

Topic 6

Surface Modeling

6.1 Identifying Customer Pockets

Geo-coding based on customer address is a powerful capability in most desktop mapping systems. It automatically links customers to digital maps like old pushpins on a map on the wall. Viewing the map provides insight into spatial patterns of customers. Where are the concentrations? Where are customers sparse? See any obvious associations with other map features, such as highways or city neighborhoods?

The spatial relationships encapsulated in the patterns can be a valuable component to good business decisions. However, the “visceral viewing” approach is more art than science and is very subjective. Grid-based map analysis, on the other hand, provides tools for objectively evaluating spatial patterns. Last month’s column discussed a *Competition Analysis* procedure that linked travel-time of customers to a retail store. This month’s discussion will focus on characterizing the spatial pattern of customers.

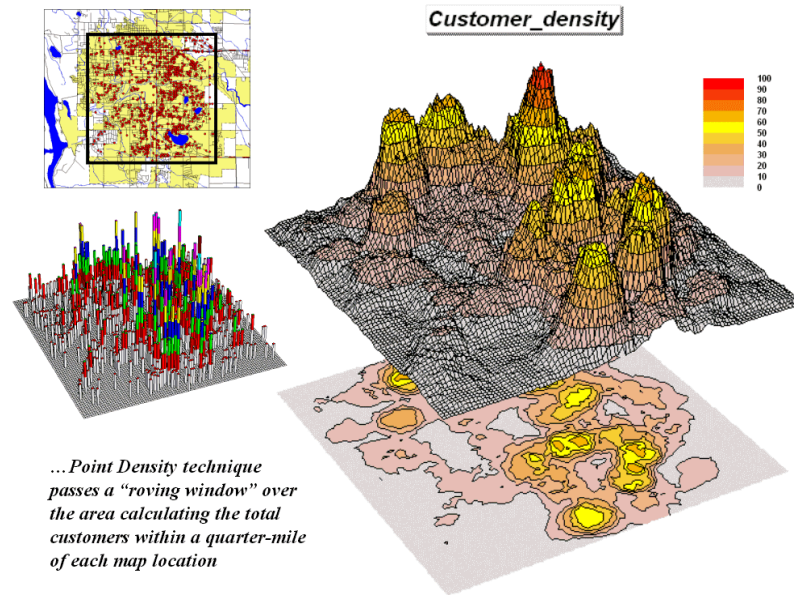


Figure 6-1. Point Density analysis identifies the number of customers with a specified distance of each grid location.

The upper left inset in figure 6-1 identifies the location of customers as red dots. Note that the dots are concentrated in some areas (actually falling on top of each) while in other areas there are very few dots. Can you locate areas with unusually high concentrations of customers? Could you delineate these important areas with a felt-tip marker? How confident would you be in incorporating your sketch map into critical marketing decisions?

The lower left inset identifies the first step in a quantitative investigation of the customer pattern—*Point Density* analysis. An analysis grid of 100 columns by 100 rows (10,000 cells) is superimposed over the project area and the number of customers falling into each cell is aggregated. The higher “spikes” on the map identify greater customer tallies. From this perspective your eyes associate big bumps with greater customer concentrations.

The map surface on the right does what your eyes were attempting to do. It summarizes the number of customers within the vicinity of each map location. This is accomplished by moving a “roving window” over the map that calculates the total number of customers within a six-cell reach (about a quarter of a mile). The result is obvious peaks and valleys in the surface that tracks customer density.

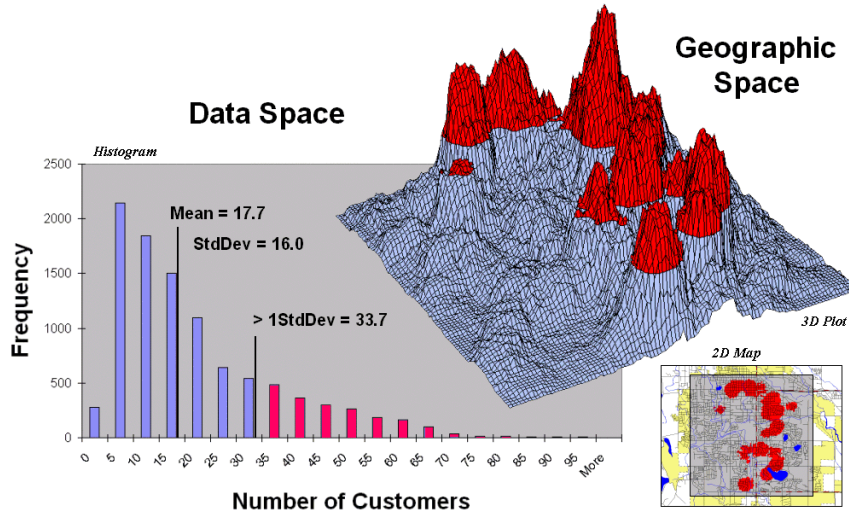


Figure 6-2. Pockets of unusually high customer density are identified as more than one standard deviation above the mean.

Figure 6-2 shows a process to identify pockets of unusually high customer density. The mean and standard deviation of the customer density surface are calculated. The histogram plot on the left graphically shows the cut-off used—more than one standard deviation above the mean ($17.7 + 16 = 33.7$). *Aside—since the customer data isn't normally distributed it might be better to use Median plus Quartile Range for the cut-off.* The red-capped peaks in the surface map on the right spatially locate these areas. The lower-right inset shows the areas transferred to a desktop mapping system. How well do you think your visual delineations would align?

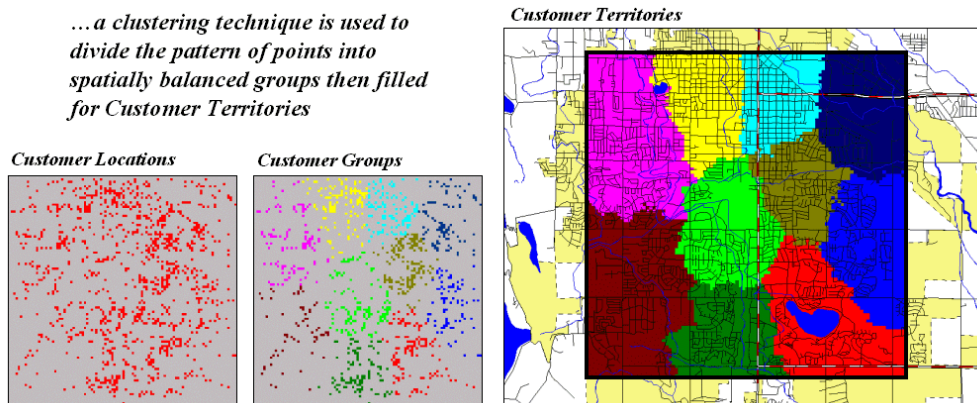


Figure 6-3. Clustering on the latitude and longitude coordinates of point locations identify customer territories.

Another grid-based technique for investigating the customer pattern involves **Point Territories** assignment. This procedure looks for inherent spatial groups in the data and assigns customers to contiguous areas. In the example, you might want to divide the customer locations into ten groups for door-to-door contact on separate days.

The two small inserts on the left of figure 6-3 show the general pattern of customers then the partitioning of the pattern into spatially balanced groups. This initial step was achieved by applying a *K-means* clustering algorithm to the latitude and longitude coordinates of the customer locations. In effect this procedure maximizes the differences between the groups while minimizing the differences within each group. There are several alternative approaches that could be applied, but K-means is an often-used procedure that is available in all statistical packages and a growing number of GIS systems.

The final step to assign territories uses a *nearest neighbor* interpolation algorithm to assign all non-customer locations to the nearest customer group. The result is the customer territories map shown on the right. The partitioning based on customer locations is geographically balanced, however it doesn't consider the number of customers within each group—that varies from 69 in the lower right (maroon) to 252 (awful green) near the upper right. We'll have to tackle that twist in a future beyond mapping column.

6.2 From Point Samples to Map Surfaces

Most of us are familiar with the old “bell-curve” for school grades. You know, with lots of C's, fewer B's and D's, and a truly select set of A's and F's. Its shape is a perfect bell, symmetrical about the center with the tails smoothly falling off toward less frequent conditions.

However the *normal distribution* (bell-shaped) isn't as normal (typical) as you might think. For example, *Newsweek* recently noted that the average grade at a major ivy-league university isn't a solid C with a few A's and F's sprinkled about as you might imagine, but an A- with a lot of A's trailing off to lesser amounts of B's, C's and (heaven forbid) the very rare D or F.

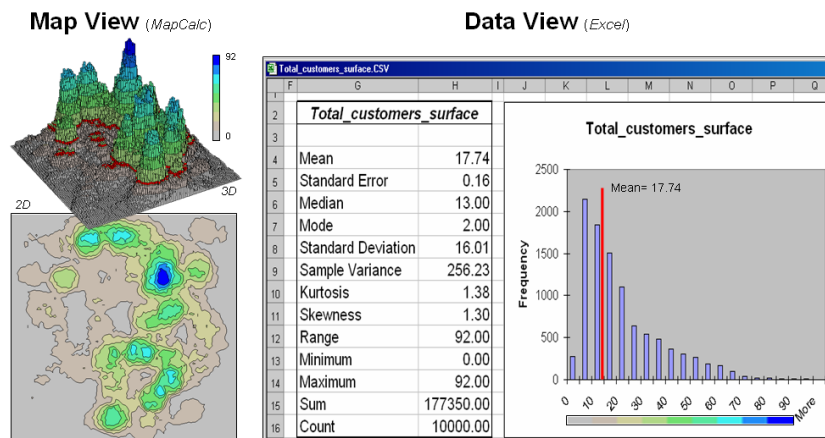


Figure 6-4. Mapped data are characterized by their geographic distribution (maps on the left) and their numeric distribution (descriptive statistics and histogram on the right).

The frequency distributions of mapped data also tend toward the *ab-normal* (formally termed *asymmetrical*). For example, consider the customer density data shown in the figure 6-4. The geographic distribution of the data is characterized in the *map view* by the 2D contour map and 3D surface on the left. Note the distinct pattern of the terrain with bigger bumps (higher customer density) in the central portion of the project area. As is normally the case with mapped data, the map values are neither uniformly nor randomly distributed in geographic space. The unique pattern is the result of complex spatial processes determining where people live that are driven by a host of factors—not spurious, arbitrary, constant or even “normal” events.

Now turn your attention to the numeric distribution of the data depicted in the right side of the figure. The *data view* was generated by simply transferring the grid values in the analysis frame to Excel, then applying the *Histogram* and *Descriptive Statistics* options of the Data Analysis add-in tools. The mechanics used to plot the histogram and generate the statistics were a piece-of-cake, but the real challenge is to make some sense of it all.

Note that the data aren't distributed as a normal bell-curve, but appear shifted (termed skewed) to the left. The tallest spike and the intervals to its left, match the large expanse of grey values in the map view—frequently occurring values. If the surface contained a disproportionately set of high value locations, there would be a spike at the high end of the histogram. The red line in the histogram locates the mean (average) value for the numeric distribution. The red line in the 3D map surface shows the same thing, except its located in the geographic distribution.

The mental exercise linking geographic space with data space is a good one, and some general points ought to be noted. First, there isn't a fixed relationship between the two views of the data's distribution (geographic and data). A myriad of geographic patterns can result in the same histogram. That's because spatial data contains additional information—*where*, as well as *what*—and the same data summary of the “what's” can reflect a multitude of spatial arrangements (“where's”).

But is the reverse true? Can a given geographic arrangement result in different data views? Nope, and it's this relationship that catapults mapping and geo-query into the arena of mapped data analysis. Traditional analysis techniques assume a functional form for the frequency distribution (histogram shape), with the standard normal (bell-shaped) being the most prevalent.

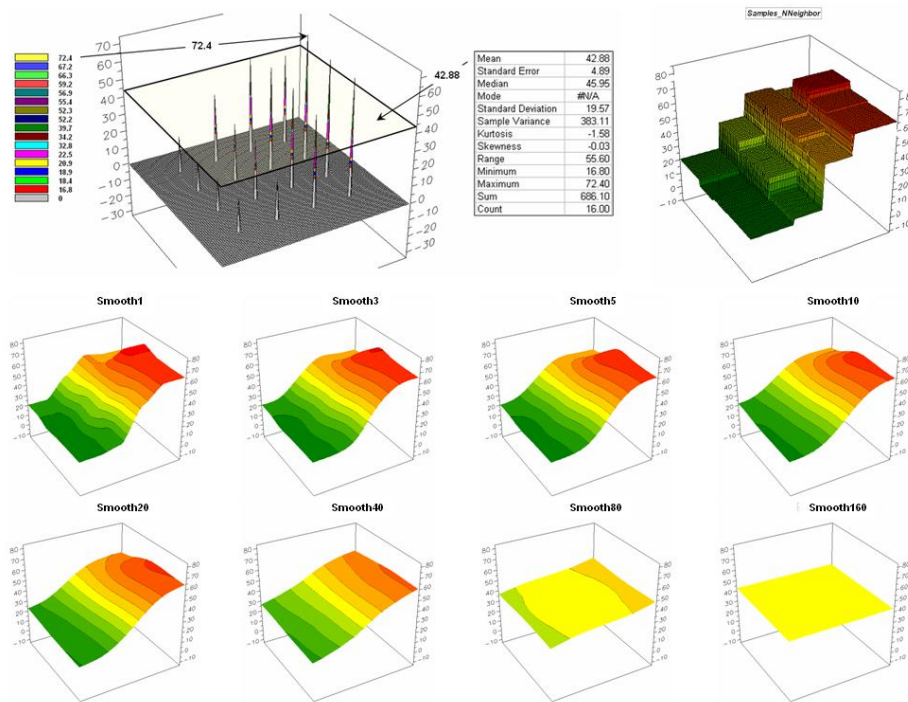


Figure 6-5. The spatial distribution implied by a set of discrete sample points can be estimated by iterative smoothing of the point values.

Figure 6-5 offers yet another perspective of the link between numeric and geographic distributions. The upper-left inset identifies the spatial pattern formed by 16 samples of the “percent of home equity loan limit” for the Smallville project area—ranging from 16.8 to 72.4. The table reports the numerical pattern of the data—mean= 42.88 and standard deviation= 19.57. The coefficient of variation is 45.6% ((19.57/42.88) * 100= 45.6%) suggesting a fairly large unexplained variation among the data values.

In a geographic context, the mean represents a horizontal plane hovering over the project area. However, the point map suggests a geographic trend in the data from values lower than the mean in the west toward higher values in the east. The inset in the upper-right portion of figure 6-5 shows a “nearest neighbor” surface generated by assigning the closest sample value to all of the non-sampled locations in the project area. While the distribution is blocky it

serves as a first order estimate of the geographic distribution you see in the point map—lower in the east and higher in the west.

The series of plots in the lower portion of Figure 6-5 shows the results of *iteratively smoothing* the blocky data. This process repeatedly passes a “roving window” over the area that calculates the average value within quarter-mile. The process is analogous to chipping away at the stair steps with the rubble filling in the bottom. The first smoothing still retains most of the sharp peak and much of the angular pattern in the blocky surface. As the smoothing progresses (left to right) the surface takes on the general geographic trend of the data (Smooth10).

Eventually the surface is eroded to a flat plane—the mean of the data. The progressive series of plots illustrate a very important concept in surface modeling—the *geographic distribution maps the variance*. Or, in other words, a map surface uses the geographic pattern in a data set to help explain the data’s variation.

6.3 The Keystone Concept

The basic concept of surface modeling involves *spatial autocorrelation*, referring to the degree of similarity among neighboring sample points. If they exhibit a lot similarity, or *spatial dependence*, they ought to derive a good map. If they are spatially independent, then expect a map of pure, dense gibberish to be generated from the sample points.

So how can we measure whether “*nearby things are more related than distant things*”—the first law of geography.

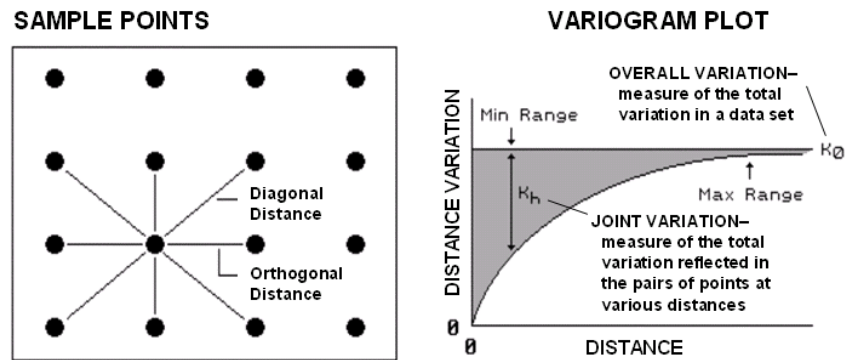


Figure 6-6. A variogram plot depicts the relationship between distance and measurement similarity (spatial autocorrelation).

Common sense suggests that more similarity exists among the neighboring samples (lines in the left side of figure 6-6) than among sample points farther away. Natural and social systems tend to form niches that transition over geographic space instead of random or check board patterns. Since all surface modeling techniques use nearby data values (roving window), if there is a lot of spatial autocorrelation in a set of samples, expect a good map; if not, expect a map of pure, dense gibberish.

An index of spatial autocorrelation in a data set compares the differences between nearby sample pairs with those from the average of the entire data set. One would expect a sample point to be more like its neighbor than it is to the overall average. The larger the disparity between the nearby and average figures the greater the spatial dependency and the likelihood of a good interpolated map.

The mathematical procedure for calculating spatial autocorrelation index is simple—calculate the difference between each sample value and its nearest neighbor ($|Value - NN_Value|$), then compare the differences to those based on the typical condition ($|Value - Average|$). If the Nearest Neighbor and Average differences are about the same, little spatial dependency exists. If the nearby differences are substantially smaller than the typical differences, then strong positive spatial dependency is indicated and it is safe to assume that nearby things are more related.

The spatial autocorrelation index, however, is limited as it merely assesses the closest neighbor, regardless of its distance. That’s where a *variogram* comes in (right side of figure 6-6). It is a plot of the similarity among all

sample values based on the distance between them. Instead of simply testing whether close things are related, it shows how the degree of dependency relates to varying distances between locations.

The distance between a pair of points is calculated by the Pythagorean Theorem and plotted along the X-axis. A normalized difference between sample values (termed *semi-variance*) is calculated and plotted along the Y-axis. Each point-pair is plotted and the pattern of the points analyzed. The origin of the plot at $0,0$ is a unique case. The distance between samples is zero; therefore, there is no dissimilarity (data variation = 0) as at location is exactly the same as itself.

As the distance between points increase, subsets of the data are scrutinized for their dependency. The shaded portion in the idealized plot shows how quickly the spatial dependency among points deteriorates with distance. The maximum range (*Max Range*) position identifies the distance between points beyond which the data values are considered independent. This tells us that using data values beyond this distance for interpolation do not help in producing a map (and actually could mess-up things).

The minimum range (*Min Range*) position identifies the smallest distance contained in the actual data set and is determined by the sampling design used to collect the data. If a large portion of the shaded area falls below this distance, it tells you there is insufficient spatial dependency in the data set to warrant interpolation. If you proceed with the interpolation, a nifty colorful map will be generated, but likely questionable.

Spatial autocorrelation exists if the differences between sample values systematically increase as the distances between sample points becomes larger. The shape and consistency of the pattern of points in the plot characterize the degree of similarity. In the figure, an idealized upward curve is indicated. If the remaining point-pairs continue to be tightly clustered about the curve considerable spatial autocorrelation is indicated. If they are scattered throughout the plot without forming a recognizable pattern, minimal autocorrelation is present. Worse yet, if the variogram of the sample data plots as a straight line or circle, no spatial dependency exists and the map will be worthless. The “goodness of fit” of the points to the curve serves as a measure of the spatial dependency—a good fit indicates strong spatial autocorrelation.

6.4 IDW and Krig Techniques

Statistical sampling has long been at the core of business research and practice. Traditionally point-sampled data were analyzed by non-spatial statistics to identify the “typical” level of sales, housing prices, customer income, etc. throughout an entire city, state or region. Considerable effort was expended to determine the best single estimate and assess just how good the “average” estimate was in typifying the extended geographic area.

However non-spatial techniques fail to make use of the geographic patterns inherent in the data to refine the estimate—the typical level is assumed everywhere the same throughout a project area. The computed variance (or standard deviation) indicates just how good this assumption is—the larger the standard deviation the less valid is the assumption “*everywhere the same.*”

Spatial interpolation, on the other hand, utilizes the spatial patterns in a data set to generate localized estimates throughout the sampled area. Conceptually it “*maps the variance*” by using geographic position to help explain the differences in the sample values. In practice, it simply fits a continuous surface (kind of like a blanket) to the point data spikes (figure 6-7).

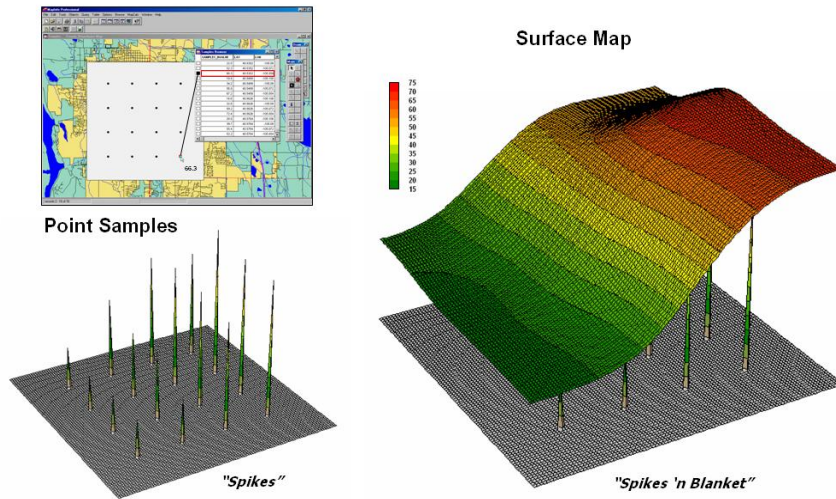


Figure 6-7. Spatial interpolation involves fitting a continuous surface to sample points.

While the extension from non-spatial to spatial statistics is quite a theoretical leap, the practical steps are relatively easy. The left side of figure 6-7 shows 2D and 3D *point* maps of the loan samples described in the previous section. This highlights the primary difference from traditional sampling—each sample must be *geo-referenced* as it is collected. In addition, the *sampling pattern and intensity* are often different than traditional sampling methods to maximize spatial information within the data collected (see author’s note).

The surface map on the right side of figure 6-7 translates pattern of the “spikes” into the peaks and valleys of the surface map. The traditional, non-spatial approach, if mapped, would be a flat plane (average everywhere) aligned within the yellow zone. Its “everywhere the same” assumption fails to recognize the patterns of larger levels (reds) and smaller levels (greens). A decision based on the average level (42.88%) would be ideal for the yellow zone but would likely be inappropriate for most of the project area as the data vary from 16.8 to 72.4 percent.

The simplest spatial interpolation technique is the iteratively smoothed one previously described—go to a location and simply average the data within a specified distance. All spatial interpolation techniques establish a "roving window" that

- moves to a location in a field,
- calculates an estimate (guess) based on the point samples around it,
- assigns the estimate to the center of the window and
- moves to the next location.

The extent of the window (both size and shape) affects the result, regardless of the summary technique. In general, a large window capturing a bunch of values tends to "smooth" the data. A smaller window tends to result in a "rougher" surface with abrupt transitions.

Three factors affect the window's extent: its reach, the number of samples, balancing. The *reach*, or search radius, sets a limit on how far the computer will go in collecting data values. The *number of samples* establishes how many data values should be used. If there are more than enough values within a specified reach, the computer uses just the closest ones. If there aren't enough values, it uses all that it can find within the reach. *Balancing* attempts to eliminate directional bias by ensuring that the values are selected in all directions around window's center. Once a window is established, the summary technique comes into play.

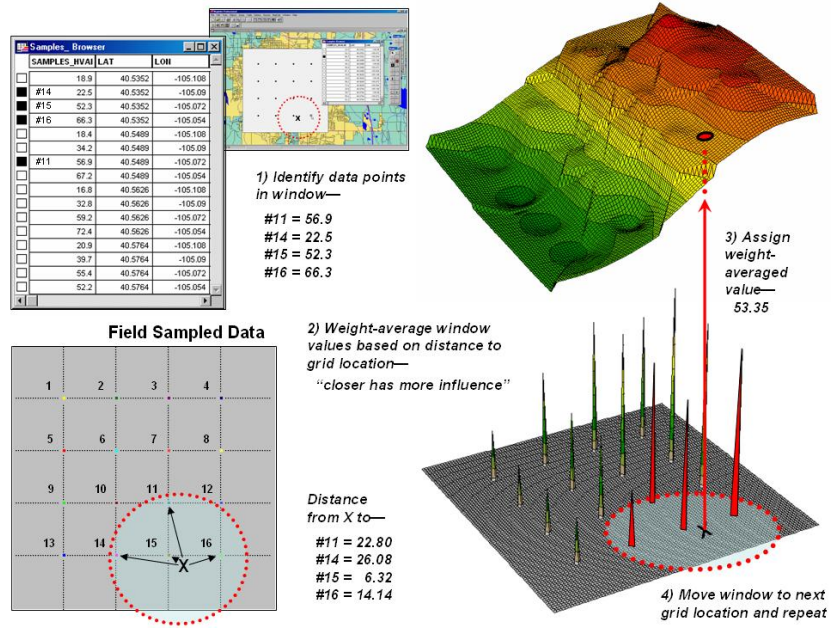


Figure 6-8. Inverse distance weighted interpolation weight-averages sample values within a roving window.

Inverse distance is an easy spatial interpolation technique to conceptualize (see figure 6-8). It estimates the value for a location as an average of the data values within its vicinity. The average is weighted in a manner that decreases the influence of the surrounding sample values as the distance increases. In the figure, the estimate of 53.35 is the "inverse distance-squared ($1/d^2$) weighted average" of the four samples in the window. Sample #15 (the closest) influences the average a great deal more than sample #14 (the farthest away). Table 6-1 shows the specific calculations.

Table 6-1. Example Calculations for Inverse Distance Squared Interpolation

Sample	Row	Column	Value	Distance	Weight ($1/D^2$)	Value * Weight
#11	41	60	56.90	22.80	0.0019	0.11
#14	21	40	22.50	26.08	0.0015	0.03
#15	21	60	52.30	6.32	0.0250	1.31
#16	21	80	66.30	14.14	<u>0.0050</u>	<u>0.33</u>
				Sum=	0.03	1.78
X Grid Location	19	66	???	X value ???=	$1.77/0.03=$	53.35

Because this is a static averaging method, the estimated values can never exceed the range of values in the original field data. Also, inverse distance tends to "pull-down peaks and pull-up valleys" in the data. The technique is best suited for data sets with random samples that are relatively independent of their surrounding locations (i.e., minimal regional trend).

The right portion of figure 6-8 contains three-dimensional (3-D) plots of the point sample data and the inverse distance-squared surface generated. The estimated value in the example can be conceptualized as "sitting on the surface," 19 units above the base (zero). Note that the surface generated by the inverse distance technique is sensitive to sampled locations and tends to put bumps around the sampled areas. Greater smoothing would occur by using a larger window.

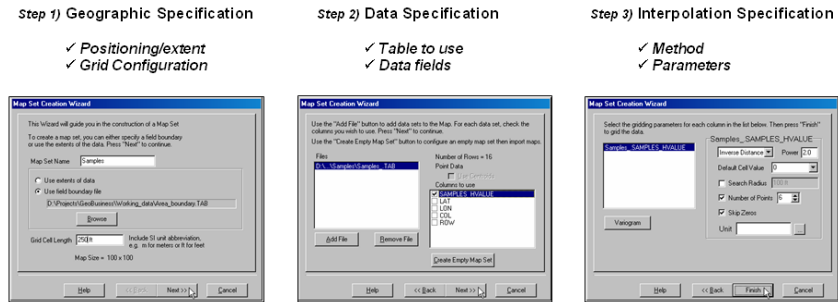


Figure 6-9. A wizard interface guides a user through the necessary steps for interpolating sample data.

The mechanics of generating an interpolated surface involves three steps relating to *geography*, *data* and *interpolation methods* (figure 6-9). The first step establishes the geographic position, project area extent and grid configuration to be used. While these specifications can be made directly, it is easiest to simply reference an exiting map of the boundary of the project area (*Positioning* and *Extent*) then enter the grid spacing (*Configuration*—250 feet on the side of a cell).

The next step identifies the table and data fields to be used. The user navigates to the file (*Data Table*) then simply checks the maps to be interpolated (*data field*—Sample_values). The final step establishes the interpolation method and necessary factors. In the example, the default *Inverse Distance Squared* ($1/D^2$) method was employed using the six nearest sample points.

Other methods, such as Kriging, could be specified resulting in a somewhat different surface as discussed in the next section when interpolation accuracy is evaluated—whether the map is a good one or a bad one. Kriging is a sophisticated that uses the variogram to summarize spatial patterns in the data to establish the window size and sample weights. In data that exhibit a trend Kriging usually produces better interpolation surfaces than IDW.

6.5 Benchmarking Interpolation Results

For some, the previous discussion on generating maps from samples might have been too simplistic—enter a few things then click on a data file and, *viola*, you have a loan percentage surface artfully displayed in 3D with a bunch of cool colors draped all over it.

Actually, it is that easy to create one. The harder part is figuring out if the map generated makes sense and whether it is something you ought to use in analysis and important business decisions. This section discusses the relative amounts of information provided by the non-spatial arithmetic average and site-specific maps by comparing the average and two different interpolated map surfaces. Discussion in the next section describes a procedure to quantitatively assess whether a particular map is a keeper.

The top-left inset in figure 6-10 shows the map of the loan data’s average. It’s not very exciting and looks like a pancake but that’s because there isn’t any information about spatial variability in an average value—assumes 42.88 percent is everywhere in the project area.

The non-spatial estimate simply adds up all of the sample values and divides by the number of samples to get 42.88 percent. Since the procedure fails to consider where the different samples were taken, it can’t map the variations in the measurements. It suggests that the average is everywhere, plus or minus the standard deviation. But there is no spatial guidance where values might be more, or where they might be less.

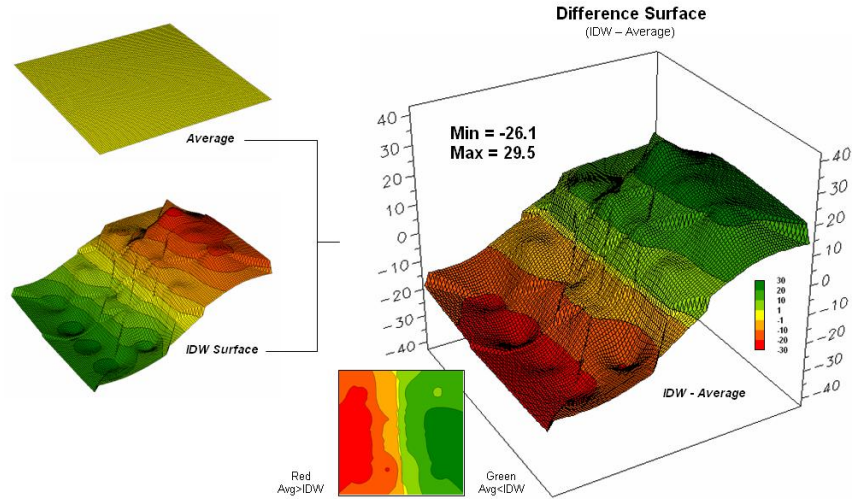


Figure 6-10. Spatial comparison of the project area average and the IDW interpolated surface.

The spatially based estimates are shown in the map surface just below the pancake. As described in the last section, spatial interpolation looks at the relative positioning of the samples values as well as their measure of loan percentage. In this instance the big bumps were influenced by high measurements in that vicinity while the low areas responded to surrounding low values.

The map surface in the right portion of figure 6-10 objectively compares the two maps simply by subtracting them. The colors were chosen to emphasize the differences between the whole-field average estimates and the interpolated ones. The yellow band indicates the no difference while the progression of green tones locates areas where the interpolated map estimated higher values than the average. The progression of red tones identifies the opposite with the average estimate being more than the interpolated ones.

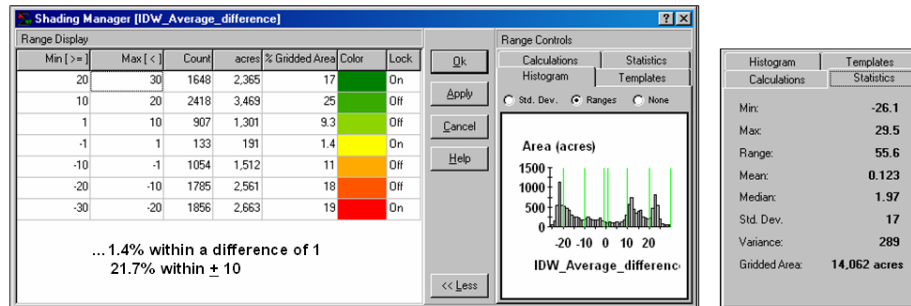


Figure 6-11. Statistics summarizing the difference between the IDW surface and the Average.

The information in figure 6-11 shows that the difference between the two maps ranges from -26.1 to +29.5 percent. If one assumes that +/- 10 percent difference won't significantly alter a decision, then about one-quarter of the area (9.3+1.4+11= 21.7%) is adequately covered by the overall average of the sample data. But that leaves about three-fourths of the area that is well-below the average (18 + 19 = 37%) or well-above (25+17 = 42%). Using the sample data average in either of these areas could lead to poor decisions as the assumption that the average is typical does not hold.

Now turn your attention to figure 6-12 that compares maps derived by two different interpolation techniques—IDW (inverse distance-weighted) and Krig. Note the similarity in the peaks and valleys of the two surfaces. While subtle differences are visible the general trends in the spatial distribution are similar.

The difference map on the right confirms the coincident trends. The narrow band of yellow identifies areas that are nearly identical (within +/- 1.0). The light red locations identify areas where the IDW surface estimates a bit lower than the Krig ones (within -10); light green a bit higher (+10). Applying the same assumption about +/- 10 percent difference being negligible for decision-making the maps are effectively identical.

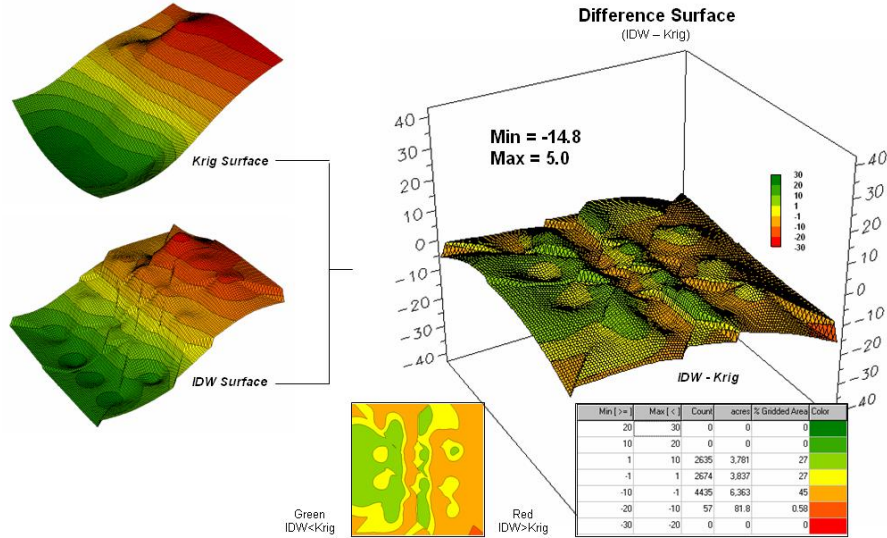


Figure 6-12. Spatial comparison of IDW and Krig interpolated surfaces.

So what’s the bottom line? That there are substantial differences between an arithmetic average and interpolated surfaces—at least for this data set. It suggests that quibbling about the best interpolation technique isn’t as important as using an interpolated surface (any surface) over the project area average. However, what needs to be addressed is whether an interpolated surface (any surface) actually reflects the real spatial distribution. That weighty question is the focus of the next section.

6.6 Assessing Interpolation Performance

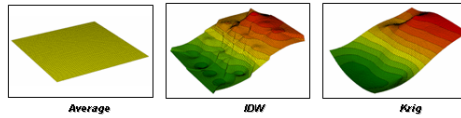
The previous discussion compared the assumption of the field average with map surfaces generated by two different interpolation techniques for phosphorous levels throughout a field. While there was considerable difference between the sample average and the IDW surface, there was relatively little difference between the IDW and Krig surfaces.

But which surface best characterizes the spatial distribution of the sampled data? The answer to this question lies in **Residual Analysis**—a technique that investigates the differences between *estimated* and *measured* values throughout an area. It is common sense that one shouldn’t simply accept interpolated surface without checking out its accuracy. Cool graphics just aren’t enough.

Ideally, one designs an appropriate sampling pattern and then randomly locates a number of “test points” to assess interpolation performance. So which estimate—average, IDW or Krig—did a better job in estimating the measured loan percent levels in the test set?

The table in figure 6-13 reports the results for twelve randomly positioned test samples. The first column identifies the sample ID and the second column reports the actual measured value for that location. Column C simply depicts estimating the project area average (42.88) at each of the test locations. Column D computes the difference of the “estimate minus actual”—formally termed the *residual*. For example, the first test point (ID#1) estimated the average of 42.88 but was actually measured as 55.2 so -12.32 is the residual (42.88-55.20= -12.32) ...quite a bit off. However, point #6 is a lot better (42.88-49.40= -6.52).

...Residual Analysis is used to evaluate interpolation performance (Krig at .03 is best)



	A	B	C	D	E	F	G	H
Test Set (randomly selected)								
ID		Actual	Average	Avg-Actual	IDW	IDW-Actual	Krig	Krig-Actual
1		55.20	42.88	-12.32	53.50	-1.70	53.60	-1.60
2		31.40	42.88	11.48	29.50	-1.90	30.60	-0.80
3		50.30	42.88	-7.42	50.60	0.30	49.20	-1.10
4		17.90	42.88	24.98	21.90	4.00	18.20	0.30
5		65.10	42.88	-22.22	63.90	-1.20	67.50	2.40
6		49.40	42.88	-6.52	51.20	1.80	50.30	0.90
7		17.80	42.88	25.08	20.40	2.60	16.40	-1.40
8		68.50	42.88	-25.62	66.50	-2.00	69.00	0.50
9		33.70	42.88	9.18	34.20	0.50	34.40	0.70
10		67.80	42.88	-24.92	64.90	-2.90	70.90	3.10
11		22.40	42.88	20.48	23.60	1.20	21.30	-1.10
12		55.60	42.88	-12.72	54.90	-0.70	57.40	1.80
Average values=		44.59	42.88		44.59		44.90	
Residual sum=				-20.54		0.00		3.70
Average error=				16.91		1.73		1.31
Normalized error=				0.38		0.04		0.03

Figure 6-13. A residual analysis table identifies the relative performance of average, IDW and Krig estimates.

The residuals for the IDW and Krig maps are similarly calculated to form columns F and H, respectively. First note that the residuals for the project area average are generally larger than either those for the IDW or Krig estimates. Next note that the residual patterns between the IDW and Krig are very similar—when one is off, so is the other and usually by about the same amount. A notable exception is for test point #4 where IDW dramatically over-estimates.

The rows at the bottom of the table summarize the residual analysis results. The *Residual sum* row characterizes any bias in the estimates—a negative value indicates a tendency to underestimate with the magnitude of the value indicating how much. The -20.54 value for the whole-field average indicates a relatively strong bias to underestimate.

The *Average error* row reports how typically far off the estimates were. The 16.91 figure for area average is about ten times worse than either IDW (1.73) or Krig (1.31). Comparing the figures to the assumption that a plus/minus 10 difference is negligible in decision-making it is apparent that the project area estimate is inappropriate to use and that the accuracy differences between IDW and Krig are very minor.


The *Normalized error* row simply calculates the average error as a proportion of the average value for the test set of samples ($1.73/44.59 = .04$ for IDW). This index is the most useful as it allows you to compare the relative map accuracies between different maps. Generally speaking, maps with normalized errors of more than .30 are suspect and one might not want to make important decisions using them.

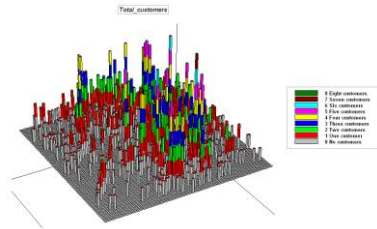
So what’s the bottom-bottom line? That Residual Analysis is an important component of geo-business data analysis. Without an understanding of the relative accuracy and interpolation error of the base maps, one can’t be sure of the recommendations and decisions derived from the interpolated data. The investment in a few extra sampling points for testing and residual analysis of these data provides a sound foundation for management decisions. Without it, the process can become one of blind faith and wishful thinking.


6.6 Exercises

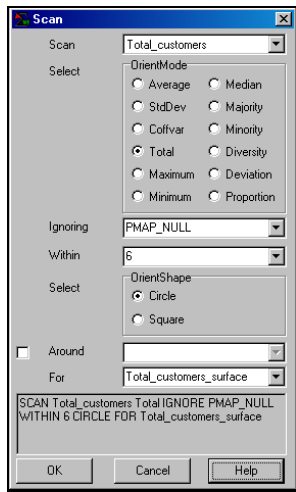
Access *MapCalc* using the *GB_Smallville.rgs* by selecting **Start** → **Programs** → **MapCalc Learner** → **MapCalc Learner** → **Open existing map set** → *GB_Smallville.rgs*. The following set of exercises utilizes this database.

6.6.1 Deriving Customer Density

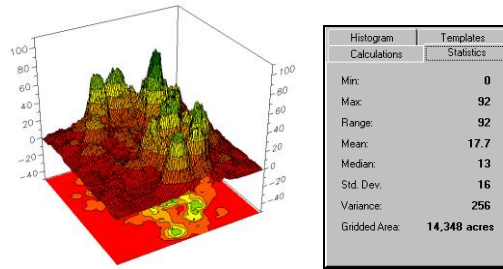
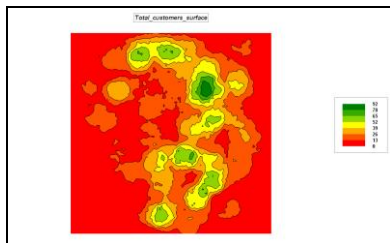
 Use the *View* tool to display the **Total_customers** map.



 Use **Neighbors** → **Scan** to create a density surface that totals the number of customers within a quarter-mile (Within 6 cells).




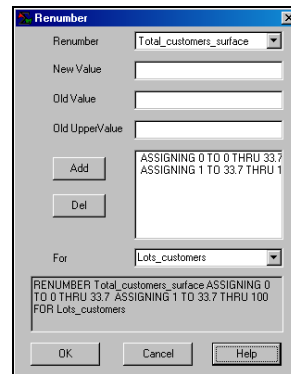
SCAN **Total_customers**
Total
 IGNORE PMAP_NULL
 WITHIN 6 CIRCLE
 FOR **Total_customers_surface**



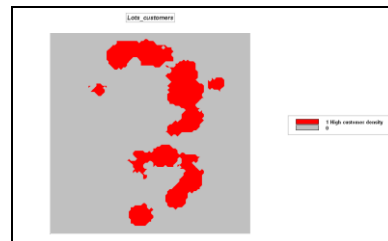
Note that customer density varies from 0 (no customers within 6 cells—1500 feet) to a high of 92 customers in the northeastern quadrant.

Right-click on the **Total_Customers_surface** map, select **Properties** then **Statistics** tab to identify the Mean and Standard Deviation. The break point for unusually high customer density is 17.7 plus 16 equals 33.7.

 Use **Reclassify** → **Renumber** to isolate the locations with unusually high customer density.



RENUMBER **Total_customers_surface**
 ASSIGNING 0 TO 0 THRU 33.7
 ASSIGNING 1 TO 33.7 THRU 100
 FOR **Lots_customers**



6.6.2 Identifying Customer Territories

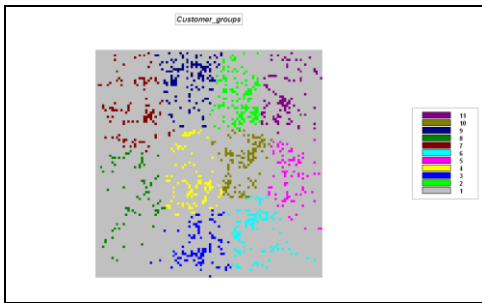
 Complete the following commands to create a map of customer groups—

Reclassify → *Renumber...*
RENUMBER Total_customers
ASSIGNING 1 TO 1 THRU 10
FOR Customer_mask

Overlay → *Compute...*
COMPUTE Customer_mask
Times Grid_latitude
FOR Customer_latitude

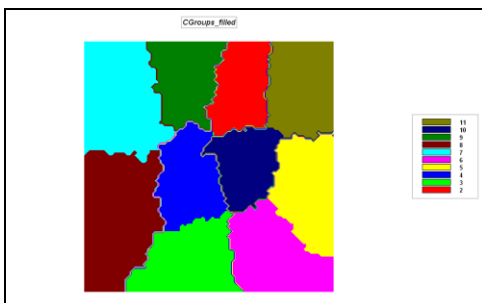
Overlay → *Compute...*
COMPUTE Customer_mask
Times Grid_longitude
FOR Customer_longitude

Statistics → *Cluster...*
CLUSTER Customer_latitude
WITH Customer_longitude
USING 11
FOR Customer_groups




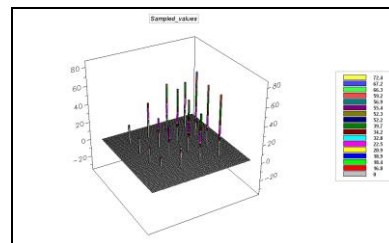
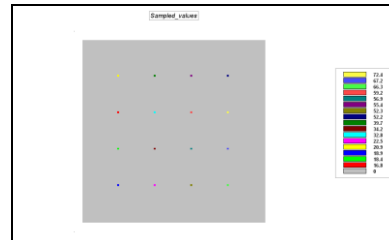
Note the distinct clusters of customers (color groupings).

Neighbors → *Interpolate...*
INTERPOLATE Customer_groups
WITHIN 20
USING 1
Discretely Maximum Retaining IGNORE 1
FOR CGroups_filled

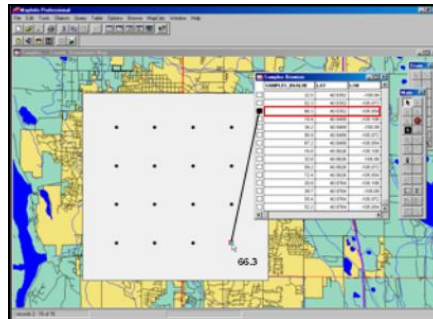



6.6.3 Creating Interpolated Surfaces

 Use the *View* tool to generate 2D and 3D displays of the the **Sampled_values** map.



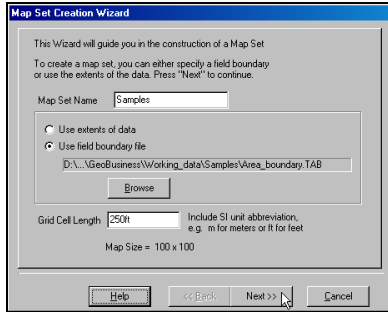
These data also are stored in MapInfo files identifying the project area boundary and the point data values as shown below.



 Press the *Create a New File* button and respond **No**, you do not “Want to save changes to GB_Smallville.rgs” the current set of maps.

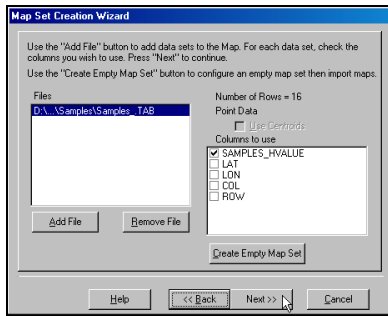
When the map *Set Creation Wizard* pops-up, specify the geographic input as—

- ✓ Map Set name of **Samples**
 - ✓ Browse to **Area Boundary.TAB**
 - ✓ Grid Cell Length = **250 ft.**
- ...then press **Next** (see below).

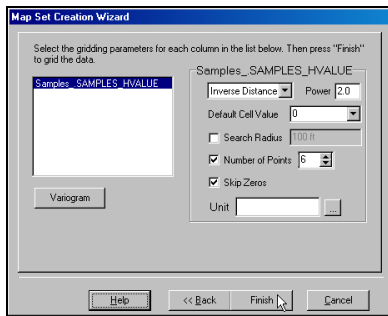


Specify the data input by—

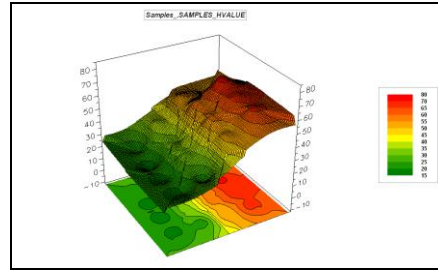
- ✓ Pressing the **Add File** button (respond **OK** to acknowledge that the data is in Lat/Lon WGS84) and select the **Samples.TAB** file.
 - ✓ Check the **SAMPLES_HVALUE** box as the data column to use.
- ...then press **Next** (see below).



In the interpolation method window, press **Next** to accept the IDW defaults (see below).



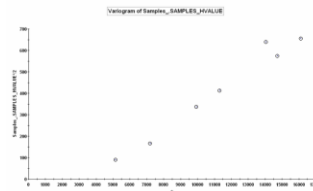
Use the **Shading Manager** to display the surface as a 3D lattice with **12 User Defined Ranges** from **15 to 80** (results in a contour interval of 5) and a color ramp from **green to red** with a mid-range color inflection of **yellow** (see below).



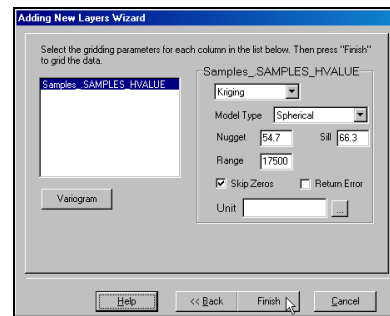
Use the **View** tool to rename the map from *Samples._SAMPLE_HVALUES* (automatically assigned name) to **Samples_IDW_default**.

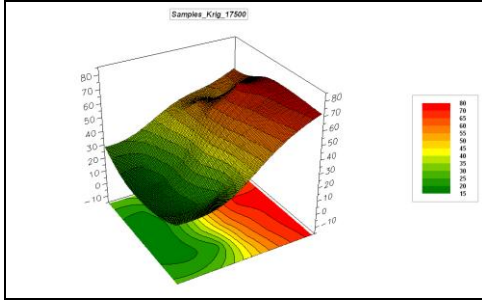
From the main menu select **Map Set → Add New Layers...** repeat the process except this time specify Kriging using the default parameters.


Select **Kriging** from the pull-down list. Set the **Range** to **17500**. Press the **Variogram** button to generate a variogram plot of the data.



The variogram plot has a consistent linear form so the data is appropriate for interpolation.



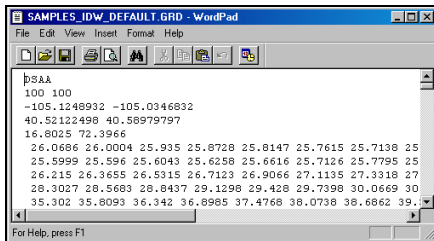



 Press the *Map Analysis* button and select **Import/Export** → **Export** then complete the following dialog box to generate a grid file (Surfer(Ascii)) of the **Samples_IDW_default** map.




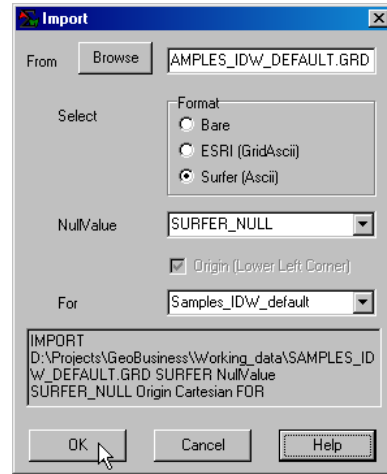
Repeat to export the **Samples_Krig_17500** map.


Use an editor (WordPad) or word processor to view the contents of the exported files. Note that the file contains standard Ascii text with the first five lines forming a descriptive header followed by a long listing of grid cell values.

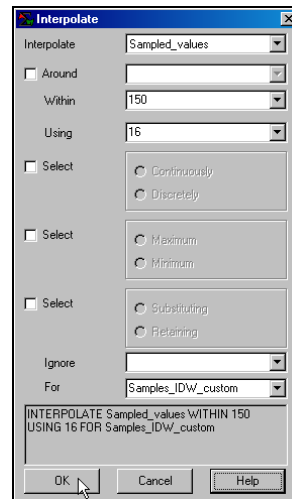



 Press the *Open an Existing File* button and respond **No**, you do not “Want to save changes to Samples.rgs” the current set of maps (unless you want to). Re-open the **GB_Smallville.rgs** data set.

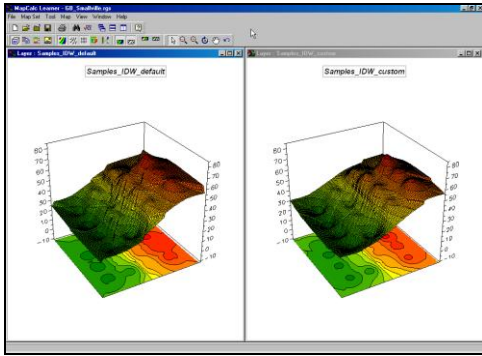
 Press the *Map Analysis* button and select **Import/Export** → **Import** then complete the following dialog box to import the maps you exported—**Samples_IDW_default.grd** and **Samples_Krig_17500.grd**. Ignore the error message about projection considerations.



 Press the *Map Analysis* button and select **Neighbors** → **Interpolate** then complete the following dialog box to generate a third surface using custom parameters for the internal IDW procedure—within 150 cell window (entire map extent) and using all 16 sample values in the roving window.



 Press the *Tile Vertically* button to display the surfaces side-by-side in 3D Lattice form.

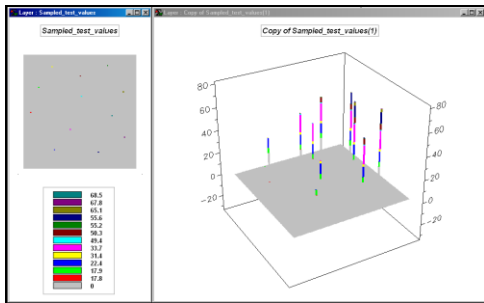


Note the additional smoothing in the IDW custom surface.

6.6.4 Assessing Interpolation Performance



Use the *View* tool to generate 2D and 3D displays of the the **Sampled_test_values** map.



Implement the following processing to determine the residuals for the three interpolated surfaces.

- 1) Use **Reclassify** → **ReNUMBER** to create a binary map of the test sample locations—
 RENUMBER Sampled_test_values
 ASSIGNING -9999 TO 0
 ASSIGNING 0 TO 1 THRU 100
 FOR Sampled_test_mask
- 2) Use **Overlay** → **Compute** to determine the difference between the Samples_IDW_default estimates and the **Sampled_test_values** map—
 COMPUTE Samples_IDW_default
 Minus Sampled_test_values
 FOR IDW_default_test_difference
- 3) Use **Overlay** → **Cover** to mask information for just the test points—

COVER IDW_default_test_difference
 WITH Sampled_test_mask IGNORE 0
 FOR IDW_default_test_residuals

4) Repeat steps 2 and 3 to determine the residuals for **IDW_custom_test_residuals** and **Krig_17500_test_residuals** maps.

Complete the following table to summarize the residuals—

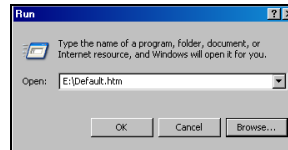
IDW_default residuals	IDW_custom residuals	Krig_17500 residuals
Avg=	Avg=	Avg=

Based on your results, which interpolated surface is best? _____







6.6.5 Using Surfer for Surface Modeling

Surfer is an advanced system for contouring, gridding, and surface mapping. It contains several additional spatial interpolation procedures beyond IDW and Kriging mentioned in these exercises. You are encouraged to complete the Surfer Tutorials.

A printer formatted version is available on the **Geo-Business CD** accompanying this book. Insert the CD and access the main menu by pressing **Start** → **Run** → **E:\Default.htm**.



<enter your CD drive>

Analyzing Geo-Business Data	
A Hands-on Case Study in Spatial Analysis and Data Mining	
<small>Note: Materials on this CD and companion hardcopy workbook are not for further distribution without prior written consent of Joseph K. Berry (jberr@uconn.edu)</small>	
Item	Description
   	<p>This item contains installation guidelines for installing—</p> <p>ArcExplorer GIS data viewer software for display and query of GIS data primarily used in exercises accompanying Topic 2, “Desktop Mapping.”</p> <p>Surfer demonstration software for contouring, gridding, and surface mapping primarily used in the exercises accompanying Topic 6, “Surface Modeling.”</p> <p>MapCalc Learner evaluation software for grid-based map analysis used in exercises throughout the book.</p> <p>SnagIt evaluation software for screen capture and creation of composite presentation graphics in exercises throughout the book.</p> <p>See Topic 1.3.1, “Installing Companion Software” for detailed instructions.</p>
<p>Example Applications</p> 	<p>This item accesses several annotated descriptions and example applications of grid-based map analysis. See Topic 1.3.1, Example of Analysis Capabilities for more information.</p> <p>Click on any of the links in the listing to access the examples. All of the example applications were developed using MapCalc Learner and can be replicated using the tutorial database identified with each example.</p>
<p>Text Figures</p> 	<p>This item contains a full color PowerPoint slide set of all of the workbook figures. Permission to copy is granted. Reference as “Figure x-xx from Analyzing Precision Ag Data, J.K. Berry, 2002.”</p>

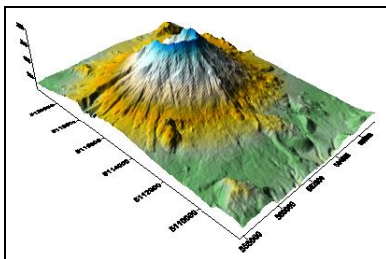
From the “Installing Surfer...” section, click on the link for the tutorial exercises. Print a copy of the exercises.

Installing Surfer Demonstration Software <small>(Surfer Version 8)</small>
<p>Close all open programs, click here then follow the on-screen instructions. The program, database and support materials require approximately 20MB free disk space to install. The demonstration program will not expire but some of the functionality (e.g., plotting) is not supported. See http://www.goldensoftware.com/ for information on the full, single-seat version (\$599.00). (February 2003)</p> <p>Click here for a general set of tutorial exercises that accompanies the Surfer demonstration software (.pdf format).</p>

Install the Surfer system if you have not already—see Topic 1.4.3. To access the system click on **Start** → **Programs** → **Golden Software Surfer 8** → **Surfer 8**.

From Surfer’s main menu, select **Map** → **Surface** and specify ...**Samples\Helens2.grd** as the grid layer to open. Press the **Open** button to plot the data.

If Surfer has been properly installed a three-dimensional plot of Mt. St. Helens will appear (see below).



You are now ready to complete the Surfer Tutorials.