W3 Exercise: Classification schemes in choropleth mapping

40 points – Choropleth mapping, color classification

Objective

The purpose of this exercise if to familiarize yourself with different methods of classing attribute data, and to determine the impact of varying number of classes and type of classing schemes on the visual presentation of your data. At the end of this exercise, you will be making choropleth maps of some quantitative theme, processed and classified with the CLASSIT application. You will make four maps in ArcPro using four different methods of classing your data, specifically: Classit-based, Equal Interval, Quantiles and Geometric.

PART 1

1 Accessing data and preprocessing:

You can use the data on cancer rate at the county level that I have posted on canvas, or you can use your own data, or access data from the web (at the census for instance). However, for the purpose of this exercise I will demonstrate the exercise using lung cancer rates for males in western counties of the US.

Please note: For choropleth maps, you should always use standardized data. Just for a moment think about standardization. To map population at the county level for instance, you can standardize by county size, and you can get this information using a geometrical attribute in ArcPro. You may want to standardize by some other measure, in some cases for example number of physicians per square mile is not logical, but number of physicians per 1000 population is logical. Data on ethnic breakdowns is conventionally standardized by total population, data on household economic measure is standardized by total number of households, and so forth. Make sure that you have access to the field you will normalize (standardize) the data with. For instance, if you are interested by a particular race (say, Hispanic), make sure you also download the Total Population field, so you can normalize.

1.1 Data preprocessing in Excel

To use the classification software CLASSIT, you will need to have your data in a tab-delimited (text file) format. The best is to do this from Excel. You can open the dbf file from your shapefile for instance, and arrange your data from there.
The format we need to import into the CLASSIT includes 3 data columns:

1. A polygon identifier (POLID)
2. The variable you want to find optimal categories for

3. A standardizing variable i.e. polygon size in square kilometers, total population, or any other numeric standardizing information. (If your data is already standardized, e.g. percent, density, etc... you do not need to add a standardizing variable- just replace the third column with values of 1).

Make sure that the variable labels do not have spaces in them (keep them short, max. 8 characters). When the file is complete, go to SAVE AS . . . and save it as Text Tab delimited. Give it a new file name to represent the variable you have chosen to work with. For example, if your variable is Single Head of Household, you may want to call it HHOLDS.txt.

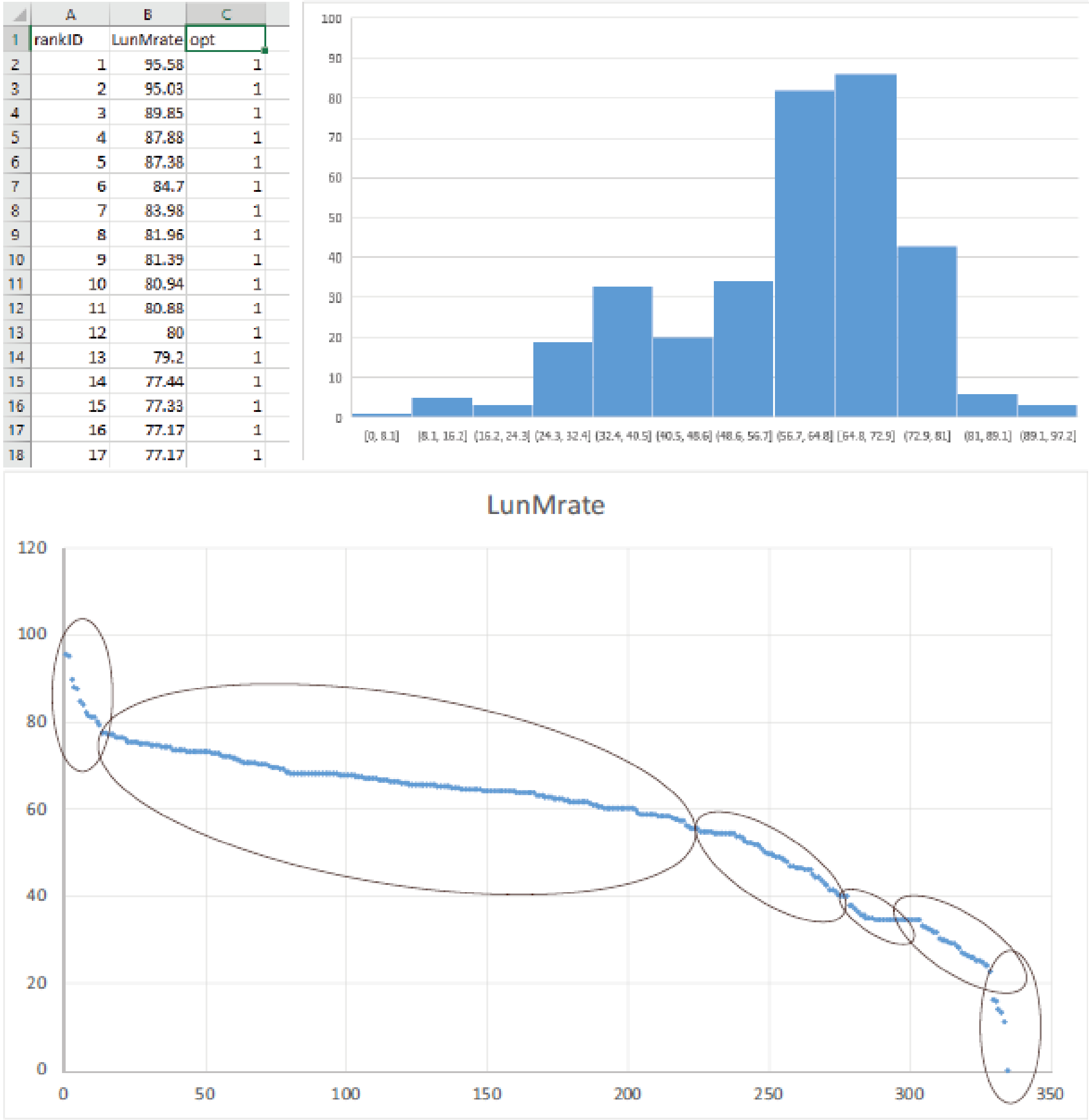
1.2 Distribution of your chosen variable.

We would like to find out what classification scheme to use, and how many classes the map should contain, to most appropriately represent our data. There are many ways to go about this, and it depends mostly on the characteristics of your data, and purpose of the map. Here are a couple of questions to guide your decision making: How is the data distributed? Does it follow a normal distribution, or is it unevenly distributed over the geographic area? Do you want to emphasize differences or clusters in your data? To answer these questions, it is a good idea to visually inspect the data distribution using a graphing tool. In addition, we need to get an idea of how many classes to use.

<u>Visual Inspection using Excel</u>: We will apply a visual inspection approach for this lab. What you will need to do in Excel is to standardize the data if it is not done already. Sort your data (in descending order) and create a new field that will serve as rank. It will basically rank your data, depending on the percentage of farmers for each county. Make a chart of this sorted list by clicking on the Chart Wizard in the toolbar (button with a little chart icon and a wand on top of it). Make a simple column chart. Now look at your chart, where do you see natural breaks occurring? You can resize the chart to inspect the breaks better. An example is given in Figure 1.
How many breaks can you identify? Determine the number of breaks that represents your own data best, and take note of it. Use the labeling tool in Excel, and put the class number you just identified on the chart (e.g. number of classes = X, where X is the number of your choice). You will need this information again for the second part of the exercise to set the number of classes for your maps. **Do a print-screen of the chart, include it in your write-up and give some explanations.**

1.3 Using CLASSIT to analyze your data distribution.

Another way to analyze your data distribution is by using a small data classification program, called CLASSIT and developed by B. Buttenfield (Coulder, Colorado). In order to do this, you will use the data you are working with, but this data must be in a text file! As I mentioned earlier, it is important that this file only contains 3 columns. If your data is already standardized, the 3rd column just contains 1s, heading is = optional. You can find the application in the zipped file from the assignment. CLASSIT will optimize for deviations from the mean and from the median. You can work with either one of them. Copy the program CLASSIT to your own directory or USB drive where you newly data file resides.



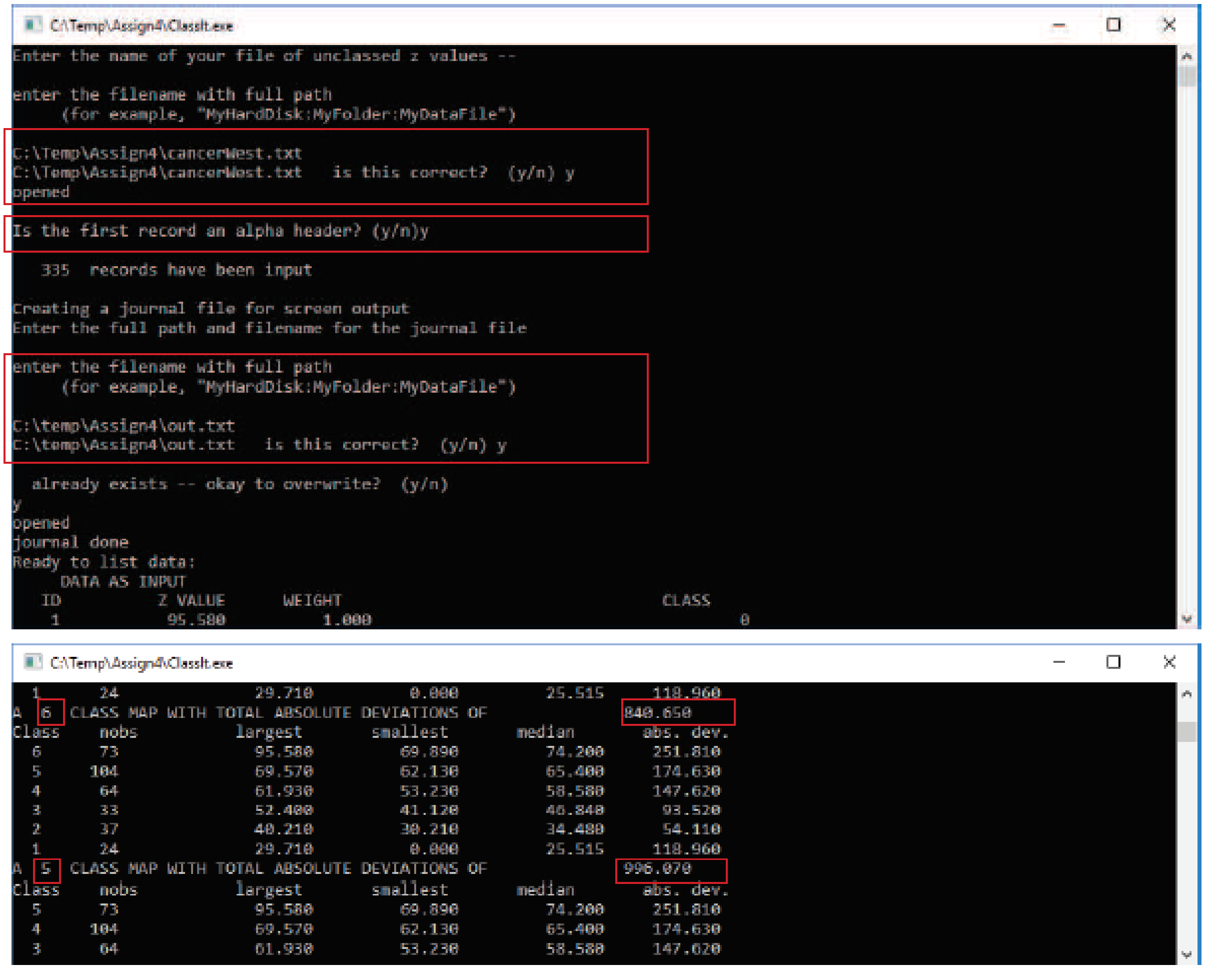
	A	B	C
1	rankID	LunMrate	opt
2	1	95.58	1
3	2	95.03	1
4	3	89.85	1
5	4	87.88	1
6	5	87.38	1
7	6	84.7	1
8	7	83.98	1
9	8	81.96	1
10	9	81.39	1
11	10	80.94	1
12	11	80.88	1
13	12	80	1
14	13	79.2	1
15	14	77.44	1
16	15	77.33	1
17	16	77.17	1
18	17	77.17	1

Figure 1: Analyzing the distribution of your data.

The windows interface is not accessible from CLASSIT, so you have to remember your own filenames. Do not use the path example shown within CLASSIT. A correct file path on Windows is for instance:

C:\GeoVis\HWK3\cancer.txt

The top bar of the application window will show an example of the corrected path. CLASSIT will ask you what file to open (Figure 2), it will ask you if your first record is an alpha header is an alpha header (probably yes). It will also ask you what to call your journal file -this is a log of program results (output). For example, if you input file is cancer.txt, call your journal file cancer_JNL.txt orcancer_LOG.txt. Or something to associate the input data with the journal log. Request a 10 class solution (median or mean), and the Brief option. When you are asked at the end of the program if you want to save information for these classes, respond no. After running the program, open Microsoft Word, then open the journal file and include a copy in your write-up. Look at the table of solutions that lists class breaks with variance or with absolute deviations for the 10,9,8 etc... class solutions. For each number of classes, a total deviation value is printed. Open Excel again. Open a new file, and enter all variance (total deviation) for each class starting from 2 from your CLASSIT output file. Column one should show the classes 3-10 and column two show the variance. Make a small line graph (in Excel) plotting the magnitude of the total deviation (y-axis) against the number of classes (3-10) (Figure 3).

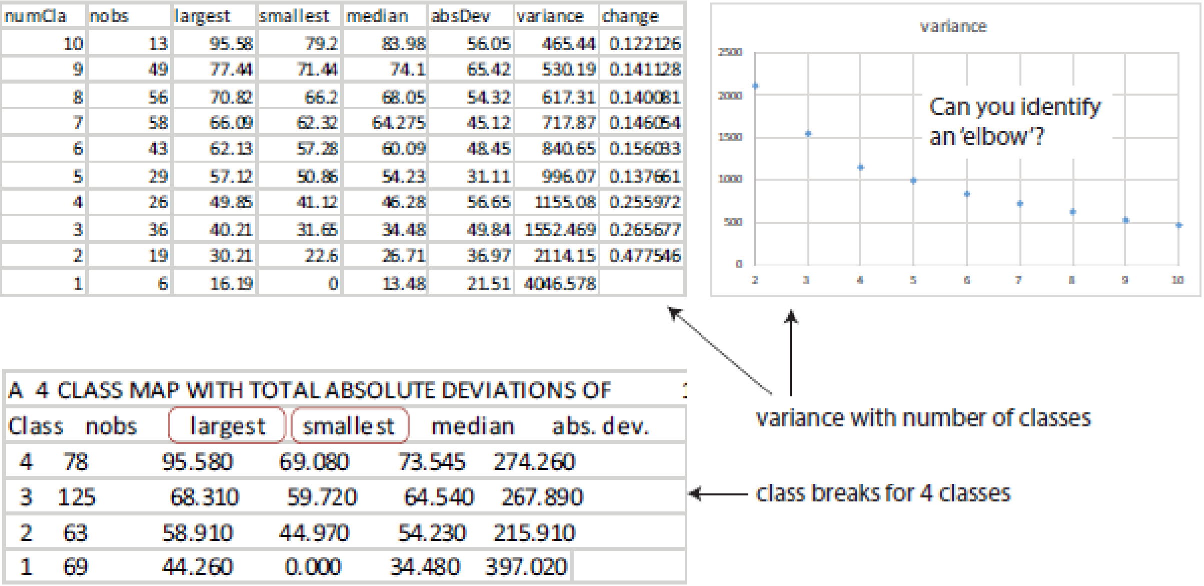


```
Enter the name of your file of unclassed z values --

enter the filename with full path
     (for example, "MyHardDisk:MyFolder:MyDataFile")

C:\Temp\Assign4\cancerWest.txt
C:\Temp\Assign4\cancerWest.txt   is this correct?  (y/n) y
opened

Is the first record an alpha header? (y/n)y

   335  records have been input

Creating a journal file for screen output
Enter the full path and filename for the journal file

enter the filename with full path
     (for example, "MyHardDisk:MyFolder:MyDataFile")

C:\temp\Assign4\out.txt
C:\temp\Assign4\out.txt   is this correct?  (y/n) y

  already exists -- okay to overwrite?  (y/n)
y
opened
journal done
Ready to list data:
     DATA AS INPUT
   ID           Z VALUE      WEIGHT                                 CLASS
    1            95.580          1.000                                       0
```

```
  1      24               29.710          0.000        25.515     118.960
A  6  CLASS MAP WITH TOTAL ABSOLUTE DEVIATIONS OF                 840.650
Class    nobs          largest        smallest       median       abs. dev.
  6      73               95.580         69.890        74.200     251.810
  5     104               69.570         62.130        65.400     174.630
  4      64               61.930         53.230        58.580     147.620
  3      33               52.400         41.120        46.840      93.520
  2      37               40.210         30.210        34.480      54.110
  1      24               29.710          0.000        25.515     118.960
A  5  CLASS MAP WITH TOTAL ABSOLUTE DEVIATIONS OF                 996.070
Class    nobs          largest        smallest       median       abs. dev.
  5      73               95.580         69.890        74.200     251.810
  4     104               69.570         62.130        65.400     174.630
  3      64               61.930         53.230        58.580     147.620
```

Figure 2: The command-line classit software.

Save your excel file, and include a screenshot of your 'elbow' graph in your write-up, and explain how many classes you will end up choosing, and why. As in the figure, variance will increase as fewer classes are formed, always (Figure 3). Notice that the graph does not include variance for the one-class solution -the magnitude of this value would obscure the pattern shown in the graph below, so it is left off. What you are looking for is a jog, or elbow. This provides an indication that maybe a 4 or 5 classes solution might be the most desirable choice for optimizing class breaks. Think about it this way.



numCla	nobs	largest	smallest	median	absDev	variance	change
10	13	95.58	79.2	83.98	56.05	465.44	0.122126
9	49	77.44	71.44	74.1	65.42	530.19	0.141128
8	56	70.82	66.2	68.05	54.32	617.31	0.140081
7	58	66.09	62.32	64.275	45.12	717.87	0.146054
6	43	62.13	57.28	60.09	48.45	840.65	0.156083
5	29	57.12	50.86	54.23	31.11	996.07	0.137661
4	26	49.85	41.12	46.28	56.65	1155.08	0.255972
3	36	40.21	31.65	34.48	49.84	1552.469	0.265677
2	19	30.21	22.6	26.71	36.97	2114.15	0.477546
1	6	16.19	0	13.48	21.51	4046.578	

Can you identify an 'elbow'?

variance with number of classes

A 4 CLASS MAP WITH TOTAL ABSOLUTE DEVIATIONS OF

Class	nobs	largest	smallest	median	abs. dev.
4	78	95.580	69.080	73.545	274.260
3	125	68.310	59.720	64.540	267.890
2	63	58.910	44.970	54.230	215.910
1	69	44.260	0.000	34.480	397.020

class breaks for 4 classes

Figure 3: Results from the optimization classification scheme using classit. Look for an elbow when you map the variance.

PART 2

2 *Choropleth mapping*

2.1 *Four data frames*

In order to make 4 maps and compare them, you will need to create four data frame. Each frame is essentially a new layout in ArcPro, and I will demonstrate an easy way to get to this. Rename each map based on the technique you are going to use for the classification scheme. The four techniques are Classit, Equal Interval, Quantiles and Geometric. For the first one, we will define class breaks based on the results from CLASSIT (look your journal log). Remember that the number of classes must be the same, based on your results from CLASSIT. Eventually, copy the shapefile (right-click, copy), and right-click on the data frame, and select Paste layer.

2.2 *Making the maps*

Select an appropriate sequential color scheme for your map. It can be gray, or colorful, check colorbrewer for judicious choices (you can also use the color brewer style that I have posted on canvas). Remember that a darker color is associated with a higher value (higher percentage), and a lighter color a lower percentage. You will make 4 maps. The color scale and the number of classes - defined in Part1!- must not change, so that you can visually compare the results of the different partitioning schemes. Obviously, the class breaks will change as a function of the method you will use. Additionally, make sure that the number of decimals does not exceed 2 digits after the comma.

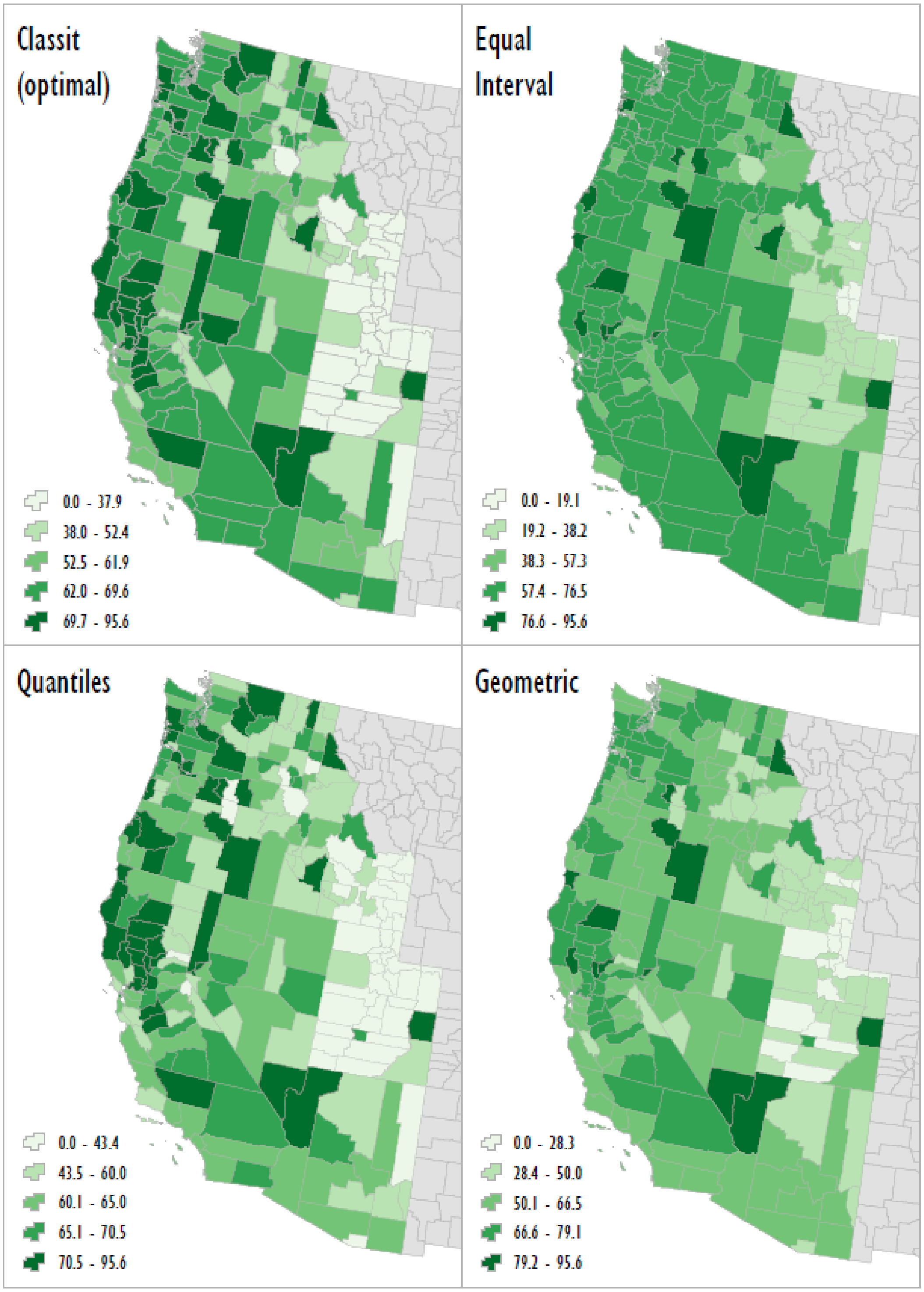


Figure 4: Lung cancer rates among males in western US counties, for four different classification schemes.

3 Layout

Once you have made the different maps, you will create a layout that will contain the 4 maps, and a legend. Go under FILE>MAP PROPERTIES and select Landscape. Arrange the 4 maps in a mosaic format. Note that you need to be consistent in the scale (all maps need to be at the same scale). Additionally, for each data frame insert the legend (but only the different classes). Right click on the legend and select convert to graphics. Then right-click again on the legend and select UNGROUP. Get rid of the legend title and the header text. Select then only the remaining elements (i.e. the little rectangles with their values to the right) and place it appropriately. Check lecture on cartographic design for advices. Repeat the same procedure (insert legend, etc...) for the 3 other maps. Once you are done, export the map in JPEG, and include it in the word document

4 Map analysis

Last step, now. Try to decide what is the advantage or disadvantage of each map classing scheme. What can you say about them? What do you know now about the geography of your variable that you did not know before? Type a 10lines essay answering these questions