02.221 – Lab 5: Data Classification

The data needed for this lab can be found on the course website or downloaded from Dropbox: https://www.dropbox.com/s/gko8z0g0piziil1/02221-Lab5.zip. Extract the data from the zip file to an appropriate location within your documents.

Goals

The primary goal of this lab to learn about the different ways to symbolize and classify thematic map data. We will do so by creating two separate map products. The first is a series of maps of global GDP per population using a sequential color scheme and different modes of determining class breaks. For the second map product, we will compare Twitter usage in Singapore in 2015 with usage in 2012 and we will use a diverging color scheme with manual class breaks that are appropriate for the underlying data/indicator.

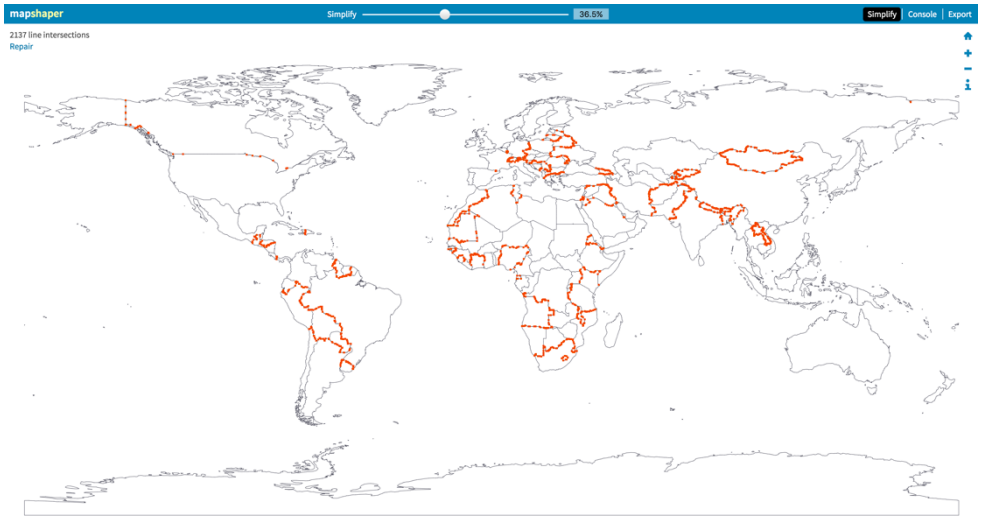
Line Simplification

Start by downloading the Natural Earth Admin 0 1:50 country data (or use the downloaded data from one of your previous assignments). With secondary spatial data, like the country polygons here, often the lines or polygons have much more complexity than needed for a particular visualization. For example, detailed coastlines are not relevant in most thematic world maps – yet, the details on rugged coastlines may distract the map reader's eye.

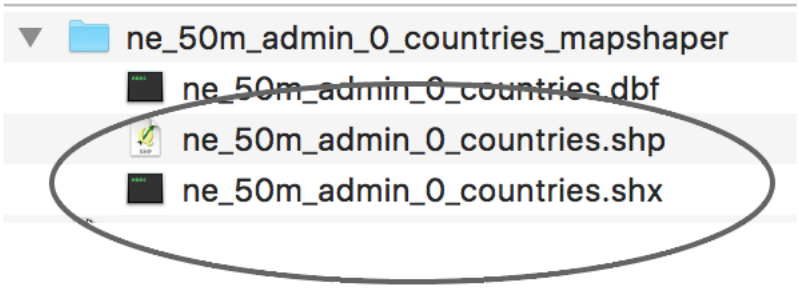
To circumvent this issue, we can simplify the line work prior to visualizing. This should only be done for visual reasons and *after* other analyses have been completed. This is because simplification will alter the geometric attributes (e.g. area) of each spatial object.

Although simplification can be performed within QGIS, it does not provide any 'live' visual feedback. Instead, we use the website mapshaper.org to complete the simplification as it has immediate visual feedback, making it much easier to reach a balance between detail and simplification. Go to mapshaper.org and select or drag the

shapefile onto the website. *Try to understand why you can just use the .shp file in this case and do not need any of the other sidecar files.* After the file is added, click simplify and play around with the parameters until you hit a good balance between simplification and maintaining essential detail. There are no specific rules here: what is appropriate simplification depends on your data, the goal of your map and your own preferences.



Export and download the resulting file in Shapefile format. You can now replace the 'old' spatial features with the new simplified features and their index.



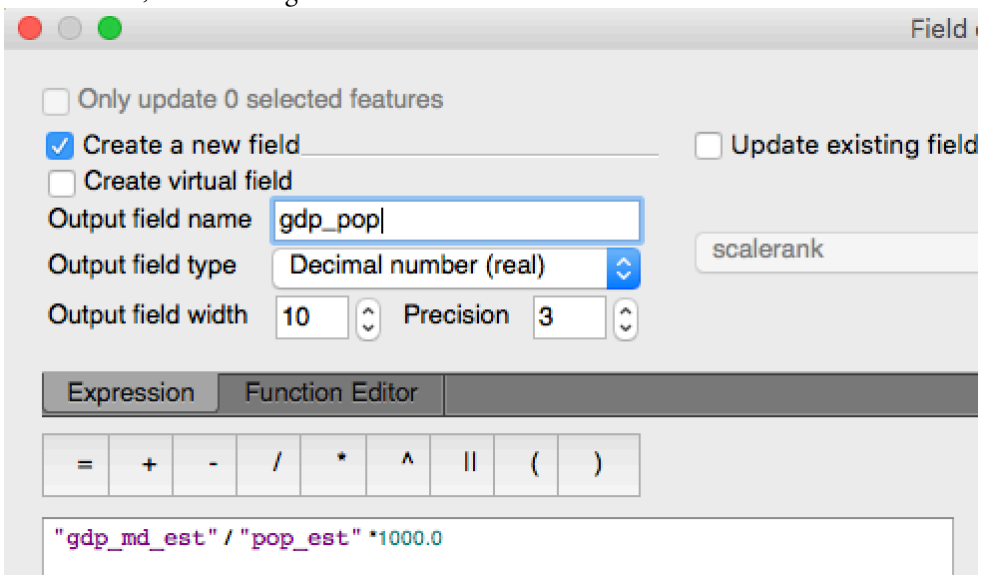
Use these two files to replace the 'old' shp/shx files

You have now created your own, custom simplified country layer. Check out the results in QGIS.

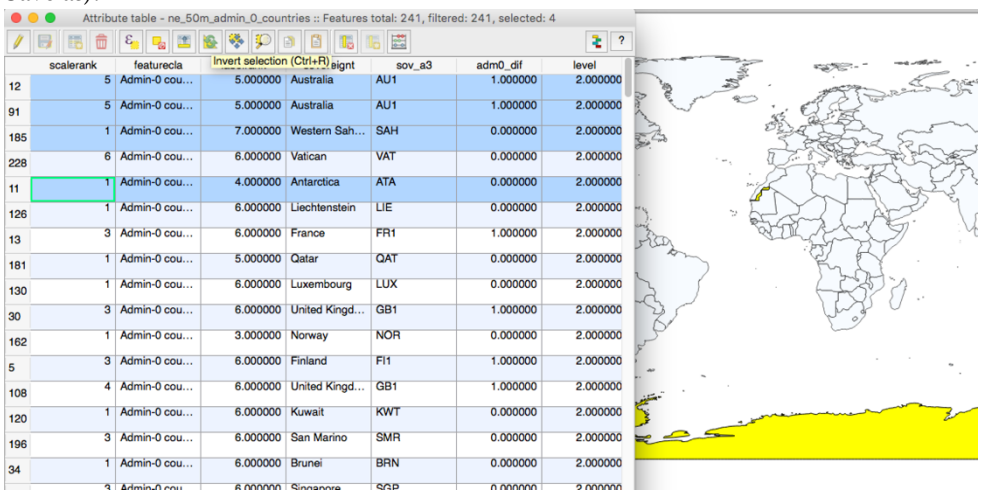
Data Classification: Sequential Data

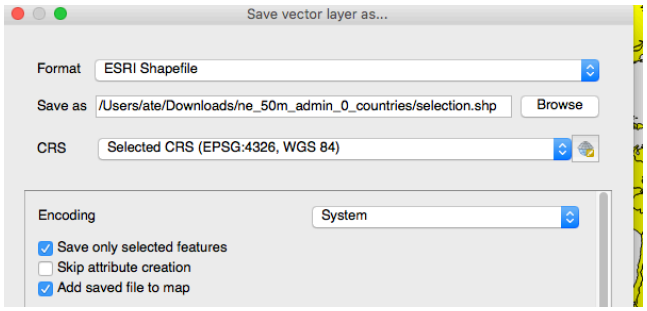
We will now use this country layer to create a series of maps of global GDP. As you know, the Natural Earth data has a variable that contains the GDP in each country. The GDP is a good indicator for a country's economic prosperity. However, it is an absolute amount of dollars that circulate in each country's economy in a certain year. As such, larger countries with a larger population naturally have higher GDPs. To compensate for this, it makes more sense to look at each country's GDP normalized

for population. Use the field calculator to create a new field that holds the GDP per 1000 people. N.B. if you only have integer fields in your calculation, QGIS might convert the result of your expression to integer as well. To prevent that, you can multiply by 1000.0 instead of 1000. As long as one variable in your expression is a real number, the resulting variable will be as well.



If you visualize the results of your field calculation as graduated color or look at the attribute table, you will see that there a few results that are questionable. For example, Western Sahara and Antarctica are not really countries but have really high values for the GDP per population. In this instance, it would be better to exclude them from your map or, alternatively, set their values to zero. You can, for example, do so by selecting these outliers in the attribute table and using the 'Invert Selection' tool. Subsequently, you can save the resulting layer with just the selection (Right-click | Save as).





When symbolizing a variable with graduated colors, QGIS allows you to use different numbers of classes and different ways of deriving class breaks. In the next section, you will experiment with class breaks and create four maps of the GDP per population variable with different class breaks. Pay attention to the following and use Making Maps pp. 155-163 as a reference:

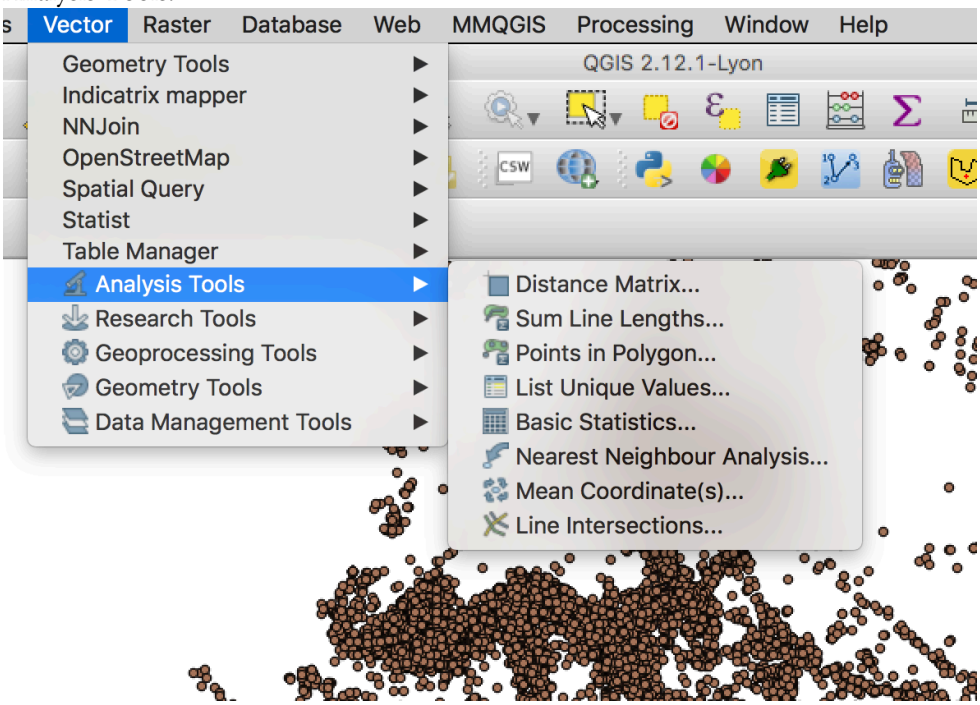
- Pick an appropriate color scheme for the data. You can use colorbrewer.org or create one yourself but it has to represent the underlying data appropriately.
- Pick an appropriate number of classes (p. 155)
- Create three maps with quantile, equal interval and natural breaks. For the fourth map, use the histogram functionality in QGIS and create your own custom class breaks that you think are appropriate.
- Combine all four maps into a single PDF document in Illustrator and make sure you have appropriate titles, legends and other necessary map elements in place (p. 108-109). Add a short descriptive paragraph that discusses the four different class breaks, how you constructed the custom class breaks and why you would use one particular mode of determining class breaks over another for this particular data.

Data Classification: Diverging Data

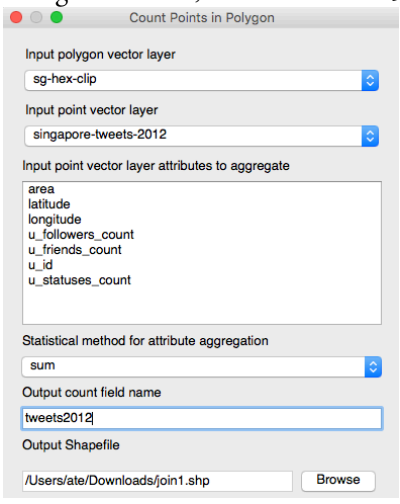
In this section, we will use two datasets that contain a random sample of tweets sent from Singapore in 2012 and 2015, respectively. The goal is to compare the density of tweeting in 2012 with that in 2015 and determine whether any (spatial) shifts have happened between the two time periods. To do so effectively, you will need to complete a number of steps, many of which you are already familiar with from previous labs.

1. Reading Data
   The data comes as .csv file and luckily it contains both the latitude and the longitude for each point. Use the 'Create layer from delimited text file' function to read the data into QGIS. After the data is successfully read in, save each layer as shapefile. Remember, a shapefile always has a spatial index (.shx) associated with it: this will help us speed up the next step where we need to join all these points to polygons in order to visualize and compare the densities. Repeat the procedure for both 2012 and 2015 and give the resulting Shapefiles appropriate filenames.
2. Spatial Join

To compare the densities in the two time periods, it is easiest to aggregate the individual points up to larger polygons. In this case, we can use the hexagon file we used in previous assignments and use a similar procedure as we did with the dengue maps. We have previously spatially joined polygons to polygons using the MMQGIS tools. However, for points to polygons there exists a more efficient/fast tool called 'Points in Polygons' under the Vector | Analysis Tools.



Use that tool to join the 2012 data to the hexagon layer and subsequently join the 2015 data to the result of the first join. The final resulting file is a hexagon layer with two fields added: one for the number of tweets in each hexagon in 2012, and one for 2015.



3. Odds Ratio
   To look at the change over time, we could simply divide the 2012 variable by the 2015 variable. However, since the total number of tweets might have increased as well (or the sampling size could have changed), it is better in these cases to calculate an 'odds ratio'. This way, we look at the ratio in each individual hexagon while also controlling for the total number of data points in each dataset. For reference, please read the paper here: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2513938 (you can focus on the section that specifically deals with the odds ratio).
   In this case, the odds ratio can simply be calculated as such: (tweets_2012/sum(tweets_2012))/ (tweets_2015/sum(tweets_2015)). First, use Vector | Analysis Tools | Basic Statistics to calculate the sum for both variables. You can then use the field calculator to calculate the odds ratio in each hexagon using the total sum you calculated using the Basic Statistics tool.
4. Color Scheme and Class Breaks
   As you can read in the paper referenced above, the odds ratio has a specific midpoint at an OR of 1. In other words, the odds ratio is not a sequential variable (e.g. a count) but it *diverges* at 1. Anything higher means there's a more of a certain variable than expected, and vice versa. Think back on our discussion on ratio versus interval variables and how that relates to the odds ratio. Naturally, this has consequences for both the color scheme and the class breaks that are most appropriate for the data. You will need to pick or create a *diverging* color scheme and manually create appropriate class breaks to create a thematic map of Twitter use in Singapore over time.
5. Illustrator Post-Production
   If all the previous steps are successful, you will see that there is a specific geographic shift in areas that sent a lot of tweets in 2015, as compared to 2012. Can you figure out what these areas are and what they have in common? Try to think what might explain this shift in behavior on Twitter. Export your map to Illustrator and make sure it is properly finished with an appropriate legend, title and some highlighting of significant areas of change. For the latter you can use leader lines or the approach Krygier and Wood use on page 162 of their book. Include a short description within your map that explains why these areas might have seen this change.

Assignment

On the class website, you will find the assignment for this lab. It consists of the two map products you created above. The assignment needs to be submitted as one or two PDF files. Please make sure you submit the assignment by **March 1.**