README Document: Carbon Dioxide Pipeline Analysis

Contact: Julia Davis; jd3td@virignia.edu

Last Updated: March 26, 2025

Overview

This project analyzes demographic and environmental factors related to CO_2 pipeline networks, including existing, proposed, and projected pipelines for 2030. The analysis utilizes statistical modeling and visualization techniques to assess disparities and environmental justice implications.

Dataset Information

The study uses various datasets, including:

- **Demographic Data (2017-2021)**: Socioeconomic and racial/ethnic data at the census tract level. We used data from IPUMS-NHGIS's 2017-2021 five-year estimates of the American Community Survey.
- **Existing Pipeline Data (2023)**: Spatial information on existing CO_2 pipelines, as of November 2023. We used data from the US Pipeline and Hazardous Materials Safety Administration.
- **Proposed Pipeline Data (2023)**: Spatial information on proposed CO_2 pipelines, as of November 2023. We used data from FracTracker Alliance.
- **Projected Pipeline Data (2030)**: Future projection of CO_2 pipeline infrastructure. We used the 2030 projection from the 2021 Net Zero American Final Report, from Princeton University.

Dependencies

This analysis requires the following R packages:

```
library(readr)
library(dplyr)
library(glmnet)
library(caret)
library(ggplot2)
library(MASS)
library(factoextra)
```

```
library(openxlsx)
library(coefplot)
library(scales)
library(stringr)
library(tibble)
library(patchwork)
library(extrafont)
library(marginaleffects)
library(knitr)
library(margins)
library(boot)
```

Data Preprocessing

1. **Load and Clean Data**:
    - Remove observations with zero population or zero land area.
    - Merge demographic and pipeline datasets based on GEO_ID.
    - Create new variables as descriptors of demographic groups. These variables relate to population, income, race, housing, and education level characteristics of a census tract observation.
2. **Filter Data**:
    - For each pipeline network, keep only census tracts in counties that contain a CO_2 pipeline.

Statistical Analysis

1. **Machine Learning for Feature Selection**
    a. Applied LASSO regression to identify significant predictors of pipeline presence.
    b. Used cross-validation to select the optimal lambda value.
    c. Assessed variable importance and visualized results.
2. **Logistic Regression Models**
    a. Model 1: Uses our selected demographic variables of interest from our LASSO regression model with the existing CO_2 pipeline network.
    b. Model 2: Uses our selected demographic variables of interest from our LASSO regression model with the proposed CO_2 pipeline network.
    c. Model 3: Uses our selected demographic variables of interest from our LASSO regression model with the 2030 projected CO_2 pipeline network.
3. **Marginal Effects Calculation**

a. Computed to interpret the influence of key demographic factors on pipeline placement likelihood for each pipeline network.

Data Visualization

- Bar Charts: Highlight variable importance in LASSO regression.
- Coefficient Plots: Display model coefficients with confidence intervals.

How to Run the Code

1. Ensure all required datasets are available and properly linked.
2. Install necessary R packages if not already installed.
3. Run the scripts sequentially, for each CO_2 pipeline network:
    - **Data Cleaning & Merging**
    - **Descriptive Statistics**
    - **Machine Learning Feature Selection (LASSO)**
    - **Logistic Regression Analysis**
    - **Visualization & Interpretation**

Contact

For questions, please reach out to Julia Davis, jd3td@virginia.edu.