Unveiling Vulnerability: Exploratory Data Analysis for the Social Vulnerability 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).

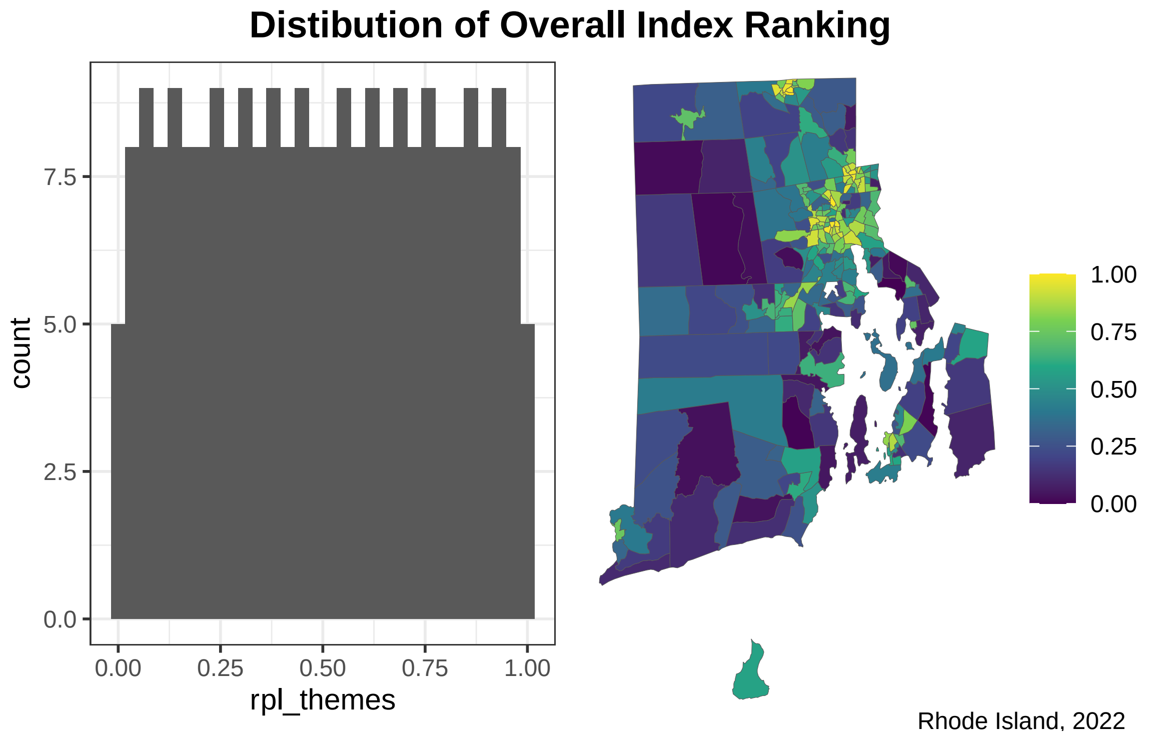
The EDA includes: distribution of variables, correlation matrix, and LOESS fitted scatter plots. Additionally, the examination of the spatial components include: Moran's I calculation with Monte Carlos Simulation, Moran's Scatterplot, Local Spatial Autocorrelation with Getis-Ord local G_i^*, and Hot/Cold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

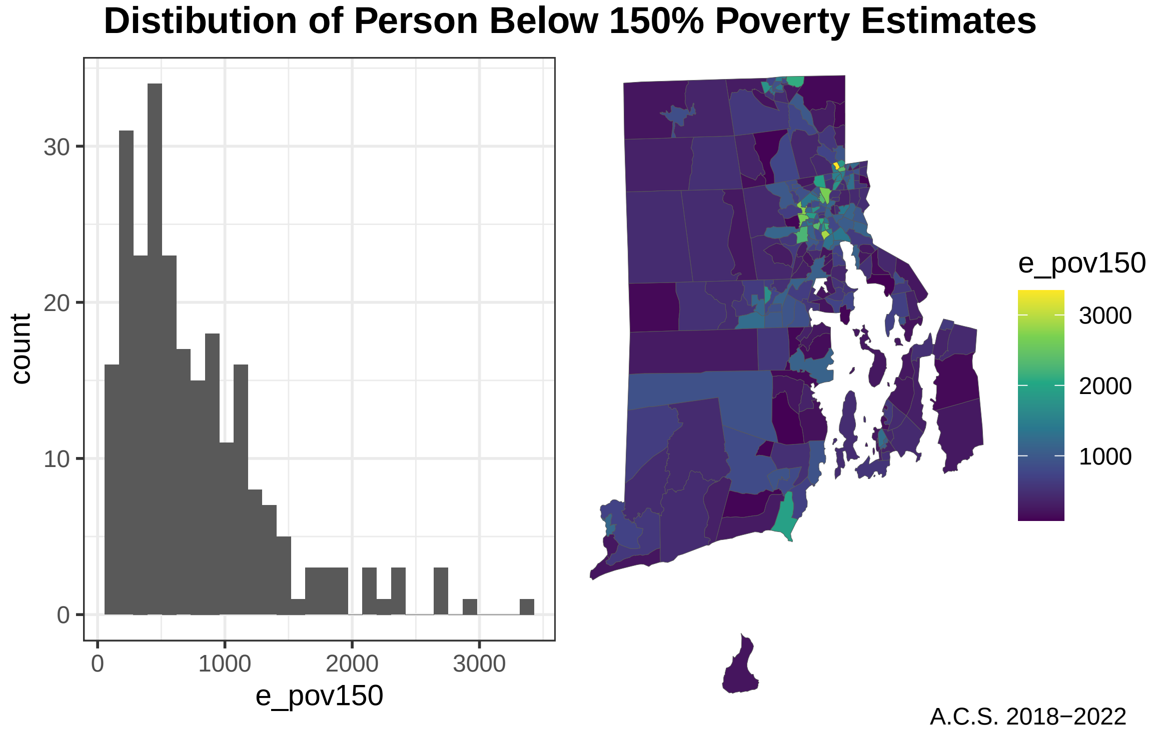
Definitions:

- **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.

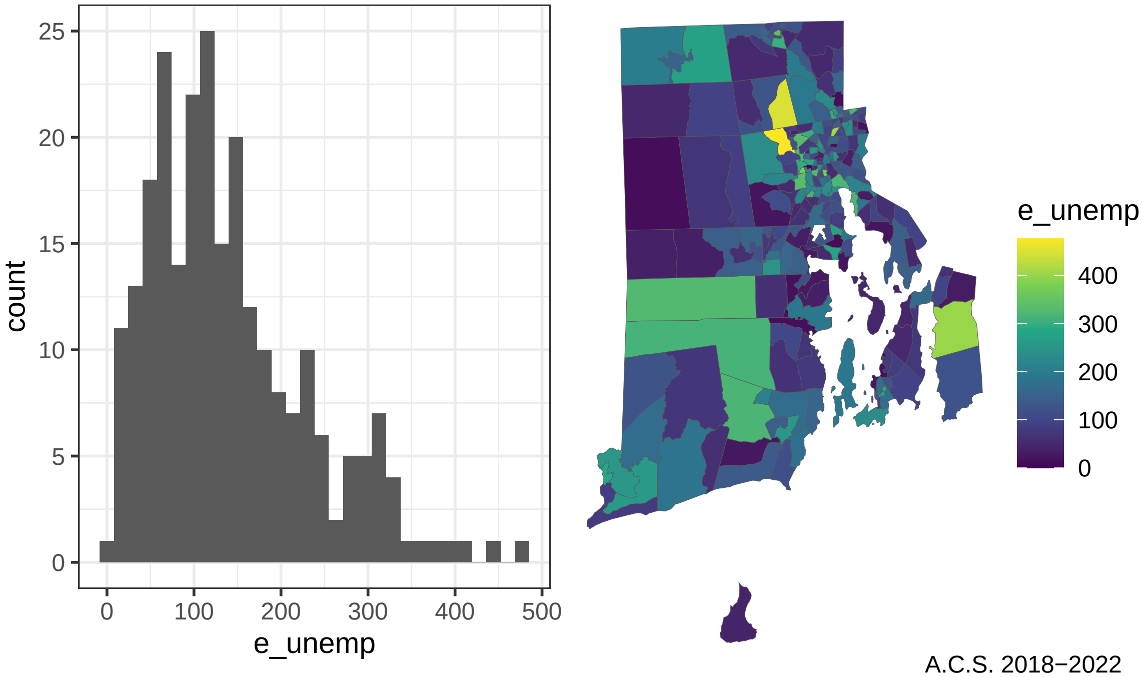
Distribution of Variables

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



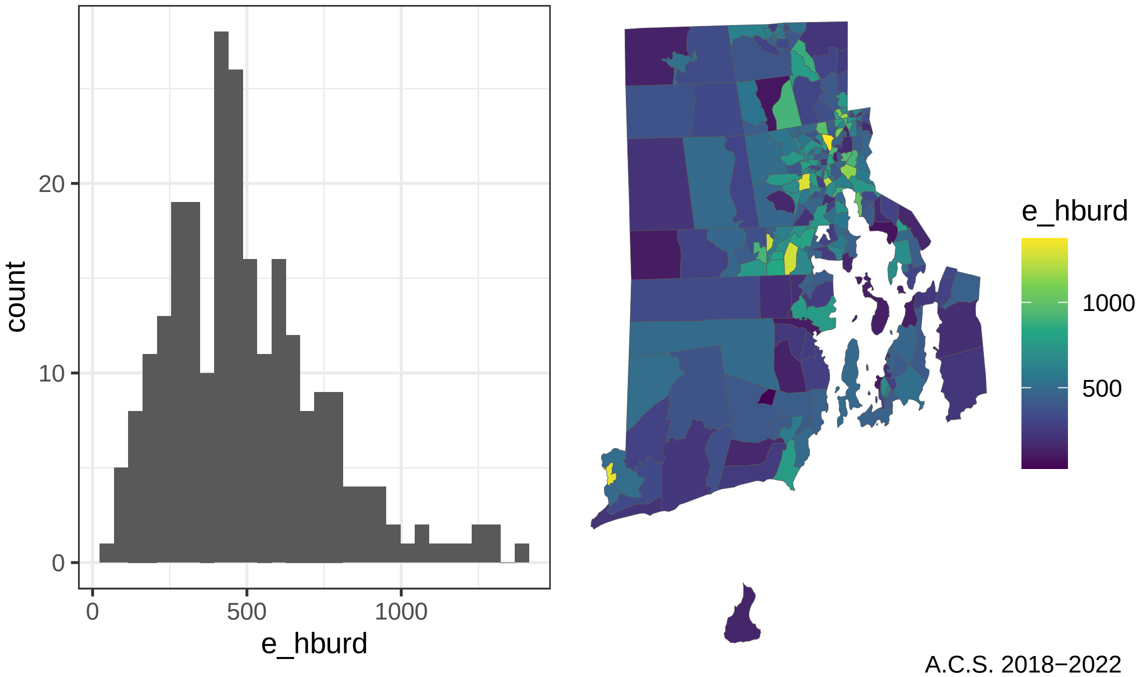


Distibution of Civilian (16 yrs+) Unemployed Estimates



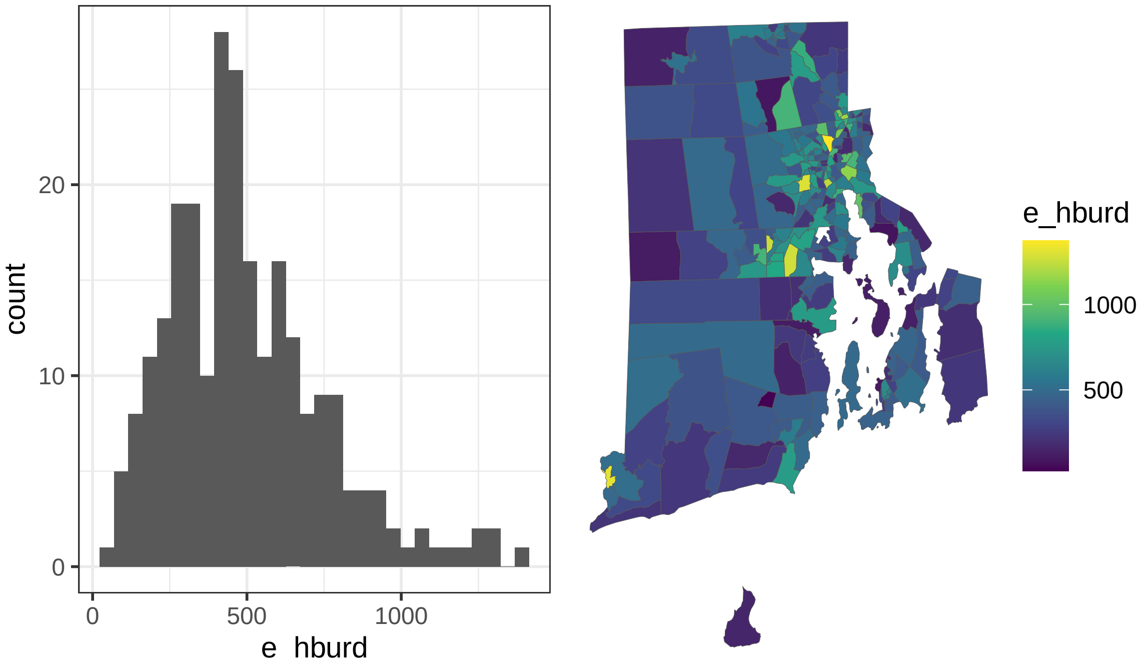
A.C.S. 2018−2022

Distibution of Cost−Burdened Occupied Housing United (< $75k)



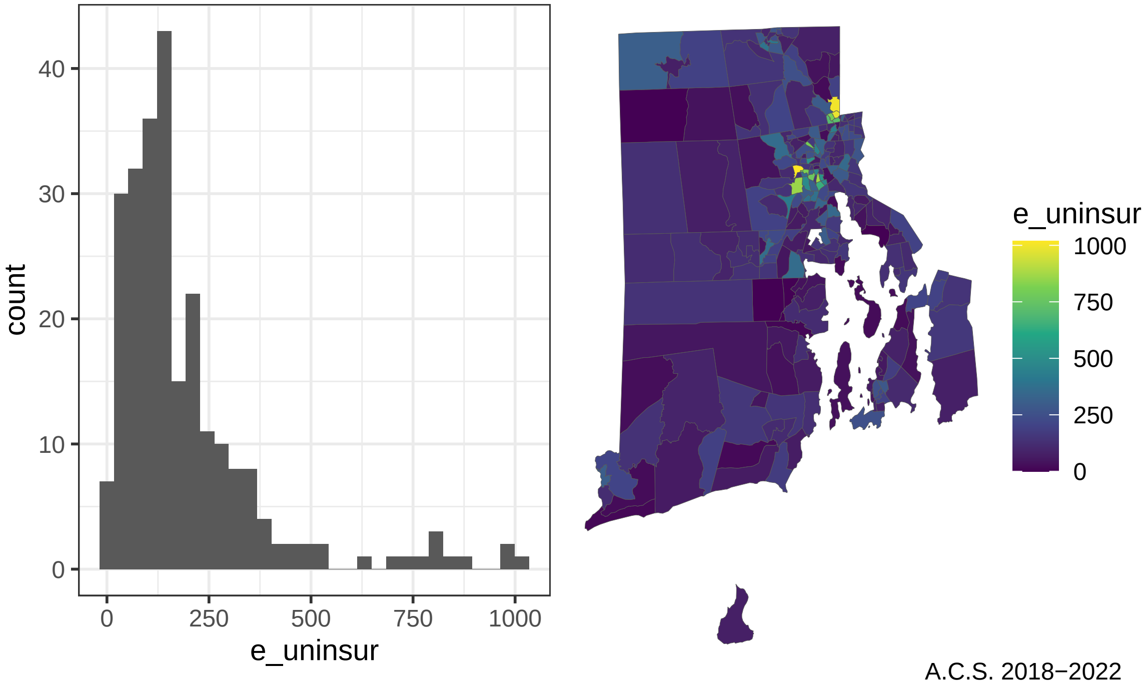
A.C.S. 2018−2022

Distibution of Persons (25 yrs+) With No High School Diploma



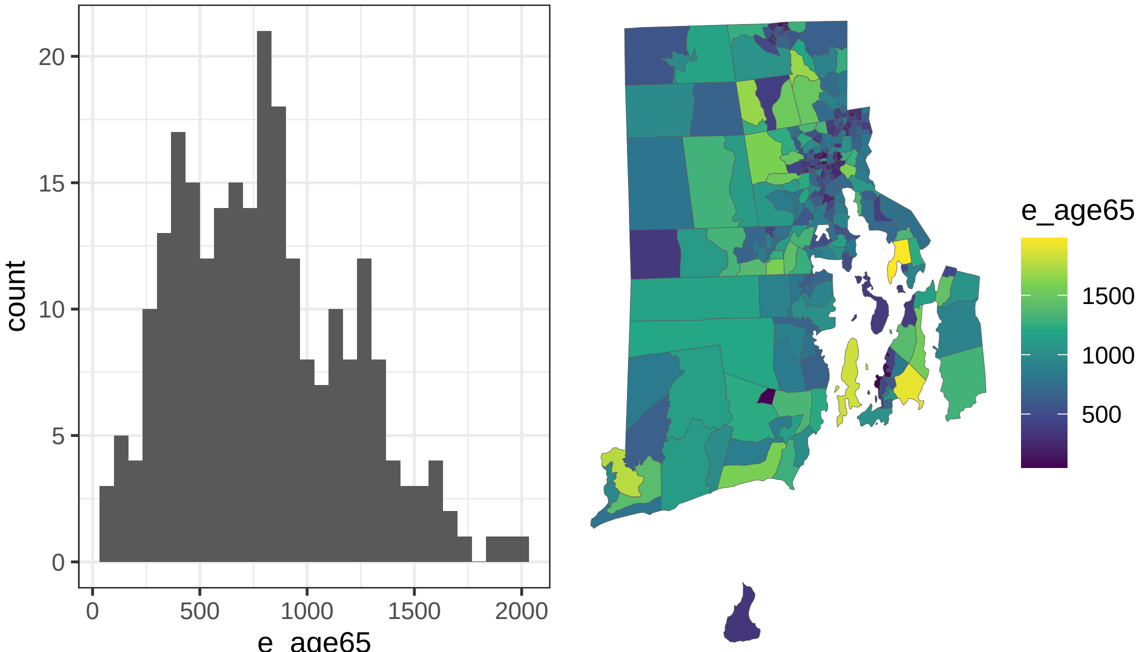
A.C.S. 2018−2022

Distibution of Uninsured in Total Civilian Non−Institutionalized Population Estimate



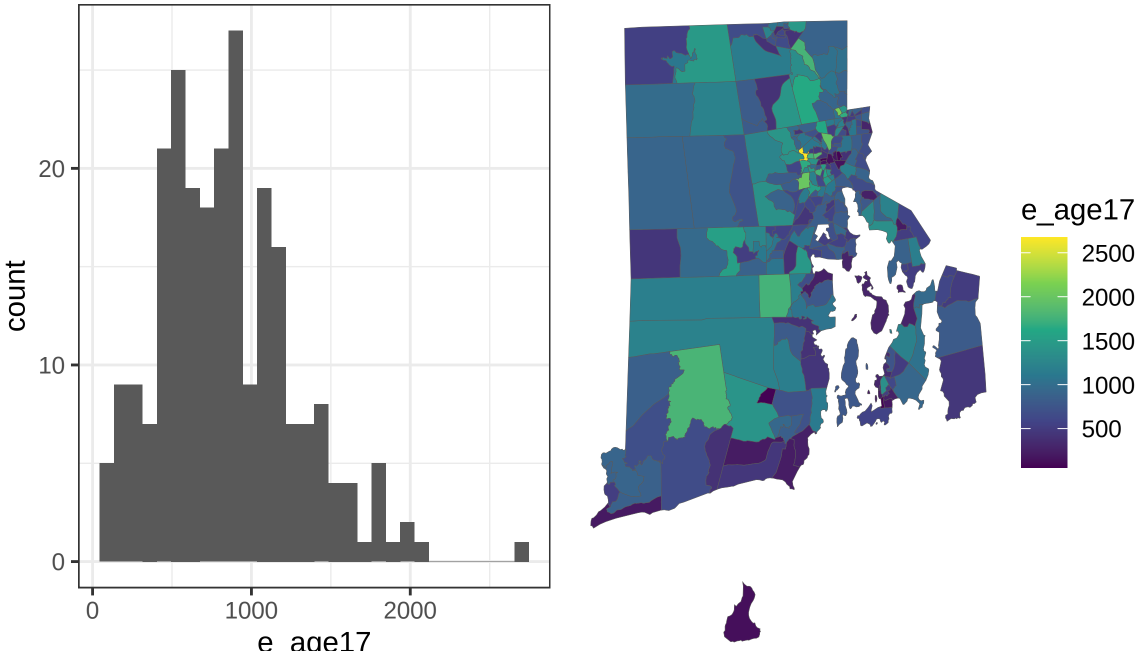
A.C.S. 2018−2022

Distibution of Persons Aged 65 or Older Estimate



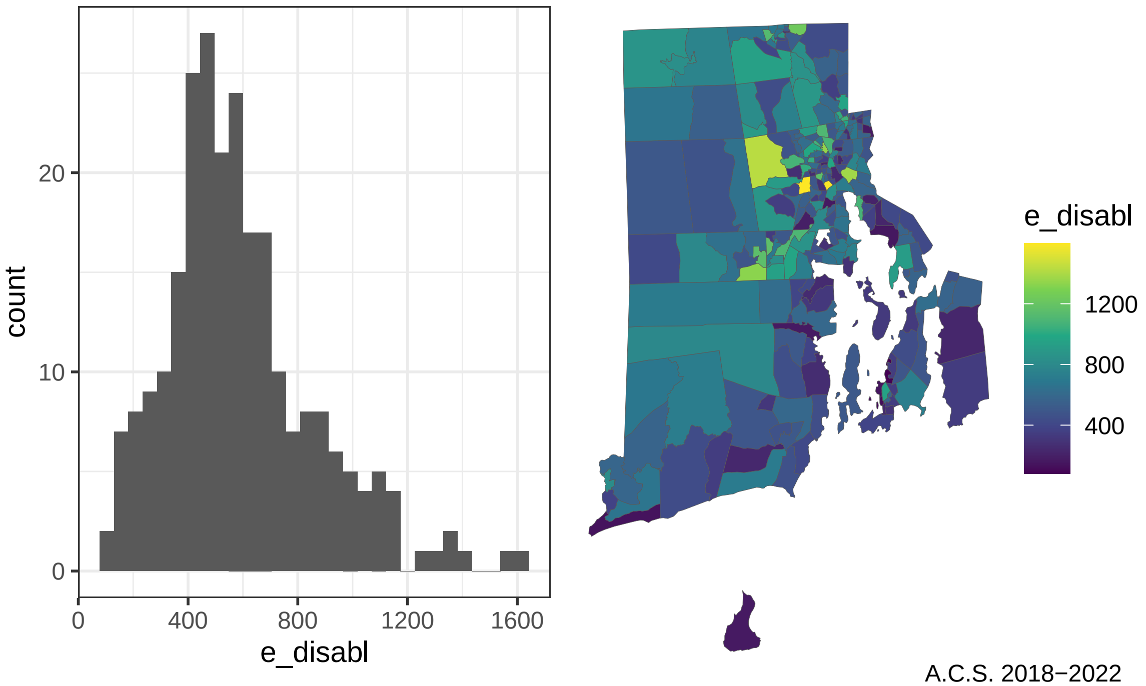
A.C.S. 2018−2022

Distibution of Persons Aged 17 or Younger Estimate



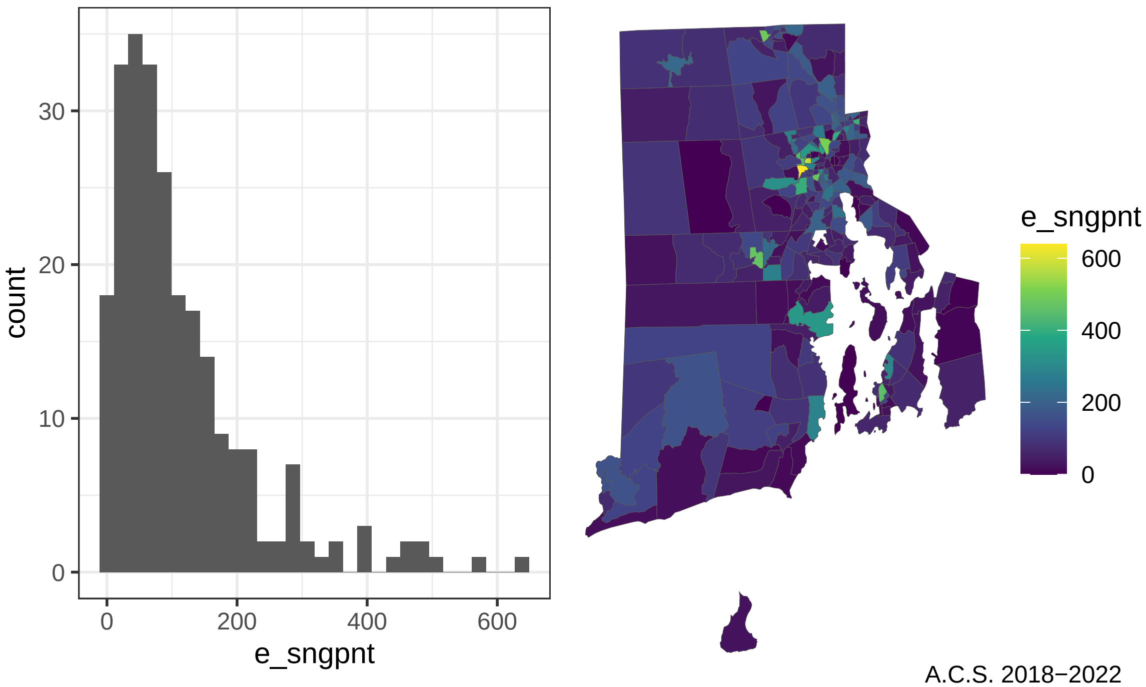
A.C.S. 2018−2022

Distibution of Total Civilian Non−Institutionalized Population with a Disability Estimate



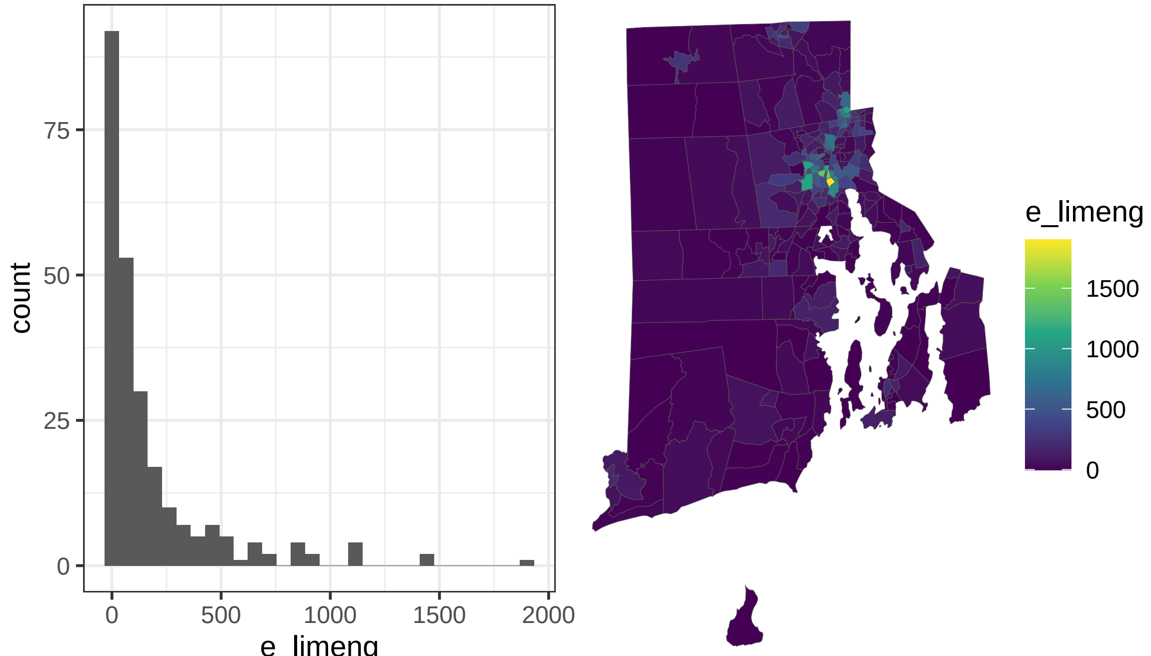
A.C.S. 2018−2022

Distribution of Single Parent Households with Children Less than 18 Estimate



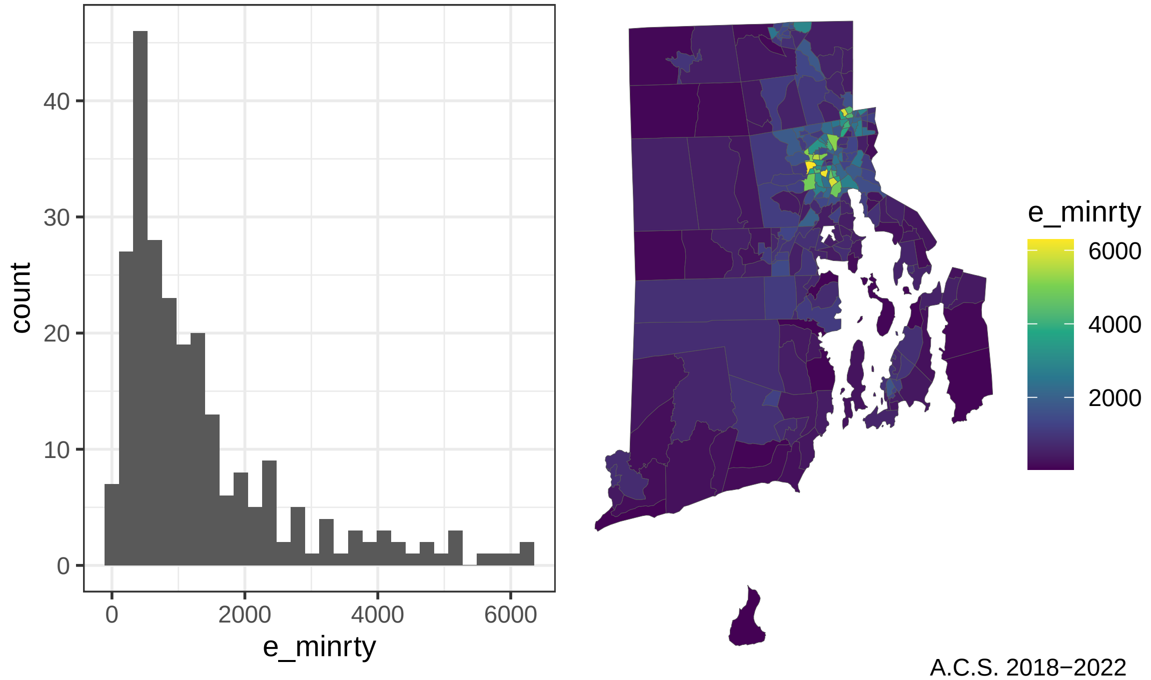
A.C.S. 2018−2022

Distribution of Person (5 yrs+) Who Speak English 'less than well' Estiamte



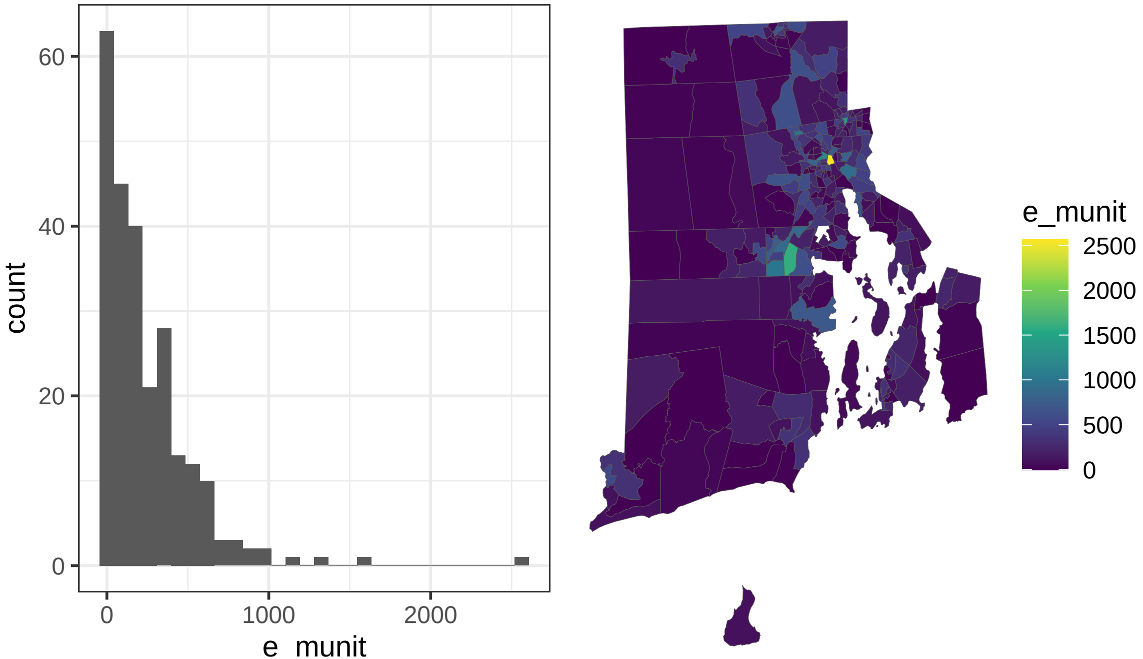
A.C.S. 2018−2022

Distribution of Minority Persons Estimate



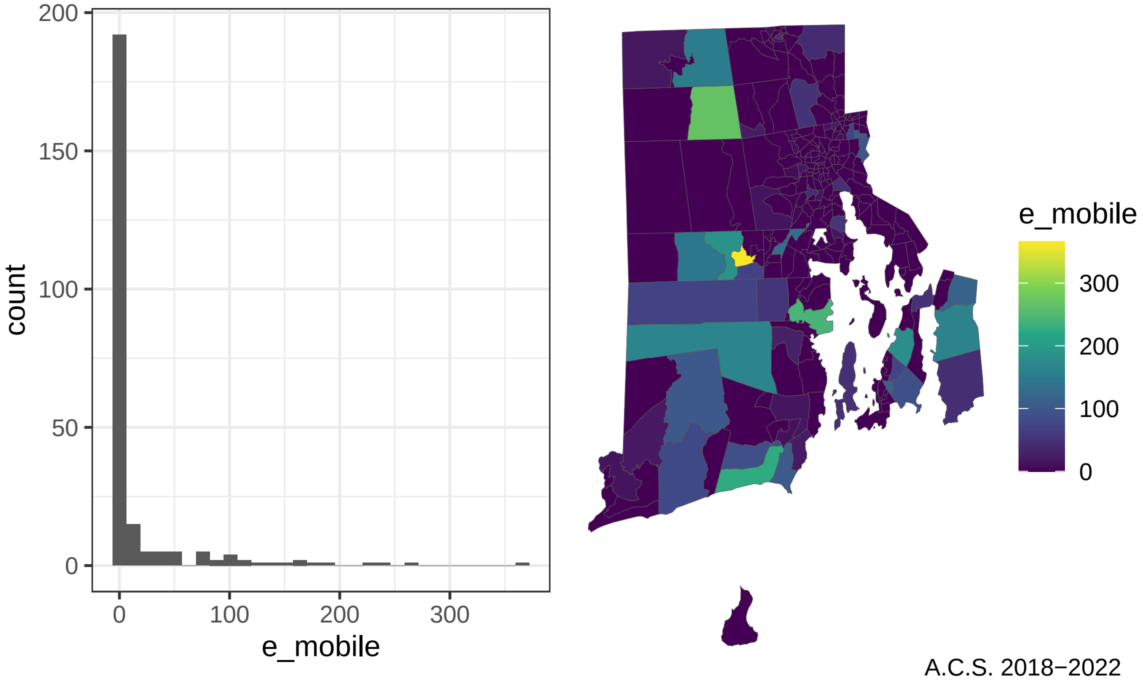
A.C.S. 2018−2022

Distribution of Housing in Structures with 10 or More Units Estimate



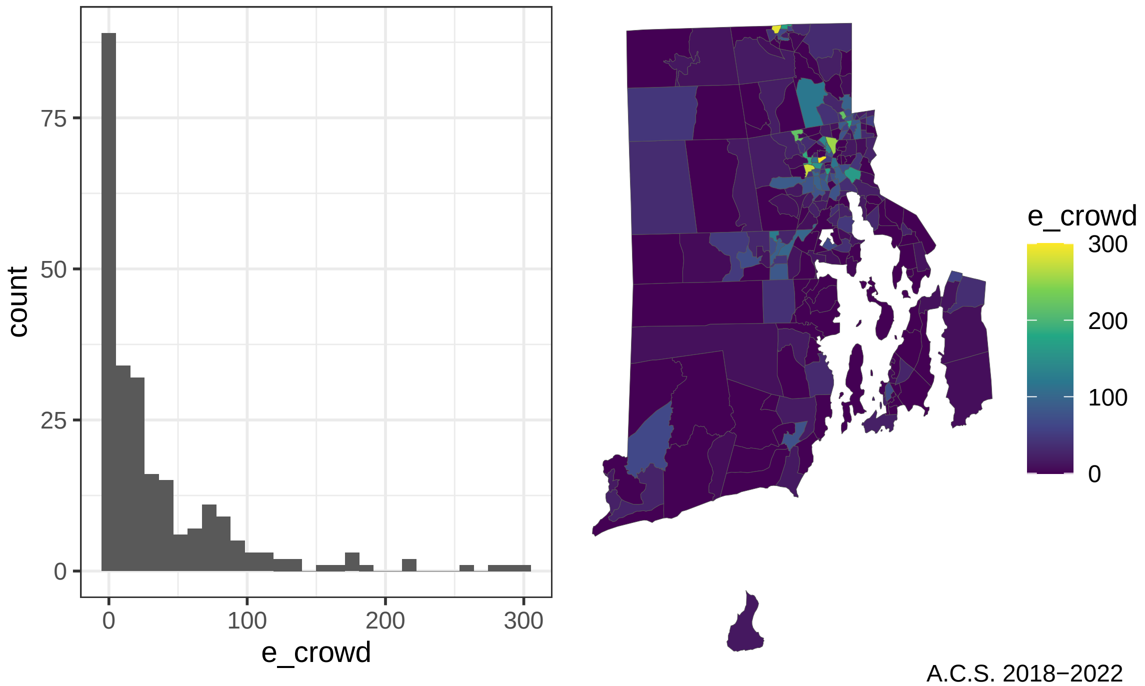
A.C.S. 2018−2022

Distribution of Mobile Homes Estimate



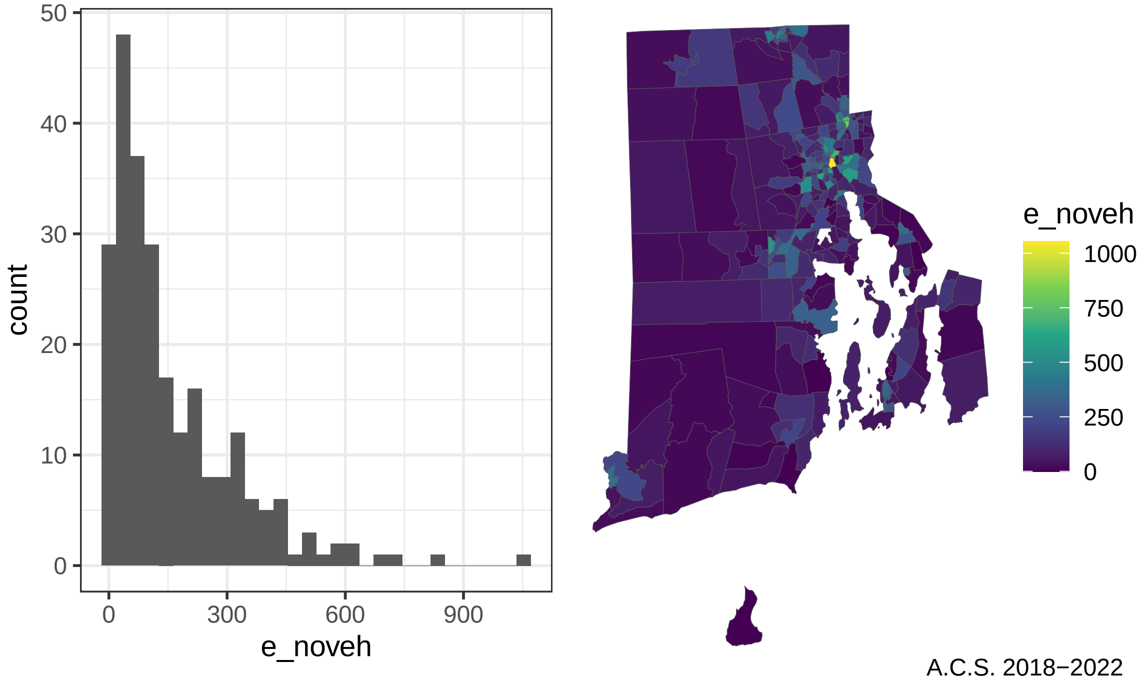
A.C.S. 2018−2022

Distribution of Occupied Housing Units With More People Than Rooms Estimate

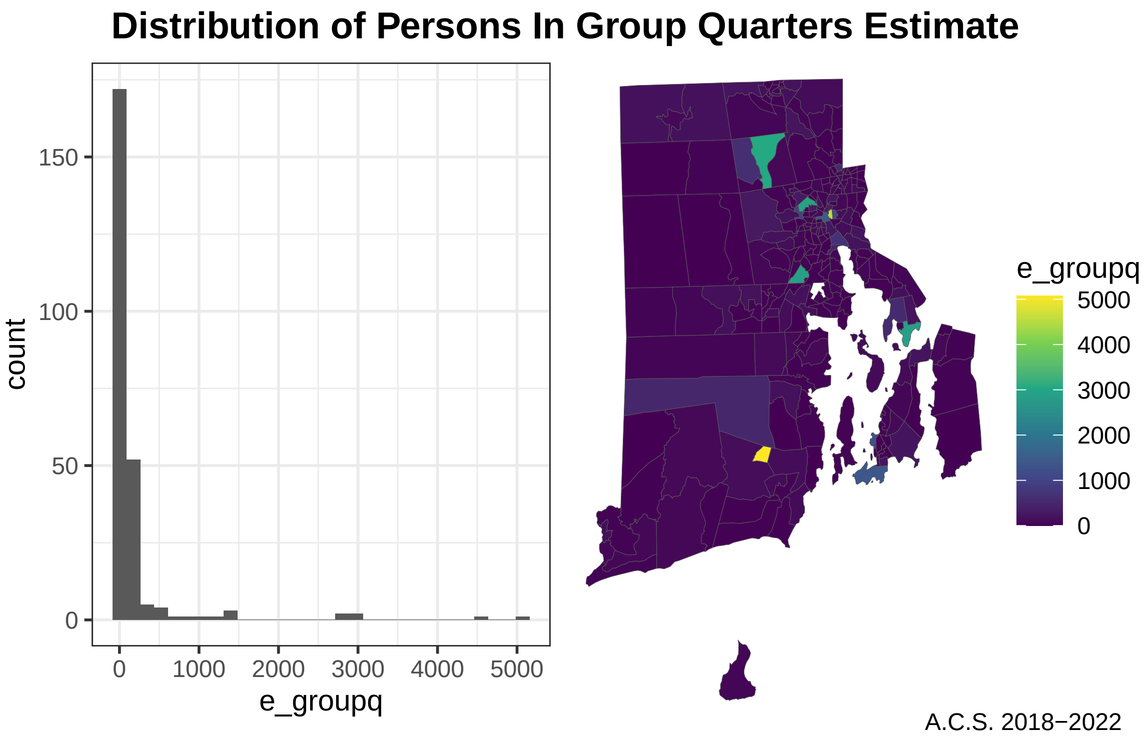


A.C.S. 2018−2022

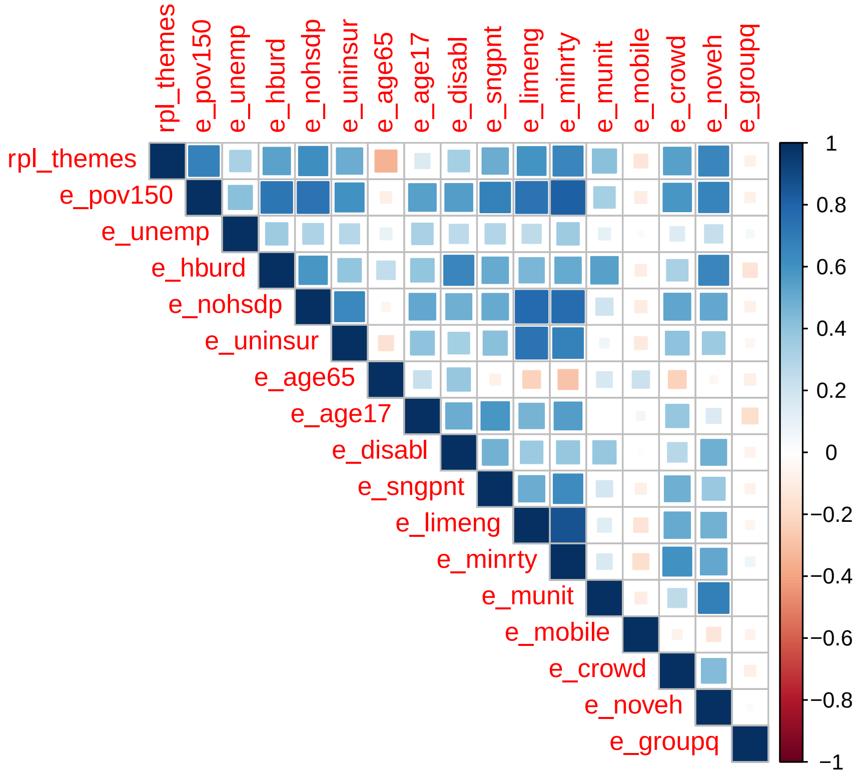
Distribution of Households With No Vehicles Avaiable Estimate



A.C.S. 2018−2022



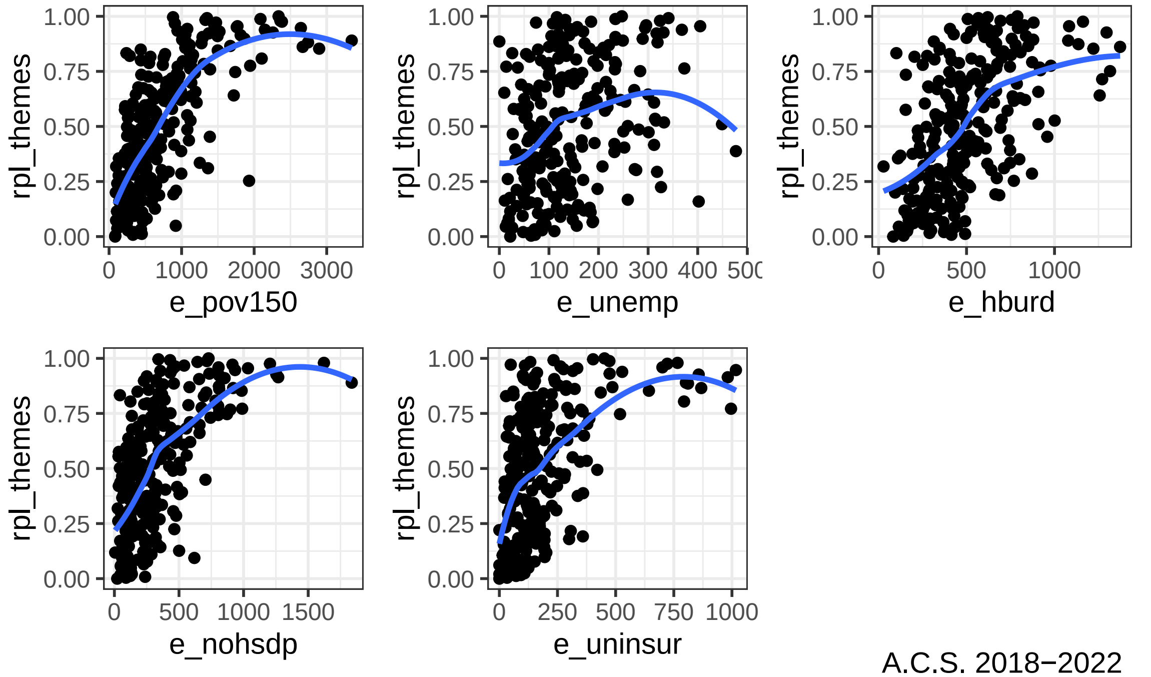
Correlation Matrix



Associations Between Response and Predictor Variables

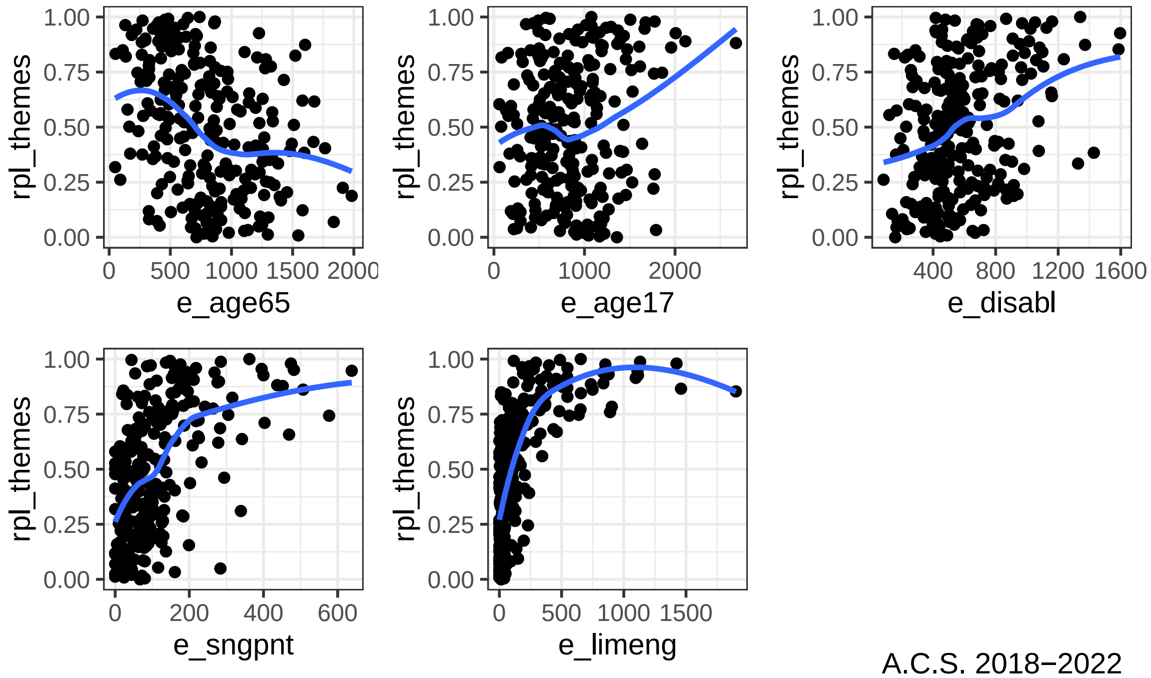
The following section illustrates the need for non-linear regression models, with regression lines being fitted with LOESS. The Social Vulnerability Index covers 4 themes: Socioeconomic Status, Household Characteristics, Racial & Ethnic Minority Status, and Housing Type & Transportation Status.

Theme: Socioeconomic Status

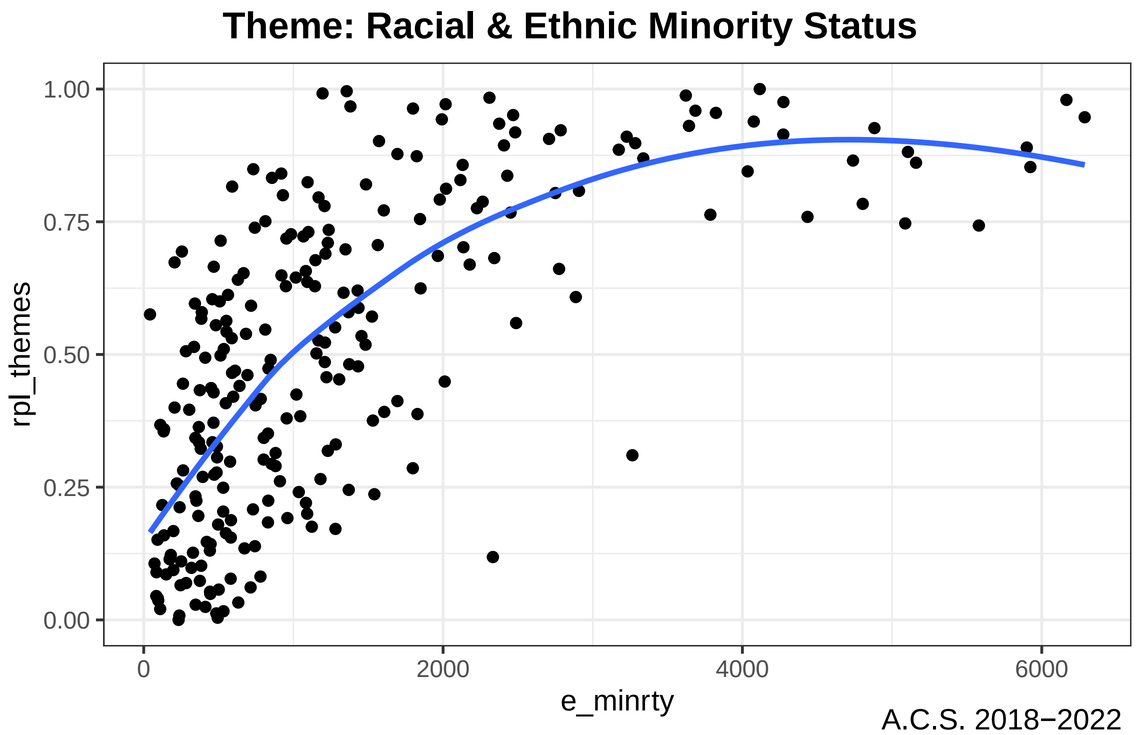


A.C.S. 2018−2022

Theme: Household Characteristics

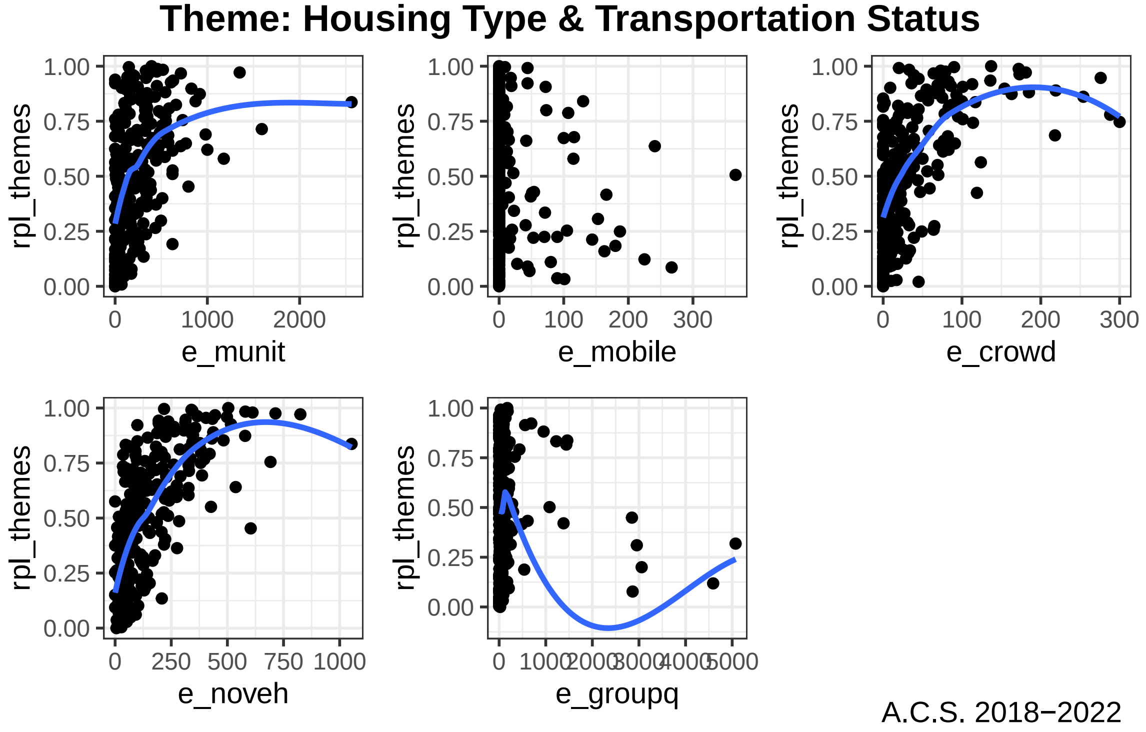


A.C.S. 2018−2022



Theme: Racial & Ethnic Minority Status

A.C.S. 2018−2022



Theme: Housing Type & Transportation Status

A.C.S. 2018−2022

Moran's I Calculation and Scatter Plot

At $\alpha = 0.05$, all variables are statistically significant. Variables that had a statistic close to zero include the estimated number of unemployed individuals (16 yrs +) (`e_unemp`) and the estimated number of persons in group quarters (`e_groupq`). Meanwhile, the variables with the top 3 highest moran's i value include the number of minorities (`e_minrty`), the number of individuals who self report having limited english (`e_limeng`), and the SVI Summary Values (`rpl_themes`). All other statistically significant variables demonstrated weaker presence of spatial autocorrelation, and the only variable to be statistically insignificant (i.e., demonstrate no spatial autocorrelation) is the proportion of individuals who are unemployed (`e_unemp`).

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.

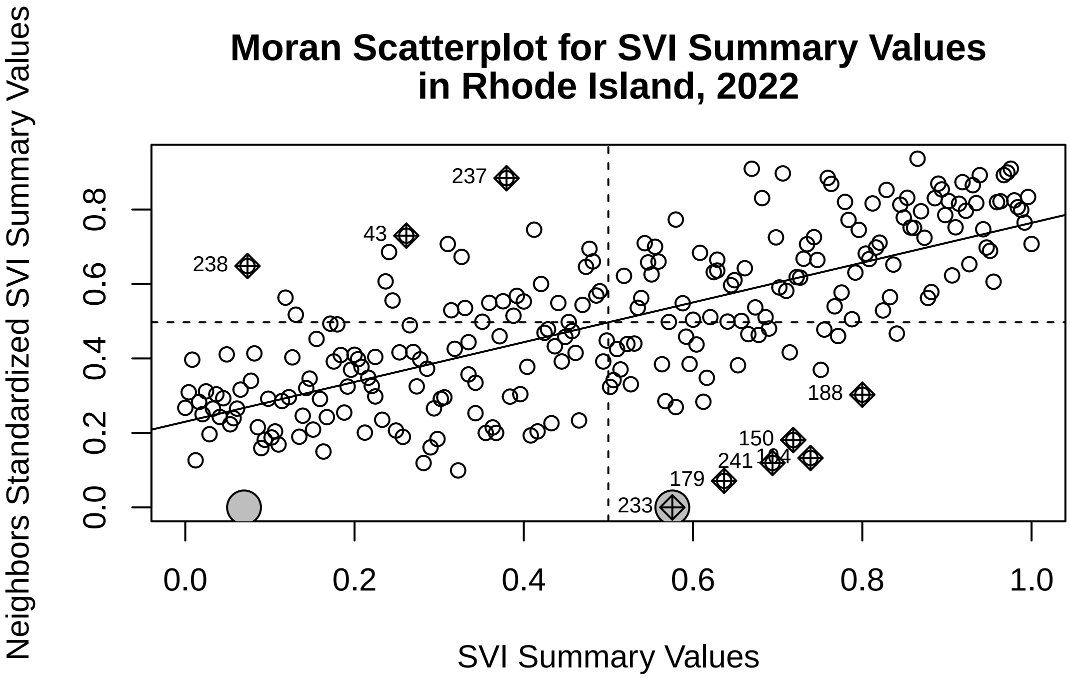
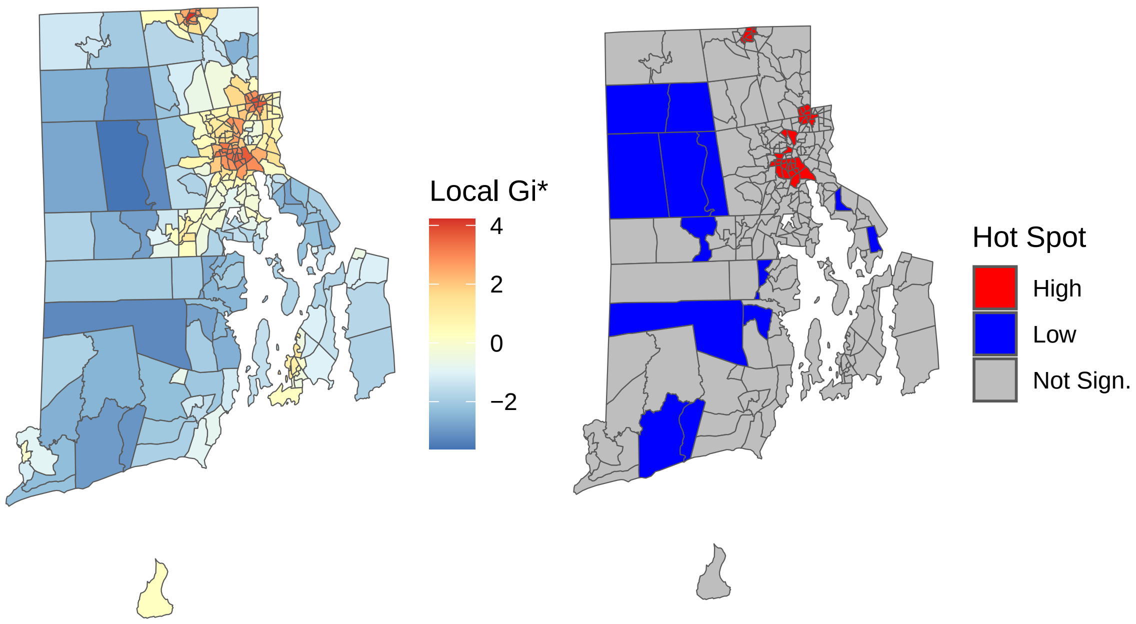


Table 1: Moran's I with Monte Carlo Simulations (nsim = 999)

variable	statistic	pvalue
e_minrty	0.586	0.001
e_limeng	0.579	0.001
rpl_themes	0.525	0.001
e_nohsdp	0.400	0.001
e_pov150	0.346	0.001
e_uninsur	0.316	0.001
e_noveh	0.287	0.001
e_age65	0.279	0.001
e_crowd	0.228	0.001
e_age17	0.203	0.001
e_hburd	0.193	0.001
e_munit	0.180	0.001
e_mobile	0.156	0.001
e_sngpnt	0.142	0.001
e_disabl	0.126	0.001
e_groupq	0.069	0.051
e_unemp	0.057	0.064

Local Spatial Autocorrelation with G_i^* & Hot 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.