# **Scientific Data Management Plan**

Below details the steps, code, and datasets required to re-run the analyses on the relationship between disaster resilience and chronic conditions in 10,149 ZCTAs and the simulation study.

Additional details on the methods and results of the analyses can be found in the associated manuscript, ‘A spectral confounder adjustment for spatial regression with multiple exposures and outcomes’ located in: https://arxiv.org/abs/2506.09325.

All files can be found at https://github.com/snprim/mv\_spectral\_confounder.

# **Step 1: Generate datasets**

*Purpose:* This step generates datasets with ZCTA level information on health outcomes (2017-2019), SVI exposures (2018), and adjusting covariates (population from 2015 and urbanicity from 2010). hese datasets are used for analysis in Step 2.

*Code:*

|  |  |
| --- | --- |
| **File** | **Information** |
| data\_preparation.R | This script:   * Reads in health outcomes, SVI, population, and urbanicity * Uses cross walk between ZCTA, state, and region to only keep data from southern states * Uses shapefile data to generate adjacency matrix and its eigen decomposition, which is used for projecting spatial data into spectral domain * Combine all the data into three matrix for data analysis:   + X\_star: exposures   + Y\_star: outcomes   + Z\_star: adjusting covariates * Other input needed for analyses:   + Qmat\_south, W\_south   + These are provided in outputs or can also be generated from the script. |
| ACS2015\_DataDownload.R | This script downloads census data, and the output is used in data\_preparation.R. The output (US\_ZCTA\_ACSvars\_2015) is also provided. |

*Inputs:*

|  |  |
| --- | --- |
| **File** | **Information** |
| data/inputs/nhgis0015\_ds172\_2010\_zcta.csv | This input file is needed for data\_generation.R to provide urbanicity information. Data information is available in nhgis0015\_ds172\_2010\_zcta\_codebook.txt. |
| data/inputs/zcta\_state\_region.csv | This input is used in data\_generation.R to link ZCTAs to state and region. Only the south region is used in the data analysis. |
| tl\_2023\_us\_zcta520 | The zip file can be downloaded at <https://www2.census.gov/geo/tiger/TIGER2023/ZCTA520/>. After the files are extracted, the corresponding folder contains shapefiles used to generate the adjacency matrix and allow us to make maps. |
| data/inputs/acs5\_variables1.csv  data/inputs/acs5\_variables2.csv | These files include lists of variables for ACS2015\_DataDownload.R to download from the R package tidycensus. |
| Medicare data for incident rates of chronic conditions (2017-2019)  "health\_data.csv" | Not provided, but this is needed for data analysis |

*Outputs:*

|  |  |
| --- | --- |
| **File** | **Information** |
| data/outputs/US\_ZCTA\_ACSvars\_2015.csv | Output from ACS2015\_DataDownload.R. This is used in data\_generation.R. Note that the only variable used from this file for the analysis is population (“TOTAL”). The ZCTA numbers (“ZCTA”) is used for merging with other files. |
| data/outputs/Qmat\_south.rds data/outputs/adj.mat\_south.rds | These are outputs from data\_preparation.R. The eigen decomposition of Qmat\_south derives Gamma\_south and W\_south.rds. They are needed for analyses. |
| data/outputs/zcta\_order.rds | This is an output from data\_preparation.R. This contains the ZCTAs included in the analyses in the set order. The adjacency-generated matrices depend on the order of the ZCTAs, so this file is to ensure the order is correct, if any of the data files are used. |

# **Step 2: Run statistical analyses**

*Purpose:* This step runs the analyses on the relationship between disaster resilience and health impact. The outputs from this analysis include south region and ZCTA-level information on the incident rates of chronic conditions in 2017-2019.

*Code:*

|  |  |
| --- | --- |
| **File** | **Information** |
| model\_selection.R | This script trains the models on 80% of the data and tests on 20% of the data to select the best model among parameter combinations of L={5,10} and K={5,10}. The code also calculates the effect estimates from tensor margin estimates and makes trace plots to check for convergence. This script needs outputs from data\_generation.R. |
| final\_analyses.R | This script runs the final analyses (L=10,K=10) with longer chain and also runs the spatial (“naïve”) and non-spatial (OLS) models. |
| cross\_validation.R | This script contains functions for data analysis that allows for splitting data into training and test sets. |
| functions\_sim.R | This script contains mainly functions for simulation. |
| functions\_analysis.R | This script contains mainly functions for data analysis. |

# **Step 3: Compile results & generate figures/tables**

*Purpose:* This step uses the outputs from the analyses in Step 2 to generate figures and tables.

*Code:*

|  |  |
| --- | --- |
| **File** | **Information** |
| final\_results\_plots.R | This script reads in the output from final analyses and runs OLS models. It then calculates effect estimates and makes summary plots. |
| tensor\_summary.R | This script reads in the data analysis result and calculates the gamma tensor. It then uses the R package rTensor to conduct one CP decomposition and plots it. |

*Outputs:*

|  |  |
| --- | --- |
| **File** | **Information** |
| plots/ | All plots unless otherwise specified. |
|  |  |

# **Step 4: Simulation studies**

*Purpose:* This step generates data, runs simulations, and create plots and tables.

*Code:*

|  |  |
| --- | --- |
| **File** | **Information** |
| simulation.R | * main simulation: K=5, L=10, 100 replicates with two parameters ( and ) varying between 1 and 2. * sensitivity test for K={2,5,10}, L={5,10,20}, 20 replicates * needs B1.RData and B2.RData |
| sim\_result.Rmd | This script reads in simulation results and makes summary plots and tables. |

*Inputs:*

|  |  |
| --- | --- |
| **File** | **Information** |
| B1.RData | B1 matrix needed for simulation |
| B2.RData | B2 matrix needed for simulation |

*Outputs:*

|  |  |
| --- | --- |
| **File** | **Information** |
| plots/allres.png | This is the summary plot for the simulation study. |