict estimated yield values for the two map locations under study:

Yield [42,44] measured= 193bu/ac  
 Yield Red=  $210 - 1.3 * 23.3 = 179.7 \text{bu/ac}$  (13.3 under)  
 Yield NIR=  $100 + 0.64 * 128 = 181.9 \text{bu/ac}$  (11.1 under)

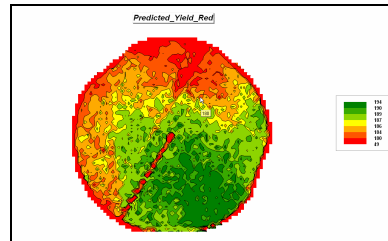
Yield [45,54] measured= 232bu/ac  
 Yield Red=  $210 - 1.3 * 22.1 = 181.3 \text{bu/ac}$  (50.7 under)  
 Yield NIR=  $100 + 0.64 * 150 = 196.0 \text{bu/ac}$  (36 under)

A map-mathematical solution of the equations for the entire field can be calculated by entering...

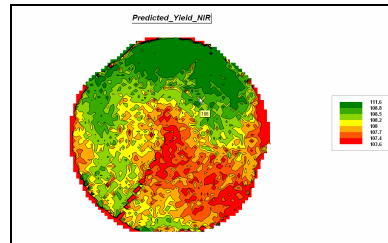
**Map Analysis** → **Overlay** → **Calculate**  
 Calculate  $210 - (1.3 * 2000\_Image\_8\_30\_RED)$  FOR *Predicted\_Yield\_Red*

**Map Analysis** → **Overlay** → **Calculate**  
 Calculate  $100 + (0.64 * 2000\_Image\_8\_30\_NIR)$  FOR *Predicted\_Yield\_NIR*

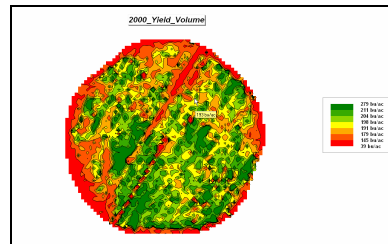
...and display the resultant maps using the **Equal Count** mode for calculating the ranges (*Shading Manager*).



*Predicted\_Red*



*Predicted\_NIR*



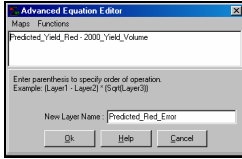
*Actual*

Visually comparing the spatial patterns on the two prediction maps with the actual yield map shows that the Red predictions have a similar relative pattern—generally higher yields in the southeastern portion of the field. The NIR predictions displays an inconsistent pattern—higher yields in the northwest portion.

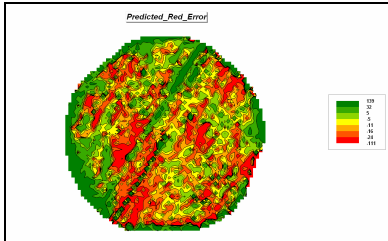
The visual evaluation is consistent with the R-squared statistic for the equations— $R^2 = 26\%$  for Red and  $R^2 = 4.7\%$  for NIR. The very small r-squared value for the NIR equation indicates that it is a very poor predictor of yield.

### 6.4.2 Calculating Error

Select **Map Analysis** → **Overlay** → **Calculate** to generate error maps of the Red and NIR predictions. Display the results using **Equal Count** mode for the ranges.

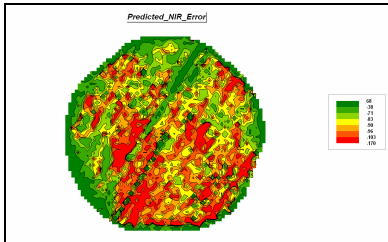


Calculate Predicted\_Yield\_Red - 2000\_Yield\_Volume FOR Predicted\_Red\_Error

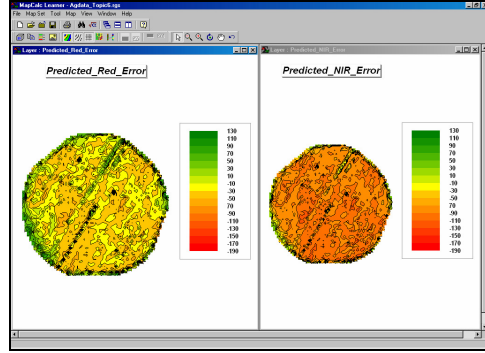


Repeat the error calculation using the yield predictions based on NIR data.

**Map Analysis** → **Overlay** → **Calculate**  
 Calculate Predicted\_Yield\_NIR - 2000\_Yield\_Volume  
 FOR Predicted\_NIR\_Error

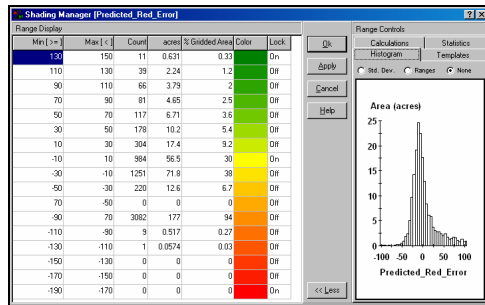


Right-click on the Red error map, select the **Shading Manager** then the **Statistics** tab. Note the min and maximum error values— -139 (under) to 111 (over). repeat for the NIR min/max error values— -68 to 170. Based on this information, construct a common display setting using 17 Equal Count ranges with an interval of 20 ranging from -190 through 150. Display the two error maps side-by-side.

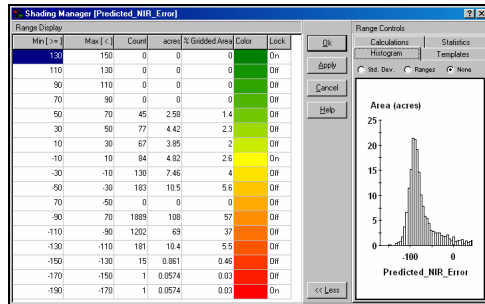


Note the reddish-tone of the NIR error map indicating a large proportion of under estimates. The green-tones of the Red error map suggests more over estimates. The larger proportion of yellow on the Red error map indicates more nearly correct estimates.

Tabular statistics is included in their Shading Manager summary tables.



Predicted\_Red\_Error



Predicted\_NIR\_Error

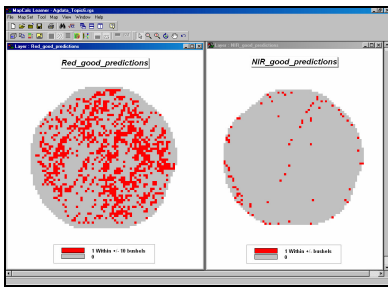
Notice that the distribution of error for the Red predictions is centered on 0. The NIR errors are centered on -75. In addition, 30% of Red predictions are within plus or minus 10 bushels, whereas NIR predictions have only 2.6%.

-10	10	984	56.5	30	On	30.0% within +/- 10bu
-10	10	84	4.82	2.6	On	2.6% within +/- 10bu

Isolate the areas of good estimates (+/-10bu) on both the Red and NIR prediction maps.

**Map Analysis**→ **Reclassify**→ **ReNUMBER**  
*RENUMBER Predicted\_Red\_Error ASSIGNING 0 TO -190 THRU -10 ASSIGNING 1 TO -10 THRU 10 ASSIGNING 0 TO 10 THRU 150 FOR Red\_good\_predictions*

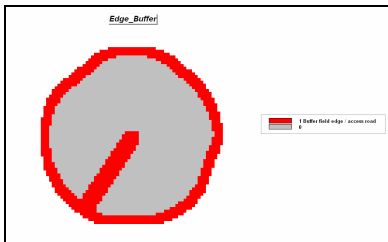
**Map Analysis**→ **Reclassify**→ **ReNUMBER**  
*RENUMBER Predicted\_NIR\_Error ASSIGNING 0 TO -190 THRU -10 ASSIGNING 1 TO -10 THRU 10 ASSIGNING 0 TO 10 THRU 150 FOR NIR\_good\_predictions*



Overall, the Red predictions appear considerably better than the NIR predictions. However, other prediction models, such as stratified for edges, might be better.

### 6.4.3 Deriving a Stratified Model

Generate a display of the **Edge\_buffer** map.

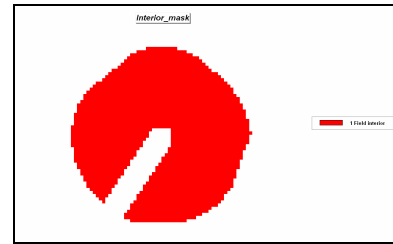


This map identifies areas along the field edge and access road (see Topic 7 exercises for how it was created). The yield measurements within the edges are thought to be highly variable due to growing conditions and measurement error. The interior portion, however, might prove to be a better predictor.

The following steps “masks” the Yield and Red values for field interior locations.

**Map Analysis**→ **Reclassify**→ **ReNUMBER**

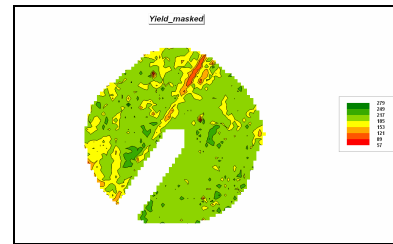
*RENUMBER Edge\_Buffer ASSIGNING PMAP\_NULL TO 1 ASSIGNING 1 TO 0 FOR Interior\_mask*



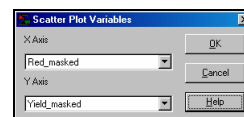
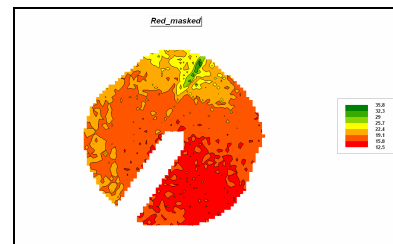
Note: The PMAP\_NULL is a special value that indicates areas not be considered in map analysis processing. Locations assigned PMAP\_NULL are ignored in calculations and displays.

Use the “mask” to identify just the interior Yield and Red data.

**Map Analysis**→ **Overlay**→ **Calculate**  
*CALCULATE 2000\_Yield\_Volume \* Interior\_mask FOR Yield\_masked*

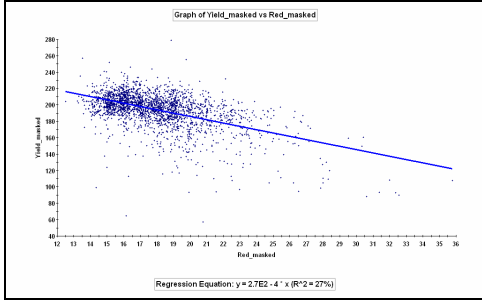


**Map Analysis**→ **Overlay**→ **Calculate**  
*CALCULATE 2000\_Image\_8\_30\_RED \* Interior\_mask FOR Red\_masked*



Generate scatter plots and regression equations for the masked Yield versus Red data by selecting **Map Set**→ **New Graph**→ **Scatter Plot**.



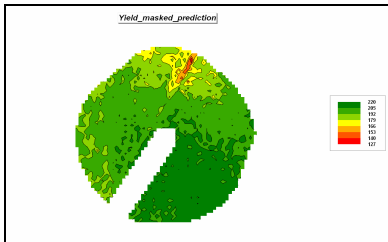


Evaluate the regression equation...

$$\text{Yield\_masked} = 270.0 - 4.0 * \text{Red\_masked}$$

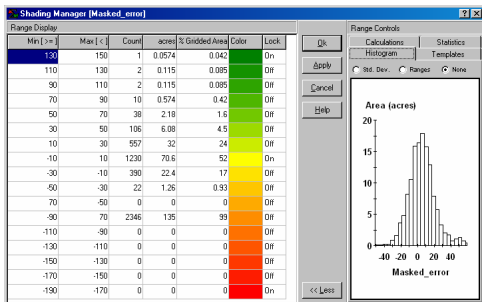
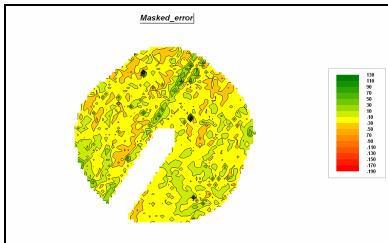
... using the Red masked data.

**Map Analysis** → **Overlay** → **Calculate**  
 CALCULATE 270.0 + 4.0 \* Red\_masked FOR Yield\_masked\_predicted



Now generate an error map and display using the same settings as the previous two error maps...

**Map Analysis** → **Overlay** → **Calculate**  
 CALCULATE Yield\_masked\_prediction - 2000\_Yield\_Volume FOR Yield\_masked FOR Masked\_error

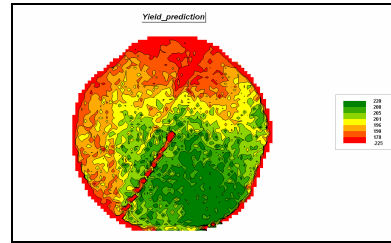


-10, 10, 1230, 70.6, 52, On 52.0% within +/- 10bu

It appears that the masked prediction model is fairly good at predicting yield for the interior portion of the field.

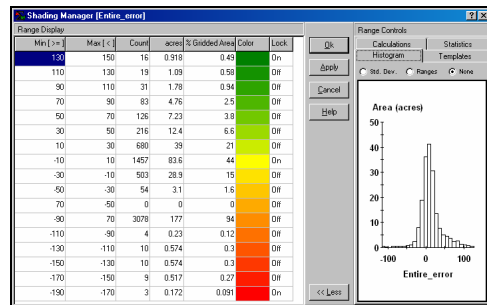
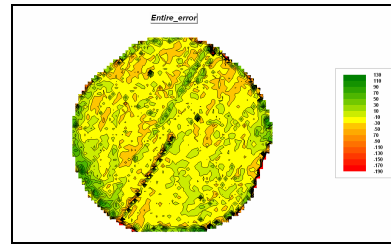
But is it as good as or better than the unmasked prediction model at predicting yield for the whole field? Evaluate the masked regression equation for the whole field...

**Map Analysis** → **Overlay** → **Calculate**  
 CALCULATE 270.0 + 4.0 \* 2000\_Image\_8\_30\_RED FOR Yield\_predicted



Calculate the error map...

**Map Analysis** → **Overlay** → **Calculate**  
 CALCULATE Yield\_prediction - 2000\_Yield\_Volume FOR Entire\_error



-10, 10, 1457, 83.6, 44, On 44.0% within +/- 10bu

It appears that the masked regression model is a better predictor (44% versus 30%) than the equation developed for the entire field.

You can exit the program by selecting **File** → **Exit** or by clicking on the "X" in the upper-right corner of the MapCalc program window. If you want to save

your work, specify a new file name, such as *AgData\_Topic6\_exercises.rgs*. Each exercise set assumes you will start with the basic *AgData.rgs* data set and this database will become cluttered with exercise maps if you save your results to it each time.