Galen Kanazawa

Geog 777 Project #1

3/7/22

**Scenario**

For Project 1, we are asked to imagine that we work for the Wisconsin DNR, and they have asked us to conduct a spatial analysis exploring the levels of nitrates in a sample collection of drinking water wells and rates of cancer occurrences. The well sampling locations are randomly distributed across the state, and the cancer occurrences are in a separate dataset, collected statewide over a 10-year period and aggregated at the census tract level. Our job is to explore to what extent high levels of nitrates in drinking water are correlated with cancer rates.

**Implementation**

Estimating nitrate levels for the entire state requires spatial interpolation from the well sample points. While there are several methods we could use, inverse distance weighting (IDW) is perhaps the most appropriate for this scenario. IDW assigns weight to a location as a function of inverse distance, so locations farther away from a known sample point receive less weight. IDW also involves the use of a distance decay coefficient *k*, which determines how quickly the weight diminishes as distance increases. As there is no real way to determine what the “correct” value for k should be, we are tasked with enabling the user to experiment with different k-values to see how the output maps and regression reports are affected.

I chose to run the analysis workflow in ArcPy and ArcGIS Pro, with the interface developed using TkInter. Pillow is another Python library which allows for manipulating the images in which the output maps were stored. I divided the project into two main .py files: the first executes the analysis, mapping, and related report output tasks, and the second is the main executable to launch the interface.

**Data Analysis and Mapping**

The workflow for the analysis and mapping script is as follows:

1. Once the project workspace and environment is prepared, the first step is to run the IDW tool. This tool interpolates groundwater nitrate levels from the well sample locations across the whole state. The tool includes a Power parameter, which we are populating with the k-value input by the user.
2. Since the IDW output is a raster surface, further processing is needed before the nitrate levels can be joined with the cancer tracts shapefile. This is accomplished via the Zonal Statistics as Table tool. However, with the default settings for cell size, I found that quite a few of the smaller census tracts were missed by the tool, resulting in null values in the output table. This was fixed by adjusting the cell size to be smaller, minimizing the number of census tracts missed.
3. Next, the table generated by the Zonal Statistics tool needs to be joined to the cancer tracts layer in order to get the average nitrates in the same layer as the cancer rates per tract. This can be accomplished a couple different ways, but there is an extra step caused by a schema lock that needs to be worked around before the script can proceed with the analysis.
4. With the table joined, the next step is to run a linear regression analysis to determine what the relationship between groundwater nitrates and cancer actually is. I used ArcPy’s Ordinary Least Squares tool on the cancer tracts layer, opting to output a report file that would be able to be viewed and saved by the user.
5. Analyzing the OLS tool’s results requires checking for spatial autocorrelation in the residuals. ArcPy’s tool for this runs the results through the Moran’s I algorithm and generates an HTML report file which can be viewed by the user.
6. Finally, the ArcGIS Pro project is accessed, which contains maps and layouts for the IDW nitrate interpolation and OLS linear regression. The maps are updated according to the K-value entered by the user before being exported to PNG image files. (Or, that’s what is supposed to happen. Sadly, while I was able to get the K-value text label to update, I wasn’t able to get the map layers to update via the layer.updateConnectionProperties method, or any other for that matter – as of now the maps are based on the map and layout contents in my Pro project.)

**Graphical User Interface**

I designed the GUI using TkInter. The layout is minimal, containing a frame with some introductory information and a text entry box where the user can enter the K-value, a button for running the analysis, and an informational dropdown menu where more contextual info on the analysis functions can be gathered. As the analysis script takes some time to complete, I added a progress bar to inform the user on how far along the script was as it runs through the functions. When the analysis is complete, the user may view either of the exported maps in the window by toggling buttons, and there are buttons for exporting the maps to PDF files, viewing the OLS regression report, and viewing the Moran’s I report.

**Results**

The OLS report contains the results of the analysis from a statistical standpoint. At its core, Ordinary Least Squares regression is a prediction model which compares the fit of the observed values to what the model’s predicted values were for how well the explanatory variable(s) explain the dependent variable. The OLS report indicates the probability value is statistically significant; in other words, there is a positive relationship between well water nitrates and cancer rates. However, the R-squared value, which runs on a scale from 0 (no relationship to the cancer rates) to 1 (perfect explanation for cancer rates), hovers around 0.02 depending on the K-value input. In other words, high groundwater nitrate levels account for about 2% of the observed cancer occurrences in Wisconsin. Clearly, other explanatory variables would need to be included in future studies to build a more robust correlation.