**Integration, Analysis, and clustering of FERC Region Data**

**Andrew Igdal**

**12/9/24**

**1. Introduction**

FERC Order 1000 encourages regional coordination in transmission planning and cost allocation. Understanding the structure of these regions—both their geographic extents and the attributes of their transmission infrastructure—is essential for informed policy and investment decisions. Additionally, linking these regions to demographic and socio-economic factors enables researchers and planners to consider how infrastructure meets community needs.

The initial proposal aimed to:

1. Approximate FERC 1000 region shapefiles from balancing authority areas.
2. Merge these approximated regions with transmission infrastructure data and demographic data.
3. Create a user-friendly analytical tool, potentially with an interactive interface (e.g., Streamlit) and consider using resources like TACC for computation.
4. Perform unsupervised learning (e.g., clustering) on states and regions to identify patterns.
5. Produce dendrograms and other visualizations to facilitate future research.

**Changes from the original plan:** While the scripts successfully generated processed data, performed clustering, and produced dendrograms and heatmaps, the step involving TACC resources proved unnecessary. Additionally, integrating an online interactive tool (e.g., via Streamlit) was not completed.

**2. Methodology**

**Data and Tools:**

* **Geospatial data:**
  + Control Areas (Control\_\_Areas.geojson) representing balancing authorities.
  + A county shapefile (US\_COUNTY\_SHPFILE/US\_county\_cont.shp) for spatial reference.
  + HIFLD transmission lines data.
* **Demographic data:** ACS 2019 data fetched via censusdata Python package.
* **Processing Libraries:** GeoPandas for spatial operations, Pandas and NumPy for data manipulation, and Matplotlib/Seaborn for visualization.

**Script Summaries:**

**SCRIPT 1:**

* Approximates FERC Order 1000 regions by aggregating control areas.
* Cleans and merges balancing authority data (Control\_\_Areas.geojson) with a CSV (BA\_FERC1000.csv) containing FERC 1000 region mappings.
* Filters states of interest, dissolves and merges shapefiles, and manually resolves overlaps between regions.
* Produces a final FERC\_1000\_Regions.geojson representing each aggregated region.
* Creates a color-coded map of regions overlaid on state borders.

**SCRIPT 2:**

* Loads the finalized FERC\_1000\_Regions.geojson and extracts individual regions (e.g., CAISO, ERCOT, ISO-NE, SE, etc.).
* Saves each region as its own GeoJSON (e.g., caisogeometry.geojson) for future spatial joins.
* Processes HIFLD transmission lines data, cleaning and ensuring consistent CRS.
* Performs spatial joins with each FERC region to isolate transmission lines per region.
* Plots transmission lines and region boundaries to verify alignment.
* Exports cleaned per-region transmission line data to GeoJSON files for further analysis.

**SCRIPT 3:**

* Takes the previously created per-region transmission line files and reprojects them.
* Cleans up column names, adds a year column based on SOURCEDATE, and estimates power capacity for AC lines.
* Removes invalid or negative voltage lines, inspects data distributions, and provides various summary statistics and visualizations (histograms, KDEs, boxplots).
* Finally, merges lines within each region if appropriate, recalculates line lengths and power capacity, and exports merged results.

**SCRIPT 4:**

* Integrates ACS demographic data at the county level and links it to each region through spatial joins.
* Aggregates demographic and transmission attributes at the region level, computing population-weighted averages of median age and income, and calculating demographic percentages.
* Collects unique transmission line types per region.
* Creates a summary DataFrame with demographic and infrastructural attributes for each FERC region.
* Standardizes numeric features and performs hierarchical clustering (Ward’s method) to produce a dendrogram of regions.
* Generates a correlation matrix heatmap and a clustered heatmap to visualize relationships among variables and similarities between regions.

**3. Results & Conclusions**

**Outputs:**

* **Geospatial Products:**
  + FERC\_1000\_Regions.geojson: Aggregated polygon geometries approximating the FERC Order 1000 planning regions.
  + Individual region geometry GeoJSONs
  + Region-specific transmission line GeoJSON files
* **Analysis Products:**
  + **Numeric Columns (e.g., VOLTAGE, YEAR, LOG\_POWER\_CAPACITY, LINE\_LENGTH\_MILES):**
    - Histograms with KDE overlays for distribution understanding.
    - Boxplots for median, quartiles, and outlier detection.
    - KDE plots for smooth density estimation.
    - CDF plots to understand the cumulative distribution.
  + **Categorical Columns (e.g., STATUS, TYPE):**
    - Count plots (bar charts) to visualize frequency distribution.
    - Pie charts to visualize proportional breakdown of categories.
  + **Geospatial Visualizations:**
    - Plots transmission lines for each region on a map.
    - Overlays region boundaries (in red) to show the spatial extent of the regions in relation to the transmission infrastructure.
  + A consolidated DataFrame (summary\_df in SCRIPT 4) associating each region with demographic attributes and transmission infrastructure metrics.
  + Hierarchical clustering output: Dendrogram and identification of clusters among FERC regions.
  + Dendrogram showing how regions group based on standardized features.
  + Correlation heatmap highlighting relationships between variables like median income, total power capacity, and demographic percentages.
  + Clustered heatmap depicting region-to-region similarity patterns.

**4. Future Work**

* **Interactive Tool Development:**Build a Streamlit-based application to allow users to explore and visualize processed data interactively, eventually integrating the tool into a personal website.
* **Exporting and Using Probability Distribution Functions (PDFs):**Export PDFs for variables like power capacity and line lengths for use in modeling.
* **Expanding Data Integration:**Incorporate additional variables such as renewable energy capacity, weather patterns, etc.
* **Advanced Clustering:**Conduct more granular clustering analyses by state within FERC regions
* **Policy and Modeling Applications:**Use the processed datasets and insights for research on transmission expansion, grid reliability, and renewable energy integration to inform energy policy decisions.