

# **Exploring the relationship between housing cost burden and measures of social vulnerability**

**Ariel Anderson**

## **Introduction**

This project examines the relationship between housing cost and two measures of social vulnerability from the Center for Disease Control Social Vulnerability Index (SVI). Measures of SVI are used to determine a population's risk in the event of an emergency disaster. The purpose of this project is to explore potential variables that may contribute to housing cost burden by examining measures of social vulnerability in the state of Maine. Examining the relationship between housing cost burden and social vulnerability may help to guide better economic policy to help individuals secure affordable housing. Since the pandemic, Maine has seen a 39% increase in rental costs, severely impacting low-income individuals (Olsen, 2022).

Ordinary Least Squares will be used to determine the relationship between housing cost burden and two measures of social vulnerability, and to determine possible spatial autocorrelation using Moran's I analysis. Housing cost burden is defined as households that make less than \$70,000 a year, and spend at least 30% of household income on housing. The two covariates explored in this project, or explanatory variables, are number of households living up to 150% of the poverty line, and number of individuals living with a documented disability. Because higher SVI measures contribute to a population's risk (CDC, 2022), it is hypothesized that housing burden increases with higher measures of social vulnerability and that positive spatial autocorrelation is present.

## **Methods**

Spatial data was obtained from the CDC Social Vulnerability database for Maine census tracts. The independent variable, or response variable, for each part of the analysis is number of households with housing cost burden. As a preliminary step, the explanatory and response variable will be mapped in R studio using six natural jenks to give a visual impression of the covariates and response variable. The two covariates explored in this project, or explanatory variables, are number of households living up to 150% of the poverty line, and number of individuals living with a documented disability.

Linear regression will be calculated for each covariate individually. Then, Ordinary Least Squares (OLS) will examine the relationship between housing cost burden and the two covariates. Exploration of individual linear regression prior to implementation of OLS follows the methods outlined in Olakunde's spatial analysis of HIV infection in at risk populations (Olakunde, et al., 2020).

This project uses a spatial regression model using R Studio similar to a study at Utah State University that examines park quality as a response variable to socioeconomic status covariates. Moran's I analysis of the residuals will be computed based on queen's case spatial weight matrix (Chen, et al. 2019). Additionally, the GeoDa workbook suggests that queen's case weights have more neighbors than the rook's case. Therefore, a rook's case weight matrix will have neighboring observations with more influence. Based on previous studies (Chen, et al. 2019), and the GeoDa workbook, a queen's case weight matrix was determined. This analysis will be conducted using both GeoDa and R Studio.

If spatial autocorrelation is found for housing cost burden, then the OLS model may not be suitable. Lagrange Multiplier Statistics will be used to determine if a Spatial Lag or Spatial Error model will be necessary (Chen, et al. 2019). The AIC, BIC, and Loglikelihood values will be used to compare models. In general, a lower AIC and higher Loglikelihood is associated with a better fit model (Plant, 2019, p. 260).

## Results

Preliminary maps were made using R studio with 6 natural jenks to give a visual impression of distribution of housing cost burden, rates of poverty, and rates of disability in the state of Maine. This is depicted in figure 1.

Figure 1.



Figure 1. Top Left: Map of Individuals with Documented Disability, Top Right: Map of Households Living at 150% or Less of the Poverty Line, Bottom: Number of Households with Housing Cost Burden

Linear regression of housing burden as the response variable to explanatory variable rate of poverty had a y-intercept estimated at 155.69, and poverty coefficient of 0.32, resulting in a linear equation of:

$$\text{Housing Burden} = 155.69 + 0.32(\text{Rate of Poverty})$$

The residual standard error of 142.8 implies an average -error of 142.8. The R Squared value of 0.4066 indicates 40.66% of the variance in data can be attributed to the linear regression model, with an adjusted R square value of 0.4051 or 40.51%. The p-value of 2e-16 indicates statistical significance.

Linear regression of housing burden as the response variable to explanatory variable rate of disability had a y-intercept estimated at 110.66, and disability coefficient of 0.47, resulting in a positive correlation of:

$$\text{Housing Burden} = 110.6 + 0.47(\text{Rate of Disability})$$

The residual standard error of 145.6 implies an average -error of 145.6. The R Squared value of 0.3828 indicates 38.28% of the variance in data can be attributed to the linear regression model, with an adjusted R square value of 38.12 or 38.12%. The p-value of 2e-16 indicates statistical significance.

Figure 2.

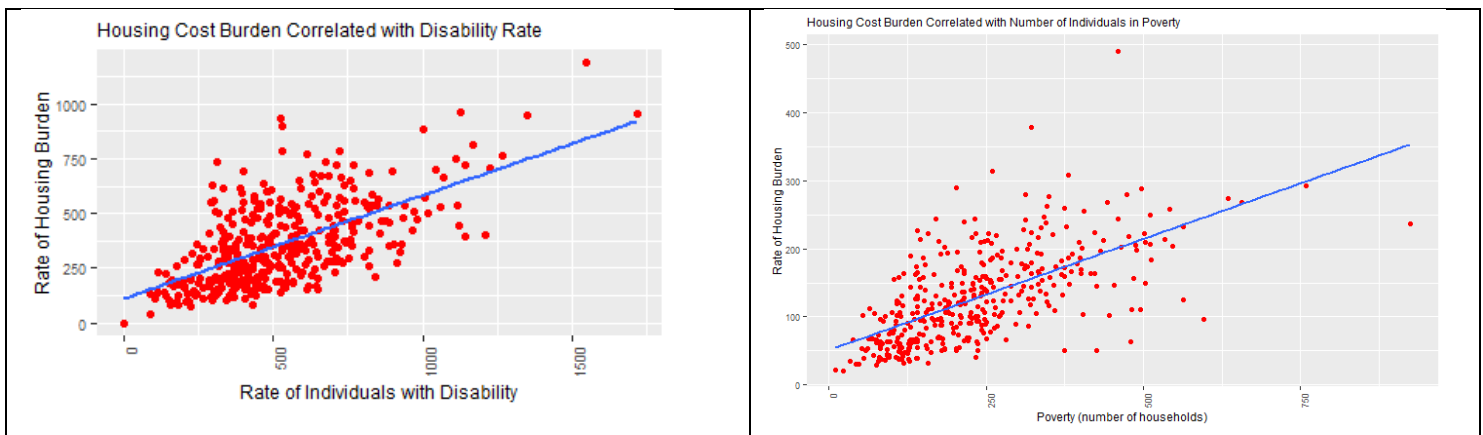


Figure 2. The linear scatter plots shows rate of housing burden as the response and rates of poverty and disability as the explanatory variables.

The summary of the OLS model indicates an intercept of 90.04, and the coefficient of poverty as 0.21 and the coefficient of disability as 0.27, resulting in a positive correlation equation of:

$$\text{Housing Burden} = 90.04 + 0.21(\text{Poverty}) + 0.27(\text{Disability})$$

The equation implies that housing burden increases as rates of poverty and disability increase. The p value for y intercept was 5.06e08, the p value for the poverty coefficient was 1.49e-15, and the p value for the disability coefficient was 4.24e-12, implying statistical significance. The residual standard error of 134.6 implies that the OLS regression model predicts violent crime with an average error of 134.6. The R Squared value of 0.4741 indicates 47.41% of the variance in data can be attributed to the OLS linear regression model, with an adjusted R square value of 0.4715 or 47.15%. The p-value of 2.2e-16 indicates statistical significance.

Figure 3.

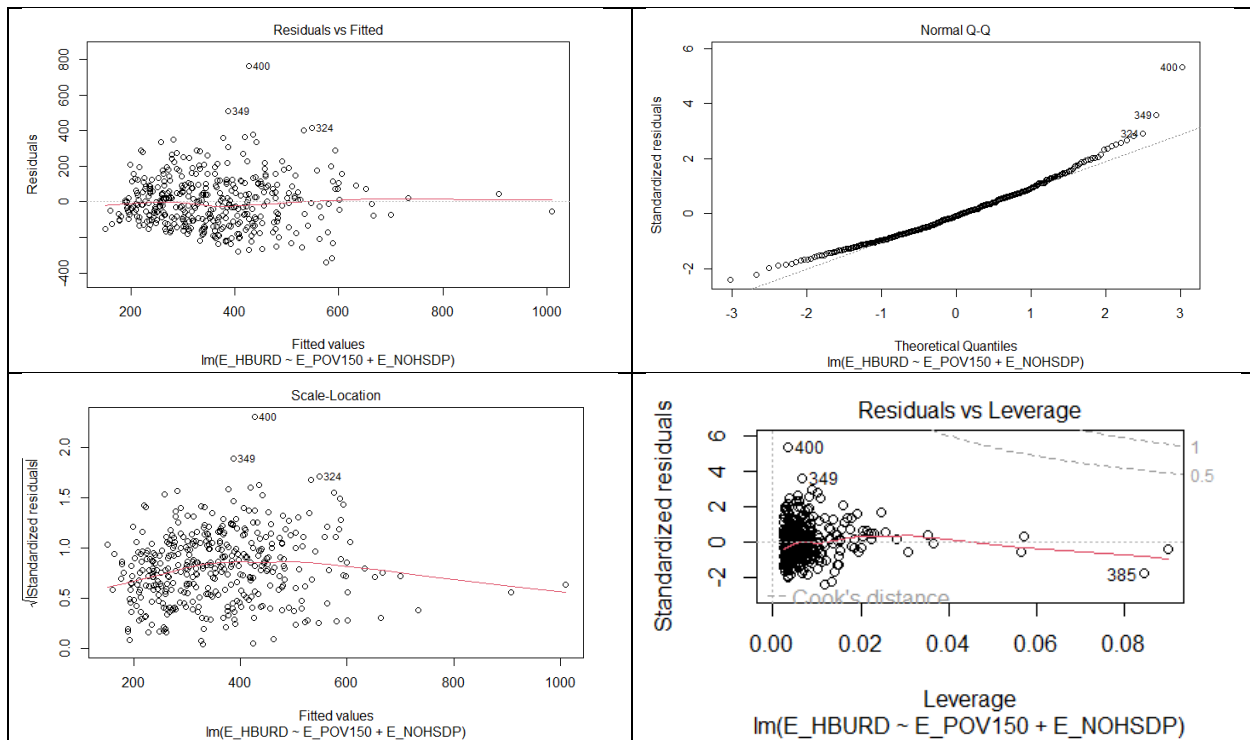


Figure 3. Plots of OLS Residuals

In the residuals versus fitted graph there is slight deviance from the center, which indicates change in variance. There may be a better fit model for this data. The QQ model shows a long right tail, which generates high kurtosis. The p-value of  $2.2e-16$  in the JB test supports this interpretation, as it indicates the null hypothesis of normality should not be rejected. The scale location graph indicates that variance in the data is not constant, suggesting non linearity in the data. The residuals versus leverage graph does not show any points outside of Cook's distance. However, outliers can be noted which may have influence on variance in the data. This may indicate a there is a better fit model for the data. The Breusch-Pagan test p-value of 0.036 indicates that the null hypothesis of homoskedasticity is rejected and heteroskedasticity is not assumed.

Figure 4.

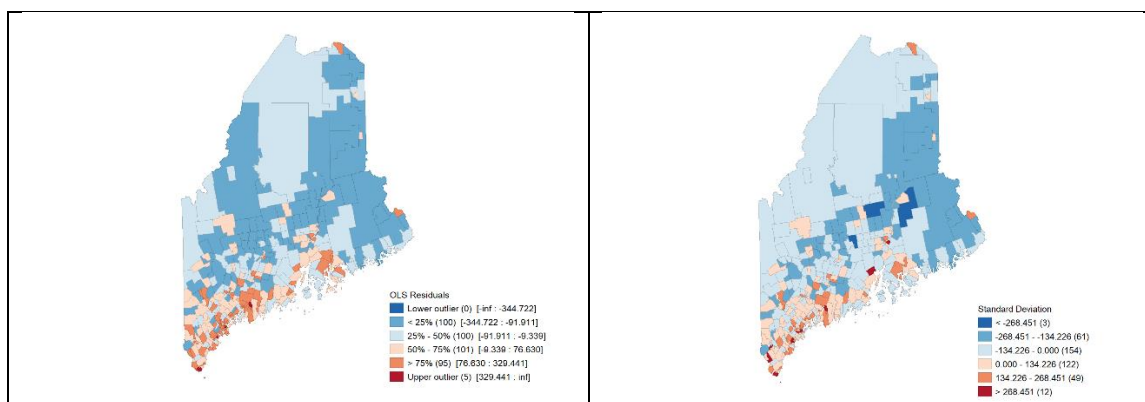
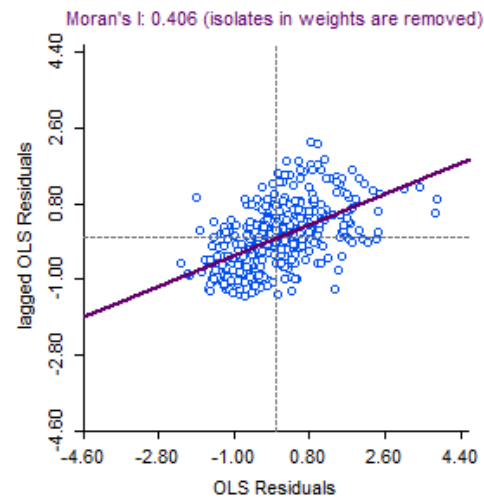


Figure 4. Left: OLS Residuals, Right: OLS Residuals with respect to standard deviation

Figure 4 depicts maps the OLS residuals and maps the residuals with respect to standard deviation. Visually, autocorrelation may be suggested if there are clusters of positive and negative standard deviations and regions. The maps shows some clusters of positive and negative, with more frequent negative clusters.

Figure 5.

Figure 5. Local Univariate Moran's I Analysis, GeoDa



Moran's I analysis was run in both GeoDa and R Studio. In R Studio Global Moran's I was observed at 0.36 with a p-value of  $<2.2e-16$  and in GeoDa Local Univariate Moran's I was observed at 0.406 with a p-value of 0.01. The Monte Carlo test of Moran's I was run in R Studio indicating a value of 0.36 and p-value of 0.01. This suggests the null hypothesis can be rejected. The alternative hypothesis suggests positive autocorrelation. The Moran's I value of 0.36 is between -1 and +1, which suggests significance as well as positive auto correlation.

Figure 6.

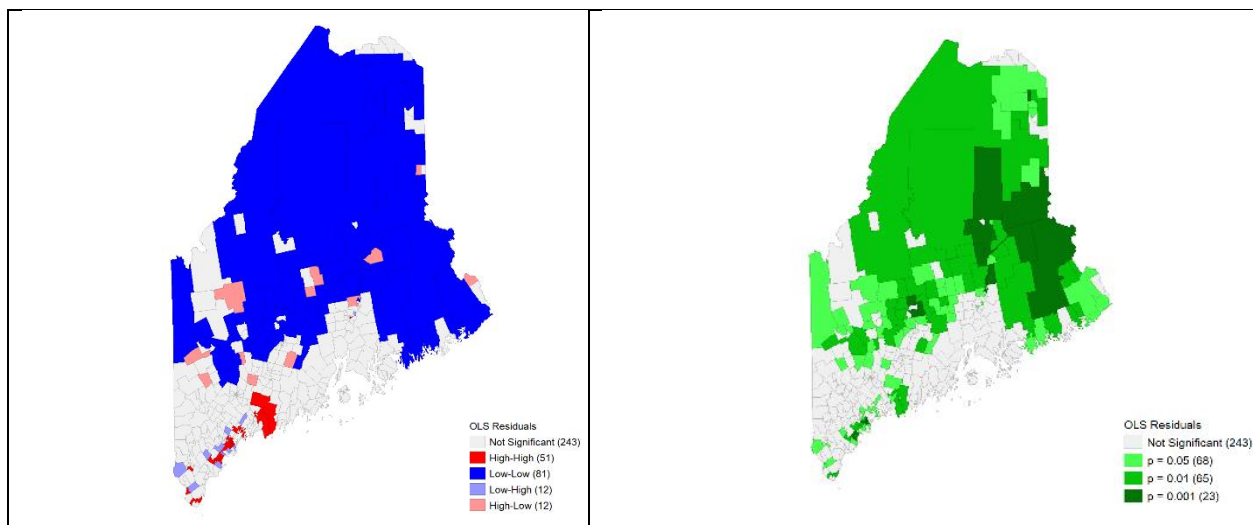


Figure 6. Left: Local Univariate Moran's I Cluster Map, Right: Significance Map Moran's I

Figure 6 shows Local Univariate Moran I Cluster and Significance Maps made on GeoDa. The Local Univariate Moran I Cluster Map with respect to residuals corresponds with the significance map. Significant Local Moran I values are highlighted in green on the significance map. The cluster map shows 51 High-High and 81 Low-Low areas, with 12 Low-High, and 12 High-Low outliers, consistent with a positive spatial autocorrelation.

Because spatial autocorrelation may be present Lagrange Multiplier Statistics was run in R Studio to determine if a Spatial Lag or Spatial Error model would be a better fit for the data.

Figure 7.

Diagnostic Test	Statistical Parameter	P-value
LM Error	166.997	<2.2e-16
LM Lag	62.357	2.887e-15
Robust LM Error	116.640	<2.2e-16
Robust LM Lag	12.001	0.00053

Figure 7. Lagrange Multiplier Statistics

Figure 7 indicates the results of Lagrange Multiplier Statistics diagnostics. Both the Robust LM Error and Robust LM Lag test were significant. Therefore, a Spatial Lag and Spatial Error Model were run.

Figure 8.

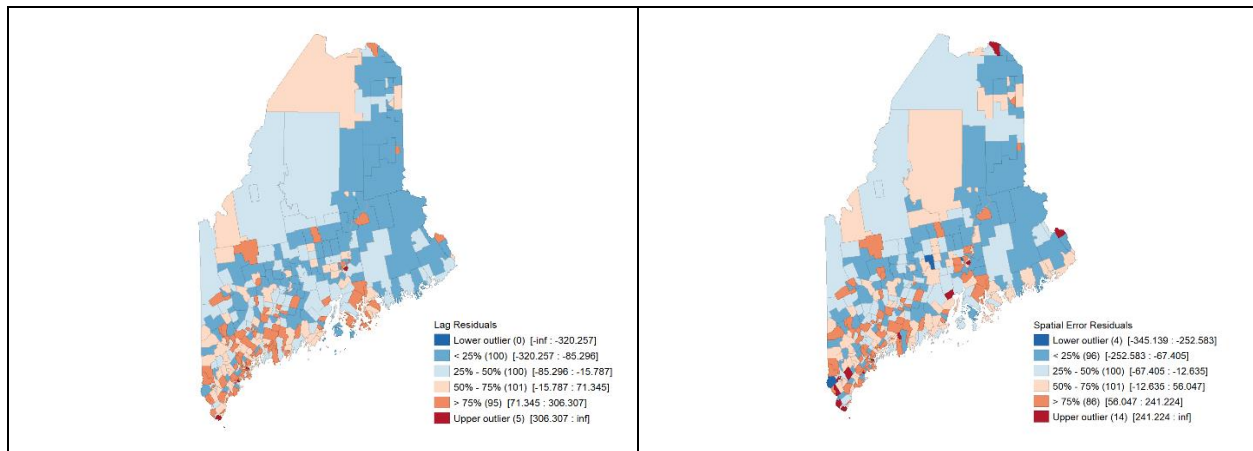


Figure 8. Left: Spatial Lag Residuals, Right: Spatial Error Residuals

Figure 8 depicts the Spatial Lag and Spatial Error Residuals, showing a fairly even distribution of positive and negative regions. The summary of the Spatial Lag model indicates an intercept of -38.50, and the coefficient of poverty as 0.20 and the coefficient of disability as 0.27, resulting in a positive correlation. For the Spatial Lag test the p value for y intercept was 0.06 the p value for the poverty coefficient was <2.2e-16, and the p value for the disability coefficient was 9.1e-15. The coefficients were statistically significant, but the y intercept was not. The p value for the model was 5.01e-14.

The summary of the Spatial Error model indicates an intercept of 65.11, and the coefficient of poverty as 0.24 and the coefficient of disability as 0.0003, resulting in a positive correlation. For the Spatial Lag test the p value for y intercept was 0.06 the p value for the poverty coefficient was <2.2e-16, and the p value for the disability coefficient was <2.26e-16. The p value for the model was 4.45e-8.

Because the y intercept and coefficients were statistically significant the Spatial Error model may be a better fit model.

*Figure 9.*

Model	AIC	BIC (Schwartz)	Log Likelihood
OLS	5074.404	5090.380	-2533.202
Spatial Lag	5025.002	5044.972	-2507.501
Spatial Error	4967.624	4987.594	-2478.812

*Figure 9. Analysis of Spatial Regression Models*

Figure 9 shows the AIC, BIC, and Log Likelihood of the OLS, Spatial Lag, and Spatial Error Model. A higher Log Likelihood and lower AIC and BIC, value indicates a better fit model. Based on the AIC, BIC, and Log Likelihood values, the Spatial Error model is likely the best fit model. This is consistent with the statistical significance of the Spatial Error model.

### **Further Considerations**

The results of this project indicate a positive correlation between housing cost burden and the covariates measuring poverty and disability. Additionally, there may be positive spatial autocorrelation between housing burden cost and the two explanatory variables of poverty and disability. Based on analysis, the best fit model for this data is the Spatial Error Model based on Lagrange Multiplier Statistics, and the accompanying AIC, BIC, and Loglikelihood values.

One issue in this analysis was the weight matrix used for Moran's I. The weight matrix for Moran's I used in this project was a queen's case weight matrix. This was based on previous methods in subsequent studies (Chen, et al. 2019). However, because Maine has populated islands, a k-nearest neighbors weight matrix may have been a better fit for the data. The k-nearest neighbors algorithm may have successfully captured island data. A comparison of both methods may be useful to determine how spatial autocorrelation may change based on weight matrix.

Examining the relationship between housing cost burden and measures of social vulnerability may help to guide policy making. According to the Maine State Housing Authority, 31,009 households at risk of eviction received help to pay rent and utilities in 2020 (Brennan, 2022). Policy can be best targeted by isolating risk factors associated with housing cost burden. Further spatial analysis can aid with the process of determining how to best support households in need of assistance.

## Sources:

Brennan, D. (2022, September 13). Housing in Maine: An Overview. *Maine State Housing Authority*.  
<https://legislature.maine.gov/doc/8866>

CDC SVI Documentation 2020. (2020, August 5). Agency for Toxic Substances and Disease Registry.  
[https://www.atsdr.cdc.gov/placeandhealth/svi/documentation/SVI\\_documentation\\_2020.html](https://www.atsdr.cdc.gov/placeandhealth/svi/documentation/SVI_documentation_2020.html)

Chen, S., Sleipness, O.R., Christensen, K.M. *et al.* Environmental justice and park quality in an intermountain west gateway community: assessing the spatial autocorrelation. *Landscape Ecol* **34**, 2323–2335 (2019). <https://doi.org/10.1007/s10980-019-00891-y>

GeoDa: Answers to Technical Questions. GeoDa Workbook.  
[https://aqs.epa.gov/aqsweb/airdata/download\\_files.html](https://aqs.epa.gov/aqsweb/airdata/download_files.html)

Olsen, S. (2022, July 21). Rent cost are increasing nationwide, and Maine is no exception. *News Center Maine*. <https://www.newscentermaine.com/article/money/economy/rent-costs-are-increasing-nationwide-and-maine-is-no-exception-bangor-portland-apartment-list-redfin-costar-group-prices/97-27bf2f5e-3b5e-4ac5-92e3-50616bde31dd>

Olakunde, B.O., Pharr, J.R., Adeyinka, D.A., Conserve, D.F., & Duncan, D.T., Spatial analysis of HIV infection and the associated correlates among transgender persons in the United States, *AIDS Care*, **34**, 1000-1007 (2022). DOI: [10.1080/09540121.2021.1929817](https://doi.org/10.1080/09540121.2021.1929817)

Plant, R.E. (2019). *Spatial Data Analysis in Ecology and Agriculture Using R* (Second Edition).