Topic 7

Spatial Data Mining

7.1 Characterizing Data Groups

How often have you seen a GIS 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 background scenery for the laser show.

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

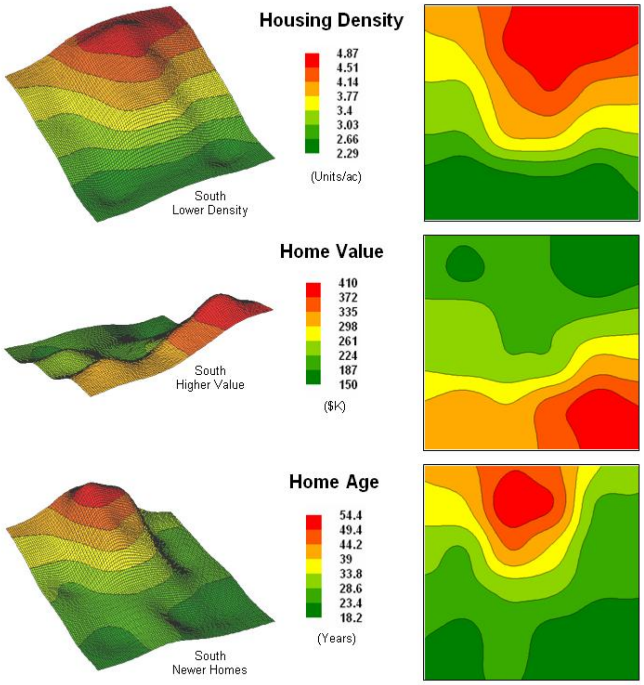


Figure 7-1. Map surfaces identifying the spatial distribution of housing density, value and age.

Consider the three maps shown in figure 7-1— 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? Or how about a northeastern section? The answers to these questions are far too complex for visual analysis and certainly beyond the geo-query and display procedures of standard desktop mapping packages.

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

The difficulty in visual analysis is two-fold— remembering the color patterns and calculating the difference. Quantitative map analysis does the same thing except it uses the actual map values in place of 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 (10,000 grid cells in this example) in a second or two.

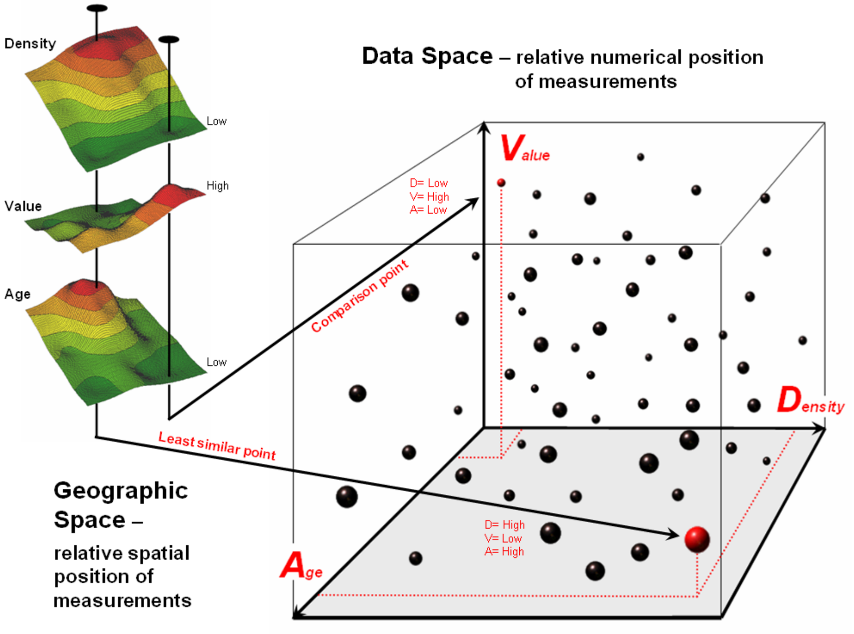


Figure 7-2. Conceptually linking geographic space and data space.

The upper-left portion of figure 7-2 illustrates capturing the data patterns for comparing two map locations. The "data spear" at map location 52column, 85row identifies the housing Density as 2.1 units/ac, Value as $407,000 and Age as 18.3 years. This step is analogous to your eye noting a color pattern of green, red, and green. The other speared location (56, 84) locates the least similar data pattern with Density= 4.8 units/ac, Value= $190,000 and Age= 51.2 years …or as your eye sees it, a color pattern of red, green and red.

The right side of figure 7-2 schematically depicts how the computer determines similarity in the data patterns for the two locations by analyzing them in three-dimensional "numeric" space. Similar data patterns plot close to one another with increasing distance indicating less similarity. The realization that mapped data can be expressed in both geographic space and data space is paramount to working knowledge of how a computer quantitatively analyses interrelationships among mapped data.

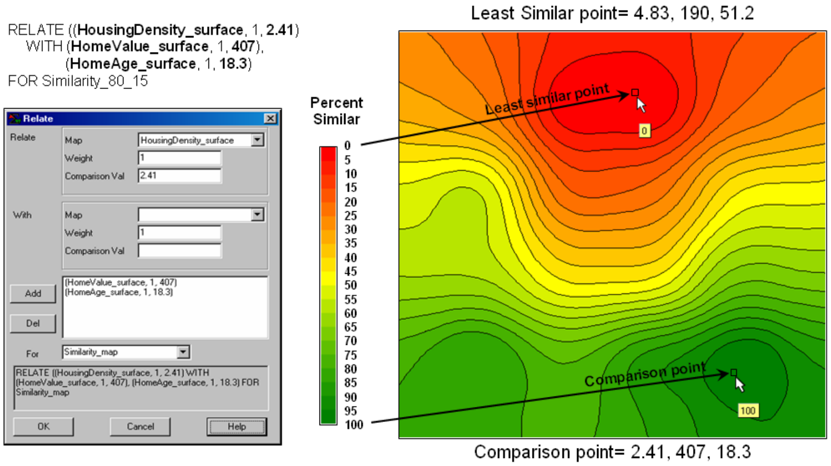
Geographic space uses coordinates, such as latitude and longitude, to locate things in the real world. The geographic expression of the complete set of measurements depicts their spatial distribution in familiar map form. ***Data space***, on the other hand, is a bit less familiar. While you can't stroll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

In the example, the three axes defining the extent of the box correspond to housing Density (D), Value (V) and Age (A). The floating balls represent data patterns of the grid cells defining the geographic space—one "floating ball" (data point) for each grid cell. The data values locating the balls extend from the data axes—2.41, 407.0 and 18.3 for the *comparison point* identified in figure 7-2. The other point has considerably higher values in D and A with a much lower V values (4.83, 51.2 and 190.0 respectively) so it plots at a different location in data space.

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, and high, high, high in the upper-right corner of the box. Points that plot in data space close to each other are similar; those that plot farther away are less similar. Data distance is the way computers "sees" what you see in the 3D plot of the values. The real difference in the graphical and quantitative approaches is in the details—the tireless computer "sees" extremely subtle differences between all points and can generate a detailed map of similarity in less than a second.

7.2 Calculating Map Similarity

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



```
RELATE ((HousingDensity_surface, 1, 2.41)
    WITH (HomeValue_surface, 1, 407),
         (HomeAge_surface, 1, 18.3)
FOR Similarity_80_15
```

Relate

Relate — Map: HousingDensity_surface; Weight: 1; Comparison Val: 2.41

With — Map: ; Weight: 1; Comparison Val:

(HomeValue_surface, 1, 407)
(HomeAge_surface, 1, 18.3)

For: Similarity_map

```
RELATE ((HousingDensity_surface, 1, 2.41) WITH
(HomeValue_surface, 1, 407), (HomeAge_surface, 1, 18.3) FOR
Similarity_map
```

Figure 7-3. A similarity map identifies how related locations are to a given point.

The similarity map shown in figure 7-3 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 locations having fairly similar D, V and A levels to the comparison location—with the darkest green identifying locations with identical D, V and A levels (100% similar). It is interesting to note that most of the very similar locations are in the southern portion of the project area. The light-green to red tones indicate increasingly dissimilar areas in the north.

A similarity map can be an extremely valuable tool for investigating spatial patterns in any complex set of mapped data. The similarity calculations can handle any number of input maps, yet humans are unable to even conceptualize more than three variables (data space box). Also, the different map layers can be weighted to reflect relative importance in determining overall similarity. For example, housing Value could be specified as 10 times more important in assessing similarity. The result would be a different map than the one shown in figure 7-3—how different depends on the data patterns themselves.

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 to a consistent and quantifiable.

7.3 Identifying Data Zones

The 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.

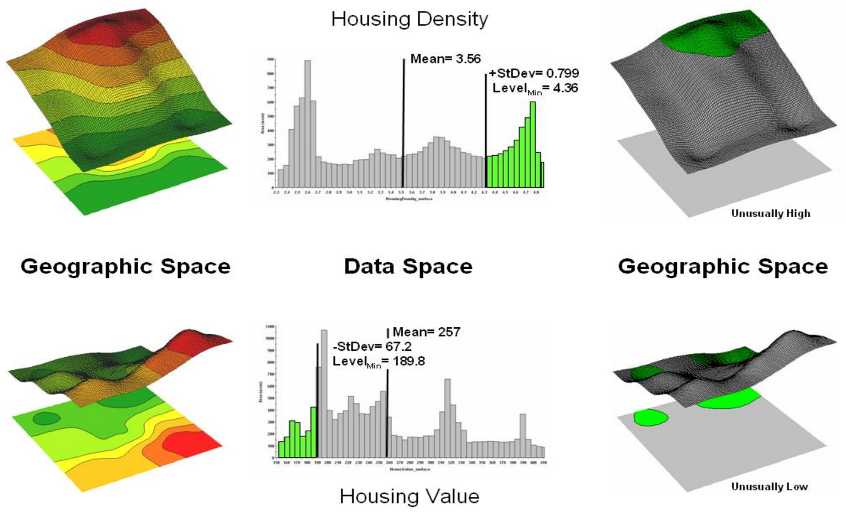


Figure 7-4. Identifying areas of unusually high measurements.

The link between *Geographic Space* and *Data Space* is the key. As shown in figure 7-4, 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 housing density within the project area. Note the areas of high housing Density along the northern edge coincide with low home Values.

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—relative level of housing Density. In this case, the spikes in the graph indicate measurements that occur more frequently. Note the high occurrence of density values around 2.6 and 4.7 units per acre.

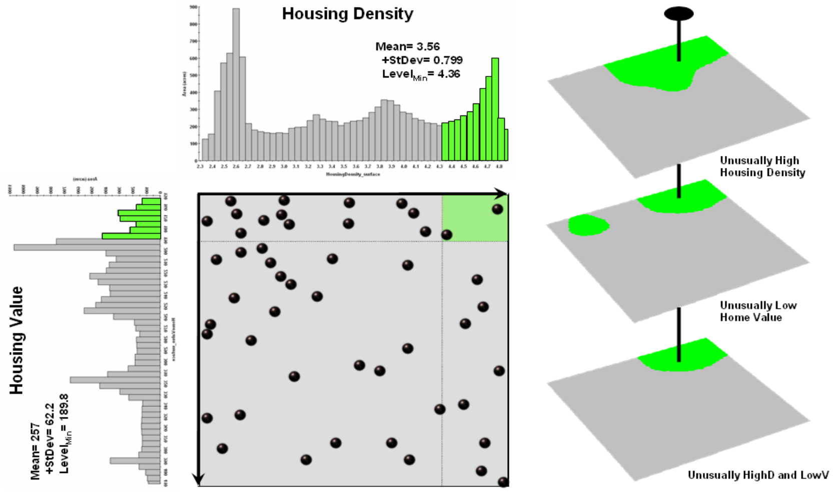
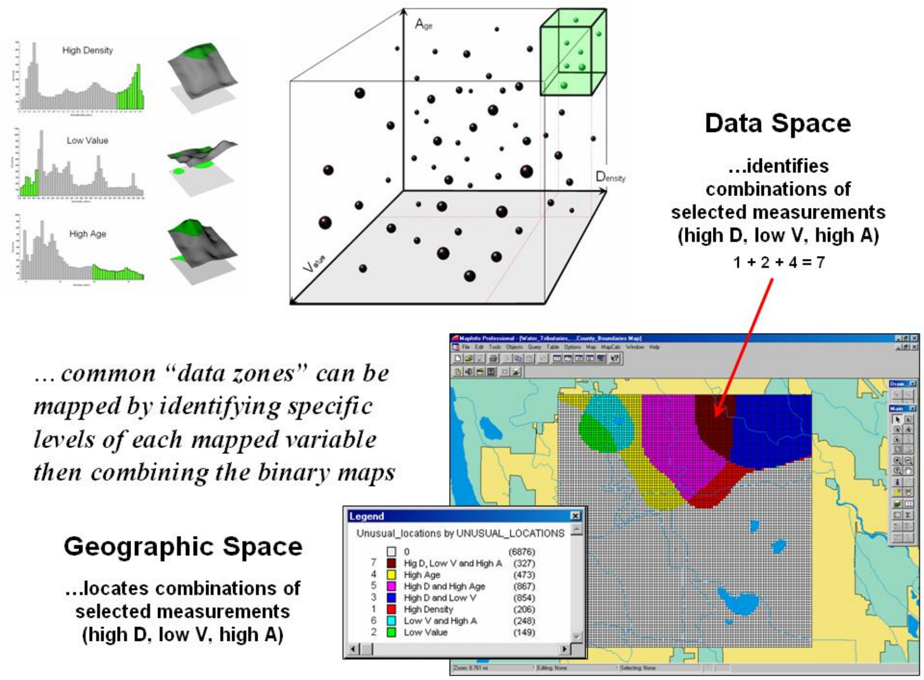


Figure 7-5 Identifying joint coincidence in both data and geographic space.

Figure 7-5 schematically illustrates combining the housing Density and Value data to locate areas that have high measurements in both. The graphic on the left is termed a scatter plot that 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 housing Density and Value measurements for one of the map locations—10,000 in all for the actual example data set. The balls schematically shown in the shaded area of the diagram identify locations that have high D and low V.

The aligned maps on the right side of the figure show the geographic solution for the high D, low V areas. A simple map-*ematical* way to generate the solution is to assign 1 to all locations of low values in the D and high values in the V map layers (bright green). Zero is assigned to locations that fail to meet the conditions. When the two binary maps (0/1) are multiplied, a zero on either map computes to zero. Locations that meet the conditions on both maps equate to one (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 in the map stack, then multiply.


High Density
Low Value
High Age
Age
Density
Value
Data Space
...identifies combinations of selected measurements (high D, low V, high A)
1 + 2 + 4 = 7
... common "data zones" can be mapped by identifying specific levels of each mapped variable then combining the binary maps
Geographic Space
...locates combinations of selected measurements (high D, low V, high A)
Legend

Unusual_locations by UNUSUAL_LOCATIONS

	□	0	(6876)
7	■	Hig D, Low V and High A	(327)
4	■	High Age	(473)
5	■	High D and High Age	(867)
3	■	High D and Low V	(854)
1	■	High Density	(206)
6	■	Low V and High A	(248)
2	■	Low Value	(149)

Figure 7-6. Level-slice classification using three map variables.

Figure 7-6 depicts level slicing for areas that are unusually low housing Density, high Value and low Age. In this instance the data pattern coincidence is a box in 3-dimensional scatter plot space (upper-right corner). However a slightly different map-*ematical* trick was employed to get the detailed map solution shown in the figure.

On the individual maps, areas of high Density were set to D= 1, low Value to V=2 and high Age to A=4, then the binary map layers were added together. The result is a range of coincidence values from zero (0+0+0= 0; gray= no coincidence) to seven (1+2+4= 7; combination not present in the example). The map values in between identify the areas meeting other combinations of the conditions. For example, the green areas contain the value 3 indicating high D and low V but not high A (1+2+0= 3) which represents 13 percent of the project area (1327/10000= 13.27%). 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 all possible combinations of the conditions specified.

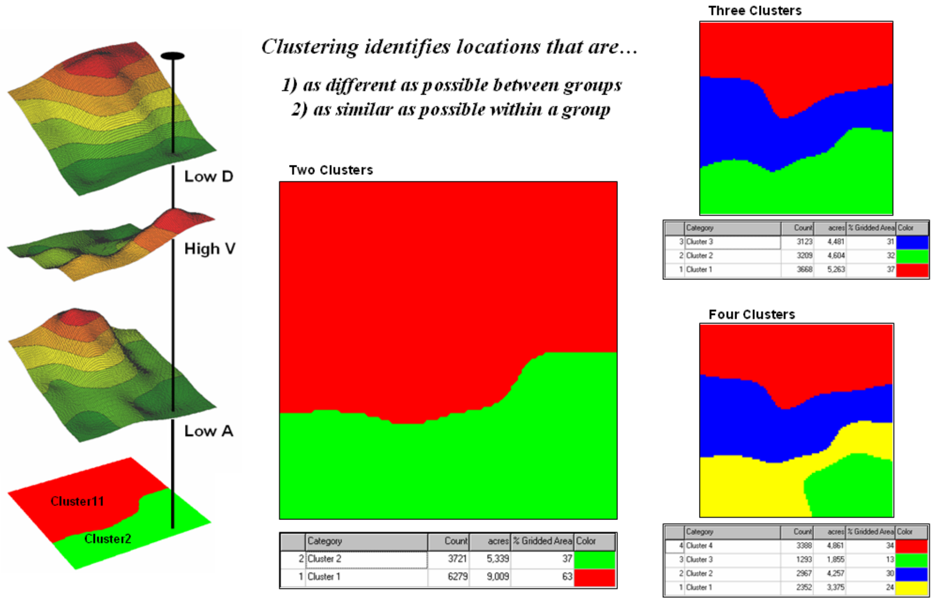
While level-slicing isn't a very sophisticated classifier, it does illustrate the useful link between data space and geographic space. Yet this fundamental concept forms the basis for most geo-statistical analysis—including map clustering discussed in the next section.

7.4 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).

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

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


Clustering identifies locations that are…
1) as different as possible between groups
2) as similar as possible within a group
Low D
High V
Low A
Cluster11
Cluster2
Two Clusters
Three Clusters
Four Clusters

	Category	Count	acres	% Gridded Area	Color
2	Cluster 2	3721	5,339	37	🟩
1	Cluster 1	6279	9,009	63	🟥

	Category	Count	acres	% Gridded Area	Color
3	Cluster 3	3123	4,481	31	🟦
2	Cluster 2	3209	4,604	32	🟩
1	Cluster 1	3668	5,263	37	🟥

	Category	Count	acres	% Gridded Area	Color
4	Cluster 4	3388	4,861	34	🟥
3	Cluster 3	1293	1,855	13	🟩
2	Cluster 2	2967	4,257	30	🟦
1	Cluster 1	2352	3,375	24	🟨

Figure 7-7. Examples of map clustering.

Figure 7-7 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 ones used in previous examples and identify the geographic and numeric distributions of housing Density, home Value and house Age levels throughout the project area.

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

The two smaller maps on the right show the division of the data set into three and four clusters. In all three of the cluster maps, red is assigned to the cluster with relatively high Density, low Value and high Age (less wealthy) responses and green to the one with the most opposite conditions (wealthy areas). Note the encroachment on these marginal groups by the added clusters that are formed by data patterns at the classification boundaries. The procedure is effectively dividing the project area into "neighborhoods" based on relative D, V and A values throughout the map area. Whereas traditional neighborhoods usually are established by historical legacy, cluster partitions respond to mapped data and can be useful in establishing insurance zones, sales areas and marketing clusters.

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

The schematic in figure 7-8 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, tiny red ball in the upper-right corner depicts a map location in the less wealthy part of town (high D, low V and high A). The large green ball appearing closest to you depicts a location in the wealthier part (low D, high V and low A). It seems sensible that these two extreme responses would belong to different data groupings (clusters 1 and 2 respectively).

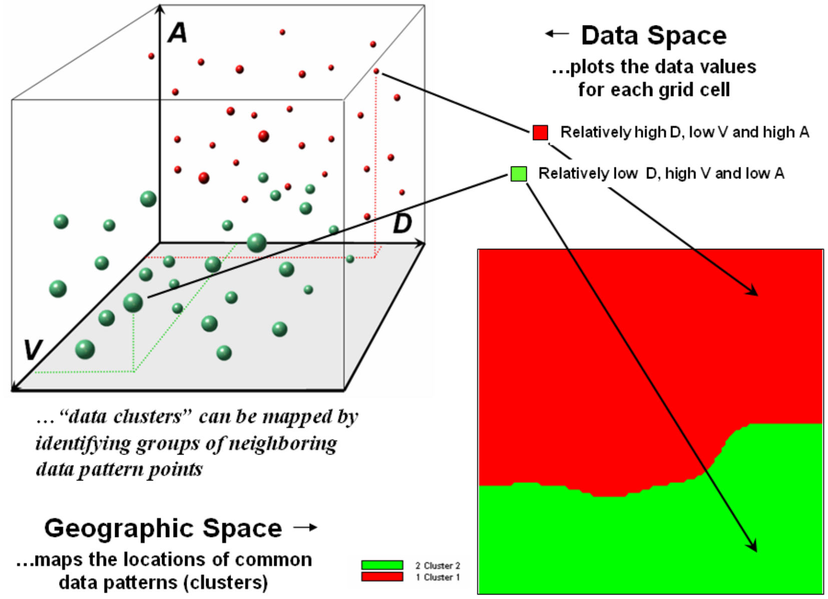

A
D
V
← Data Space
…plots the data values for each grid cell
Relatively high D, low V and high A
Relatively low D, high V and low A
… "data clusters" can be mapped by identifying groups of neighboring data pattern points
Geographic Space →
…maps the locations of common data patterns (clusters)
2 Cluster 2
1 Cluster 1

Figure 7-8. Data patterns for map locations are depicted as floating balls in data space.

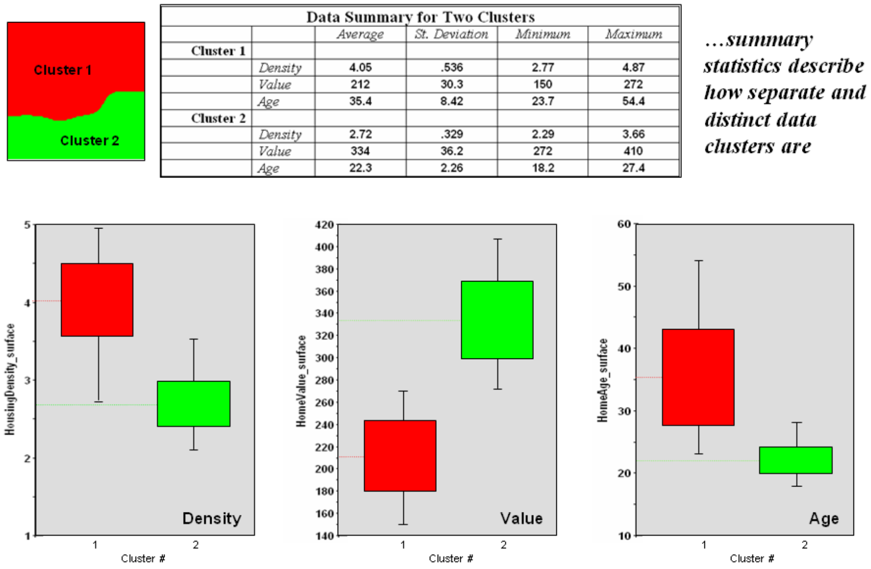
While the specific algorithm used in clustering is beyond the scope of this discussion, it suffices to note that data distances between the floating balls are used to identify cluster membership—groups of balls that are relatively far from other groups and relatively close to each other form separate data clusters. In this example, the red balls identify relatively less wealthy locations while green ones identify wealthier sections. The geographic pattern of the classification (wealthier in the south) 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, such as T- and F-statistics used in comparing sample populations.

Figure 7-9 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 (width) with the whiskers drawn to the minimum and maximum values to provide a visual sense of the data range. If the plots for the two clusters overlap, it suggests that the clusters are not very distinct—significant overlapping membership.



Data Summary for Two Clusters

		Average	*St. Deviation*	*Minimum*	*Maximum*
Cluster 1					
	Density	4.05	.536	2.77	4.87
	Value	212	30.3	150	272
	Age	35.4	8.42	23.7	54.4
Cluster 2					
	Density	2.72	.329	2.29	3.66
	Value	334	36.2	272	410
	Age	22.3	2.26	18.2	27.4

…summary statistics describe how separate and distinct data clusters are

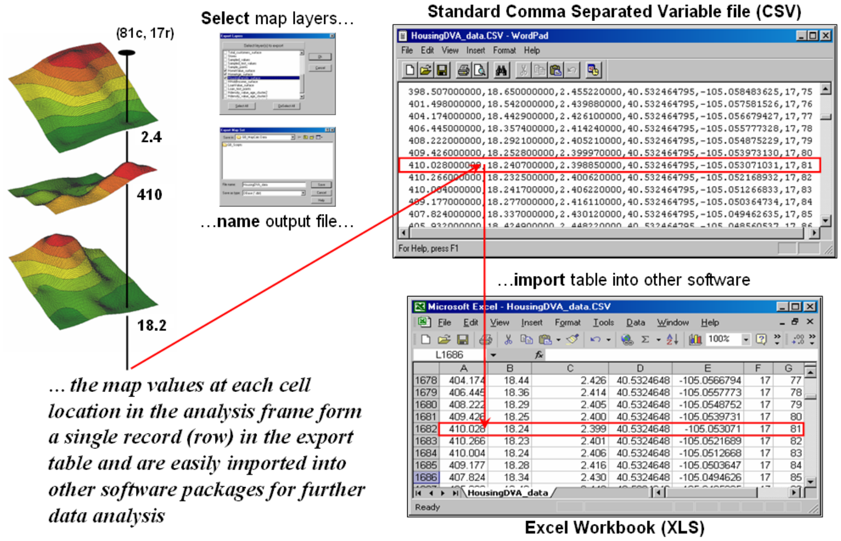
Figure 7-9. Clustering results can be roughly evaluated using basic statistics.

The separation between the boxes in all three of the data layers of the example suggests good distinction between the two clusters. Given the results, it appears that the clustering classification in the example is acceptable… and hopefully statisticians will accept in advance my apologies for such an introductory and visual treatment of a complex topic.

7.5 Exporting Mapped Data

Keep in mind that all of the grid layers in a data set have same analysis grid configuration—identical number of geo-registered columns and rows. This configuration greatly facilitates export of mapped data to other software systems.

Figure 7-10 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 positsion is contained in each record. In the example, the export file is brought into Excel for further processing.


(81c, 17r)
2.4
410
18.2
Select map layers…
…name output file…
HousingDVA_data.CSV - WordPad
File Edit View Insert Format Help
For Help, press F1
Microsoft Excel - HousingDVA_data.CSV
File Edit View Insert Format Tools Data Window Help
100%
L1686
HousingDVA_data
Ready

Standard Comma Separated Variable file (CSV)

```
398.507000000,18.650000000,2.455220000,40.532464795,-105.058483625,17,75
401.498000000,18.542000000,2.439880000,40.532464795,-105.057581526,17,76
404.174000000,18.442900000,2.426100000,40.532464795,-105.056679427,17,77
406.445000000,18.357400000,2.414240000,40.532464795,-105.055777328,17,78
408.222000000,18.292100000,2.405210000,40.532464795,-105.054875229,17,79
409.426000000,18.252800000,2.399970000,40.532464795,-105.053973130,17,80
410.028000000,18.240700000,2.398850000,40.532464795,-105.053071031,17,81
410.266000000,18.232500000,2.400620000,40.532464795,-105.052168932,17,82
410.004000000,18.241700000,2.406220000,40.532464795,-105.051266833,17,83
409.177000000,18.277000000,2.416110000,40.532464795,-105.050364734,17,84
407.824000000,18.337000000,2.430120000,40.532464795,-105.049462635,17,85
```

…**import** table into other software

	A	B	C	D	E	F	G
1678	404.174	18.44	2.426	40.5324648	-105.0566794	17	77
1679	406.445	18.36	2.414	40.5324648	-105.0557773	17	78
1680	408.222	18.29	2.405	40.5324648	-105.0548752	17	79
1681	409.426	18.25	2.400	40.5324648	-105.0539731	17	80
1682	410.028	18.24	2.399	40.5324648	-105.053071	17	81
1683	410.266	18.23	2.401	40.5324648	-105.0521689	17	82
1684	410.004	18.24	2.406	40.5324648	-105.0512668	17	83
1685	409.177	18.28	2.416	40.5324648	-105.0503647	17	84
1686	407.824	18.34	2.430	40.5324648	-105.0494626	17	85

Excel Workbook (XLS)

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

Figure 7-10. The map values at each grid location form a single record in the exported table.

The ability to "walk" between grid maps and databases, spreadsheets and statistical packages provide a host of extended analytical operations. The key is the consistent partitioning formed by the analysis grid as each "sample" represents an equivalent geographic area. In vector-based systems, map features are irregular objects of varying size that inhibits analytical analysis requiring complex area-weighting procedures. In a sense, the analysis grid is like traditional statistical sampling of a variable, except in grid-based map analysis the samples cover the entire area of interest—no troublesome gaps between samples.

The consistent data format coupled with modern 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 new revelation providing a foothold for site-specific decisions.

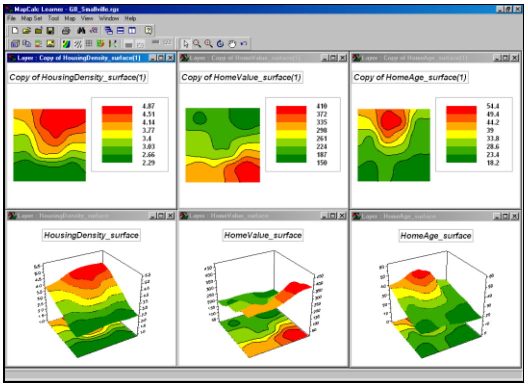
7.6 Exercises

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

7.6.1 Characterizing Map Similarity

Based on your previous experience in displaying maps generate the following side-by-side 2D and 3D displays of the **Housingdensity_surface**, **Homevalue_surface** and **HomeAge_surface** maps.

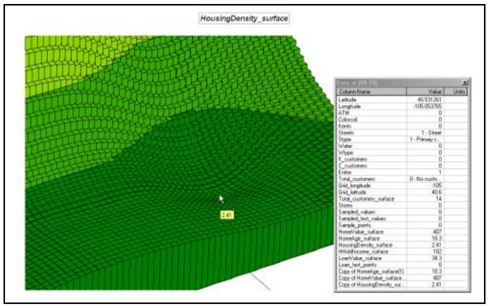
*Hint: Use the View button (binocular icon) to clone the D, V and A maps. Use the 3D Toggle button to switch to 2D display for the cloned map displays. Select **Map→ Legend→ None** for the 3D plots. Click on the "Title Vertically" button then click/drag the display windows to arrange the six displays as shown.*


MapCalc Learner - GB_Smallville.rgs
File Map Set Tool Map View Window Help
Layer : Copy of HousingDensity_surface(1)
Copy of HousingDensity_surface(1)
4.87
4.51
4.14
3.77
3.4
3.03
2.66
2.29
Copy of HomeValue_surface(1)
410
372
335
298
261
224
187
150
Copy of HomeAge_surface(1)
54.4
49.4
44.2
39
33.8
28.6
23.4
18.2
HousingDensity_surface
HomeValue_surface
HomeAge_surface

Make a mental note of any similarities in the D, V, and A values depicted in the graphical displays. For example you might "see" generally low D and A values coinciding with high V values in the southern portion of the project area.

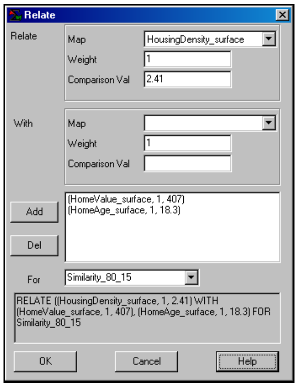
Maximize the **HousingDensity_surface** window, press the *Use Cells* button to switch to grid display format and then press the *Layer Mesh* button to turn on the analysis grid for reference.

Zoom in on the southeastern portion of the map. Move the cursor to the field location Column=**80**, row= **15** using the coordinate reference in lower left corner of the screen. Double-click at this location and note the *HousingDensity_surface*, *HomeValue_surface* and *HomeAge_surface* values—**2.41**, **407**, and **18.3**, respectively—that represent the "data pattern" for the location.



Column Name	Value	Units
Latitude	40.531261	
Longitude	-105.053765	
ATM	0	
Colossal	0	
Kents	0	
Streets	1 - Street	
Stype	1 - Primary s...	
Water	0	
Wtype	0	
K_customers	0	
C_customers	0	
Entire	1	
Total_customers	0 - No custo...	
Grid_longitude	-105	
Grid_latitude	40.6	
Total_customers_surface	14	
Stores	0	
Sampled_values	0	
Sampled_test_values	0	
Sample_points	0	
HomeValue_surface	407	
HomeAge_surface	18.3	
HousingDensity_surface	2.41	
HHoldIncome_surface	102	
LoanValue_surface	34.3	
Loan_test_points	0	
Copy of HomeAge_surface(1)	18.3	
Copy of HomeValue_surface...	407	
Copy of HousingDensity_sur...	2.41	

Press the **Map Analysis** button to access the analytical operations, select **Statistics→ Relate** and complete the dialog box shown below to generate a similarity map to the data pattern at field location column=80, row=15.



Relate — Map: HousingDensity_surface; Weight: 1; Comparison Val: 2.41

With — Map: ; Weight: 1; Comparison Val:

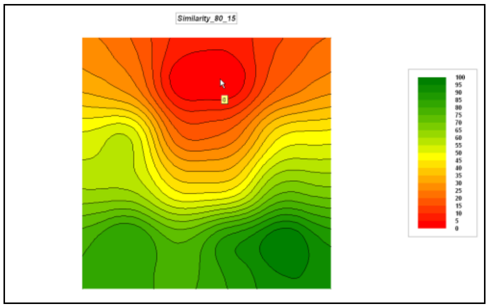
(HomeValue_surface, 1, 407)
(HomeAge_surface, 1, 18.3)

For: Similarity_80_15

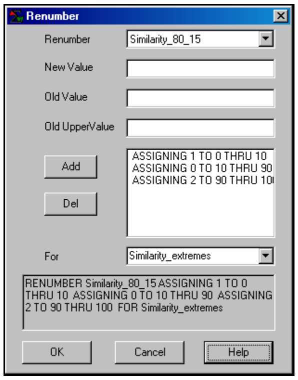
RELATE ((HousingDensity_surface, 1, 2.41) WITH (HomeValue_surface, 1, 407), (HomeAge_surface, 1, 18.3) FOR Similarity_80_15

RELATE ((HousingDensity_surface, 1, 2.41) WITH (HomeValue_surface, 1, 407), (HomeAge_surface, 1, 18.3) FOR Similarity_80_15

The resulting map shows how similar (red= low to green= highly similar; 20 Equal Count intervals) all map locations are to the data pattern that was entered—the northern portion of the project area has the most dissimilar data patterns (0% similar at 56c, 84r is least similar).



Select **Map Analysis→ Reclassify→ 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: Similarity_80_15

New Value:

Old Value:

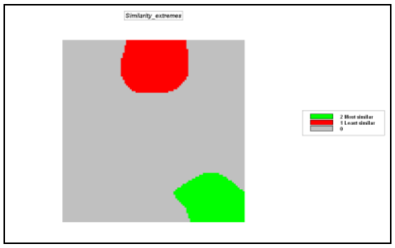
Old UpperValue:

ASSIGNING 1 TO 0 THRU 10
ASSIGNING 0 TO 10 THRU 90
ASSIGNING 2 TO 90 THRU 10(

For: Similarity_extremes

RENUMBER Similarity_80_15 ASSIGNING 1 TO 0 THRU 10 ASSIGNING 0 TO 10 THRU 90 ASSIGNING 2 TO 90 THRU 100 FOR Similarity_extremes

RENUMBER Similarity_80_15 ASSIGNING 1 TO 0 THRU 10 ASSIGNING 0 TO 10 THRU 90 ASSIGNING 2 TO 90 THRU 100 FOR Similarity_extremes

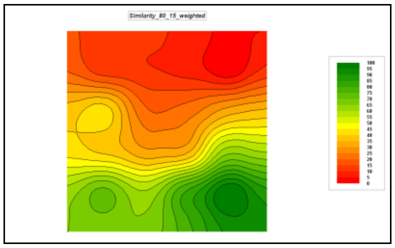


D, V and A conditions are very different in the northern portion of the area with very similar locations localized around the comparison location in the southeast.

Repeat the similarity analysis weighting the HomeValue_surface map as ten times more important in determining similarity…

HousingDensity_surface= **1**
HomeValue_surface= **10**
HomeAge_surface= **1**

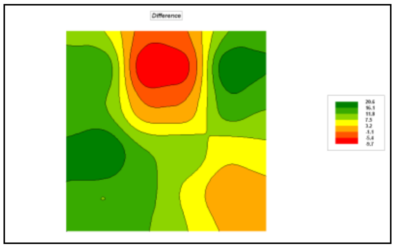
…to generate a weighted similarity map—



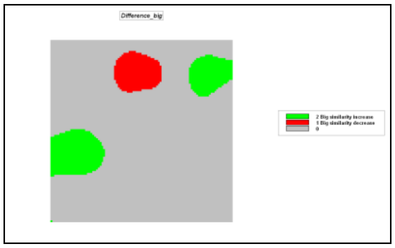
Do you "see" a significant difference between the weighted and un-weighted surfaces? Where did most of the change occur?

Use **Overlay→ Calculate** to create a percent difference surface that quantifies the changes between the weighted and un-weighted maps.

Calculate Similarity_80_15 - Similarity_80_15_weighted FOR Difference

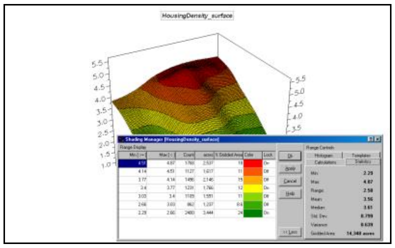


Isolate the locations that had the greatest change in similarity (-5.4 to -9.7 and +16.1 to +20.6)—



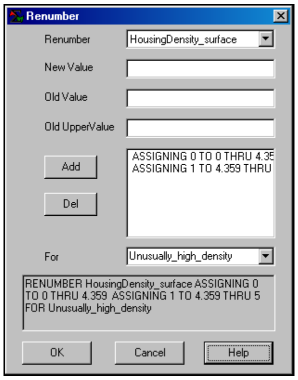
6.5.2 Identifying Data Zones

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



Note that the Mean value of housing density is 3.56 units/ac with a standard deviation of 0.799. The cutoff for unusually high density levels is 4.359 (3.56 + 0.799).

Select **Map Analysis→ Reclassify→ Renumber** to generate a binary map of unusually high housing density by completing the following.



Renumber

Renumber: HousingDensity_surface

New Value:

Old Value:

Old UpperValue:

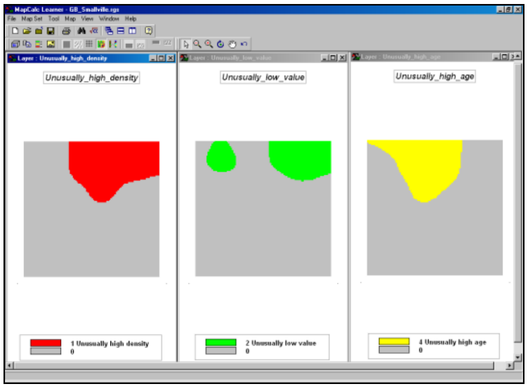
ASSIGNING 0 TO 0 THRU 4.35
ASSIGNING 1 TO 4.359 THRU

For: Unusually_high_density

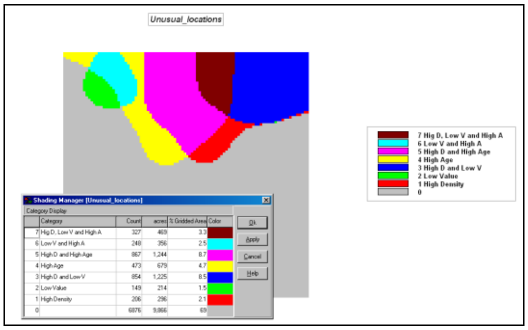
RENUMBER HousingDensity_surface ASSIGNING 0 TO 0 THRU 4.359 ASSIGNING 1 TO 4.359 THRU 5 FOR Unusually_high_density

RENUMBER HousingDensity_surface ASSIGNING 0 TO 0 THRU 4.359 ASSIGNING 1 TO 4.359 THRU 5 FOR Unusually_high_density

Repeat the same processing (cutoff= mean + StDev) to isolate the areas of *low home value* as **2** and *high home age* levels as **4**.


MapCalc Learner - GB_Smallville.rgs
File Map Set Tool Map View Window Help
Layer : Unusually_high_density
Layer : Unusually_low_value
Layer : Unusually_high_age
Unusually_high_density
Unusually_low_value
Unusually_high_age
1 Unusually high density
0
2 Unusually low value
0
4 Unusually high age
0

Select **Map Analysis→ Overlay→ Calculate** and add the three binary maps together.



Shading Manager [Unusual_locations]

Category Display

	Category	Count	acres	% Gridded Area	Color
7	Hig D, Low V and High A	327	469	3.3	🟫
6	Low V and High A	248	356	2.5	🟦
5	High D and High Age	867	1,244	8.7	🟪
4	High Age	473	679	4.7	🟨
3	High D and Low V	854	1,225	8.5	🟦
2	Low Value	149	214	1.5	🟩
1	High Density	206	296	2.1	🟥
0		6876	9,866	69	⬜

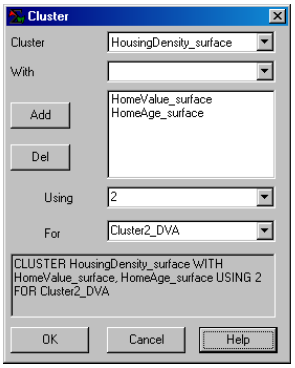
The summed values identify unique combinations of the three maps—

7= *High Density, Low Value and High Age*
6= *Low Value and High Age*
5= *High Density and High Age*
4= *High Age*
3= *High Density and Low Value*
2= *Low Value*
1= *High Density*
0= *not unusual*

Most of the project area (69% grey) does not contain areas with unusual conditions. Note that all three conditions coincide in a small area in the northern part of the project area (3.3% magenta). How might you use this information in site-specific land use, insurance, or sales planning?

6.5.3 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 project area using the housing D, V and A maps by selecting **Map Analysis→ Statistics→ Cluster** and entering—

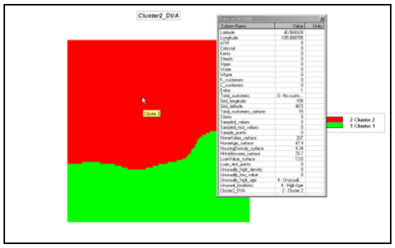


Cluster: HousingDensity_surface
With:
HomeValue_surface
HomeAge_surface
Using: 2
For: Cluster2_DVA

```
CLUSTER HousingDensity_surface WITH
HomeValue_surface, HomeAge_surface USING 2
FOR Cluster2_DVA
```

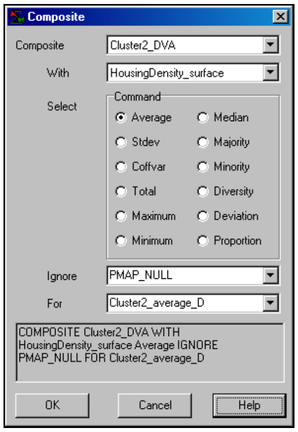
CLUSTER HousingDensity_surface WITH HomeValue_surface, HomeAge_surface USING 2 FOR Cluster2_DVA

Use the "Data Drill-down" feature (double-click on the *Cluster2_DVA* map) and note how the D, V and A values change within the two clusters by moving the cursor around the map. What was the typical range of housing Density values for Cluster 1? For cluster 2?



Column Name	Value	Units
Latitude	40.568328	
Longitude	-105.088785	
ATM	0	
Colossal	0	
Kents	0	
Streets	0	
Stype	0	
Water	0	
Wtype	0	
K_customers	0	
C_customers	0	
Extras	1	
Total_customers	0 - No custs...	
Grid_longitude	-105	
Grid_latitude	40.5	
Total_customers_surface	15	
Stores	0	
Sampled_values	0	
Sampled_test_values	0	
Sample_points	0	
HomeValue_surface	207	
HomeAge_surface	47.9	
HousingDensity_surface	4.34	
HHoldIncome_surface	70.7	
LoanValue_surface	13.6	
Loan_test_points	0	
Unusually_high_density	0	
Unusually_low_value	0	
Unusually_high_age	4 - Unusual...	
Unusual_locations	4 - High Age	
Cluster2_DVA	2 - Cluster 2	

Select **Map Analysis→ Overlay→ Composite** …



Composite: Cluster2_DVA
With: HousingDensity_surface
Select — Command:
● Average ○ Median
○ Stdev ○ Majority
○ Coffvar ○ Minority
○ Total ○ Diversity
○ Maximum ○ Deviation
○ Minimum ○ Proportion
Ignore: PMAP_NULL
For: Cluster2_average_D

```
COMPOSITE Cluster2_DVA WITH
HousingDensity_surface Average IGNORE
PMAP_NULL FOR Cluster2_average_D
```

COMPOSITE Cluster2_DVA WITH HousingDensity_surface Average IGNORE PMAP_NULL FOR Cluster2_average_D

… to calculate the average D value for both clusters. This is a "compute-heavy" operation so it might take a few moments—be patient.

Repeat the procedure to calculate the standard deviation in density levels in both clusters.

COMPOSITE Cluster2_DVA WITH HousingDensity_surface StdDev IGNORE PMAP_NULL FOR Cluster2_StDev_D

Based on the avgP and stdP results is there a "substantial" difference between the two clusters in their housing density levels? Hint: *avg + std* ranges should not overlap very much.

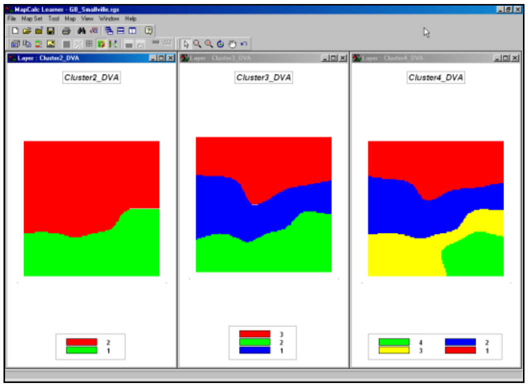
Note the careful use of the word "substantial" difference as the simple range test does not allow the statement that there is or isn't a statistical "significant" difference. A statistical test, such as an T-test or F-test, between the two data groups would be required (see *Appendix B, Exchanging Data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

If time and energy persists, repeat the analysis for the home Value and home Age levels in the two clusters. Is there a "substantial" difference in the levels within the two clusters for either the Value or Age data?

Now create cluster maps containing three and four clusters—

CLUSTER HousingDensity_surface WITH HomeValue_surface, HomeAge_surface USING 3 FOR Cluster3_DVA

CLUSTER HousingDensity_surface WITH HomeValue_surface, HomeAge_surface USING 4 FOR Cluster4_DVA



—see any consistency in the partitioning as the number of clusters gets larger?

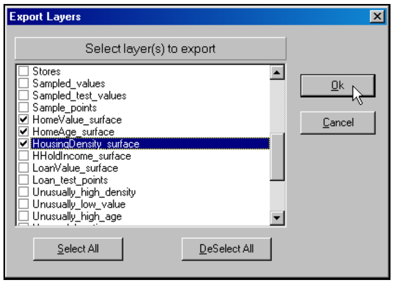
If time and energy persists, use the *Composite* command to calculate the average housing Density, Value and Age within each of the four clusters depicted on the Cluster4_DVA map.

COMPOSITE Cluster4_DVA WITH HousingDensity_surface Average IGNORE PMAP_NULL FOR Cluster4_average_DVA

COMPOSITE Cluster4_DVA WITH HousingDensity_surface StdDev IGNORE PMAP_NULL FOR Cluster4_StDev_DVA

Does there appear to be a progression in the average density among the four clusters? Which clusters are most distinct? Which clusters are least distinct? (Hint: need to consider overlap of cluster means +/- their StDev values).

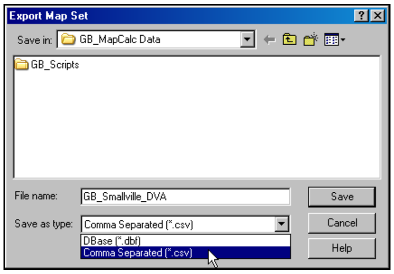
6.5.4 Exporting Data for Further Analysis


Export Layers

Select layer(s) to export

- □ Stores
- □ Sampled_values
- □ Sampled_test_values
- □ Sample_points
- ☑ HomeValue_surface
- ☑ HomeAge_surface
- ☑ HousingDensity_surface
- □ HHoldIncome_surface
- □ LoanValue_surface
- □ Loan_test_points
- □ Unusually_high_density
- □ Unusually_low_value
- □ Unusually_high_age

From the *Main Menu* select **File → Export → Data** to access the wizard for exchanging data. Press the *DeSelect All* button then click on the boxes next to the **HousingDensity_surface**, **HomeValue_surface** and **HomeAge_surface** map layers. Press **OK** to move to the next wizard window.


Export Map Set

Save in: GB_MapCalc Data

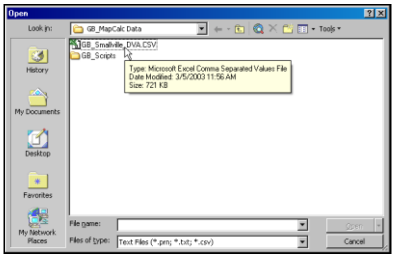
GB_Scripts

File name: GB_Smallville_DVA

Save as type: Comma Separated (*.csv)

DBase (*.dbf)
Comma Separated (*.csv)

Specify the file name as **GB_Smallville_DVA**, the file type as "**CSV**" then press the **Save** button to export the data to the default …\MapCalc data\ folder.


Open

Look in: GB_MapCalc Data

GB_Smallville_DVA.CSV
GB_Scripts

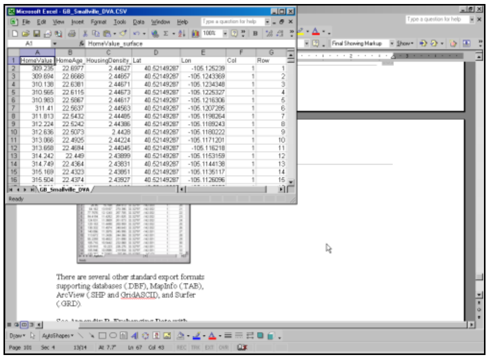
Type: Microsoft Excel Comma Separated Values File
Date Modified: 3/5/2003 11:56 AM
Size: 721 KB

File name:

Files of type: Text Files (*.prn; *.txt; *.csv)

Access *Excel* by clicking on **Start → Programs → Microsoft Excel → File → Open →** browse to the …*MapCalc Data*\ folder → specify **Text Files (*.prn, *.txt, *.csv)** as the file type → click on the **GB_Smallville_DVA.csv** file → and press the **Open** button.

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


Microsoft Excel - GB_Smallville_DVA.CSV
File Edit View Insert Format Tools Data Window Help
A1
HomeValue_surface
GB_Smallville_DVA
Ready

	A	B	C	D	E	F	G
1	HomeValue	HomeAge_	HousingDensity_	Lat	Lon	Col	Row
2	309.235	22.6977	2.44627	40.52149287	-105.125239	1	1
3	309.694	22.6668	2.44657	40.52149287	-105.1243369	1	2
4	310.138	22.6381	2.44671	40.52149287	-105.1234348	1	3
5	310.565	22.6115	2.44673	40.52149287	-105.1225327	1	4
6	310.983	22.5867	2.44617	40.52149287	-105.1216306	1	5
7	311.41	22.5637	2.44563	40.52149287	-105.1207285	1	6
8	311.813	22.5432	2.44485	40.52149287	-105.1198264	1	7
9	312.224	22.5242	2.44386	40.52149287	-105.1189243	1	8
10	312.636	22.5073	2.4428	40.52149287	-105.1180222	1	9
11	313.066	22.4925	2.44224	40.52149287	-105.1171201	1	10
12	313.658	22.4684	2.44045	40.52149287	-105.116218	1	11
13	314.242	22.449	2.43899	40.52149287	-105.1153159	1	12
14	314.749	22.4364	2.43831	40.52149287	-105.1144138	1	13
15	315.169	22.4323	2.43851	40.52149287	-105.1135117	1	14
16	315.504	22.4374	2.43927	40.52149287	-105.1126096	1	15

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

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).