READ ME

Summary

The following scripts aim to analyze the partisan leaning of U.S congressional districts and examine the dissimilarity index between districts. The latter index measures the evenness with which two groups (in this case Democrats and Republicans) are distributed across component geographic areas that make up a larger area. Thus, the dissimilarity calculations can be utilized to potentially quantify the level of gerrymandering-related "packing" in a particular district. This project will use zip code level campaign contributions made to Democratic and Republican candidates during the 2020 Presidential election to quantify a district's partisan leaning, and will apply this analysis on districts in Colorado. A similar analysis can be run for any other state.

Input data

The following input files will first need to be downloaded:

1. The state of interest's zip code boundary GIS shapefile. This project specifically uses a Colorado zip code boundary GIS shapefile: https://data-cdphe.opendata.arcgis.com/datasets/6b6091f299204e4c9c406a624baf43e6_10?geometry=-113.968%2C37.512%2C-97.280%2C40.499
2. 116th U.S Congressional District GIS shapefile from U.S Census website: https://www.census.gov/cgi-bin/geo/shapefiles/index.php. A query call on the shapefile will need to be performed to select the state of interest.
3. 2019 U.S States GIS shape file from the U.S Census website: https://www.census.gov/cgi-bin/geo/shapefiles/index.php. A query call on the shapefile will need to be performed to select the state of interest.
4. Two files, *contrib_clean.pkl* and *com_cand_info.csv*, which were both created as part of a previous class assignment. The first file is a pickled version of aggregated individual campaign contribution data across party for each U.S zip code. The second file links campaign contribution amounts to committees, candidates, and political parties. The scripts needed to generate these files as well as the corresponding README document outlining how to create them, are listed under the sub-folder entitled "G-19".

Outputs

The script ***co_by_party.py*** creates the csv file ***CO_by_party.csv,*** which contains the campaign contributions per political party across Colorado zip codes. The script ***co_join.py*** creates a geopackage file entitled ***co.gpkg*** with state and zip, and zip_by_party layers. The ***co_zip_boundary_calc.py*** script updates the Colorado zip code boundary shapefile by adding total area calculations for each zip code, producing a new ***co_zip_boundaries.gpkg*** file. The ***zip_calcs.py*** script creates a csv file entitled ***zip_pieces_contrib.csv*** that lists campaign contributions per party for the zip pieces of Colorado zip codes, aggregates contributions on the congressional district level, and creates a new layer in the ***co.gpkg*** file. Finally, the ***co_dissim.py*** script calculates the dissimilarity index for each Colorado congressional district, exports this data to a new ***dissim.gpkg*** geopackage file, and creates a bar graph ***co_dissim.png*****.**

Detailed Script Overview (by run order)

co_by_party.py

- This script involves a join with respect to two input files. These are namely the *contrib_clean.pkl* file, which lists 2020 U.S Presidential election political committees for candidates and the candidate's respective party affiliation, as well as the *com_cand_info.csv* file, which lists the amounts of campaign contributions, by state and zip, made in the same election cycle in to specific political committees.
- These two datasets are first joined by their shared committee ID, and contributions from a state of interest by zip code (in this case Colorado) are selected. Then across all committees, the total amount of contributions to each political party are calculated. Primarily, this script is concerned with contributions made to the two largest parties, the Democrats and Republicans. Finally, the share of contributions in each zip code made to Democrats are calculated.
- These steps generate a csv file listing, by zip code in a state, the contributions made to each party and the share of contributions made to Democrats. This csv file will be subsequently used to map contributions by a state's zip codes in QGIS.

co_join.py

- The csv file generated in the previous step, *CO_by_party.csv*, which lists, by zip code in a state, the contributions made to each party and the share of contributions made to Democrats. In order to map this information in QGIS, these contributions by zip code need to be merged with a shapefile of Colorado zip codes.
- The script first reads in shapefiles for Colorado zip codes and U.S state boundaries, and writes the two variables as "zip" and "state" layers respectively in a *co.gpkg* file. Subsequently the "zip" layer and the contribution information in *CO_by_party.csv* are merged on their shared zip codes.
- The resulting dataset is written to a new "zip_by_party" layer in the *co.gpkg* file.

co_zip_boundary_calc.py

- As downloaded, the Colorado zip codes shapefile does not provide any information regarding the total land area for each zip code. However, this information could be useful for data verification purposes.
- In order to calculate the areas, this script reads in the Colorado zip code shapefile, and sets its projection to the recommended UTM Zone 13N, which is also known as EPSG:32613. Next, a "total area" column is created in the Colorado zip code file using the ".area" function on the file.
- This new file with zip code areas is written to a new *co_zip_boundaries.gpkg* geopackage file.

QGIS Intersect Function

- The data generated so far lists political contributions on a zip code level. However, this information needs to be aggregated on a congressional district level. The boundaries of zip codes and congressional districts oftentimes do not match. For example, a zip code may be within multiple congressional districts.
- To solve this issue, the zip codes in a state that are located within multiple congressional districts will need to be broken into multiple zip pieces. This task is performed by using QGIS's "Intersect" function on the "zip" layer of the *co.gpkg* file as the input layer and the

U.S Congressional Districts shapefile as the intersect layer (use UTM Zone 13N as the projection type). A query call will need to be used for the U.S Congressional District shapefile for the state of interest. This operation creates multiple zip pieces for zip codes that cross congressional district boundaries. Next, the areas of each zip piece are calculated using the Vector > Geometry Tools > Add Geometry Attributes function in QGIS. This resulting layer is then saved in a new "zip_pieces_area" layer in the *co.gpkg* file.

- Subsequently, the zip pieces are used to allocate the contributions in proportion to each piece's area in relation to the entire zip code. This step is performed in the next script.

zip_calcs.py

- The following script allocates campaign contributions to Democrats and Republicans in each zip piece in proportion to their area relative to the entire zip code. It first calculates the total area of each zip code by grouping the "zip_pieces_area" layer of the *co.gpkg* file by zip code and summing each zip code's area. It additionally aggregates the contributions of zip pieces by party for the congressional district each piece is located in
- As a side note, the total areas for each zip code can be matched with the total area zip code calculations in the *zip_code_boundaries.gpkg* file to verify the accuracy of the zip code total area calculations.
- The allocated zip piece contribution results are written out to a *zip_pieces_contrib.csv* file, which will be used for dissimilarity calculations in the next script.
- The contribution information aggregated by congressional district is then joined on the U.S Congressional District shapefile (with a query call to select the state of interest), and is saved in a new "distr_contrib" layer in the *co.gpkg* file.

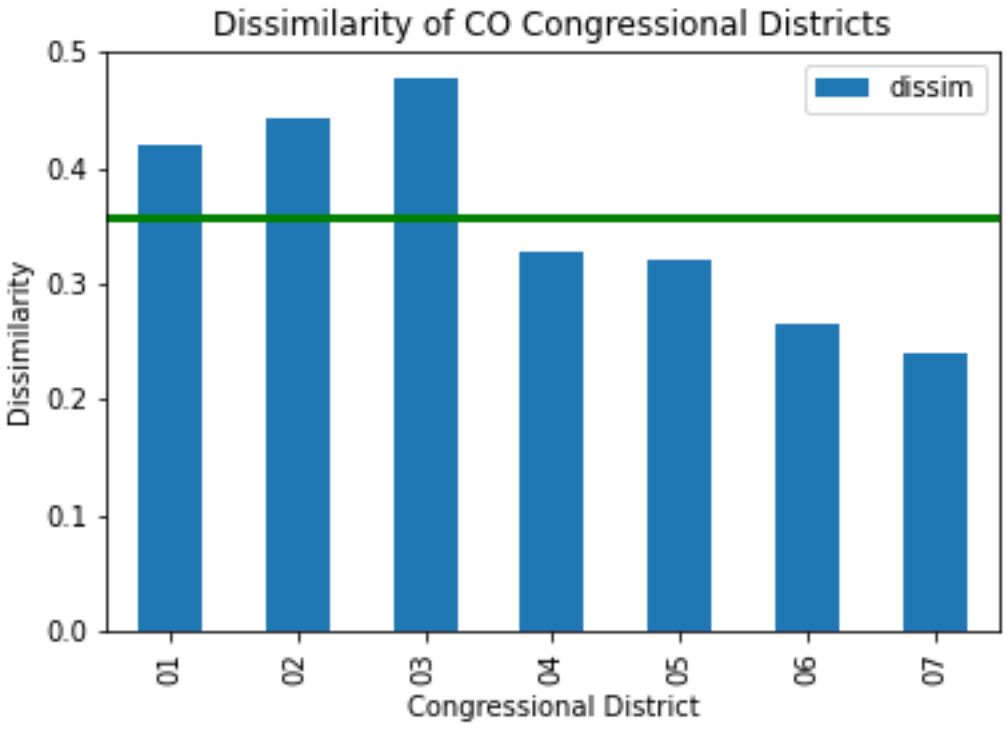
co_dissim.py

- The purpose of this script is to calculate each congressional district's dissimilarity index values. The index is calculated using the share of contributions made to Democrats and Republicans respectively in each zip piece. The zip piece contributions are listed in the *zip_pieces_contrib.csv* file previously created.
- The dissimilarity index calculations for each congressional district is then merged onto the "distr_contrib" layer in the *co.gpkg* file, and is saved as a new *dissim.gpkg* file.

- Finally, a bar graph is created to visually display the dissimilarity index value for each congressional district in the state, as well as the state's average dissimilarity level. This graph is saved as *co_dissim.png*

Results

- The patterns of Colorado contributions by zip code can be visualized by selecting the "zip_by_party" layer in the *co.gpkg* file. Then, set its style to "Graduated" using "d_share" as the value, set the color ramp to RdBu, and set the mode to "Equal Interval" using 5 classes.
- The patterns of Colorado contributions by congressional district can be visualized by selecting the "distr_contrib" layer in the *co.gpkg* file. Then, set its style to "Graduated" using "d_share" as the value, set the color ramp to RdBu, and set the mode to "Equal Count" using 5 classes.
- The patterns of Colorado dissimilarity index values by congressional district can be visualized by selecting the *dissim.gpkg* file. Then, set its style to "Graduated" using "dissim" as the value, set the color ramp to OrRd, and set the mode to "Equal Count" using 5 classes.



The horizontal line in the bar graph of dissimilarity index values for Colorado's congressional districts illustrates that the average dissimilarity level for the entire state is about 0.36. Of the seven

congressional districts in the state, three districts (the 1st, 2nd, and 3rd districts) all boast dissimilarity level above the state average. The 3rd congressional district, for example, holds an index value of about 0.48. This indicates that about 48% of the population in the 3rd district would have to move for true equality. The 2nd congressional district holds the second highest dissimilarity level of about 0.44, meaning that about 44% of the population in that district would have to move to achieve true equality.

A high level of dissimilarity in a district may potentially indicate that a district was drawn in order to "pack" certain voters. Specifically, this may mean concentrating one party's voting power in one district to reduce their voting power in other districts. However, this method may not be foolproof. Increasingly, Americans are geographically segregating themselves based on political preferences, with more liberal voters concentrating in cities and other urban areas, while conservatives may tend to live in more rural regions. Therefore, a district may also hold a high degree of dissimilarity due to the "natural" geographic distributions of individuals rather than due to the "packing" of voters.