Topic 5

Characterizing Data Groups

5.1 Calculating Field Similarity

How often have you seen a precision agriculture presenter "lasso" a portion of a map with a laser pointer and boldly state "…see how similar this area is to the locations over here and here…" as the pointer rapidly moves about the map. More often than not, there is a series of side-by-side maps serving as the background scenery for the laser show.

But just how similar is one field location to another? Really similar or just a little bit similar? And just how dissimilar are all of the other areas? While visceral analysis can identify broad relationships it takes a quantitative map analysis approach to handle the detailed scrutiny demanded by site-specific management.

Consider the three maps shown in figure 5-1— 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?

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 data in the example shows the relative amounts of phosphorous, potassium and nitrogen distributions throughout the case study field.

In visual analysis you move your focus among the maps to summarize the color assignments at different locations. The difficulty in this approach 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 colors. In addition, the computer doesn't tire as easily as you and completes the comparison for all of the locations throughout the map window (3,289 grid cells in this example) in a couple seconds.

The upper-left portion of figure 5-2 illustrates capturing the data patterns of two locations for comparison. The "data spear" at map location 45column, 18row identifies the P-level as 11.0, the K-level as 177.0 and N-level as 32.9ppm. This step is analogous to your eye noting a color pattern of burnt-red, dark-orange and light-green. The other location for comparison (32c, 62r) has a data pattern of P= 53.2, K= 412.0 and N= 27.9 …or as your eye sees it, a color pattern of dark-green, dark-green and yellow.

The right side of figure 5-2 conceptually depicts how the computer calculates a similarity value for the two response patterns. The realization that mapped data can be expressed in both geographic space and data space is key to understanding quantitative analysis similarity.
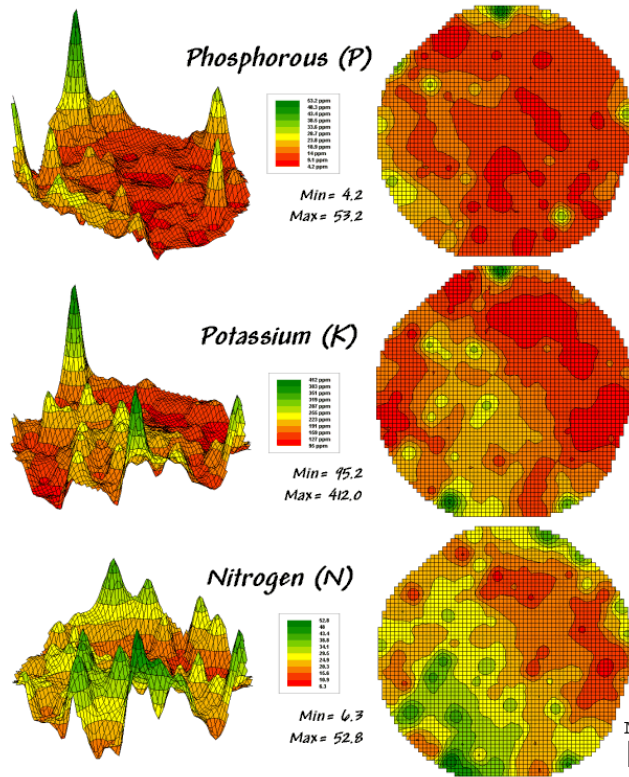

Figure 5-1. Map surfaces identifying the spatial distribution of P,K and N throughout a field.

Geographic space uses coordinates, such latitude and longitude, to locate things in the real world—such as the southeast and extreme north points identified in the example field. The geographic expression of the complete set of measurements depicts their spatial distribution in familiar map form.
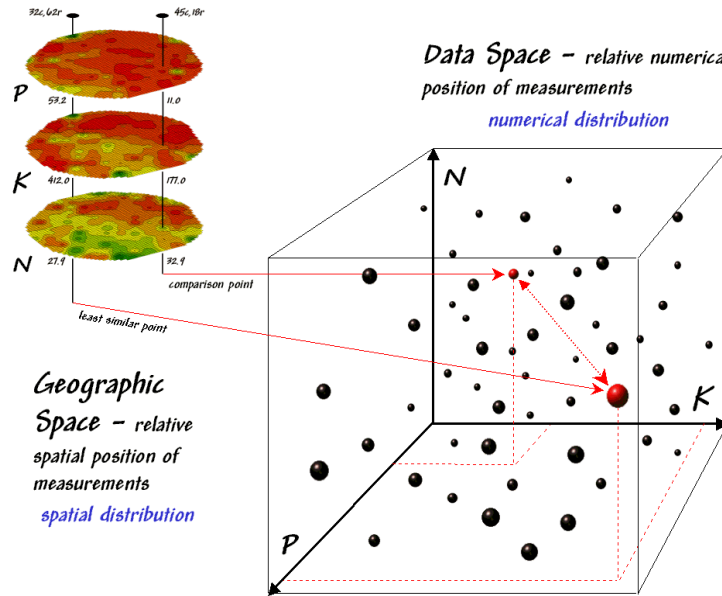

Figure 5-2. Conceptually linking geographic space and data space.

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 the P, K and N levels measured in the field. The floating balls represent grid locations defining the geographic space—one for each grid cell. The coordinates locating the floating balls extend from the data axes—11.0, 177.0 and 32.9 for the comparison point. The other point has considerably higher values in P and K with slightly lower N (values 53.2, 412.0, 27.9 respectively) so it plots at a different location in data space.

while high, high, high is 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.

In the example, the floating ball closest to you is least similar (greatest 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 that equates to the highest similarity value (100% similar).

The similarity map shown in figure 5-3 applies the similarity scale to the data distances calculated between the comparison point and all of the other points in data space. The green tones indicate field locations with fairly similar P, K and N levels to the comparison location in the field—with the darkest green identifying locations with identical P, K and N levels (100% similar). It is interesting to note that most of the very similar locations are in the western portion of the field. The light-green to red tones indicate increasing dissimilar areas.
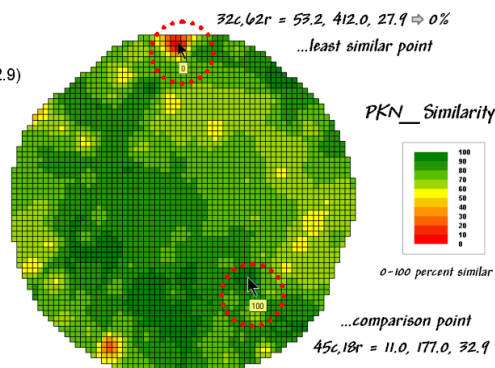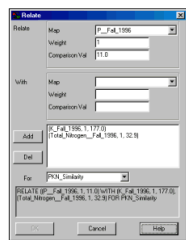
Figure 5-3. A similarity map identifies how related locations are to a given point.

The bottom line for data space analysis is that the position of a point identifies its numerical pattern—low, low, low in the back-left corner,

A similarity map can be an invaluable tool for investigating spatial patterns in any complex set of mapped data. The calculation of similarity can handle any number of input maps, yet humans are unable to conceptualize more than three variables (the data space box). Also,

the different map layers can be weighted to reflect relative importance in determining overall similarity.

In effect, a similarity map replaces a lot of laser-pointer waving and subjective suggestions of similar/dissimilar areas with a concrete, quantitative measurement at each map location. The technique moves map analysis well beyond the old *"…I'd never have seen, it if I hadn't believed it…"* mode of visual map interpretation.

5.2 Identifying Data Zones

Last section introduced the concept of **Data Distance** as a means to measure similarity within a map. One simply mouse-clicks 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 map displays with an exact quantitative measure 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 to the area of interest.

The link between **Geographic Space** and **Data Space** is key. As shown in figure 5-4, spatial data can be viewed as a map or a histogram. While a map shows us "where is what," a histogram summarizes "how often"

measurements occur (regardless where they occur).

The top-left portion of the figure shows a 2D/3D map display of the relative amount of phosphorous (P) within a farmer's field. Note the spikes of high measurements along the edge of the field, with a particularly big spike in the north portion.

The histogram to the right of the map display depicts a different perspective of the data. Rather than positioning the measurements in geographic space it summarizes their relative frequency of occurrence in data space. The X-axis of the graph corresponds to the Z-axis of the map—amount of phosphorous. In this case, the spikes in the graph indicate measurements that occur more frequently. Note the high occurrence of phosphorous values around 11ppm.
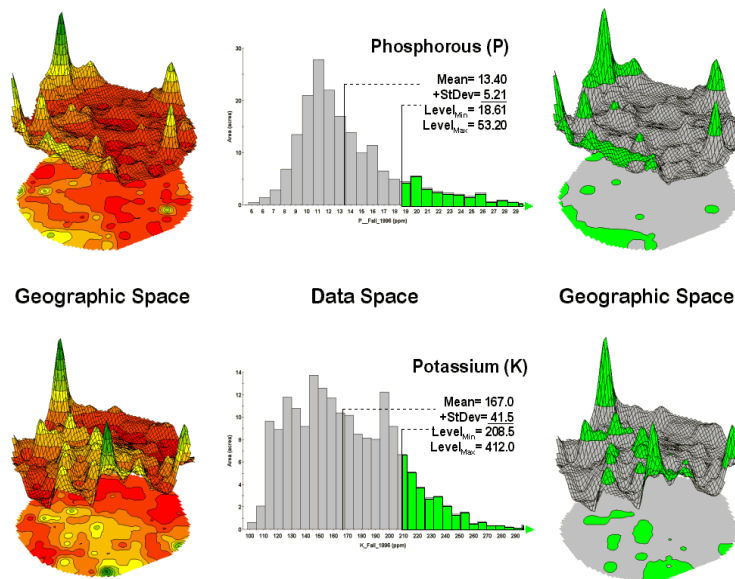

Figure 5-4. Identifying areas of unusually high measurements.

The top-right portion of the figure puts the data and geographic space link to use. The shaded area in the histogram view identifies field phosphorous values that are unusually high—more than one standard deviation above the mean. This statistical cutoff is used to isolate locations of high concentrations as shown in the map on the right. The procedure is repeated for the potassium (K) map surface to identify its locations of unusually high measurements.

Figure 5-5 illustrates combining the P and K data to locate areas in the field that have high measurements in both. The graphic on the left is termed a scatter diagram or plot. It graphically summarizes 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 plot identifies the P and K measurements at that location. The balls plotting in the shaded area of the diagram identify field locations that have both high P and high K. The upper-left partition identifies joint conditions

in which neither P nor K are high. The off-diagonal partitions in the scatter plot identify locations that are high in one nutrient but not the other.

The aligned maps in the right inset show the geographic solution for areas that are high in both of the soil nutrients. A simple map-*ematical* way to generate the solution is to assign 1 to all locations of high values in the P and K map layers (bight green). Zero is assigned to locations that aren't high (light gray). When the two binary maps (0/1) are multiplied, a zero on either map computes to zero.

Figure 5-5 Identifying joint coincidence in both data and geographic space.

However a map-*ematical* trick was employed to get the detailed map solution shown in the figure. On the individual maps, high areas were set to P= 1, K= 2 and N= 4, then the maps were added together.

The result is a range of coincidence values from zero (0+0+0= 0; gray= no high areas) to seven (1+2+4= 7; red= high P, high K, high N). The map values in between identify the map layers having high values. For example, the yellow areas with the value 3 have high P and K but not N (1+2+0= 3).

If four or more maps are combined, the areas of interest are assigned increasing binary progression values (…8, 16, 32, etc)—the sum will always uniquely identify the combinations.

Locations that are high on both maps equate to 1 (1*1 = 1). In effect, this "level-slice" technique maps any data pattern you specify—just assign 1 to the data interval of interest for each map variable.

Figure 5-6 depicts level slicing for areas that are unusually high in P, K and N (nitrogen). In this instance the data pattern coincidence is a box in 3-dimensional scatter plot space.

Figure 5-6. Level-slice classification using three map variables.

While level-slicing isn't a very sophisticated classifier, it does illustrate the useful link between data space and geographic space. This fundamental concept forms the basis for most geo-statistical analysis—including map clustering discussed in the next section.
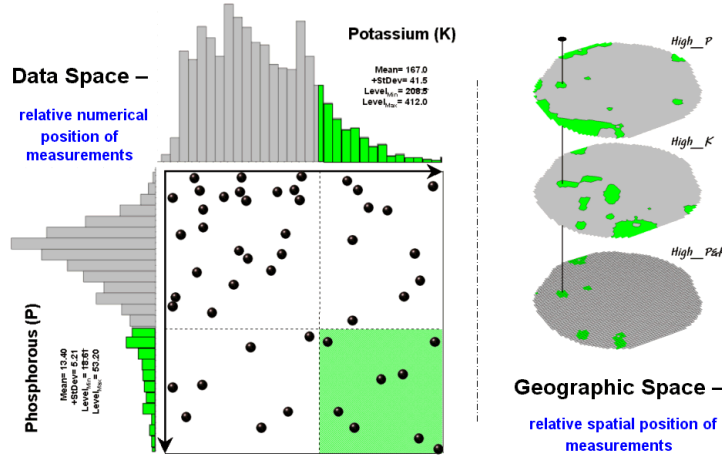
5.3 Mapping Data Clusters

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 other map locations are assigned a value from 0 (0% similar—as different as you can get) to 100 (100% similar—exactly the same data pattern).

Figure 5-7. Examples of map clustering.

Figure 5-7 shows some examples 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 P, K, and N maps identifying phosphorous, potassium and nitrogen levels throughout a cornfield that were used for the examples in the previous sections.

The other technique, *Level Slicing*, enables a user to specify a data range of interest for each map in the stack then generates 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.

The map in the center of the figure shows the results of classifying the P, K and N 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 field into two groups would be worse at balancing the two criteria.

The two smaller maps at 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 low responses and green to the one with relatively high responses. Note the encroachment on these marginal groups by the added clusters that are formed by data patterns at the boundaries.
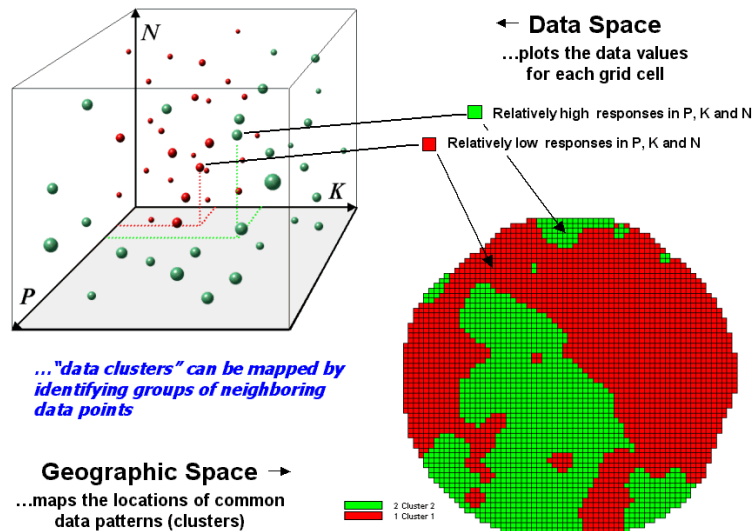
While both of these techniques are useful in examining spatial relationships, they require the user to specify data analysis parameters. But what if you don't know what Level Slice intervals to use or which locations in the field warrant Map Similarity investigation? Can the computer on its own identify groups of similar data? How would such a classification work? How well would it work?

Figure 5-8. Data patterns for map locations are depicted as floating balls in data space.

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
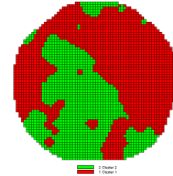
performed? What happens inside cluster's black box?

The schematic in figure 5-8 depicts the process. The floating balls identify the data patterns for each map location (geographic space) plotted against the P, K and N axes (data space). For example, the large green ball appearing closest to you depicts a location with high values on all three input maps. The tiny red ball in the opposite corner (near the plot origin) depicts a map location with small map values. It seems sensible that these two extreme responses would belong to different data groupings.

While the specific algorithm used in clustering is beyond the scope of this discussion*, it suffices to note that data distances between the floating balls are used to identify cluster membership— groups of balls that are relatively far from other groups and relatively close to each other form separate data clusters. In this example, the red balls identify relatively low responses while green ones have relatively high responses. The geographic pattern of the classification is shown in the 2D map in the lower right portion of the figure.
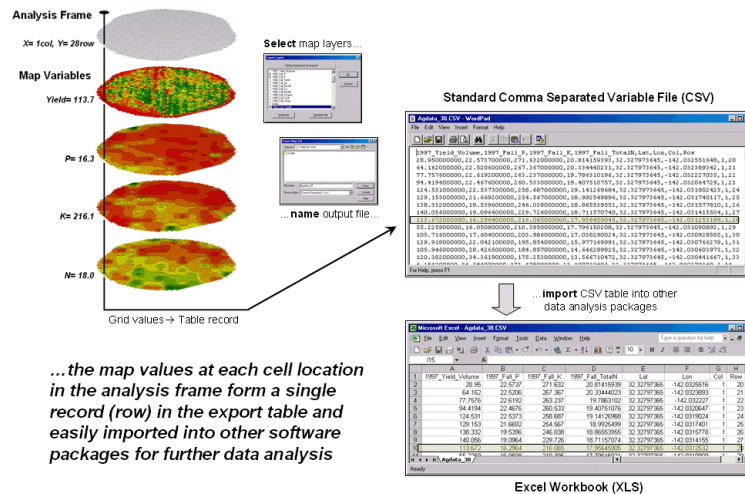
Identifying groups of

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 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.

Figure 5-9 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.

Box-and-whisker plots enable us to visualize these differences. The box is centered on the average (position) and extends above and below one standard deviation

Data Summary for Two Clusters					
		Avg	StDev	Min	Max
Cluster1					
🟥	P	12.7	4.3	5.0	32.7
	K	156.0	18.9	95.2	174.0
	N	20.4	5.3	6.3	39.8
Cluster2					
🟩	P	14.6	6.2	4.2	53.2
	K	207.0	32.9	173.0	412.0
	N	29.0	5.9	11.6	52.8

…summary statistics describe how separate and distinct the data clusters are.

Figure 5-9. Clustering results can be roughly evaluated using basic statistics.

…the map values at each cell location in the analysis frame form a single record (row) in the export table and easily imported into other software packages for further data analysis

Figure 5-10. The map values at each grid location form a single record in the exported table.

(width) with the whiskers drawn to the minimum and maximum values to provide a visual sense of the data range. When the diagrams for the two clusters overlap, as they do for the phosphorous responses, it tells us that the clusters aren't very distinct along this axis.

The separation between the boxes for the K and N axes suggests greater distinction between the clusters. Given the results a practical precision farmer would likely accept the classification results… and statisticians hopefully will accept in advance my apologies for such a introductory and terse treatment of a complex topic.

5.4 Exporting Mapped Data

Figure 5-10 shows the procedure for exporting a standard comma separated variable (CSV) file with each record containing the selected data for a single grid cell. The user selects the map layers for export and specifies the name of the output file. The computer accesses the data and constructs a standard text line with commas separating each data value. Note that the column, row of the analysis frame and its latitude, longitude earth poison is contained in each record. In the example, the export file is brought into Excel for further processing.
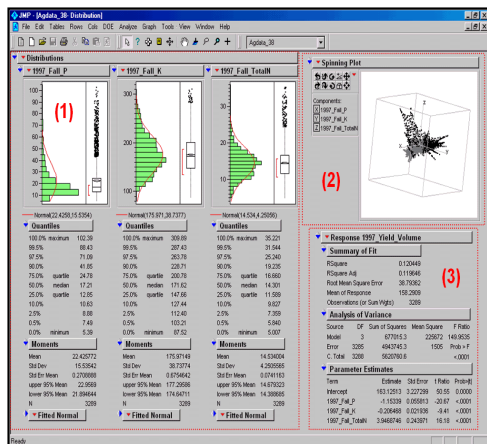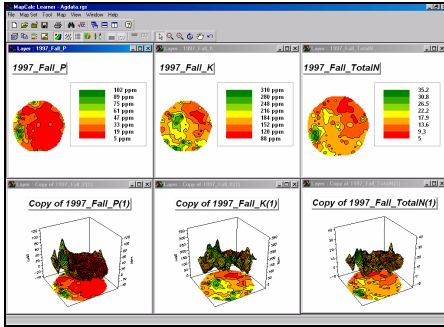

Figure 5-11. Mapped data can be imported into standard statistical packages for further analysis.

Figure 5-11 shows the agricultural data imported into the JMP statistical package (by SAS). Area (1) shows the histograms and descriptive statistics for the P, K and N map layers shown in figure 5-10. Area (2) is a "spinning 3D plot" of the data that you can rotate to graphically visualize the 3D data space of the map layers.

Area (3) shows the results of applying a multiple linear regression model to predict crop yield from the soil nutrient maps. These are but a few of the tools beyond mapping that are available through data exchange between GIS and traditional spreadsheet, database and statistical packages—a perspective that integrates maps with other technologies.

Modern statistical packages like JMP "aren't your father's" stat experience and are fully interactive with point-n-click graphical interfaces and wizards to guide you through appropriate analyses. The analytical tools, tables and displays provide a whole new view of traditional mapped data. While a map picture might be worth a thousand words, a gigabyte or so of digital map data is a whole revelation and foothold for site-specific decisions.

---

* See ***www.innovativegis.com/basis/pfprimer/***, an online book entitled <u>Precision Farming Primer</u>, Topic 4 Mapped Data Analysis Among Several Maps, Section 6 "Clustering Map Data" for further discussion of the clustering algorithm.

5.5 Exercises

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

5.5.1 Characterizing Map Similarity

Based on your previous experience in map analysis create a map of total nitrogen for the 1997 period by summing…

	1997_Fall_NH4N	
+	1997_Fall_NO3N	
	1997_Fall_TotalN	*Hint: Calculate command.*

Based on your previous experience in displaying maps generate the following side-by-side 2D and 3D displays of the **1997_Fall_P**, **1997_Fall_K** and **1997_Fall_TotalN** maps…

Hint: Use the View button (binocular icon) to clone the P, K and TotalN maps. Use the 3D Toggle button to switch to 3D display for the cloned map displays. Click on the Title Vertically button and click/drag to arrange the six displays.

Make a mental note of any similarities in the P, K, and TotalN levels in the graphical displays. For example you might "see" generally low values for all three nutrients occurring in the northeastern portion of the field.

Maximize the *1997_Fall_P window*, press the *Use Cells* button to switch to grid display format and then press the *Layer Mesh* button to turn on the analysis grid for reference. Move the cursor to the field location Column=**8**, row= **24** using the coordinate reference in lower left corner of the screen. Double-click at this location and note the *1997_Fall_P*, *_K* and *_TotalN* values—**96.5**, **160**, and **32.4**, respectively—that represent the "data pattern" for the location.

$\sqrt{\alpha}$ Press the **Map Analysis** button to access the analytical operations, select **Statistics**à **Relate** and complete the dialog box shown below to generate a similarity map to the data pattern at field location column=8, row=24.
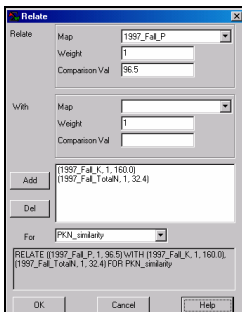
RELATE ((1997_Fall_P, 1, 96.5) WITH (1997_Fall_K, 1, 160.0), (1997_Fall_TotalN, 1, 32.4) FOR PKN_similarity

The resulting map shows how similar (red= low to green= highly similar) all map locations are to the data pattern that was entered—the western

portion of the field has the most similar data patterns.
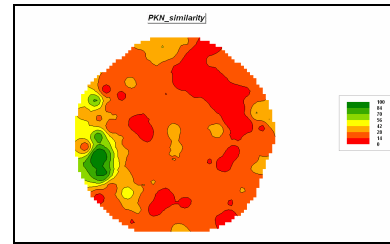
Select **Map Analysis**à **Reclassify**à **Renumber** and complete the following dialog box to isolate the locations that are very different (0-10 percent similar) and are very similar (90-100 percent similar).
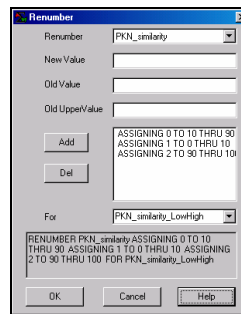
RENUMBER PKN_similarity ASSIGNING 0 TO 10 THRU 90 ASSIGNING 1 TO 0 THRU 10 ASSIGNING 2 TO 90 THRU 100 FOR PKN_similarity_LowHigh
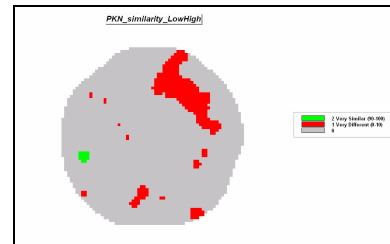

P, K and N conditions are very different in the northeast portion of the field with very similar location localized around the comparison location.
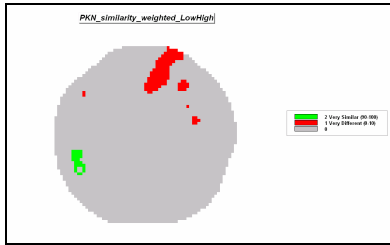
Repeat the similarity analysis but this time weighting the TotalN levels as five times more important in determining similarity than P and K levels…
  1997_Fall_P weight= 1
  1997_Fall_K weight= 1
  1997_Fall_TotalN weight= 5

…to generate a weighted similarity map—
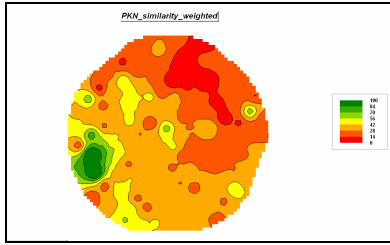
…and a map isolating Low/High similarity—

Did the change in relative weighting of the data pattern make a significant difference in the similarity maps produced? How could you quantify the difference (Hint: see Topic 3, Comparing Mapped Data)?

On your own, generate another similarity analysis (map variable weights= 1) using the normalized density vegetation index (NDVI) map series…
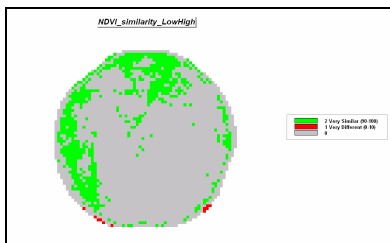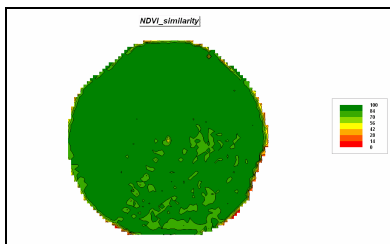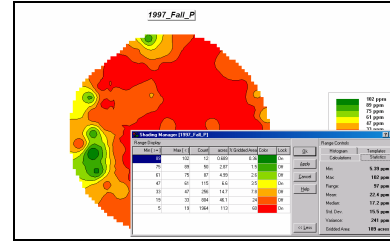 **2000_Image_7_21_NDVI**
  2000_Image_8_10_NDVI
  2000_Image_8_30_NDVI
…at field location column=**38**, row=**51**.
Complete the analysis by identifying the portions of the field that have very different and very similar data patterns.

5.5.2 Identifying Data Zones

Display the **1997_Fall_P** map, press the **Shading Manager** button and select the **Statistics** tab.

Note that the Mean value of phosphorous is 22.4ppm with a standard deviation of 15.5ppm. The cutoff for unusually high levels of phosphorous is 37.9ppm (22.4 + 15.5).

Select **Map Analysis**à **Reclassify**à **Renumber** to generate a binary map of unusually high 1997_Fall_P levels by completing the following.
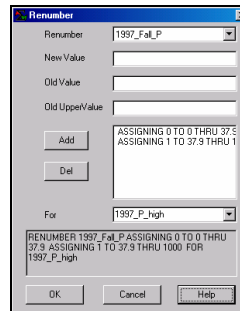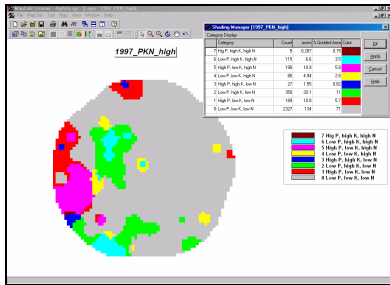

RENUMBER 1997_Fall_P ASSIGNING 0 TO 0 THRU 37.9 ASSIGNING 1 TO 37.9 THRU 1000 FOR 1997_P_high

Repeat the same processing to isolate the areas of *high K* levels in fall 1997 as **2** and *high TotalN* levels as **4**.
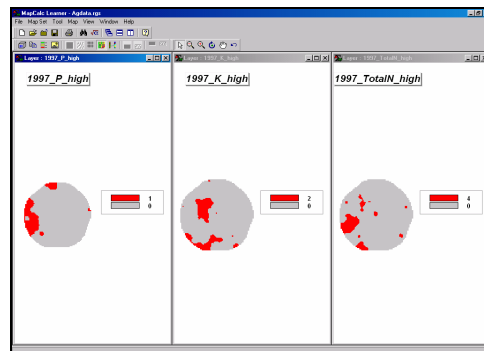
Select **Map Analysis**à **Overlay**à **Calculate** and add the three binary maps together.

The summed values identify unique combinations of the three maps—

7= *High P, high K, high N*
6= *Low P, high K, high N*
5= *High P, low K, high N*
4= *Low P, low, K, high N*
3= *High P, high K, low N*
2= *Low P, high K, low N*
1= *High P, low K, low N*
0= *Low P, low K, low N*

Most of the field (71% grey) does not contain areas with unusually high levels of nutrients. Note that most of the high nutrient coincidence occurs in the western half of the field. How might you use this information in planning a site-specific fertilization program?

5.5.3 Mapping Data Clusters

The Cluster command can be used to partition the field into inherent data pattern groups. Relative positioning in data space determines the membership of the clusters. Generate a two-cluster grouping of the field using the 1997_ Fall P, K and N maps by selecting **Map Analysis**à **Statistics**à **Cluster** and entering—
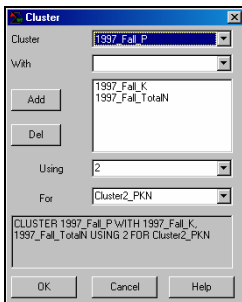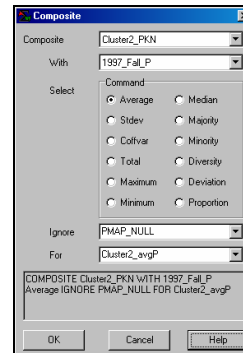

CLUSTER 1997_Fall_P WITH 1997_Fall_K, 1997_Fall_TotalN USING 2 FOR Cluster2_PKN
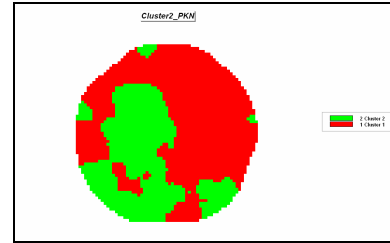
Use the "Data Drill-down" feature (double-click on the Cluster2_PKN map) and note how the P, K and N values change within the two clusters by moving the cursor around the map. What was the typical range of P values for Cluster 1? For cluster 2?
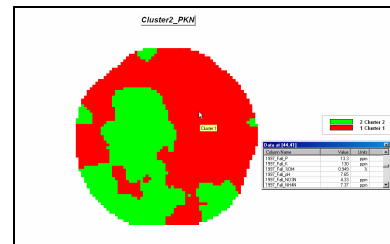
Select **Map Analysis**à **Overlay**à **Composite** …

COMPOSITE Cluster2_PKN WITH 1997_Fall_P Average IGNORE PMAP_NULL FOR Cluster2_avgP

… to calculate the average P value for both clusters. Repeat the procedure to calculate the standard deviation in phosphorous levels in both clusters. Based on the avgP and stdP results is there a "substantial" difference between the two clusters in their phosphorous levels? Hint: *avg + std* ranges should not overlap very much.

Note the careful use of the word "substantial" difference as the simple range test does not allow the statement that there is or isn't a statistical "significant" difference. A statistical test, such as an T-test or F-test, between the two data groups would be required (see *Appendix B, Exchanging Data With MapCalc* for procedures to export the data for further statistical testing in Excel or other system).

Repeat the analysis for the K and N levels in the two clusters. Is there a "substantial" difference in the levels within the two clusters for either of these nutrients?

Now create cluster maps containing three and four clusters—

*CLUSTER 1997_Fall_P WITH 1997_Fall_K, 1997_Fall_TotalN USING 3 FOR Cluster3_PKN*

CLUSTER 1997_Fall_P WITH 1997_Fall_K, 1997_Fall_TotalN USING 4 FOR Cluster4_PKN
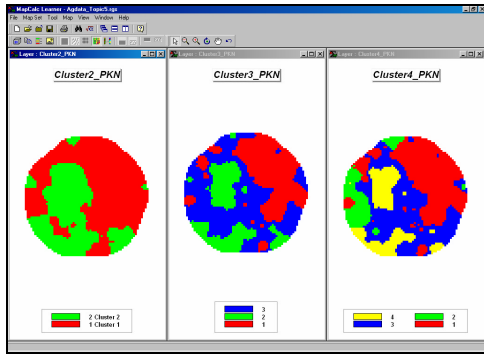
—see any consistency in the partitioning as the number of clusters gets larger?

Use the Composite command to calculate the average yield within each of the four clusters depicted on the Cluster4_PKN map.

COMPOSITE Cluster4_PKN WITH 1997_Yield_Volume Average IGNORE PMAP_NULL FOR Cluster4_avgY
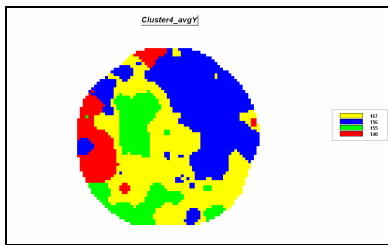
Does there appear to be a progression in the average yield among the four clusters?

Repeat the processing to determine the average yield within the clusters on the Cluster3_PKN and Cluster2_PKN maps. Can you discern any consistent patterns in the results?

5.5.4 Exporting Data for Further Analysis
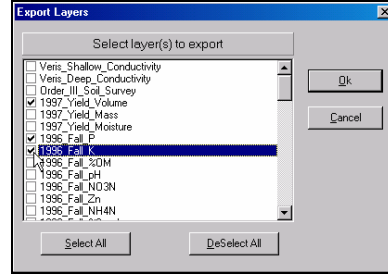
From the Main Menu select **File** à **Export** à **Data** to access the wizard for exchanging data. Press the *DeSelect All* button then click on the boxes next to the **1997_Yield_Volume**, **1997_Fall_P** and **1997_Fall_K** map layers. Press *OK* to begin the export.
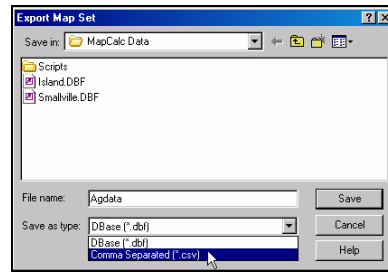
Specify "**CSV**" as the type and **Save** the file. This procedure stores the file in the default data folder.
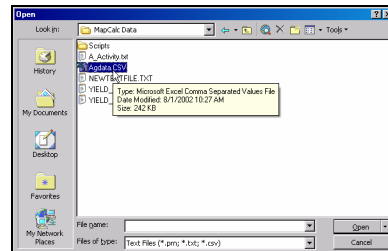
Access *Excel* by clicking on **Start** à **Programs** à **Microsoft Excel** à **File** à **Open** à browse to the *...\MapCalc Data* folder à specify **Text Files (*.prn, *.txt, *.csv)** as the file type à click on the **AgData.csv** file à and press the *Open* button.

The exported file containing the specified map layers will be opened in *Excel*.

There are several other standard export formats supporting databases (.DBF), MapInfo (.TAB), ArcView (.SHP and GridASCII), and Surfer (.GRD).

See **Appendix B, Exchanging Data with MapCalc** for information on additional procedures for exchanging data (import and export).

You can exit the program by selecting **File** à **Exit** or by clicking on the "X" in the upper-right corner of the MapCalc program window. If you want to save your work, specify a new file name, such as *AgData_Topic5_exercises.rgs*. Each exercise set assumes you will start with the basic *AgData.rgs* data set and this database will become cluttered with exercise maps if you save your results to it each time.