

Spatial Data Mining in Geo-Business

Joseph K. Berry¹

*W. M. Keck Visiting Scholar in Geosciences, Geography, University of Denver
Principal, Berry & Associates // Spatial Information Systems (BASIS), Fort Collins, Colorado
Email: jberry@innovativegis.com; Website: www.innovativegis.com*

Kenneth L. Reed²

Xtreme Data Mining, Costa Mesa, California; geatumspraec@yahoo.com

This paper was presented at GeoTec, June 2-5, 2008, Ottawa, Ontario, Canada
Available online at <http://www.innovativegis.com/basis/present/GeoTec08/> — click for a printer-friendly [.pdf version](#)

Abstract

Most traditional Geo-business applications force spatial information, such as customer location, to be aggregated into large generalized reporting units. More recently targeted marketing, retail trade area analysis, competition analysis and predictive modeling provide examples applying sophisticated spatial analysis and statistics to improve decision making and ensure sound business decisions. This paper describes a spatial data mining process for generating predicted sales maps for various products within a large metropolitan area. The discussion identifies the processing steps involved in compiling a spatially-aware customer database and then applying CART technology for analyzing relative travel-time advantage coupled with existing customer data to derive information on travel-time sensitivity and sales patterns. The conceptual keystone of this application is the concept of a grid-based analytic frame and its continuous map surfaces that underwrite the spatially aware database. The column, row index of the grid cells in the matrix is appended to each record the database and serves as a primary key for cross-walking information between the GIS and the customer database. As business interests and GIS specialists become more aware of the informational value of digital maps and the unique characteristics of business applications, the rapidly developing field of Geo-business will alter our paradigms of maps, mapped data and their use in understanding and predicting the business environment.

Twisting the Perspective of Map Surfaces

Traditionally one thinks of a map surface in terms of a postcard scene of the Rocky Mountains with peaks and valleys recording uplift and erosion over thousands of years—a down to earth view of a map surface one can stand on.

However, a geomorphological point of view of a digital elevation model (DEM) isn't the only type of map surface. For example, a *Customer Density Surface* can be derived from sales data that depicts the peaks and valleys of customer concentrations throughout a city as discussed in an earlier Beyond Mapping column (November 2005; see author's note). Figure 1 summarizes the processing steps involved—1) a customer's street address is geocoded to identify its Lat/Lon coordinates, 2) vector to raster conversion is used to place and aggregate the number of customers in each grid cell of an analysis frame (*discrete mapped data*), 3) a rowing window is used to count the total number of customers within a specified radius of each cell (*continuous mapped data*), and then 4) classified into logical ranges of customer density.

The important thing to note is that the peaks and valleys characterize the *Spatial Distribution* of the customers, a concept closely akin to a *Numerical Distribution* that serves as the foundation for traditional statistics. However in this instance, three dimensions are needed to characterize the data's dispersion—X and Y coordinates to position the data in geographic space and a Z coordinate to indicate the relative magnitude of the variable (# of customers). Traditional statistics needs only two dimensions—X to identify the magnitude and Y to identify the number of occurrences.

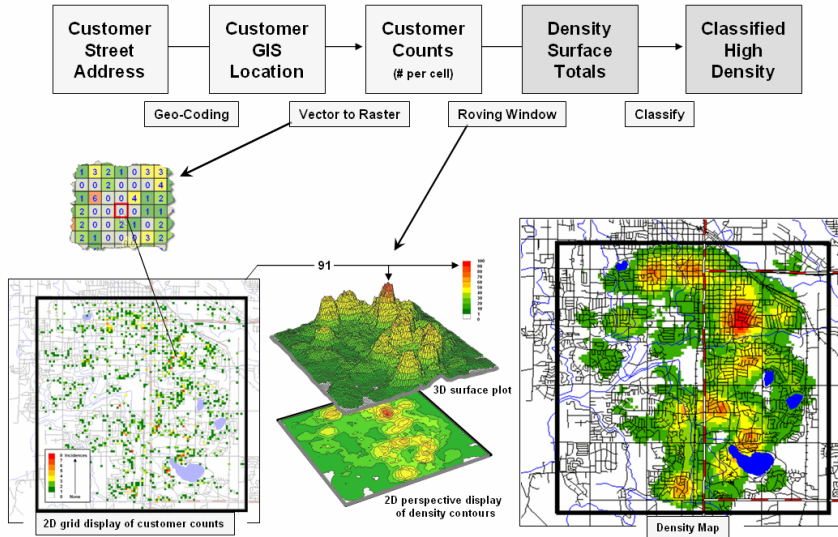


Figure 1. Geocoding can be used to connect customer addresses to their geographic locations for subsequent map analysis, such as generating a map surface of customer density.

While both perspectives track the relative frequency of occurrence of the values within a data set, the spatial distribution extends the information to variations in geographic space, as well as in numerical variations in magnitude— from just “what” to “where is what.” In this case, it describes the geographic pattern of customer density as peaks (lots of customers nearby) and valleys (not many).

Within our historical perspective of mapping the ability to plot “where is what” is an end in itself. Like Inspector Columbo’s crime scene pins poked on a map, the mere visualization of the pattern is thought to be sufficient for solving crimes. However, the volume of sales transactions and their subtle relationships are far too complex for a visual (visceral?) solution using just a Google Earth mashed-up image.

The interaction of numerical and spatial distributions provides fertile turf for a better understanding of the mounds of data we inherently collect every day. Each credit card swipe identifies a basket of goods and services purchased by a customer that can place on a map for grid-based map analysis to make sense of it all—“why” and “so what.”

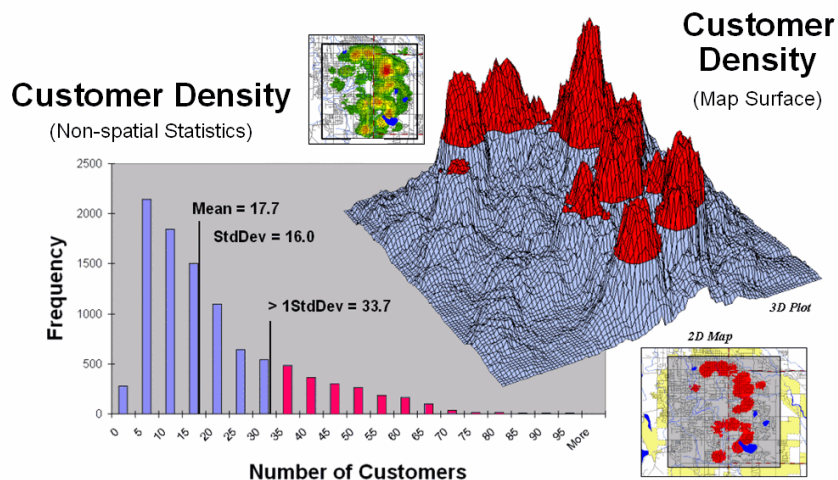


Figure 2. Merging traditional statistics and map analysis techniques increases understanding of spatial patterns and relationships (variance-focused; continuous) beyond the usual central tendency (mean-focused; scalar) characterization of geo-business data.

For example, consider the left side of figure 2 that relates the unusually high response range of customers of greater than 1 standard deviation above the mean (a numerical distribution perspective) to the right side that identifies the location of these pockets of high customer density (a spatial distribution perspective). As discussed in a previous Beyond Mapping column (May 2002; see author's note), this simple analysis uses a typical transaction database to 1) map the customer locations, 2) derive a map surface of customer density, 3) identify a statistic for determining unusually high customer concentrations ($> \text{Mean} + 1\text{SD}$), and 4) apply the statistic to locate the areas with lots of customers. The result is a map view of a commonly used technique in traditional statistics—an outcome where the combined result of an integrated spatial and numerical analysis is far greater than their individual contributions.

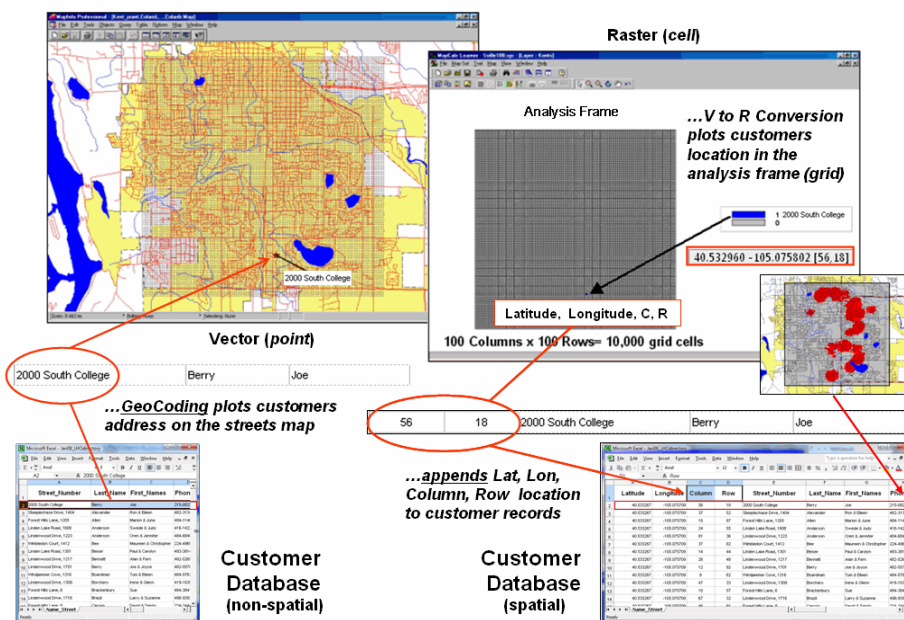


Figure 3. The column, row index of the grid cells in the analytic frame serves as a primary key for linking spatial analysis results with individual records in a database.

However, another step is needed to complete the process and fully illustrate the geo-business framework. The left side of figure 3 depicts the translation of a non-spatial customer database into a mapped representation using vector-based processing that determines Lat/Lon coordinates for each record—the mapping component. Translation from vector to raster data structure establishes the analytic frame needed for map analysis, such as generating a customer density surface and classifying it into statistically defined pockets of high customer density—then map analysis component.

The importance of the analytic frame is paramount in the process as it provides the link back to the original customer database—the spatial data mining component. The column, row index of the matrix for each customer record is appended to the database and serves as a primary key for “walking” between the GIS and the database. For example, the areas of unusually high customer density can be appended to the customer database using the column, row index in both data sets, and then used to identify individual customers within these areas.

Similarly, maps of demographics, sales by product, travel-time to our store, and the like can be used in customer segmentation and propensity modeling to identify maps of future sales probabilities. Areas of high probability can be cross-walked to an existing customer database (or zip +4 or other generic databases) to identify new sales leads, product mix, stocking levels, inventory management and competition analysis. At the core of this vast potential for geo-business applications are the analytic frame and its continuous map surfaces that underwrite a spatially aware database.

Linking Numeric and Geographic Distributions

Another grid-based technique for investigating customer patterns involves **Point Territories** assignment. This procedure looks for inherent spatial groups in the point data and assigns customers to contiguous areas. In the example in figure 4, you might want to divide the customer locations into ten groups for door-to-door contact on separate days.

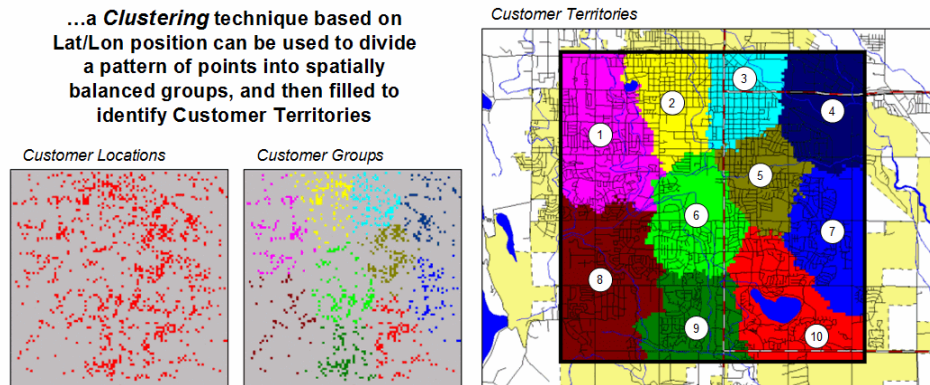


Figure 4. Clustering on the latitude and longitude coordinates of point locations can be used to identify geographically balanced customer territories.

The two small inserts on the left show the general pattern of customers, and 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 (approximately the same area in each cluster), however it doesn't consider the number of customers within each group—that varies from 69 in the lower right (Territory #8) to 252 (Territory #5) near the upper right... that twist of map analysis will be tackled in a future beyond mapping column.

However it does bring up an opportunity to discuss the close relationship between spatial and non-spatial statistics. 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* 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.

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 5 that was derived by counting the total number of customers within a specified radius of each cell (roving window). 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.

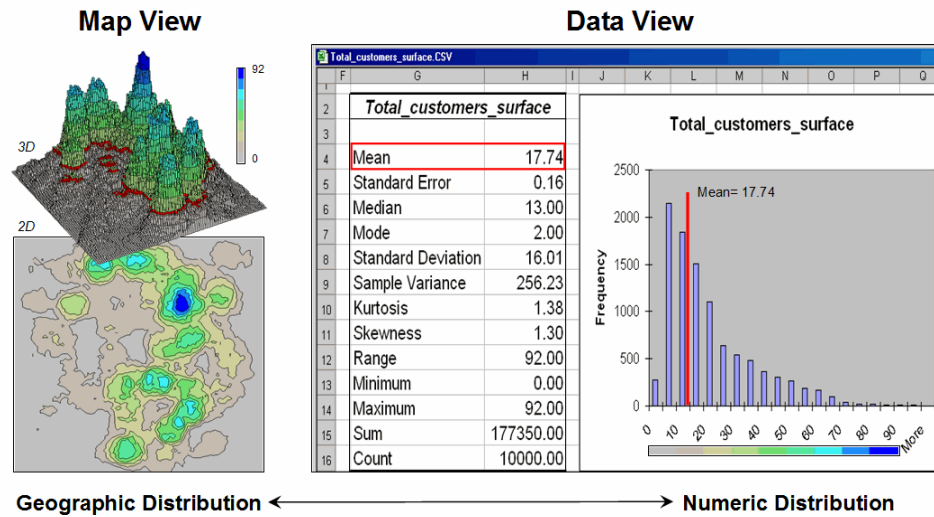


Figure 5. Mapped data are characterized by their geographic distribution (maps on the left) and their numeric distribution (descriptive statistics and histogram on the right).

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 map organizes the data as 100 rows by 100 columns (X,Y) while the non-spatial view simply summarizes the 10,000 values into a set of statistical indices characterizing the overall central tendency of the data. The mechanics used to plot the histogram and generate the statistics are 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 low customer density 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 numeric). 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.

Spatial statistics, the foundation of geo-business applications, doesn't predispose any geographic or numeric functional forms—it simply responds to the inherent patterns and relationships in a data set. The next several sections will describe some of the surface modeling and spatial data mining techniques available to the venturesome few who are willing to work “outside the lines” of traditional mapping and statistics.

Interpolating Spatial Distributions

Statistical sampling has long been at the core of business research and practice. Traditionally data analysis used non-spatial statistics to identify the “typical” level of sales, housing prices, customer income, etc. throughout an entire neighborhood, city 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.*” But no information is provided as to where values might be more or less than the computed typical value (average).

Spatial Interpolation, on the other hand, utilizes 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).

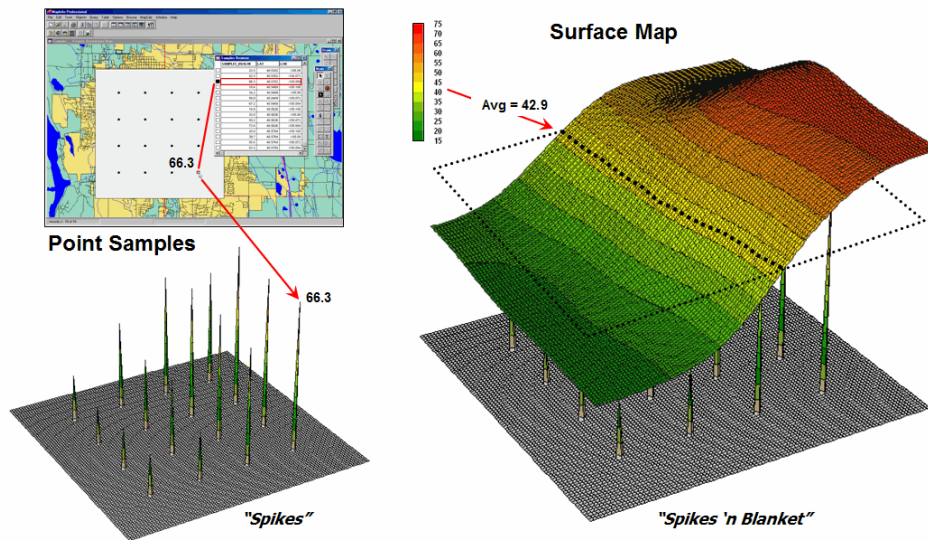


Figure 6. Spatial interpolation involves fitting a continuous surface to sample points.

While the extension from non-spatial to spatial statistics is a theoretical leap, the practical steps are relatively easy. The left side of figure 1 shows 2D and 3D “point maps” of data samples depicting the percentage of home equity loan to market value. Note that the samples are *geo-referenced* and that the *sampling pattern* and *intensity* are different than those generally used in traditional non-spatial statistics and tend to be more regularly spaced and numerous.

The surface map on the right side of figure 6 translates pattern of the “spikes” into the peaks and valleys of the surface map representing the data’s spatial distribution. The traditional, non-spatial approach when mapped is 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 process of converting point-sampled data into continuous map surfaces representing a spatial distribution is termed *Surface Modeling* involving density analysis and map generalization (discussed

last month), as well as spatial interpolation techniques. All spatial interpolation techniques establish a "roving window" that—

- moves to a grid location in a project area (analysis frame),
- calculates an estimate based on the point samples around it (roving window),
- assigns the estimate to the center cell of the window, and then
- moves to the next grid 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 larger number of values tends to "smooth" the data. A smaller window tends to result in a "rougher" surface with more 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 is more than the specified number of values within a specified reach, the computer uses just the closest ones. If there are not enough values, it uses all that it can find within the reach. *Balancing* of the data attempts to eliminate directional bias by ensuring that the values are selected in all directions around window's center.

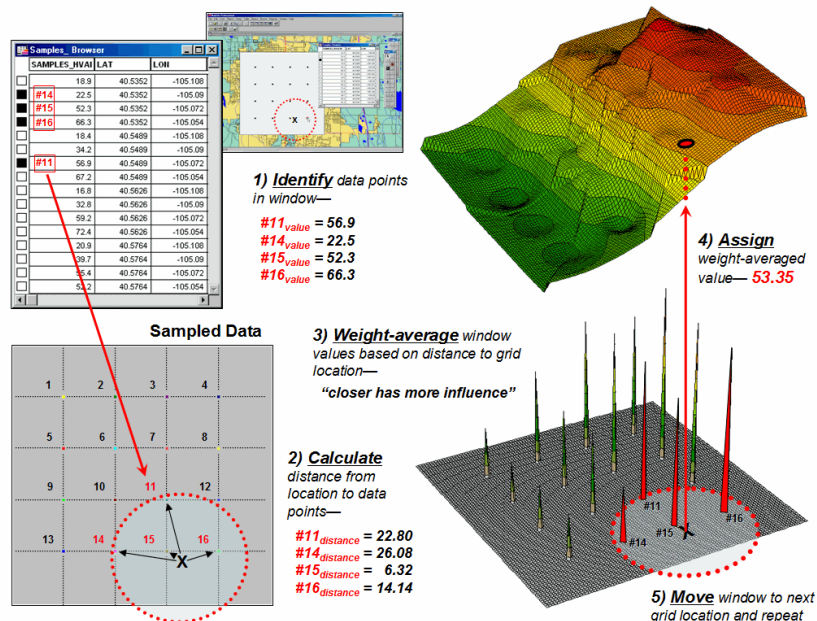


Figure 7. Inverse distance weighted interpolation weight-averages sample values within a roving window.

The right portion of figure 7 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," 53.35 units above the base (zero).

Figure 8 shows the weighted average calculations for spatially interpolating the example location in figure 7. The Pythagorean Theorem is used to calculate the *Distance* from the Grid Location to each of the data *Samples* within the summary window. The distances are converted to *Weights* that are inversely proportional ($1/D^2$; see example calculation in the figure). The sample *Values* are multiplied by their computed *Weights* and the "sum of the products" is divided by the "sum of the weights" to calculate the weighted average value (53.35) for the location on the interpolated surface.

Sample	Row	Column	Value	Distance $\text{SQRT}[A^{**2} + B^{**2}]$	Weight $(1/D^{**2})$	Value times Weight
#11	41	60	56.80	22.80	0.0019	0.1094
#14	21	40	22.50	26.08	0.0015	0.0331
#15	21	60	52.30	6.32	0.0250	1.3075
#16	21	80	66.30	14.14	0.0050	0.3315
				Sum =	0.0334	1.7815
"X" Grid Location	19	66		For example... $\text{SQRT} [(41 - 19)^{**2} + (60 - 66)^{**2}]$ = 22.80	Weighted Average = $(1.7815 / 0.0334)$	53.35

Figure 8. Example Calculations for Inverse Distance Squared Interpolation.

Because the inverse distance procedure is a fixed, geometric-based method, the estimated values can never exceed the range of values in the original field data. Also, IDW tends to "pull-down peaks and pull-up valleys" in the data, as well as generate "bull's-eyes" around sampled locations. The technique is best suited for data sets with samples that exhibit minimal regional trends.

However, there are numerous other spatial interpolation techniques that map the spatial distribution inherent in a data set. Next month's column will focus on benchmarking interpolation results from different techniques and describe a procedure for assessing which is best.

Interpreting Interpolation Results

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

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 month's column discusses the relative amounts of information provided by the non-spatial arithmetic average versus site-specific maps by comparing the average and two different interpolated map surfaces. The discussion is further extended to describe a procedure for quantitatively assessing interpolation performance.

The top-left inset in figure 9 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—it assumes 42.88 percent is everywhere. The non-spatial estimate simply adds up all of the sample values and divides by the number of samples to get the average disregarding any geographic pattern.

The spatially-based estimates comprise the map surface just below the pancake. As described last month, *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 9 compares the two maps by simply subtracting them. The colors were chosen to emphasize the differences between the whole-field average estimates and the interpolated ones. The thin yellow band indicates 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 condition with the average estimate being larger than the interpolated ones.

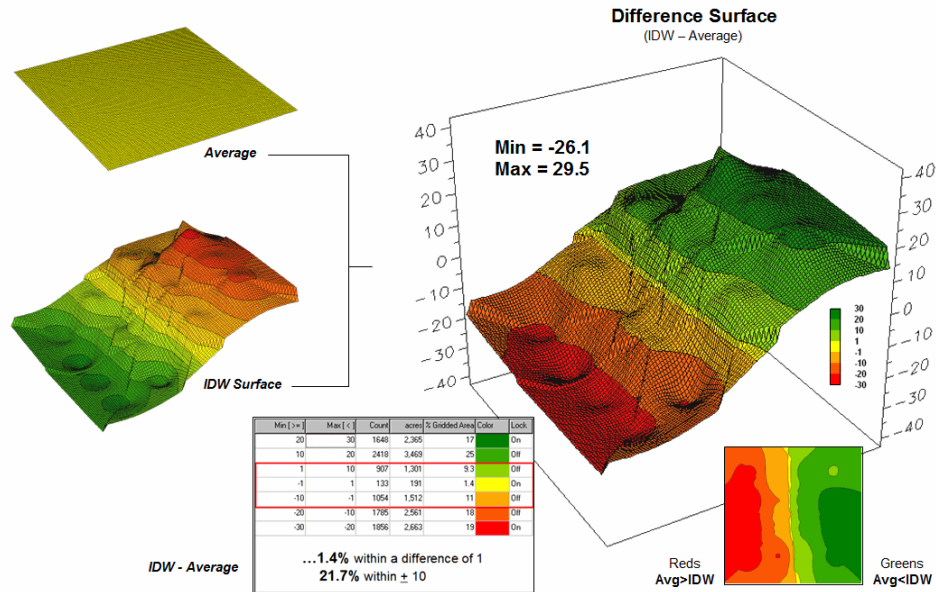


Figure 9. Spatial comparison of the project area average and the IDW interpolated surface.

The difference between the two maps ranges from -26.1 to +29.5. If one assumes that a difference of +/- 10 would not significantly alter a decision, then about one-quarter of the area (9.3+1.4+11 = 21.7%) is adequately represented by the overall average of the sample data. But that leaves about three-fourths of the area that is either well-below the average (18 + 19 = 37%) or well-above (25+17 = 42%). The upshot is that using the average value in either of these areas could lead to poor decisions.

Now turn your attention to figure 10 that compares maps derived by two different interpolation techniques—IDW (inverse distance weighted) and Kriging (an advanced spatial statistics technique using data trends). Note the similarity in the two surfaces; while subtle differences are visible, the overall trend of the spatial distribution is similar.

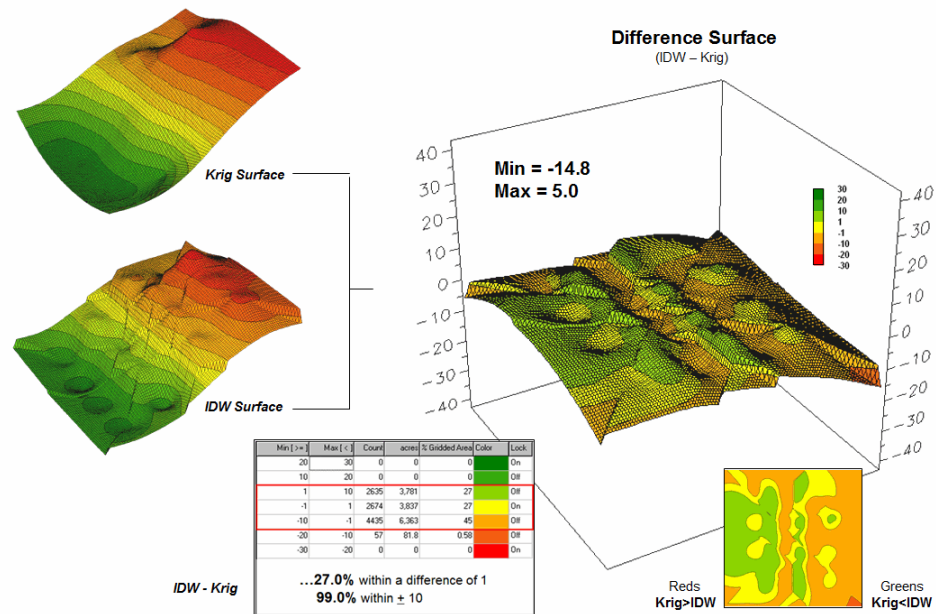


Figure 10. Spatial comparison of IDW and Krig interpolated surfaces.

The difference map on the right confirms the similarity between the two map surfaces. 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 (within +10). Applying the same assumption about “plus/minus 10 difference” being negligible for decision-making, the maps are effectively the same (99.0%).

So what’s the bottom line? First, that there are substantial differences between an arithmetic average and interpolated surfaces. Secondly, that quibbling about the best interpolation technique isn’t as important as using any interpolated surface for decision-making.

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.

The table in figure 11 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 the assumption that the project area average (42.88) represents 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).

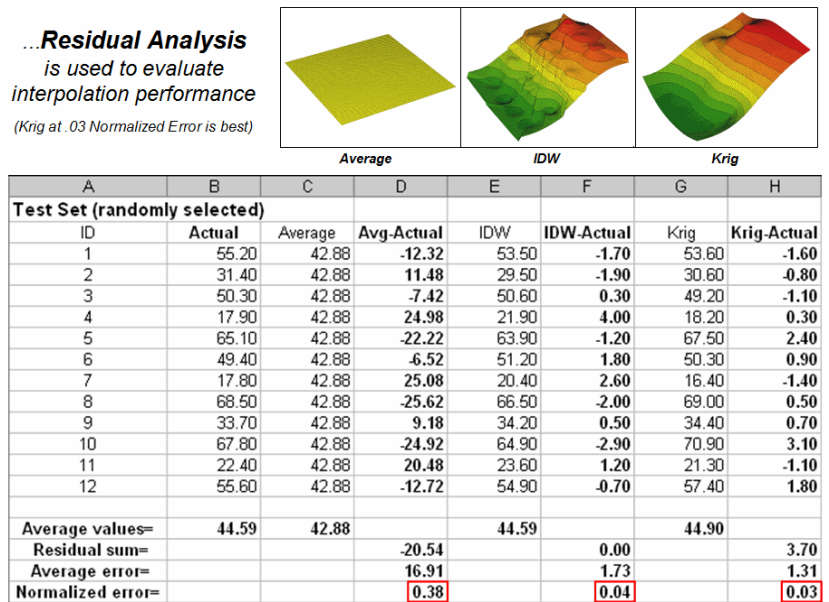


Figure 11. A residual analysis table identifies the relative performance of average, IDW and Krig estimates.

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 considerably 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 the IDW estimate is dramatically larger.

The rows at the bottom of the table summarize the residual analysis results. The *Residual Sum* 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* 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 1) the project area average is inappropriate and that 2) the accuracy differences between IDW and Krig are very minor.

The *Normalized Error* simply calculates the average error as a proportion of the average value for the test set of samples ($1.73/44.59 = 0.04$ 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 use them for important decisions.

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 cannot 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 business decisions. Without it, the process becomes one of blind faith and wishful thinking with colorful maps.

Characterizing Data Groups

One of the most fundamental techniques in map analysis is the comparison of a set of maps. This usually involves staring at some side-by-side map displays and formulating an impression about how the colorful patterns do and don't appear to align.

But just how similar is one location to another? Really similar, or just a little bit similar? And just how dissimilar are all of the other areas? While visual (visceral?) analysis can identify broad relationships, it takes quantitative map analysis to generate the detailed scrutiny demanded by most Geo-business applications.

Consider the three maps shown in figure 12— what areas identify similar data patterns? If you focus your attention on a location in the southeastern portion, how similar are all of the other locations? Or how about a northeastern section? The answers to these questions are far too complex for visual analysis and certainly beyond the geo-query and display procedures of standard desktop mapping packages.

The mapped data in the example show the geographic patterns of housing density, value and age for a project area. In visual analysis you move your focus among the maps to summarize the color assignments (2D) or relative surface height (3D) at different locations. In the southeastern portion the general pattern appears to be low Density, high Value and low Age— low, high, low. The northeastern portion appears just the opposite—high, low, high.

The difficulty in visual analysis is two-fold— remembering the color patterns and calculating the difference. Quantitative map analysis does the same thing except it uses the actual map values in place of discrete color bands. In addition, the computer doesn't tire as easily as you and completes the comparison throughout an entire map window in a second or two (10,000 grid cells in this example).

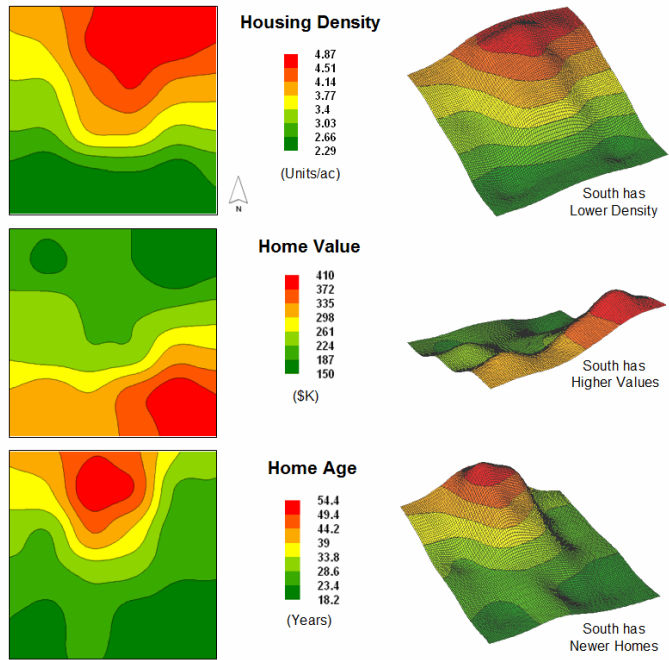


Figure 12. Map surfaces identifying the spatial distribution of housing density, value and age.

The upper-left portion of figure 13 illustrates capturing the data patterns for comparing two map locations. The “data spear” at Point #1 identifies the housing Density as 2.4 units/ac, Value as \$407,000 and Age as 18.3 years. This step is analogous to your eye noting a color pattern of green, red, and green. The other speared location at Point #2 locates the least similar data pattern with housing Density of 4.8 units/ac, Value of \$190,000 and Age of 51.2 years— or as your eye sees it, a color pattern of red, green, red.

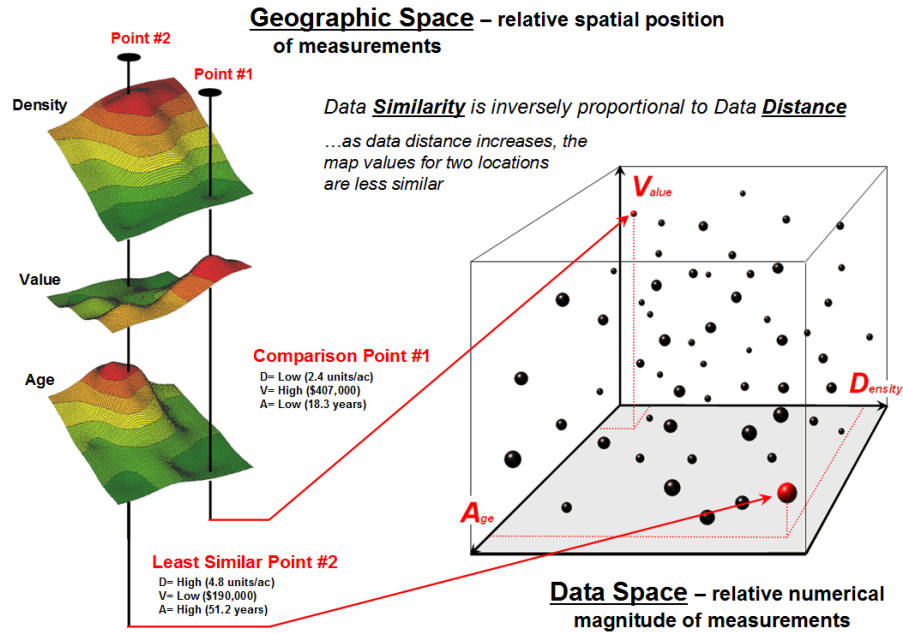


Figure 13. Conceptually linking geographic space and data space.

The right side of the figure schematically depicts how the computer determines similarity in the data patterns by analyzing them in three-dimensional “data space.” Similar data patterns plot close to one another with increasing distance indicating decreasing similarity. The realization that mapped data can be expressed in both geographic space and data space is paramount to understanding how a computer quantitatively analyses numerical relationships among mapped data.

The right side of the figure schematically depicts how the computer determines similarity in the data patterns by analyzing them in three-dimensional “data space.” Similar data patterns plot close to one another with increasing distance indicating decreasing similarity. The realization that mapped data can be expressed in both geographic space and data space is paramount to understanding how a computer quantitatively analyses numerical relationships among mapped data.

Geographic space uses earth coordinates, such as latitude and longitude, to locate things in the real world. The geographic expression of the complete set of measurements depicts their spatial distribution in familiar map form. **Data space**, on the other hand, is a bit less familiar. While you can’t stroll through data space you can conceptualize it as a box with a bunch of balls floating within it.

In the example, the three axes defining the extent of the box correspond to housing Density (D), Value (V) and Age (A). The floating balls represent data patterns of the grid cells defining the geographic space—one “floating ball” (data point) for each grid cell. The data values locating the balls extend from the data axes—2.4, 407.0 and 18.3 for the *comparison point* identified in figure 2. The other point has considerably higher values in D and A with a much lower V values so it plots at a different location in data space (4.8, 51.2 and 190.0 respectively).

The bottom line for data space analysis is that the position of a point in data space identifies its numerical pattern—low, low, low in the back-left corner, and high, high, high in the upper-right corner of the box. Points that plot in data space close to each other are similar; those that plot farther away are less similar. Data distance is the way computers “see” what you see in the map displays. The real difference in the graphical and quantitative approaches is in the details—the tireless computer “sees” subtle differences between all of the data points and can generate a detailed map of similarity.

In the example in figure 13, the floating ball closest to you is least similar—greatest “data distance” from the comparison point. This distance becomes the reference for “most different” and sets the bottom value of the similarity scale (0% similar). A point with an identical data pattern plots at exactly the same position in data space resulting in a data distance of 0 equating to the highest similarity value (100% similar).

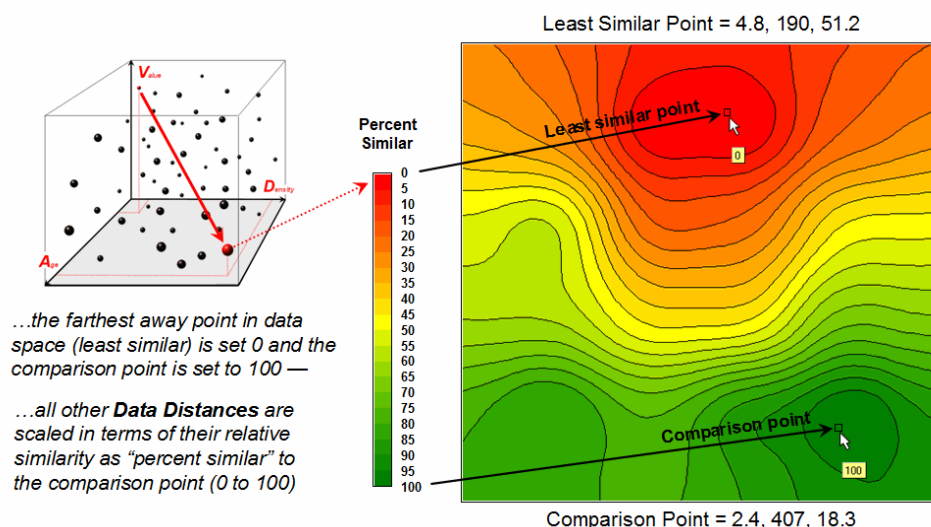


Figure 14. A similarity map identifies how related locations are to a given point.

The similarity map shown in figure 14 applies a consistent scale to the data distances calculated between the comparison point and all of the other points. The green tones indicate locations having fairly similar D, V and A levels to the comparison location—with the darkest green identifying locations with an identical data pattern (100% similar). It is interesting to note that most of the very similar locations are in the southern portion of the project area. The light-green to red tones indicate increasingly dissimilar areas occurring in the northern portion of the project area.

A similarity map can be an extremely valuable tool for investigating spatial patterns in a complex set of mapped data. The similarity calculations can handle any number of input maps, yet humans are unable to even conceptualize more than three variables (data space box). Also, the different map layers can be weighted to reflect relative importance in determining overall similarity. For example, housing Value could be specified as ten times more important in assessing similarity. The result would be a different map than the one shown in figure 14— how different depends on the unique coincidence and weighting of the data patterns themselves.

In effect, a similarity map replaces a lot of general impressions and subjective suggestions for comparing maps with an objective similarity measure assigned to each map location. The technique moves map analysis well beyond traditional visual/visceral map interpretation by transforming digital map values into to a quantitative/consistent index of percent similarity. Just click on a location and up pops a map that shows how similar every other location is to the data pattern at the comparison point— an unbiased appraisal of similarity.

Identifying Data Zones

The previous section introduced the concept of *Data Distance* as a means to measure data pattern similarity within a stack of map layers. One simply mouse-clicks on a location, and all of the other locations are assigned a similarity value from 0 (zero percent similar) to 100 (identical) based on a set of specified map layers. The statistic replaces difficult visual interpretation of a series of side-by-side map displays with an exact quantitative measure of similarity at each location.

An extension to the technique allows you to circle an area then compute similarity based on the typical data pattern within the delineated area. In this instance, the computer calculates the average value within the area for each map layer to establish the comparison data pattern, and then determines the normalized data distance for each map location. The result is a map showing how similar things are throughout a project area to the area of interest.

The link between *Geographic Space* and *Data Space* is the keystone concept. As shown in figure 15, spatial data can be viewed as either a map, or a histogram. While a map shows us “*where is what*,” a histogram summarizes “*how often*” data values occur (regardless where they occur). The top-left portion of the figure shows a 2D/3D map display of the relative housing density within a project area. Note that the areas of high housing Density along the northern edge generally coincide with low home Values.

The histogram in the center of the figure depicts a different perspective of the data. Rather than positioning the measurements in geographic space it summarizes the relative frequency of their occurrence in data space. The X-axis of the graph corresponds to the Z-axis of the map—relative level of housing Density. In this case, the spikes in the graph indicate measurements that occur more frequently. Note the relatively high occurrence of density values around 2.6 and 4.7 units per acre. The left portion of the figure identifies the data range that is unusually high (more than one standard deviation above the mean; $3.56 + .80 = 4.36$ or greater) and mapped onto the surface as the peak in the NE corner. The lower sequence of graphics in the figure depicts the histogram and map that identify and locate areas of unusually low home values.

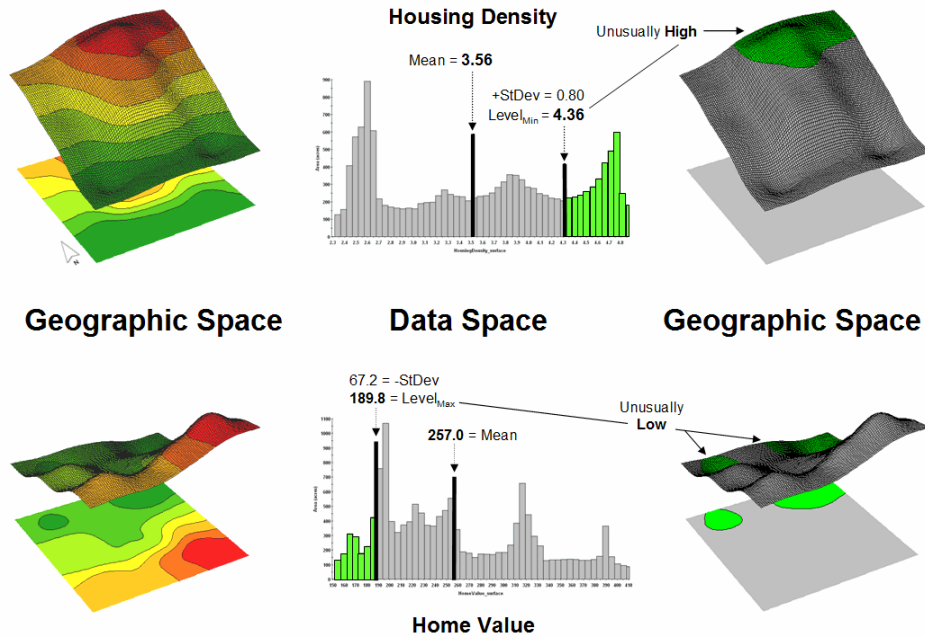


Figure 15. Identifying areas of unusually high measurements.

Figure 16 illustrates combining the housing Density and Value data to locate areas that have high measurements in both. The graphic in the center is termed a *Scatter Plot* that depicts the joint occurrence of both sets of mapped data. Each ball in the scatter plot schematically represents a location in the field. Its position in the scatter plot identifies the housing Density and home Value measurements for one of the map locations—10,000 in all for the actual example data set. The balls shown in the light green shaded areas of the plot identify locations that have high Density or low Value; the bright green area in the upper right corner of the plot identifies locations that have high Density and low Value.

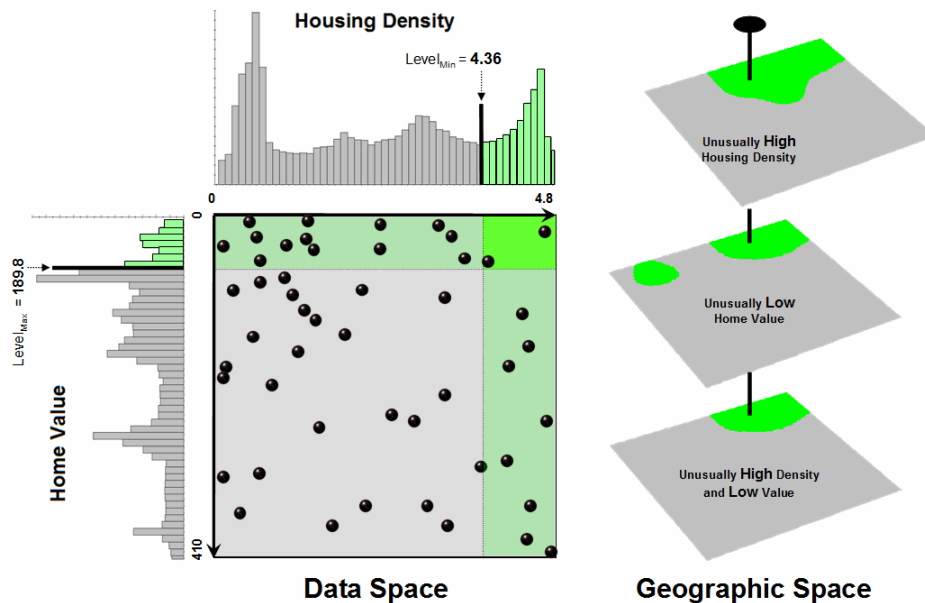


Figure 16. Identifying joint coincidence in both data and geographic space.

The aligned maps on the right side of figure 16 show the geographic solution for the high D and low V areas. A simple map-ematical way to generate the solution is to assign 1 to all locations of high Density and Value map layers (green). Zero (grey) is assigned to locations that fail to meet the conditions. When the two binary maps (0 and 1) are multiplied, a zero on either map computes to zero. Locations that meet the conditions on both maps equate to one ($1 \times 1 = 1$). In effect, this “level-slice” technique locates any data pattern you specify—just assign 1 to the data interval of interest for each map variable in the stack, and then multiply.

Figure 17 depicts level slicing for areas that are unusually low housing Density, high Value and low Age. In this instance the data pattern coincidence is a box in 3-dimensional scatter plot space (upper-right corner toward the back). However a slightly different map-ematical trick was employed to get the detailed map solution shown in the figure.

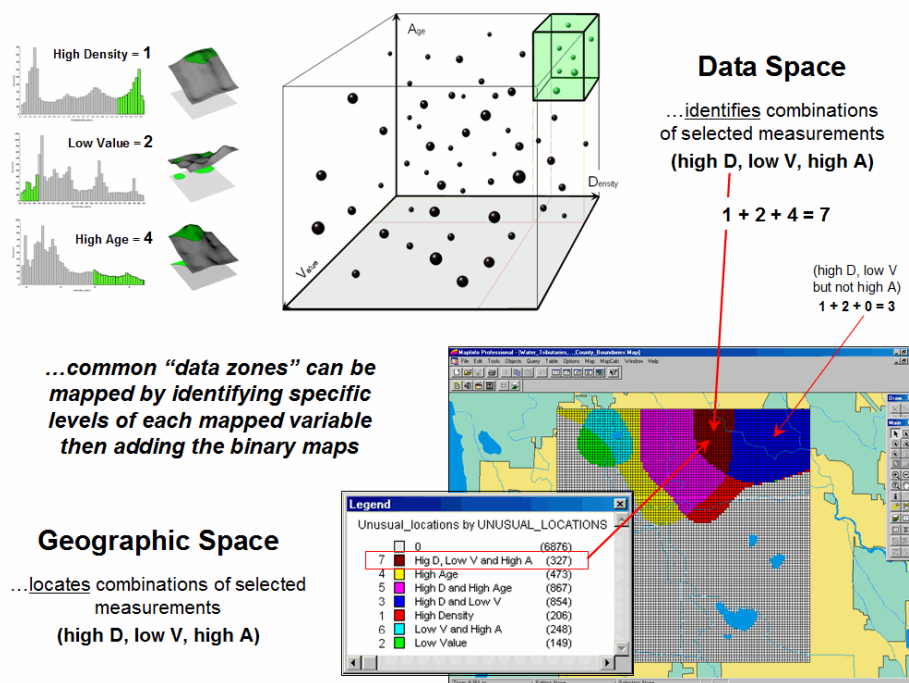


Figure 17. Level-slice classification using three map variables.

On the individual maps, areas of high Density were set to $D=1$, low Value to $V=2$ and high Age to $A=4$, then the binary map layers were added together. The result is a range of coincidence values from zero ($0+0+0=0$; gray= no coincidence) to seven ($1+2+4=7$; dark red for location meeting all three criteria). The map values in between identify the areas meeting other combinations of the conditions. For example, the dark blue area contains the value 3 indicating high D and low V but not high A ($1+2+0=3$) that represents about three percent of the project area ($327/10000=3.27\%$). If four or more map layers are combined, the areas of interest are assigned increasing binary progression values (...8, 16, 32, etc)—the sum will always uniquely identify all possible combinations of the conditions specified.

While level-slicing isn't a very sophisticated classifier, it illustrates the usefulness of the link between Data Space and Geographic Space to identify and then map unique combinations of conditions in a set of mapped data. This fundamental concept forms the basis for more advanced geo-statistical analysis—including map clustering that will be the focus of next section.

Mapping Data Clusters

The last couple of sections have focused on analyzing data similarities within a stack of maps. The first technique, termed *Map Similarity*, generates a map showing how similar all other areas are to a selected location. A user simply clicks on an area and all of the other map locations are assigned a value from zero (0% similar—as different as you can get) to one hundred (100% similar—exactly the same data pattern).

The other technique, *Level Slicing*, enables a user to specify a data range of interest for each map layer in the stack then generate a map identifying the locations meeting the criteria. Level Slice output identifies combinations of the criteria met—from only one criterion (and which one it is), to those locations where all of the criteria are met.

While both of these techniques are useful in examining spatial relationships, they require the user to specify data analysis parameters. But what if you don't know which locations in a project area warrant Map Similarity investigation or what Level Slice intervals to use? Can the computer on its own identify groups of similar data? How would such a classification work? How well would it work?

Figure 18 shows some example spatial patterns derived from *Map Clustering*. The “floating” map layers on the left show the input map stack used for the cluster analysis. The maps are the same ones used in previous examples and identify the geographic and numeric distributions of housing Density, home Value and home Age levels throughout the example project area.

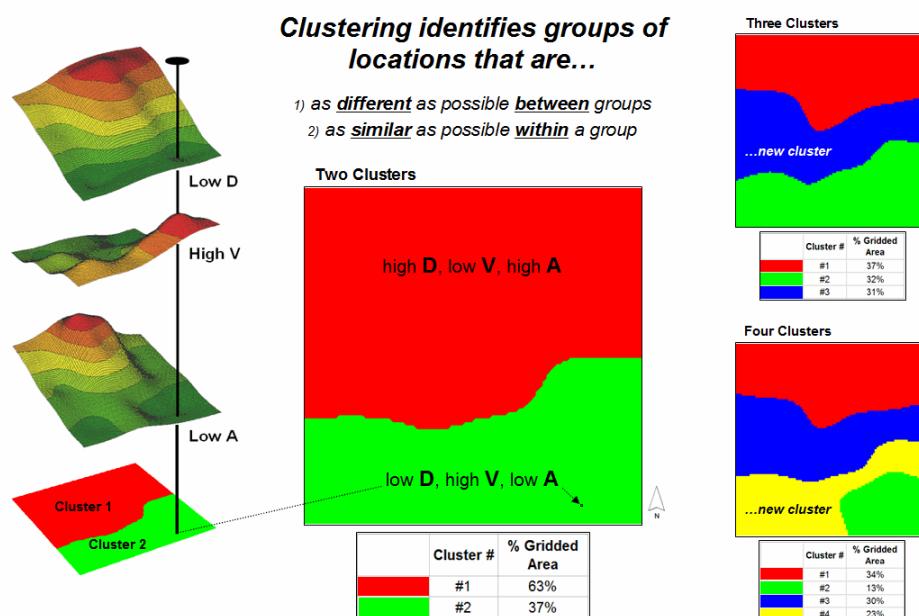


Figure 18. Example output from map clustering.

The map in the center of the figure shows the results of classifying the D, V and A map stack into two clusters. The data pattern for each cell location is used to partition the field into two groups that are 1) as different as possible between groups and 2) as similar as possible within a group. If all went well, any other division of the mapped data into two groups would be worse at mathematically balancing the two criteria.

The two smaller maps on the right show the division of the data set into three and four clusters. In all three of the cluster maps, red is assigned to the cluster with relatively high Density, low Value and high Age responses (less wealthy) and green to the one with the most opposite conditions (wealthy areas). Note the encroachment at the margin on these basic groups by the added clusters that are formed by reassigning data patterns at the classification boundaries. The procedure is effectively dividing the

project area into “data neighborhoods” based on relative D, V and A values throughout the map area. Whereas traditional neighborhoods usually are established by historical legacy, cluster partitions respond to similarity of mapped data values and can be useful in establishing insurance zones, sales areas and marketing clusters.

The mechanics of generating cluster maps are quite simple. Just specify the input maps and the number of clusters you want then miraculously a map appears with discrete data groupings. So how is this miracle performed? What happens inside clustering’s black box?

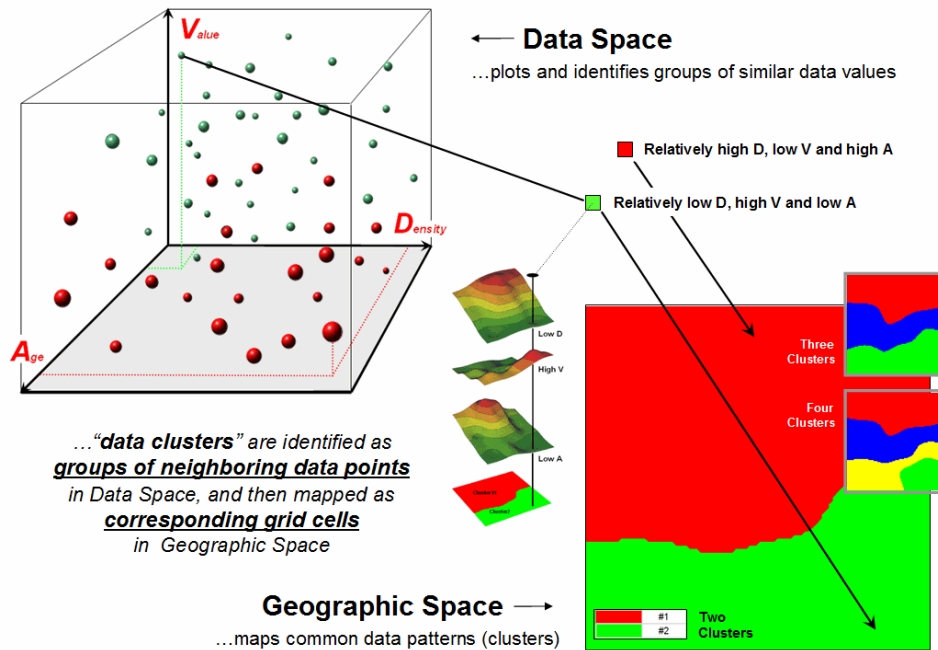


Figure 19. Data patterns for map locations are depicted as floating balls in data space with groups of nearby patterns identifying data clusters.

The schematic in figure 19 depicts the process. The floating balls identify the data pattern for each map location (in Geographic Space) plotted against the P, K and N axes (in Data Space). For example, the tiny green ball in the upper-right corner corresponds to a map location in the wealthiest part of town (low D, high V and low A). The large red ball appearing closest depicts a location in a less wealthy part (high D, low V and high A). It seems sensible that these two extreme responses would belong to different data groupings (clusters 1 and 2, respectively).

While the specific algorithm used in clustering is beyond the scope of this discussion (see author’s note), it suffices to recognize that data distances between the floating balls are used to identify cluster membership— groups of balls that are relatively far from other groups (different between groups) and relatively close to each other (similar within a group) form separate data clusters. In this example, the red balls identify relatively less wealthy locations while green ones identify wealthier locations. The geographic pattern of the classification (wealthier in the south) is shown in the 2D maps in the lower right portion of the figure.

Identifying groups of neighboring data points to form clusters can be tricky business. Ideally, the clusters will form distinct “clouds” in data space. But that rarely happens and the clustering technique has to enforce decision rules that slice a boundary between nearly identical responses. Also, extended techniques can be used to impose weighted boundaries based on data trends or expert knowledge. Treatment of categorical data and leveraging spatial autocorrelation are additional considerations.

So how do you know if the clustering results are acceptable? Most statisticians would respond, "...you can't tell for sure." While there are some elaborate procedures focusing on the cluster assignments at the boundaries, the most frequently used benchmarks use standard statistical indices, such as T- and F-statistics used in comparing sample populations.

Figure 20 shows the performance table and box-and-whisker plots for the map containing two data clusters. The average, standard deviation, minimum and maximum values within each cluster are calculated. Ideally the averages between the two clusters would be radically different and the standard deviations small—large difference between groups and small differences within groups.

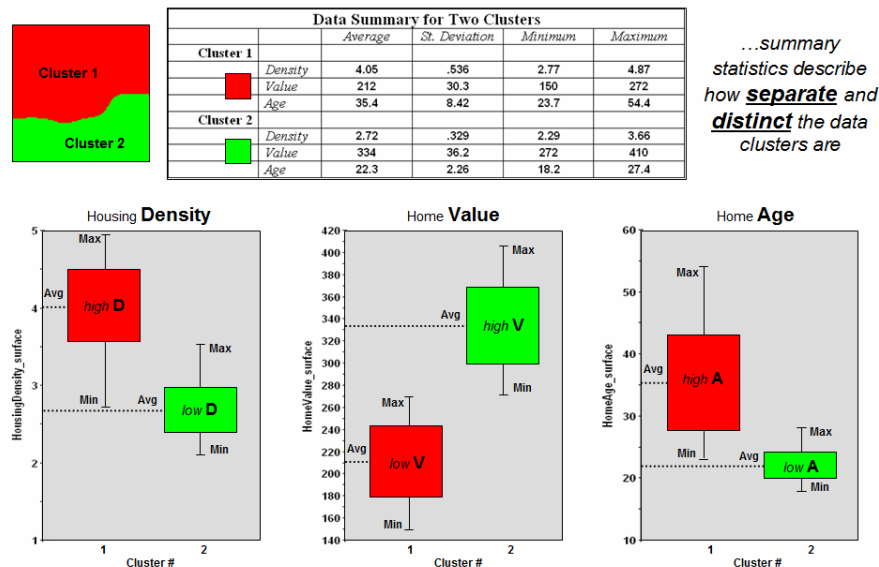


Figure 20. Clustering results can be evaluated using "box and whisker" diagrams and basic statistics.

Box-and-whisker plots enable us to visualize these differences. The box is aligned on the Average (center of the box) and extends above and below one Standard Deviation (height of the box) with the whiskers drawn to the minimum and maximum values to provide a visual sense of the data range. If the plots tend to overlap a great deal, it suggests that the clusters are not very distinct and indicates significant overlapping of data patterns between clusters.

The separation between the boxes in all three of the data layers of the example suggests good distinction between the two clusters with the Home Value grouping the best with even the Min/Max whiskers not overlapping. Given the results, it appears that the clustering classification in the example is acceptable... and hopefully the statisticians among us will accept in advance my apologies for such an introductory and visual treatment of a complex topic.

Mapping the Future

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 a non-spatial approach doesn't account for geographic patterns and the result is just summary of the overall relationship for an entire project area.

Extending predictive analysis to mapped data seems logical as maps at their core 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 columns have discussed how the computer can "see" spatial data

relationships including “descriptive techniques” for assessing *map similarity*, *data zones*, and *clusters*. The next logical step is to apply “predictive techniques” that generates mapped forecasts of conditions for other areas or time periods.

To illustrate the process, suppose a bank has a database of home equity loan accounts they have issued over several months. Standard geo-coding techniques are applied to convert the street address of each sale to its geographic location (latitude, longitude). In turn, the geo-tagged data is used to “burn” the account locations into an analysis grid as shown in the lower left corner of figure 21. A roving window is used to derive a Loan Concentration surface by computing the number of accounts within a specified distance of each map location. Note the spatial distribution of the account density—a large pocket of accounts in the southeast and a smaller one in the southwest.

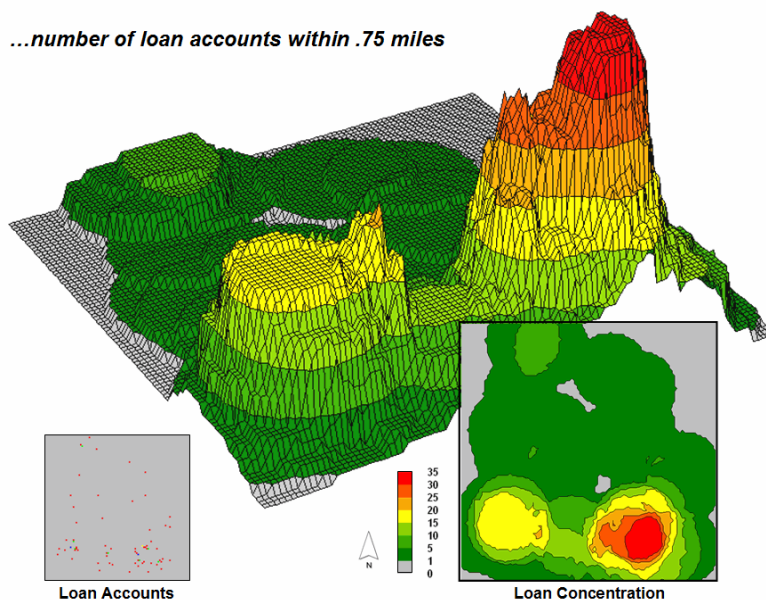


Figure 21. A loan concentration surface is created by summing the number of accounts for each map location within a specified distance.

The most frequently used method for establishing a quantitative relationship among variables involves **Regression**. It is beyond the scope of this column to discuss the underlying theory of regression; however in a conceptual nutshell, a line is “fitted” in data space that balances the data so the differences from the points to the line (termed the residuals) are minimized and the sum of the differences is zero. The equation of the best-fitted line becomes a prediction equation reflecting the spatial relationships among the map layers.

To illustrate predictive modeling, consider the left side of figure 22 showing four maps involved in a regression analysis. The loan Concentration surface at top is 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 relationship. For example, the “pin” in the figure identifies a location where high loan Concentration coincides with a low housing Density, high Value and low Age response pattern.

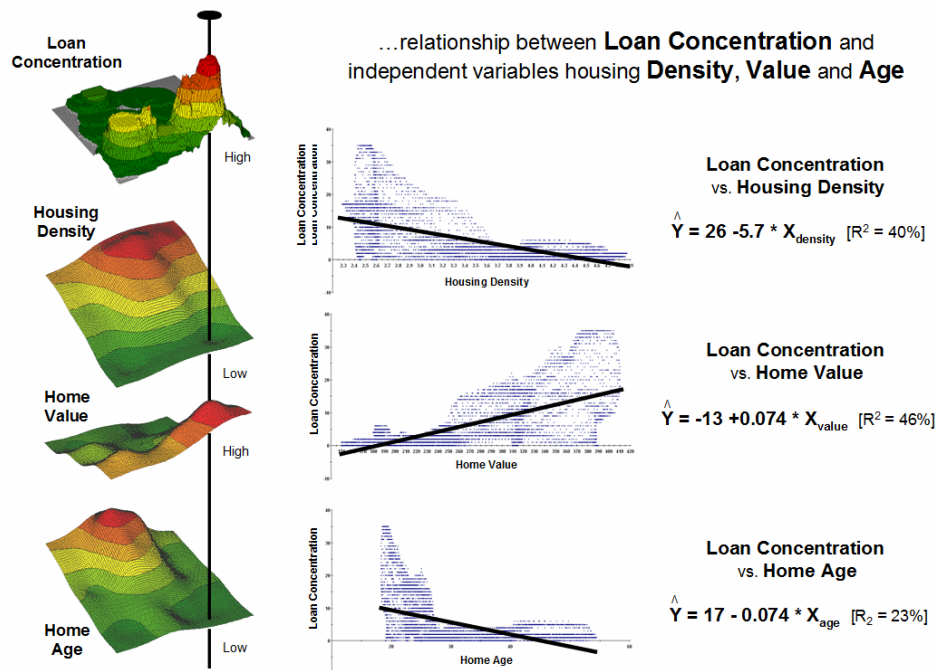


Figure 22. Scatter plots and regression results relate Loan Density to three independent variables (housing Density, Value and Age).

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). 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 indicates a *positive correlation* where an increase in X always results in a corresponding increase in Y. A downward directed line indicates a *negative correlation* with an increase in X resulting in a corresponding decrease in Y. The slope of the line indicates the extent of the relationship with a 45-degree slope indicating a 1-to-1 unit change. A vertical or horizontal line indicates *no correlation*— a change in one variable doesn't affect the other. Similarly, a circular cloud of points indicates there isn't any consistency in the changes.

Rarely does the data plot into these ideal conditions. Most often they form dispersed clouds like the scatter plots in figure 22. 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 22 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 estimated within .75 miles. 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 a map of expected loan Concentration. 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.

Multiple regression can be used to simultaneously consider all three independent map variables as a means to derive a better prediction equation. Or more sophisticated modeling techniques, such as Non-linear Regression and Classification and Regression Tree (CART) methods, can be used that often results in an R-squared index exceeding 90% (nearly perfect).

The bottom line is that predictive modeling using mapped data is fueling a revolution in sales forecasting. Like parasailing on a beach, spatial data mining and predictive modeling are affording an entirely new perspective of geo-business data sets and applications by linking data space and geographic space through grid-based map analysis.

Mapping Potential Sales

My first sojourn into geo-business involved an application to extend a test marketing project for a new phone product (nick-named “teeny-ring-back”) that 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. This pre-Paleolithic project was debuted in 1991 when phones were connected to a wall by a pair of copper wires and street addresses for customers could be used to geo-code the actual point of sale/use. 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 of the project was that a relationship existed between conditions throughout the city, such as income level, education, number in household, etc. could help explain sales pattern. The demographic data for the city was analyzed to calculate a prediction equation between product sales and census data.

The prediction equation derived from test market sales in one city could be applied to another city by evaluating exiting demographics to “solve the equation” for a predicted sales map. In turn, the predicted sales map was combined with a wire-exchange map to identify switching facilities that required upgrading before release of the product in the new city. Although GIS systems were crude at the time, the project was deemed a big success.

Now fast-forward to more contemporary times. A GeoWorld feature article described a similar, but much more thorough analysis of retail sales competition (*Beyond Location, Location, Location: Retail Sales Competition Analysis*, www.innovativegis.com/basis/present/GW06_retail/GW06_Retail.htm; GeoWorld, March 2006). Figure 23 outlines the steps for determining competitive advantage for various store locations.

Most GIS users are familiar with network analysis that accepts starting and ending locations and then determines the best route between the two points along a road network. However the complexity of retail competition analysis with tens of thousands of customers and dozens of competitor locations makes the traditional point-to-point navigational solution impractical. A more viable approach uses grid-based map analysis involving continuous surfaces (steps 1 and 2 in figure 23).

Competition Analysis – Spatial Analysis Steps

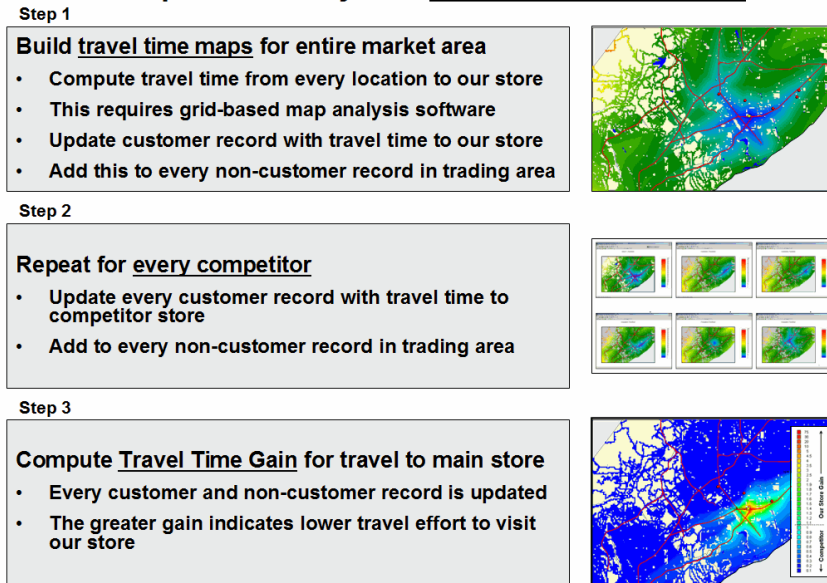


Figure 23. Spatial Modeling steps derive the relative travel time relationships for our store and each of the competitor stores for every location in the project area and links this information to customer records.

Predictive Modeling – Spatial Statistics Steps

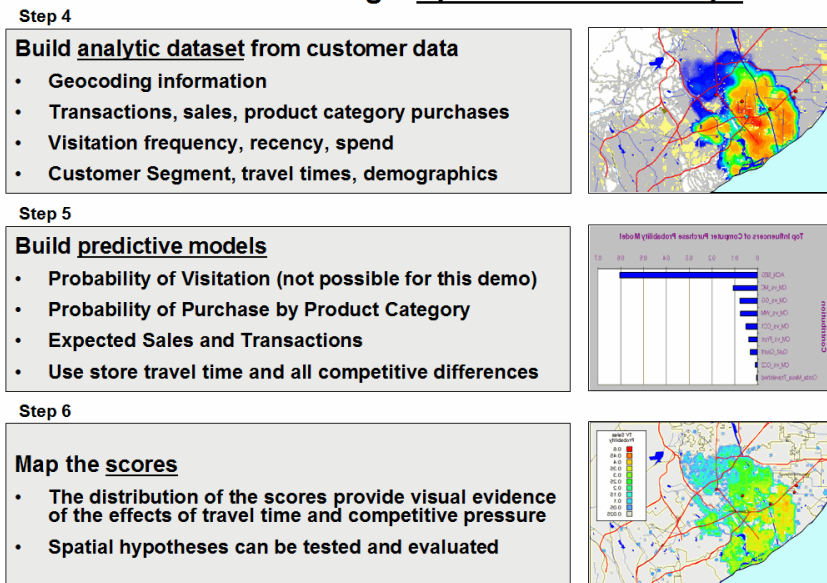


Figure 24. Predictive Modeling steps use spatial data mining procedures for relating spatial and non-spatial factors to sales data to derive maps of expected sales for various products.

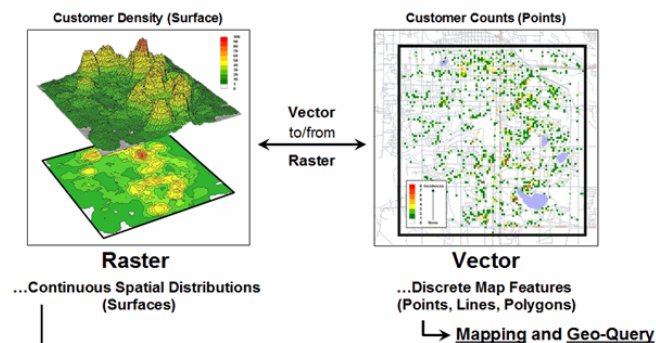
Step 1 map shows the grid-based solution for travel-time from “Our Store” to all other grid locations in the project area. The blue tones identify grid cells that are less than twelve minutes away assuming travel on the highways is four times faster than on city streets. Note the star-like pattern elongated around the highways and progressing to the farthest locations (warmer tones). In a similar manner,

competitor stores are identified and the set of their travel time surfaces forms a series of geo-registered maps supporting further analysis (Step 2).

Step 3 combines this information for a series of maps that indicate the relative cost of visitation between our store and each of the competitor stores (pair-wise comparison as a normalized ratio). The derived “Gain” factor for each map location is a stable, continuous variable encapsulating travel-time differences that is suitable for mathematical modeling. A Gain of less than 1.0 indicates the competition has an advantage with larger values indicating increasing advantage for our store. For example, a value of 2.0 indicates that there is a 200% lower cost of visitation to our store over the competition.

Figure 24 summarizes the predictive modeling steps involved in competition analysis of retail data. The geo-coding link between the analysis frame and a traditional customer dataset containing sales history for more than 80,000 customers was used to append travel-times and Gain factors for all stores in the region (Step 4).

The regression hypothesis was that sales would be predictable by characteristics of the customer in combination with the travel-time variables (Step 5). A series of mathematical models are built that predict the probability of purchase for each product category under analysis. This provides a set of model scores for each customer in the region. Since a number of customers could be found in many grid cells, the scores were averaged to provide an estimate of the likelihood that a person from each grid cell would travel to our store to purchase one of the analyzed products. The scores for each product are mapped to identify the spatial distribution of probable sales, which in turn can be “mined” for pockets of high potential sales.



Map Analysis Classes of Operations:

- **Spatial Statistics** — “numerical” relationships
 - Surface Modeling — derives spatial distribution (within a map layer)
 - Density Analysis* ♦ *Spatial Interpolation (IDW and Kriging)*
 - Spatial Data Mining — investigates spatial relationships (among map layers)
 - Similarity Map* ♦ *Level Slicing* ♦ *Clustering* ♦ *Map Regression*
- **Spatial Analysis** — “contextual” relationships
 - Reclassifying Maps — reassign map values
 - Overlaying Maps — map coincidence
 - Characterizing Distance — proximity and connectivity
 - Effective Proximity* ♦ *Travel-Time Map* ♦ *Competition Analysis*
 - Summarizing Neighbors — roving windows

Figure 25. Map Analysis exploits the digital nature of modern maps to examine spatial patterns and relationships within and among mapped data.

Targeted marketing, retail trade area analysis, competition analysis and predictive modeling provide examples applying sophisticated *Spatial Analysis* and *Spatial Statistics* to improve decision making. The techniques described in the past nine sections have focused on Map Analysis— procedures that extend traditional mapping and geo-query to map-ematically based analysis of mapped data. Figure 25 outlines the classes of operations described in the series (blue highlighted techniques were specifically discussed).

Conclusion

Recall that the keystone concept is an *Analysis Frame* of grid cells that provides for tracking the continuous spatial distributions of mapped variables and serves as the primary key for linking spatial and non-spatial data sets. While discrete sets of points, lines and polygons have served our mapping demands for over 8,000 years and keep us from getting lost, the expression of mapped data as continuous spatial distributions (surfaces) provides a new foothold for the contextual and numerical analysis of mapped data— in many ways, “thinking with maps” is more different than it is similar to traditional mapping.

References

[Analyzing Geo-Business Data](#) by Joseph K. Berry (2003; BASIS Press) ...a case study in Spatial Analysis and Data Mining in Business with software, database and “hands-on” exercises (US\$45; www.innovativegis.com/basis/Books/AnalyzingGBdata/)

[Map Analysis](#) by Joseph K. Berry (2007; GeoTec Media) ...provides a structured view of grid-based map analysis and GIS modeling with software, database and “hands-on” exercises (US\$45; www.innovativegis.com/basis/Books/MapAnalysis/)

Also see the online book [Beyond Mapping III](#) by Joseph K. Berry (BASIS Press) that is posted at www.innovativegis.com/basis

...select the following topics for more information on grid-based map analysis and procedures described in this paper:

- Topic 7, Linking Data Space and Geographic Space
- Topic 10, Analyzing Map Similarity and Zoning
- Topic 14, Deriving and Using Travel-Time Maps
- Topic 16, Characterizing Patterns and Relationships
- Topic 18, Understanding Grid-based Data
- Topic 22, Overview of Spatial Analysis and Statistics
- Topic 28, Spatial Data Mining in Geo-Business

¹ **Dr. Berry** is the Principal of Berry and Associates // Spatial Information Systems ([BASIS](#)), consultants and software developers in GIS technology. He is a leading consultant and educator in the application of GIS technology, written over two hundred papers on the theory and application of map analysis techniques, and is the author of the popular books [Beyond Mapping](#) (Wiley, 1993), [Spatial Reasoning](#) (Wiley 1995) and [Map Analysis](#) (GeoTec Media, 2007).

² **Dr. Reed** is with Xtreme Data Mining. Prior to this position he was an Associate Partner with Accenture and head of their retail sales modeling group with customers throughout the world. His geo-business projects have extended from analyzing in-store shopping patterns to modeling sales potential throughout large metropolitan areas.