

## Topic 8

# Predictive Modeling

### 8.1 Examples of Predictive Modeling

Talk about the future of geo-business—how about maps of things yet to come? Sounds a bit far fetched but spatial data mining and predictive modeling 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-spatial approach doesn't account for geographic relationships and the result is just a table of numbers addressing an entire project area.

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*, *data zones*, and *map clustering*. The next logical step is to apply “predictive techniques” to generate extrapolative maps that forecast future conditions.

This approach is the basis for a new discipline in agriculture—Precision Agriculture. In this application a farmer collects yield data on-the-fly as he harvests a field. Every second the combine checks GPS to see where it is within a couple of feet and then records the crop yield for that location. The result is a yield measurement about every six feet throughout the field. These data are combined with spatially interpolated soil nutrient samples to uncover relationships between yield and nutrient levels.

In turn the map-*ematical* relationships are used to determine the best mix of nutrients needed at each ten-meter grid location in the field. The final step employs variable rate technology (VRT) that applies the right blend of nutrients, at the right place and time as the fertilizer rig moves through the field. While this “star wars” approach to farming might seem far fetched, it is currently being on millions of acres of farmland in the U.S. The eye-opening aspect is that the map analysis application didn't even exist ten years ago.

A geo-business application to extend a test market project for a phone company might serve to introduce the basic approach within a business context. 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 areas), while in other areas sales were few and far between (low sales areas).

The assumption was that a pattern existed between conditions throughout the city, such as income level, education, number in household, etc. (analogous to soil nutrients throughout a field) that help explain sales “yield.” The demographic data for the city was analyzed to calculate a prediction equation between product sales and census 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 emerging industry in retail sales forecasting. Discovery of spatial and other relationships in product sales directly translates into key business decisions.

### 8.2 Spatial Regression

Suppose a bank has a database of home equity loan accounts they have issued in the past few months. A standard desktop mapping system can be used to geo-code these data by matching street addresses with a street map. This process uses information about the address range for each line segment in the map to estimate each loan account's geographic location (latitude, longitude).

In turn, this information is used to “burn” the account locations into an analysis grid as discussed in the previous sections of this book. The small inset in the lower left corner of figure 8-1 shows such a map of loan accounts. As described in Topic 6.1 a roving window is used to derive a Loan Density surface by computing the number of accounts within a specified distance of each map location. In the example about a 3/4-mile radius is used (within 15 cells \* 250 feet/cell= 3750 feet). Note the spatial distribution of the account density, or concentration—a large pocket in the southeast and a smaller one in the southwest.

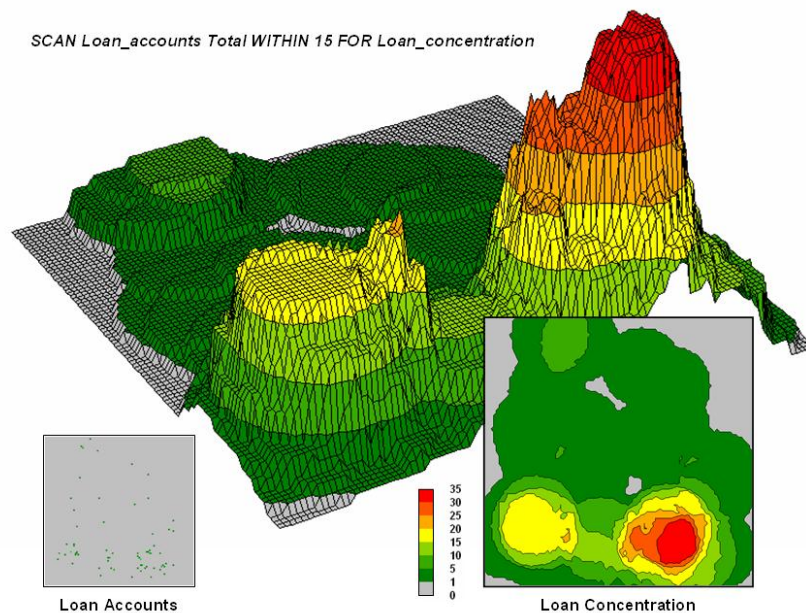


Figure 8-1. A loan concentration surface is created by summing the number of accounts for each map location within a specified distance.

Do you think there is a relationship between loan Concentration and housing Density? What about Housing Value or Age? If the relationships are adequate, relatively easily obtained assessor's data on housing Density, Value and Age might be used to predict loan Concentration throughout the city. Armed with this information one could identify areas where high rates of loan accounts ought to occur within the project area, or another portion of the city (potential branch bank) or even another city.

The focus of the following example will be on the map analysis procedures used to address these questions. Keep in mind that the data is hypothetical and “real-world” data might not behave in the same way. While the procedure worked adequately for predicting sales of a double-ring phone product from demographic data and corn yield from soil nutrient levels as discussed in the previous section, success depends on the actual relationships (geographic and numeric patterns) in an application under consideration. If the data exhibits weak relationships in either geographic space or data space then map prediction is a bust. If strong relationships exist then a useful prediction equation can be developed.

A frequently used method for establishing a quantitative relationship among variables involves regression. It is beyond the scope of this book to discuss the underlying theory of the regression procedure. However, figure 8-2 schematically depicts the basic concept. A line is “fitted” in data space that balances the data so the differences from the points to the line (termed the residual) for all the points is minimized and the sum of the differences is zero.

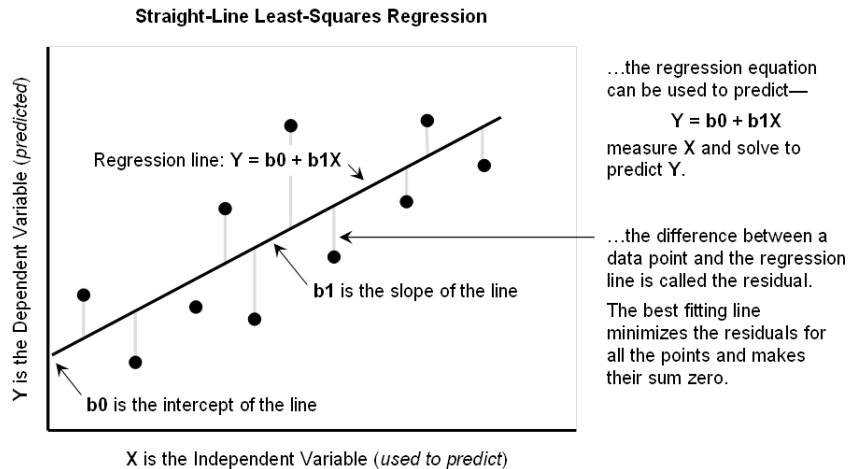


Figure 8-2. Regression derives a prediction equation based on the data pattern between variables.

The equation of the regression line is used to predict the “dependent” variable (**Y** axis) using one or more “independent” variables (**X** axis). The generalized equation,  $Y = b_0 + b_1X$ , is solved for the specific parameters of the where the line crosses the Y axis (**b0** called the intercept) and its inclination (**b1** called the slope). If more than one independent variable is used the equation is expanded to  $Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 \dots$  for as many variables as used. In the following example, three independent variables (housing Density, Value and Age) will be used to predict a derived dependent variable (loan Concentration).

To illustrate predictive modeling, consider an application to determine the relative propensity for home equity loans throughout the Smallville project area. The left side of figure 8-3 shows the four maps involved in the analysis. The loan Concentration surface at top is the same one shown in figure 8-1 and serves as the dependent map variable (to be predicted). The housing Density, Value and Age surfaces serve as the independent map variables (used to predict). Each grid cell contains the data values used to form the relationships. For example, the “pin” in the figure identifies a high loan Concentration response coinciding with a low housing Density, high Value and low Age responses.

As your eye roams over the map surfaces, do you detect any consistent patterns? Do high loan Concentration values generally coincide with low housing Density? Or is the data pattern inconsistent? What about the relationships with home Value and Age? Does it appear that increases in loan Concentration consistently follow increases in home Value? What about increases in home Age—consistently lower loan Concentration?

The scatter plots in the center of the figure graphically portray the consistency of the relationships. The Y axis tracks the dependent variable (loan Concentration) in all three plots while the X axis follows the independent variables (housing Density, Value and Age, respectively). Each plotted point represents the joint condition at one of the grid locations in the project area—10,000 dots in each scatter plot. The shape and orientation of the cloud of points characterizes the nature and consistency of the relationship between the two map variables.

A plot of a perfect relationship would have all of the points forming a line. An upward directed line would indicate perfect *positive correlation* where an increase in X always results in a corresponding increase in Y. A downward directed line would indicate *perfect negative correlation* with an increase in X always resulting in a corresponding decrease in Y. The slope of the line indicates the magnitude of the relationship with a 45-degree slope indicating a strong 1-to-1 unit change. A vertical or horizontal line indicates *no correlation*—a change in one doesn’t affect the other. Similarly, a circular cloud of points indicates there isn’t any consistency in the changes.

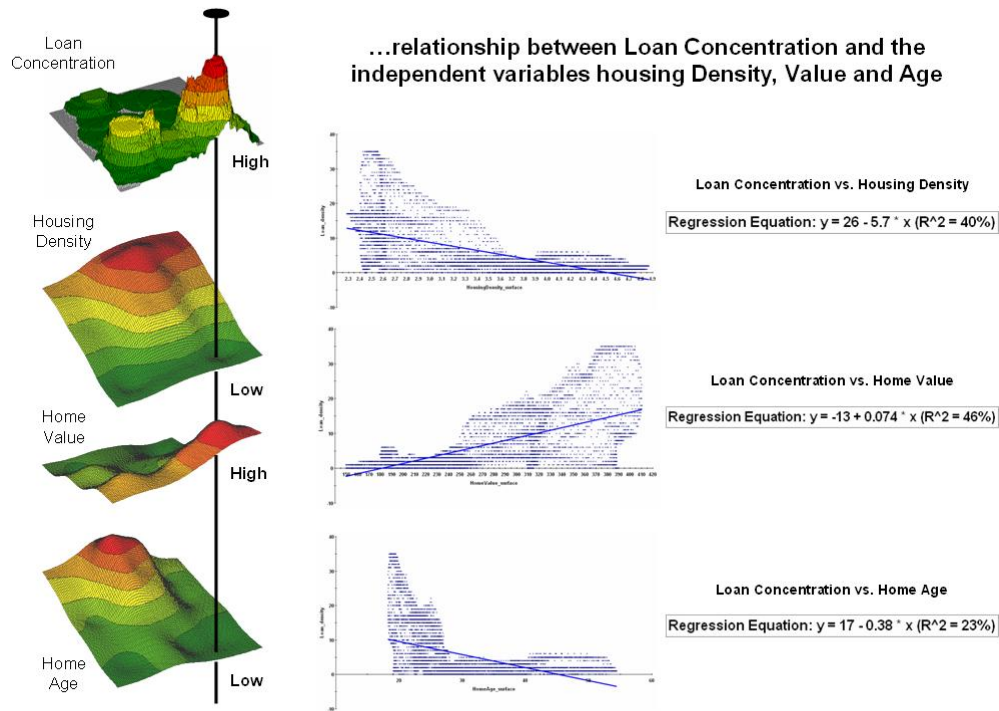


Figure 8-3. Scatter plots and regression results relate Loan Density to three independent variables (housing Density, Value and Age).

Rarely does the data plot into these extreme conditions. Most often they form dispersed clouds like the scatter plots in figure 8-3. The general trend in the data cloud indicates the amount and nature of correlation in the data set. For example, the loan Concentration vs. housing Density plot at the top shows a large dispersion at the lower housing Density ranges with a slight downward trend. The opposite occurs for the relationship with housing Value (middle plot). The housing Age relationship (bottom plot) is similar to that of housing Density but the shape is more compact.

Regression is used to quantify the trend in the data. The equations on the right side of figure 8-3 describe the “best-fitted” line through the data clouds. For example, the equation  $Y = 26.0 - 5.7X$  relates loan Concentration and housing Density. The loan Concentration can be predicted for a map location with a housing Density of 3.4 by evaluating  $Y = 26.0 - (5.7 * 3.4) = 6.62$  accounts per half-mile radius. For locations where the prediction equation drops below 0 the prediction is set to 0 (infeasible negative accounts beyond housing densities of 4.5).

The “R-squared index” with the regression equation provides a general measure of how good the predictions ought to be—40% indicates a moderately weak predictor. If the R-squared index was 100% the predicting equation would be perfect for the data set (all points directly falling on the regression line). An R-squared index of 0% indicates an equation with no predictive capabilities.

In a similar manner, the other independent variables (housing Value and Age) can be used to derive the expected loan Concentration for any location in the project area. Generally speaking it appears that home Value exhibits the best relationship with loan Concentration having an R-squared index of 46%. The 23% index for housing Age suggests it is a poor predictor of loan Concentration.

### 8.3 Assessing Prediction Model Results

A major problem is that the “R-squared” statistic for all three independent variables are fairly small ( $R^2= 40\%$ ,  $46\%$  and  $23\%$ ) which suggests that the individual prediction lines do not fit the data very well. One way to improve the predictive model might be to combine all of the information.

Figure 8-4 shows the results for multivariate regression analysis of loan concentration. The equation in the top left portion of the figure identifies the regression equation that relates housing Age, Value and Density to loan Concentration. As a means to estimate the performance of the equation a predicted map of loan Concentration can be generated by solving the equation for each map location by substituting its housing Age, Value and Density values.

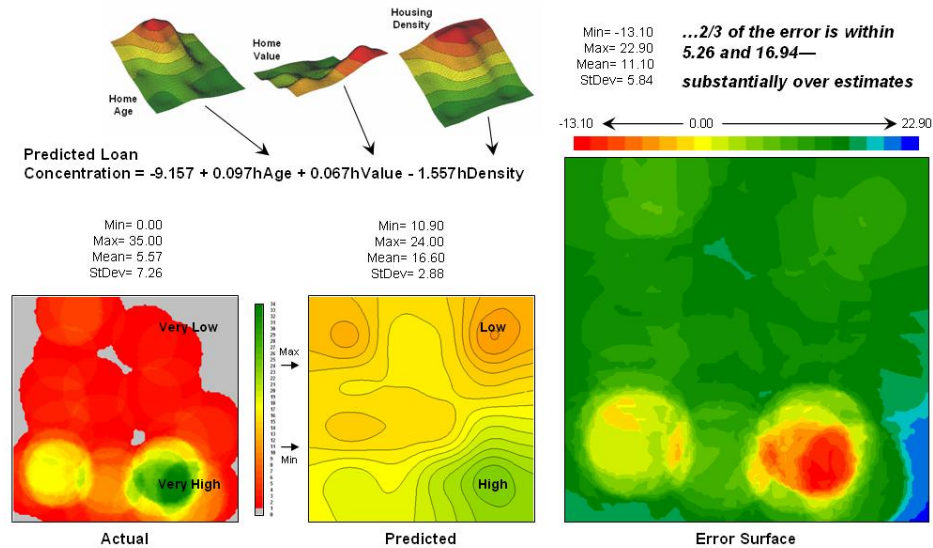


Figure 8-4. Multivariate regression prediction and error surfaces.

At first glance the *Predicted* map doesn’t appear to be very similar to the *Actual* loan concentration map. However closer inspection shows similar trends—lowest values in the northeast and highest values in the southeast with an extension toward the southwest. At least the equation didn’t predict relative responses that exhibit a “bogus” opposite trend. In fact the two maps exhibit a very similar geographic trend with most of the differences in the magnitude of the estimates. The predicted map has a data range from 10.9 to 24.0, whereas the actual range is much larger from 0 to 35.0. Like the regression lines in figure 8-3, there is a lot of dispersion in the actual data about the line but the general trend seems to be captured by the equation.

In many applications, relative spatial patterns are sufficient for decision-making. If the equation were applied to another portion of the city (or “moved” to another similar city) the areas of relatively lower or higher expected loan concentration (potential customers) could be quite valuable for marketing and sales territory decisions.

The *Error Surface* on the left side of the figure 8-4 quantifies “where and how” good the prediction is for each map location. The Error surface is calculated by subtracting the values on the Actual map from the values on the Predicted map—a negative value indicating an under estimate; a positive value an over estimate. The color ramp for the Error surface extends from -13.1 (under estimates in red tones) through 0.0 (identical in yellow) to 22.9 (over estimates in green and blue tones). It appears that the prediction equation substantially over estimates as most of the error is between 5.6 and 16.9 above the Actual values (mean of 11.10 plus/minus standard deviation of 5.84).

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 spatial patterns. The next section investigates some of refinement techniques and what information can be gleaned by analyzing the error surface.

### 8.4 Stratifying Maps for Better Predictions

The previous 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 or more, 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.

In the example discussed in the last section, regression analysis was used to relate a map of loan Concentration to maps housing Age, Value and Density. Then the equation was used to derive a map of predicted loan Concentration based on the independent map variables and the results evaluated for how well the prediction equation performed.

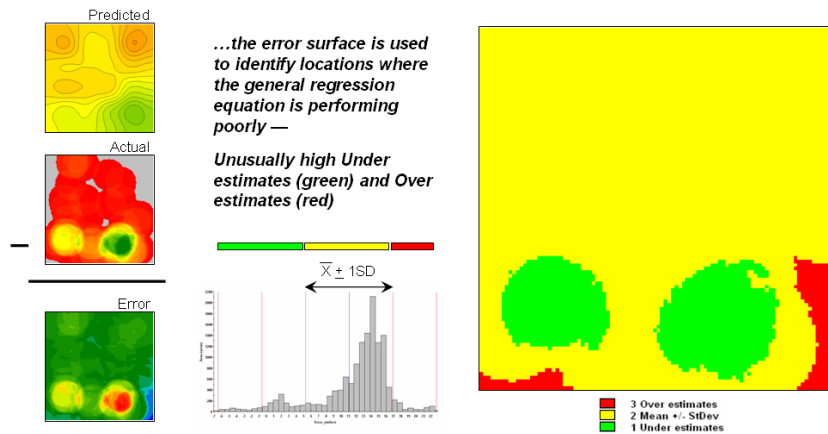


Figure 8-5. Using prediction errors to stratify.

The left side of figure 8-5 shows the evaluation procedure that subtracts the Actual loan Concentration values from the predicted ones to derive the Error surface. The previous discussions noted that the concentration “guesses” weren’t too bad—at least the general geographic trends were the same.

One way to improve the predictions, however, 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. Stratification 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 is a frequently used technique. The histogram in the center of figure 8-5 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 (green), zone 2 of typical error (yellow) and zone 3 of unusually high over-guesses (red). The map on the right of the figure locates the three strata throughout the project area.

The rationale behind the stratification is that the single general prediction equation is having trouble accounting for the radically different responses in zones 1 and 3. The assumption is that conditions within zone 1 make the equation under estimate while conditions within zone 3 cause it to over estimate. If the assumption holds, one would expect a set of tailored equations, one for each zone, would be better at predicting than the overall equation. This result generally is the case.

A couple of things should be noted about stratified prediction maps. First, that there is a myriad of ways to stratify mapped data—1) *Geographic Zones*, such as proximity to known break points; 2) *Dependent Map Zones*, such as

areas of low, medium and high dependent variable responses; 3) *Data Zones*, such as Clustering and Level-slice discussed in Topic 7; and 4) *Correlated Map Zones*, such as known neighborhood divisions. 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 geo-business are in their infancy. As the business applications embrace spatial technology, research will tailor the procedures for the unique data and situations encountered.

### 8.5 Conclusion

The *Spatial Analysis, Surface Modeling and Spatial data Mining/Prediction* operations introduced in this book identify an important compartment in the emerging geo-business toolkit. Understanding the spatial, as well as the numerical distribution of data provides considerable insight into relationships and interactions on a site-specific basis. While broad, project area averages give us a general feel for the data they are often woefully inadequate for tailoring decisions and actions in a manner that considers spatial patterns.

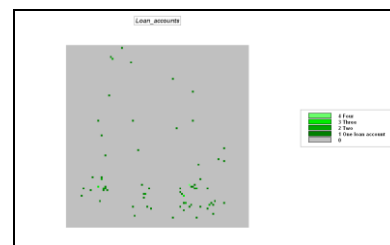
Geo-business is at a threshold akin to the concepts of the “assembly line and inter-changeable parts” in the mid-1800s and the “railhead distribution centers” in the early 1900s. The computer and digital maps form the cornerstone of an entirely new perspective on how we analyze data and formulate management actions. The vital ingredient needed is innovation by researchers and professionals who understand geo-business data and can “*think with maps.*”

The application of this new perspective within geo-business is in its infancy. At present, a full implementation of the geo-business toolbox is still in the hands of the developers, researchers and innovative businesses. Advancements in mapped data analysis and spatial modeling await contributions from a new generation of business professionals. The considerable knowledge and methodologies of current and future management science need to be reviewed for their spatial inferences and insight.

*The Analyzing Geo-Business Data discussion and hands-on experience in this book are intended to fuel this revolution.* The procedures introduced are just the tip of the spatial analysis and data mining iceberg that is rapidly changing our paradigm on the nature of business data and its infusion into spatially relevant decisions.

### 8.6 Exercises

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



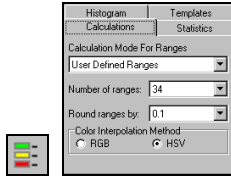
#### 8.6.1 Deriving a Dependent Map Variable



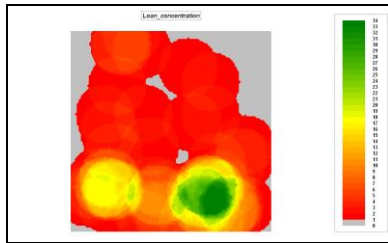
Use the *View* tool to generate a display of the **Loan\_accounts** map.

Press the *Map Analysis* button, select **Neighbors** → **Scan** and complete the following command.

*SCAN Loan\_accounts Total IGNORE 0.0 WITHIN 15 CIRCLE FOR Loan\_concentration*

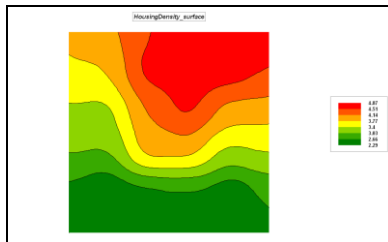


Use the *Shading Manager* to theme the map with 34 User Defined Ranges (1 unit interval) from 0 (light grey) to 1 (red) through 34 (green) with a yellow color inflection at mid range.

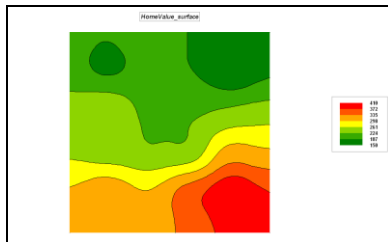


### 8.6.2 Scatter Plots of Map Correlation

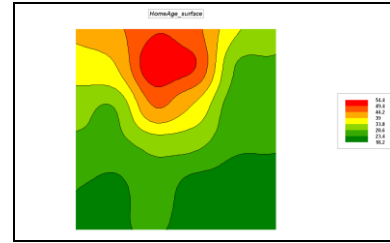
Use the *View* tool to create displays of...



*HousingDensity\_surface*

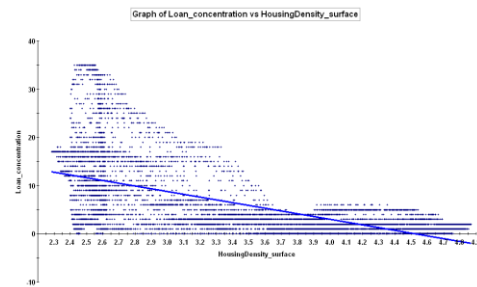
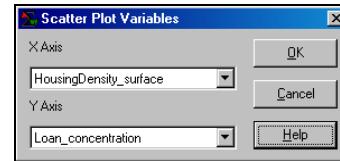


*HomeValue\_surface*

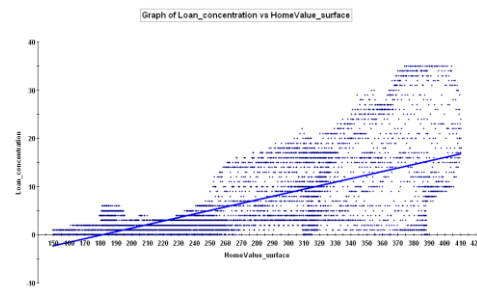


*HomeAge\_surface*

Generate a *Scatter Plot* and Regression equation for **Loan\_concentration** and each of these maps by selecting **Map Set** → **New Graph** → **Scatter Plot** and complete the dialog box as shown.

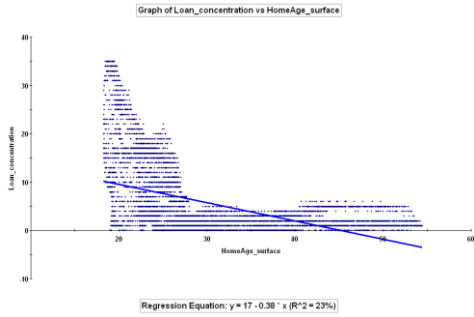


*L\_concentration vs. H\_Density*



*L\_concentration vs. H\_Value*





*L\_concentration vs. H\_Age*

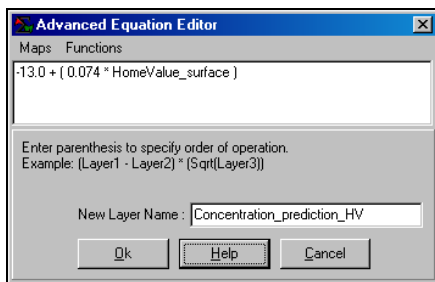
### 8.6.3 Spatial Regression

Recall the scatter plot for the *L\_concentration* vs. *HousingValue\_surface* relationship. Note the regression equation...

$$Y = -13.0 + 0.074 * X$$

..and its R-squared index (46%).

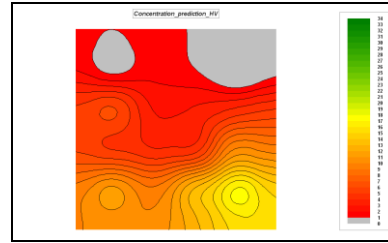
Press the *Map Analysis* button, select **Overlay** → **Calculate** and complete the following command to evaluate the univariate regression equation (one independent map variable) and generate a predicted loan Concentration surface (Y) using the home Value information (X).



Select **Reclassify** → **Renumber** and set all negative predictions to zero.

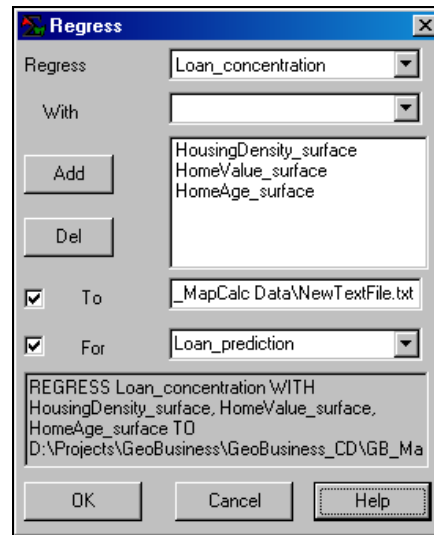
```
RENUMBER Concentration_prediction_HV
ASSIGNING 0 TO -2 THRU 0 FOR
Concentration_prediction_HV
```

Use the *Shading Manager* to apply the same legend as the **Loan\_concentration** map—34 User Defined Ranges (1 unit interval) from 0 (light grey) to 1 (red) through 34 (green) with a yellow color inflection at mid range.



Note that the regression equation severely under estimates loan concentration (no values above 18; green tones).

Now generate a multivariate regression equation (more than one independent map variable) by completing the following command.

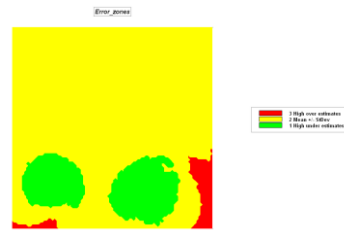
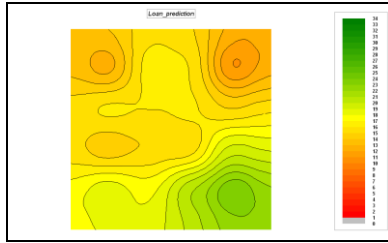


Note the regression equation...


*Least Squares Coefficients*

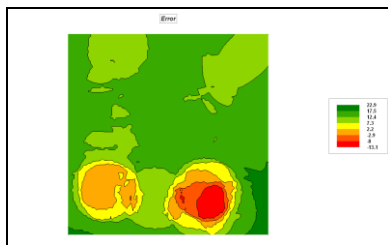
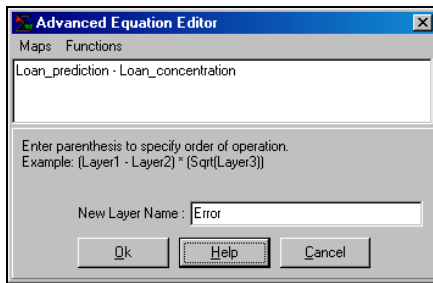
$$Y Est = -9.157 + 0.097 * HomeAge_surface + 0.067 * HomeValue_surface - 1.557 * HousingDensity_surface$$

Use the *Shading Manager* to apply the same legend as the **Loan\_concentration** map—34 User Defined Ranges (1 unit interval) from 0 (light grey) to 1 (red) through 34 (green) with a yellow color inflection at mid range.



### 8.6.4 Evaluating Regression Performance

 Press the *Map Analysis* button, select **Overlay** → **Calculate** and complete the following command to calculate the difference between the actual and predicted loan concentration maps.



Histogram	Templates
Calculations	Statistics
Min:	-13.1
Max:	22.9
Range:	36
Mean:	11.1
Median:	12.9
Std. Dev.:	5.84
Variance:	34.1
Gridded Area:	14,348 acres

Select **Reclassify** → **Renumber** and complete the following command to isolate three error zones—

- Zone 1 – Unusually high under estimate
- Zone 2 – Mean +/- 1 StDev
- Zone 3 – Unusually high over estimate