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

Shombit Roy, Christina Kim, Johan Booc

2024-04-27

Introduction

Literature Review Section

Literature Review

- 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 from the Census and TIGER Bureau, median income from Census API, air quality from NASA PM2.5 data, and unemployment rates loaded from CDC.
- Grouped data by year, 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

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

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.

Simulation

- Simulated data is used with a piecewise function dividing the geographical space into quadrants, each assigned different coefficients to model spatial heterogeneity
- A spatial autocorrelation test (Moran's I) is performed on the residuals of the GWR model

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

Picture of the Global dist of residuals

Eda plot

Results

% White demographic (Figure 3):

- Significant negative correlation in central and southeastern regions.
- Positive correlation near New Mexico and Arizona.

% African demographic (Figure 4):

- Significant positive correlation in northern central regions.
- Significant negative correlation in central to southeastern areas.

Results

% Hispanic demographic (Figure 5):

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

% Asian demographic (Figure 6):

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

Results

PM2.5 air quality (Figure 7):

- Higher levels in southern and northeastern regions correlate with higher CVD rates.
- Negative correlations in some western regions.

Results

Median income (Figure 8):

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

Unemployment (Figure 9):

- 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.?

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.

Discussion

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

1	2	3	4	5	6	7
183.22	196.12	7.24	2.56	1.15	2.44	2.62
81.32	93.80	4.14	2.53	1.77	3.39	2.65
302.16	313.78	16.89	4.06	1.27	2.42	2.09
140.83	151.46	5.25	2.13	1.10	2.35	2.73
131.77	128.39	8.33	2.67	1.07	2.34	2.32
282.66	294.23	14.14	3.77	1.24	2.44	2.22

Discussion

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-speific interventions to curb the toll of CVD.