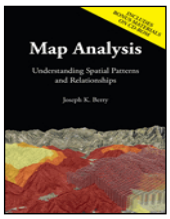
Beyond Mapping III

Topic 10 – Spatial Data Mining (Further Reading)



Map Analysis book

(***Underlying Spatial Data Mining Concepts***)

Beware the Slippery Surfaces of GIS Modeling — *discusses the relationships among maps, map surfaces and data distributions (May 1998)*
Link Data and Geographic Distributions — *describes the direct link between numeric and geographic distributions (June 1998)*
Explore Data Space — *establishes the concept of "data space" and how mapped data conforms to this fundamental view (July 1998)*
Identify Data Patterns — *discusses data clustering and its application in identifying spatial patterns (August 1998)*

(***Advanced Map Comparison Techniques***)

Compare Maps by the Numbers — *describes several techniques for comparing discrete maps (September 1999)*
Use Statistics to Compare Map Surfaces — *describes several techniques for comparing continuous map surfaces (October 1999)*

(***Approaches Used in Deriving Prediction Maps***)

Use Scatterplots to Understand Map Correlation — *discusses the underlying concepts in assessing correlation among maps (November 1999)*
Can Predictable Maps Work for You? — *describes a procedure for deriving a spatial prediction model (December 1999)*
Spatial Data Mining Allows Users to Predict Maps — *describes the basic concepts and procedures for deriving equations that can be used to derive prediction maps (January 2002)*
Stratify Maps to Make Better Predictions — *illustrates a procedure for subdividing an area into smaller more homogenous groups prior to generating prediction equations (February 2002)*

<Click here> for a printer-friendly version of this topic (.pdf).

(***Back to the Table of Contents***)

(***Underlying Spatial Data Mining Concepts***)

Beware the Slippery Surfaces of GIS Modeling

(GeoWorld, May 1998)

(return to top of Topic)

GIS means different things to different people. My favorite characterization is that "GIS technology is as different as it is similar to traditional mapping and data analysis." That statement leaves some room for both conventional and unconventional views of what GIS is, what it's good for, and who uses it for what.

For example, traditional mapping focuses on *points*, *lines* and *polygons* as fundamental map features. Spatial database management systems (desktop mapping packages) extend this view by linking the discrete "map objects" to descriptive information about them. The nature of the maps, as well as the linkage, is as familiar as the map on the wall and the file cabinet beside it. The paradigm also fits nicely into standard office software with only minimal education and training in unfamiliar spatial reasoning.

The problem is the traditional paradigm doesn't fit a lot of reality. Sure the lamppost, roadway and the parking lot at the mall are physical realities of a map's set of points, lines and polygons. Even very real (legally), though non-physical, property boundaries conceptually align with the traditional paradigm. But not everything in space can be so discretely defined, nor is it as easily put in its place.

Obviously meteorological gradients, such as temperature and barometric pressure, don't fit the P, L and P mold. They represent phenomena that are constantly changing in geographical space, and are inaccurately portrayed as contour lines, regardless how precisely the lines are drawn. By its very nature, a continuous spatial variable is corrupted by artificially imposing abrupt transitions. The contour line is simply a mechanism to visually portray a 2-dimensional rendering of a 3-dimensional gradient. It certainly isn't a basic map feature, nor should it be a data structure for storage of spatially complex phenomena.

Full-featured GIS packages extend the basic P, L and P features to map *surfaces* that treat geographic space as a continuum. The most familiar surface feature is the terrain you walk on, composed of hills and valleys and everything in between. One of the unique characteristics of an elevation surface is well known to hikers— terrain steepness or slope. While hikers hate steep slopes, avalanches love them. Another unique characteristic of surfaces determines the orientation for a portion of the surface— aspect. Elk seek out southern exposures in the winter, but thirsty trees wouldn't be caught dead there in a moisture-limited climate.

While a contour map of elevation lets your eye assess slope (closeness of lines) and aspect (downhill direction), it leaves the computer without a clue, as all it can see is a pile of numbers (digital data), not an organized set of lines (analog image).

Figure 1 links different views of surface data. Keep in mind, that the computer doesn't "see" any of them; they're for human viewing enjoyment. The traditional *2-D map view* chooses a map value, then identifies all of the locations (mathematically implied) having that value. That begs the traditional GIS 101 question, "Does a set of contours form point, line or area features?" Actually, arguments can be made for any of the discrete feature types depending on the application.

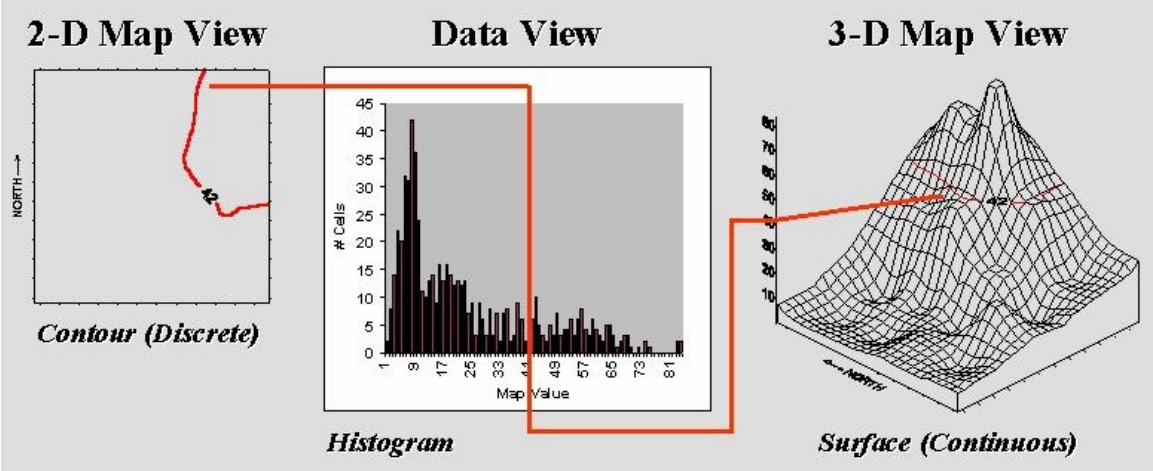


Figure 1. A histogram (numerical distribution) is linked to a map surface (geographic distribution) by a common axis of map values.

However, it's the discrete nature of a contour map and the irregular features spawned that restrict its ability to effectively represent continuous phenomena. The *data view* uses a histogram to characterize a continuum of map values in "numeric" space. It summarizes the number of times a given value occurs in a data set (relative frequency).

The ***3-D map view*** forms a continuous surface by introducing a regular grid over an area. The X and Y axes position the grid cells in geographic space, while the Z axis reports the numeric value at that location. Note that the data and surface views share a common axis— map value. It serves as the link between the two perspectives. For example, the data view shows number of occurrences for the value 42 and the relative numerical frequency considering the occurrences for all other values. Similarly, the surface view identifies all of the locations having a value of 42 and the relative geographic positioning considering the occurrences for all other values.

The concept of an "aggregation interval" is shared as well. In constructing a histogram, a constant data step is used. In the example, a data interval of 1.0 unit was used, and all fractional values were rounded, then "placed" in the appropriate "bin." In an analogous fashion, the aggregation interval for constructing a surface uses a constant geographic step, as well as a constant data step. In the example, an interval of 1.0 hectare was used to establish the constant partitioning in geographic space. In either case, the smaller the aggregation interval the better the representation.

Both perspectives characterize data dispersal. The ***numeric distribution*** of data (a histogram's shape of ups/downs) forms the cornerstone of traditional statistics and determines appropriate data analysis procedures. In a similar fashion, a map surface establishes the ***geographic distribution*** of data (a surface's shape of hills/valleys) and is the cornerstone of spatial statistics. The assumptions, linkages, similarities and differences between these two perspectives of data distribution are the focus of the next few columns… we'll dust off the old stat book together.

Link Data and Geographic Distributions

(GeoWorld, June 1998)

(return to top of Topic)

Some of the previous Beyond Mapping columns might have found you reaching for your old Stat 101 textbook. Actually, the concepts used in mapped data analysis are quite simple— it's the intimidating terminology and "picky, picky" theory that are hard. The most basic concept involves a ***number line*** that is like a ruler with tic-marks for numbers from small to large. However, the units aren't always inches, but data units like number of animals or dollars of sales. If you placed a dot for each data measurement (see top of figure 1), there would be a ***minimum*** value on the left (#animals = 0) extending to a ***maximum*** value on the right. The rest of the points would fall on top of each other throughout the data ***range***.

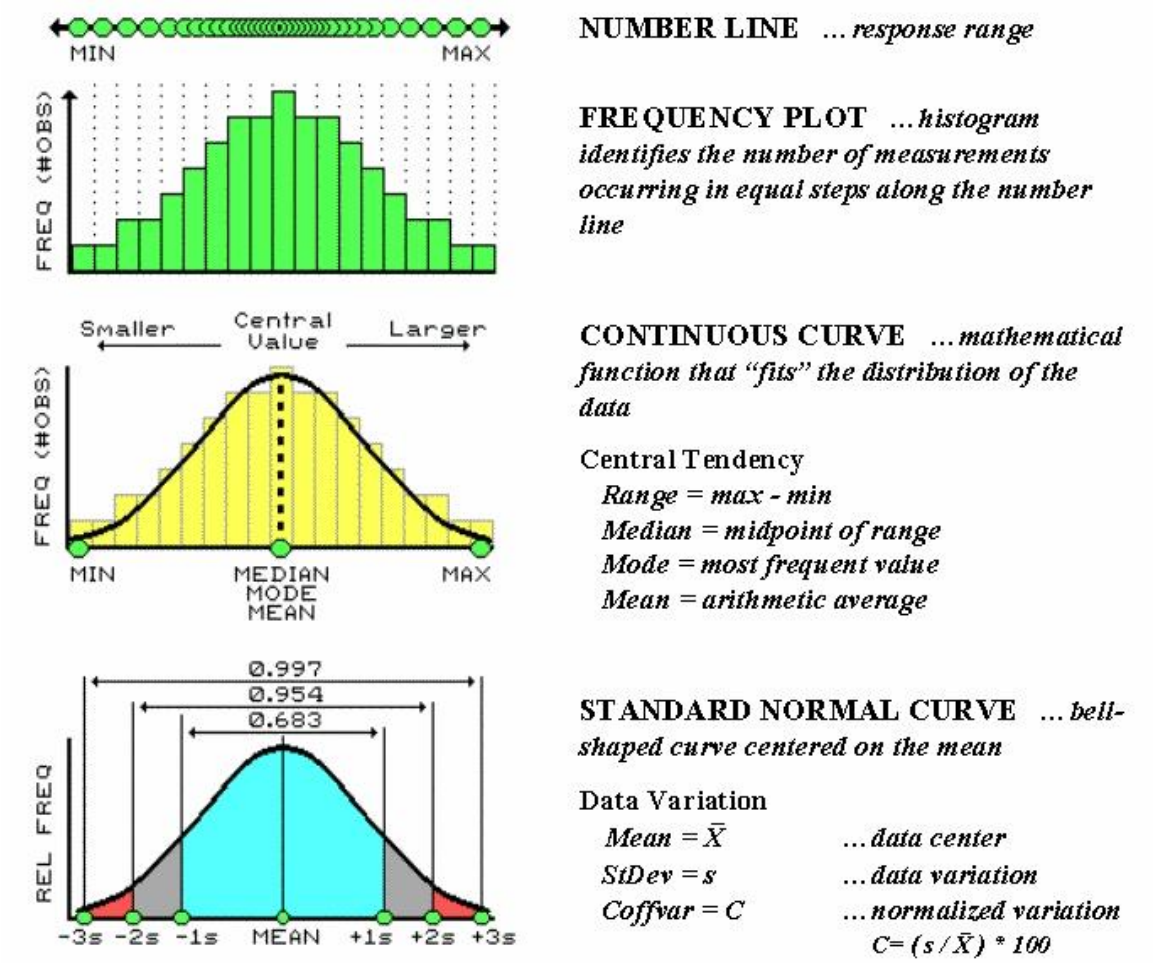


Figure 1. The distribution of measurements in "data space" is described by its histogram and summarized by descriptive statistics.

To visualize these data, we can look at the number line from the side and note where the measurements tend to pile up. If the number line is divided into equally spaced "shoots" (like in pinball machine) the measurements will pile up to form a ***histogram*** plot of the data's

distribution. Now you can easily see that most of the measurements fell about midrange.

In statistics, several terms are used to describe this plot and its "central tendency." The ***median*** identifies the "midway" value with half of the distribution below it and half above it, while the ***mode*** identifies the most frequently occurring value in the data set. The ***mean***, or *average*, is a bit trickier as it requires calculation. The total of all the measurements is calculated and then divided by the number of measurements in the data set.

Although the arithmetic is easy (for a tireless computer), its implications are theoretically deep. When you calculate the mean and its standard deviation you're actually imposing the "standard normal curve" assumption onto the histogram. The bell-shaped curve is symmetrical with the mean at its center. For the "normally distributed" data shown in the figure, the fit is perfect with exactly half of the data on either side. Also note that the mean, mode and median occur at the same value for this idealized distribution of data.

Now let's turn our attention to the tough stuff— characterizing the data variation about the mean. When considering variation one must confront the concept of a ***standard deviation*** (StDev). The standard deviation describes the dispersion, or spread, of the data around the mean. It's a consistent measure of the variation, as one standard deviation on either side of the mean "captures" slightly more than two-thirds of the data (.683 of the total area under the curve to be exact). Approximately 95% of all the measurements are included within two standard deviations, and more than 99% are covered by three.

The larger the standard deviation, the more variable is the data, indicating that the mean isn't very typical. In GIS applications, a small standard deviation tells you there isn't much variation in an area of interest. However, a large standard deviation indicates a lot of variability and using a simple average to characterize the area is likely misleading.

So what determines whether a standard deviation is large or small? That's the role of the ***coefficient of variation*** (Coffvar). This semantically-challenging mouthful simply "normalizes" the variation in the data by expressing the standard deviation as a percent of the mean— if it's large, say over 50%, then there is a lot of variation and the mean is a poor estimator of what's happening in a mapped area. Keep this in mind the next time you assign an average value to map features, such as the average tree diameter for each forest parcel, or the average home value for each county.

A large portion of the variation can be "explained" through its spatial distribution. Figure 2 shows a technique that brings statistics down to earth by mapping the standard normal curve in geographic space. The procedure first calculates the mean and standard deviation for the typical response in a data set. The data is then spatially interpolated into a continuous geographic distribution. A ***standard normal variable surface*** is derived by subtracting the mean from the map value at each location (deviation from the typical), then dividing by the standard deviation (normalizing to the typical variation) and multiplying by a hundred to form a percent. The result is that every map location gets a number indicating exactly how "typical" it is.

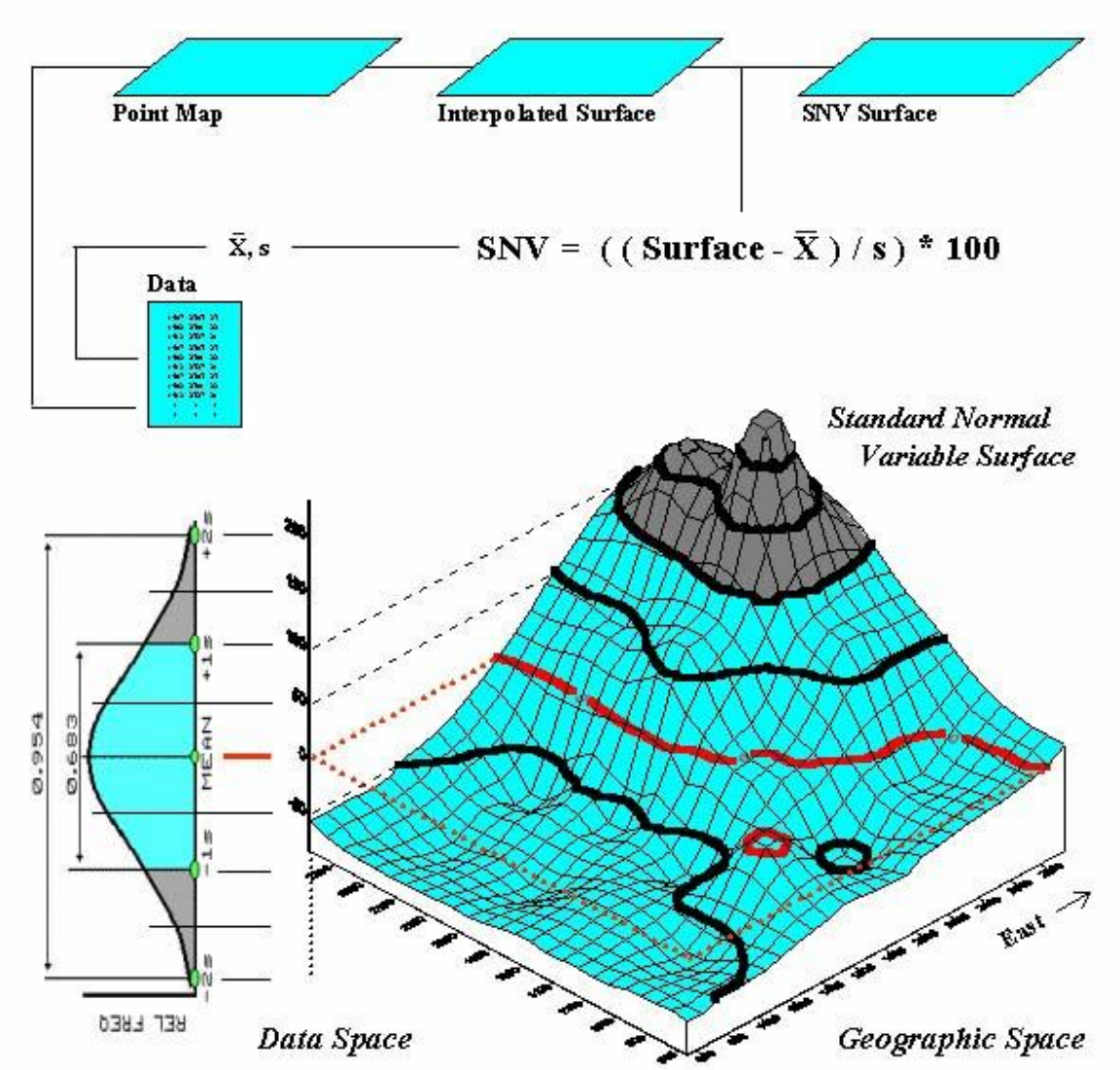


Figure 2. A "standard normal surface" identifies how typical every location is for an area of interest.

The contour lines draped on the SNV surface in the figure show half-StDev steps. Locations that are exactly the same as the mean equate to zero (typical areas). The high peak in the northeast portion locates an area of unusually high response (>200% of a standard deviation above the mean). The entire western portion of the map is characterized by responses below the typical, however there are no unusually low areas in this case. The geographic distribution balances at zero (just as many locations above the mean as below), but the data distribution is not balanced (no unusually low responses). This condition (asymmetrical data) muddles the standard normal procedure, but that discussion is reserved for techy-types via *www.innovativegis.com/basis*, select *Column Supplements*.

Explore Data Space

(GeoWorld, July 1998)

(return to top of Topic)

The previous discussion described a histogram plot and the old bell curve as a computer's view of an individual map. Now let's extend those concepts to how a computer "visualizes" several maps at a time. Fundamental to this perspective is recognition that digital maps are "numbers first, pictures later." In a computer, each map is a long list of numbers, with each value identifying a characteristic or condition at a particular location.

Following my usual recipe for journalistic suicide let's consider the fundamental concepts of the dismal discipline of spatial statistics within the context of production agriculture. For example, a map of phosphorous in the top layer of soil (0-5cm) in a farmer's field contains values ranging from 22 to 140 parts per million. The spatial pattern of these data is characterized by the relative positioning of the data values within a reference grid of cell locations. The set of cells forming the analysis grid identifies the spatial domain (***geographic space***), while the map values identifies the data domain (***data space***).

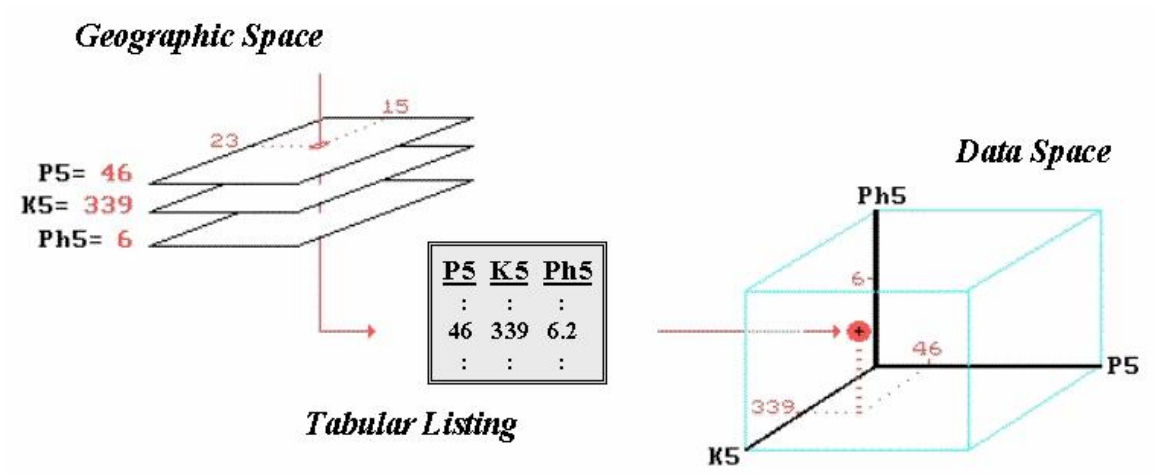


Figure 1. The map values for a series of maps can be simultaneously plotted in data space.

A dull and tedious ***tabular listing***, as shown in the center of Figure 1, is the traditional human perspective of such data. We can't consume a long list of numbers, so we immediately turn the entire column of data into a single "typical" value (average) and use it to make a decision. For the soil phosphorous data set, the average is 48. A location in the center of the field (column 15, row 23 of the analysis grid) has a phosphorous level of 46 that is close to the average value (in a data sense, not a geographic sense). But recall that the data range tells us that somewhere in the field there is at least one location that is less than half (22) and another that is nearly three times the average (140), so the average value doesn't tell it all.

Now consider additional map surfaces of potassium levels (K5) and soil acidity (Ph5), as well as phosphorous (P5), for the field. As humans we could "see" the coincidence of these data sets by aligning their long columns of numbers in a spreadsheet or database. Specific levels for all three

of the soil measurements at any location in the field are identified as rows in the table. However, the combined set of data is even more indigestible, with the only "humane" view of map coincidence being the assumption that the averages are everywhere— 48, 419, 6.2 in this example. The center location's ***data pattern*** of 46, 339, and 6.0 is fairly similar to the pattern of the field averages, but exactly how similar? Which map locations in the field are radically different?

Before we can answer these questions, we need to understand how the computer "sees" similarities and differences in multiple sets of data. It begins with a three-dimensional plot in data space as shown on the right side of figure 1. The data values along a row of the tabular listing are plotted resulting in each map location being positioned in the imaginary box based on its data pattern. Similarity among the field's soil response patterns are determined by their relative positioning in the box— locations that plot to the same position in the data space box are identical; those that plot farther away from one another are less similar.

How the computer "measures" the relative distance becomes the secret ingredient in assessing data similarity. Actually it's quite simple if you revisit a bit of high school geometry, but I bet you thought you had escaped all that awful academic fluff when you entered the colorful, fine arts world of computer mapping and geo-query.

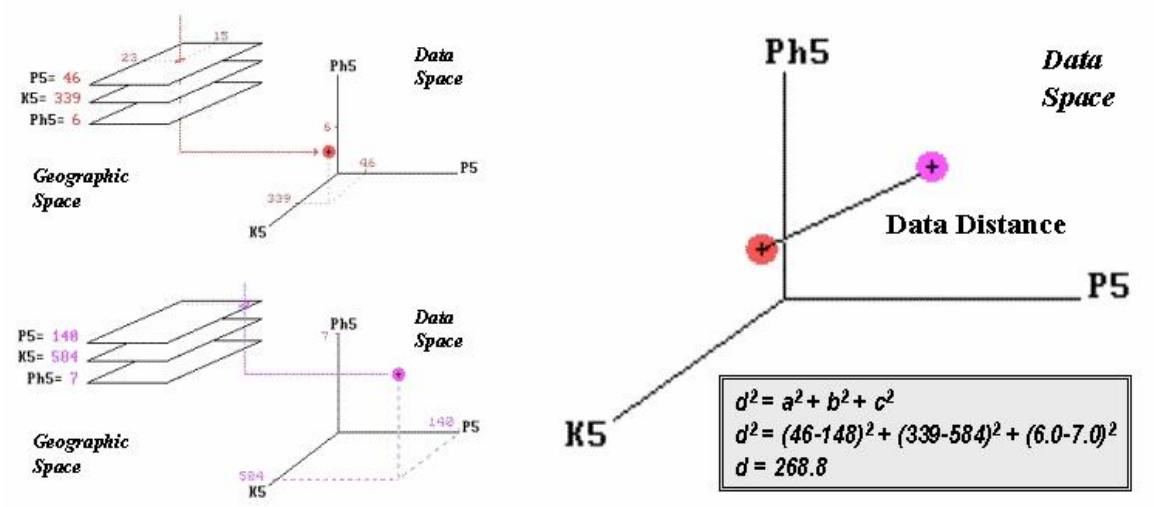


Figure 2. Similarity is determined by the data distance between two locations and is calculated by expanding the Pythagorean Theorem.

The left side of figure 2 shows the data space plots for soil conditions at two locations in the farmer's field. The right side of the figure shows a straight line connecting the data points whose length identifies the data distance between the points. Now for the secret— it's the old Pythagorean Theorem of $c^2 = a^2 + b^2$ (I bet you remember it).

However, in this case it looks like $d^2 = a^2 + b^2 + c^2$ as it has to be expanded to three dimensions to accommodate the three maps of phosphorous, potassium and acidity (P5, K5 and Ph5 axes in the figure). All that the Wizard of Oz (a.k.a., computer programmer) has to do is subtract the values for each condition between two locations and plug them into the equation. If there are

more than three maps, the equation simply keeps expanding into *hyper*-data space which as humans we can no longer plot or easily conceptualize.

The underlying principle is that the smaller the data distance, the greater the similarity between locations. That's the basics, but the subtle nuances, such as normalizing the axes, provide fodder for the next sectuion's discussion of mapping similarity and clustering data into similar ***spatial patterns***— a useful tool down on the farm, but just as useful for retailers, land planners, foresters, real estate agents and just about any GIS user.

Identify Data Patterns

(GeoWorld, August 1998)

(return to top of Topic)

The previous section introduced the concept of data distance. While most of us are comfortable with the concept of distance in geographic space, things get a bit abstract when we move from feet and meters in the real world to data units in data space.

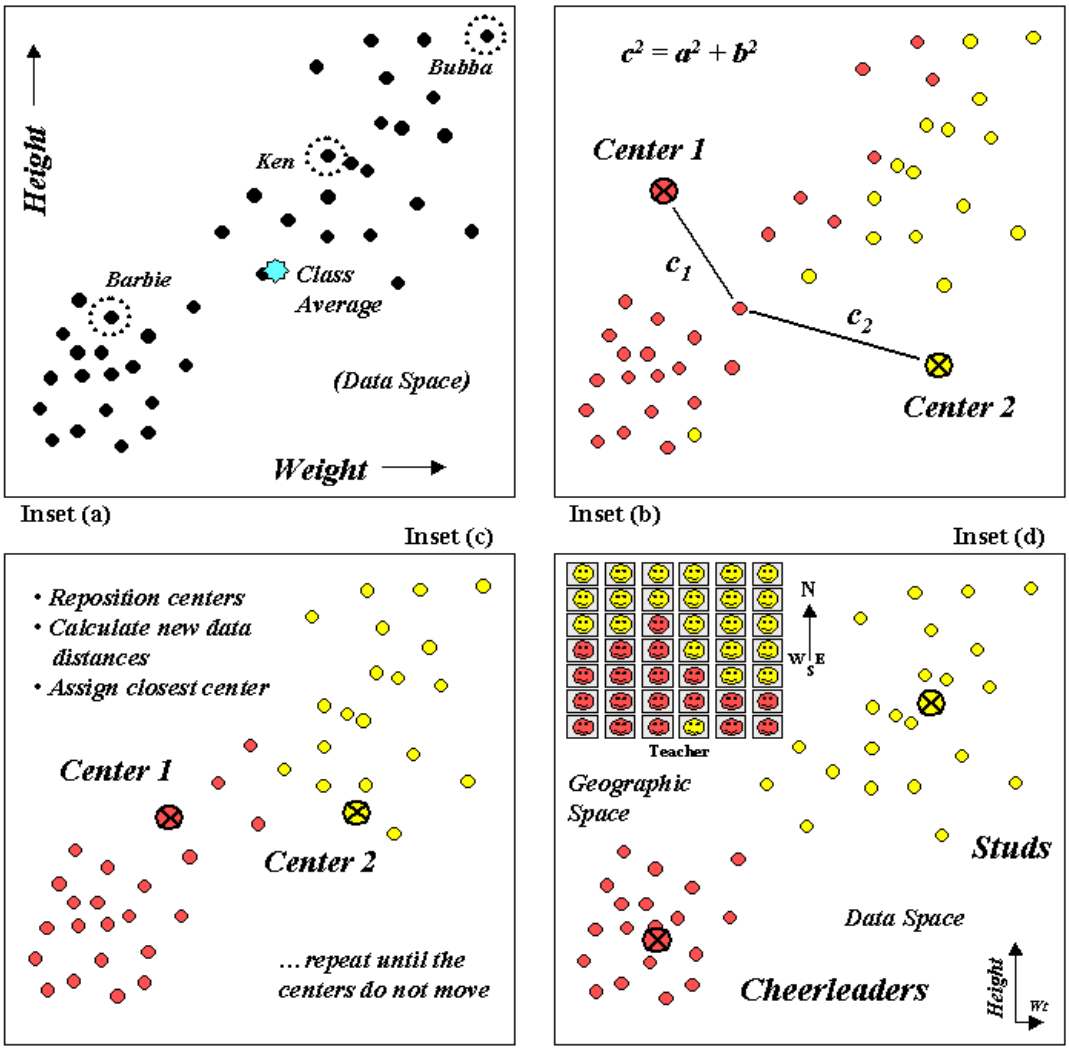


Figure 1. Clustering uses repeated data distance calculations to identify numerical patterns in a data set.

Recall that data space is formed by the intersection of two or more axes in a typical graph. If you measured the weight and height of several students in your old geometry class, each of the paired measurements for a person would plot as a dot in XY data space locating their particular weight (X axis) and height (Y axis) combination. A plot of all the data looks like a shotgun blast and is termed a ***scatter plot***.

The scatter plots for a lot of data sets form clusters of similar measurements. For example (see figure 1), two distinct groups might be detected in the geometry class's data— the demure cheerleaders (low weight and height) and the football studs (high weight and height). Traditional (non-spatial) data analysis stops at identifying the groupings. Spatial statistics, however, extends the analysis to geographic contexts.

The scatter plot in inset (a) of figure 1 shows weight/height data that might have been collected in your old geometry class. Note that all of the students do not have the same weight/height measurements and that many vary widely from the class average. Your eye easily detects two groups (low/low and high/high) in the plot but the computer just sees a bunch of numbers. So how does it identify the groups without seeing the scatter plot?

If seating "coordinates" accompanied the classroom data you might detect that most of the cheerleaders were located in one part of the room, while the studs were predominantly in another. Further analysis might show a spatial relation between the positioning of the groups and proximity of the teacher— cheerleaders in front and studs in back.

The linking of traditional statistics with spatial analysis capabilities (such as proximity measurement) provides insight into the spatial context and relationships inherent in the data. The only prerequisite is "tagging" geographic coordinates to the measurements. Until recently, this requirement presented quite a challenge and spatial coordinates were rarely included in most data sets. With the advent of GPS and Geo-Coding based on street address, geographic tagging has become a whole lot easier.

The new hurdle, however, isn't so much technical as it is social. Most data analysis types aren't familiar with spatial concepts, while most GIS types aren't familiar with data mining and knowledge discovery procedures. With GIS moving beyond mapping and geo-query, understanding data analysis concepts becomes as important as understanding geographic concepts. With this in mind, let's see one way a computer might identify ***numerical patterns*** within a set data.

One approach (termed *k-means clustering*) arbitrarily establishes two cluster centers in the data space (inset (b)). The data distance to each weight/height measurement pair is calculated and the point is assigned to the closest cluster center. Recall from the previous section that the Pythagorean Theorem of $c^2 = a^2 + b^2$ is used to calculate the ***data distance*** and can be extended to more than just two variables (hyper-data space). It should be at least some comfort to note that the geometry you learned in high school holds for the surreal world of data space, as well as the one you walk on. In the example, c_1 is smaller than c_2 therefore that student's measurement pair is assigned to cluster center 1. The remaining student assignments are identified in the scatter plot by their color codes.

The next step calculates the average weigh/height of the assigned students and uses these coordinates to reposition the cluster centers (inset (c)). Other rounds of data distances, cluster assignments and repositioning are made until the cluster membership does not change (i.e., the centers do not move).

Inset (d) shows the final groupings with the big folks (high/high) differentiated from the smaller folks (low/low). By passing these results to a GIS the data space pattern (clumps of similar measurements) can be investigated for geographic space patterns.

The positioning of these data in the real world classroom (upper left portion of inset (d)) shows a distinct ***spatial pattern*** between the two groups— smaller folks in front and bigger folks in the

rear. Like before, you simply see these things but the computer has to derive the relationships (distance to similar neighbor) from a pile of numbers.

What is important to note is that analysis in data space and geographic space has a lot in common. In the example, both "spaces" are represented as XY coordinates— weight/height measurements in data space (characteristics) and longitude/latitude in geographic space (positioning). Data distance is used to partition the measurements into separate groups (data pattern). Geographic distance is used to partition the locations into separate groups (spatial pattern). In both instances the old Pythagorean Theorem served as the procedure for measuring distance.

So who cares? We have gotten along for years with data analysts and mapmakers doing their own thing. Maps are maps and data are data, right? Not exactly, at least not any more… with the advent of the digital map, maps are data (not pictures). Information on the relative positioning and coincidence among mapped variables extend traditional data analysis. Likewise, the digital nature of maps provides data analysis tools that enable us to "see" geographic space in abstract terms (decision-making) beyond traditional descriptions of precise placement of physical features (inventory).

(Advanced Map Comparison Techniques)

Compare Maps by the Numbers

(GeoWorld, September 1999)

(return to top of Topic)

I bet you've seen and heard it a thousand times— a speaker waves a laser pointer at a couple of maps and says something like "see how similar the patterns are." But what determines similarity? A few similarly shaped gobs appearing in the same general area? Do all of the globs have to kind of align? Do display factors, such as color selection, number of classes and thematic break-points, affect perceived similarity? What about the areas that misalign?

Like abstract art, the patterns formed by map features are subject to subjectivity. The power of suggestion plays an important role and the old adage that "I wouldn't have seen it, if I hadn't believed it" often holds. Suggestion can influence map interpretation. But in visual map comparison, it dominates.

So how can we objectively assess map similarity? Like most things GIS, it's in the numbers. Consider the three maps on the right side of figure 1. Are they similar? Is the top map more similar to the middle one, than the map on the bottom? If so, how much more similar?

Actually the three maps were derived from the same map surface. Map1 identifies low response (lightest tone) as values below 65, medium as values in the range 65 through 70, and high as values over 70 (darkest tone). Map2 extends the mid-range to 62.5 through 72.5, while Map3 increases it even further to 60 through 75. In reality all three maps are supposed to be tracking the same spatial variable. But the categorized renderings appear radically different; or are they

surprisingly similar? What's your visceral vote?

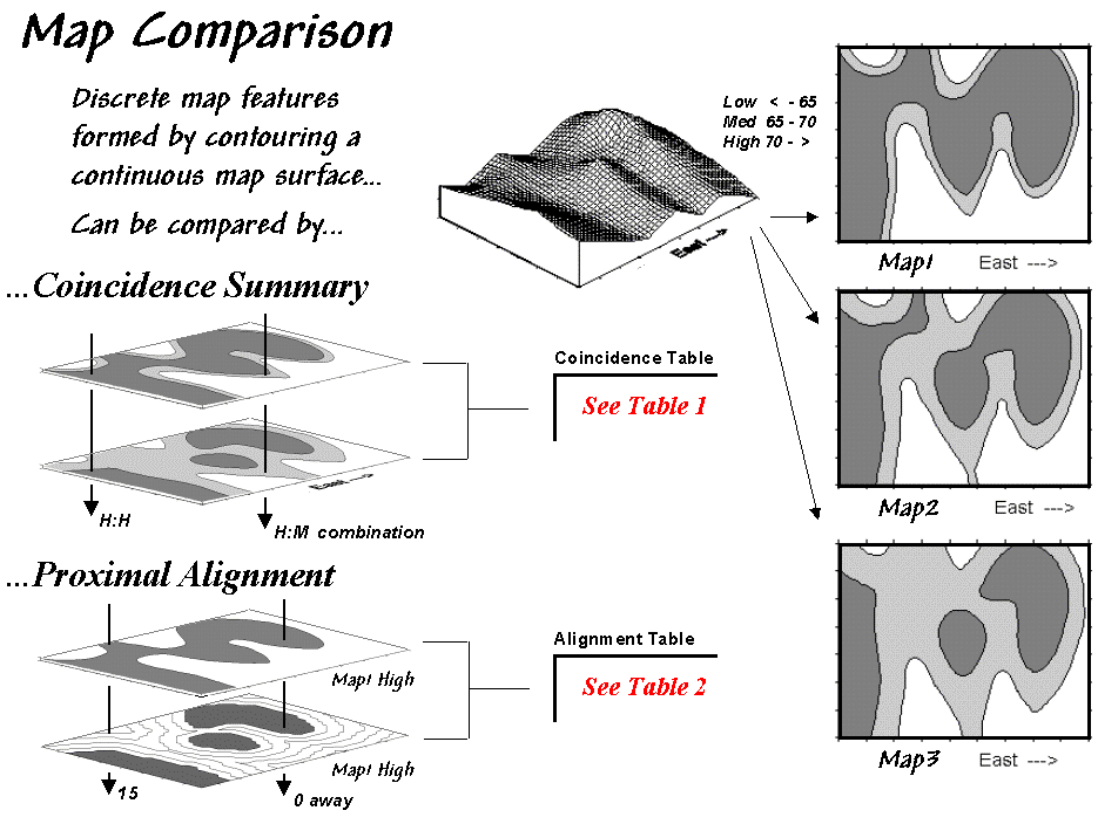


Figure 1. Coincidence Summary and Proximal Alignment can be used to assess the similarity between maps.

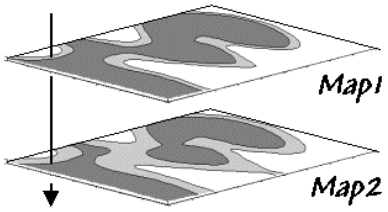
One way to find out for certain is to overlay the two maps and note where the classifications are the same and where they are different. At one extreme, the maps could perfectly coincide with the same conditions everywhere (identical). At the other extreme, the conditions might be different everywhere. Somewhere in between these extremes, the high and low areas could be swapped and the pattern inverted—similar but opposite.

Coincidence Summary generates a cross-tabular listing of the intersection of the two maps. In vector analysis the two maps can be "topologically" overlaid and the areas of the resulting son/daughter polygons aggregated by their joint condition. Another approach establishes a systematic or random set of points that uses a "point in polygon" overlay to identify/summarize the conditions on both maps.

Raster analysis uses a similar approach but simply counts the number of cells within each category combination as depicted by the arrows in the figure. In the example, a 39 by 50 grid was used to generate a comprehensive sample set of 1,950 locations (cells). Table 1 reports the coincidence summaries for the top map with the middle and bottom maps.

The highlighted counts along the diagonals of the table report the number of cells having the same classification on both maps. The off-diagonal counts indicate disagreement. The percent values in parentheses report relative coincidence.

The overall coincidence between Map1 and Map2 is 83%... fairly similar



		Map1			
Map2		Low	Medium	High	Totals
	Low	631	0	0	631 **(100%)**
	Medium	104	297	225	626 **(47%)**
	High	0	0	693	563 **(100%)**
	Totals	735 **(86%)**	297 **(100%)**	918 **(75%)**	1950 **(83%)**

		Map1			
Map3		Low	Medium	High	Totals
	Low	475	0	0	475 **(100%)**
	Medium	260	297	355	912 **(33%)**
	High	0	0	563	563 **(100%)**
	Totals	735 **(65%)**	297 **(100%)**	918 **(61%)**	1950 **(68%)**

The overall coincidence between Map1 and Map3 is 68%... not too similar

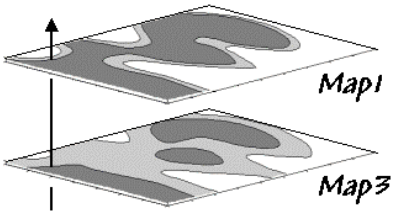


Table 1. Coincidence Summary.

For example, the 100% in the first row indicates that all of "Low" areas on Map2 coincide with "Low" areas on Map1. The 86% in the first column, however, notes that not all of the "Low" areas on Map1 are classified the same as the same as those on Map2. The lower portion of the table summarizes the coincidence between Map1 and Map2.

So what do all the numbers mean— in user-speak? First, the "overall" coincidence percentage in the lower right corner gives you a general idea of how well the maps match; 83% is fairly similar, while 68% is not too similar. Inspection of the individual percentages gives you a handle on which categories are, or are not lining-up. A perfect match would have 100% for each category; a complete mismatch would have 0%.

But simple coincidence summary just tells you whether things are the same or different. One extension considers the thematic difference. It notes the disparity in mismatched categories with a "Low/High" combination considered even less similar than a "Low/Medium" match.

Another procedure investigates the spatial difference, as shown in table 2. The technique, termed ***Proximal Alignment***, isolates one of the map categories (the dark-toned areas on Map3 in this case) then generates its proximity map. The proximity values are "masked" for the corresponding feature on the other map (enlarged dark-toned area on Map1 High). The highlighted area on the Masked Proximity Map identifies the locations of the greatest misalignment. Their relative occurrence is summarized in the lower portion of the tabular listing.

	Cell Count	Percent	Cumulative Percent
Align Exactly	*563*	**61.33**	--
Misalign	*355*	**38.67**	--
Within 1	*171*	**48.17**	**48.17**
2	*56*	**15.77**	**63.94**
3	*33*	**9.30**	**73.24**
4	*15*	**4.23**	**77.46**
5	*9*	**2,54**	**80.00**
6	*7*	**1.97**	**81.97**
7	*10*	**2.82**	**84.79**
8	*12*	**3.38**	**88.17**
9	*9*	**2.54**	**90.70**
10	*8*	**2.25**	**92.96**
11	*8*	**2.25**	**95.21**
12	*6*	**1.69**	**96.90**
13	*5*	**1.41**	**98.31**
14	*5*	**1.41**	**99.72**
15	*1*	**0.28**	**100.00**
Misalign Total	*355*	--	--

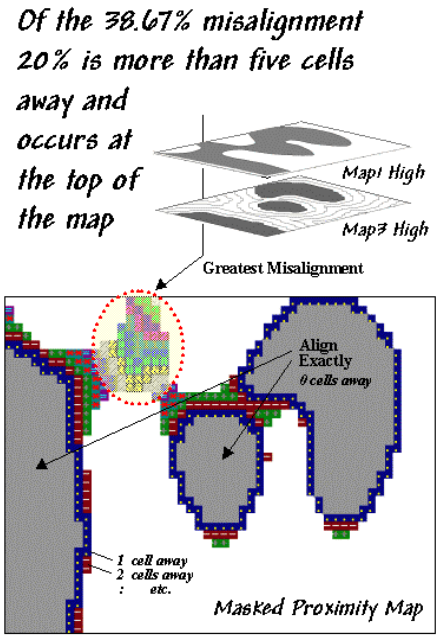


Table 2. Proximal Alignment.

So what does all this tell us— in user-speak? First, note that twenty percent of the total mismatches occurs more than five cells away from the nearest corresponding feature, thereby indicating fairly poor overall alignment. A simple measure of misalignment can be calculated by weight-averaging the proximity information—1*171 + 2*56 + … / 15 = 3.28. Perfect alignment would result in 0, with larger values indicating progressively more misalignment. Considering the dimensionality of the grid (39 x 50), a generalized proximal alignment index can be calculated — 3.28 / (39*50)**.5 = .074.

So what's the bottom line? If you're a GIS software provider, you should include a "big button" for comparing maps with your next release. If you're a GIS user, you should report coincidence percentages and alignment indices before passing comparative judgment. If you're a laser-waving presenter implying your visceral interpretation, your days are numbered.

Use Statistics to Compare Map Surfaces

(GeoWorld, October 1999)

(return to top of Topic)

While the human brain is good at lot of things, objective and detailed comparison among maps isn't one of them. Quantitative techniques provide a foothold for map comparison beyond waving a laser-pointer over a couple of maps and boldly stating "see how similar (or dissimilar) the patterns are."

The previous section identified a couple of techniques for comparing maps composed of discrete map "objects"— *Coincidence Summary* and *Proximal Alignment*. Comparing map "surfaces" involves similar approaches, but employs different techniques taking advantage of the more robust nature of continuous data.

Consider the two map surfaces shown on the left side of figure 1. Are they similar, or different? Where are they more similar or different? Where's the greatest difference? How would you know? In visual comparison, your eye looks back-and-forth between the two surfaces attempting to compare the relative "heights" at corresponding locations on the map. After about ten of these eye-flickers your patience wears thin and you form a hedged opinion— "not too similar."

In the computer, the relative "heights" are stored as individual map values (in this case, 1380 numbers in an analysis grid of 46 rows by 30 columns). One thought might be to use statistical tests to analyze whether the data sets are "significantly different."

Since map surfaces are just a bunch of spatially registered numbers, the sets of data can be compared by spatial coincidence (comparing corresponding values on two maps) and spatial partitioning (dividing the mapped data into subsets, then comparing the partitioned areas within one surface or between two surfaces).

In this approach, GIS is used to "package" the data and pass it to standard statistical procedures for analysis of differences within and between data groups. The packaging can be done in a variety of ways including systematic/random *sampling*, specified administration/management *zones*, or inferred spatial *groupings*. For example, a scientist could randomly sample values for terrain in two watersheds then test to see if their means are significantly different. Or a power company could investigate for significant change in household energy consumption for a neighborhood between two billing periods.

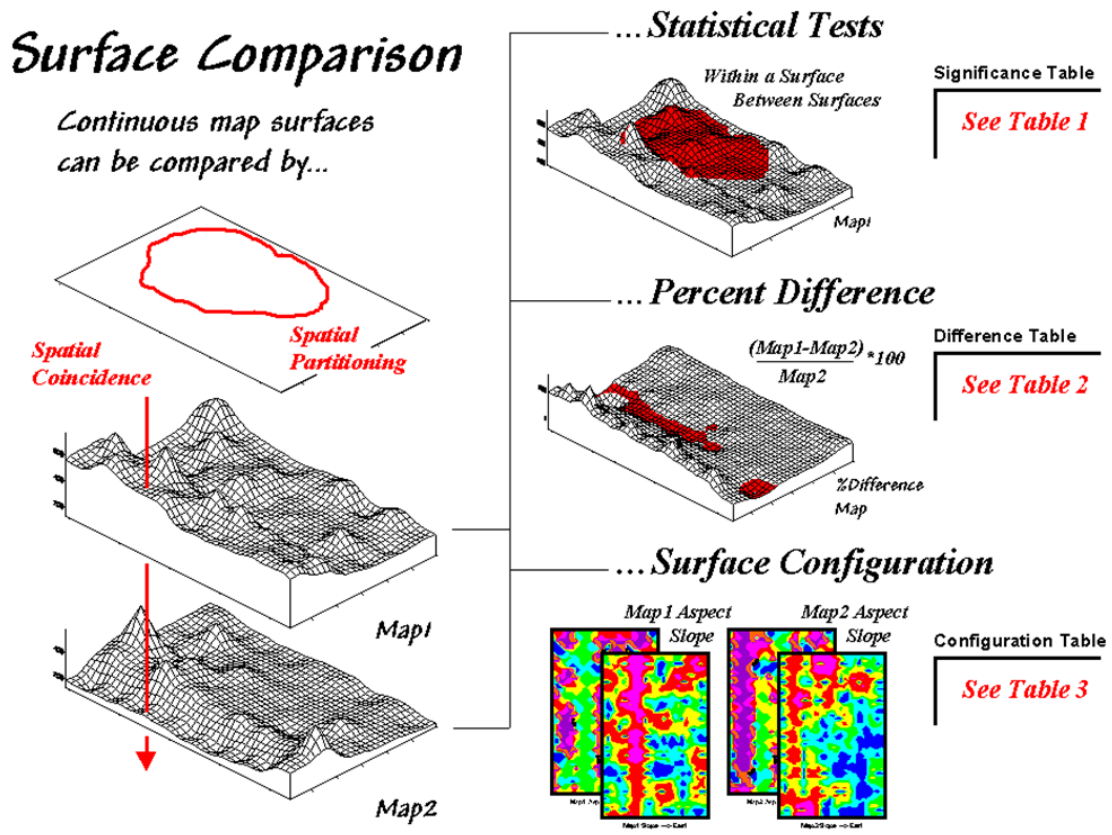
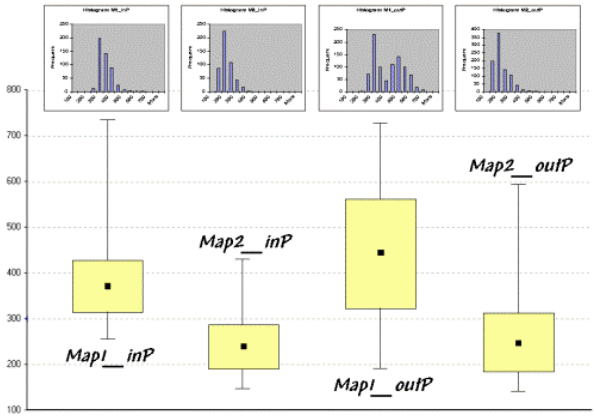


Figure 1. Map surfaces can be compared by statistically testing for significant differences in data sets, differences in spatial coincidence, or surface configuration alignment.

Or a farmer could test whether there is a significant difference in the topsoil versus substrata potassium levels for a portion of a field. Actually, this is the case depicted in figure 1 (Map1 = topsoil; Map2 = substrata potassium) and summarized in table 1. The dark red area on the surface locates the partitioned area in the field.



t-Test: Paired Two Sample for Means (Map1_inP vs. Map2_inP)		
	Map1_inP	Map2_inP
Minimum	269	152
Maximum	723	428
Mean	372.53	241.21
Variance	3827.93	2417.31
Standard Deviation	61.87	49.17
Observations	479	479
Pearson Correlation	0.72	
Hypoth. Mean Difference	0	
Degrees of Freedom	478	
t Stat	66.56	
P(T<=t) one-tail	3.3E-244	
t Critical one-tail	1.65	
P(T<=t) two-tail	6.7E-244	
t Critical two-tail	1.96	

...t-Test results using Excel

Table 1. Statistical Tests.

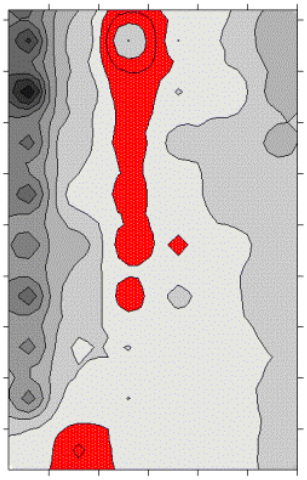
The "box-and-whisker" plots in the table identify the *mean* (dot), *+/- standard deviation* (shaded box) and *min/max* values (whiskers) for each of the four data sets (Maps1&2 and in&out of the Partition). Generally speaking, if the boxes tend to align there isn't much of a difference between data groups (e.g., *Map2_inP* and *Map2_outP* surfaces). If they don't align (e.g., *Map1_inP* and *Map2_inP* surfaces), there is a significant difference. The plots provide useful pictures of data distributions and allow you to eyeball the overall differences among a set of map surfaces.

The most commonly used statistical method for evaluating the differences in the means of two data groups is the *t*-test. The right side of table 1 shows the results of a *t*-test comparing the partitioned data between *Map1_inP* and *Map2_inP* (the first and second box-whisker plots).

While a full explanation of statistical tests is beyond the scope of this discussion, it is relative safe to say the larger the "*t stat*" value the greater the difference between two data groups. The values for the "one- and two-tail" tests at the bottom of the table suggest that "*the means of the two groups appear distinct and there is little chance that there is no difference between the groups*."

As with all things statistical, there are a lot of preconditions that need to be met before a *t*-test is appropriate— the data must be independent and normally distributed. The problem is that these conditions rarely hold for mapped data. While the *t*-test example might serve as a reasonable instance of "blindly applying" non-spatial, statistical tests to mapped data, it suggests this approach is a bit shaky as it seldom provides a reliable test like it does in traditional, non-spatial statistics (see Author's Notes).

In addition to data condition problems, statistical tests ignore the explicit spatial context of the data. Comparison using ***percent difference***, on the other hand, capitalizes on this additional information in map surfaces. Table 2 shows a categorized rendering and tabular summary of the percent difference between the Map1 and Map2 surfaces at each grid location. Note that the average difference is fairly large (76% +/- 49%), while two identical surfaces would compute to 0% average difference with +/- 0% standard deviation.



% Difference	# of Cells	% of Map
-99 to -67	1	0.07
-66 to -33	8	0.58
-32 to 0	17	1.23
0 to 32	128	9.28
33 to 66	580	47.03
67 to 99	363	26.16
100 to 132	116	8.41
133 to 166	67	4.86
167 to 199	61	4.42
200 to 232	24	1.74
233 to 266	8	0.58
267 to 299	6	0.43
300 to 332	1	0.07
76%	← Average % Difference	
49%	← Standard Deviation	

Table 2. Percent Difference.

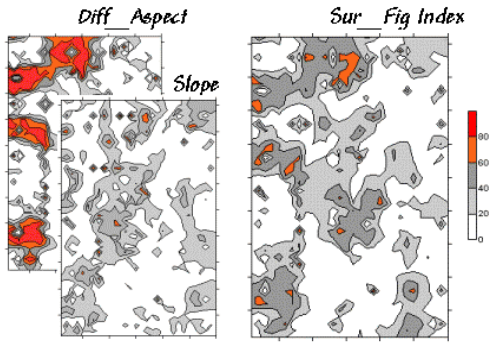
The dark red areas along the center crease of the map correspond to the highlighted rows in the table identifying areas within +/- 33 percent difference (moderate). That conjures up the "thirds rule of thumb" for comparing map surfaces—*if two-thirds of the map area is within one-third (33 percent) difference, the surfaces are fairly similar; if less than one-third of the area is within one-third difference, the surfaces are fairly different*—generally speaking that is. In this case only about 11% of the area meets the criteria so the surfaces are "considerably" different.

Another approach termed ***surface configuration***, focuses on the differences in the localized trends between two map surfaces instead of the individual values. Like you, the computer can "see" the bumps in the surfaces, but it does it with a couple of derived maps. A *slope* map indicates the relative steepness while an aspect map denotes the orientation of locations along the surface. You see a big bump; it sees an area with large slope values at several aspects. You see a ridge; it sees an area with large slope values at a single aspect.

So how does the computer see differences in the "lumpy-bumpy" configurations of two map surfaces? Per usual, it involves map-ematical analysis, but in this case some fairly ugly trigonometry is employed (see equations at end of chapter). Conceptually speaking, the immediate neighborhood around each grid location identifies a small plane with steepness and orientation defined by the slope and aspect maps. The mathematician simply solves for the normalized difference in slope and aspect angles between the two planes (see Author's Notes).

For the rest of us, it makes sense that locations with flat/vertical differences in inclination (Slope_Diff = 90^{o}) and diametrically opposed orientations (Aspect_Diff = 180^{o}) are as different as different can get. Zero differences for both, on the other hand, are as similar as things can get (exactly the same slope and aspect). All other slope/aspect differences fall somewhere in between on a scale of 0-100.

The two superimposed maps at the left side of table 3 show the normalized differences in the slope and aspect angles (dark red being very different). The map of the overall differences in surface configuration (Sur_Fig) is the average of the two maps. Note that over half of the map area is classified as low difference (0-20) suggesting that the "lumpy-bumpy" areas align fairly well overall. The greatest differences in surface configuration appear in the northwest portion.



Difference Level	Slope % of Map	Aspect % of Map	Sur_Fig % of Map
80-100 (High)	0.22	8.99	0.07
60-80	2.32	8.84	4.20
40-60	9.78	8.48	16.01
21-40	28.99	17.83	27.39
0-20 (Low)	58.70	55.87	52.32
Average	19.89	27.21	23.55
StDeviation	16.22	29.32	17.83

Table 3. Surface Configuration.

Does all this analysis square with your visual inspection of the *Map1* and *Map2* surfaces in figure 1? Sort of big differences in the relative values (surface height comparison summarized by *percent difference* analysis) with smaller differences in surface shape (bumpiness comparison summarized by *surface configuration* analysis). Or am I leading the "visually malleable" with quantitative analysis that lacks the comfort, artistry and subjective interpretation of laser-waving map comparison?

***Author's Notes**: An extended discussion by William Huber of Quantitative Decisions on the validity of statistical tests and an Excel workbook containing the equations and computations leading to the t-test, percent difference and surface configuration analyses are available online at the "Column Supplements" page at http://www.innovativegis.com/basis.*

(***Approaches Used in Deriving Prediction Maps***)

Use Scatterplots to Understand Map Correlation

(GeoWorld, November 1999)

(*return to top of Topic*)

A continuing theme of the Beyond Mapping columns has been that "*GIS maps are numbers first, pictures later*." The previous discussions have gone so far as to suggest that there are several ways to quantitatively compare spatial information— *Coincident Summary* and *Proximal Alignment* for discrete maps (September, 1999 BM column) and *Statistical Tests*, *Percent Difference* and *Surface Configuration* for continuous map surfaces (October, 1999 BM column). That brings us to the next big "bungy-jump" in map analysis involving correlation and predictive modeling.

In traditional statistics there is a wealth of procedures for investigating correlation, or "the relationship between variables." The most basic of these is the ***scatterplot*** that provides a graphical peek at the joint responses of paired data. It can be thought of as an extension of the histogram used to characterize the data distribution for a single variable.

For example, the *x*- and *y*-axes in figure 1 summarize the data described in the previous section. Recall that Map1 identifies the spatial distribution of potassium in the topsoil of a farmer's field, while Map2 tracks the concentrations in the root zone. Admittedly, this example is "a bit dirty" but keep in mind that a wide array of mapped data from resource managers to market forecasters can be used.

The histograms and descriptive statistics along the axes show the individual data distributions for the partitioned area (dark red "glob" draped on the map surfaces). It appears the topsoil concentrations in Map1 are generally higher (note the *positioning* of the histogram peaks; compare the means) and a bit more variable (note the *spread* of the histograms; compare the standard deviations). But what about the "joint response" considering both variables at the same time? Do higher concentrations in the root zone tend to occur with higher concentrations in the

topsoil? Or do they occur with lower concentrations? Or is there no discernible relationship at all?

These questions involve the concept of ***correlation*** that tracks *the extent that two variables are proportional to each other*. At one end of the scale, termed *positive correlation*, the variables act in unison and as values of one variable increase, the values for the other make similar increases. The other ends, termed negative correlation, the variables are mirrored with increasing values for one matched by decreases in the other. Both cases indicate a strong relationship between the variables just one is harmonious (positive) while the other is opposite (negative). In between the two lies *no correlation* without a discernable pattern between the changes in one variable and the other.

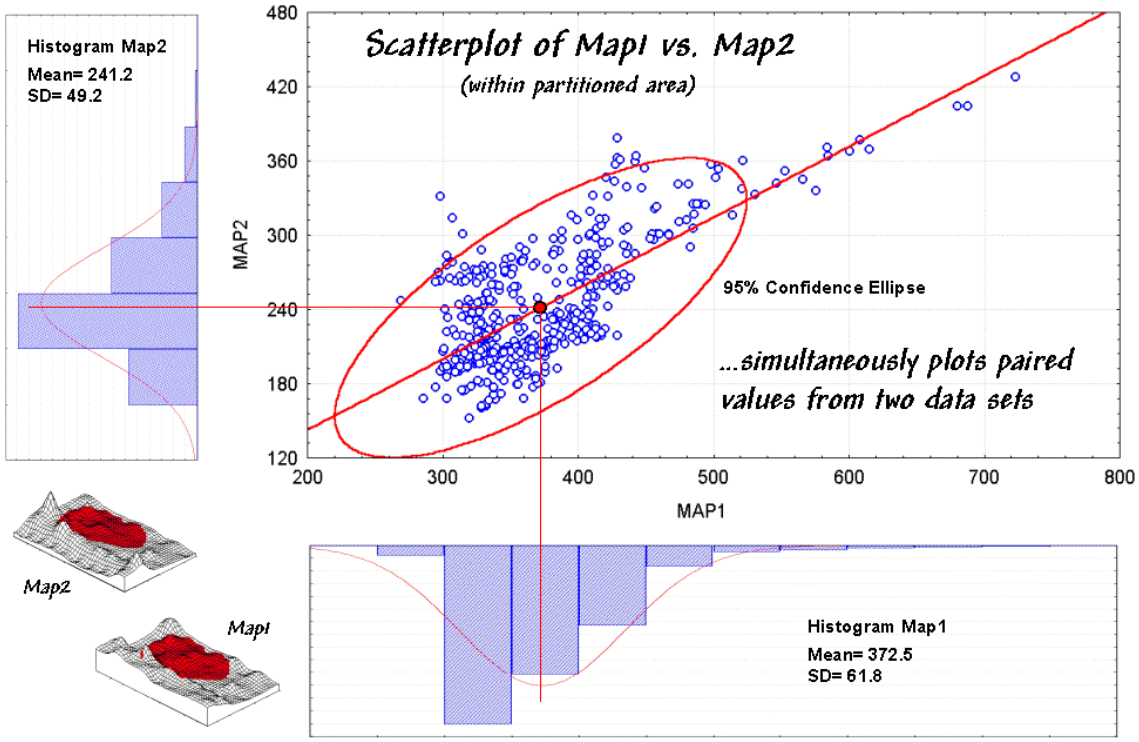


Figure 1. A scatterplot shows the relation between two variables by plotting the paired responses.

Now turn your attention to the scatterplot in figure 1. Each of the data points (small blue circle) represents one of the 479 grid locations within the partitioned area. The general pattern of the points provides insight into the joint relationship. If there is an upward linear trend in the data (like in the figure) positive correlation is indicated. If the points spread out in a downward fashion there's a negative correlation. If they seem to form a circular pattern or align parallel to either of the axes, a lack of correlation is noted.

Now let's apply some common sense and observations about a scatterplot. First the "strength" of a correlation can be interpreted by 1) the degree of alignment of the points with an upward (or downward) line and 2) how dispersed the points are around the line. In the example, there

appears to be fairly strong positive correlation (tightly clustered points along an upward line), particularly if you include the scattering of points along the right side of the diagonal.

But should you include them? Or are they simply "outliers" (abnormal, infrequent observations) that bias your view of the overall linear trend? Accounting for outliers is more art than science, but most approaches focus on the dispersion in the vicinity of the ***joint mean*** (i.e., statistical "balance point" of the data cloud). The joint mean in figure 30.3 is at the intersection of the lines extended from the Map1 and Map2 averages. Now concentrate on the bulk of points in this area and reassess the *alignment* and *dispersion*. Doesn't appear as strong, right?

A quantitative approach to identifying outliers involves a ***confidence ellipse***. It is conceptually similar to standard deviation as it identifies "typical" paired responses. In the figure, a 95% confidence ellipse is drawn indicating the area in the plot that accounts for the vast majority of the data. Points outside the ellipse are candidates for exclusion (25 of 479 in this case) in hopes of concentrating on the overall trend in the data. The orientation of the ellipse helps you visualize the linear *alignment* and its thickness helps you visualize the *dispersion* (pretty good on both counts).

In addition to assessing *alignment*, *dispersion* and *outliers* you should look for a couple of other conditions in a scatterplot— distinct groups and nonlinear trends. ***Distinct group*** bias can result in a high correlation that is entirely due to the arrangement of separate data "clouds" and not the true relationships between the variables within each data group. ***Nonlinear trends*** tend to show low "linear" correlation but actually exhibit strong curvilinear relationships (i.e., tightly clustered about a bending line). Neither of these biases is apparent in the example data.

Now concentrate on the linkage between the scatterplot and the map surfaces. The analysis grid structures the linkage and enables you to "walk" between the maps and the plot. If you click on a point in the scatterplot its corresponding cell location on both surfaces are highlighted. If you click on a location on one of the maps its scatter plot point is highlighted.

That's set the stage for interactive data analysis. One might click on all of the outlier points and see if they are scattered or grouped. If they tend to form groups there is a good chance a geographic explanation exists— possibly explained by another data layer.

Another investigative procedure is to delineate sets of points on the scatterplot that appear to form "fuzzy globs." The globs indicate similar characteristics (data pattern) while the map plays out their spatial pattern. In a sense, manually delineating data globs is analogous to the high-tech, quantitative procedure termed *data clustering* (see Author's Notes). In fact quantitative expression of the scatterplot's correlation information forms the basis for predictive modeling…but that's next month's story.

Author's Notes*: An extended discussion of data grouping and a online version of "Identifying Data Patterns" (GIS World, August 1998) are available online at the "Column Supplements" page at http://www.innovativegis.com/basis.*

Can Predictable Maps Work for You?

The last section discussed map correlation as viewed through a scatterplot. Recall that the orientation of the "*data cloud*" indicated the nature of the relationship between the values on two map surfaces, while its shape showed the strength of the relationship.

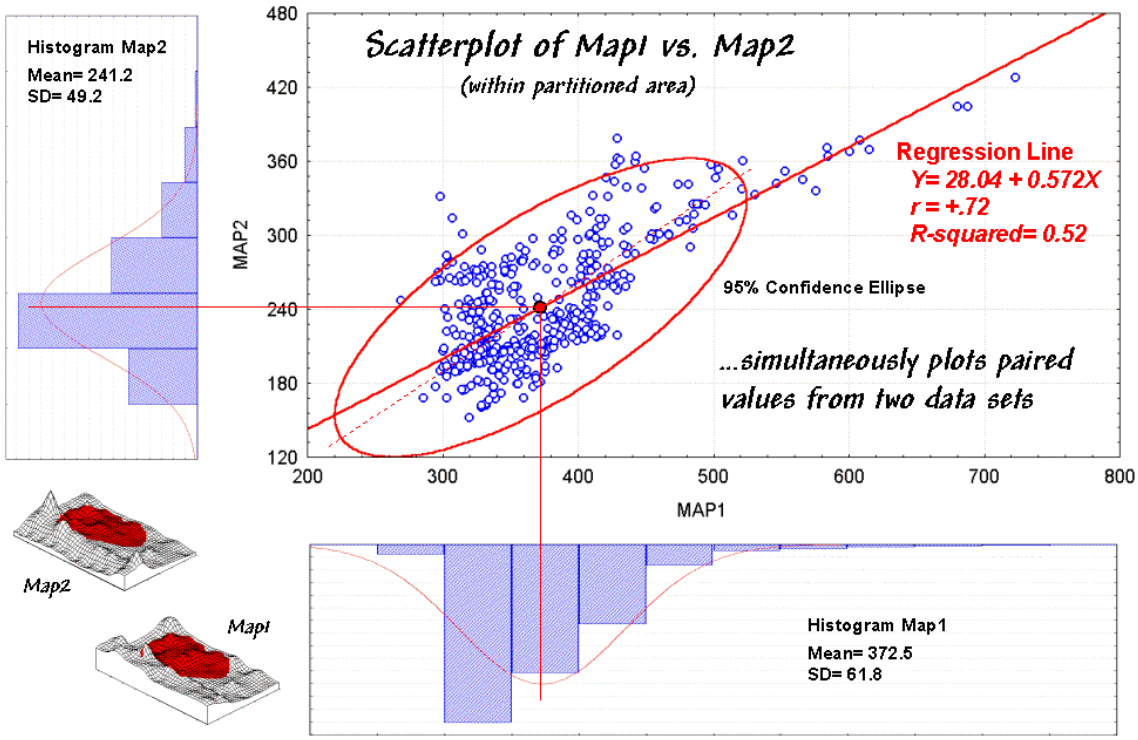


Figure 1. Scatterplot with correlation and regression information identified.

Figure 1 should rekindle the concepts, but note the addition of the information about the regression line. That brings us to the tough (more interesting?) tuff— quantitative measures of correlation and predictive modeling.

While a full treatise of the subject awaits your acceptance to graduate school, discussion of some basic measures might be helpful. The ***correlation coefficient*** (denoted as "*r*") represents the linear relationship between two variables. It can range from +1.00 (perfect positive correlation) to -1.00 (perfect negative correlation), with a value of 0.00 representing no correlation.

Calculating "*r*" involves finding the "best-fitting line" that minimizes the sum of the squares of distances from each data point to the line, then summarizing the deviations to a single value. When the correlation coefficient is squared (referred to as the "***R-squared***" value), it identifies the "proportion of common variation" and measures the overall strength of the relationship.

The examples in figure 2 match several scatterplots with their R-squared values. The inset on the left shows four scatterplots with increasing correlation (tighter linear alignment in the data clouds). The middle inset depicts data forming two separate sub-groups (***distinct group*** bias). In

this instance the high R-squared of .81 is misleading. When the data groups are analyzed separately, the individual R-squared values are much weaker (.00 and .04).

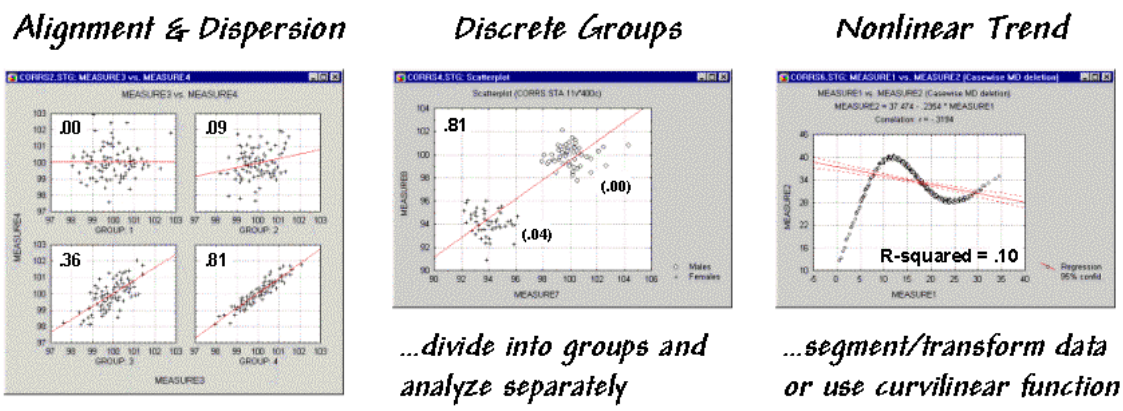


Figure 2. Example scatterplots depicting different data relationships.

The inset on the right is an example of a data pattern that exhibits low linear correlation, but has a strong curvilinear relationship (***nonlinear trend***). Unfortunately, dealing with nonlinear patterns is difficult even for the statistically adept. If the curve is continuously increasing or decreasing, you might try a logarithmic transform. If you can identify the specific function, use it as the line to fit. Or, if all else fails, break the curve into segments that are better approximated by a series of straight lines.

Regression analysis extends the concept of correlation by focusing on the best-fitted line (solid red lines in the examples). The equation of the line generalizes the linear trend in the data. Also, it serves as a predictive model, while its correlation indicates how good the data fit the model. In the case of figure 1, the regression equation is Map2(*estimated*) = 28.04 + 0.572 * Map1, with an R-squared value of .52. That means if you measure a potassium level of 500 in the topsoil expect to find about 314 in the root zone (28.04 + 0.572 * 500 = 314.04). But how good is that guess in the real world?

One way to evaluate the model is to "play-it-back" using the original data. The left-side of figure 3 shows the results for the partitioned area. As hoped, the predicted surface is very similar to the actual data with an average error of only 2% and over 98% of the area within 33% difference.

Model validation involves testing it on another set of data. When applied outside the partition the regression model didn't fair as well— an average error of 19% and only 25% of the area within 33% difference. The difference surface shows you that the model is pretty good in most places but really blows it along the western edge (big ridge of over estimation) and part of the northern edge (big depression of under estimation). Maybe those areas should be partitioned and separate prediction models developed for them? Or, more likely your patience has ebbed.

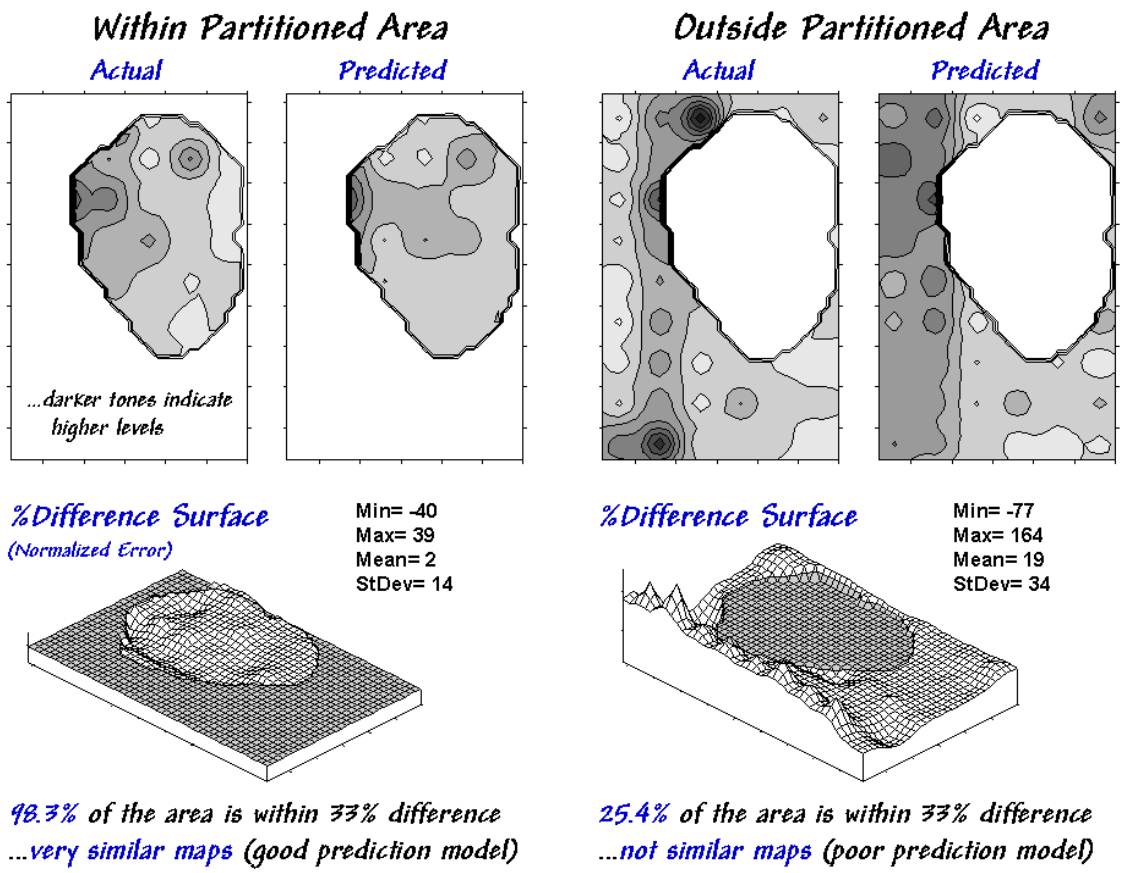


Figure 3. Results of applying the predictive model.

A few final concepts should wrap things up. First, data analysis rarely uses raw data. As discussed last month "outliers" are identified and eliminated. In figure 3, the dotted axis through the confidence ellipse suggests a somewhat steeper regression line "better fits" the bulk of the data. Possibly this equation is a better predictor.

Some data analysts use a "roving window" (e.g., values in a 3x3 adjacent neighborhood) to derive a neighborhood-weighted average for each grid location before deriving a prediction model. This "smoothing" addresses slight misalignment in the data layers and salt-and-pepper conditions in some data sets. Another school of thought suggests sampling the data such that the distance between the samples is larger than the spatial autocorrelation as determined by variogram analysis (see "Uncovering the Mysteries of Spatial Autocorrelation," GIS World, June, 1997, page ??). The sampling interval addresses concern for dependence in the data.

For most of us, however, the bottom-line lies not in debatable statistical theory but in the results. Regardless of technique, if model validation yields predictions are better than current guesses, then "predictable maps" could work for you.

***Author's Notes**: An Excel workbook extending this discussion to segmented and localized regression is available online at the "Column Supplements" page at http://www.innovativegis.com/basis.*

Using trigonometric relationships to establish differences in surface configuration

Note: Data preparation was completed in MapCalc using an analysis grid configured as 46 rows by 30 rows (1380 map values). Slope (rate of change) and Aspect (direction of change) maps were derived for both the top and bottom soil potassium maps.

Calculate the "normalized" difference in slope:

Most grid-based GIS packages calculate % slope.
Percent slope can be converted to degrees slope by ***DEGREES(ARCTAN(%slope/100)****.*
The difference between the two slopes is obtained by ***ABS(M1_DegSlope - M2_DegSlope)****.*
The difference in slope angles can be normalized between 0 to 100 by
(((Diff_DegSlope - min) * 100) / (max - min))
where min = 0 and max = 90 for degree_slope possible range.

Calculate the "normalized" difference in azimuth:

Most grid-based GIS packages calculate precise aspect in degrees azimuth.
Degrees azimuth must be converted to radians by **RADIANS (Deg_Azimuth)**.
The difference between two azimuths can be calculated in degrees by
DEGREES(ACOS(SIN(Map1_RadAzimuth) * SIN(Map2_RadAzimuth) +
COS(Map1_RadAzimuth) * COS(Map2_RadAzimuth)))

Normalized between 0 to 100 by **(((Diff_DegAzimuth - min) * 100) / (max - min))**
where *min = 0 and max = 180 for degree_azimuth possible range.*

Spatial Data Mining Allows Users to Predict Maps

(GeoWorld, January 2002)

(return to top of Topic)

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

Extending predictive analysis to mapped data seems logical. After all, maps are just organized sets of numbers. And GIS 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— "descriptive techniques" for assessing *map similarity*, *classification units* and *data zones*. The next logical step is to apply "predictive techniques" to generate extrapolative maps.

If fact, the first time I used prediction mapping was in 1991 to extend a test market project for a phone company. The customer's address was used to geo-code sales of a new product that enabled two numbers with distinctly different rings to be assigned to a single phone— one for the kids and one for the parents. Like pushpins on a map, the pattern of sales throughout the city emerged with some areas doing very well, while in other areas sales were few and far between.

The demographic data for the city was analyzed to calculate a prediction equation between product sales and census block data. The prediction equation derived from the test market sales in one city was applied to another city by evaluating exiting demographics to "solve the equation" for a predicted sales map. In turn the predicted map was combined with a wire-exchange map to identify switching facilities that required upgrading before release of the product in the new city.

To illustrate the data mining procedure, the approach can be applied to the cornfield data that has been focus for the past several sections. The top portion of figure 1 shows the yield pattern for the field varying from a low of 39 bushels per acre (red) to a high of 279 (green). Corn yield, like "sales yield," is termed the *dependent map variable* and identifies the phenomena we want to predict.

The *independent map variables* depicted in the bottom portion of the figure are used to uncover the spatial relationship— ***prediction equation***. In this instance, digital aerial imagery will be used to explain the corn yield patterns. The map on the left indicates the relative reflectance of red light off the plant canopy while the map on the right shows the near-infrared response (a form of light just beyond what we can see).

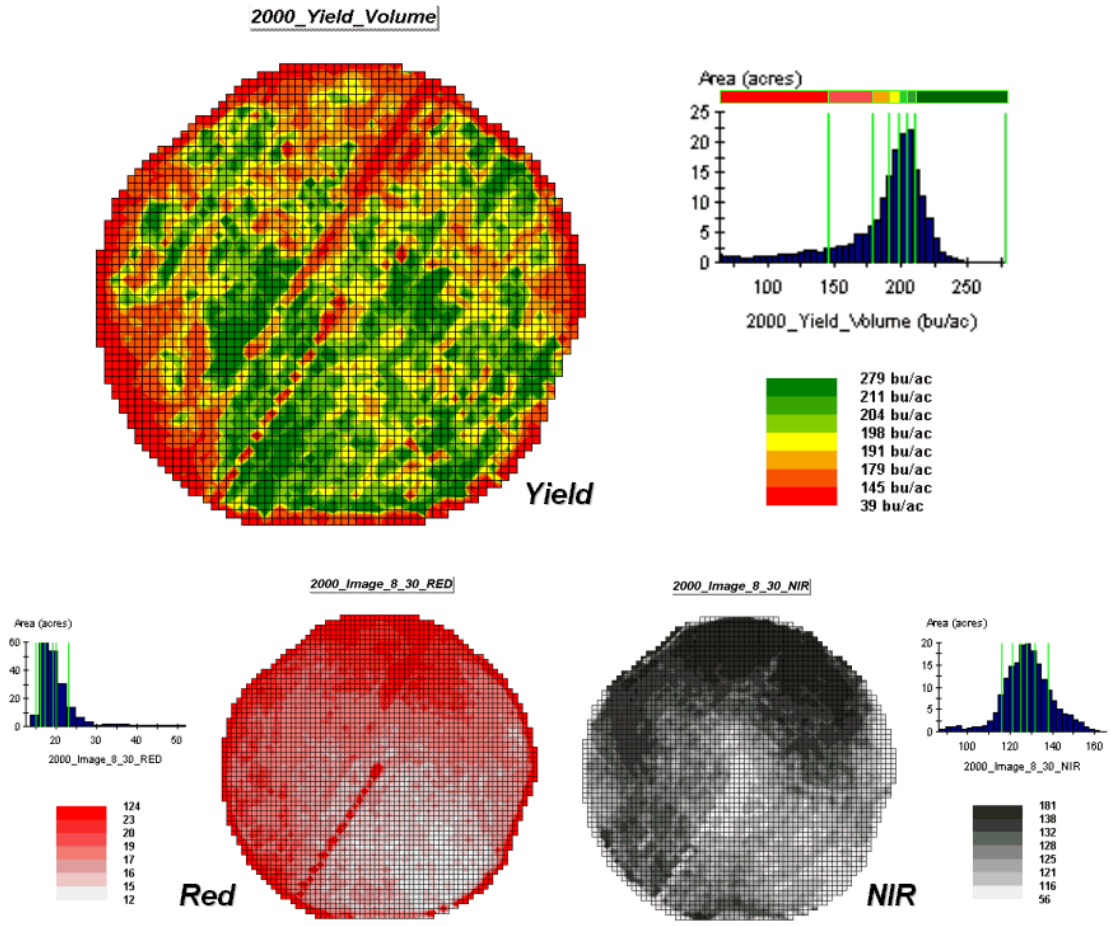


Figure 1. The corn yield map (top) identifies the pattern to predict; the red and near-infrared maps (bottom) are used to build the spatial relationship.

While it is difficult for you to assess the subtle relationships between corn yield and the red and near-infrared images, the computer "sees" the relationship quantitatively. Each grid location in the analysis frame has a value for each of the map layers— 3,287 values defining each geo-registered map covering the 189-acre field.

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

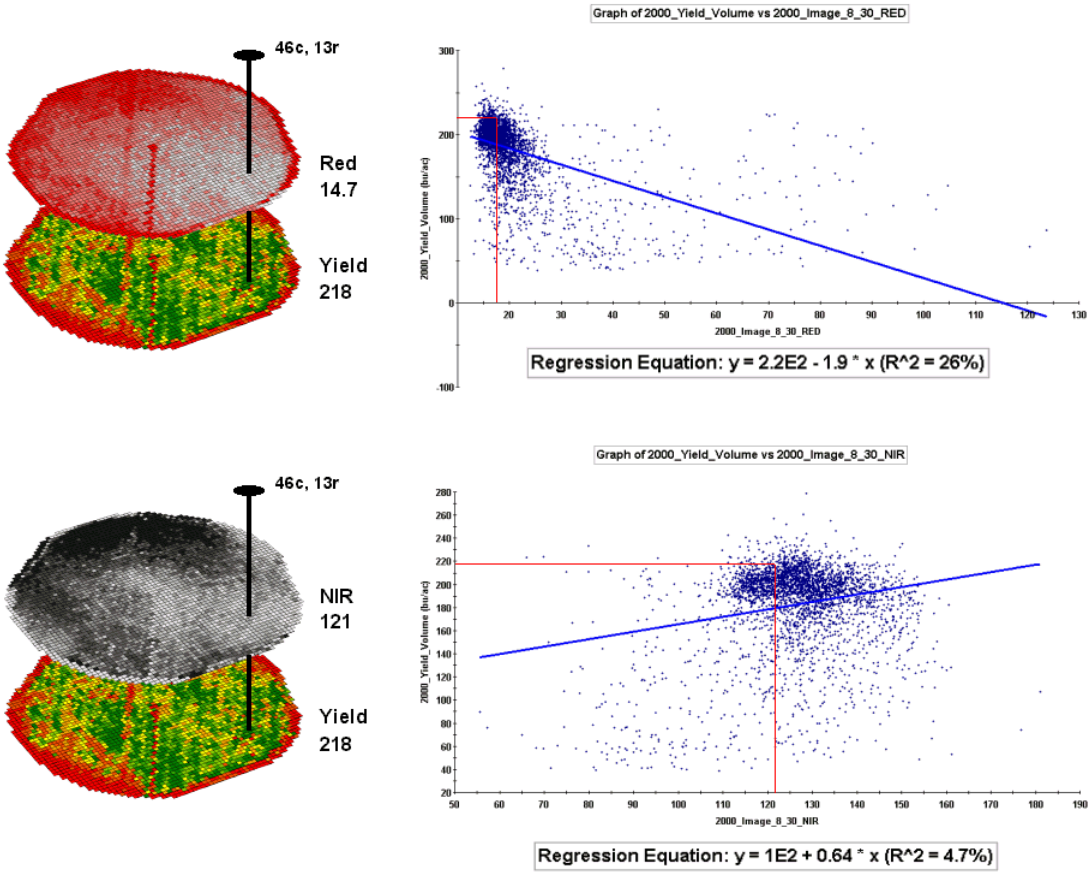


Figure 2. The joint conditions for the spectral response and corn yield maps are summarized in the scatter plots shown on the right.

In fact the set of "blue dots" in both of the scatter plots represents data pairs for each grid location. The blue lines in the plots represent the prediction equations derived through regression analysis. While the mathematics is a bit complex, the effect is to identify a line that "best fits the data"— just as many data points above as below the line.

In a sense, the line sort of identifies the average yield for each step along the X-axis (red and near-infrared responses respectively). Come to think of it, wouldn't that make a reasonable guess of the yield for each level of spectral response? That's how a regression prediction is used… a value for red (or near-infrared) in another field is entered and the equation for the line is used to predict corn yield. Repeat for all of the locations in the field and you have a prediction map of yield from an aerial image… but alas, if it were only that simple and exacting.

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

Figure 3 shows the process for calculating NDVI for the sample grid location— ((121-14.7) / (121 + 14.7))= 106.3 / 135.7= .783. The scatter plot on the right shows the yield versus NDVI plot and regression line for all of the field locations. Note that the R^^2 value is a higher at 30% indicating that the combined index is a better predictor of yield.

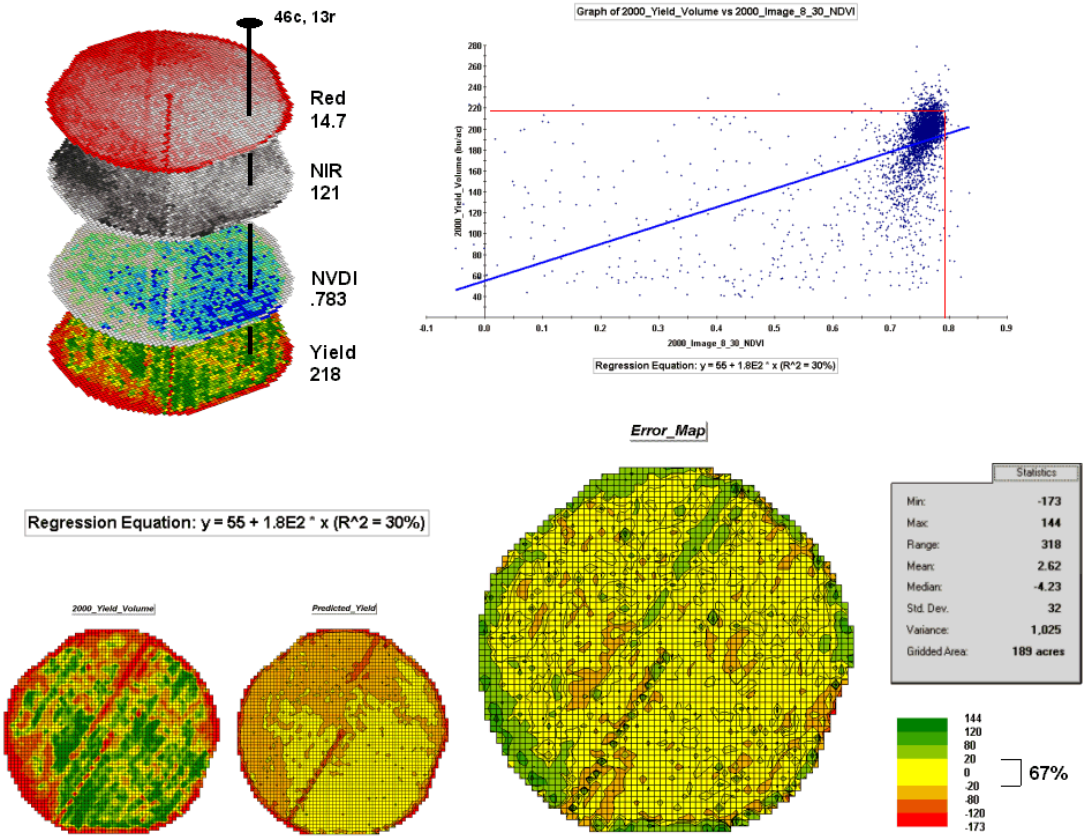


Figure 3. The red and NIR maps are combined for NDVI value that is a better predictor of yield.

The bottom portion of the figure evaluates the prediction equation's performance over the field. The two smaller maps show the actual yield (left) and predicted yield (right). As you would

expect the prediction map doesn't contain the extreme high and low values actually measured. However the larger map on the right calculates the error of the estimates by simply subtracting the actual measurement from the predicted value at each map location.

The error map suggests that overall the yield "guesses" aren't too bad— average error is a 2.62 bu/ac over guess; 67% of the field is within 20 bu/ac. Also note that most of the over estimating occurs along the edge of the field while most of the under estimating is scattered along curious NE-SW bands.

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

Stratify Maps to Make Better Predictions

(GeoWorld, February 2002)

(return to top of Topic)

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

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

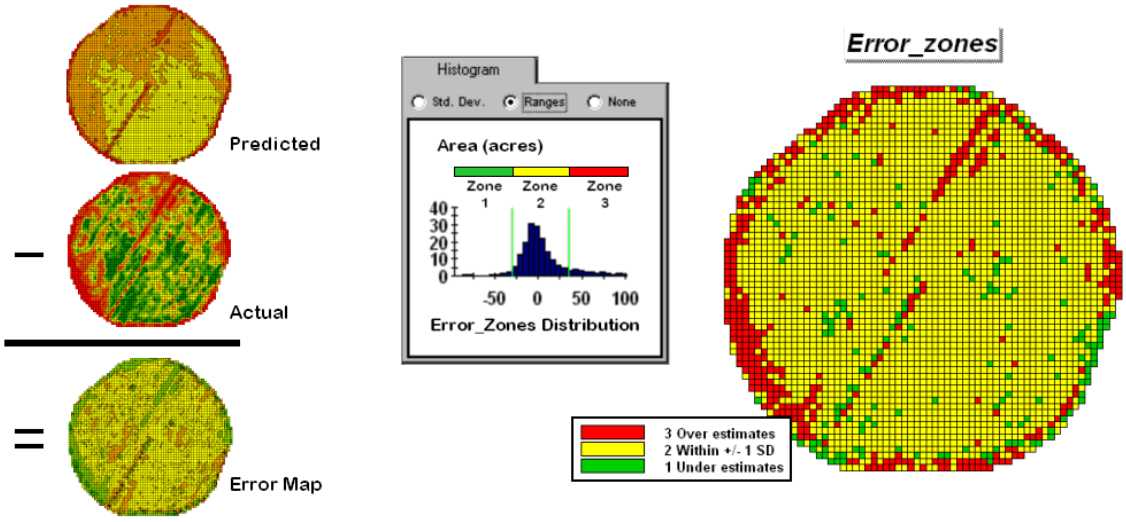


Figure 1. A field can be stratified based on prediction errors.

The left side of figure 1 shows the evaluation procedure. Subtracting the actual yield values from the predicted ones for each map location derives an Error Map. The previous discussions noted that the yield "guesses" weren't too bad— average error of 2.62 bu/ac with 67% of the estimates within 20 bu/ac of the actual yield. However, some locations were as far off as 144 bu/ac (over-guess) and –173 bu/ac (under-guess).

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

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

The rationale behind the stratification is that the whole-field prediction equation works fairly well for zone 2 but not so well for zones 1 and 3. The assumption is that conditions within zone 1 make the equation under estimate while conditions within zone 3 cause it to overestimate. If the assumption holds, one would expect a tailored equation for each zone would be better at predicting than an overall equation. Figure 2 summarizes the results of deriving and applying a set of three prediction equations.

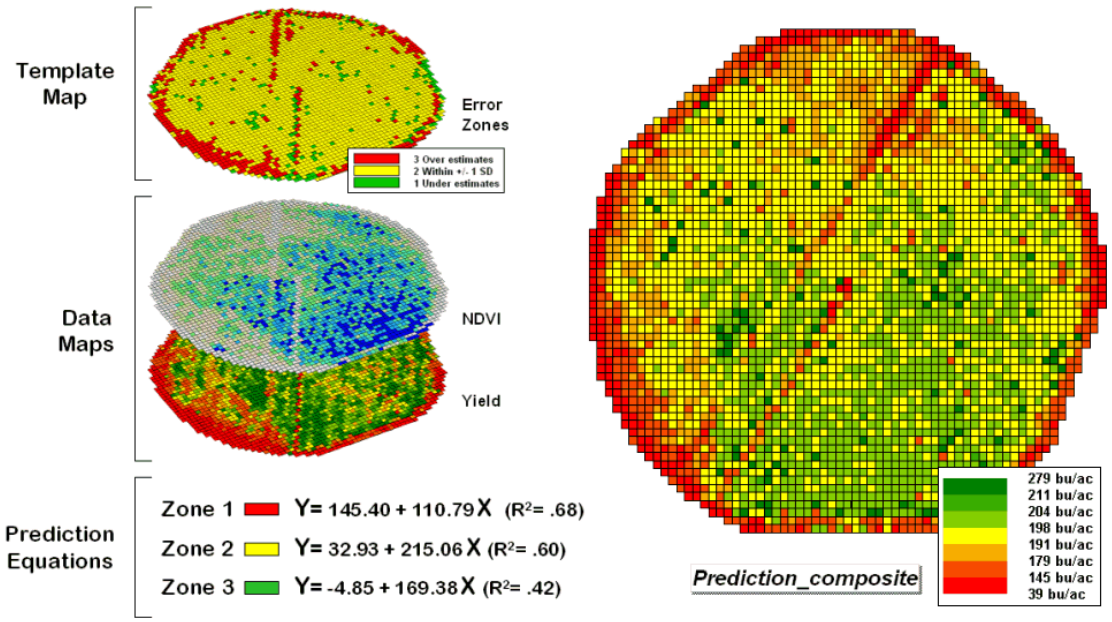


Figure 2. After stratification, prediction equations can be derived for each element.

The left side of the figure illustrates the procedure. The Error Zones map is used as a template to identify the NDVI and Yield values used to calculate three separate prediction equations. For each map location, the algorithm first checks the value on the Error Zones map then sends the data to the appropriate group for analysis. Once the data has been grouped a regression equation is generated for each zone. The "r-squared" statistic for all three equations (.68, .60, and .42 respectively) suggests that the equations fit the data fairly well and ought to be good predictors.

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

The left side of figure 3 provides a visual comparison between the actual yield and predicted maps. The "stratified prediction" shows detailed estimates that more closely align with the actual yield pattern than the "whole-field" derived prediction map.

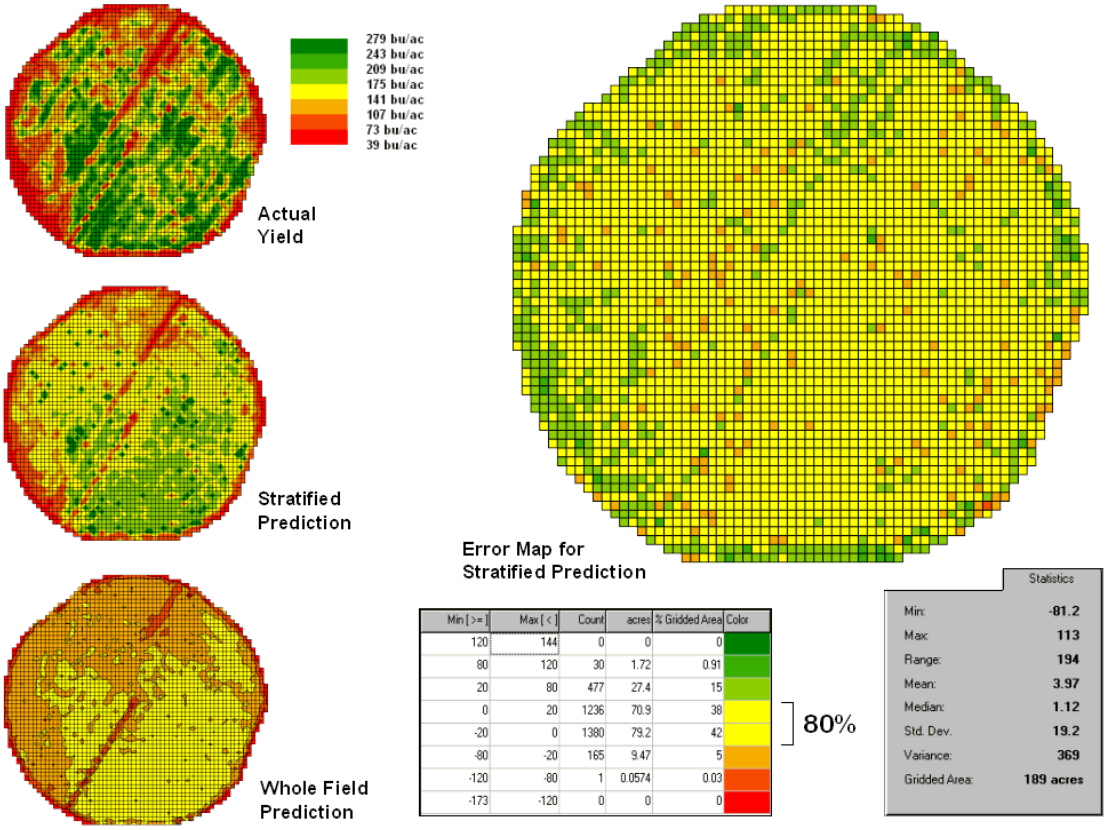


Figure 3. Stratified and whole-field predictions can be compared using statistical techniques.

The error map for the stratified prediction shows that eighty percent of the estimates are within +/- 20 bushels per acre. The average error is only 4 bu/ac with maximum under- and over-estimates of –81.2 and 113, respectively. All in all, not bad guessing of yield based on a remote sensing shot of the field nearly a month before the field was harvested.

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

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

Finally, one needs to recognize that spatial data mining is not restricted to precision agriculture but has potential for analyzing relationships within almost any set of mapped data. For example,

prediction models can be developed for geo-coded sales from demographic data or timber production estimates from soil/terrain patterns. The bottom line is that maps are increasingly seen as organized sets of data that can be map-ematically analyzed for spatial relationships… we have only scratched the surface.

(return to top of Topic)

(Back to the Table of Contents)