GIS 563
Local Statistical Modeling
Final Project

Nathan A. Nguyen

10 December 2024

Introduction

This write-up serves as the second part submission for the final project in the class.

For this project, an empirical application of MGWR was performed and compared with a global OLS model. The response variable was the estimated median household income in 2021, and the spatial units were United States counties. Details about the dataset and any preprocessing that occurred will be discussed. The methods section will cover the models explored and the software implementation as well as what R packages were used for mapping for those who are interested. Finally this write-up will close with brief discussions of the results and any room for improvements should this be a real academic project.

The model presented in this write-up will deviate slightly from the model presented in part 1. The modifications were made to accommodate critiques while presenting – namely the use of poverty level as a predictor. This variable was replaced with the percentage of the population in a respective county that are considered in an urban area. During the initial presentation, the Monte Carlo test was still running, so no results were available.

Due to time constraints, and unexpected events, a Monte Carlo test for the existence of spatial variability was not performed for the model presented in this write-up. The Monte Carlo test for the first version of this model did complete eventually, and it suggested that the only variable with evidence for spatial variability was the all age poverty levels in 2021. I included this variable initially because although poverty is obviously associated with income, I wanted to see whether or not the effects of poverty on median income were uniform across space or if they changed based on geography.

That being said, a rigorous test for spatial variance will not be provided for this second model presented.

Data Details

The dataset used for analysis is an amalgamation of various datasets from the United States Census Bureau/United States Department of Commerce, the American Community Survey, the United States Department of Agriculture's Economic Research Service, and from a 2022 paper by Fotheringham et al.,[1] .

The area of study were United States counties, and only mainland counties were intended to be retained in the dataset. For transparency, most of Connecticut is missing and this issue was not observed until after-the-fact. The missingness is attributed to the non-standardization of FIP codes among the various datasets used. Some locations in Connecticut are not considered true counties and are "county equivalents", which was not known a-priori.

After preprocessing, the final dataset consisted of 3,100 locations. Some R functions were defined in order to assist the preprocessing step – namely cleaning of FIP codes and joining all of the datasets together.

The chosen response variable was the estimated median household income in the year 2021 and was provided by the Census Bureau/Department of Commerce[6]. Nine predictors were included in the models:

1. Gini Index (1-year estimate; 2021)[5]
2. Population Density (natural logged)[1]
3. Percent of Households with Internet Access (5-year estimate; 2017-2022)[2]
4. Percent of Population with Bachelors Degree or Higher[3]
5. Percent of Population Living in an Urban Area(1-year estimate; 2020)[6]
6. Sex Ratio (Male-to-Female) (5-year estimate; 2017-2022)[4]
7. Median Age (5-year estimate; 2017-2022)[4]
8. Percent Population that is Black (5-year estimate; 2017-2022)[4]
9. Percent Population that is Hispanic or Latino (5-year estimate; 2017-2022)[4]

Table 1: Summary Statistics for Selected Variables

Variable	Min	Mean	Median	Max
Median Income (21)	25653.00	58741.99	56465.50	153716.00
Gini Index (17-21)	0.25	0.45	0.44	0.73
Population Density (Natural Log)	-1.93	3.78	3.78	10.77
% Internet Access (21)	35.97	82.78	83.89	100.00
% with Bachelor's Degree or Higher (18–22)	0.00	23.44	20.90	78.90
% Population in Urban Area (20)	0.00	35.95	33.41	100.00
Sex Ratio (Male:Female, 17–21)	76.90	101.93	99.60	221.30
Median Age (17–21)	22.40	41.52	41.30	68.10
% Black (17–21)	0.00	9.03	2.26	87.12
% Hispanic or Latino (17–21)	0.00	9.82	4.49	98.22

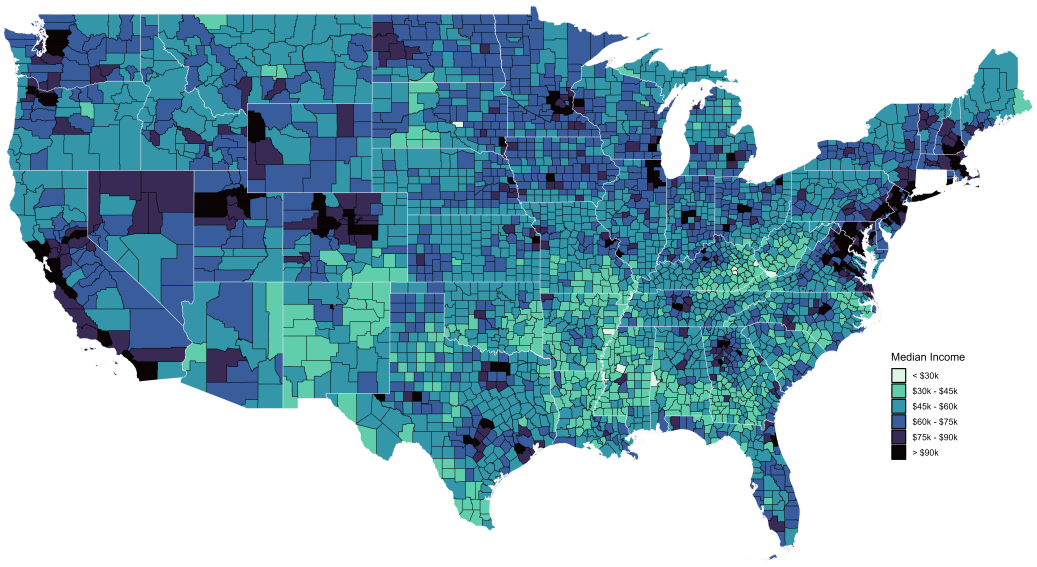

Figure 1: Estimated Median Household Income (2021)

Figure 1 shows the estimated median household income in 2021. By inspection, there appears to be some clustering of this random variable. For example, we see that the median household income is generally higher in the north-east coast of the United States, indicated by the darker coloring, when compared to the deep south and Appalachia, indicated by the lighter shading. The west coast, generally, has higher income as well.

This is likely due to the fact that the north-east and west-coast are more developed regions of the country with high paying industries like finance and technology while the regions with lower income are more rural and have declining industries e.g., manufacturing and mining.

Furthermore, the regions with higher median household income generally have higher educational attainment as well. Figure 2 is evidence of this:

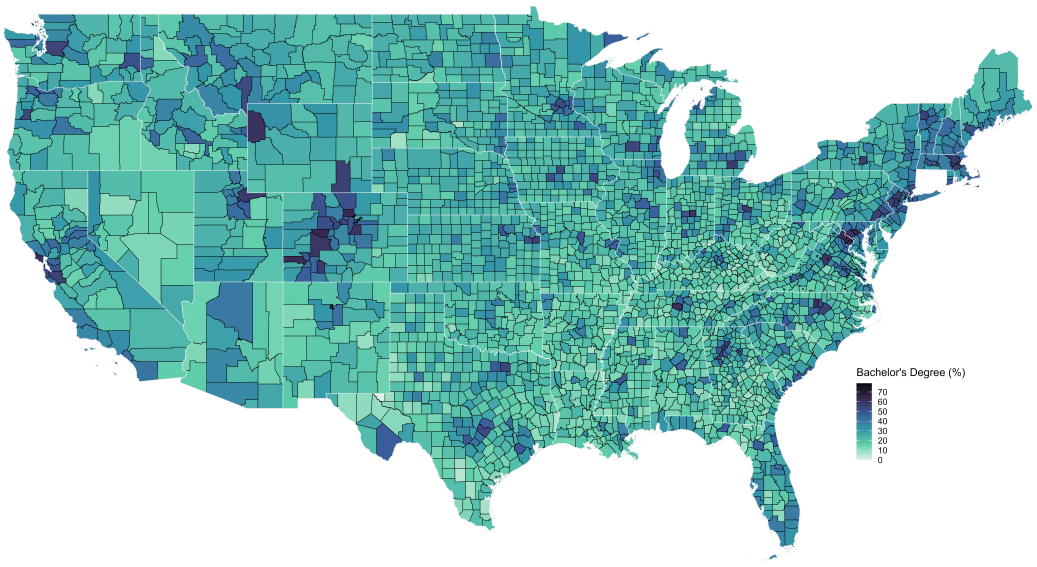

Figure 2: Percent of Population Having Bachelors Degree or Higher

Methods

A global OLS and MGWR model was fit on the dataset and compared to one another. The variance explained in both models were compared as well as the AICc. All variables were standardized to have mean zero and and variance one, and standardization was performed in R. All features in the model are of order one, and no interaction terms were considered.

A linear model was fit in R using the `lm()` function while the MGWR model was fit using the MGWR 2.2 GUI software. I also calibrated an MGWR with the `mgwr` python module, but I ran into issues while extracting some data.

Ad-hoc tests were used in place of a Monte Carlo test:

$$IQR_k > 2 \times SE_{k-global}$$

Corrected α-values were computed following:

$$\alpha_j = \frac{\alpha^*}{ENP_j}$$

where $\alpha^* = 0.05$ and the ENP_j were obtained from the `txt` file from the MGWR 2.2 session.

Global OLS Model

Table 2: Global OLS Results

Variable	**Estimate**	**Std. Error**	**t-value**	**p-value**
Intercept	5.87e-17	1.00e-02	0.000	1.000
Gini Index (17–21)	-0.259	0.011	-22.595	<2e-16 ***
Population Density (Log)	0.179	0.016	11.468	<2e-16 ***
% Internet Access (21)	0.237	0.015	15.627	<2e-16 ***
% Bachelor's Degree or Higher (18–22)	0.579	0.014	41.660	<2e-16 ***
% Population in Urban Area (20)	-0.124	0.018	-6.994	3.26e-12 ***
Sex Ratio (Male:Female, 17–21)	0.351	0.107	3.295	0.000997 ***
Median Age (17–21)	0.099	0.118	0.845	0.398
% Black (17–21)	-0.059	0.012	-4.929	8.70e-07 ***
% Hispanic or Latino (17–21)	0.080	0.012	6.975	3.73e-12 ***
Residual Standard Error: 0.5578 on 3090 degrees of freedom				
Multiple R-squared: 0.6897, *Adjusted R-squared:* 0.6888				
F-statistic: 763.3 on 9 and 3090 DF, *p-value:* <2.2e-16				

Signif. Codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

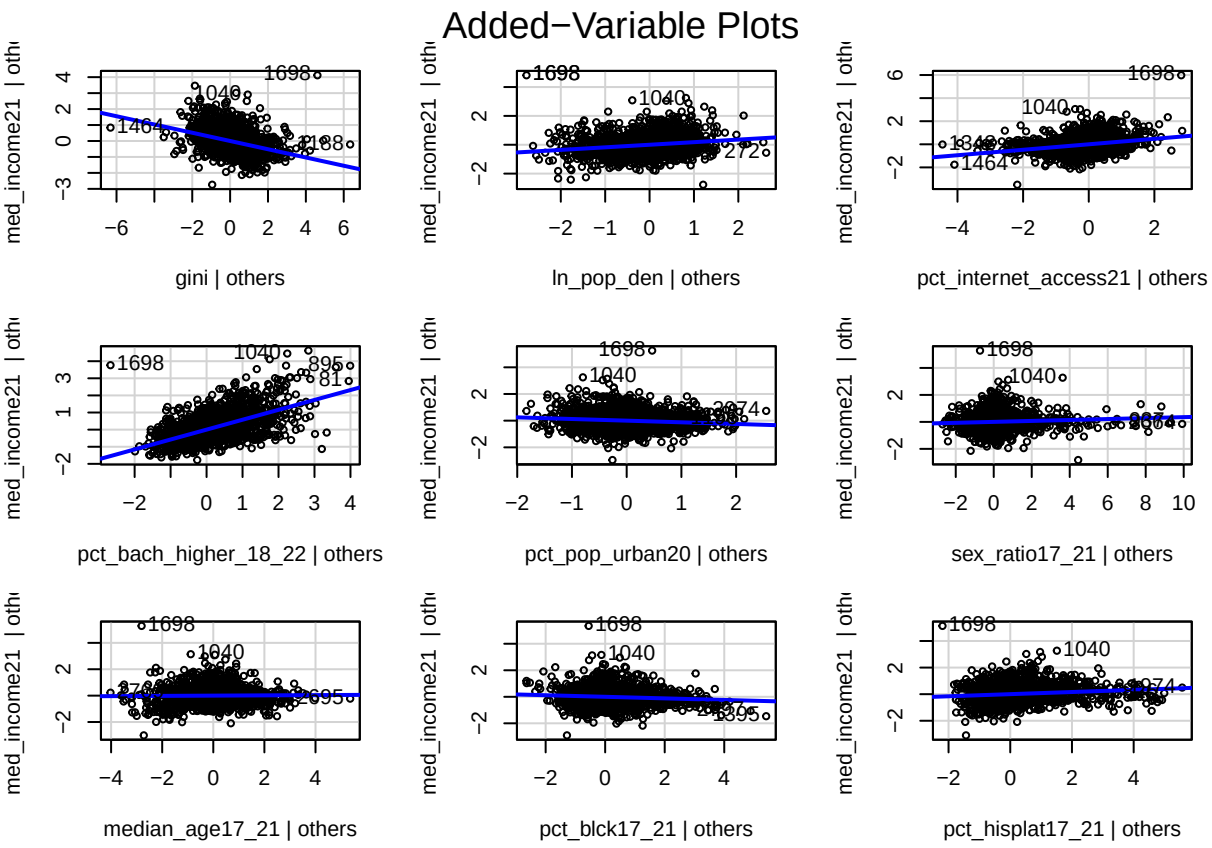

Figure 3: Global OLS Added Variable Plots

A summary of the global OLS model results is contained in Table 2. The global model is able to explain approximately 69% of the variance of standardized median income (adjusted R^2). Among all predictors, all are statistically significant at $\alpha = 0.05$ except for median age. A side comment is that with sufficiently large sample sizes, any non-trivial effects will be statistically significant. Effect sizes might be a better measure of model quality in the future.

Educational attainment had the largest positive effect on median income as seen in Table 2 as well as the added variable plots (Figure 3). A one standard deviation increase in the percentage of the population having a bachelors degree or higher is associated with a 0.579 standard deviation increase in the median household income with all other variables held constant. This result is non-surprising and is well supported in modern socioeconomic theories; however, it underscores the importance of education attainment and earning potential. Higher paying industries like technology, engineering, and so on oftentimes require at least a bachelors degree to be considered "qualified" for a role.

The Gini Index, a measure of income inequality, had the largest negative effect on median income. A one standard deviation increase the index is associated with a -0.259 standard deviation decrease in the median income with all other variables held constant. This can also be observed in Figure 3.

The percent with internet access and population density variables both have positive effects on the median income. This might reflect that counties with more developed infrastructure and are more densely populated have higher median incomes, which is a logical conclusion. If there's a large, and densely, populated area, then there's an incentive to invest in infrastructure. Interestingly though is that the percent of the population living in an urban area has a negative effect on the median income.

The sex ratio also has a relatively large and positive effect on median income. As the number of males

increase in the population, the median income increases by about 0.351 standard deviations. This could be explained by the known so-called gender pay -gap, but could also be largely attributed to what industry someone works in. After all, this dataset is very aggregated.

An increase in the population being composed of Black individuals is associated with a -0.059 standard deviation decrease in median income, while an increasing in the Hispanic or Latino population is associated with a 0.08 standard deviation increase in the median income. In both cases, the effects seem marginal. In both cases, the dominant reference group is the White population.

Although the global model has strong explanatory power ($R^2_{adj} \approx 69\%$), it is not without limitations. The diagnostic plots (Figure 4) indicate potential violations in the assumptions of linear regression – name heteroskedasticity (non-constant variance) and the distribution of the residuals being non-normal. Figure 4 is evidence of heteroskedasticity as there is a cone structure in the residuals. The residuals get larger for larger predicted values of Y. Furthermore, it can be observed that the distribution of residuals has fatter right-tails and skinnier-left tails. If the residuals were distributed normally, then the standardized residuals would be more symmetric and it would hug the theoretical line more tightly.

Figure 5 also suggests that higher-order terms, or at least some transformation on the raw variables, might be warranted. The component-residual plot for the percent with internet access variable shows curvature in the data. This indicates a non-linear specification of this variable might be the proper functional form. All other graphs are relatively linear.

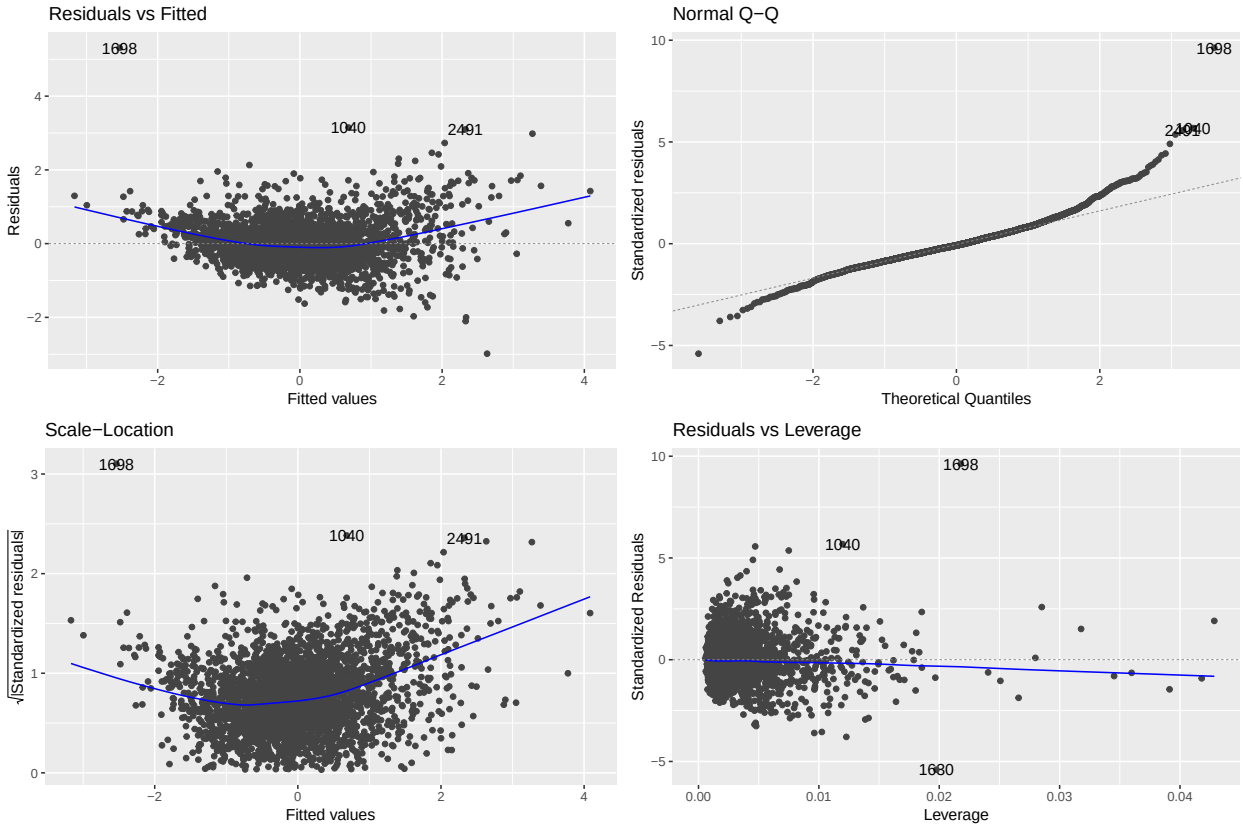

Figure 4: Global OLS Diagnostics

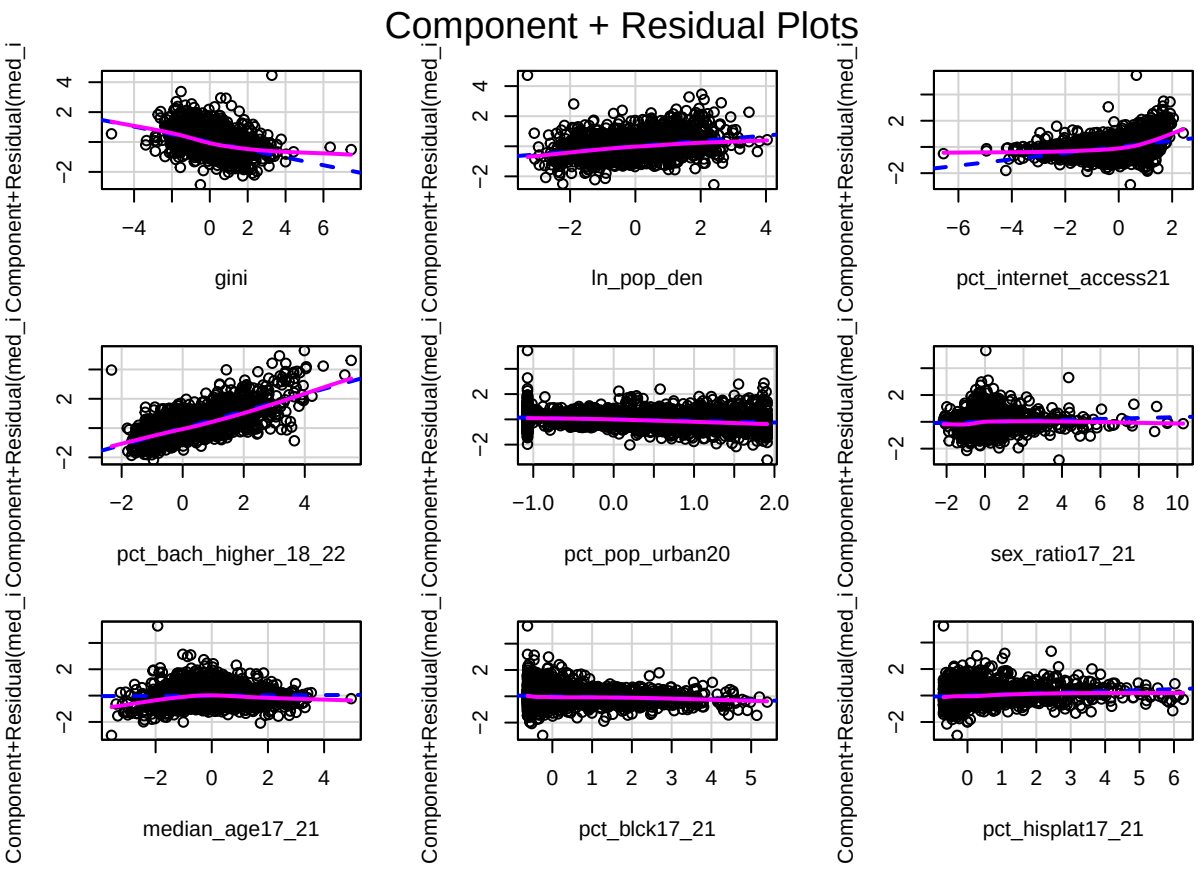

Figure 5: Global OLS Component Regression Plots

On the issue of heteroskedasticity, it is possible that spatial autocorrelation might be the factor. It's reasonable to suspect that locations are more similar to one another. In this case, counties with higher median income might be clustered together. To assess this, a Moran's Test was implemented on the OLS residuals. Queen's contiguity was used to define neighbors, and row-standardized weights were chosen to give all neighbors equal weights.

The results of the Moran's test are:

Table 3: Global OLS Residual Moran's I Test Results

Statistic	**Value**
Moran's I Statistic	0.3087
Expectation	-0.0003
Variance	0.0001
Standard Deviate	28.675
p-value	< 2.2e-16
Alternative Hypothesis	Greater
Notes: Moran's I test under randomization.	

The test is significant at the level of 0.05, and so the null hypothesis is rejected. There is sufficient evidence to suggest that positive spatial autocorrelation exists.

Another way to look at this is to plot the response variable and its lagged counterparts (similarly for the OLS residuals) seen in Figure 6-7:

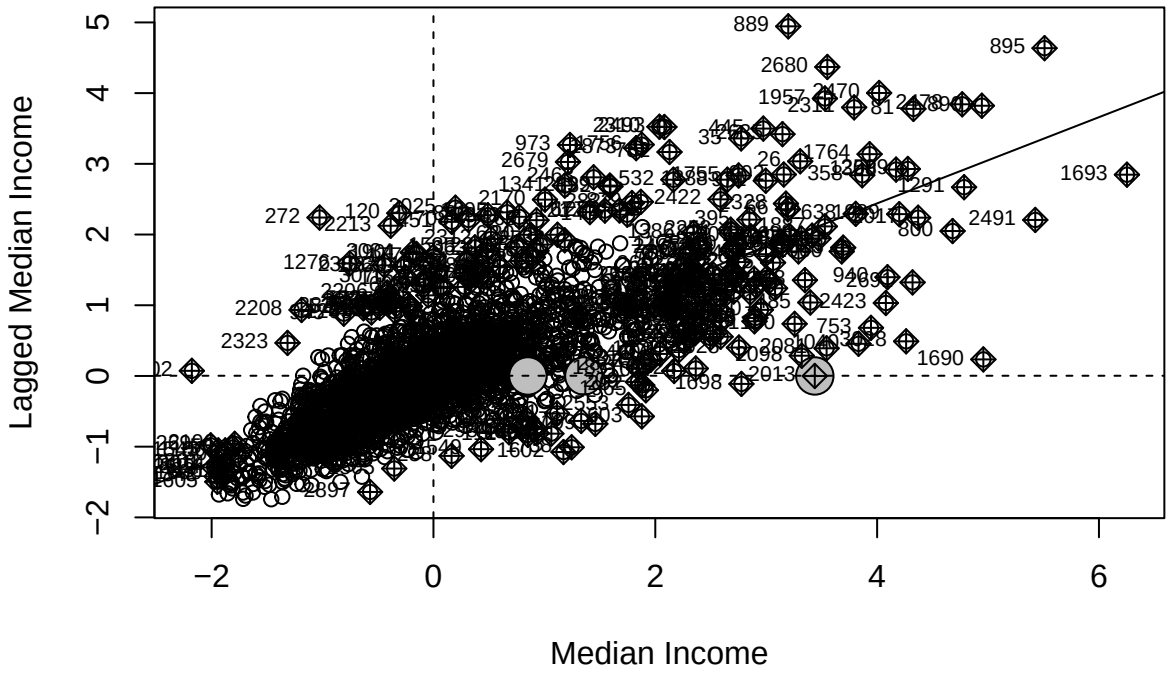

Figure 6: Moran Plot for Median Income

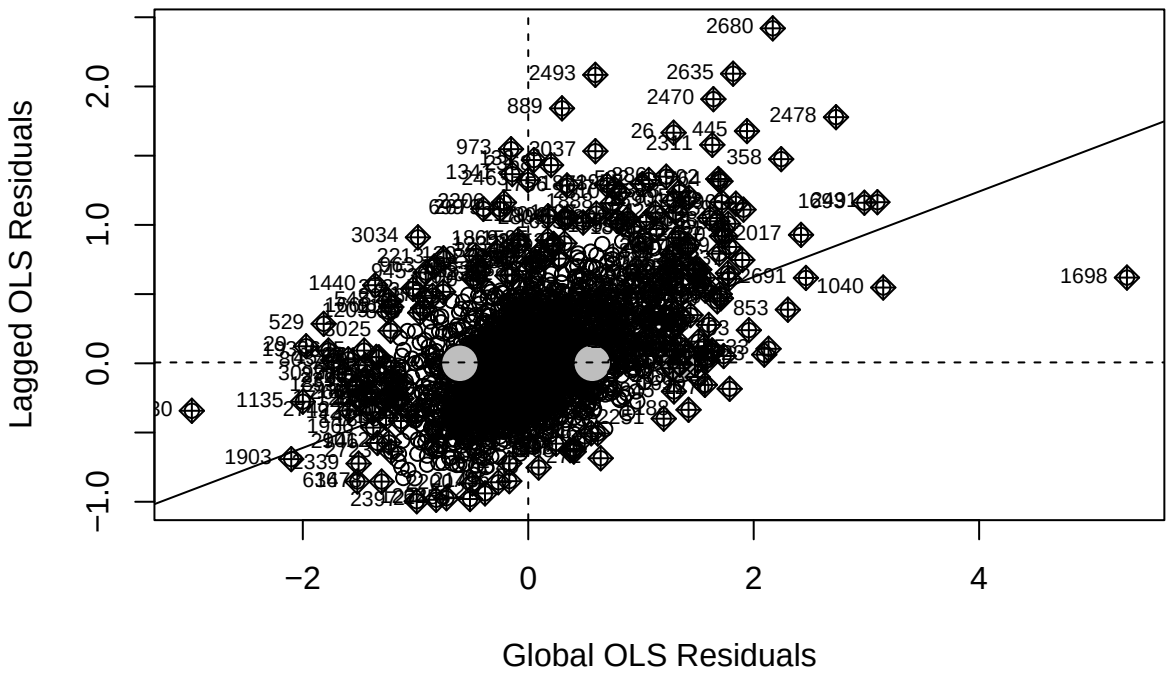

Figure 7: Moran Plot for Global OLS Residuals

If spatial autocorrelation did not exist, the slope of the diagonal line would be approximately zero. In this case, there is clearly a positive slope, and in fact – the slope of the line is the value of Moran's I in Table 3.

The implication is that maybe a global OLS model is insufficient in explaining the data generating process and that a local modeling approach might better capture the underlying processes. This is where MGWR comes into play.

MGWR Model

Table 4: MGWR Model Summary

Variable	Min	Mean	Median	Max	Bandwidth (95% CI)
Intercept	-0.584	0.007	-0.018	0.949	44 [44, 44]
Gini Index (17–21)	-0.568	-0.206	-0.193	0.055	92 [82, 107]
Population Density (Log)	-0.172	0.040	0.063	0.196	588 [488, 764]
% Internet Access (21)	-0.390	0.269	0.230	1.143	44 [44, 46]
% Bachelor's Degree or Higher (18–22)	-0.137	0.471	0.477	0.971	52 [48, 57]
% Population in Urban Area (20)	-0.071	-0.049	-0.050	-0.028	2263 [1932, 2654]
Sex Ratio (Male:Female, 17–21)	-0.004	0.032	0.029	0.136	626 [488, 764]
Median Age (17–21)	-0.297	0.041	0.060	0.258	142 [132, 172]
% Black (2017–2021)	-0.228	-0.226	-0.226	-0.225	3098 [2378, 3098]
% Hispanic or Latino (17–21)	-0.195	0.044	0.067	0.252	473 [423, 594]
Metric					
Residual Sum of Squares					328.055
Log-Likelihood					-917.446
AIC					3023.785
AICc					3306.439
BIC					6613.743
R^2					0.894
Adjusted R^2					0.869
Degree of Dependency (DoD)					0.492

Table 5: IQR (ad-hoc) Results

Variable	IQR	SE (Global)	Threshold	Significant
Intercept	0.3380	0.010	0.020	True
Gini Index (17–21)	0.1420	0.011	0.022	True
Population Density (Log)	0.1780	0.016	0.032	True
Median Age (17–21)	0.1290	0.012	0.024	True
% Bachelor's Degree or Higher (18–22)	0.2470	0.014	0.028	True
% Black (2017–2021)	0.0017	0.012	0.024	False
% Hispanic or Latino (17–21)	0.1670	0.012	0.024	True
% Internet Access (21)	0.2650	0.015	0.030	True
% Population in Urban Area (20)	0.0256	0.018	0.036	False
Sex Ratio (Male:Female, 17–21)	0.0312	0.011	0.022	True

Table 4 provides a summary of the MGWR calibration. This local model is able to explain about 86% of the variance in the data (adjusted R^2). The IQR ad-hoc procedure was performed instead of the recommended Monte Carlo test due to time constraints. All variables except for the percent of the population being Black and what percent of the population living in an urban area showed evidence for spatial variability under this method.

The intercept (location if all other variables were homogeneous and zero), Gini Index, percent with internet access, percent with a bachelors degree or higher, and median age all have relatively small bandwidths and narrow bandwidth confidence intervals when compared to the overall number of locations, $N = 3,100$. This indicates that the effects of these variables are very local i.e., the effects of these variables are not uniform across space, which a global OLS incorrectly assumes. **<insert figure numbers>**. If we were to plot these variables' significant local parameter estimates, we'd expect to see clusters and a lot of variation in parameter surface. Whereas more regional and uniform effects, indicated by larger bandwidths would have a more uniform/smooth coloring on the entire parameter surface.

Population density, sex ratio, and the percent of the population that are Hispanic or Latino have larger bandwidths than the prior variables discussed, but they are not large enough to say they have global effects. For these variables, we argue that these variables have more of a regional effect. For these regional effects, the coloring of the parameter surface will be more smooth when compared to the very local effects, but not completely uniform which would be the case for global effects (large bandwidths).

The percentage of the population living in an urban area and what percent of the population is Black are considered to have global effects. Their point estimate for the bandwidth make up over 70% of the overall number of locations. The parameter surface for these variables are uniform, and so no clustering will be observed.

The effects of internet access are regional, given that the bandwidth is essentially as large as the number of locations (3,100). Having access to internet indicates higher median income likely an indication of more developed localities.

The effects of having a bachelor's degree or higher are very local given its small bandwidths and clustering observed in the map. Having more education has a much stronger positive impact on one's median income in Colorado, New Jersey, and some parts of Ohio wen compared to Arizona and Idaho for example. It's also interesting to note the clustering in the Appalachia area as well. Some pockets with higher education attainment have higher income levels compared to neighbors in the area.

<blank text>

<placeholder>

The following corrected α-values were used for plotting significant local parameter estimated:

Table 6: Corrected Alpha Levels for MGWR Variables

Variable	**ENP_j**	**Alpha Corrected**
Intercept	164.854	0.000303
Gini Index (17–21)	76.129	0.000657
Population Density (Log)	6.687	0.007477
% Internet Access (21)	149.197	0.000335
% Bachelor's Degree or Higher (18–22)	125.310	0.000399
% Population in Urban Area (20)	1.920	0.026042
Sex Ratio (Male:Female, 17–21)	11.088	0.004509
Median Age (17–21)	48.350	0.001034
% Black (2017–2021)	1.027	0.048685
% Hispanic or Latino (17–21)	8.884	0.005628

Comparing Global OLS vs. MGWR

Table 7: Comparison: Global OLS vs. MGWR

Variable	**OLS Estimate**	**MGWR Mean**	**MGWR Median**
Intercept	0.000	0.007	-0.018
Gini Index (17–21)	-0.259	-0.206	-0.193
Population Density (Log)	0.179	0.040	0.063
% Internet Access (21)	0.237	0.269	0.230
% Bachelor's Degree or Higher (18–22)	0.578	0.471	0.477
% Population in Urban Area (20)	-0.124	-0.049	-0.050
Sex Ratio (Male:Female, 17–21)	0.035	0.032	0.029
Median Age (17–21)	0.010	0.041	0.060
% Black (2017–2021)	-0.060	-0.226	-0.226
% Hispanic or Latino (17–21)	0.080	0.044	0.067
Metric	**Global OLS**	**MGWR**	
Residual Sum of Squares	961.468	328.055	
Log-Likelihood	-2584.131	-917.446	
AIC	5188.262	3023.785	
AICc	5190.347	3306.439	
BIC	N/A	6613.743	
R^2	0.690	0.894	
Adjusted R^2	0.689	0.869	
Degree of Dependency (DoD)	N/A	0.492	

Table 7 is a summary comparing the global OLS and MGWR models. The sign of the parameter estimates are consistent between the global OLS and MGWR framework indicating that the global model does agree with the direction of the effects that the predictor variables have when compared to MGWR.

MGWR has superior explanatory power when compared to global OLS ($R^2_{adj-MGWR} \approx 87\% > R^2_{adj-OLS} \approx 69\%$). Despite having far more parameters than OLS, MGWR's AICc is smaller than OLS' AICc ($AICc_{MGWR} \approx 3306 < AICc_{OLS} \approx 5190$). So while being a more complicated model, it does a better job at explaining the data when compared to OLS, and so the trade-off is worthwhile.

The existence of spatial autocorrelation was again explored, but now with the MGWR generated residuals, and the same procedure for the Moran's test was implemented on the MGWR residuals. The test failed to reject the null hypothesis, and so there is insufficient evidence to suggest any sort of spatial autocorrelation in the MGWR residuals. Refer to Table 8 for details. Furthermore, figure 8 is a mapping of the OLS and MGWR residuals. Although subtle, the clustering of the residuals has been mediated with MGWR.

Table 8: MGWR Residual Moran's I Test Results

Statistic	**Value**
Moran's I statistic	0.0023
Expectation	-0.0003
Variance	0.0001
Standard Deviate	0.2432
p-value	0.4039
Alternative Hypothesis	Greater
Notes: Moran's I test under randomization.	

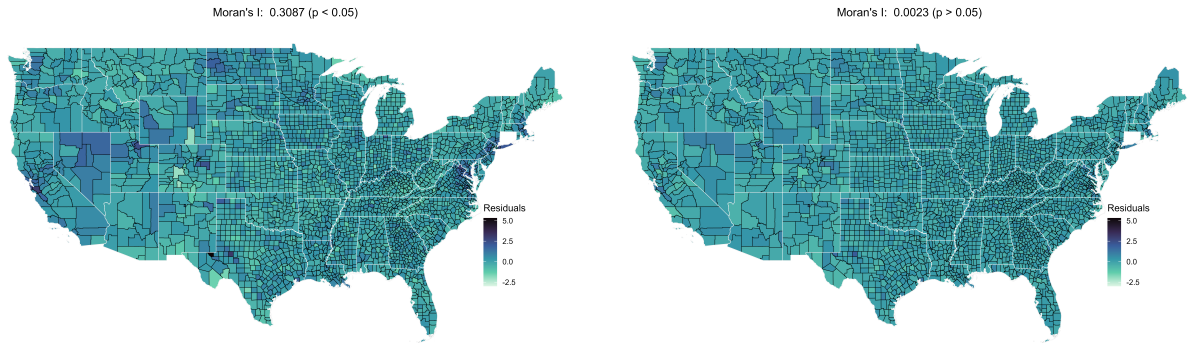

Figure 8: Global OLS Residuals vs. MGWR Residuals

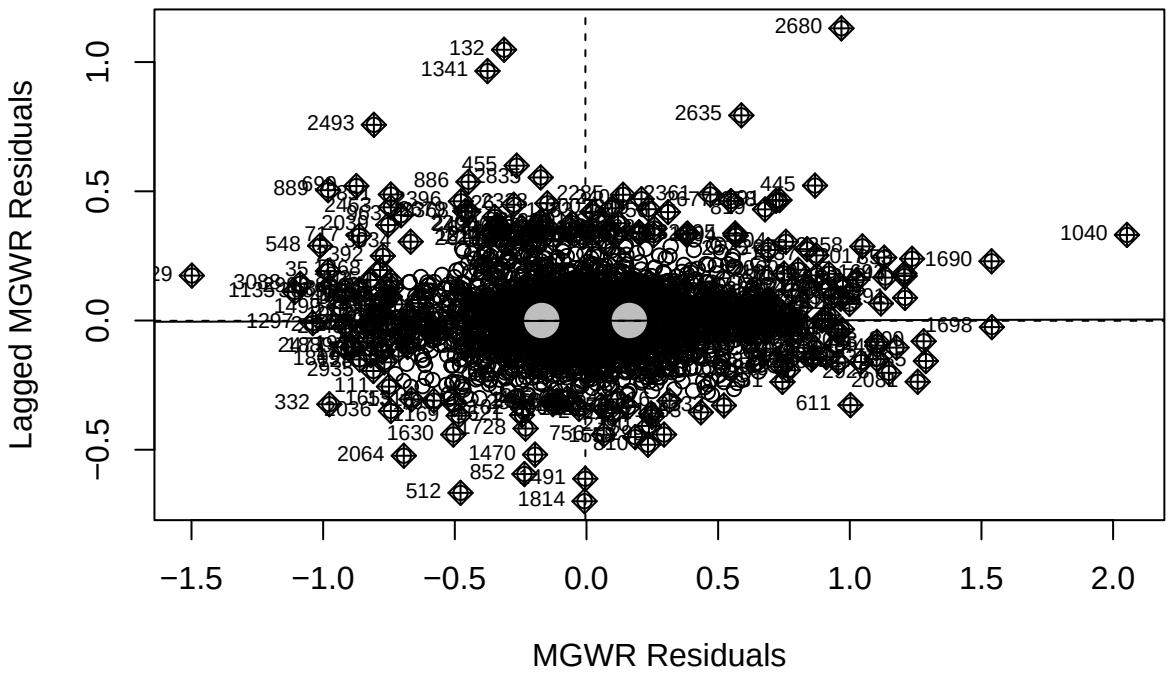

Figure 9: Moran Plot for MGWR Residuals

A Moran plot was also generated for the MGWR residuals (figure 9). The slope of the line is now zero, indicating no spatial autocorrelation.

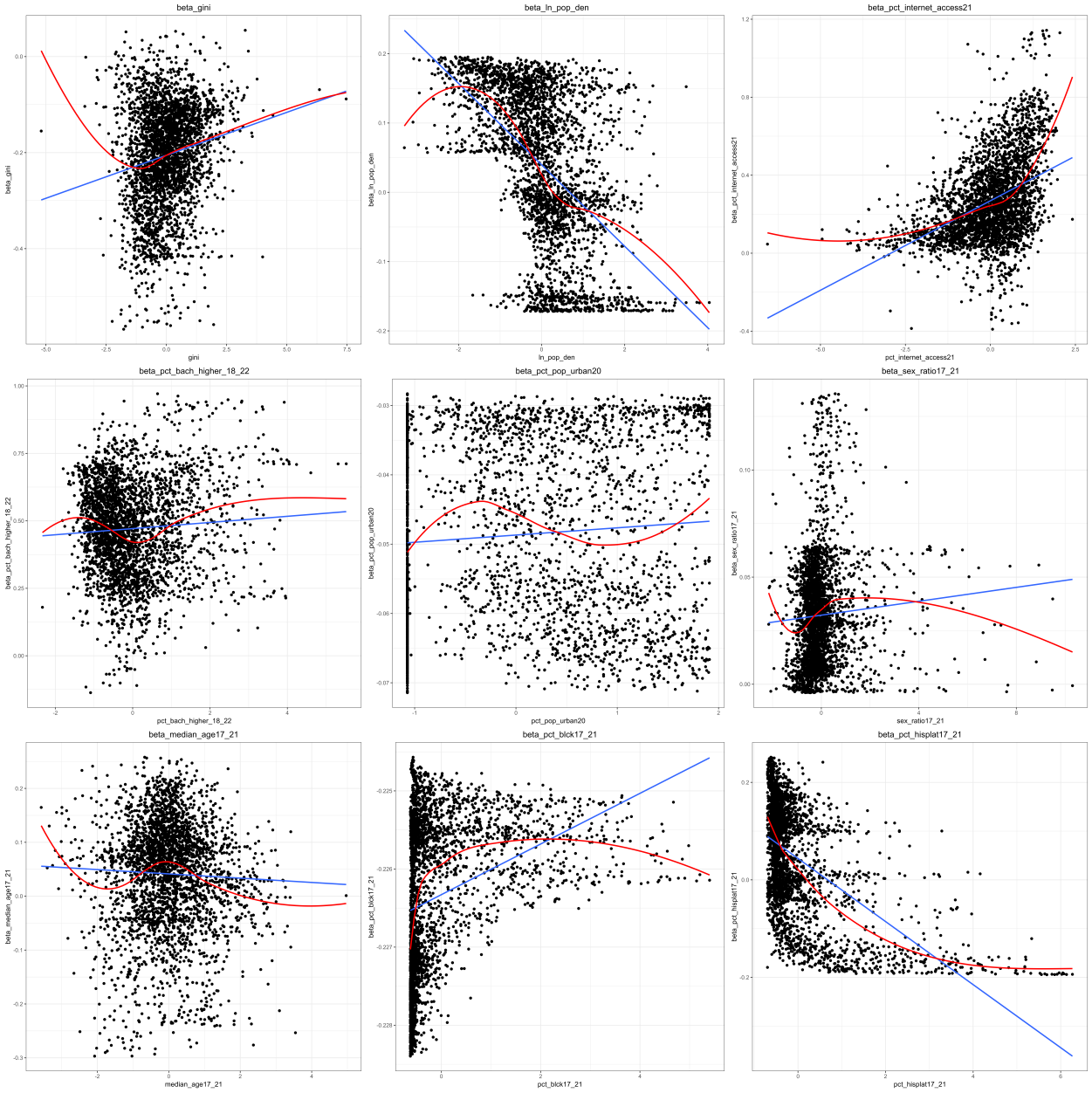

Figure 10: MGWR Diagnostic Plots

Figure 10 shows various plots of the local $\hat{\beta}$s against the the standardized values for each predictor. Upon inspection, first-order specifications for population density, percent with internet access, and percent of population that are Hispanic or Latino may not be appropriate. Polynomial terms, a transformation of the raw variables before standardization, or interaction terms might mediate non-linearity issues, but that is for future work.

Conclusions

Improvements

If I were to write a real paper on something like this, I'd probably change my response variable to something that reflects discretionary income more e.g., the median household income after adjusting for housing costs. While people in California, New York, and Washington might make a lot more than say, someone living in Alabama, people living in California, New York, or Washington likely have a much larger housing cost when compared to someone living in Alabama.

A transformation of some of the predictor variables might be warranted given some non-linearity observed in both OLS and MGWR diagnostic plots.

Finally, a thorough literature review would have been beneficial in order to understand the unique socioeconomic profiles of the US counties. This would have aided in better understanding the results of MGWR and/or helped validated existing theories in the literature. A review would have also provided a better foundation for variable selection in the models e.g., including employment industry variables, and so on.

For those who are interesting in creating the maps in R, please refer to my Github Repo: https://github.com/loafing-cat/gis563-local-stat-model-example.

The following are the core R libraries for mapping:

```
library(tidyverse)
library(sf)
library(tigris)
library(colorspace)
```

References

1. Li, Z., & Fotheringham, A. S. (2022). The spatial and temporal dynamics of voter preference determinants in four U.S. presidential elections (2008– 2020). Transactions in GIS, 26, 1609– 628. https://doi.org/10.1111/tgis.12880

2. U.S. Census Bureau. (2022). Internet Subscriptions in Household. American Community Survey, ACS 5-Year Estimates Detailed Tables, Table B28011. Retrieved December 7, 2024, from https://data.census.gov/table/ACSDT5Y2022.B28011?q=Telephone, Computer, and Internet Access&g=010XX00US$0500000.

3. United States Department of Agriculture, Economic Research Service. (2022). County-Level Data Sets: Poverty estimates. Retrieved from https://www.ers.usda.gov/data-products/county-level-data-sets/county-level-data-sets-download-data/

4. U.S. Census Bureau. (2021). ACS DEMOGRAPHIC AND HOUSING ESTIMATES. American Community Survey, ACS 5-Year Estimates Data Profiles, Table DP05. Retrieved December 9, 2024, from https://data.census.gov/table/ACSDP5Y2021.DP05?q=Density&t=Populations and People&g=010XX00US$0500000.

5. U.S. Census Bureau. (2021). GINI INDEX OF INCOME INEQUALITY. American Community Survey, ACS 5-Year Estimates Detailed Tables, Table B19083. Retrieved December 9, 2024, from https://data.census.gov/table/ACSDT5Y2021.B19083?q=gini&g=010XX00US$0400000.

6. United States Census Bureau. (2023). County-level 2020 Census Urban and Rural Information for the U.S., Puerto Rico, and Island Areas sorted by state and county FIPS codes [Data file]. Retrieved from https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural.html