Environmental Citizen Complaints Main README

Overview

This repository contains administrative data on environmental pollution complaints reported to the Texas Commission on Environmental Quality (TCEQ) from 2003-2019, obtained via public information request (PIR 21-65675). The data provides comprehensive coverage of citizen-reported environmental concerns across Texas, the second-largest U.S. state by population and land area.

This README provides comprehensive information on all of the data and code required to replicate the figures and tables in the main manuscript and all appendices. Text mining replication steps and information are included separately in the text_mining folder.

Description of datasets

Data Sources

Primary Data

- **Source**: Texas Commission on Environmental Quality (TCEQ)
  - **Description: Pollution Concerns Data**: Administrative records of all pollution concerns reported to the TCEQ
  - **Collection Method**: Public Information Request (PIR 21-65675) from the TCEQ website: https://www.tceq.texas.gov/agency/data/records-services/reqinfo.html. To obtain public information, one must submit an online or written request and pay a fee, which may range from small to large.

Supplementary Data

- **Source**: Texas Commission on Environmental Quality (TCEQ)
  - **Description: Administrative Data**: Administrative records on labor inputs to investigations
  - **Collection Method**: Public Information Request (PIR 23-83999) from the TCEQ website: https://www.tceq.texas.gov/agency/data/records-services/reqinfo.html. To obtain public information, one must submit an online or written request and pay a fee, which may range from small to large.
- **Source**: Meng et al. (2019)
  - **Description: Air Quality**: Average annual PM2.5 estimates at zip-code level (Meng et al., 2019; following Colmer et al., 2020 methodology)
  - **Collection Method**: Publicly available datasets at https://sites.wustl.edu/acag/datasets/historical-pm2-5-across-north-america/#:~:text=This%20dataset%20is%20the%20estimated,suspended%20particles%20(TSP)%20measurements.
- **Source**: U.S. Census Bureau
  - **Description: Demographic Data**: 2000 Decennial Census at zip-code level
  - **Collection Method**: Publicly available via Social Explorer (https://www.socialexplorer.com/data/C2000#)
- **Source**: U.S. Census Bureau
  - **Description: Demographic Data**: Cartographic Boundary Files
  - **Collection Method**: Publicly available at the Census website (https://www.census.gov/geographies/mapping-files/time-series/geo/cartographic-boundary.2019.html)
- **Source**: U.S. Census Bureau
  - **Description: Demographic Data**: Socioeconomic and demographic variables at the zip code level from 5-year 2012 American Community Survey (ACS)
  - **Collection Method**: Publicly available via Social Explorer (https://www.socialexplorer.com/tables/ACS2012_5yr)
- **Source**: EPA
  - **Description: State Data**: Information on state-delegated compliance and enforcement
  - **Collection Method**: Publicly available at https://www.epa.gov/state-review-framework

Author-Generated Survey Data

- **Source**: Authors
  - **Description**: Qualtrics survey regarding environmental complaints
  - **Collection Method**: Qualtrics

Dataset list

Name	Data Files	Source	Citation	Notes

Name	Data Files	Source	Citation	Notes
Public Information Requests (PIR 21-65675 and PIR 23-83999)	PIR56575*.csv, TCEQ Workplan Definitions_July2022.csv, PIR 21-56575*.csv, All Investigations*.dta	Texas Commission on Environmental Quality (TCEQ)	Texas Commission on Environmental Quality. "Administrative Records of Pollution Complaints, 2003–2019." Obtained via Public Information Request PIR 21-65675. https://www.tceq.texas.gov/agency/data/records-services/reqinfo.html., Texas Commission on Environmental Quality. Obtained via Public Information Request PIR 23-83999. https://www.tceq.texas.gov/agency/data/records-services/reqinfo.html.	
PM 2.5 Data by Zip Code	V4NA03_PM25_NA_201101_201112-RH35*.csv	Meng et al. 2019.	Meng, X., Glen, G., Chang, K. H., Van Donkelaar, A., Martin, R. V., and Dominici, F. 2019. "Estimating Historical PM2.5 Concentrations in the United States Using Satellite Data, Chemical Transport Models, and Ground-Based Monitors: A Hybrid Modeling Approach." *Environmental Science & Technology* 53(9) 5071–5079. https://doi.org/10.1021/acs.est.8b06875.	
Zip Code Decennial Census Demographics	R12907124.dct, R13696125.dct, R12907124_SL860.txt, R13696125_SL860.txt, nhgis0002_ts_geog2010_zcta.csv	U.S. Census	U.S. Census Bureau. "2000 Decennial Census." Prepared by Social Explorer. https://www.socialexplorer.com/data/C2000#.	Codebook included in deposit under: `nhgis0002_ts_geog2010_zcta_codebook.txt`
Demographic and Socioeconomic Information by Zip Code	R13696237.dct, R13696237_SL860.txt	U.S. Census	U.S. Census Bureau. "American Community Survey 5-Year Data 2012." Prepared by Social Explorer. https://www.socialexplorer.com/tables/ACS2012_5yr.	
Cartographic Boundary Files	cb_2019_us_state_500k.shp, cb_2019_us_zcta_510_500k.shp	U.S. Census	U.S. Census Bureau. 2019. "Cartographic Boundary Files." https://www.census.gov/geographies/mapping-files/time-series/geo/cartographic-boundary.2019.html.	
EPA SRF	Full_Download_of_SRF_Findings_Nov_3_2023.csv	U.S. EPA	U.S. Environmental Protection Agency. "State Review Framework." https://www.epa.gov/state-review-framework.	
Author-Generated Survey	Citizen_Complaints_Survey_June_7_2022_08.40.csv	Self-generated	Colmer, J., Evans, M. F., and Shimshack, J. 2024. "Environmental Citizen Complaints." Forthcoming, *Journal of Political Economy: Microeconomics*. Survey conducted May 26–June 7, 2022.	

Description of programs

The code folder in the Replication package contains all code necessary for the replication. The following scripts are used in main.do:

- 01_setup.do: sets global directories and installs Stata packages
- 02_data_preparation.do: processes raw data to create intermediate data sets and final analysis data files that are used in figure and table construction.
- 03_main_figure_and_table_construction.do: creates Figures 1-7 and Table 1 from the main section of the manuscript
- 04_appendix: creates remaining appendix figures and tables

The code folder also contains 3 subfolders:

- data_prep_code: contains all data preparation code. These scripts are called in 02_data_preparation.do
- main_figure_and_table_code: contains all figure and table code for the main manuscript. These scripts are called in 03_main_figure_and_table_construction. Note that, within each script, figure panels may not be generated sequentially; this is due to the fact that one figure as a whole may require two separate datasets, one for part of the panels and another for the other part.
- appendix_figure_and_table_code: contains all scripts that are needed for appendix figure and table construction. These scripts are called in 04_appendix.do.

Computational Requirements

This code was run on a 16 GB RAM Windows laptop. It took ~7 hours to run from start to finish. The following are software and package requirements

For Stata:

- Stata 19.5
- packages:
  - reghdfe 6.12.3: Released August 8, 2023
  - ftools 2.49.1: Released August 8, 2023
  - unique 1.2.4: Released June 16, 2023
  - coefplot 1.8.8: Releaesd August 22, 2025
  - estout 3.31: Released April 26, 2022
  - distinct 1.2.1: Released March 1, 2012
  - regsave 1.4.9: Released March 4, 2021

For R:

- R 4.1.3, to be compatible with rgdal package
- packages:
  - sf 1.0.12: Released June 29, 2022
  - dplyr 1.1.4: Released January 24, 2024
  - ggplot2 3.5.2: Released July 8, 2024
  - showtext 0.9.7: Released September 3, 2022
  - tidyverse 2.0.0: Released October 24, 2023
  - readstata13 0.11.0: Released October 24, 2023
  - sp 1.6.0: Released October 31, 2022
  - rgdal 1.6.6: Released October 31, 2022
  - raster 3.6.20: Released July 19, 2023
  - rgeos 0.6.2: Released June 24, 2022
  - maptools 1.1.6: Released June 17, 2023
  - exactextractr 0.9.1: Released June 23, 2022
  - haven 2.5.2: Released Aigust 23, 2022

For Python:

- Python: 3.13.5
- packages:
  - pandas 2.3.: Released May 25, 2024

Instructions for Replicators

- To run, first unzip change the root directory in main.do
- All code other than code for text mining is run via Stata, which incorporates R and Python code using `Rshell` and `Python` commands. To ensure compatibility with your device, open the code folder and select 02_data_preparation.do and 04_appendix.do. In these files, after Rshell is called, adjust the Rscript path so that Rscript.exe is specified for the correct folder on your device.
- Then, run main.do to create all of the main figures and tables and most of the appendix figures and tables.
- To create the wordclouds (appendices C and E), first adjust paths in paths.py and run this script. The script is located in the text_mining/Environmental_Complaints/Full_Sample/Python_Code. Then, in the same folder, run main_text_mining_LDA.py

> [NOTE]: In order to run, replicator must have the Replication folder (with data populated inside) available and downloaded on device.

> [NOTE]: Though the package contains Stata, R, and Python, all code other than text mining code is run via Stata. Stata calls R and Python and sets paths dynamically. Text mining code, due to its more expansive nature, is run separately.

> [NOTE]: If the replicator wants to use processed data to generate all of the figures and the tables, some of the processed_data files were zipped in the file upload process. Please unzip any files in the processed_data folder before running.

Mappings of Figures/Tables to Programs

Name of Figure/Table	Name(s) and Location(s) for Figure/Table	Do File(s) that Create Figure/Table
Figure 1	output/figures/Figure_1_Panel_A.pdf, output/figures/Figure_1_Panel_B.pdf, output/figures/Figure_1_Panel_C.pdf, output/figures/Figure_1_Panel_D.pdf	code/main_figure_and_table_code/Figure_1.do
Figure 2	output/figures/Figure_2_Panel_A.pdf, output/figures/Figure_2_Panel_B.pdf, output/figures/Figure_2_Panel_C.pdf	code/main_figure_and_table_code/Figure_2.do
Figure 3	output/figures/Figure_3_Panel_A.pdf, output/figures/Figure_3_Panel_B.pdf, output/figures/Figure_3_Panel_C.pdf, output/figures/Figure_3_Panel_D.pdf	code/main_figure_and_table_code/Figure_3.do
Figure 4	output/figures/Figure_4_Panel_A.pdf, output/figures/Figure_4_Panel_B.pdf, output/figures/Figure_4_Panel_C.pdf, output/figures/Figure_4_Panel_D.pdf	code/main_figure_and_table_code/Figure_4.do
Figure 5	output/figures/Figure_5_Panel_A.pdf, output/figures/Figure_5_Panel_B.pdf, output/figures/Figure_5_Panel_C.pdf, output/figures/Figure_5_Panel_D.pdf	code/main_figure_and_table_code/Figure_5.do

Name of Figure/Table	Name(s) and Location(s) for Figure/Table	Do File(s) that Create Figure/Table
Figure 6	output/figures/Figure_6_Panel_A.pdf, output/figures/Figure_6_Panel_B.pdf, output/figures/Figure_6_Panel_C.pdf, output/figures/Figure_6_Panel_D.pdf, output/figures/Figure_6_Panel_E.pdf, output/figures/Figure_6_Panel_F.pdf	code/main_figure_and_table_code/Figure_6.do
Figure 7	output/figures/Figure_7_Panel_A.pdf, output/figures/Figure_7_Panel_B.pdf, output/figures/Figure_7_Panel_C.pdf, output/figures/Figure_7_Panel_D.pdf, output/figures/Figure_7_Panel_E.pdf, output/figures/Figure_7_Panel_F.pdf	code/main_figure_and_table_code/Figure_7.do
Table 1	output/tables/Table_1_Panel_A.tex, output/tables/Table_1_Panel_B.tex, output/tables/Table_1_Panel_C.tex	code/main_figure_and_table_code/Table_1.do
Figure A1	output/figures/Figure_A1.pdf	code/appendix_figure_and_table_code/Figures_A1_C4.R
Figure A2	output/figures/Figure_A2.pdf	code/appendix_figure_and_table_code/Figure_A2.do
Figure A3	output/figures/Figure_A3.pdf	code/appendix_figure_and_table_code/Figure_A3.do
Figure A4	output/figures/Figure_A4_Panel_A.pdf, output/figures/Figure_A4_Panel_B.pdf, output/figures/Figure_A4_Panel_C.pdf	code/appendix_figure_and_table_code/Figure_A4.do
Figure A5	output/figures/Figure_A5_Panel_A.pdf, output/figures/Figure_A5_Panel_B.pdf	code/appendix_figure_and_table_code/Figure_A5.do
Figure A6	output/figures/Figure_A6_Panel_A.pdf, output/figures/Figure_A6_Panel_B.pdf	code/appendix_figure_and_table_code/Figure_A6.do
Figure A7	output/figures/Figure_A7.pdf	code/appendix_figure_and_table_code/Figure_A7.do
Figure A8	output/figures/Figure_A8_Panel_A.pdf, output/figures/Figure_A8_Panel_B.pdf, output/figures/Figure_A8_Panel_C.pdf, output/figures/Figure_A8_Panel_D.pdf	code/appendix_figure_and_table_code/Figure_A8.do
Figure A9	output/figures/Figure_A9_Panel_A.pdf, output/figures/Figure_A9_Panel_B.pdf	code/appendix_figure_and_table_code/Figure_A9.do
Figure A10	output/figures/Figure_A10_Panel_A.pdf, output/figures/Figure_A10_Panel_B.pdf	code/appendix_figure_and_table_code/Figure_A10.do
Table A1	output/tables/Table_A1.tex	code/appendix_figure_and_table_code/Table_A1.do
Figure B1	output/figures/Figure_B1_Panel_A.pdf, output/figures/Figure_B1_Panel_B.pdf, output/figures/Figure_B1_Panel_C.pdf, output/figures/Figure_B1_Panel_D.pdf, output/figures/Figure_B1_Panel_E.pdf, output/figures/Figure_B1_Panel_F.pdf	code/appendix_figure_and_table_code/ Figures_B1_B2_Table_B1.do
Figure B2	output/figures/Figure_B2_Panel_A.pdf, output/figures/Figure_B2_Panel_B.pdf, output/figures/Figure_B2_Panel_C.pdf, output/figures/Figure_B2_Panel_D.pdf, output/figures/Figure_B2_Panel_E.pdf, output/figures/Figure_B2_Panel_F.pdf	code/appendix_figure_and_table_code/ Figures_B1_B2_Table_B1.do
Table B1	output/figures/Table_B1_Panel_A.tex, output/figures/Table_B1_Panel_B.tex, output/figures/Table_B1_Panel_B_politics.tex	code/appendix_figure_and_table_code/ Figures_B1_B2_Table_B1.do
Figure C1	Manually generated	N/A
Figure C2	Manually generated	N/A
Figure C3	output/figures/Figure_C3.pdf	code/appendix_figure_and_table_code/Figure_C3.do
Figure C4	output/figures/Figure_C4_map.png	code/appendix_figure_and_table_code/Figures_A1_C4.R
Figure C5	output/figures/Figure_C5.pdf	code/appendix_figure_and_table_code/Figure_C5.do
Figure C6	output/figures/Figure_C6_Panel_A.pdf, output/figures/Figure_C6_Panel_B.pdf, output/figures/Figure_C6_Panel_C.pdf, output/figures/Figure_C6_Panel_D.pdf, output/figures/Figure_C6_Panel_E.pdf, output/figures/Figure_C6_Panel_F.pdf	code/appendix_figure_and_table_code/Figure_C6.do
Figure C7	output/figures/Figure_C7_Panel_A.pdf, output/figures/Figure_C7_Panel_B.pdf, output/figures/Figure_C7_Panel_C.pdf	code/appendix_figure_and_table_code/Figure_C7.do
Appendix C Wordclouds	text_mining/Environmental_Complaints/Air_Subset/wordclouds text_mining/Environmental_Complaints/Waste_Subset/wordclouds text_mining/Environmental_Complaints/Water_Subset/wordclouds	text_mining/Environmental_Complaints/Air_Subset/ Python_Code/Air_LDA_Lemmatize_Topics_10_seed_3.py text_mining/Environmental_Complaints/Waste_Subset/ Python_Code/Waste_LDA_Lemmatize_Topics_10_seed_3.py text_mining/Environmental_Complaints/Water_Subset/ Python_Code/Water_LDA_Lemmatize_Topics_10.py
Figure C8	output/figures/Figure_C8_Panel_A.pdf, output/figures/Figure_C8_Panel_B.pdf, output/figures/Figure_C8_Panel_C.pdf	code/appendix_figure_and_table_code/Figure_C8.do
Figure D1	output/figures/Figure_D1_Panel_A.pdf, output/figures/Figure_D1_Panel_B.pdf, output/figures/Figure_D1_Panel_C.pdf	code/appendix_figure_and_table_code/Figure_D1.do
Figure D2	output/figures/Figure_D2_Panel_A.pdf, output/figures/Figure_D2_Panel_B.pdf, output/figures/Figure_D2_Panel_C.pdf	code/appendix_figure_and_table_code/Figure_D2.do

Name of Figure/Table	Name(s) and Location(s) for Figure/Table	Do File(s) that Create Figure/Table
Figure D3	output/figures/Figure_D3_Panel_A.pdf, output/figures/Figure_D3_Panel_B.pdf, output/figures/Figure_D3_Panel_C.pdf	code/appendix_figure_and_table_code/Figure_D3.do
Figure E1	text_mining/Environmental_Complaints/Air_Subset/wordclouds/10_Topics_seed_3	text_mining/Environmental_Complaints/Air_Subset/Python_Code/Air_LDA_Lemmatize_Topics_10_seed_3.py
Figure E2	text_mining/Environmental_Complaints/Waste_Subset/wordclouds/10_Topics_seed_3	text_mining/Environmental_Complaints/Waste_Subset/Python_Code/Waste_LDA_Lemmatize_Topics_10_seed_3.py
Figure E3	text_mining/Environmental_Complaints/Water_Subset/wordclouds/10_Topics	text_mining/Environmental_Complaints/Water_Subset/Python_Code/Water_LDA_Lemmatize_Topics_10.py
Table E1	Manually generated, but uses output text from many files	text_mining/Environmental_Complaints/Air_Subset/Python_Code/Air_LDA_Preprocess_Lemmatize.py text_mining/Environmental_Complaints/Air_Subset/Python_Code/Air_LDA_Lemmatize_Topics_10_seed_3.py text_mining/Environmental_Complaints/Waste_Subset/Python_Code/Waste_LDA_Preprocess_Lemmatize.py text_mining/Environmental_Complaints/Waste_Subset/Python_Code/Waste_LDA_Lemmatize_Topics_10_seed_3.py text_mining/Environmental_Complaints/Water_Subset/Python_Code/Water_LDA_Preprocess_Lemmatize.py text_mining/Environmental_Complaints/Water_Subset/Python_Code/Water_LDA_Lemmatize_Topics_10.py
Figure F1	Manually generated	N/A
In-text values, pages 15, 16, and 26	In code via comments and print statements	code/main_figures_and_tables/In_Text_Numbers.do

> [NOTE] For instances where in-text numbers stem from figures, corresponding point estimate files have been populated in the output/point_estimates folder

References

Colmer, Jonathan, Evans, Mary F., and Shimshack, Jay. 2024. "Environmental Citizen Complaints." Forthcoming, Journal of Political Economy: Microeconomics. Survey conducted May 26–June 7, 2022.

Meng, X., Glen, G., Chang, K. H., Van Donkelaar, A., Martin, R. V., and Dominici, F. 2019. "Estimating Historical PM2.5 Concentrations in the United States Using Satellite Data, Chemical Transport Models, and Ground-Based Monitors: A Hybrid Modeling Approach." *Environmental Science & Technology* 53 (9): 5071–5079. https://doi.org/10.1021/acs.est.8b06875.

Texas Commission on Environmental Quality. "Administrative Records of Pollution Complaints, 2003–2019." Obtained via Public Information Request PIR 21-65675. https://www.tceq.texas.gov/agency/data/records-services/reqinfo.html.

Texas Commission on Environmental Quality. Obtained via Public Information Request PIR 23-83999. https://www.tceq.texas.gov/agency/data/records-services/reqinfo.html.

U.S. Census Bureau. "2000 Decennial Census." Prepared by Social Explorer. https://www.socialexplorer.com/data/C2000#.

U.S. Census Bureau. 2019. "Cartographic Boundary Files." https://www.census.gov/geographies/mapping-files/time-series/geo/cartographic-boundary.2019.html.

U.S. Census Bureau. "American Community Survey 5-Year Data 2012." Prepared by Social Explorer. https://www.socialexplorer.com/tables/ACS2012_5yr.

U.S. Environmental Protection Agency. "State Review Framework." https://www.epa.gov/state-review-framework.