DOVER'S HANDS-ON MAPPING EXAMPLE, CHPATER 1

Techniques covered: Creating a choropleth map
Using a histogram to help determine classification ranges
Joining tables
Experimenting with different classification methods

Step 1: Convert incident counts to rates

Dover knew that mapping involves a series of choices. First, one must select the map type that will most appropriately display the data. To show aggregate-level data, Dover created a choropleth map, which represents areas, such as police jurisdictions, shaded according to their statistical values, such as counts, percentages, or rates. It is better to show **rates per population** rather than counts on a choropleth map because it allows the analyst to compare variables (in this case, vehicle burglaries) with beats that have a different population base. Counts could be misleading because they could show jurisdictions that are similar in the number of crimes, but vastly different in population, with the same shading.

The department maintained current population figures for each police beat, so Dover copied the figures for vehicle burglaries in the spreadsheet. He then divided the variable under study (vehicle burglaries) by the population for each beat. Finally, he multiplied the results by a standardized population base. Dover decided to use 1,000 as his population base. The worksheet contained a column for counts of thefts from vehicles, a column for the population, and a column for the rate per 1,000 persons.

	Vehicle Burglaries	Resident Population	Rate per 1,000 persons
Beat 1	7	3,105	2.3
Beat 2	27	3,171	8.5
Beat 3	16	2,965	5.4
Beat 4	6	3,272	1.8
Beat 5	6	2,901	2.1
Beat 6	13	2,731	4.8
Beat 7	91	1,961	46.4
Beat 8	10	1,845	5.4
Beat 9	10	2,401	4.2
Totals	**186**	**24,352**	**7.6**

Step 2: Examine the various classification methods in the mapping program

Dover's next step was to bring the data into the mapping program and select a **classification method.** A classification method determines how the data will be divided or distributed. Dover knew that the selection of a classification method, along with the number of classifications, was crucial in determining how the data would be viewed on a map. He always found this step a little confusing because there were many options from which to choose. His mapping program presented him with the following classification choices:

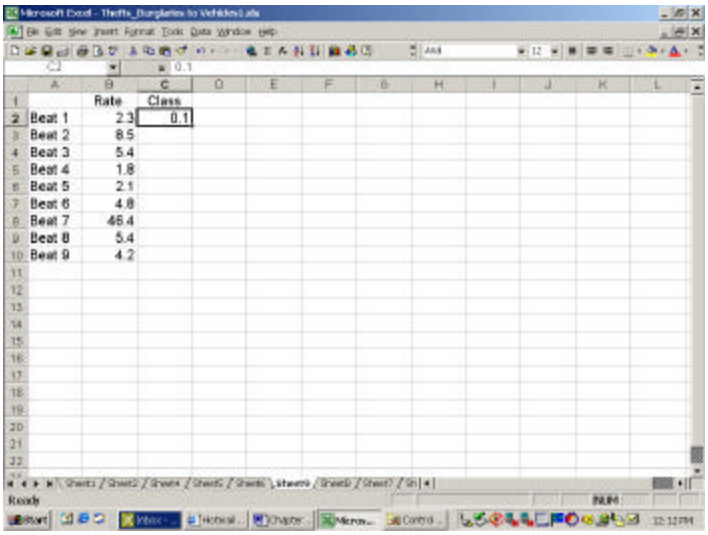
Natural Breaks
Equal Interval
Defined Interval
Quantile
Standard Deviation
Manual

Step 3: Using a histogram to help determine appropriate classification ranges

While the mapping program made it easy to explore the differences between each method, it was left to the judgment and experience of the analyst to choose the most appropriate one. To determine the best method, Dover knew that it was best to understand what the distribution of the data looked like graphically. One of the most useful tools to accomplish this is a histogram, a bar graph of the number of cases in each of several classifications. While many GIS programs provide histogram tools, Dover discovered a way to do it with the Excel spreadsheet program. A summary of the steps are listed below, then detailed in the following pages.

1. Enter data into a column.
2. Create a class column and enter the first class.
3. Use 'Edit', 'Fill Series' to fill the column.
4. Use the function button (*fx*), then pick 'Statistical' in the left column, 'Frequency' in the right column.
5. Enter the data range and the bin or class range. Important: This formula must be entered as an array formula by highlighting the range and entering the formula the formula with 'Ctrl, 'Shift', and 'Enter' at the same time.
6. Mark the 'Class' and 'Count' columns, then use the chart button.
7. Under chart type, select 'Column' graph.
8. Increment the 'Use first X' column button.
9. Turn legends off and label the graph.

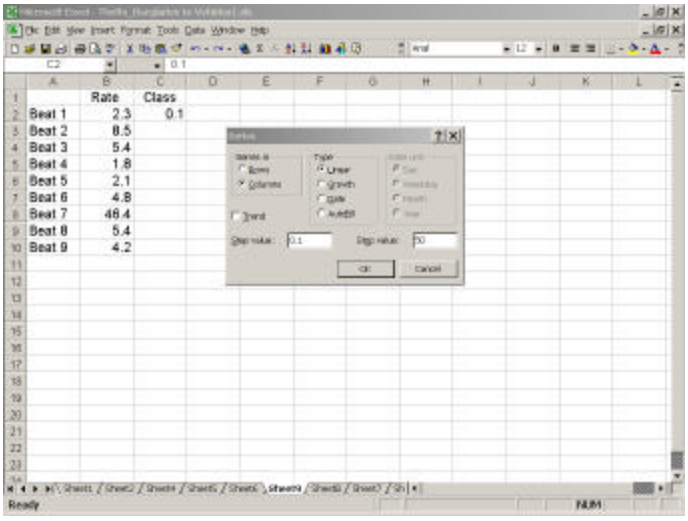
Dover copied the rates for each police beat into a new worksheet. The next step involved creating a class column and entering the first class.



	Rate	Class
Beat 1	2.3	0.1
Beat 2	8.5	
Beat 3	5.4	
Beat 4	1.8	
Beat 5	2.1	
Beat 6	4.8	
Beat 7	46.4	
Beat 8	5.4	
Beat 9	4.2	

By setting the first class at 0.1, Dover indicated where he wanted the X-axis of the graph to begin. He could have set the beginning of the X-axis at 1.8, because that was the lowest rate, but Dover wanted the first number to be offset to make it more visible. In addition, the class number must be in the same format as the rate figures. For those reasons, Dover used a number that was one place to the right of the decimal point.

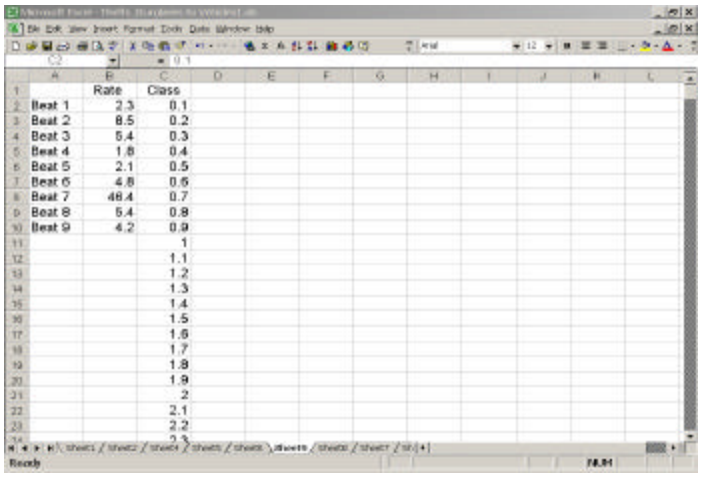
The next step was to create the remaining classes. In this process, Dover would be creating classes from the beginning value to the last value (from 1.8 to 46.4). But Dover wanted to create an offset distance from both ends of the graph. He decided to create classes from 0.1 to 50. Dover selected the 'Edit,' 'Fill,' 'Series' option on the menu bar at the top, which opened this dialog box:



	Rate	Class
Beat 1	2.3	0.1
Beat 2	8.5	
Beat 3	5.4	
Beat 4	1.8	
Beat 5	2.1	
Beat 6	4.8	
Beat 7	46.4	
Beat 8	5.4	
Beat 9	4.2	

In this dialog box, Dover selected 'Columns' in the 'Series in' frame, 'Linear' in the 'Type' frame, set the 'Step value' at 0.1, and the 'Stop value' at 50. If he were to use whole numbers, Dover would have set the 'Step value' accordingly, with a 1 or 2 rather than a decimal. After

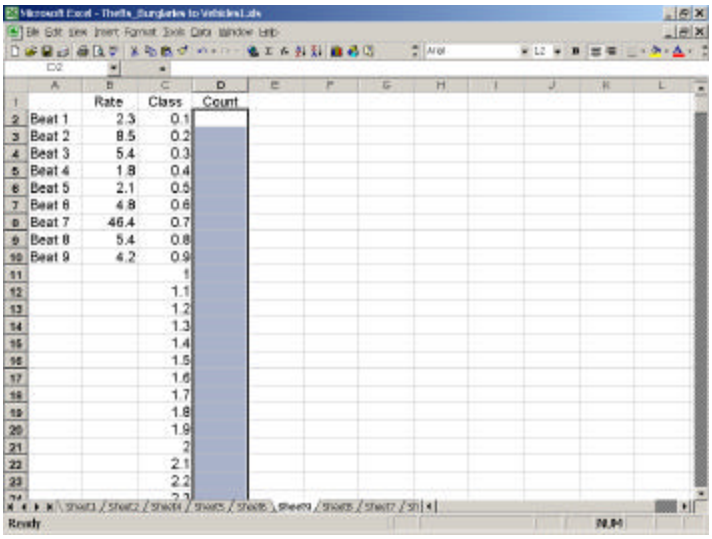
modifying the options according to his specifications, Dover hit ‘OK’. His table appeared with classes from 0.1 to 50, or in spreadsheet terminology, cells C2:C501 (as shown below).



	A	B	C
1		Rate	Class
2	Beat 1	2.3	0.1
3	Beat 2	8.5	0.2
4	Beat 3	5.4	0.3
5	Beat 4	1.8	0.4
6	Beat 5	2.1	0.5
7	Beat 6	4.8	0.6
8	Beat 7	46.4	0.7
9	Beat 8	5.4	0.8
10	Beat 9	4.2	0.9
11			1
12			1.1
13			1.2
14			1.3
15			1.4
16			1.5
17			1.6
18			1.7
19			1.8
20			1.9
21			2
22			2.1
23			2.2

The next step was to calculate the frequency, or counts, of cases for each class. In this step, Dover knew that every class would be given a count of zero, except for the classes that coincided with his rate values. Among those values, there was only one repeating value, the rate of 5.4 in Beat 3 and Beat 8, so that would be given a count of 2, while the rest of the values would be given a count of 1.

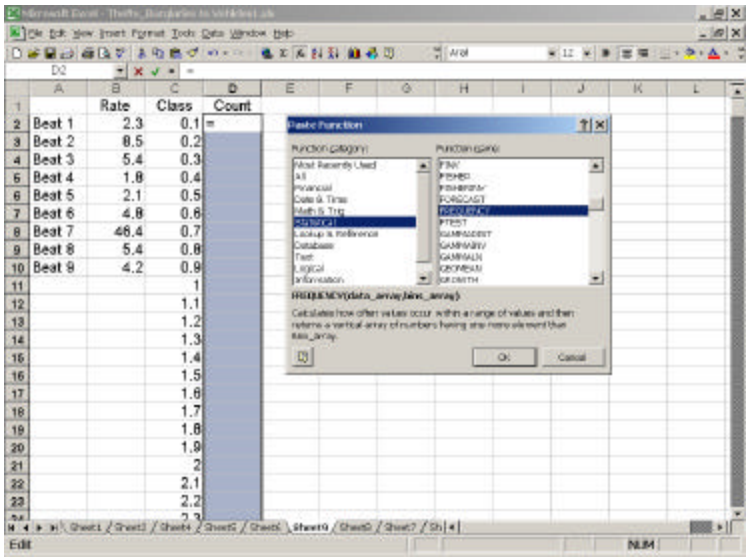
To calculate the frequency would require a statistical function. But before opening one, Dover knew there was one tricky step in calculating a frequency using the spreadsheet program. The formula had to be entered as an array formula – he had to enter the formula into all cells at the same time with ‘Control’, ‘Shift’, and ‘Enter’ on the keyboard. In a mouse-driven windows environment, this seemed outmoded to Dover. But after performing the routine a couple of times, he got the hang of it. In order to prepare for this trick, Dover first created a field in his worksheet called “Count” and then highlighted the entire range where the counts would be placed, from cells D2:D501 (as shown below).



	A	B	C	D
1		Rate	Class	Count
2	Beat 1	2.3	0.1	
3	Beat 2	8.5	0.2	
4	Beat 3	5.4	0.3	
5	Beat 4	1.8	0.4	
6	Beat 5	2.1	0.5	
7	Beat 6	4.8	0.6	
8	Beat 7	46.4	0.7	
9	Beat 8	5.4	0.8	
10	Beat 9	4.2	0.9	
11			1	
12			1.1	
13			1.2	
14			1.3	
15			1.4	
16			1.5	
17			1.6	
18			1.7	
19			1.8	
20			1.9	
21			2	
22			2.1	
23			2.2	

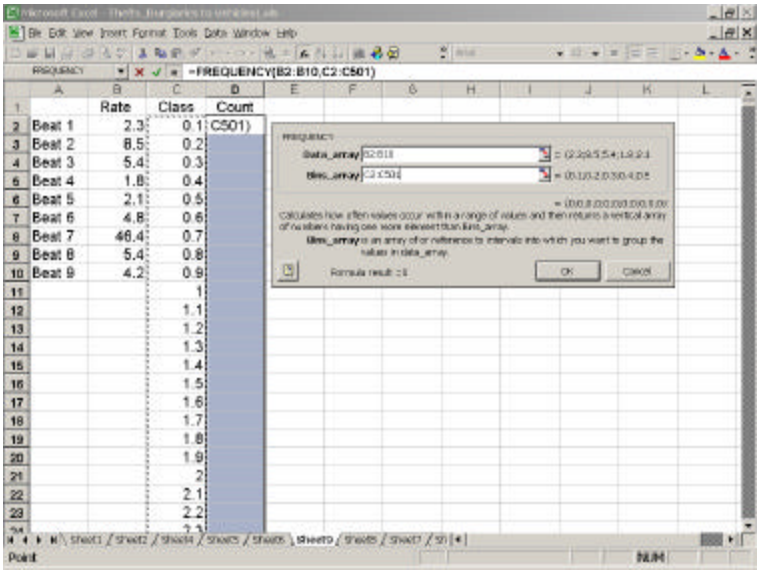
Dover was ready to calculate a frequency. He started a statistical function in his spreadsheet program by clicking on the '*fx*' button, which looks like this:

The '*fx*' button is located on the toolbar next to the sigma, '*E*'. Once the '*fx*' button was pressed, the 'Paste Function' screen opened (as shown below).



Dover chose the 'Statistical' function from the left category column, and the 'FREQUENCY' function from the right name column.

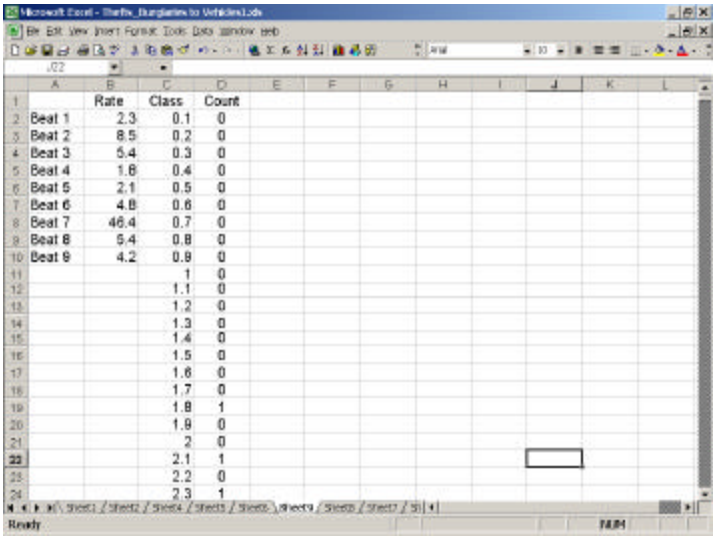
Dover hit 'OK' and the program requested the range of data to calculate the frequency (as shown below).



The range of data consists of the cell locations of the data to be arranged. In this situation, the range was read as B2:B10 and it contained the vehicle burglary rates by beat. He pressed the button located to the right of 'Data_array' box. He entered the range by highlighting the cells, and hit 'OK.'

The program then asked him for the array of reference values to group the rates, or 'Bins array'. He pressed the button to the right of that box, then highlighted the cells in the 'Class' field. It was read as C2:C501.

At this stage, Dover was ready to complete the process by entering the formula into all relevant cells in the 'Count' field (column D). Without pressing the 'OK' button, he held down the 'Shift', 'Ctrl', and 'Enter' keys on his keyboard at the same time. It produced counts for each class (as shown below). As illustrated in the table, only classes that corresponded to vehicle burglary rates contained a count greater than zero. Note the first count of 1 is positioned across from 1.8.

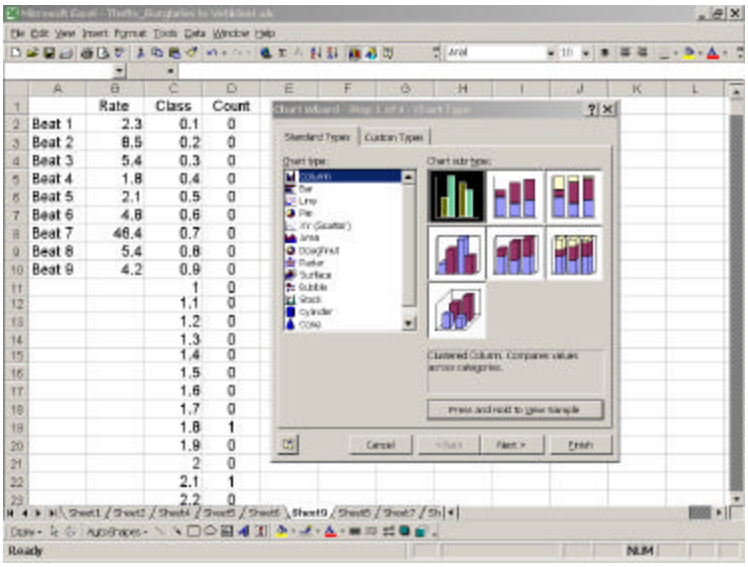


	Rate	Class	Count
Beat 1	2.3	0.1	0
Beat 2	8.5	0.2	0
Beat 3	5.4	0.3	0
Beat 4	1.8	0.4	0
Beat 5	2.1	0.5	0
Beat 6	4.8	0.6	0
Beat 7	46.4	0.7	0
Beat 8	5.4	0.8	0
Beat 9	4.2	0.9	0
		1	0
		1.1	0
		1.2	0
		1.3	0
		1.4	0
		1.5	0
		1.6	0
		1.7	0
		1.8	1
		1.9	0
		2	0
		2.1	1
		2.2	0
		2.3	1

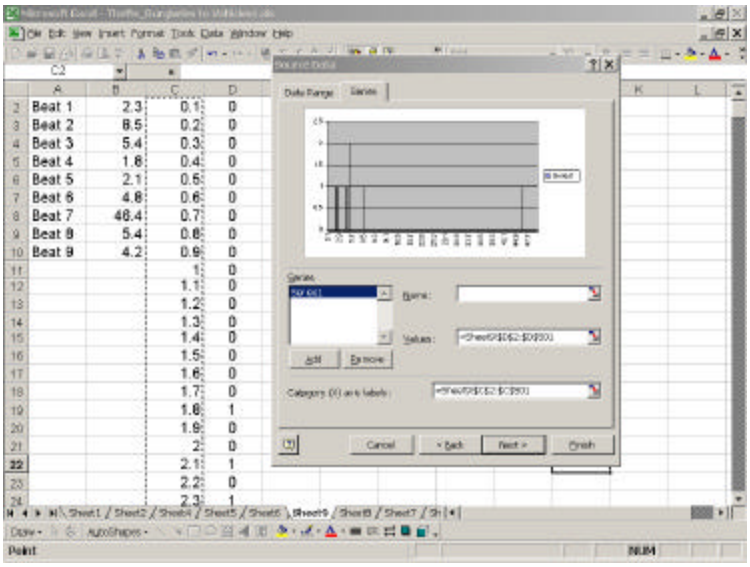
Using the data in the 'Count' field, Dover was ready to create a histogram. The process was initiated by clicking the Chart button at the top of the screen (as shown below).

This brought up the chart wizard (as shown below). The chart wizard then created a graph in four easy steps.

In step one, to create a histogram, Dover had to use the 'Column graph' option. He selected the column graph from the 'Chart type' and the 'Clustered Column' type (top left option) from the 'Chart sub-type', and hit 'Next .'

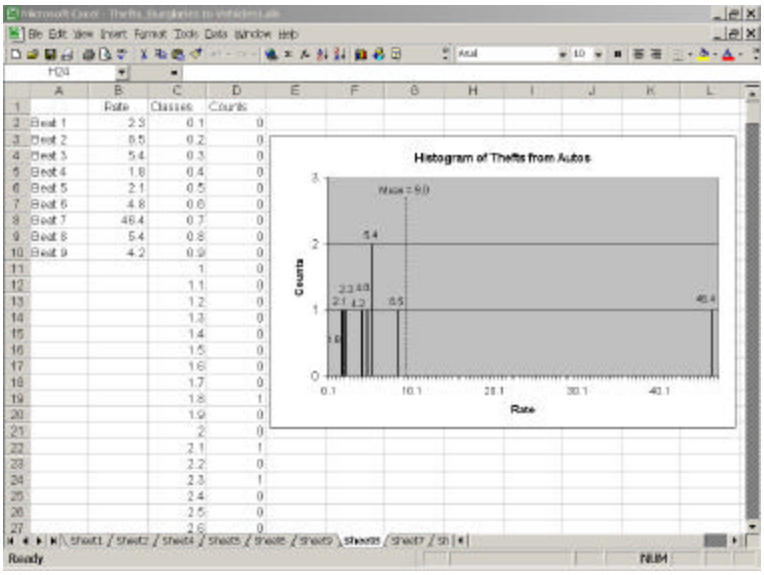


This moved him to the chart ‘Source data’ screen for step 2. In this screen, Dover selected the ‘Data Range’ and set an X-axis label. Dover selected the ‘Count’ field range and labeled the X-axis by selecting the range of values in the ‘Class’ field. At this stage, the histogram looked like this:



Dover then pressed ‘Next’ and moved to step 3, the ‘Chart Options’ screen. Here, Dover typed in a title, labels, and made other modifications to the format of the histogram.

Dover pressed the ‘Next’ button one last time, and set the location for the chart (either as a new tab or an embedded object in the spreadsheet). The finished product looked like this:



After creating the histogram, Dover analyzed it carefully, with an eye toward the best way to graphically represent the data on a map. He noticed the distribution was severely skewed in a positive direction (to the right) as a result of the high rate in Beat 7 (46.4). Beat 7 would have to be mapped in its own class. The histogram also gave him a visual aid that would help determine the other categories as well. He identified three additional groupings from the histogram based on how they clustered together along the number line in the graph. He concluded that the map should contain four groupings:

Group 1:	1.8 – 2.3
Group 2:	4.2 – 5.4
Group 3:	8.5
Group 4:	46.4

Step 4: Create choropleth map by joining tables

Armed with the results from the histogram and confident in his ability to select the most suitable classification method, Dover was ready to take on the process of mapping the data.

The next step was to display the vehicle burglary rates by police beats on a map. At this point, there was no link between the data in his spreadsheet and his map of police beats. However, a link could be established in the GIS software program by **joining** or **relating** the two tables.

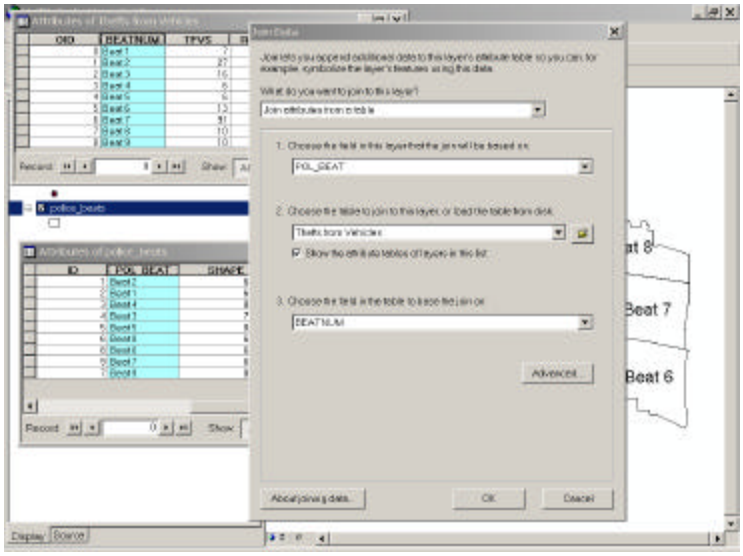
To begin this process Dover opened both the **attribute table** containing vehicle burglary rates and the attribute table for Beaufort beat boundaries in his mapping program. An attribute table is information about features on a map, stored in rows and columns. Each row relates to a single feature, and each column contains the values of a single characteristic.

In the first step, Dover determined which fields the *join* or *relate* would be based. First, they must be based on values that can be found in both tables. In this case, the common values were the beat names, such as “Beat 1”, “Beat 2”, etc. The values needed to be of the same data type. In

other words, a text field can only be joined or related to another text field, a number to a number, a string to string, and so on. In contrast, the field names do not have to be the same. Thus, Dover performed his *join* or *relate* on the 'BEATNUM' field in the attribute table containing vehicle burglary rates and the 'POL_BEAT' field in the beat attribute table (as shown below).

Next, Dover decided which type of link was appropriate, a *join* or a *relate.* Based on the technical documentation that came with his GIS program, a *join* is appropriate for a one-to-one relationship or a many-to-one relationship between the two tables, and a *relate* is appropriate for one-to-many or many-to-many relationships. Since there was only one record per beat, this illustrated a one-to-one relationship between the two tables. A *join* was the way to go.

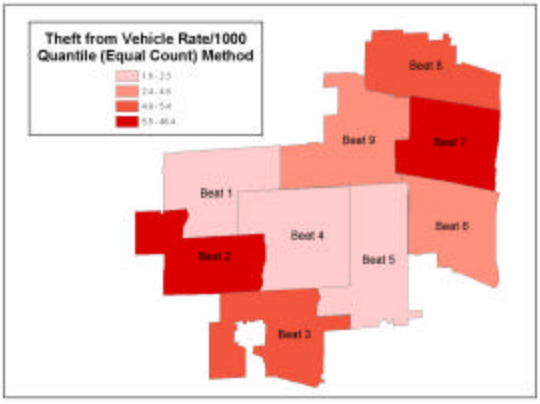
In the final step, Dover selected the beat layer with his mouse to set that table as the target destination for the *join*. He then selected the 'Join' option in his program. Next he was prompted to choose the fields on which to base the *join*, and the name of the table to join to the beat layer (as shown below).

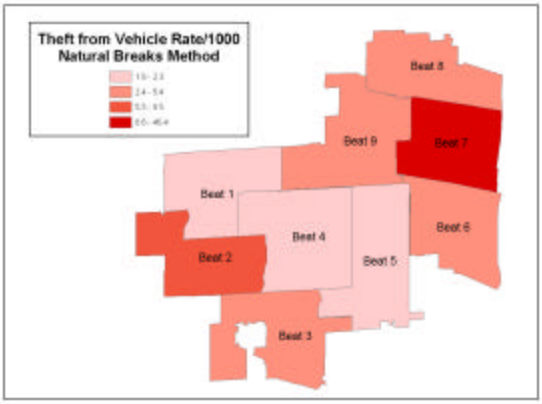


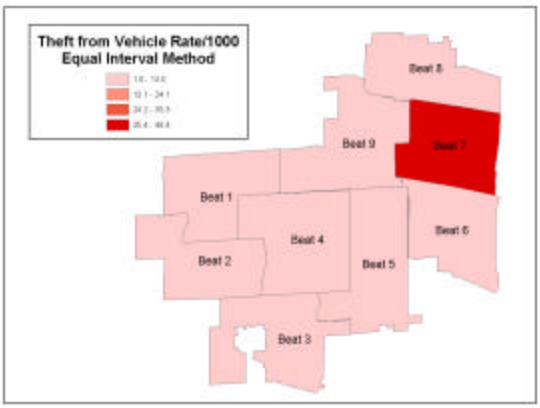
After entering the correct information, Dover hit 'OK' and the data fields in the 'Theft from Vehicles' table were appended to the 'Police beats' layer. This process extended the information about the police beats, thereby enabling Dover to symbolize the beats using vehicle burglaries.

Step 5: Selecting a classification method

Dover then experimented with the various classification methods provided by his program. One by one, he selected a method, set the number of classes and viewed the results, all the while comparing the map to the classification scheme he derived from the histogram. The style he chose showed each beat in the map with a gradation of the color red. Higher values were represented by darker red, and lower values were represented by lighter shades of red. By the end of the process, Dover had compared three different classification methods, *natural breaks, equal interval*, and ***quantile*** (equal count) and how they represented the data on a map (as shown below).







After evaluating the various options, it was evident that the *natural breaks* method was the best choice, because the categories matched the breakdowns on the histogram. The natural breaks scheme works by looking for natural breaks, or gaps, in the distribution. Over time, Dover found that *natural breaks* was usually one of the best methods when creating area maps.

Neither of the other two methods worked as well for this distribution. The *equal interval* method represented four categories in the legend, but only two categories on the map. *Equal interval* divides the range of data values into equal-sized sub-ranges using the formula *(highest value – lowest value) / number of classes.* The problem when applied to the distribution of rates for vehicle burglaries is that the first range (from 1.8 – 13.0) encompassed every data point except the highest rate in Beat 7. In one way this was good because it accurately made Beat 7 stand out from the rest, but it obscured the natural gaps between values in the remaining beats. Because vehicle burglary rates were right-skewed due to the relatively high rate in Beat 7, the *equal interval* method favored the lower values. Another problem with this method was that it created two middle ranges of non-existent data on the map. *Equal interval* works best with data sets that do not have large gaps between values.

The *quantile* method puts approximately the same number of observations in each class using the formula *number of observations / number of classes.* This scheme forced Beat 2 into a group with Beat 7. If Dover learned anything from creating the histogram, it was that Beat 7 was in a class by itself. Dover also found through experience that the *quantile* method usually worked best with ordinal (ranked) data used when he wanted to compare, for instance, the top 25 percent to the bottom 25 percent of a group.

While Dover chose the *natural breaks* classification method to be the best, he still had one problem to tackle – the legend it created was misleading. People reading the legend might get the impression that the data was continuous between the lowest value of 1.8 and the highest value of 46.4. Unlike the actual distribution, there were no gaps in the legend. This was especially problematic for Beat 2 and Beat 7, because they were individual classes with one value instead of classes with a range of values.

Step 6: Making the legend more informative

In the final step, to make the legend more informative, Dover changed the labels for each range to match the lowest and highest values that they actually contained. For Beat 2 and Beat 7, Dover simply used the single value as the label. This was the same scheme he derived from observing the histogram. That way, he reasoned, there would be no room for misinterpretation. The ranges were changed accordingly (as shown below)

Old Classification:	New Classification:
1.8 – 2.3	1.8 – 2.3
2.4 – 5.4	4.2 – 5.4
5.5 – 8.5	8.5
8.6 – 46.4	46.4