Unveiling Vulnerability: Exploratory Data Analysis for the Enviornmental Justice Index

Thesis for a Master of Public Health, Epidemiology

Nathan Garcia-Diaz

Brown University, School of Public Health

Last Modified on August 5, 2024

Contents

Note: the table of contents acts as in-document hyperlinks

Statement of Purpose

This file is an Exploratory Data Analysis of the Social Vulnerability Index. Given the nature of the data, an examination of the spatial components is required. Specifically, this document will pull from two sources to help with the writing of the code: Ch 7.6 and 7.7 in Analyzing US Census Data: Methods, Maps, and Models in R (link) and Manny Gimond's A basic introduction to Moran's I analysis in R (link).

The EDA includes: distribution of variables, correlation matrix, and LOESS fitted scatter plots. Additionally, the examination of the spatial components include: Moran's I calculation with Monte Carlos Simulation, Moran's Scatterplot, Local Spatial Autocorrelation with Getis-Ord local G_i^*, and Hot/Cold Spot Identification.

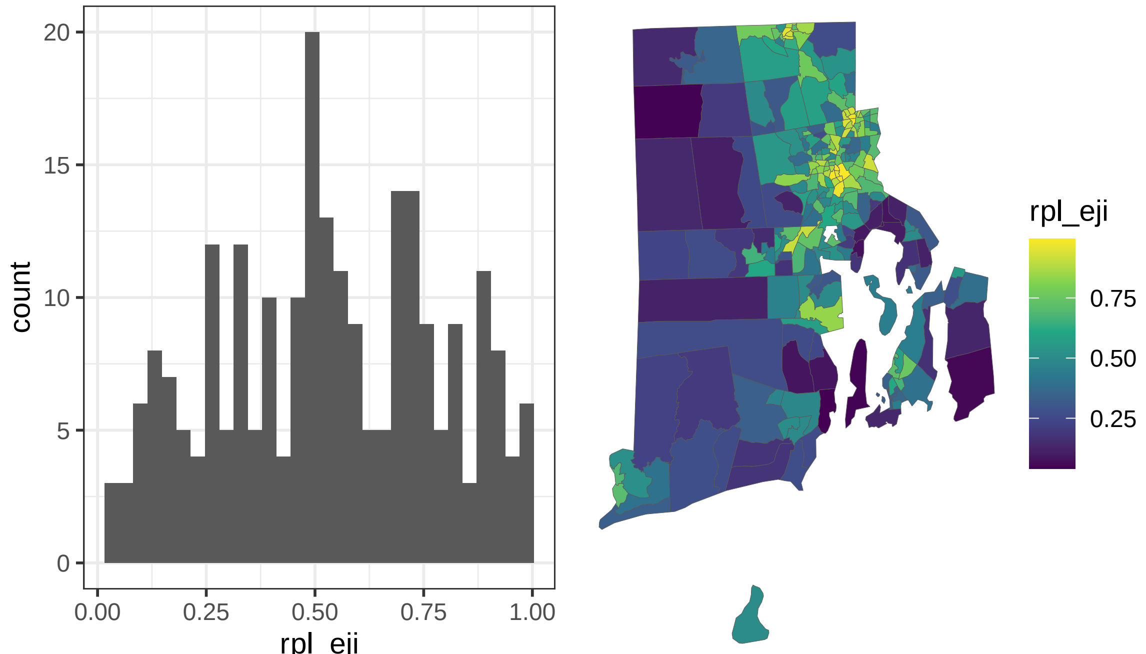
Definitions:

- **Moran's I**: Moran's I is a measure of spatial autocorrelation, quantifying the degree to which a variable is similarly distributed across neighboring geographic areas. It ranges from -1 (indicating perfect dispersion) to +1 (indicating perfect clustering), with values around 0 suggesting a random spatial pattern. It is used to detect and measure the presence of spatial autocorrelation, helping analysts understand whether the spatial distribution of a variable is clustered, dispersed, or random.
- **Monte Carlo Simulations**: Monte Carlo simulation is a computational technique that uses repeated random sampling to estimate the statistical properties of a system. It is used in tandem with Moran's I calculation to assess the significance of observed spatial autocorrelation by comparing it to the distribution of Moran's I values generated under the null hypothesis of spatial randomness. This is preformed as suggested by Gimond.
- **Moran's Scatter Plot**: Moran's scatterplot is a graphical representation that illustrates the relationship between a variable's values and the spatially lagged values of the same variable, used to visualize spatial autocorrelation. The plot typically includes a 45-degree reference line and divides the data points into four quadrants to help identify patterns of clustering or dispersion. It is used to diagnose and visualize spatial autocorrelation, helping to identify patterns of spatial clustering or dispersion in a dataset.
- **Local Spatial Autocorrelation**: Local measures of spatial autocorrelation, like the Getis-Ord local G_i^*, are used to identify clusters or "hot spots" of similar values within a spatial dataset. The Getis-Ord local G_i^* statistic specifically measures the degree of clustering of high or low values around each point, indicating areas with significant local spatial association.
  - *Positive Gi Values*: indicate areas where high values of rpl_themes are surrounded by other high values, or low values are surrounded by other low values. This suggests clustering of similar values.
  - *Negative Gi Values*: Indicate areas where high values of rpl_themes are surrounded by low values, or vice versa. This suggests spatial outliers or contrast.

Distribution of Variables

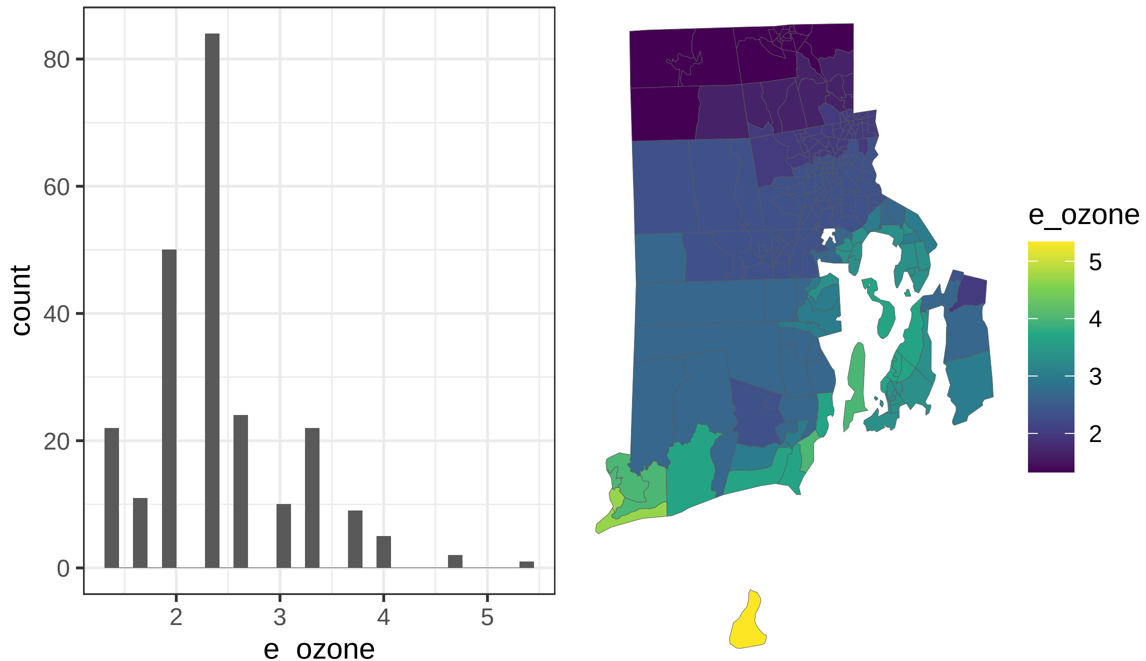
This section will produce graphs that contain both non-spatial and spatial distribution of predictor variables.

Distibution of Overall Index Ranking



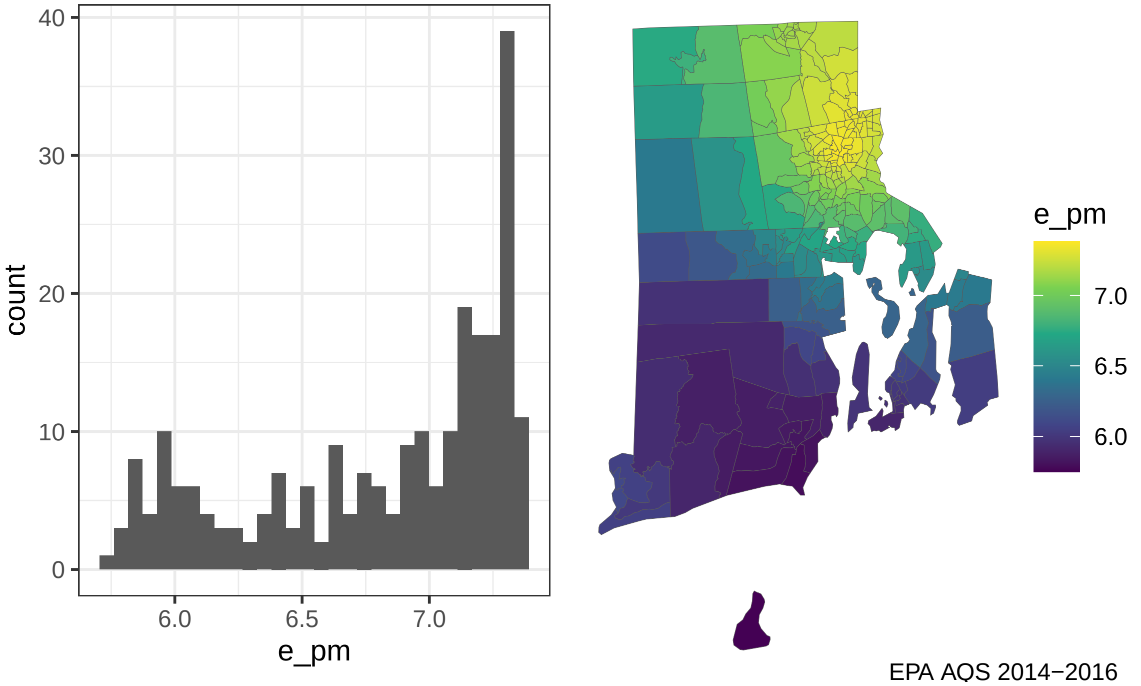
Rhode Island, 2022

Distibution of Annual mean days above O3 regulatory standard (3 yr avg)



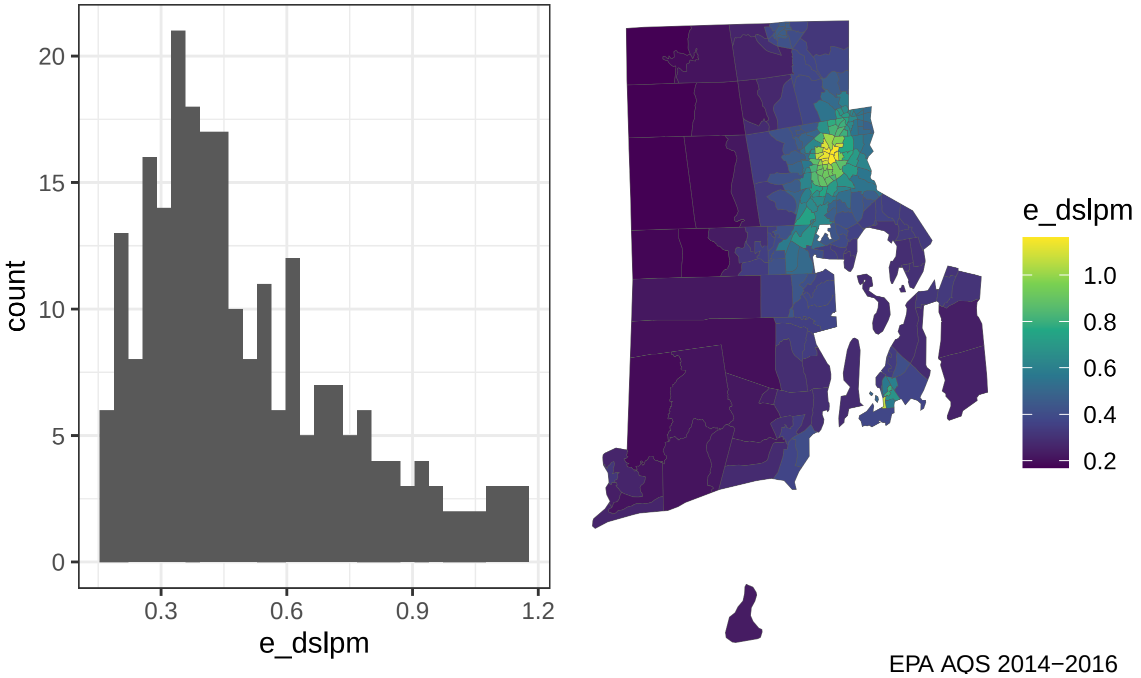
EPA AQS 2014−2016

Distibution of Annual mean days above PM2.5 regulatory standard (3 yr avg)



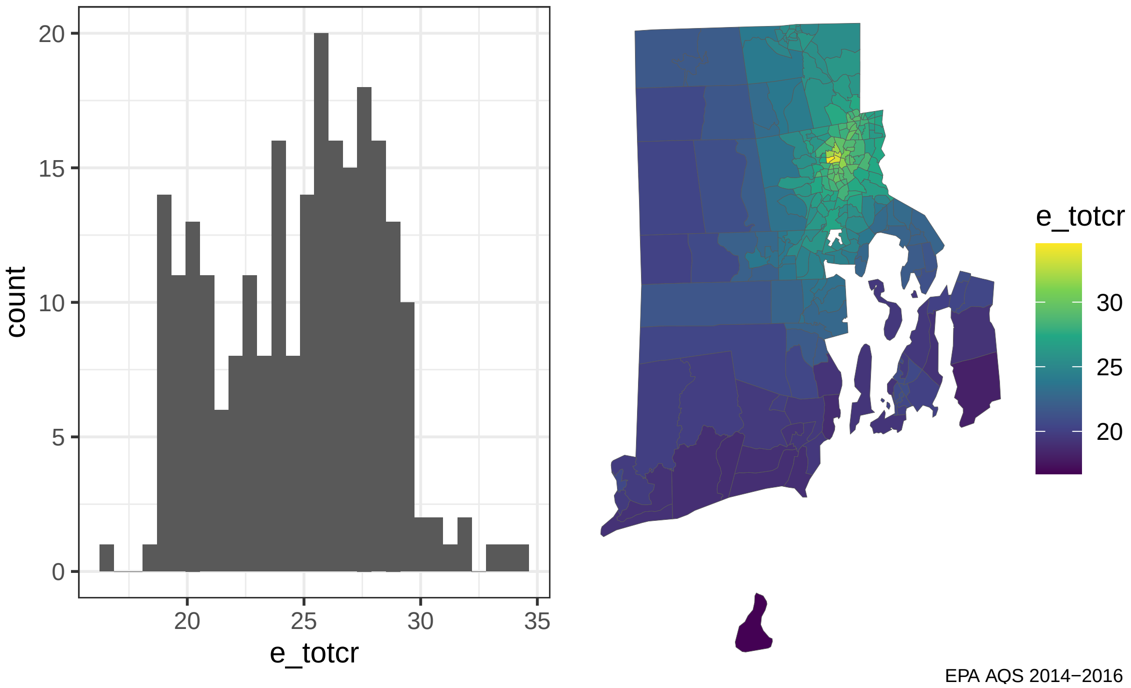
EPA AQS 2014−2016

Distibution of Ambient concentrations of Diesel (PM/m3)



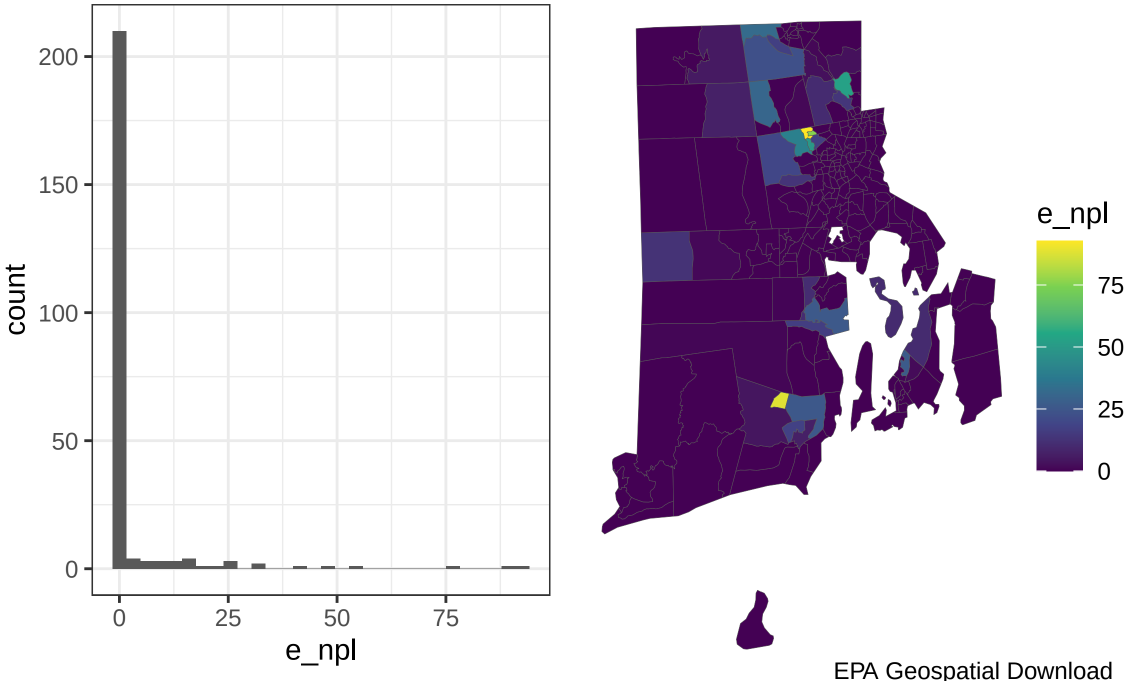
EPA AQS 2014−2016

Distibution of Probability of Contracting Cancer (assuming continuous exposure)



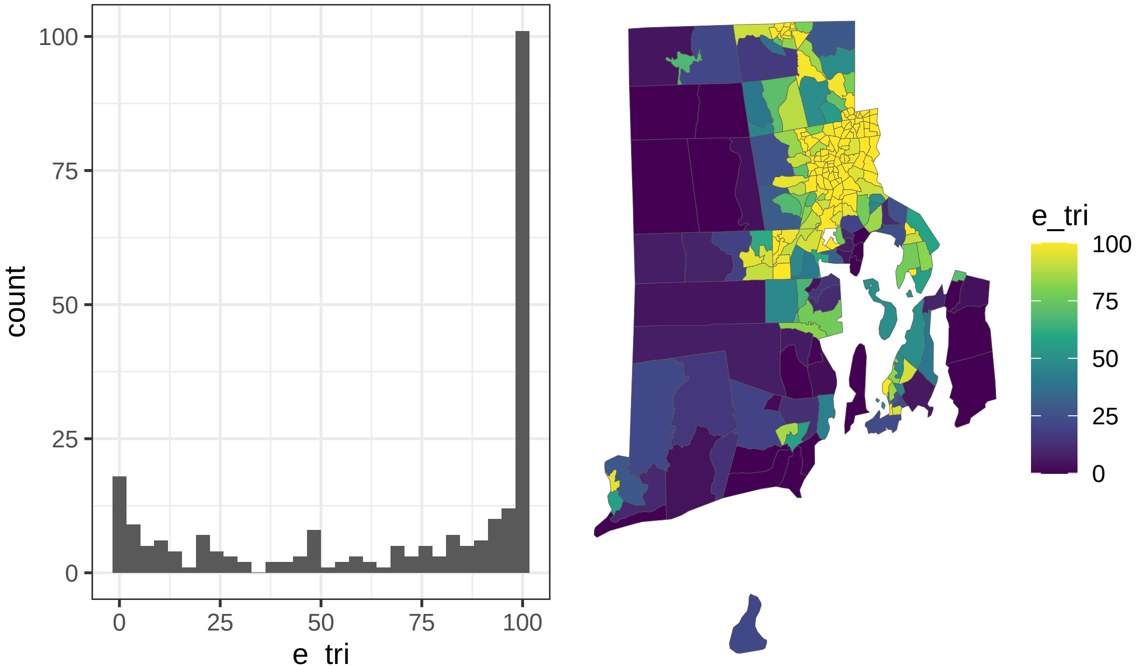
EPA AQS 2014−2016

Distibution of Proportion of Tract's Area within 1−mi buffer of EPA National Priority List Site



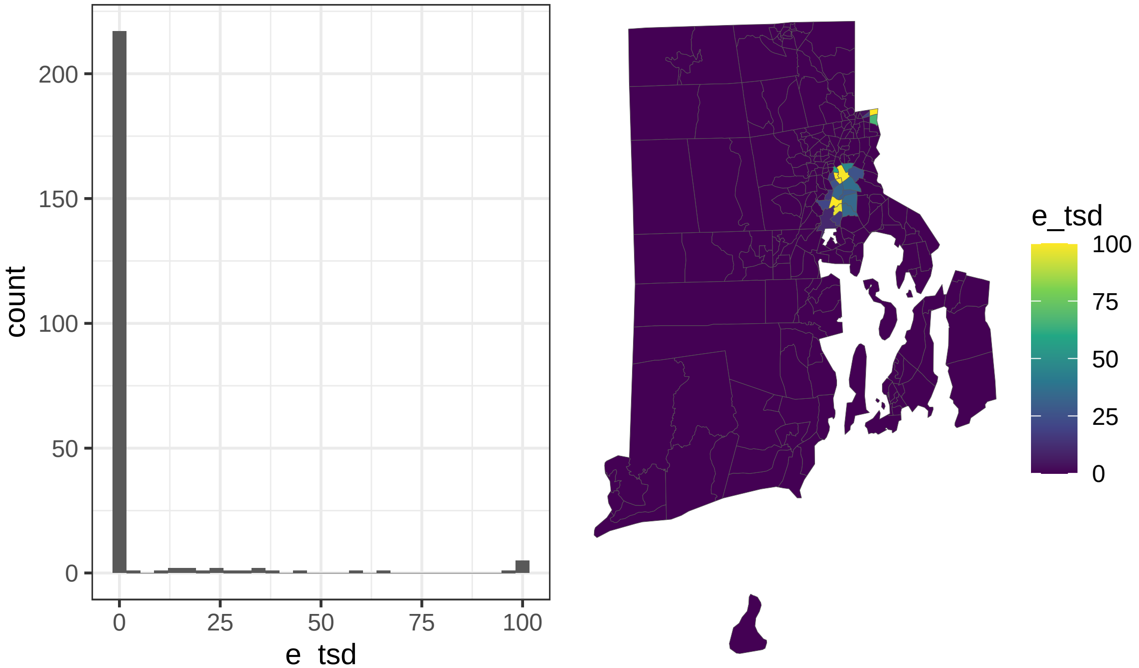
EPA Geospatial Download

Distibution of Proportion of Tract's Area within 1−mi buffer of EPA Toxic Release Inventory Site



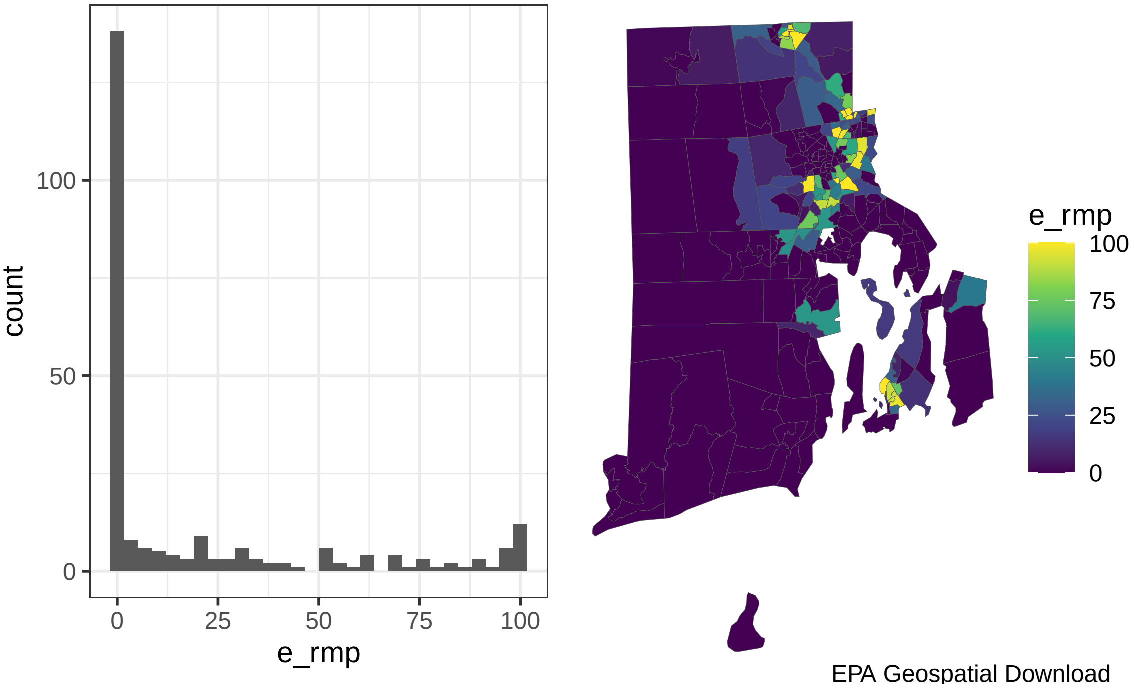
EPA Geospatial Download

Distibution of Proportion of Tract's Area within 1−mi buffer of EPA Treatment, Storage, and Disposal site



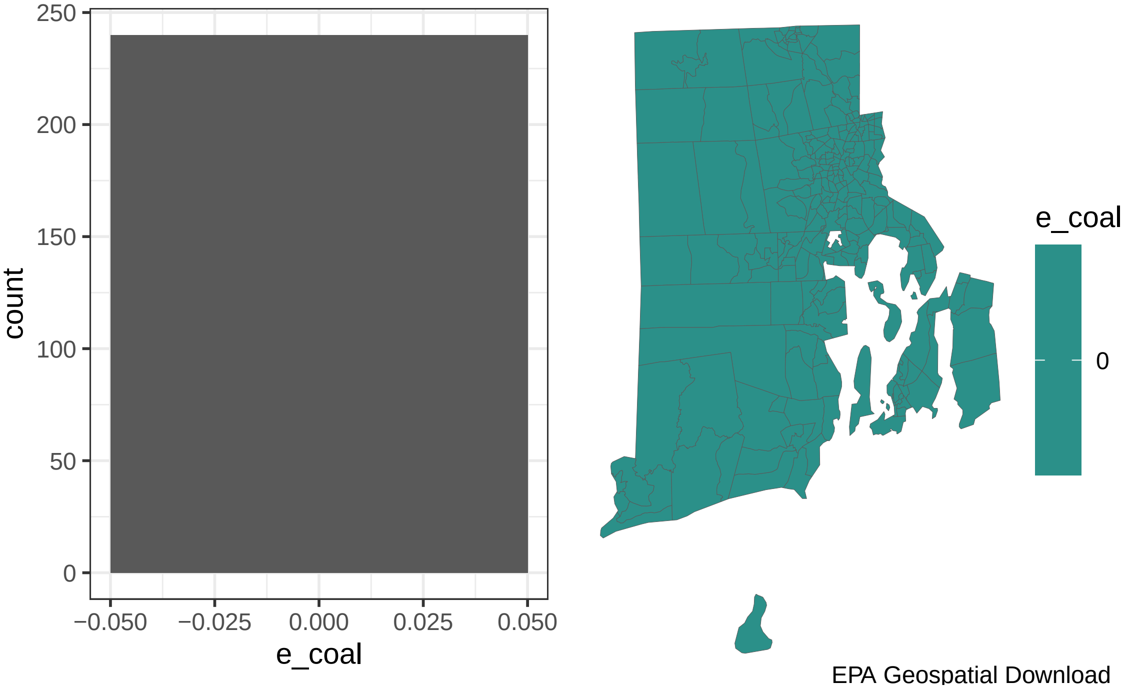
EPA Geospatial Download

Distibution of Proportion of Tract's Area within 1−mi buffer of EPA risk management plan site



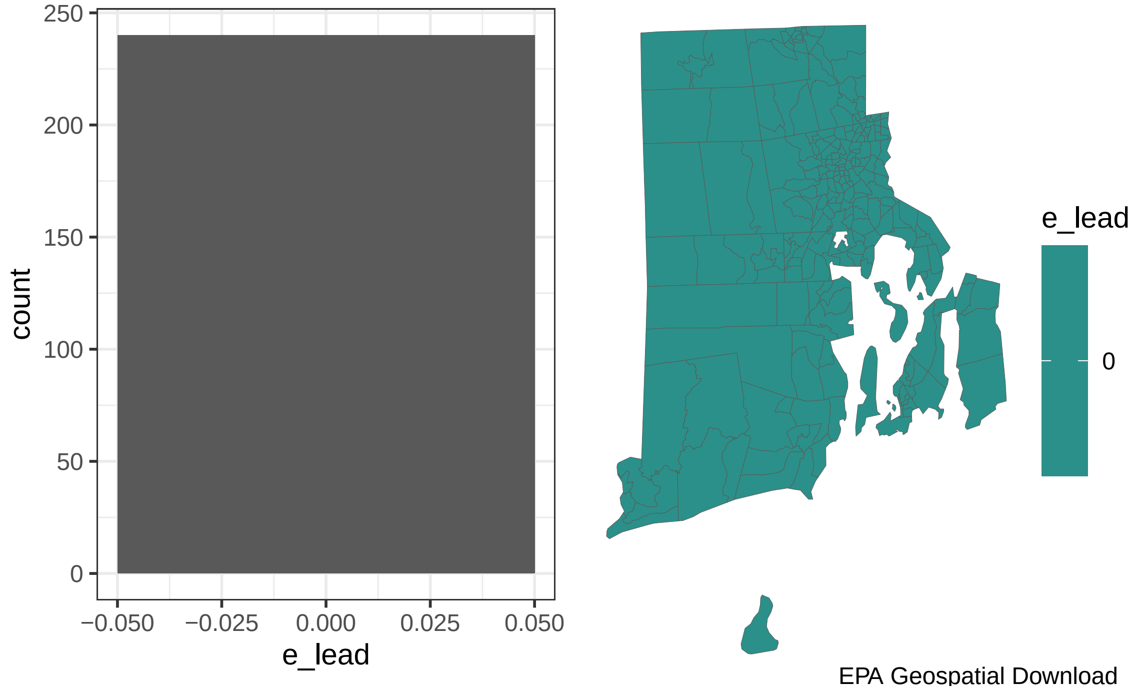
EPA Geospatial Download

Distibution of Proportion of tract's area within 1−mi buffer of coal mines



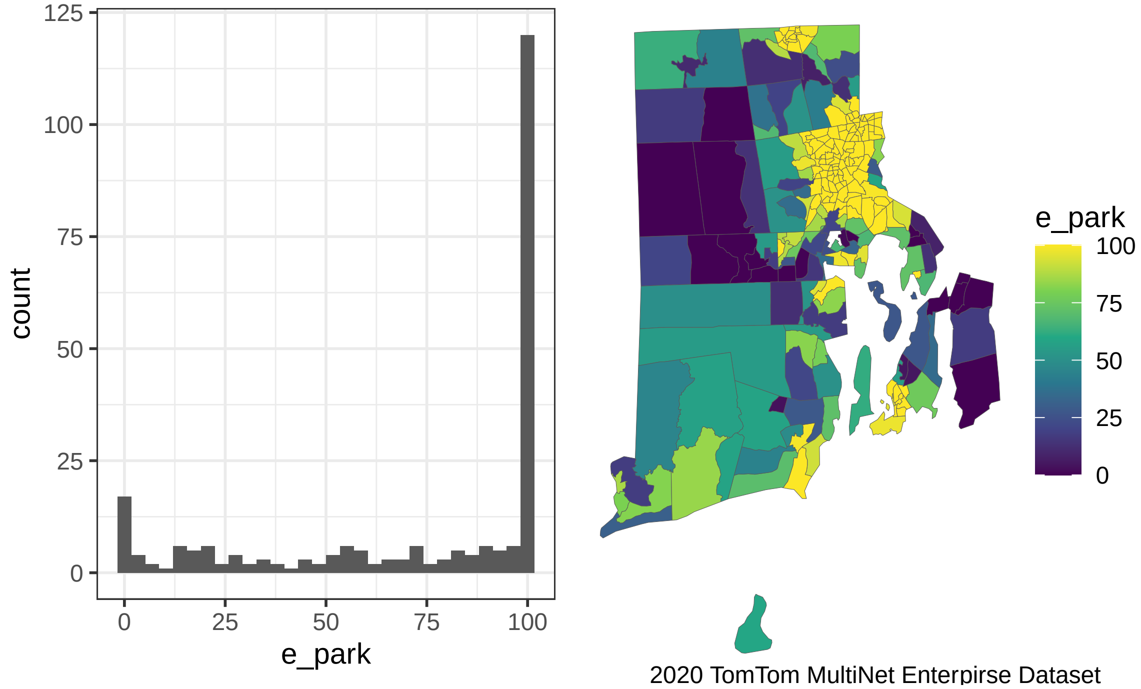
EPA Geospatial Download

Distibution of Proportion of tract's area within 1−mi buffer of lead mines



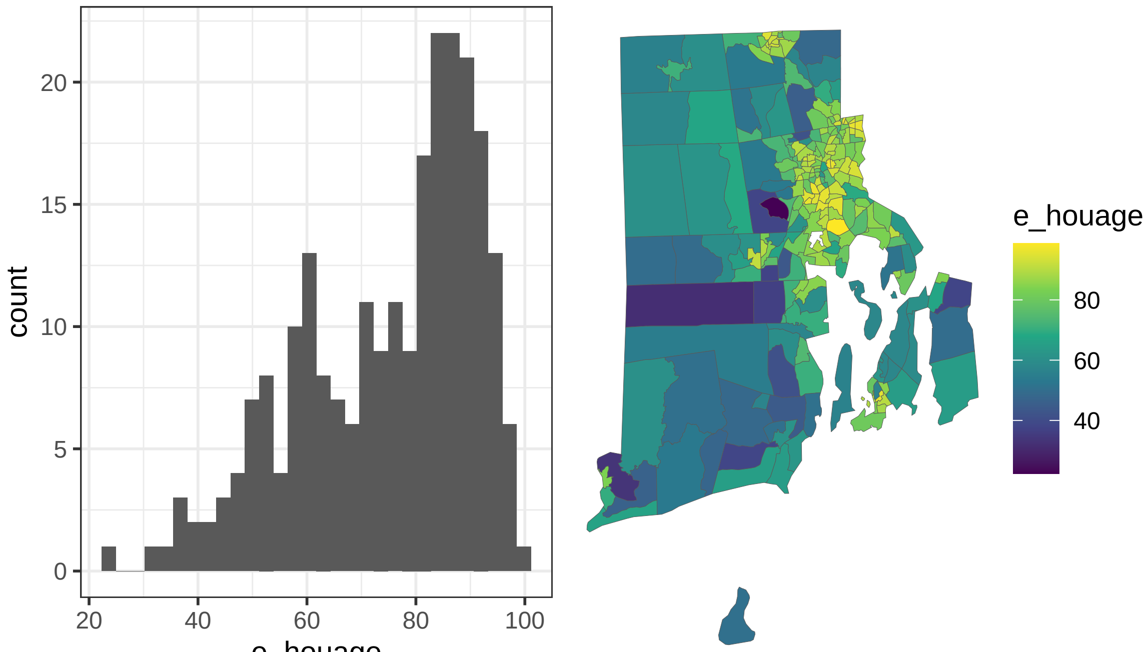
EPA Geospatial Download

Distibution of Proportion of tract's area within 1−mi buffer of green space



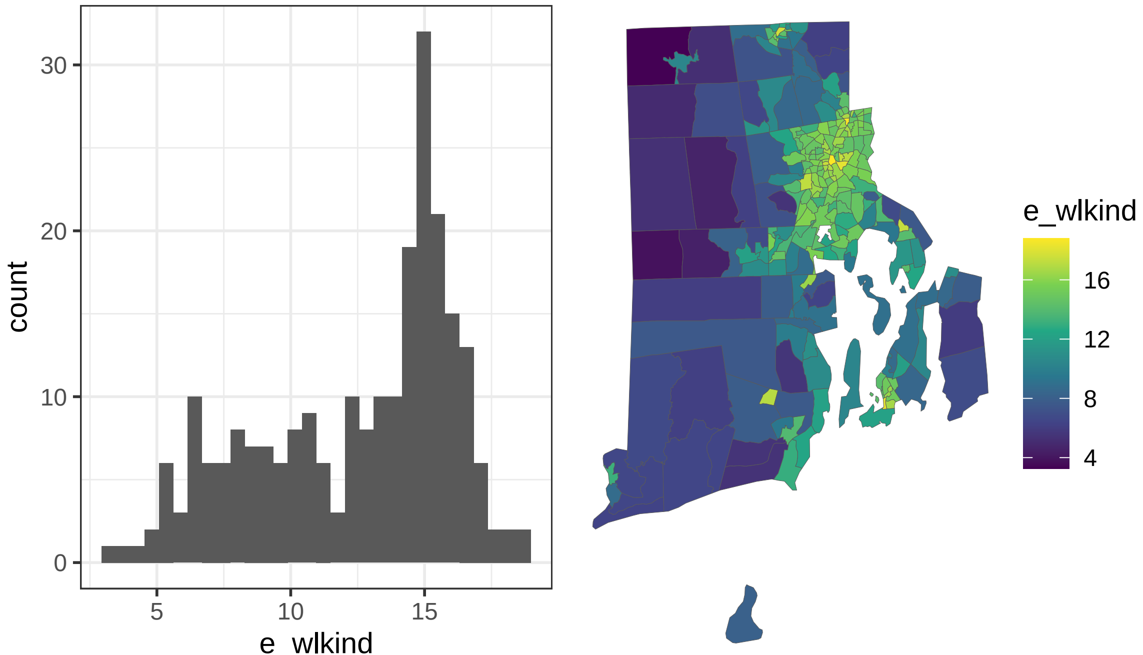
2020 TomTom MultiNet Enterpirse Dataset

Distibution of Percentage of houses built pre 1980 (lead exposure)



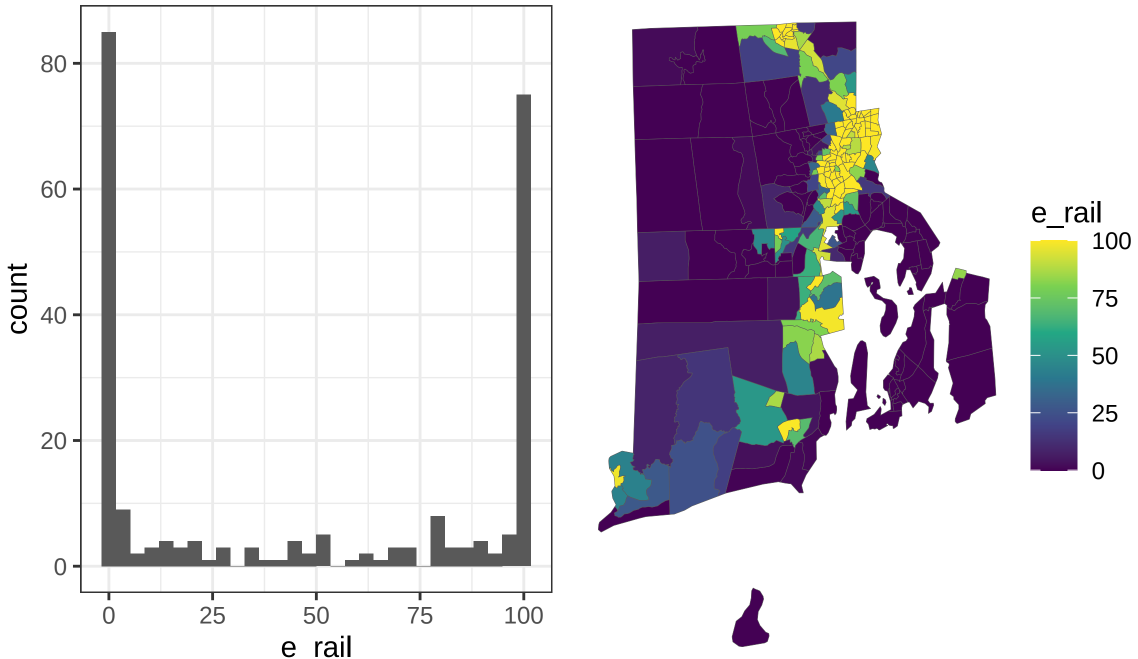
No Citation

Distibution of Walkability Values



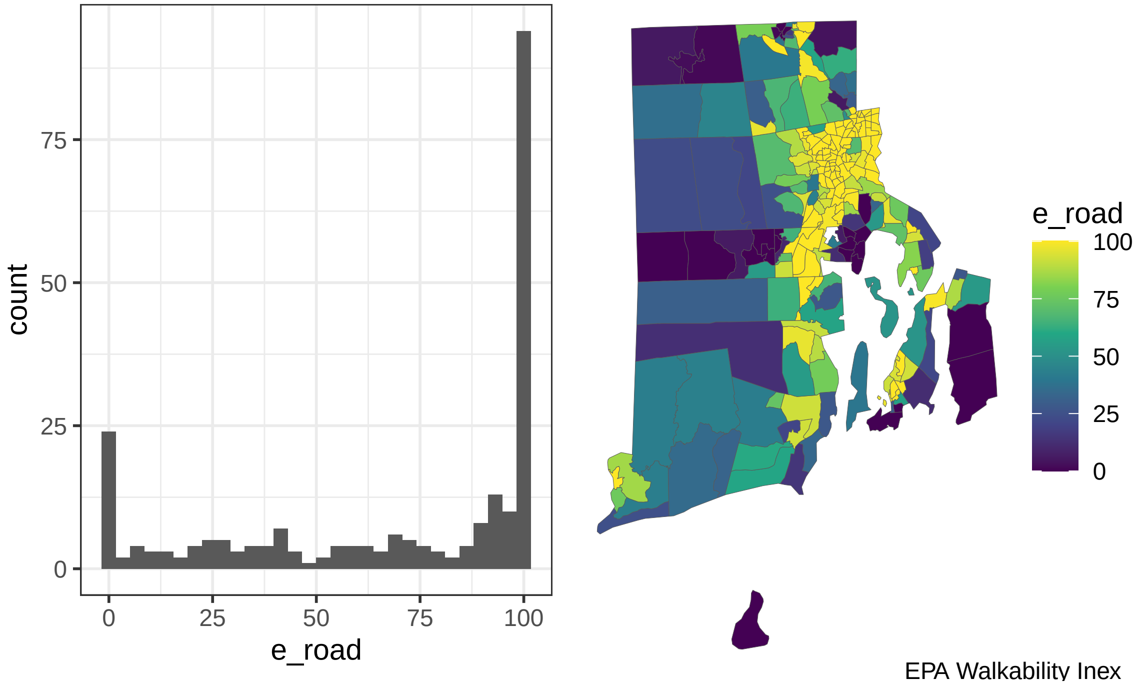
EPA Walkability Inex

Distibution of Proportion of tract's area within 1−mi buffer of railroad



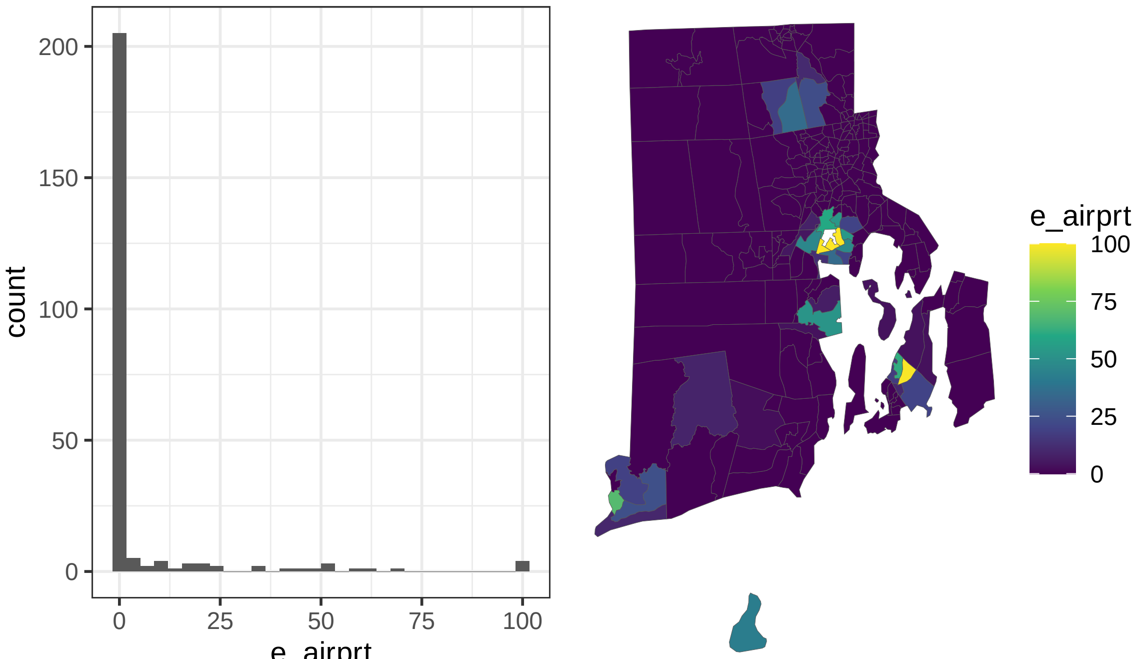
EPA Walkability Inex

Distibution of Proportion of tract's area within 1−mi buffer of high volume road or highway



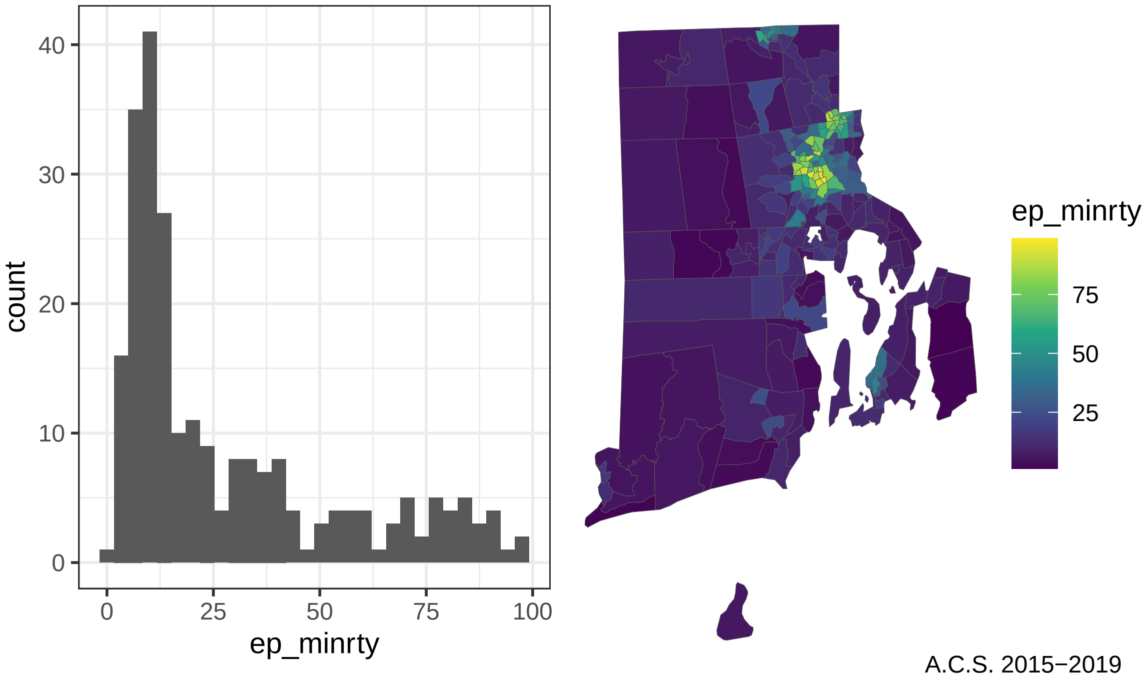
EPA Walkability Inex

Distibution of Proportion of tract's area within 1−mi buffer of airport



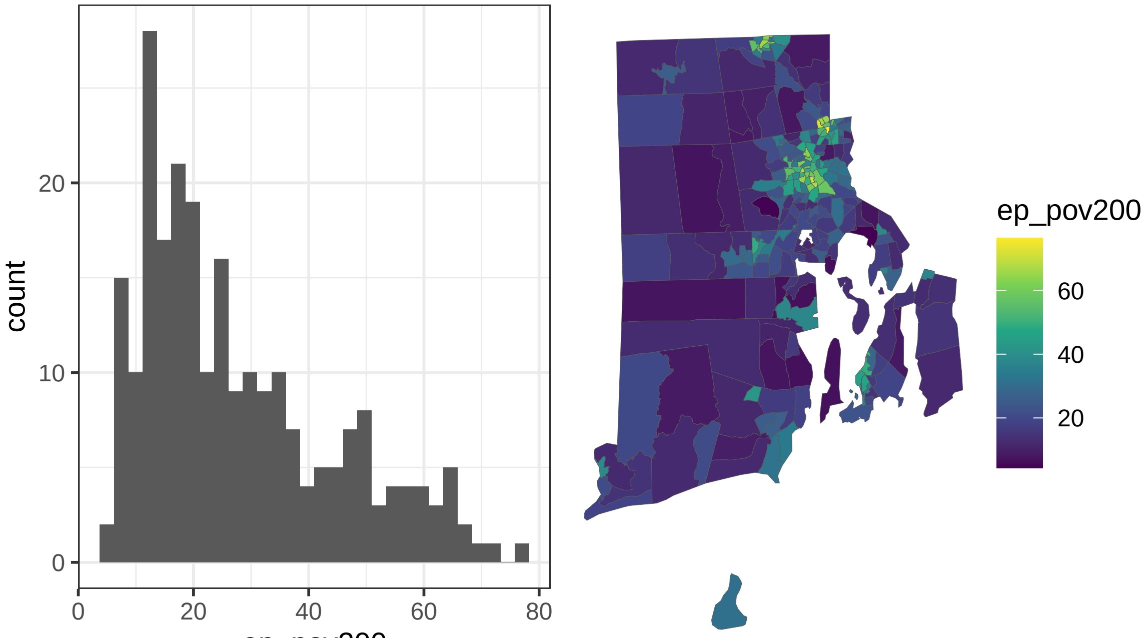
EPA Walkability Inex

Distibution of Percentage of Minority Persons



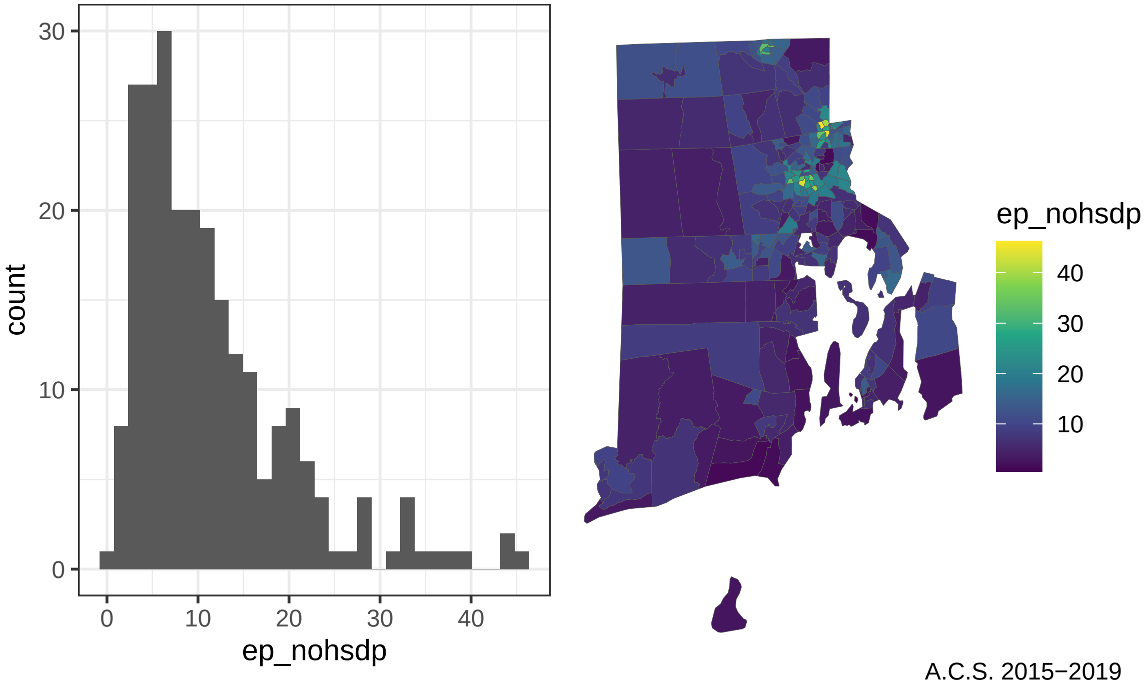
A.C.S. 2015−2019

Distibution of Percentage of Persons Below 200% Poverty

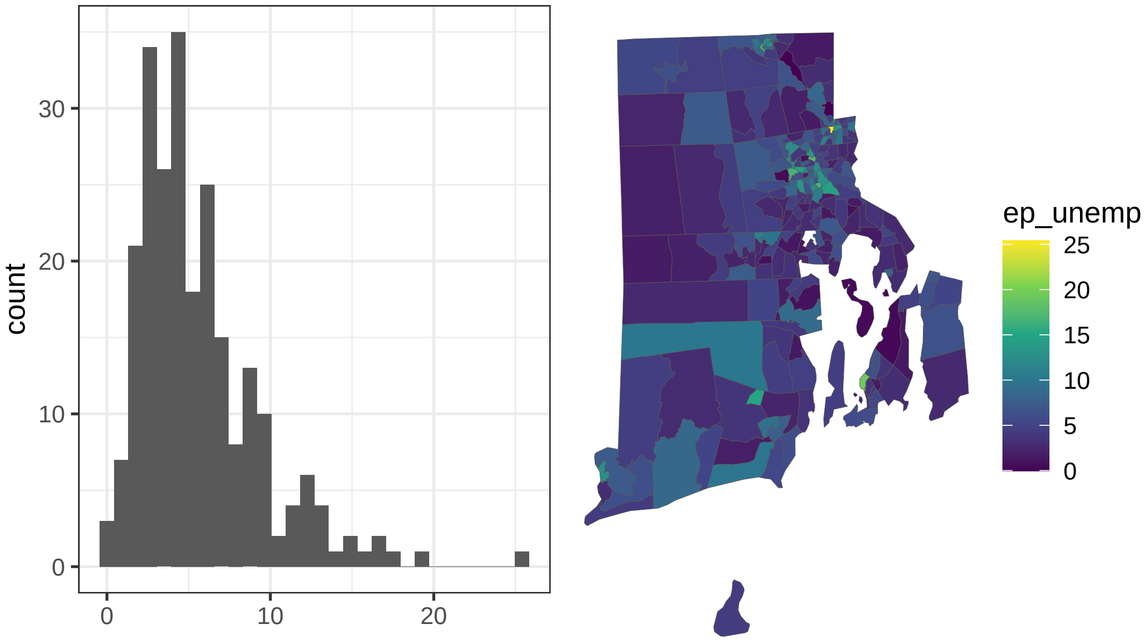


A.C.S. 2015−2019

Distibution of Percentage of Persons with no High School Diploma (age 25+)

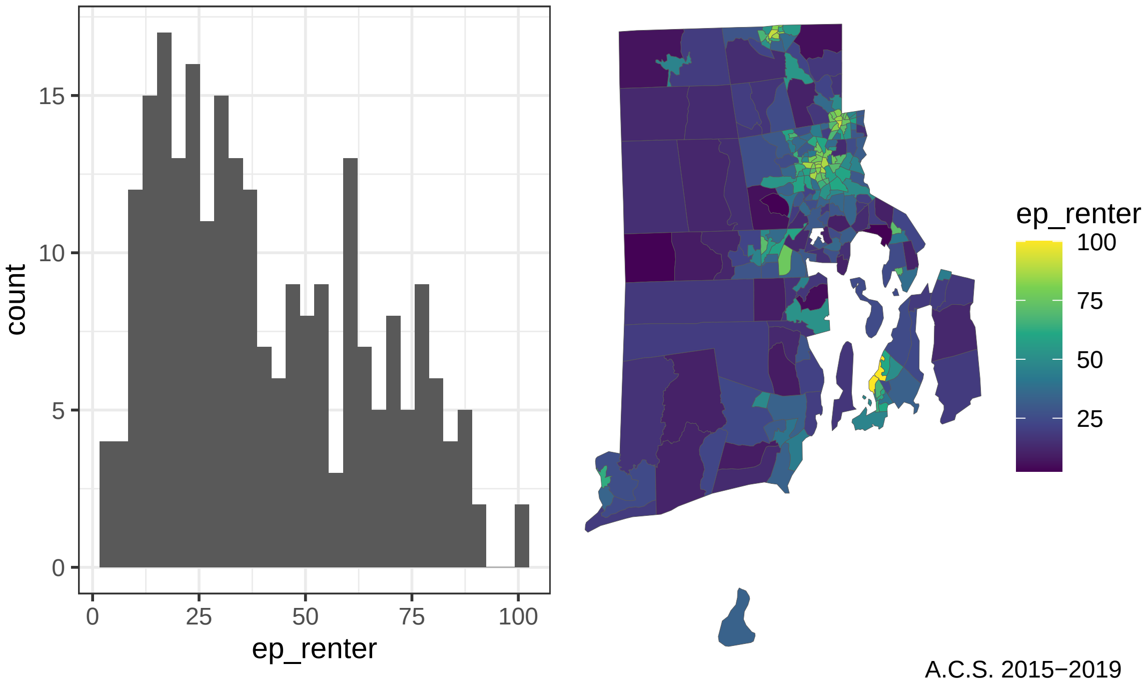


A.C.S. 2015−2019



Distibution of Percentage of Unemployed Persons

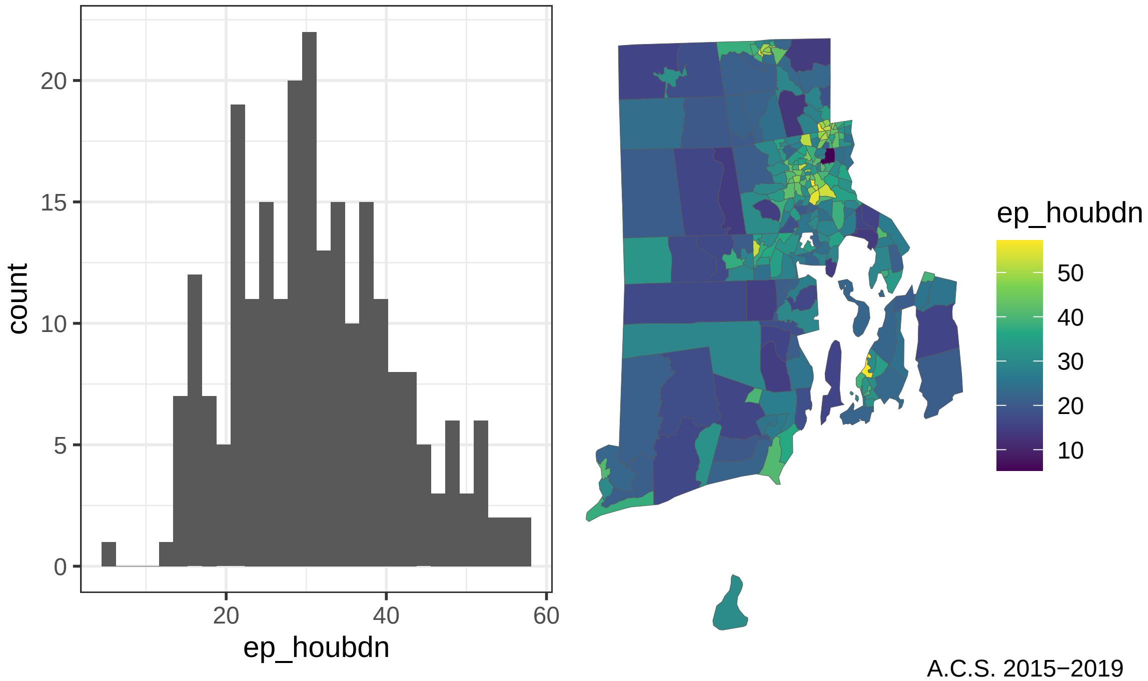
A.C.S. 2015−2019



Distibution of Percentage of Renters

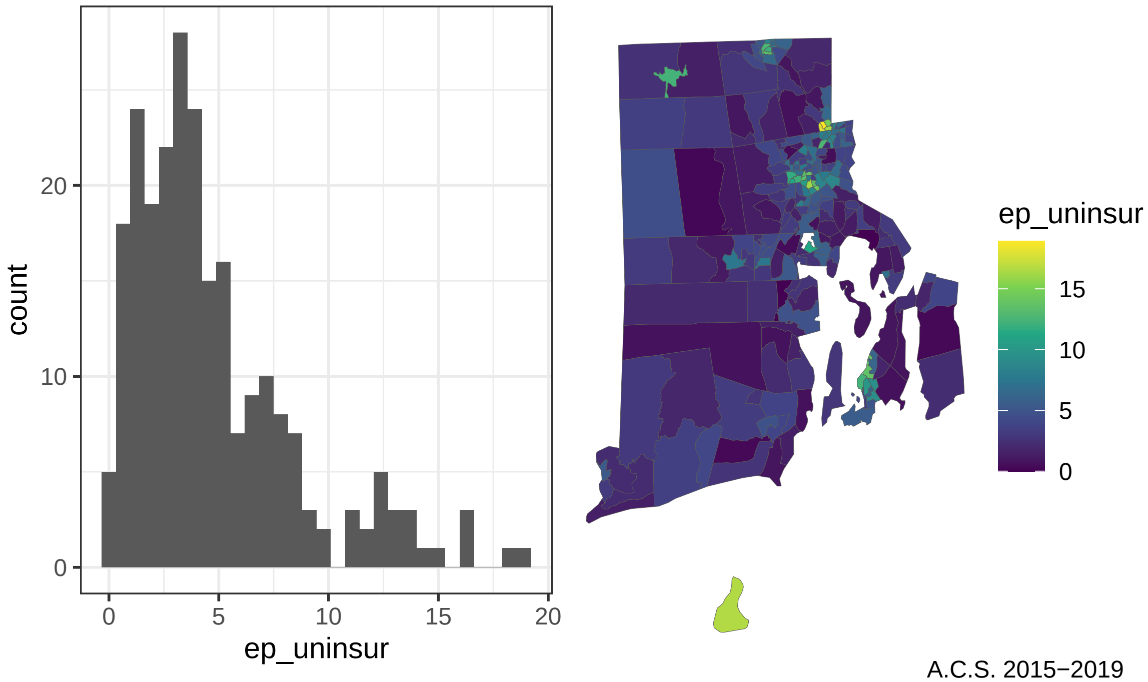
A.C.S. 2015−2019

Distibution of Percentage of households that make less than $75,000



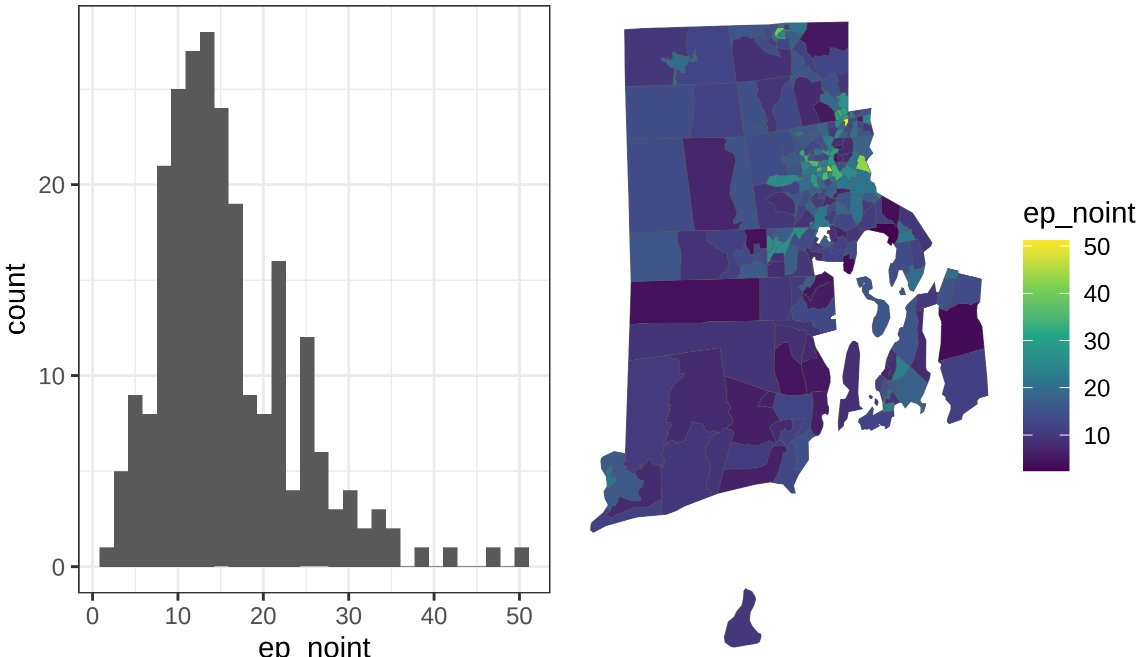
A.C.S. 2015−2019

Distibution of Percentage of Unisured Persons



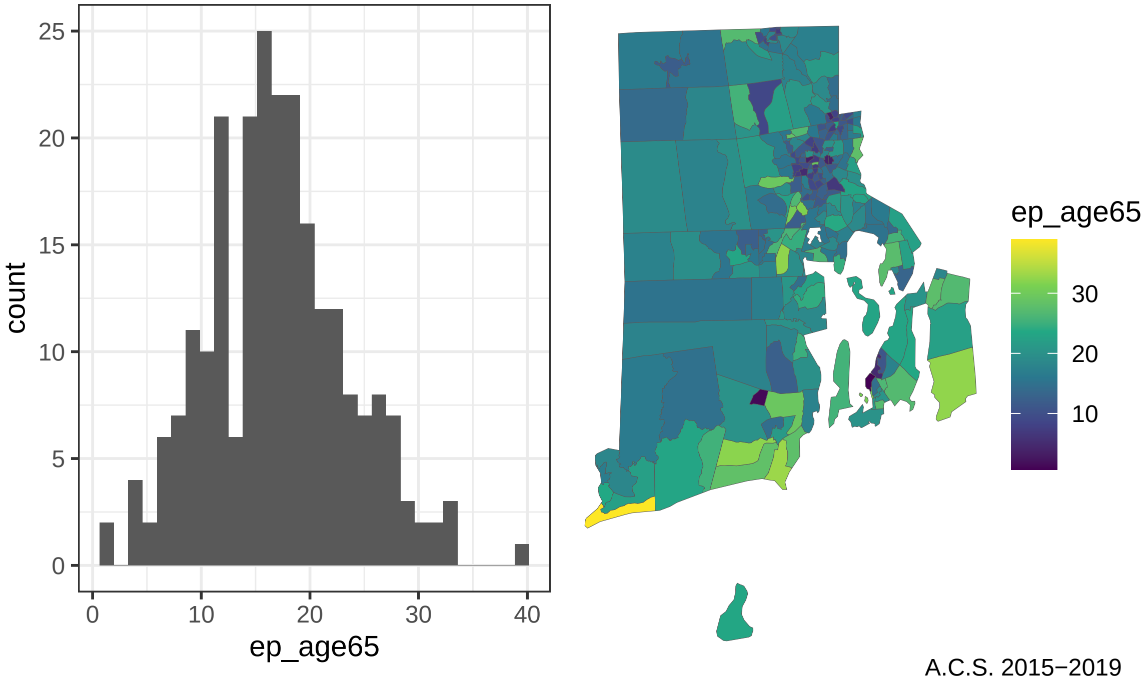
A.C.S. 2015−2019

Distibution of Percentage of Persons Without Internet



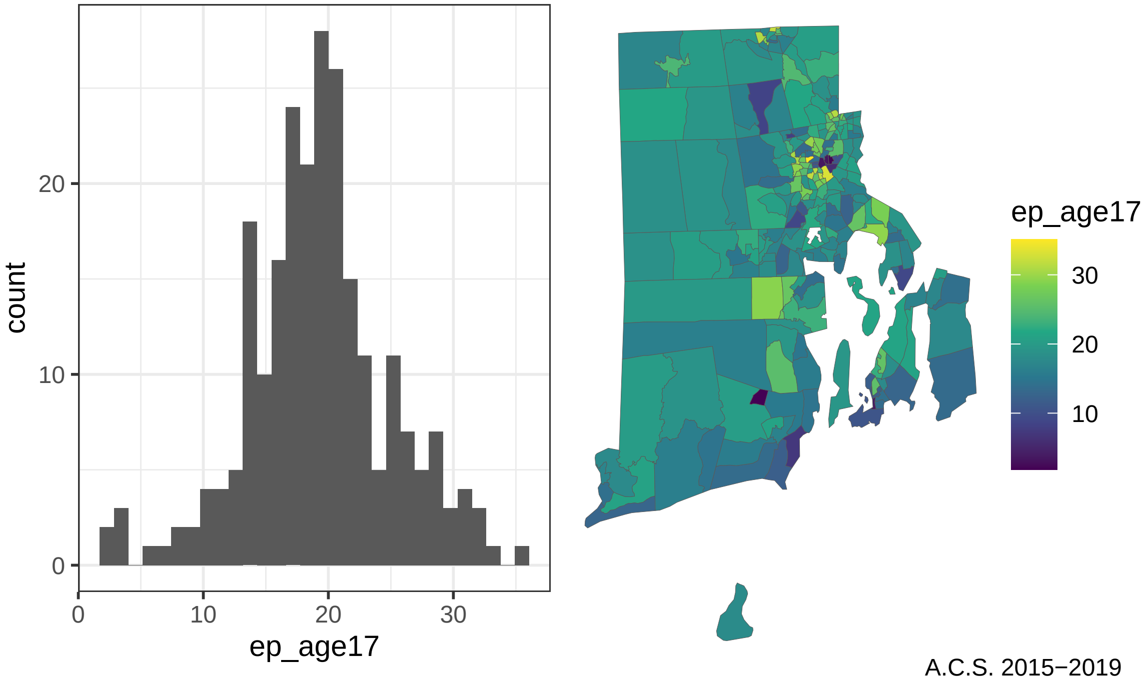
A.C.S. 2015−2019

Distibution of Percentage of persons aged 65 and older



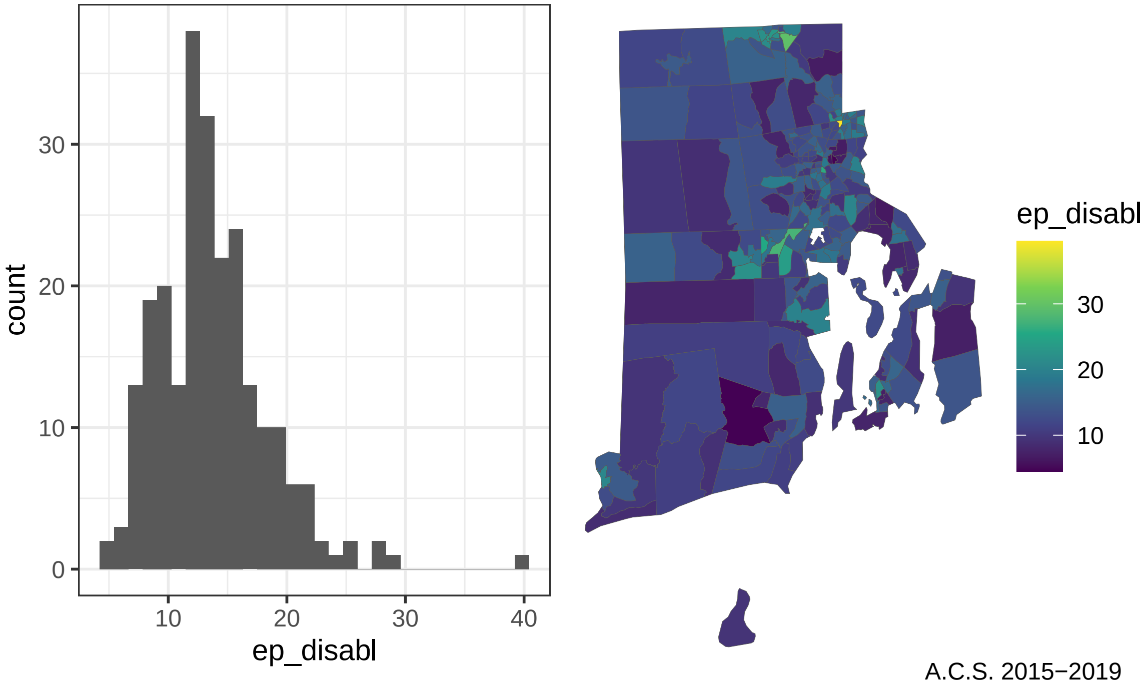
A.C.S. 2015−2019

Distibution of Percentage of persons aged 17 and younger



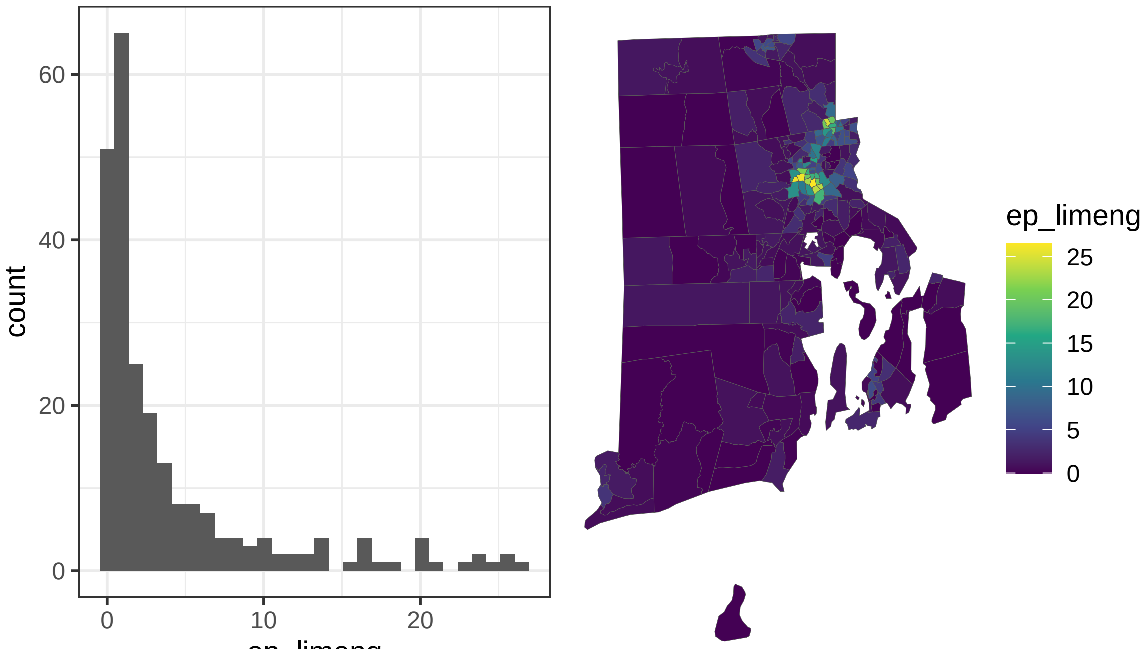
A.C.S. 2015−2019

Distibution of Percentage of civilian noninstitutionalized population with a disability



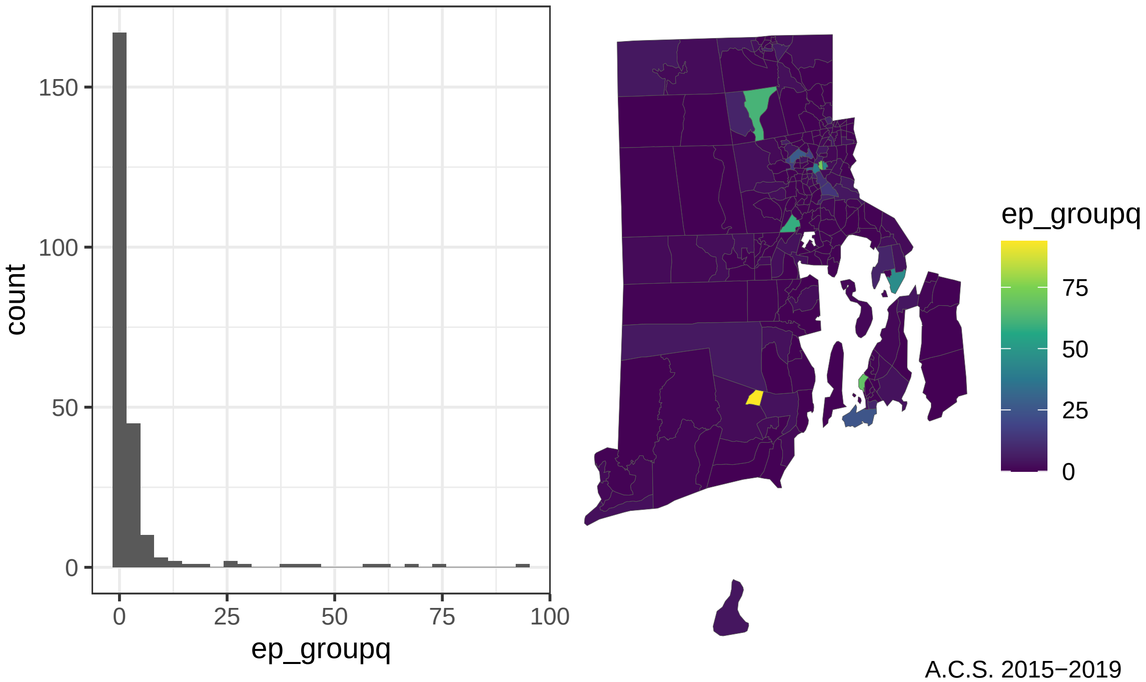
A.C.S. 2015−2019

Distibution of Percentage of persons (age 5+) who speak English 'less than well'

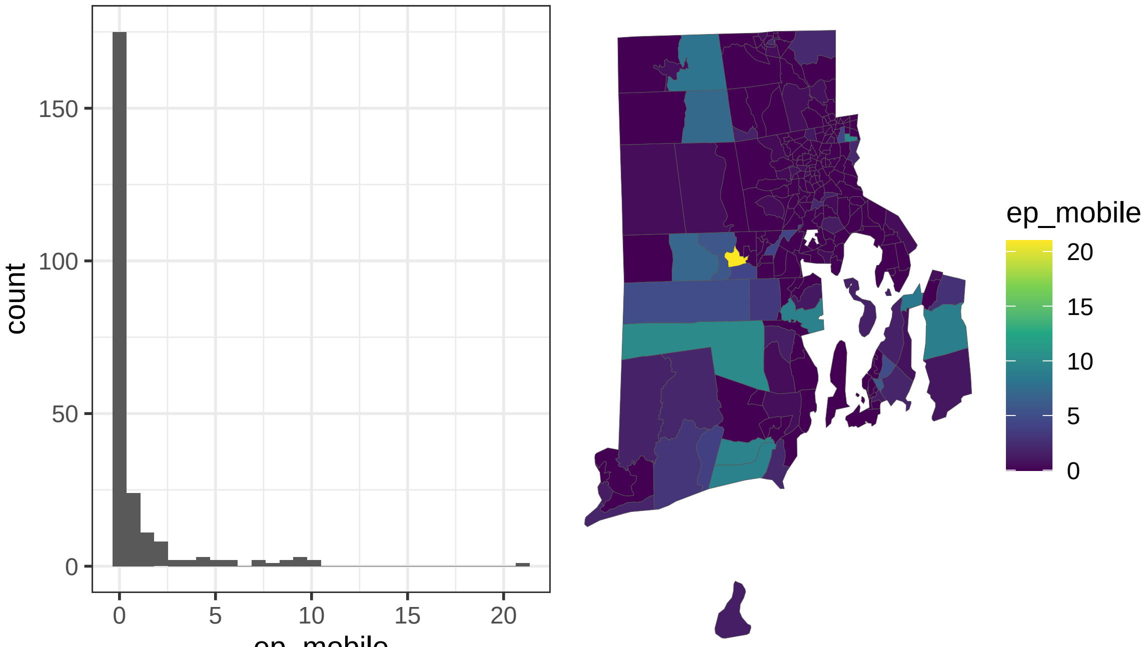


A.C.S. 2015−2019

Distibution of Percentage of persons in group quarters estimate

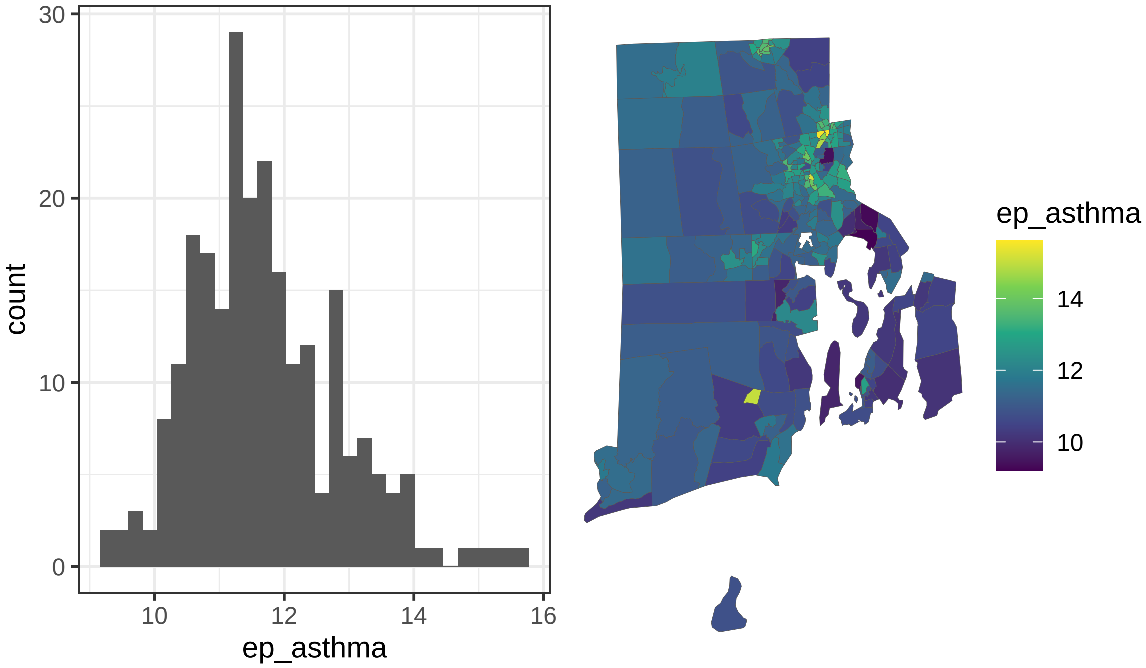


A.C.S. 2015−2019



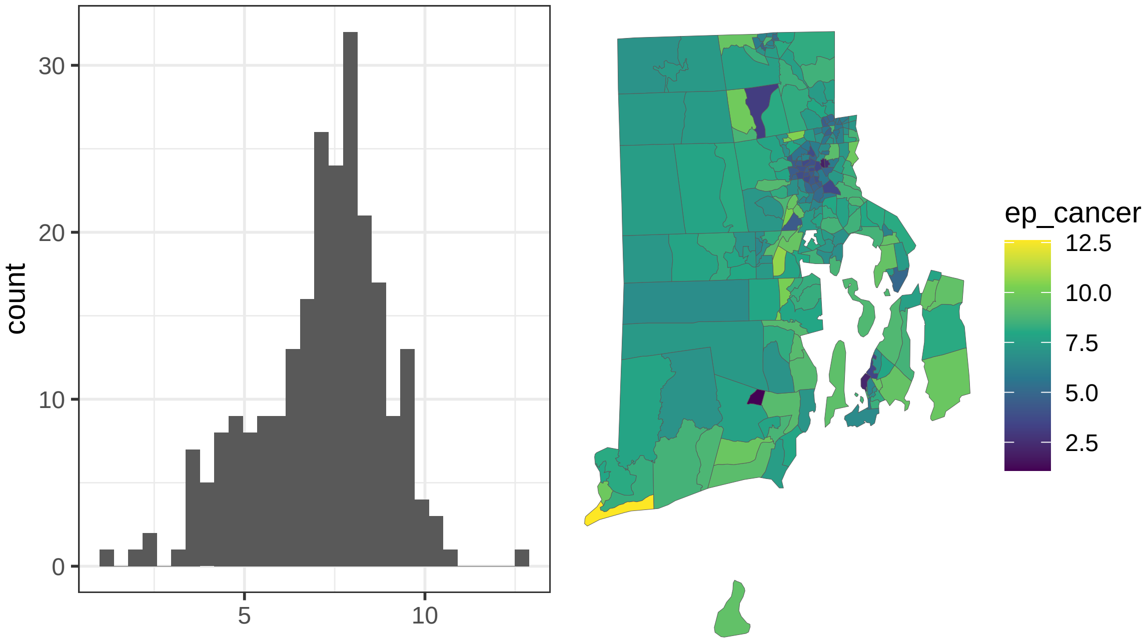
Distibution of Percentage of Mobile Homes

A.C.S. 2015−2019

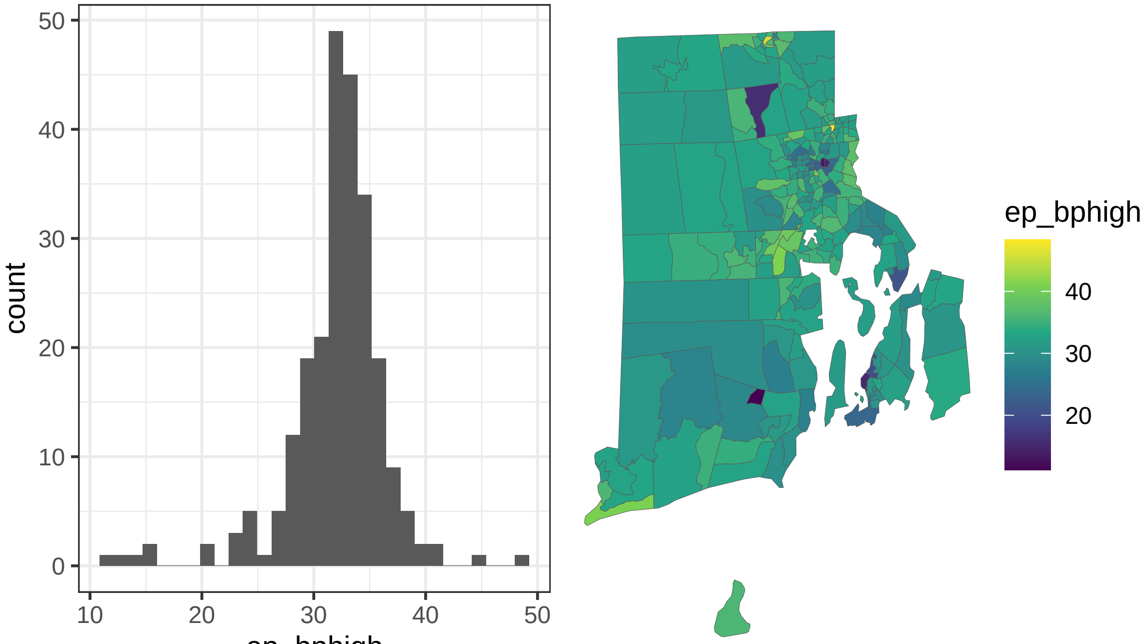


Distibution of Percentage Individuals With Asthma

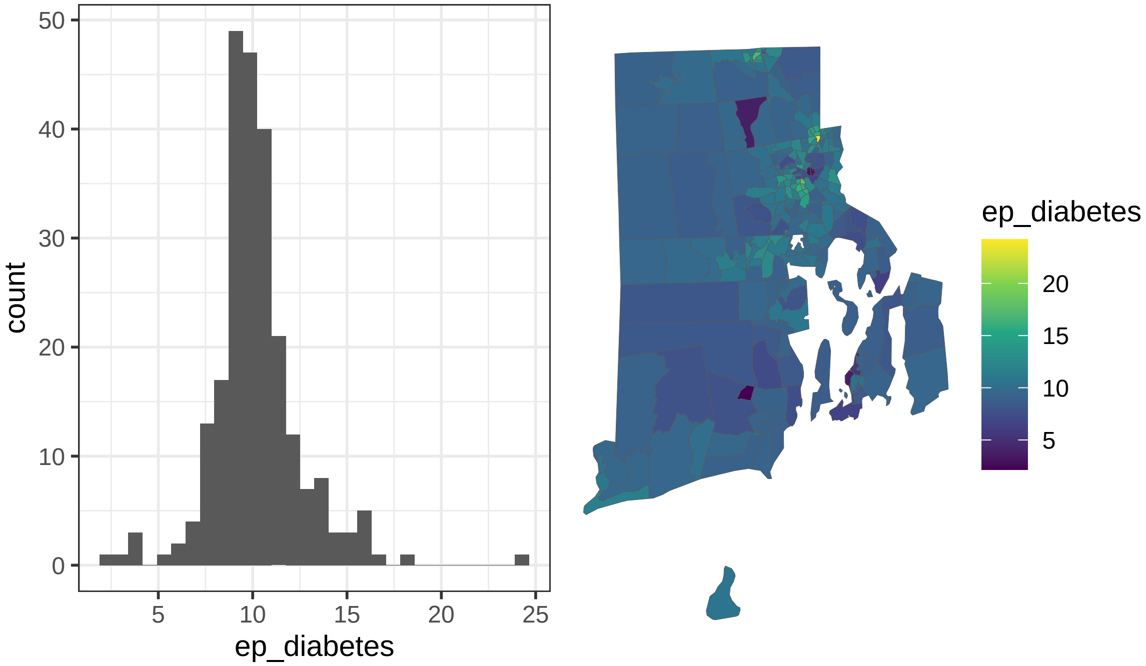
Distibution of Percentage Individuals With Cancer



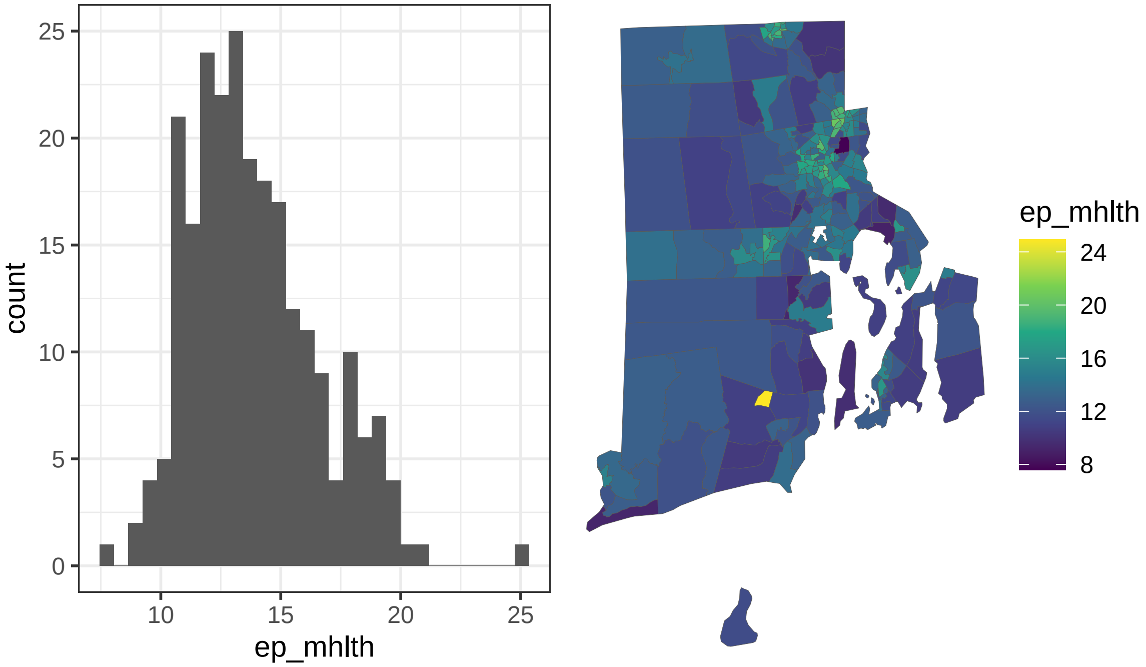
Distibution of Percentage Individuals With High Blood Pressure



Distibution of Percentage Individuals With Diabetes

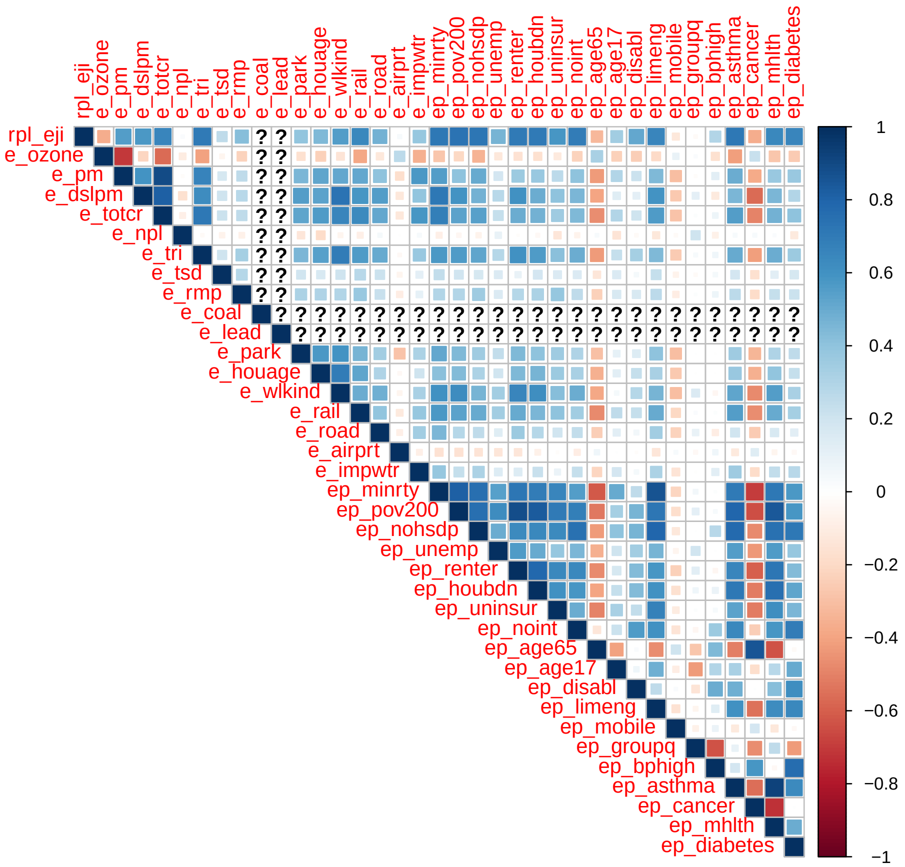


Distibution of Percentage Individuals No Reporting Good Mental Health



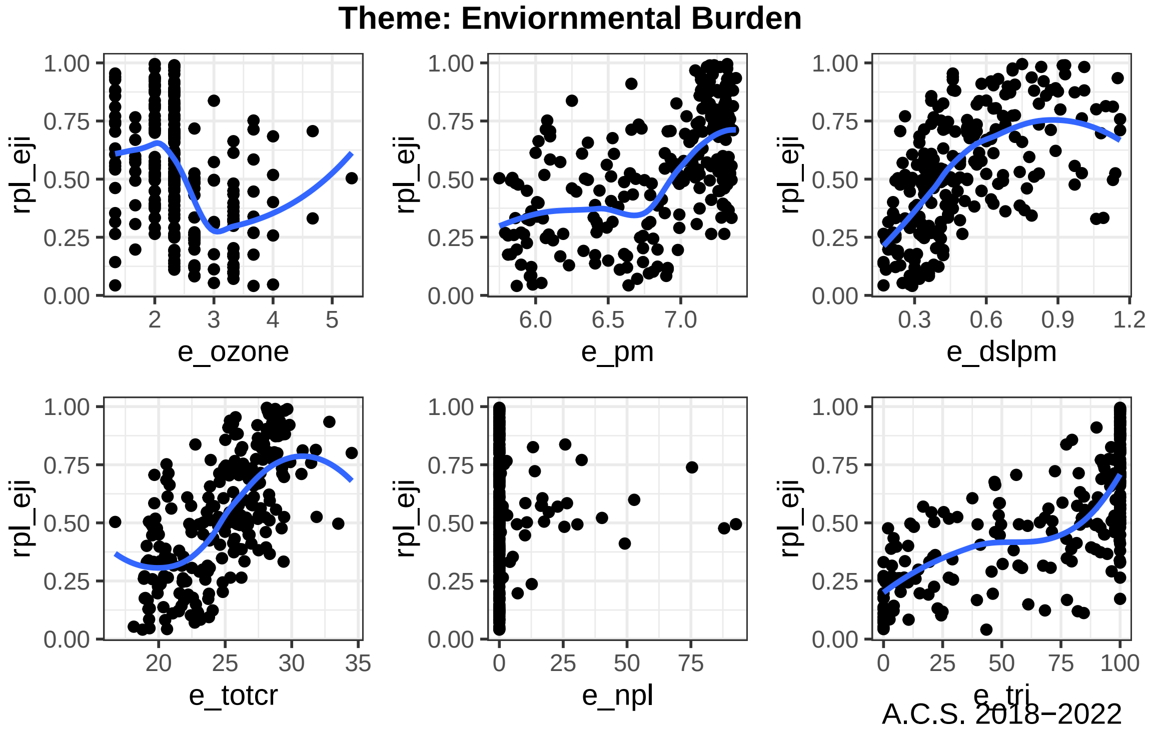
Correlation Matrix

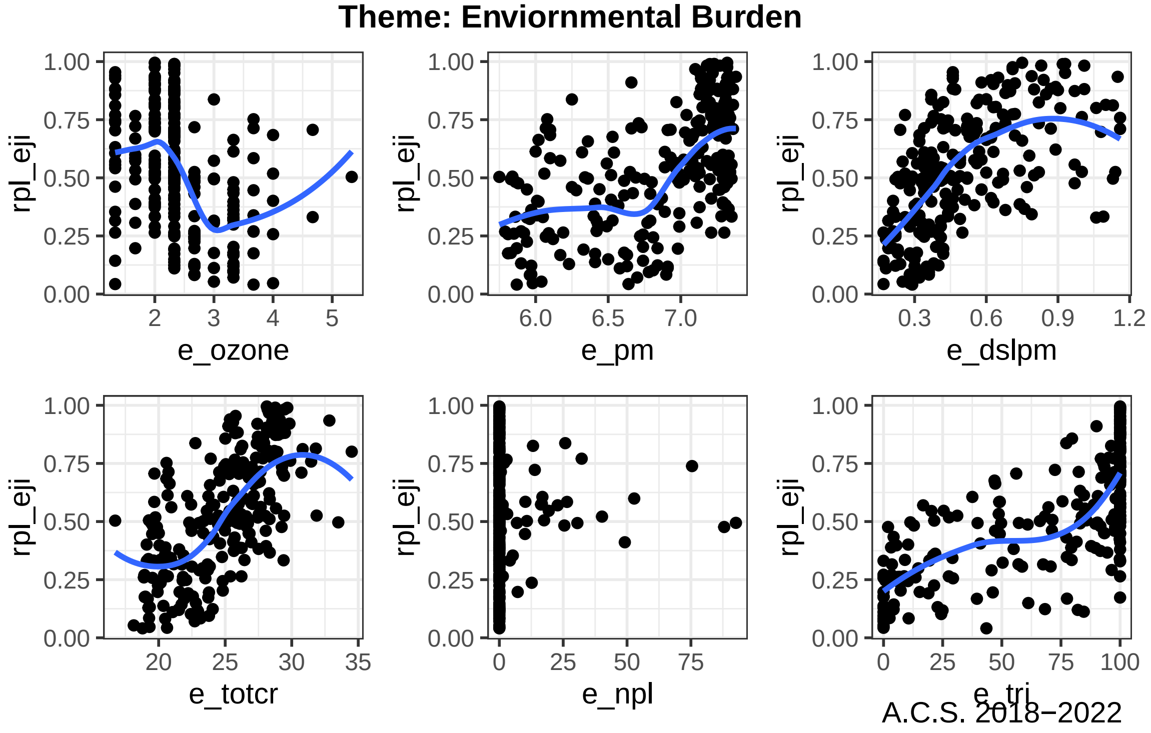
The “?” refers to the fact that these variables have a value of zero.

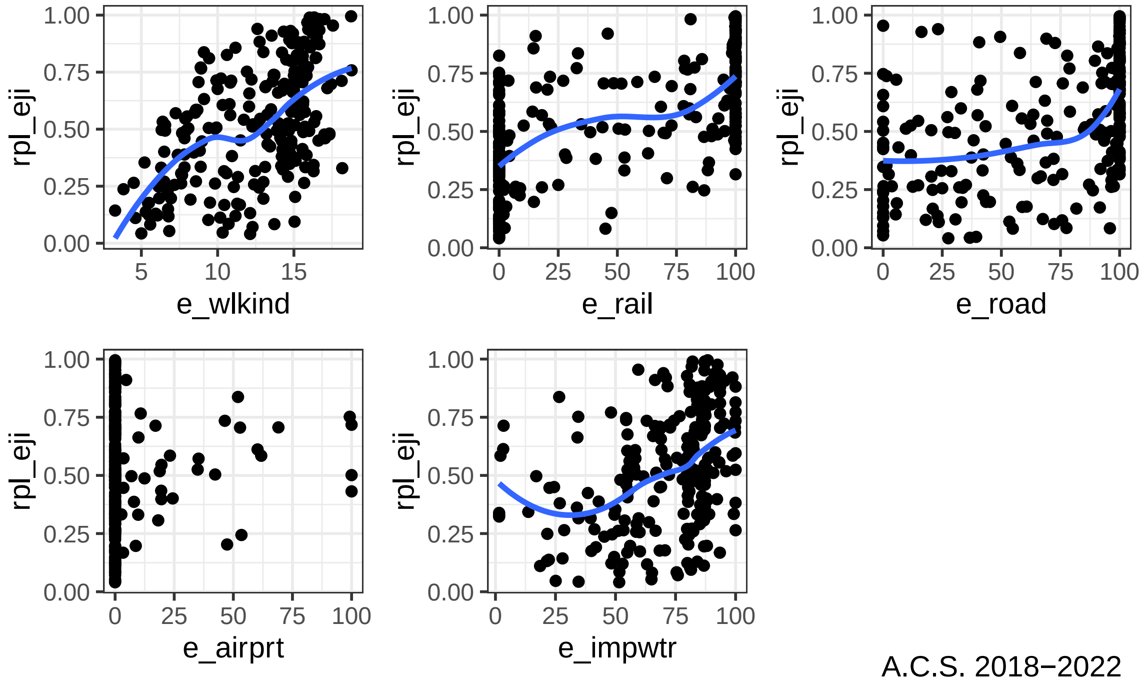


Associations Between Response and Predictor Variables

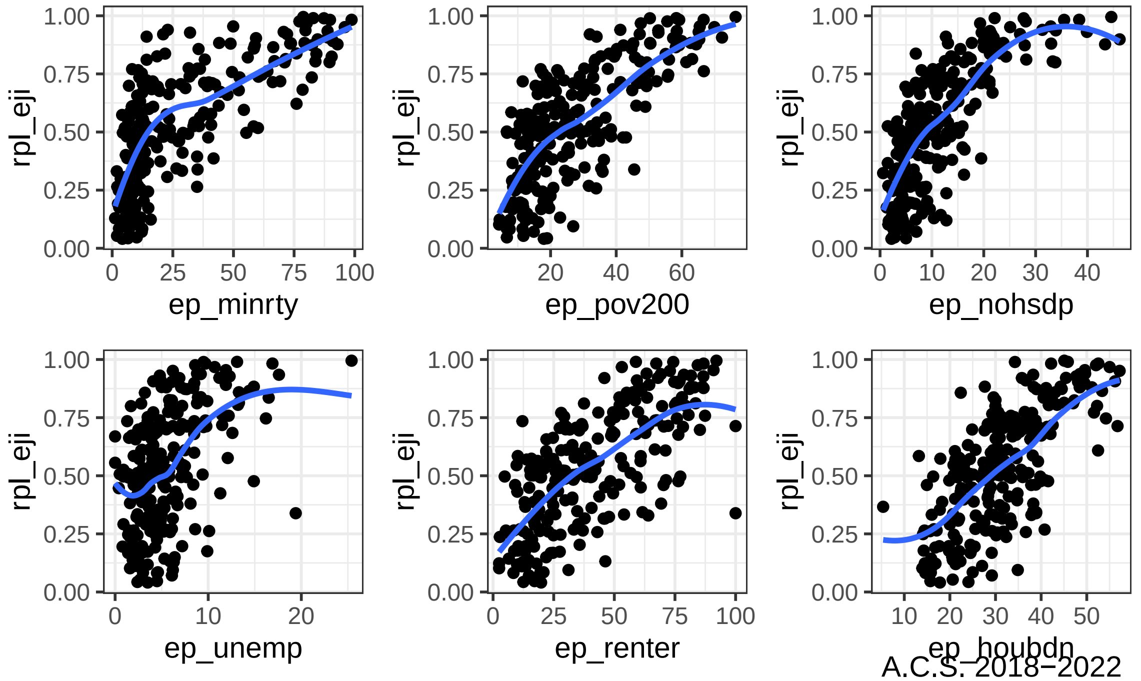
The following section illustrates the need for non-linear regression models. The Social Vulnerability Index covers 4 themes: Socioeconomic Status, Household Characteristics, Racial & Ethnic Minority Status, and Housing Type & Transportation Status.



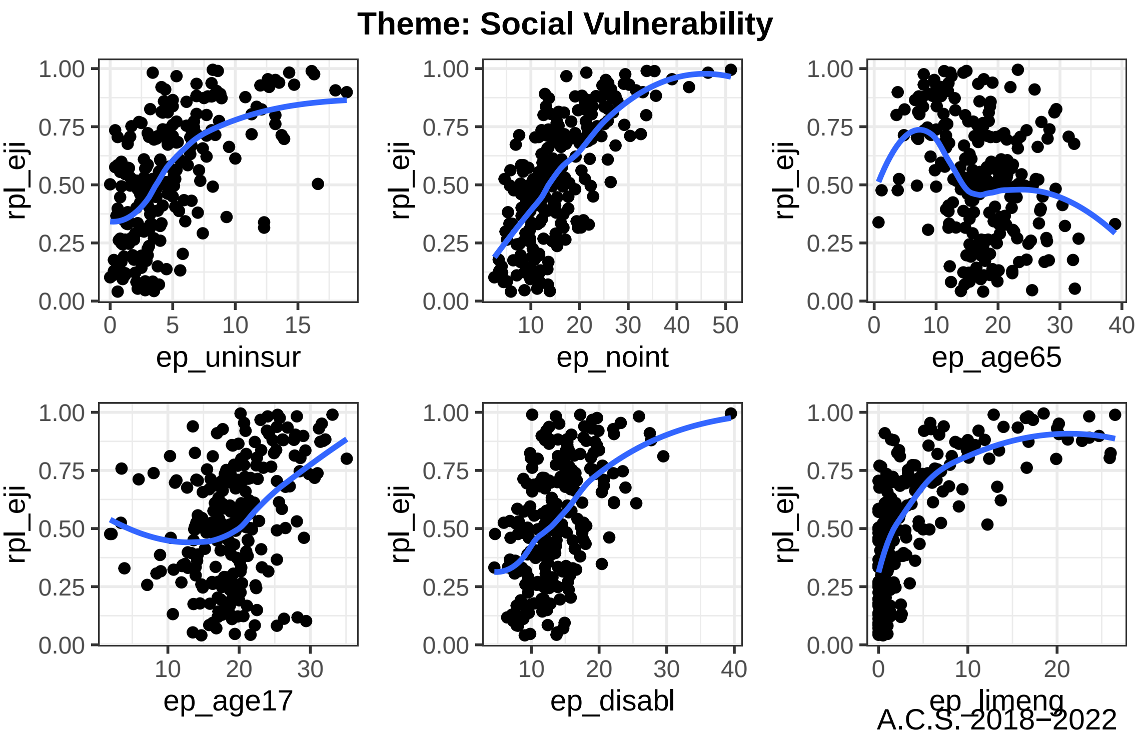



Theme: Enviornmental Burden

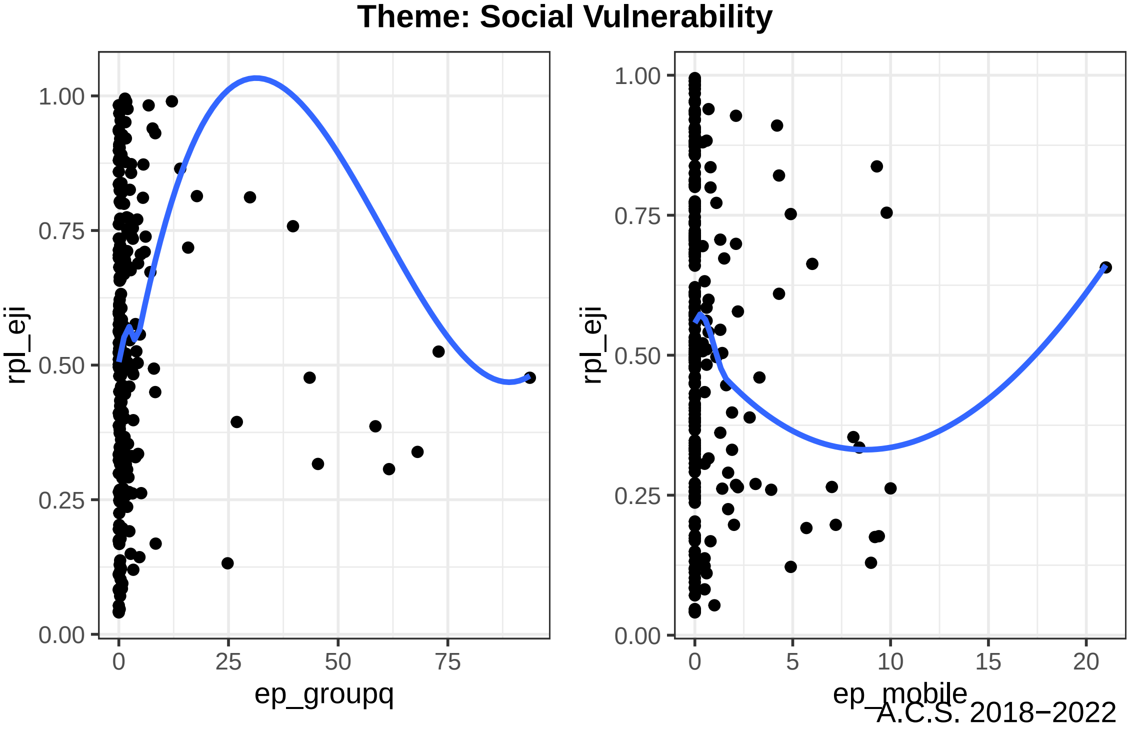
A.C.S. 2018−2022


Theme: Social Vulnerability

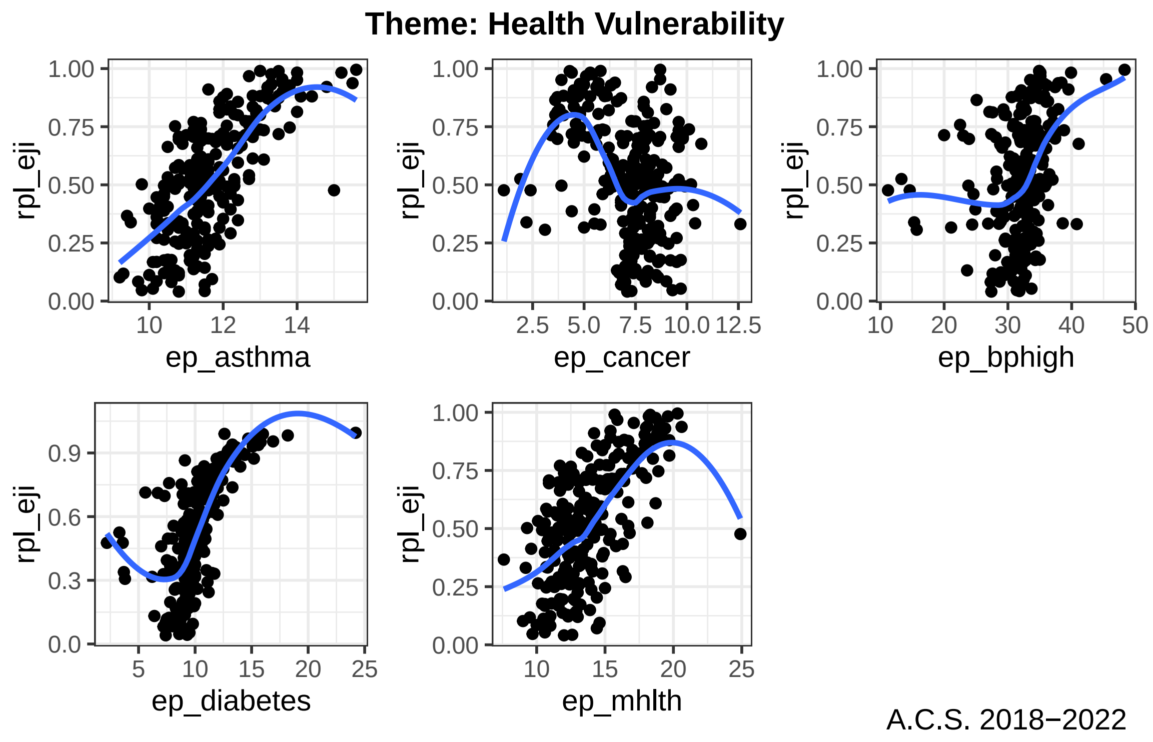
A.C.S. 2018−2022


Theme: Social Vulnerability

A.C.S. 2018−2022


Theme: Social Vulnerability

A.C.S. 2018−2022


Theme: Health Vulnerability
rpl_eji
1.00
0.75
0.50
0.25
0.00
10
12
14
ep_asthma
rpl_eji
1.00
0.75
0.50
0.25
0.00
2.5
5.0
7.5
10.0
12.5
ep_cancer
rpl_eji
1.00
0.75
0.50
0.25
0.00
10
20
30
40
50
ep_bphigh
rpl_eji
0.9
0.6
0.3
0.0
5
10
15
20
25
ep_diabetes
rpl_eji
1.00
0.75
0.50
0.25
0.00
10
15
20
25
ep_mhlth
A.C.S. 2018−2022

Moran's I Calculation and Scatter Plot

At $\alpha = 0.05$, all variables are statistically significant. Only one that had a statistic close to zero,the estimated number of persons in group quarters (`e_groupq`). All other variables had higher Moran's I values that what was foud in the SVI. The variables with the top 3 highest moran's i value include the annual mean days above PM2.5 regulatory standard (3 yr avg) (`e_pm`), the probability of Contracting Cancer (assuming continuous exposure) (`e_totcr`), and the ambient concentrations of diesel (PM/m3) (`e_dslpm`). The summary index value remained higher (`rpl_eji`) than what was found for the SVI summary index value. All other statistically significant variables demonstrated weaker presence of spatial autocorrelation, and measures related to coal and mining did were statistically insignificant.

Additionally, a moran's scatterplot was only preformed for the outcome of interest since creating graphs for all variables would provide redundant information. However, these graphs can be made available upon request. In support with the Moran's I calculation with Monte Carlo Simulations, the scatterplot suggests a positive correlation between the SVI Summary Value and its spatial lag, representative of spatial autocorrelation in the data.

All examined spatial components suggest that a geographically weighted regression model should be implemented.

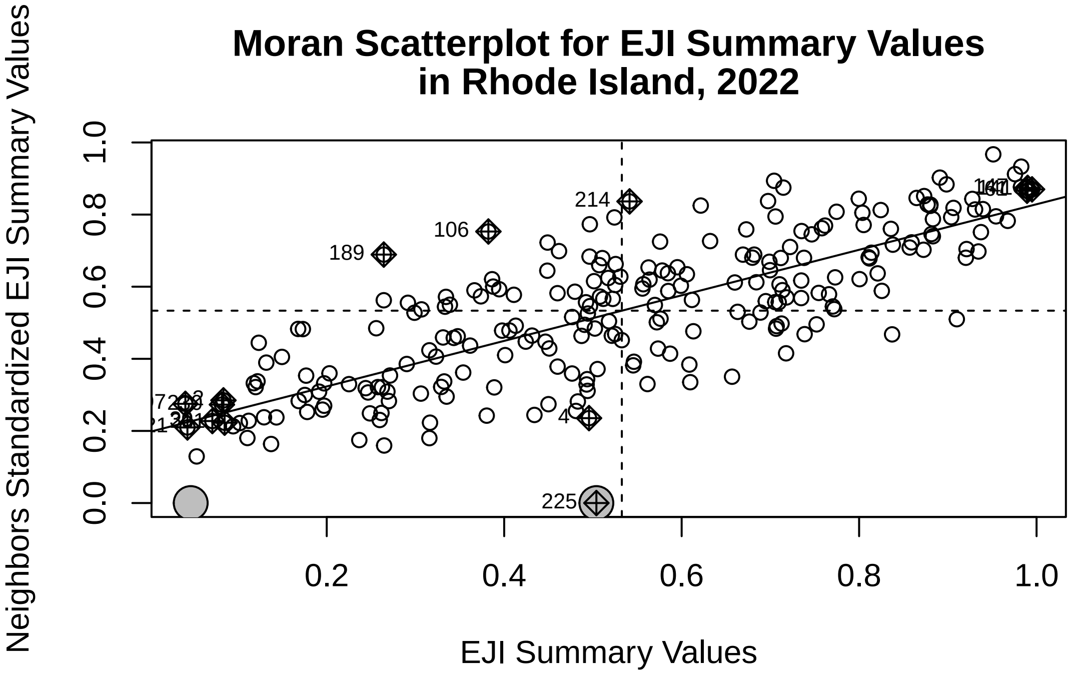
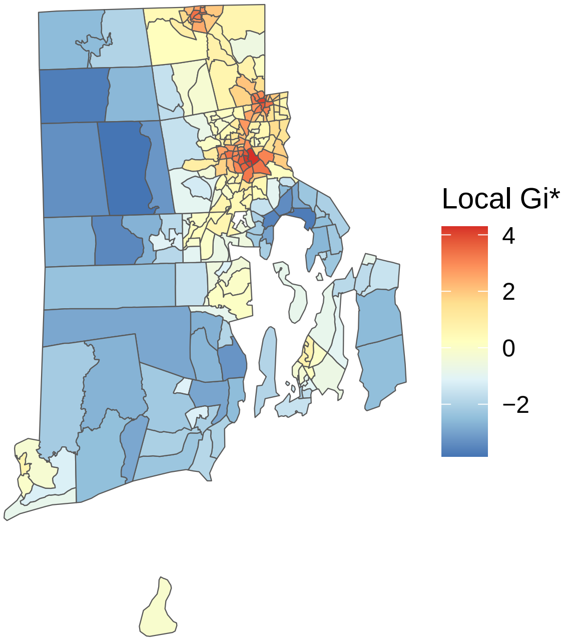


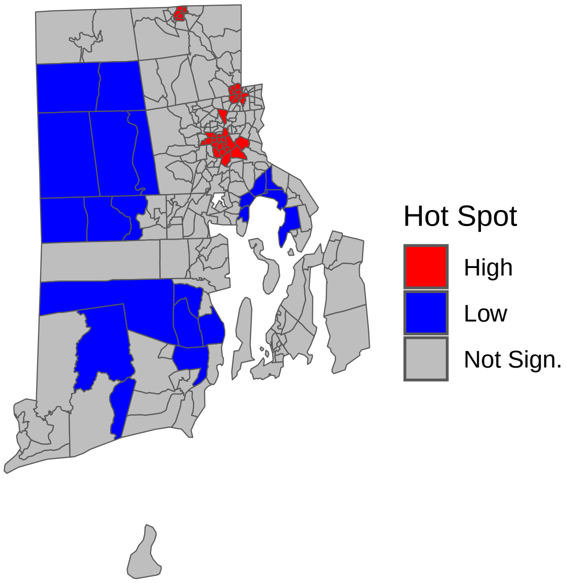
Table 1: Moran's I with Monte Carlo Simulations

variable	statistic	pvalue
e_pm	0.952	0.001
e_totcr	0.908	0.001
e_dslpm	0.877	0.001
e_ozone	0.812	0.001
ep_minrty	0.758	0.001
e_rail	0.755	0.001
e_impwtr	0.738	0.001
ep_limeng	0.707	0.001
e_tri	0.678	0.001
e_wlkind	0.667	0.001
e_park	0.666	0.001
rpl_eji	0.614	0.001
e_road	0.591	0.001
ep_pov200	0.591	0.001
ep_renter	0.591	0.001
e_houage	0.577	0.001
ep_nohsdp	0.553	0.001
ep_asthma	0.532	0.001
e_tsd	0.521	0.001
e_airprt	0.494	0.001
e_rmp	0.481	0.001
ep_uninsur	0.464	0.001
ep_mhlth	0.457	0.001
ep_diabetes	0.441	0.001
ep_houbdn	0.430	0.001
ep_cancer	0.429	0.001
ep_noint	0.386	0.001
ep_age65	0.312	0.001
ep_age17	0.307	0.001
e_npl	0.305	0.001
ep_bphigh	0.280	0.001
ep_unemp	0.233	0.001
ep_disabl	0.220	0.001
ep_mobile	0.176	0.001
ep_groupq	0.081	0.031
e_coal	NaN	0.001
e_lead	NaN	0.001

Local Spatial Autocorrelation with G_i^* & Hot/Cold Spot Identification

Both maps illustrate that Providence, Pawtucket, and Woonsocket demonstrate high values of Social Vulnerability.


Local Gi*
4
2
0
−2


Hot Spot
High
Low
Not Sign.