# Spatial Modeling of Cardiovascular Disease Incidence Positively Associated with PM2.5

Shombit Roy, Christina Kim, Johan Booc

2024-04-28

# Introduction

# Introduction

- As the leading cause of death in the US, we focused on cardiovascular disease (CVD) and its impact on the 18-44-year-old age group and the covariates, PM 2.5 concentration, and socioeconomic factors
- Our goals with this study were to:
  - Fill a gap in prior studies that focused on the Stroke Belt region of the United States
  - Answer the question: Where are the socioeconomic and environmental factors affecting CVD rates in the United States

# Some Key Terms:

- Fine Particulate Matter (PM2.5): particles with a diameter of 2.5 microns or less
- Cardiovascular Disease (CVD): group of disorders affecting the heart and blood vessels
- Cardiovascular Mortality (CVM): for the purpose of our study, the metric of CVM is calculated via deaths from CVD per 100,000 people
- Stroke Belt: region of the southeastern United States with high incidence of CVD compared to the rest of the country

# Visualization of Stroke Belt

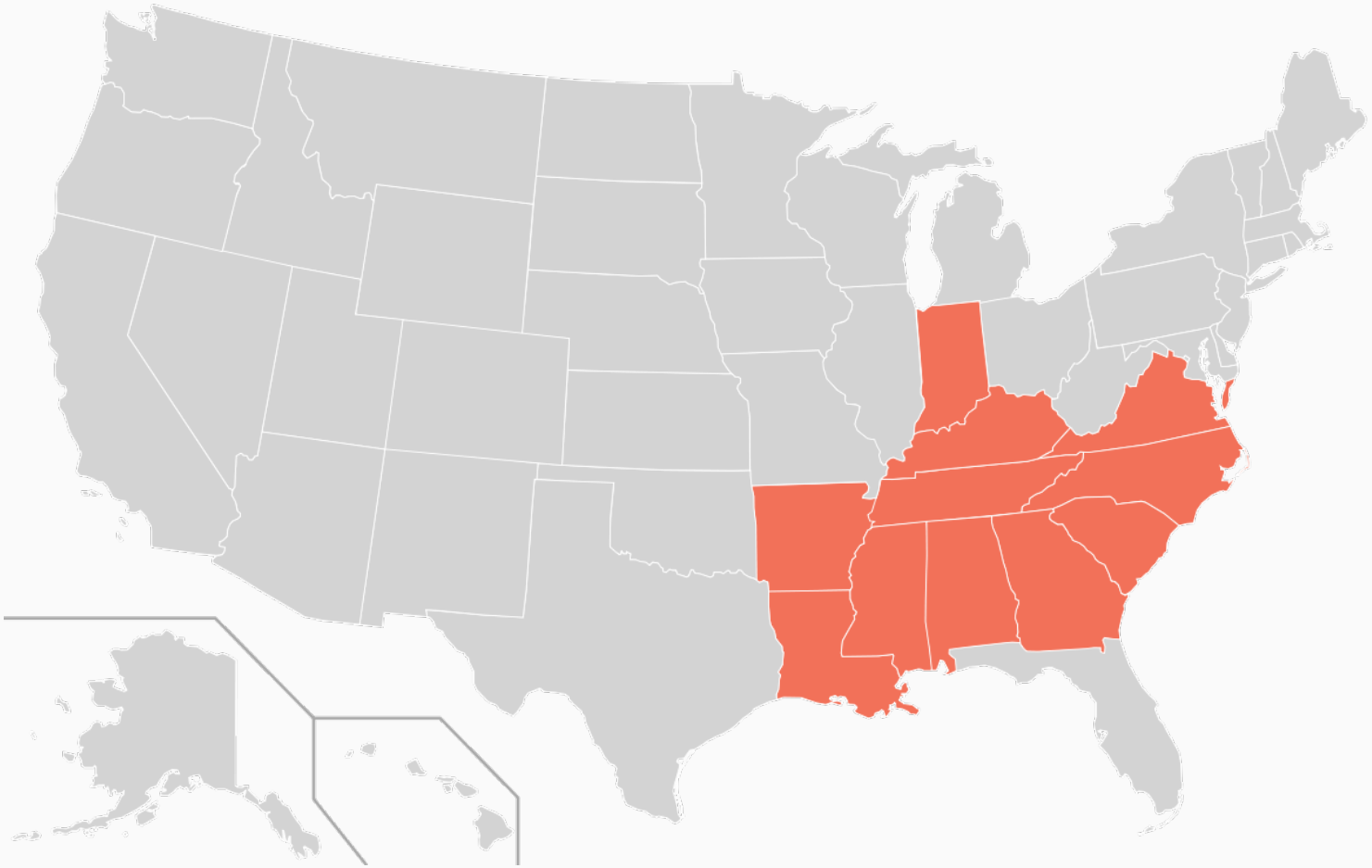

**Figure 1:** States that are a part of the Stroke Belt, highlighted in orange

# Introduction

- We chose to use the geographically weighted regression (GWR) model rather than other methods such as clustering and traditional linear regression for the following reasons:
  - The GWR model highlights the local spatial significance between the response variable and each covariate
  - We can analyze the spatial distribution, changes in percentage rates, and any correlations

# Introduction

- The concentrations of the covariates differ across the US so public health experts need to tailor to each region
- Because many factors influence CVD incidence rates, it's useful to use the GWR model to help health practitioners implement necessary intervention

# Literature Review Section

# Literature Review

- For further context on our problem, we looked towards other research papers studying CVD using spatial analysis
- Statistical regression + spatial models have been growing in popularity to analyze the distribution and factors of cardiovascular disease
- A good amount of other studies focus on socioeconomic covariates and their impact on cardiovascular disease mortality rates however, our study takes into account PM2.5 concentrations
- By analyzing the spatial distribution of our covariates, we can provide risk estimates for the year 2015

# Literature Review

- The geographically weighted model builds on the weighted least squares method and considers coefficients for each spatial unit for estimation
- We are interested in seeing the values that carry more weight because they carry a greater amount of influence

# Literature Review

- The ordinary least squares method generates global regression coefficients however it's prone to hidden spatial variability
- We use the GWR model to test and verify our results to be statistically significant, treating the OLS model as the null
- The GWR model minimizes errors between the actual and predicted values, which helps health practitioners narrow down the areas for improvement

# Methods Section

# Methods

- Medicare claims, racial/geographic data were gathered from the Census and TIGER Bureau, median income was sourced from the Census API, air quality from NASA PM2.5 data, and unemployment rates were loaded from the Center for Disease Control (CDC).
- The data was grouped by year, United States county, and coordinates for 2015; cleaned by removing entries with empty geometries and incomplete data, excluding Alaska and Hawaii, which led to Nantucket County being removed from the dataset

# Methods

- Transformed integrated data into a format suitable for Geographical Weighted Regression (GWR) by calculating centroids of the multi-polygon geometries to provide a single point to get precise location-based assessments.
- Cross-validation was utilized to estimate the optimal fixed bandwidth for the GWR model
- The Gaussian kernel was chosen for its smoothness

# Model

$$y = \beta_0 * \%White + \beta_1 * \%Black + \beta_2 * \%Hispanic + \beta_3 * \%Asian + \beta_4 * PM2.5 + \beta_5 * MedianIncome + \beta_6 * \%Unemployed$$

- A correction for the asymptotics of the t-values from the GWR model is required, as they do not follow a standard t-distribution
  - We adjusted the t-values of the GWR model, aligning the assessment of t-values with appropriate statistical significance testing.

# Methods

- Geographic plots visualize the spatial distribution of CVD death rates across different US counties
- Significance of various socioeconomic, environmental, and demographic factors on CVD death rates is analyzed regionally.

# Methods

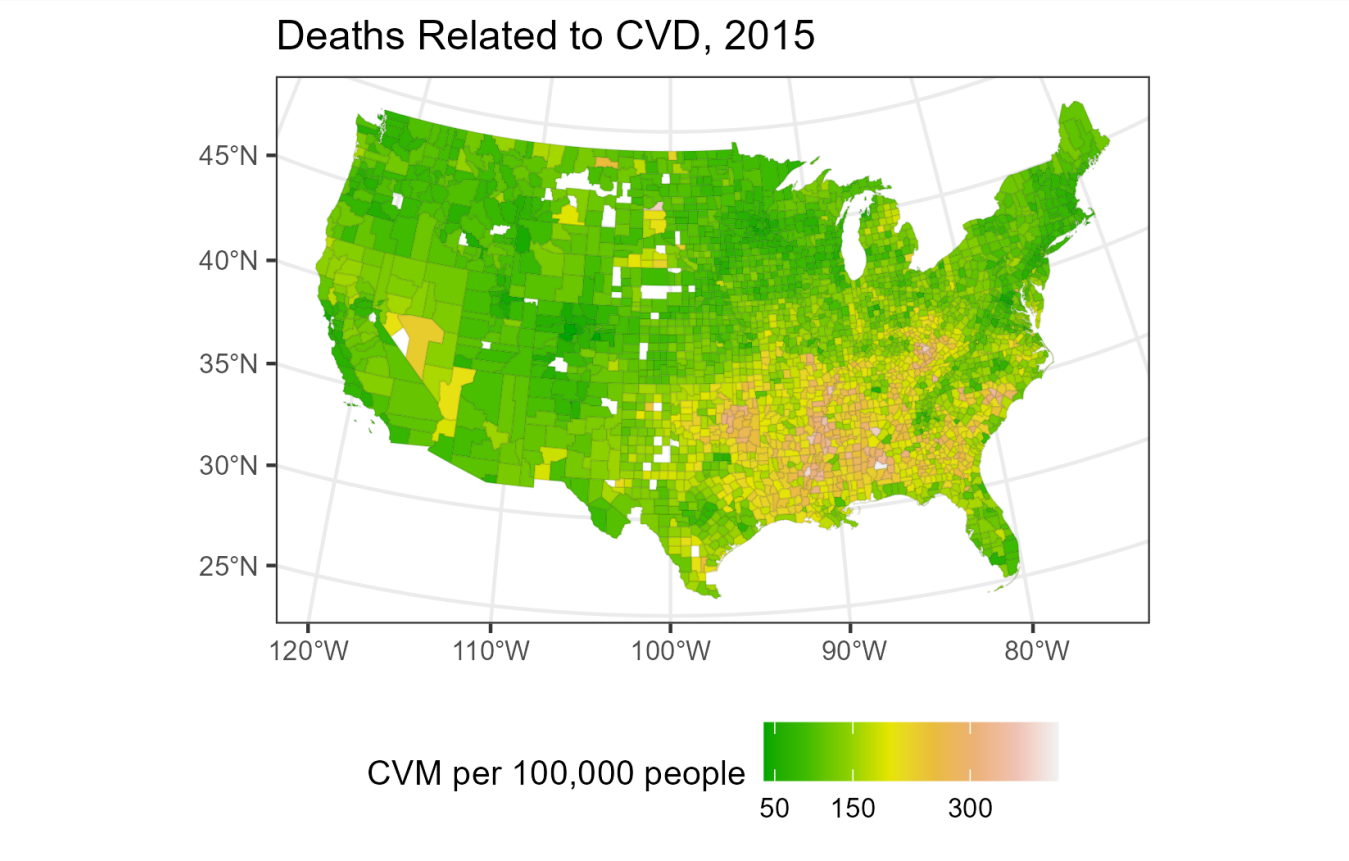


**Figure 2:** Exploratory data analysis with full model. Many deaths are shown to be concentrated in the Stroke Belt region.

# Simulation

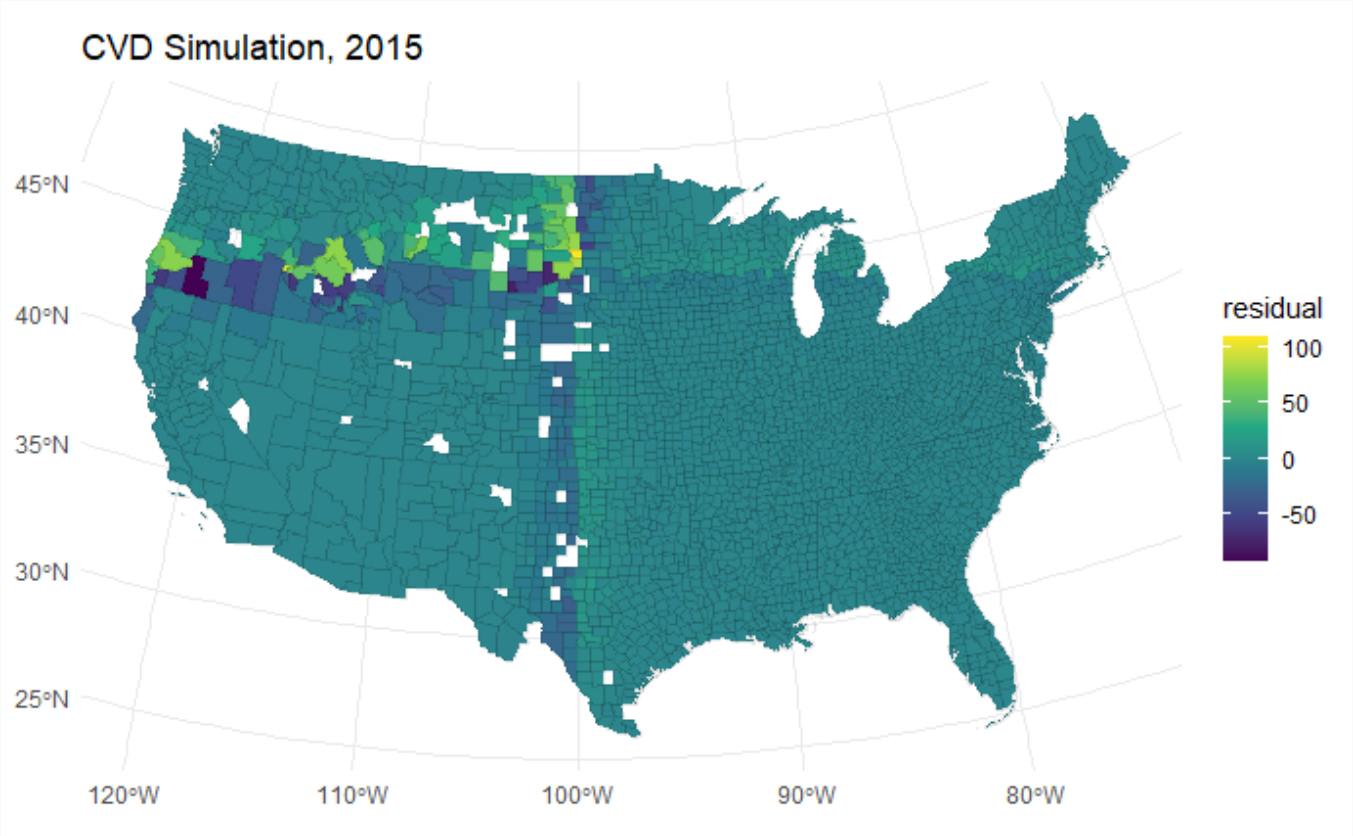


**Figure 3:** Simulated data is used with a piecewise function dividing the geographical space into quadrants, each assigned different coefficients to model spatial heterogeneity

# Simulation

**Table 1:** Moran's I, a spatial autocorrelation test, is performed on the residuals of the GWR model. Results confirm a significant spatial correlation.

| Moran I Statistic | Expectation | Variance |
|---|---|---|
| 0.3010445156 | 0.0003274394 | 0.0001178233 |

# Results Section

# Results

- The global regression model:
  - Shows all predictors as statistically significant.
  - Does not consider spatial correlation, potentially missing local variations.
  - Residual analysis reveals spatial patterns, with clusters of higher and lower residuals indicating missed spatial variation.
- The GWR model:
  - Incorporates spatial variation, addressing a critical aspect of the data.
  - Provides a more nuanced and realistic interpretation of how various factors affect death rates across different regions.

# Results

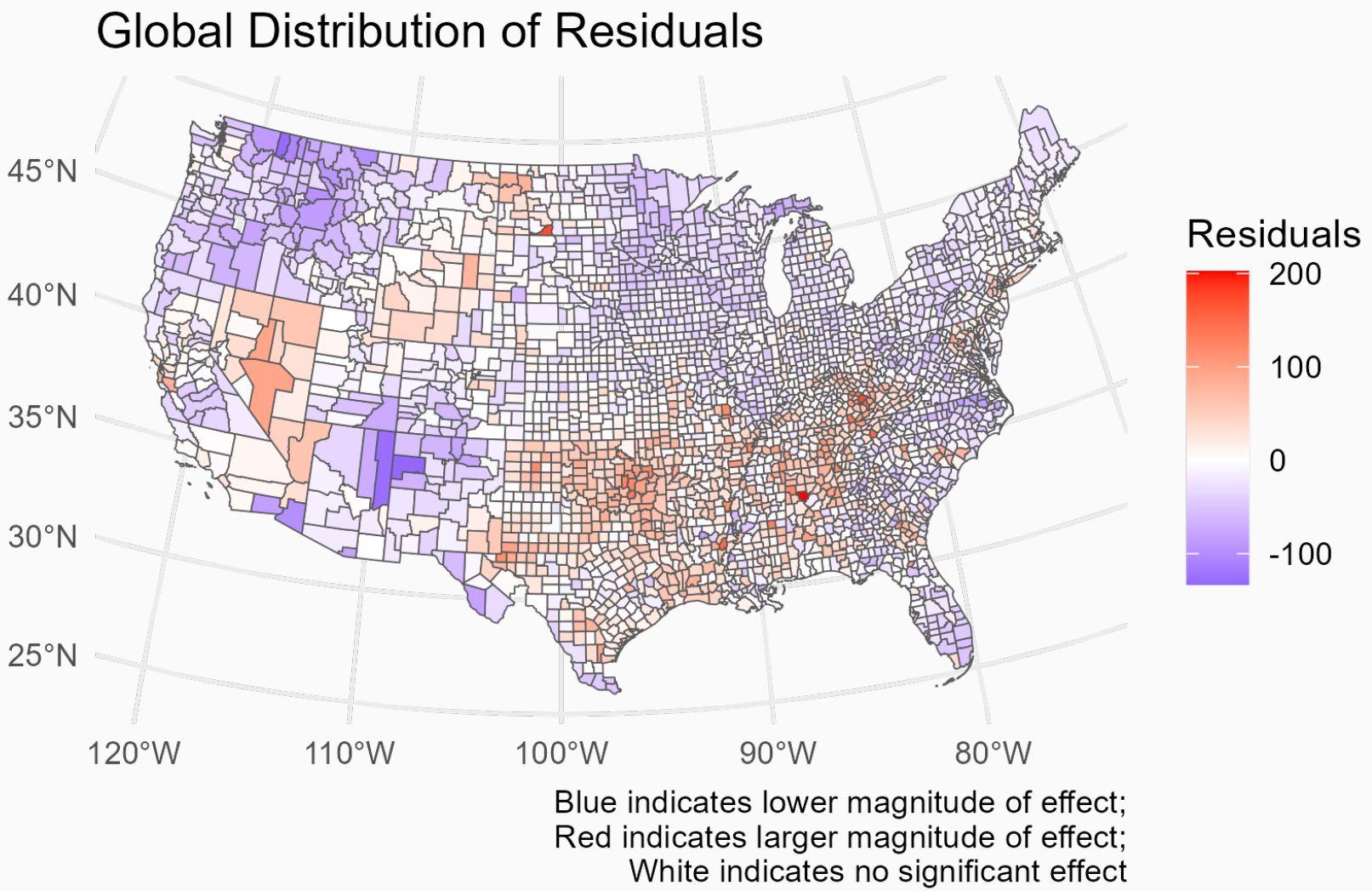


**Figure 4:** Local variations are critical for understanding the nature of our

# Results

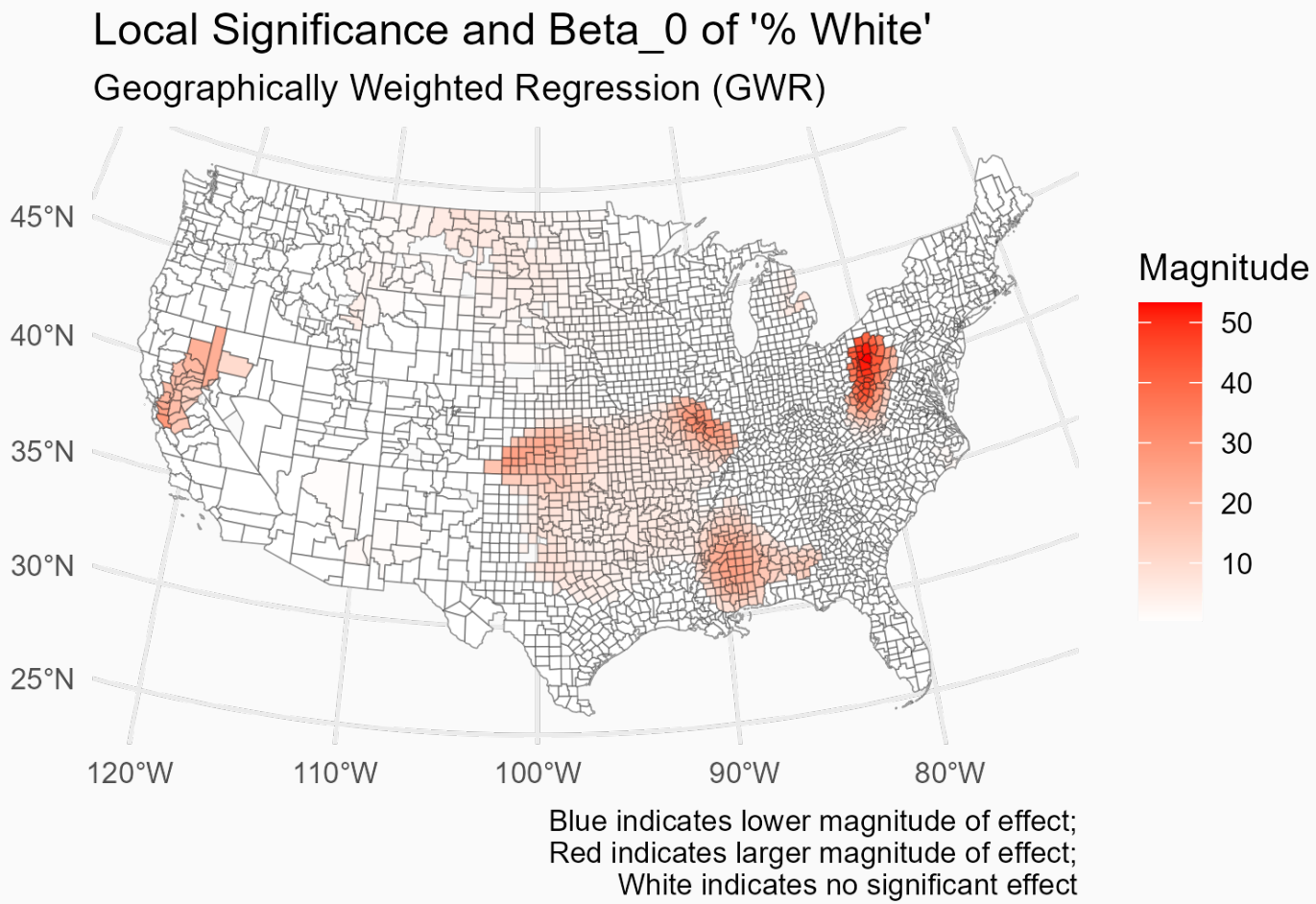


**Figure 5:** Significant negative correlation in central and southeastern regions. Positive correlation near New Mexico and Arizona.

# Results

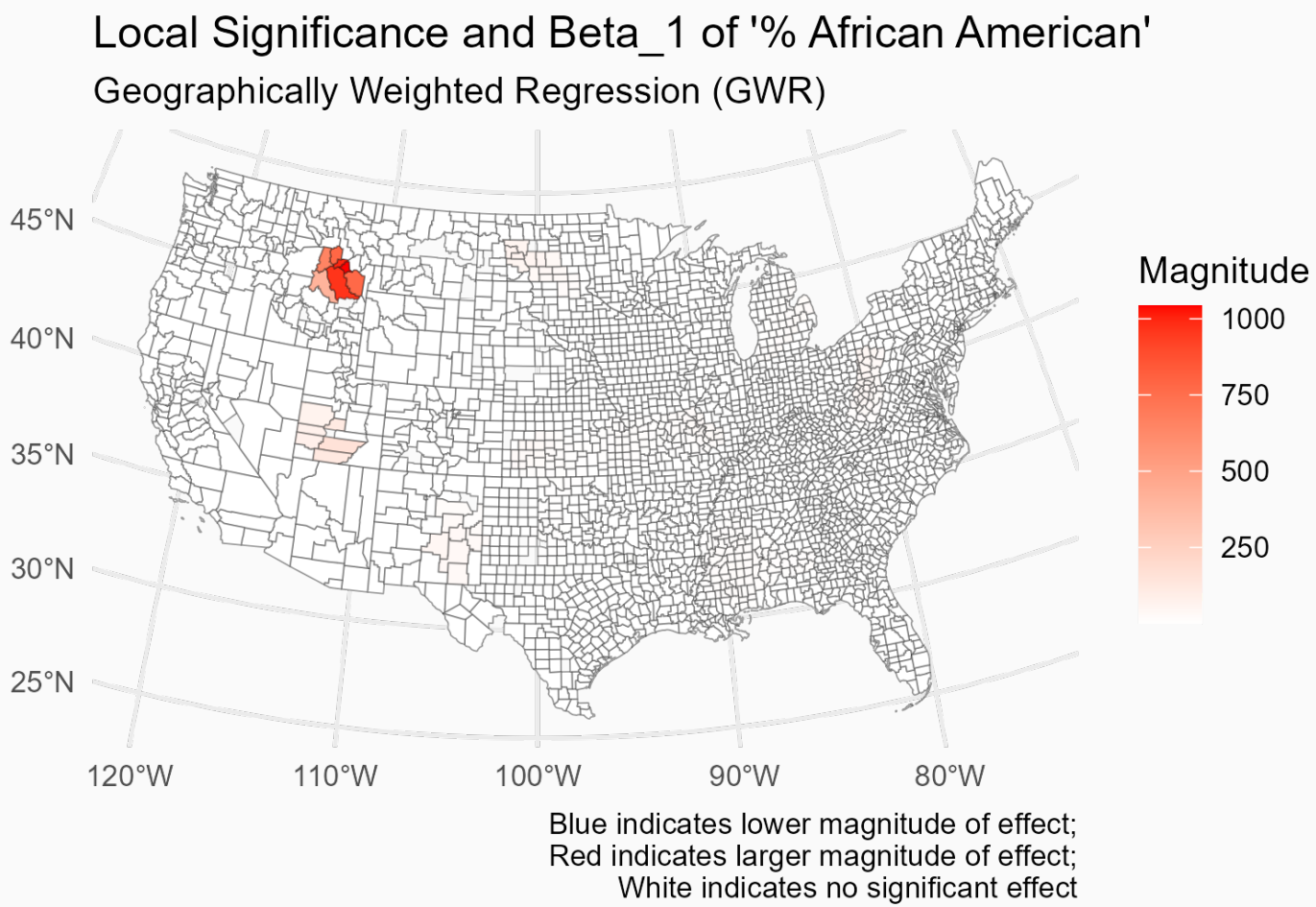


**Figure 6:** Significant positive correlation in northern central region. Significant negative correlation in central to southeastern areas.

# Results

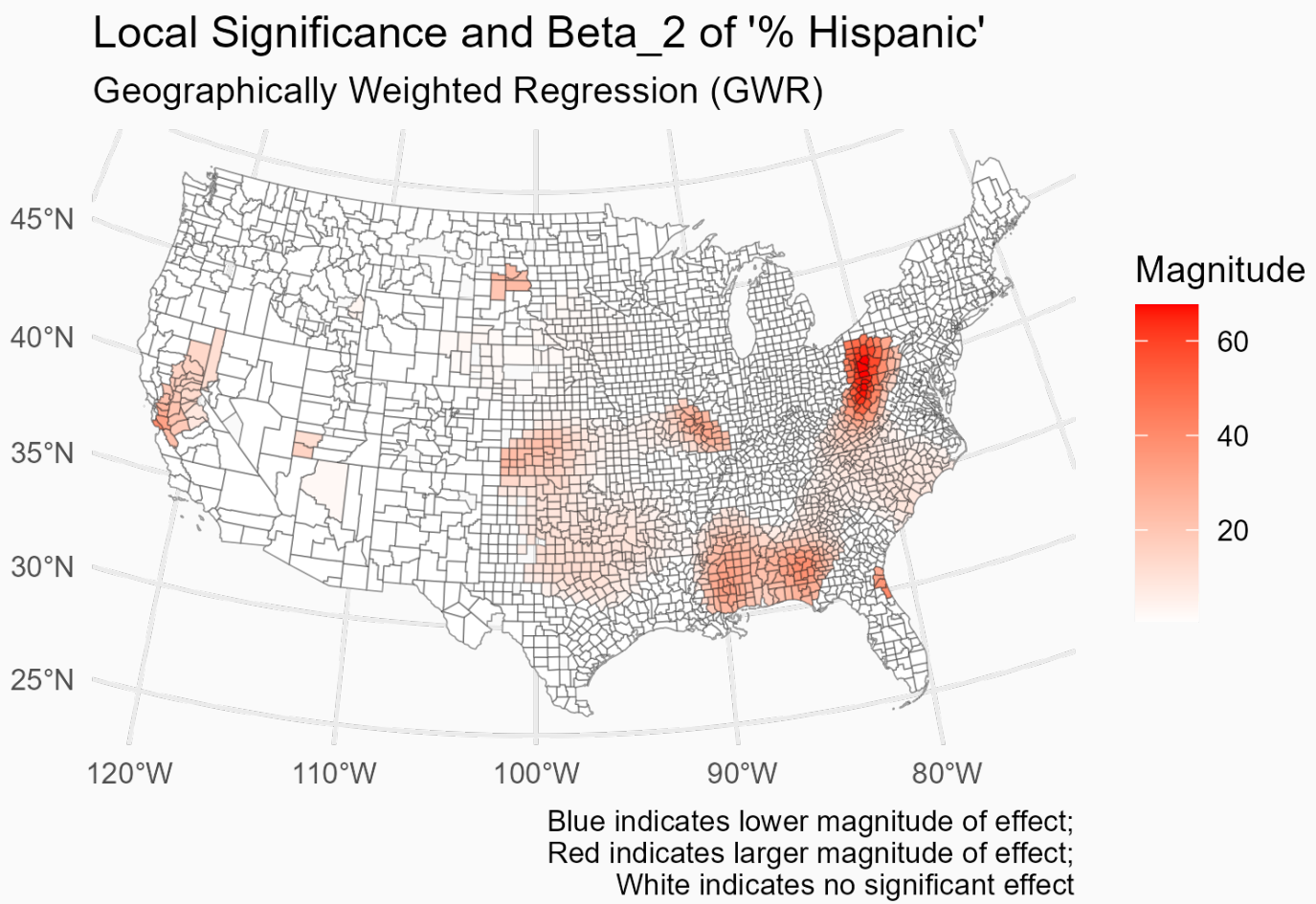


**Figure 7:** Significant negative correlation across central and southeastern regions. Isolated red patches in north-central suggest positive correlations.

# Results

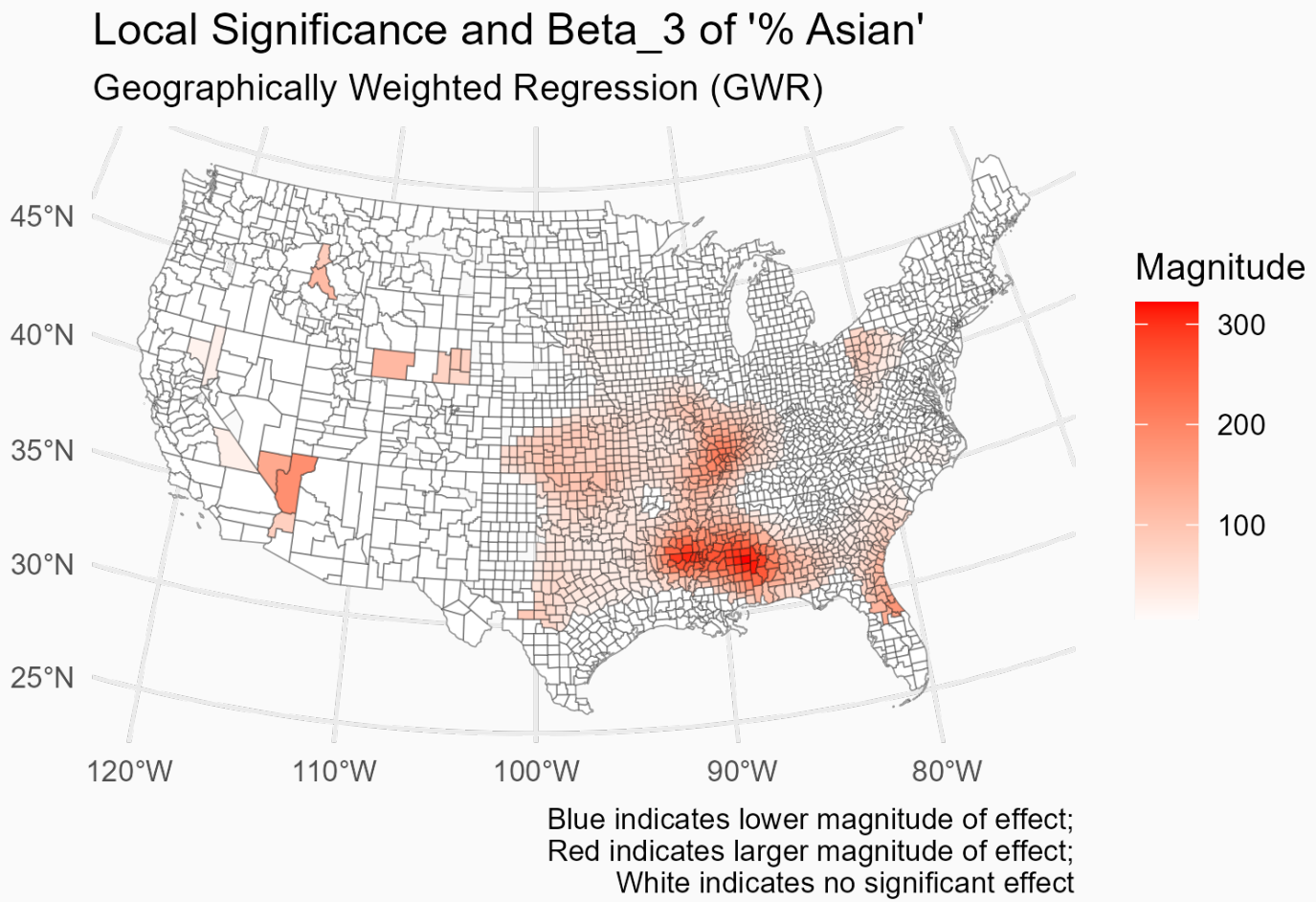


**Figure 8:** Generally, increased % Asian correlates with lower CVD outcomes in central to eastern regions. Some western regions show positive correlations.

# Results

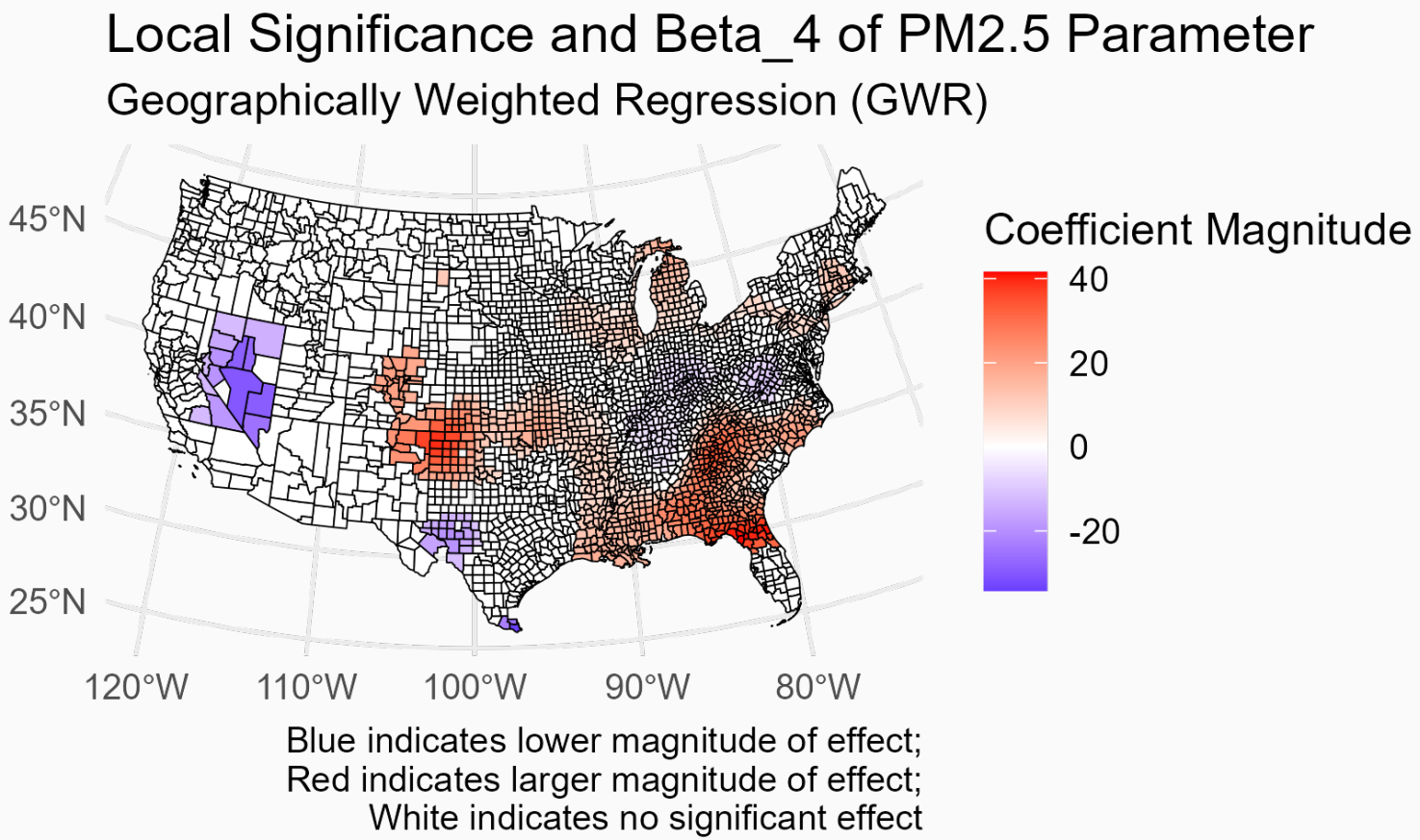


**Figure 9:** Higher levels in southern and northeastern regions correlate with higher CVD rates. Negative correlations in some western regions.

# Results

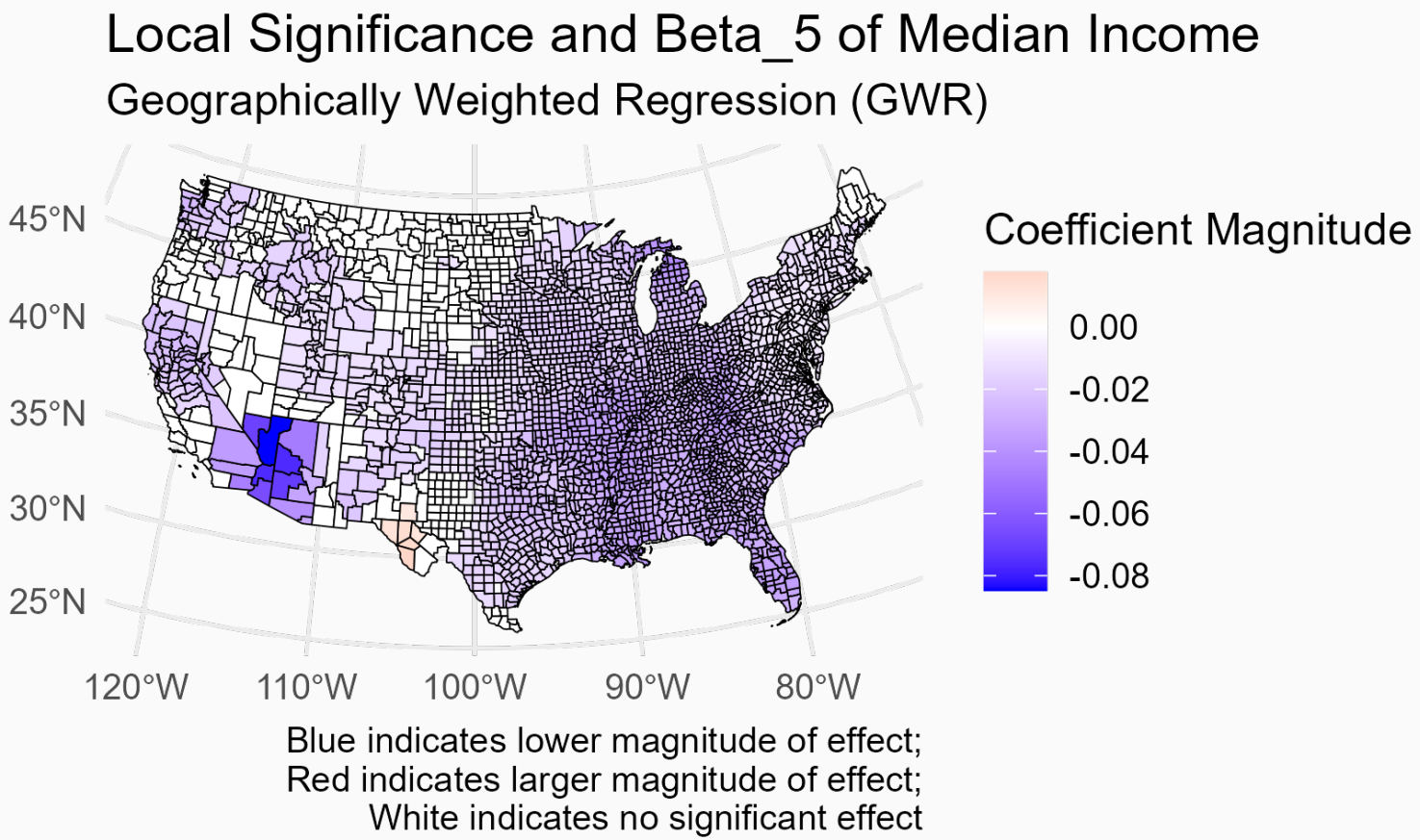


**Figure 10:** Generally, higher median income correlates with lower CVD outcomes. Notable exception in parts of Texas with a positive correlation.

# Results

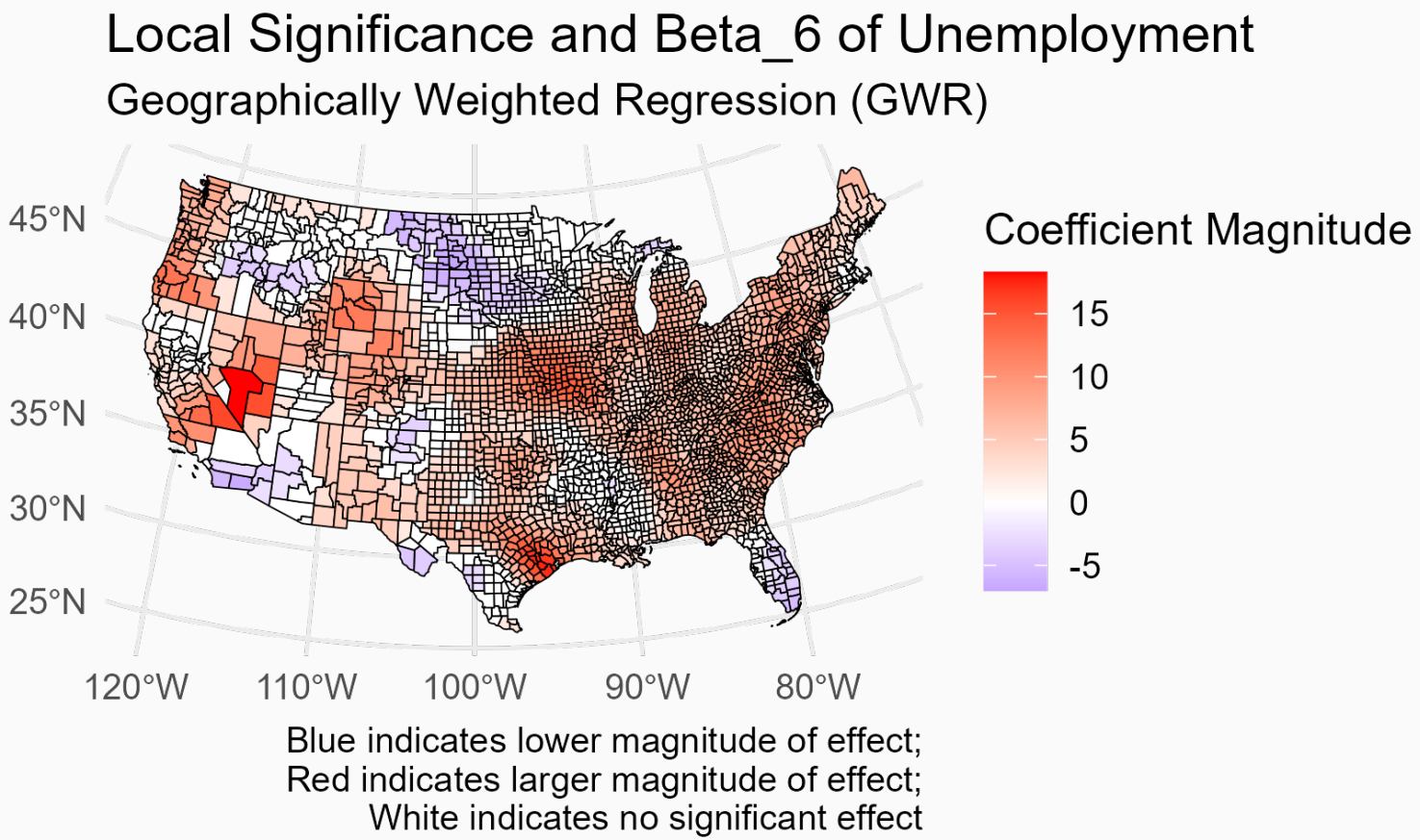


**Figure 11:** Mostly positive correlation with CVD rates in central regions. Negative correlation in southwestern regions.

# Discussion Section

# Discussion

- The purpose of the study was to fill a gap in CVD research that primarily focused on the stroke belt and adults over 55, with the goal of answering: what are the socioeconomic and environmental factors affecting CVD rates in the U.S.?

**Figure 12:** Urban smog in Chinese city

# Discussion

- The prior maps showed that factors are highly localized for most of our covariates
- Expected exception for median income, which has a constant effect regardless of region

# Discussion

- We observed localized impacts for African Americans, Hispanics, and Asians in different areas of the country
- Limitation in race percentages presented by misreporting of medical records affecting minorities (Tabb et al. 2020). Shows that intervention is needed to combat racial health disparities that were caused by disparities in the healthcare system.

# Variance Inflation Factor (VIF)

- Looking at the Variance Inflation Factor (VIF) of our GWR model:

| %WT | %BLK | %HISP | %AS | PM2.5 | Income | Unemp. |
|---|---|---|---|---|---|---|
| 183.00 | 196.00 | 7.00 | 3.00 | 1.00 | 2.00 | 3.00 |
| 81.00 | 94.00 | 4.00 | 3.00 | 2.00 | 3.00 | 3.00 |
| 302.00 | 314.00 | 17.00 | 4.00 | 1.00 | 2.00 | 2.00 |
| 141.00 | 151.00 | 5.00 | 2.00 | 1.00 | 2.00 | 3.00 |
| 132.00 | 128.00 | 8.00 | 3.00 | 1.00 | 2.00 | 2.00 |
| 283.00 | 294.00 | 14.00 | 4.00 | 1.00 | 2.00 | 2.00 |

We can see that the %WT and %BLK columns have disproportionately high values. This represents a limitation in the model as it suggests multicollinearity.

# Variance Inflation Factor

This is, however, to be expected given that the racial variables we are using sum up to one. An adjustment to the model could be made in a future study to eliminate the % White variable and aggregate the other three races into a single % Non-White variable.

# Discussion

- PM2.5 concentrations in the central and southeastern regions of the United States suggest that region-specific intervention may be needed.

**Figure 13:** Particulate smog in Houston, Texas

# Conclusion

# Conclusion

- We have shown that a GWR model can be used to analyze the relationship between CVM and socioeconomic covariates.
- The differing local significance that we detected highlights the need to identify region-specific interventions to curb the toll of CVD.