Methods: Mind the Gap
Webinar Series

Exploratory and Inferential Spatial Statistical Methods: Tools To Understand the Geography of Health Across the U.S.

Presented by:
Loni Philip Tabb, Ph.D.
Drexel University Dornsife School of Public Health

National Institutes of Health
Office of Disease Prevention

Land Acknowledgement

Drexel University is in Philadelphia, Pennsylvania, and physically exists on the ancestral lands of the Lenni-Lenape people. I acknowledge that the Lenape people are the original inhabitants of Delaware, New Jersey, Eastern Pennsylvania, and Southern New York.

I offer respect and gratitude to the Indigenous peoples' past, present and future, whose custodial responsibility we benefit from.

I also acknowledge the land stolen from these nations, whether through colonization, broken treaties, and/or forced removal, and their inherent right to self-determination.

Outline

History

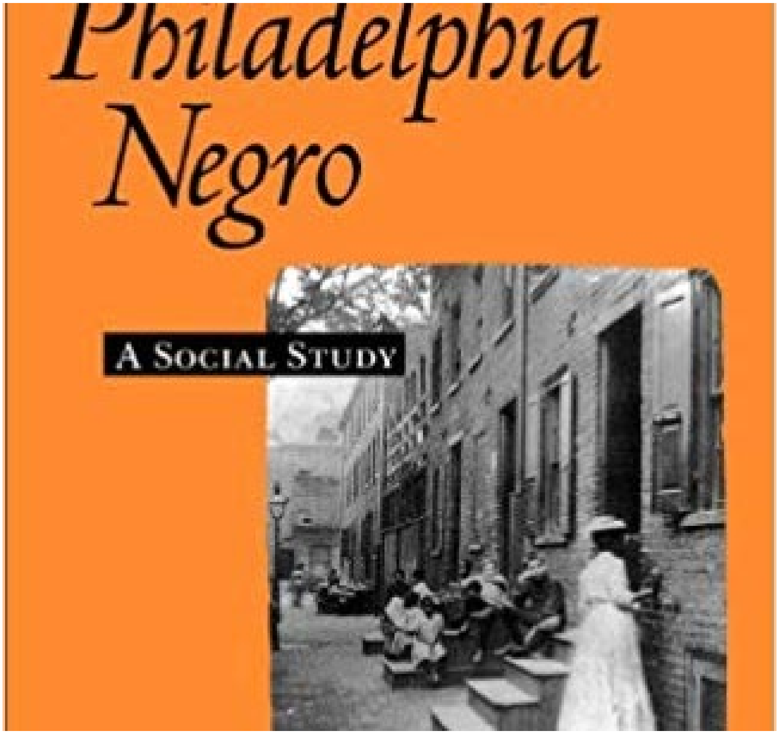
Motivation

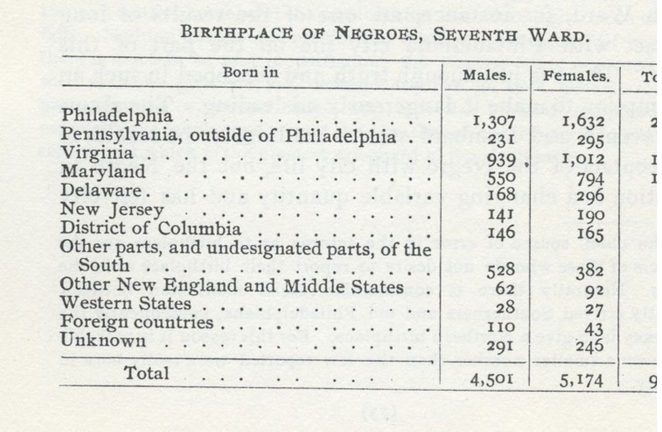
Exploratory Spatial Data Analysis
- Methods
- Empirical Examples

Inferential Spatial Data Analysis
- Methods
- Empirical Examples

Closing Thoughts…

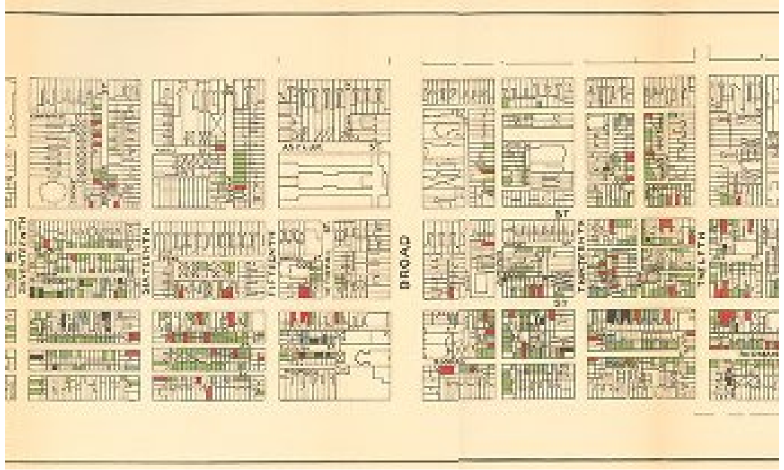
References





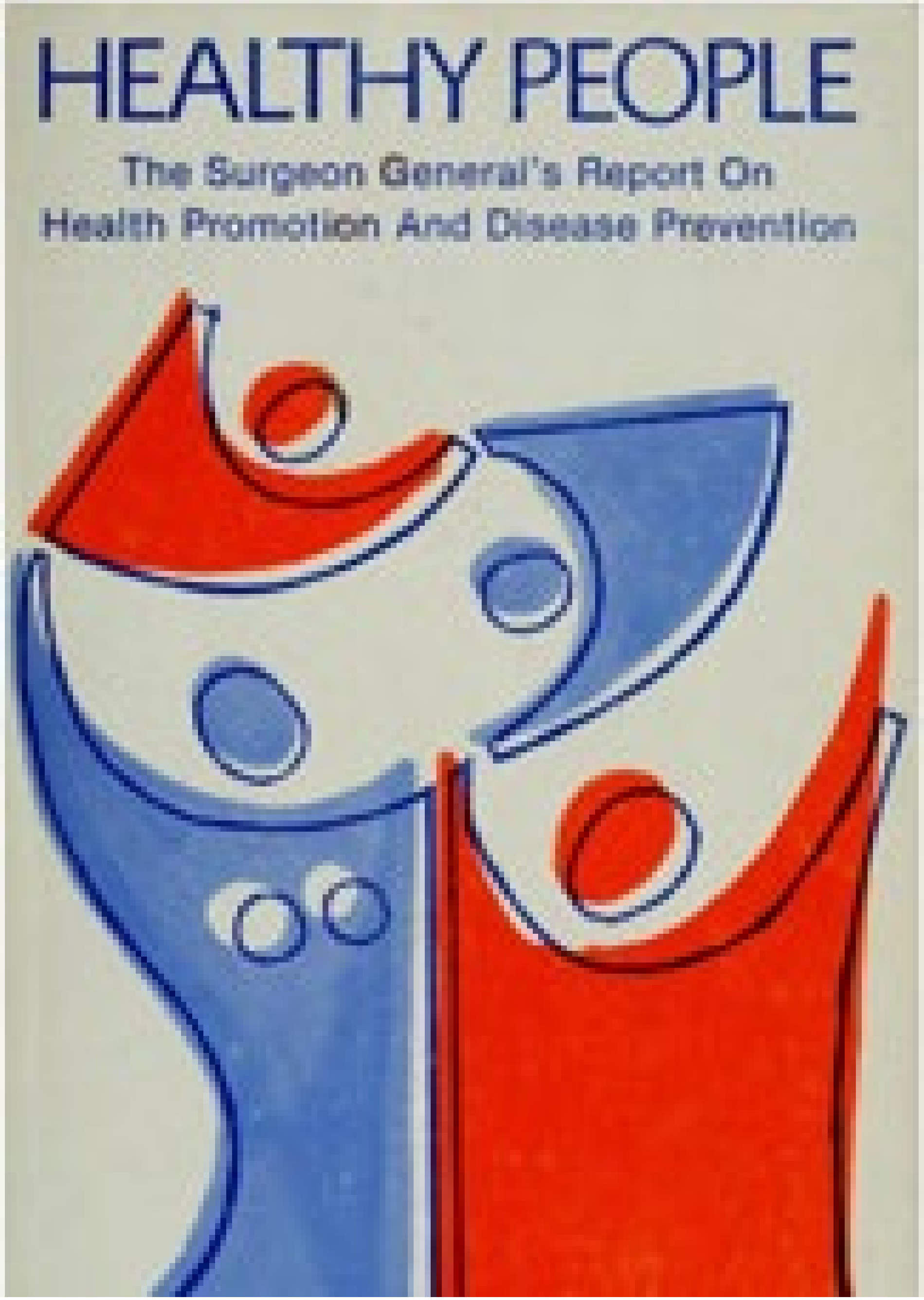
Birthplace of Negroes, Seventh Ward.

Born in	Males.	Females.
Philadelphia	1,307	1,632
Pennsylvania, outside of Philadelphia	231	295
Virginia	939	1,012
Maryland	550	794
Delaware	168	296
New Jersey	141	190
District of Columbia	146	165
Other parts, and undesignated parts, of the South	528	382
Other New England and Middle States	62	92
Western States	28	27
Foreign countries	110	43
Unknown	291	246
Total	4,501	5,174



History

IMAGE SOURCE: https://philadelphiaencyclopedia.org/essays/philadelphia-negro-the/



Healthy People Then…

Landmark report, issued by the Surgeon General

Birth of ODPHP's Healthy People 1990

- Measurable 10-year objectives for improving health and well-being nationwide
- Focus: decreasing deaths throughout the lifespan and increasing independence among older adults

And, then

- Healthy People 2000, 2010, 2020, and then…

Image Source: https://www.cdc.gov/nchs/images/healthy-people/HealthyPeople1979-250px.png?_=46082

Healthy People Now…

Builds on knowledge gained over the last 4 decades

Increased focused on:

- Health Equity
- Health Literacy
- New Focus: Well-Being
- Social Determinants of Health

Image Source: https://health.gov/our-work/national-health-initiatives/healthy-people

Motivation

SDOH = **S**ocial **D**eterminants **O**f **H**ealth

- Conditions in the environments where people are born, live, learn, work, play, worship, and age that affect a wide range of health, functioning, and quality-of-life outcomes and risks.
- 5 Domains:
  - Economic Stability
  - Education Access and Quality
  - Health Care Access and Quality
  - Neighborhood and Built Environment
  - Social and Community Context

Image Source: https://health.gov/healthypeople/priority-areas/social-determinants-health

But how…

How do we define a **neighborhood**?

How do we explore the resources and attributes in a **neighborhood**?

How do we tie the **neighborhood** to the health outcomes and/or exposures we are interested in measuring?

Exploratory Spatial Data Analysis (ESDA) Methods

ESDA Methods Introductions

- Extension of exploratory data analysis, with a focus on geographical data 3 - Tobler's First Law
- Common geographic information system (GIS) based technique
- Goals:
  - Describe and visualize spatial distributions
  - Identify atypical locations, i.e., spatial outliers
  - Discover patterns of spatial association, clusters, or hot spots
  - Assess (subjectively, objectively) spatial heterogeneity
- Software
  - GIS based software: ArcGIS (ESRI) , MapInfo , GeoDa
  - Open-source software: QGIS[7], R[8]
  - Additional: web-based mapping platforms (Google Maps), graphic designing software (Adobe Illustrator)

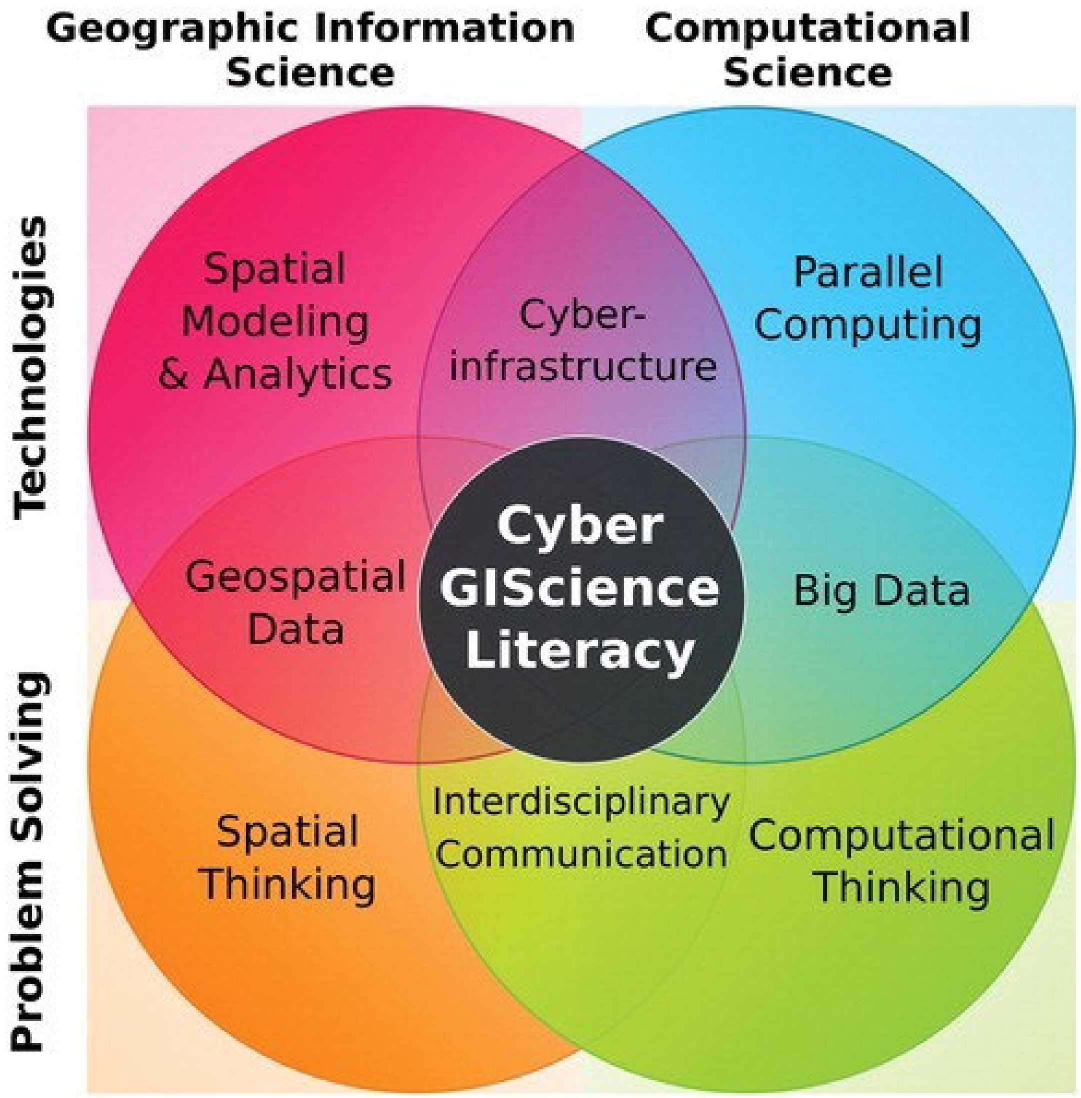
ESDA Methods
1: Choropleth Maps

What are they?

- Thematic maps that represent data by shading or coloring geographical areas of interest (i.e., census tracts, census block groups, health services areas, states,…)
- Allows for a subjective assessment of geographic patterning

What types of data are captured in them?

- Health and social outcomes at an aggregate level
- Health and social outcomes at a point level



ESDA Methods
Choropleth Maps (cont'd)

How do you create them? And what are the fundamental principles here?

- Geo-spatial data literacy
  - Foundational
  - Ability to judge the reliability of spatial data
  - Create ingenuous maps (not at the expense of good and easy-to-read maps)
- Data representative of: (1) phenomena of interest, (2) shapes/regions that capture patterns of phenomena

Image Reference: https://blogs.agu.org/onthejob/2019/02/01/cyber-literacy-for-21st-century-geospatial-professionals/

ESDA Methods
2: Moran's I Statistics/Tests

What are they?

- Statistical measure to assess spatial autocorrelation
  - Spatial Autocorrelation = Degree To Which Phenomena Values are Correlated in Space with Neighboring Locations
- Objective complement to Choropleth Maps

What types of questions can be answered with them?

- Are the high/large values of the phenomena in one region surrounded by other high/large values in a neighboring region?
- Are the high/large values of the phenomena in one region surrounded by other low/small values in a neighboring region?

ESDA Methods
2: Moran's I Statistics/Tests (cont'd)

$$I = \left(\frac{1}{s^2}\right) \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(Y_i - \overline{Y})(Y_j - \overline{Y})}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}},$$

$$s^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \overline{Y})^2.$$

How do you calculate it?

i, j: regions i and j

s^2: sample variance observed in the Y_i

w_{ij}: spatial weights⋆ between regions i and j

N: total regions being considered (i.e., mapped)

How do you judge the significance of it?

- Often, Normality based

How do you interpret it?

- Support: [-1, +1]
- -1: negative spatial correlation (inverse relationship)
- 0: no spatial correlation
- +1: positive spatial correlation

⋆ *How weights are defined will be dependent on neighborhood classification (i.e., contiguity, distance)*

ESDA Methods
3: Local Indicators of Spatial Analysis (LISA) Maps

What are they?

- Methods to identify localized regions in a map that are strongly positive or strongly negative.

Mapping Patterns:

- High-High: regions with high values surrounded by regions with also high values
- Low-Low: regions with low values surrounded by regions with also low values
- High-Low: regions with high values surrounded by regions with low values (i.e. spatial outlier = hot spot)
- Low-High: regions with low values surrounded by regions with high values (i.e. spatial outlier = cold spot)

What types of questions can be answered with them?

- Which locations/regions are making a meaningful (and significant) contribution to the global spatial patterning in the phenomena being considered?

ESDA Methods
#1 Empirical Example: Marijuana Access

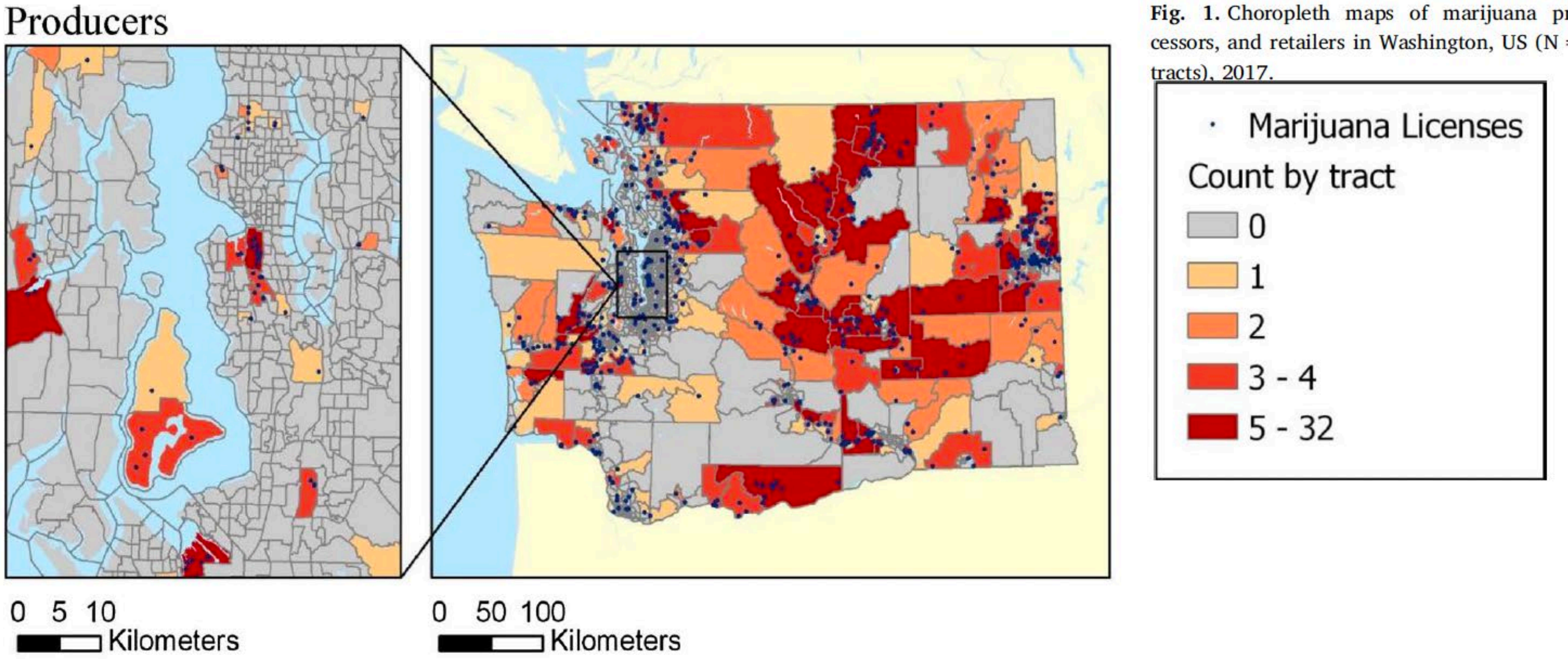
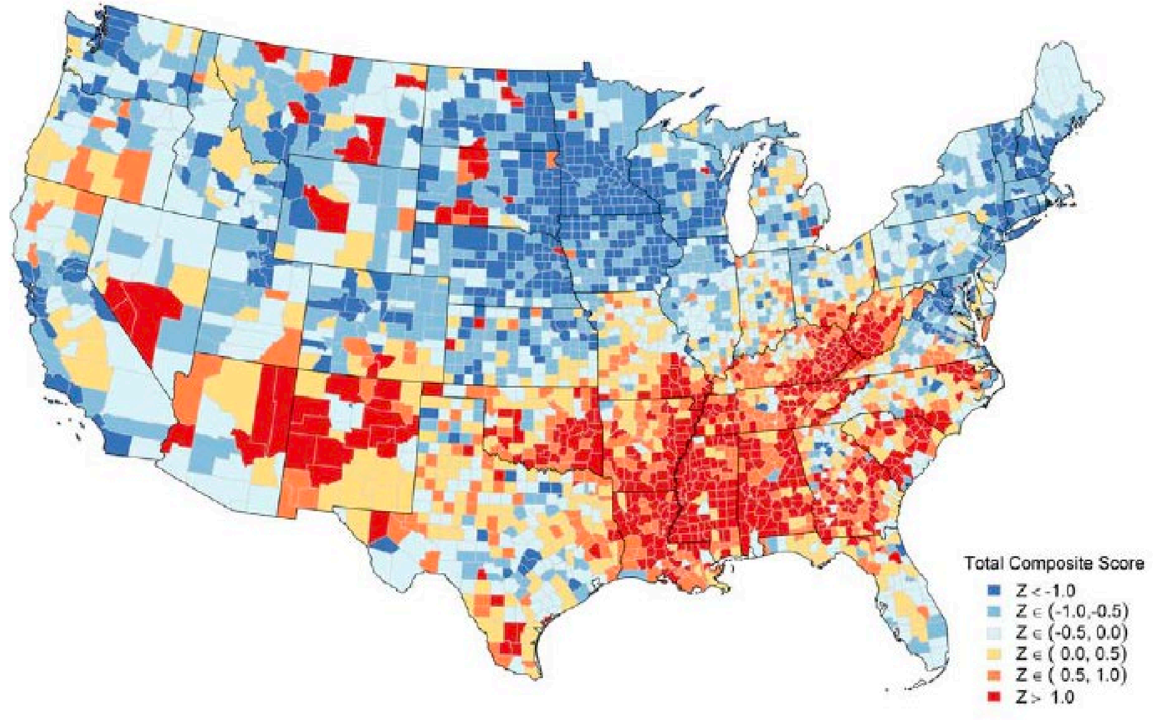


Fig. 1. Choropleth maps of marijuana producers, processors, and retailers in Washington, US (N = 1458 census tracts), 2017.

Tabb, L. P., Fillmore, C., & Melly, S. (2018). Location, location, location: assessing the spatial patterning between marijuana licenses, alcohol outlets and neighborhood characteristics within Washington state. *Drug and alcohol dependence*, *185*, 214-218.

Moran's I = 0.678, p = 0.01



(A) Health Outcomes

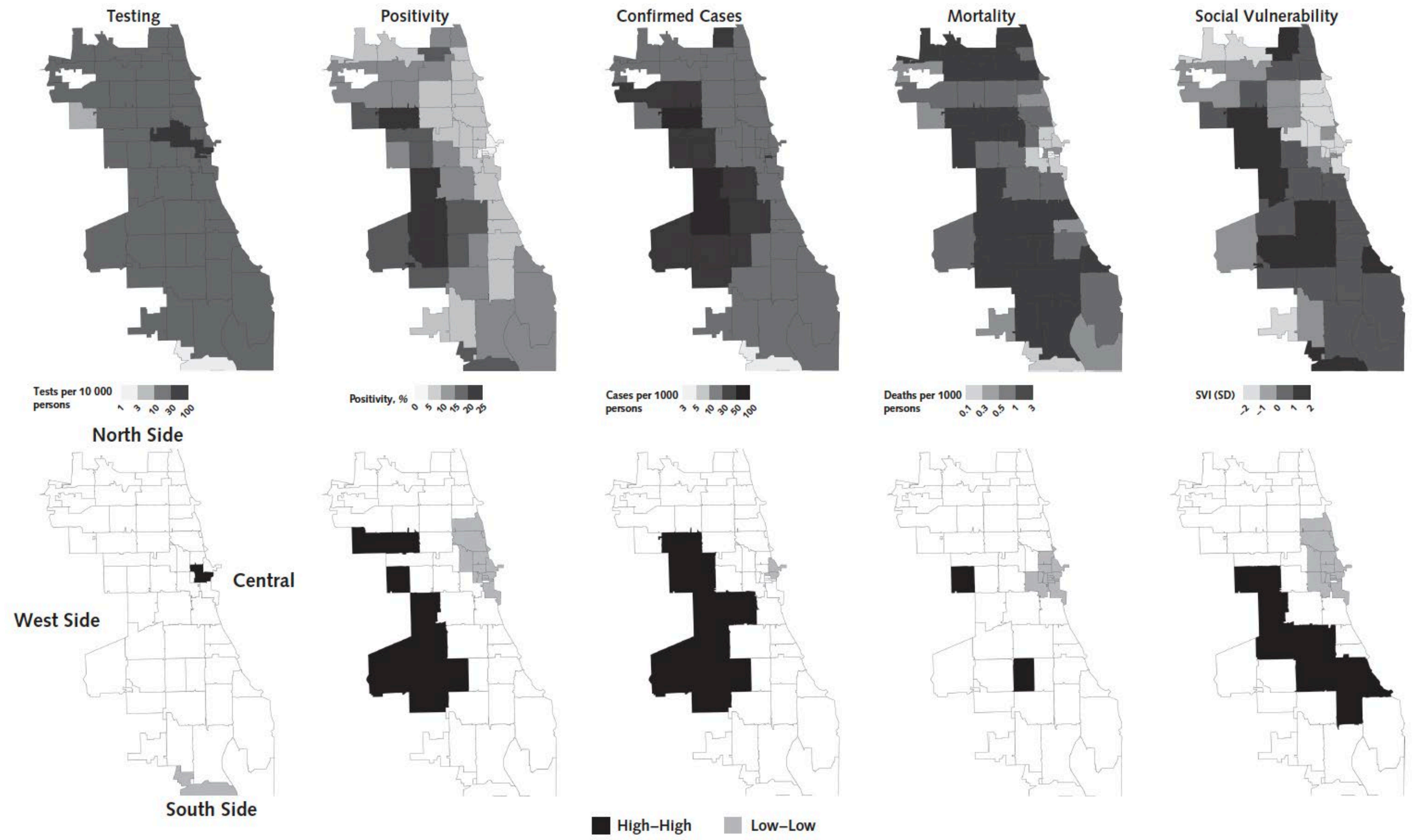
ESDA Methods #2 Empirical Example: Overall Health

Figure 1. Choropleth maps of composite scores for: (A) health outcomes, (B) health behaviors, (C) clinical care, (D) social and economic, and (E) physical environment for the US, 2017 (N = 3108 counties). Larger composite scores imply poorer health outcomes or health factors.
Not shown, maps (B), (C), (D), and (E).

Tabb, L. P., McClure, L. A., Quick, H., Purtle, J., & Roux, A. V. D. (2018). Assessing the spatial heterogeneity in overall health across the United States using spatial regression methods: the contribution of health factors and county-level demographics. *Health & place*, *51*, 68-77.

ESDA Methods #3 Empirical Example: COVID-19

Figure 1. Spatial distribution and clusters of coronavirus disease 2019 testing, positivity, confirmed cases, and mortality and social vulnerability in ZIP code tabulation areas of Chicago.



Clusters were calculated by using the local Moran *I* statistic; clusters have a *P* value < 0.05. "High-high" indicates hot spots and "low-low" indicates cold spots. SVI = Social Vulnerability Index.

Bilal, U., Tabb, L. P., Barber, S., & Diez Roux, A. V. (2021). Spatial inequities in COVID-19 testing, positivity, confirmed cases, and mortality in 3 US cities: an ecological study. *Annals of internal medicine*, *174*(7), 936-944.

Inferential Spatial Data Analysis Methods

Inferential Spatial Data Analysis Methods Introductions

- Regression based approaches to formally assess geographic based associations (and predictions, if applicable)
- Goals:
  - Test hypotheses of geographic based associations
  - Identify trends in the presence of spatial heterogeneity
- Software
  - GIS based software: ArcGIS (ESRI)5, MapInfo6 , GeoDa7
  - Open-source software: QGIS8, R9
  - Additional: web-based mapping platforms (Google Maps), graphic designing software (Adobe Illustrator)

Inferential Spatial Data Analysis Methods
Spatial Error Models

$$y = X\beta + s^*$$
$$s^* = \lambda W s + \varepsilon$$

Where:

- y: vector of dependent measures
- X: (design) matrix of independent measures
- β: vector of regression coefficients
- λ: spatial autocorrelation effect (scalar)
- W: neighborhood matrix (i.e., spatial weights)
- s : vector of spatially autocorrelated errors
- ε: vector of spatially uncorrelated errors

What is it?

- Extension of Ordinary Least Squares (OLS)
- Spatial autocorrelation is confined in the error term

Modelling assumptions

- Error term decomposed into spatially unstructured (i.e., random error) and spatially structured (i.e., autocorrelated error)

Interpretations

- Similar to OLS
- Global relationship being captured

Inferential Spatial Data Analysis Methods
Spatial Lag Models

$$y = \rho W y + X\beta + \varepsilon$$

Where:

- $y, W, X, \beta, \varepsilon$: defined as before
- $\rho W y$: spatial lag effect
  - $W y$: weighted average of the neighboring dependent measures

What is it?

- Also known as the Spatial Autoregressive Model
- Spatial autocorrelation is confined in the error term

Modelling assumptions

- Spatially lagged outcome/dependent measure

Interpretations

- Outcome in each region is affected by outcome in a neighboring region
- Global relationship being captured, but spatial dependence being estimated as well
  - ρ: strong/weak spatial dependence

Inferential Spatial Data Analysis Methods
Spatial Durbin Models

$$y = \rho W y + X\beta + W X \alpha + \varepsilon$$

Where:

- $y, W, X, \beta, \varepsilon$: defined as before
- $\rho W y$: spatial lag effect (outcome)
- $W X \alpha$: spatial lag effect (independent measures)

What is it?

- Extension of Spatial Lag model

Modelling assumptions

- Spatially lagged outcomes
- Spatially lagged independent measures

Interpretations

- Outcome in each region is affected by outcome in region and its neighboring regions
- Various relationships captured:
  - Indirect: across region comparisons
  - Direct: within region comparisons
  - Total: sum of indirect and direct
  - Spillover: magnitude of effect is the same in region and neighboring regions
  - Spatial dependence
    - ρ : strong/weak spatial dependence

Inferential Spatial Data Analysis Methods Geographically Weighted Regression (GWR)

$$y = X\beta(lat, long) + \varepsilon$$

Where:

- y, X, β, ε: defined as before
- $\beta(lat, long)$: estimated coefficients at location $(lat, long)$

What is it?

- Extension of traditional regression models, but at each location/geographic region

Modelling assumptions

- Relationships (between independent and dependent measures) vary locally

Interpretations

- Similar to OLS, except at each location

Inferential Spatial Data Analysis Methods Bayesian (Hierarchical) Regression Models

$$Y \sim \text{Likelihood}(\theta)$$

$$\theta \sim \text{Prior}(\phi)$$

$$\phi \sim \text{Hyperprior}$$

Where:

- Y: observed data, assumed to follow a distribution, based on parameter(s) θ
- θ: vector of parameters, assumed to follow a prior distribution
- ϕ: vector of hyperparameters, assumed to follow a hyperprior distribution

$$\text{Posterior}(\theta, \phi | Y) \propto \text{Likelihood}(\theta | Y) \times \text{Prior}(\phi) \times \text{Hyperprior}$$

What is it?

- Probabilistic based framework, instead of fixed/deterministic
- Allows for multilevel/hierarchically structured data
  - e.g., individuals nested within neighborhoods, and neighborhoods nested within states

Modelling assumptions

- Relationships can vary globally and locally

Inferential Spatial Data Analysis Methods #1 Empirical Example

Health & Place
Volume 51, May 2018, Pages 68-77

Assessing the spatial heterogeneity in overall health across the United States using spatial regression methods: The contribution of health factors and county-level demographics

Loni Philip Tabb [a], Leslie A. McClure [a], Harrison Quick [a], Jonathan Purtle [b], Ana V. Diez Roux [a]

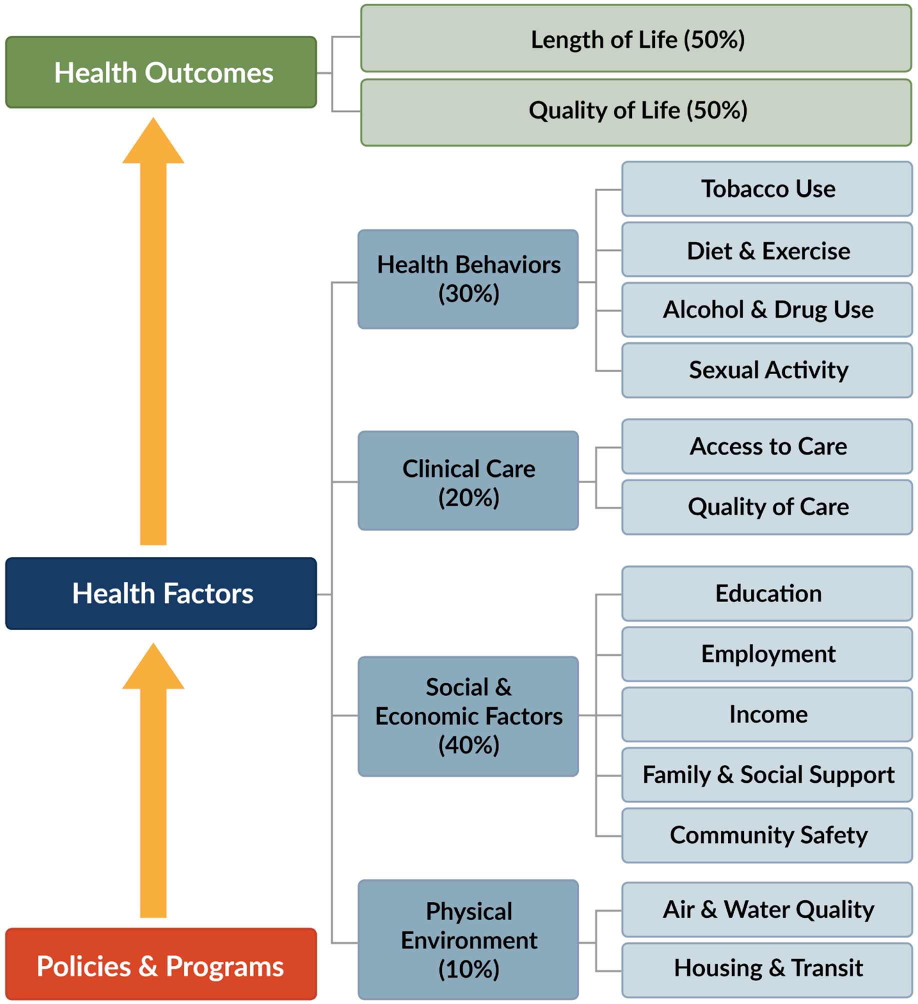
Show more

https://doi.org/10.1016/j.healthplace.2018.02.012

Study Aims

- How much spatial heterogeneity exists in health outcomes and health factors in counties across the US?
- To what extent is spatial heterogeneity in health outcomes attributable to spatial heterogeneity in health factors or to spillover effects (i.e., neighboring counties affecting each other)?

Inferential Spatial Data Analysis Methods
#1 Empirical Example



County Health Rankings model © 2014 UWPHI

Relationships of Interest

Health Outcomes = County Demographics

Health Outcomes = County Demographics + Health Factors

Where County Demographics include:
- *% White non-Hispanic*
- *% African American*
- *% Hispanic*
- *% Female*
- *% < 18 years*
- *% > 64 years*
- *% Living in rural area*
- *Population estimate*

Image Source: https://www.countyhealthrankings.org/sites/default/files/media/image/resources/CHRRmodel.png

Table 2
Mean differences in health outcome composite score per unit increase in independent variables for spatial error and spatial lag models, with direct, indirect, and total impacts for the spatial Durbin model, as well as spatial effects and model fit statistics for the US, 2017 (N = 3108 counties) [*].

	Spatial Error Estimate	SE	Spatial Lag Estimate	SE	Spatial Durbin Direct	Indirect	Total	Spillover
County Demographics								
% white non-Hispanic	**−0.110**	0.025	−0.033	0.021	**−0.131**	**0.283**	**0.152**	No
% African American	0.006	0.022	−0.006	0.017	−0.032	**0.127**	0.095	No
% Hispanic	**−0.135**	0.024	−0.023	0.019	**−0.156**	**0.293**	**0.137**	No
% female	**0.098**	0.010	**0.119**	0.010	**0.110**	**0.200**	**0.311**	Yes
% < 18 years	0.004	0.012	−0.014	0.011	−0.007	**−0.151**	**−0.158**	No
% > 64 years	**0.084**	0.013	0.023	0.012	**0.069**	**−0.278**	**−0.209**	No
% living in rural area	−0.021	0.011	−0.005	0.011	−0.004	**0.120**	**0.116**	No
Population estimate	−0.033	0.018	−0.016	0.017	**−0.028**	−0.057	**−0.084**	No
Health Factors								
Health behaviors	**0.314**	0.016	**0.268**	0.012	**0.313**	−0.025	**0.287**	No
Clinical care	**0.127**	0.013	**0.150**	0.011	**0.105**	**0.233**	**0.338**	Yes
Social and economic factors	**0.406**	0.017	**0.366**	0.013	**0.402**	**0.098**	**0.500**	Yes
Physical environment	−0.001	0.010	−0.016	0.009	−0.009	−0.021	−0.030	No
Spatial Effects								
λ (Spatial Error)	**0.66**							
ρ (Spatial Lag, Spatial Durbin)			**0.34**		**0.53**			
Fit Statistics								
R^2	0.85		0.85		0.86			
AIC	2865.10		3028.90		2659.90			
Moran's I	−0.05		**0.10**		−0.04			

[*] Bold font indicates statistically significant estimates, p-value < 0.05; SE = standard error; AIC = Akaike Information Criterion.

Inferential Spatial Data Analysis Methods
1 Empirical Example: Overall Health

Tabb, L. P., McClure, L. A., Quick, H., Purtle, J., & Roux, A. V. D. (2018). Assessing the spatial heterogeneity in overall health across the United States using spatial regression methods: the contribution of health factors and county-level demographics. *Health & place*, *51*, 68-77.

Inferential Spatial Data Analysis Methods #2: Empirical Example

Spatial and Spatio-temporal Epidemiology 40 (2022) 100473

Contents lists available at ScienceDirect

Spatial and Spatio-temporal Epidemiology

journal homepage: www.elsevier.com/locate/sste

ELSEVIER

Check for updates

Spatially varying racial inequities in cardiovascular health and the contribution of individual- and neighborhood-level characteristics across the United States: The REasons for geographic and racial differences in stroke (REGARDS) study

Loni Philip Tabb [a,*], Ana V. Diez Roux [a], Sharrelle Barber [a], Suzanne Judd [b], Gina Lovasi [a], Andrew Lawson [c], Leslie A. McClure [a]

[a] Department of Epidemiology and Biostatistics, Dornsife School of Public Health, Drexel University, 3215 Market Street, Philadelphia, PA 19104, United States of America

[b] School of Public Health, Department of Biostatistics, University of Alabama at Birmingham, United States of America

[c] College of Medicine, Medical University of South Carolina, United States of America

Study Aims

- What is the impact of both individual- and neighborhood-level risk factors on the ***spatially varying*** Black-White differences in CVH?
  - Global effects?
  - Locally varying effects?
- And how does structural racism play a role in these differences?

METHODS:
individual-level factors

Age (years)

Sex (male/female)

Income ($/year)

Education (# of years)

Marital Status (yes/no)

METHODS:
neighborhood-level factors

Neighborhood = census tract (CT)

Social Engagement Venues

Favorable Food Stores

Physical Activity Resources

Walkability

Stroke Belt Residency

Racialized Economic Segregation

- Index of Concentration at the Extremes (ICE)
  - Low-income Black vs. High-income White

Inferential Spatial Data Analysis Methods Empirical Example: Cardiovascular Health

Tabb, L. P., Roux, A. V. D., Barber, S., Judd, S., Lovasi, G., Lawson, A., & McClure, L. A. (2022). Spatially varying racial inequities in cardiovascular health and the contribution of individual-and neighborhood-level characteristics across the United States: The REasons for geographic and racial differences in stroke (REGARDS) study. *Spatial and Spatio-temporal Epidemiology*, *40*, 100473.

Table 2.

Mean differences in total CVH score associated with race and individual- and neighborhood-level covariates and variances of random components for all REGARDS participants.*

	Model 1 [a]				**Model 2** [b]				**Model 3** [c]			
	Mean	**SD**	**2.5%**	**97.5%**	**Mean**	**SD**	**2.5%**	**97.5%**	**Mean**	**SD**	**2.5%**	**97.5%**
Fixed Effects												
Individual-Level												
Race (black vs. White)	−0.97	0.05	−1.06	−0.87	−0.71	0.05	−0.81	−0.61	−0.54	0.05	−0.65	−0.44
Age (years)	0.01	0.02	−0.02	0.04	0.12	0.02	0.09	0.15	0.12	0.02	0.09	0.15
Sex (male vs. female)	0.09	0.03	0.03	0.15	−0.10	0.03	−0.16	−0.04	−0.09	0.03	−0.15	−0.03
Income (reference: < $20,000/year)												
$20,000–34,999					0.23	0.05	0.13	0.33	0.22	0.05	0.12	0.32
$35,000–74,999					0.49	0.05	0.39	0.60	0.46	0.05	0.36	0.56
> $75,000					0.83	0.06	0.70	0.95	0.76	0.06	0.64	0.88
Refused					0.52	0.06	0.41	0.64	0.50	0.06	0.38	0.61
Education (# of years)					0.28	0.02	0.24	0.31	0.27	0.02	0.23	0.30
Marital status (yes vs. no)					0.19	0.04	0.12	0.26	0.18	0.04	0.11	0.25
Neighborhood-Level												
Physical activity resources density									0.00	0.02	−0.03	0.04
Walkability density									0.01	0.03	−0.04	0.06
Favorable food stores density									0.02	0.02	−0.01	0.05
Social engagement density									−0.02	0.03	−0.07	0.04
Living in the Stroke Belt (yes vs. no)									0.07	0.16	−0.24	0.39
Index of concentration at the extremes									0.15	0.02	0.11	0.19
Random Effects												
Precision for the total CVH scores $\left(\tau = 1/\sigma^2\right)$	0.24	0.00	0.24	0.25	0.26	0.00	0.25	0.26	0.26	0.00	0.25	0.26
Precision for unstructured state-level random effects $\left(\tau_u = 1/\sigma^2_u\right)$	13.45	4.08	6.87	22.75	14.26	4.26	7.53	24.06	14.02	4.28	7.50	24.13
Precision for structured state-level random effects $\left(\tau_v = 1/\sigma^2_v\right)$	8.68	3.42	3.81	17.04	9.39	3.55	4.10	17.87	9.40	3.59	4.13	17.99
Precision for structured state-level varying race coefficients $\left(\tau_\delta = 1/\sigma^2_\delta\right)$	11.31	3.81	5.34	20.14	12.09	4.02	5.90	21.51	12.45	4.14	6.05	22.12

*REGARDS, Reasons for Geographic and Racial Differences in Stroke; CVH, cardiovascular health; SD = standard deviation; 2.5% and 97.5% are the lower and upper 95% credible interval limits, respectively.

a Model 1 includes race + age + sex.

b Model 2 is Model 1 + income + education + marital status.

c Model 3 is Model 2 + physical activity resources + walkability + favorable food stores + social engagement + living in the Stroke Belt + index of concentration at the extremes.

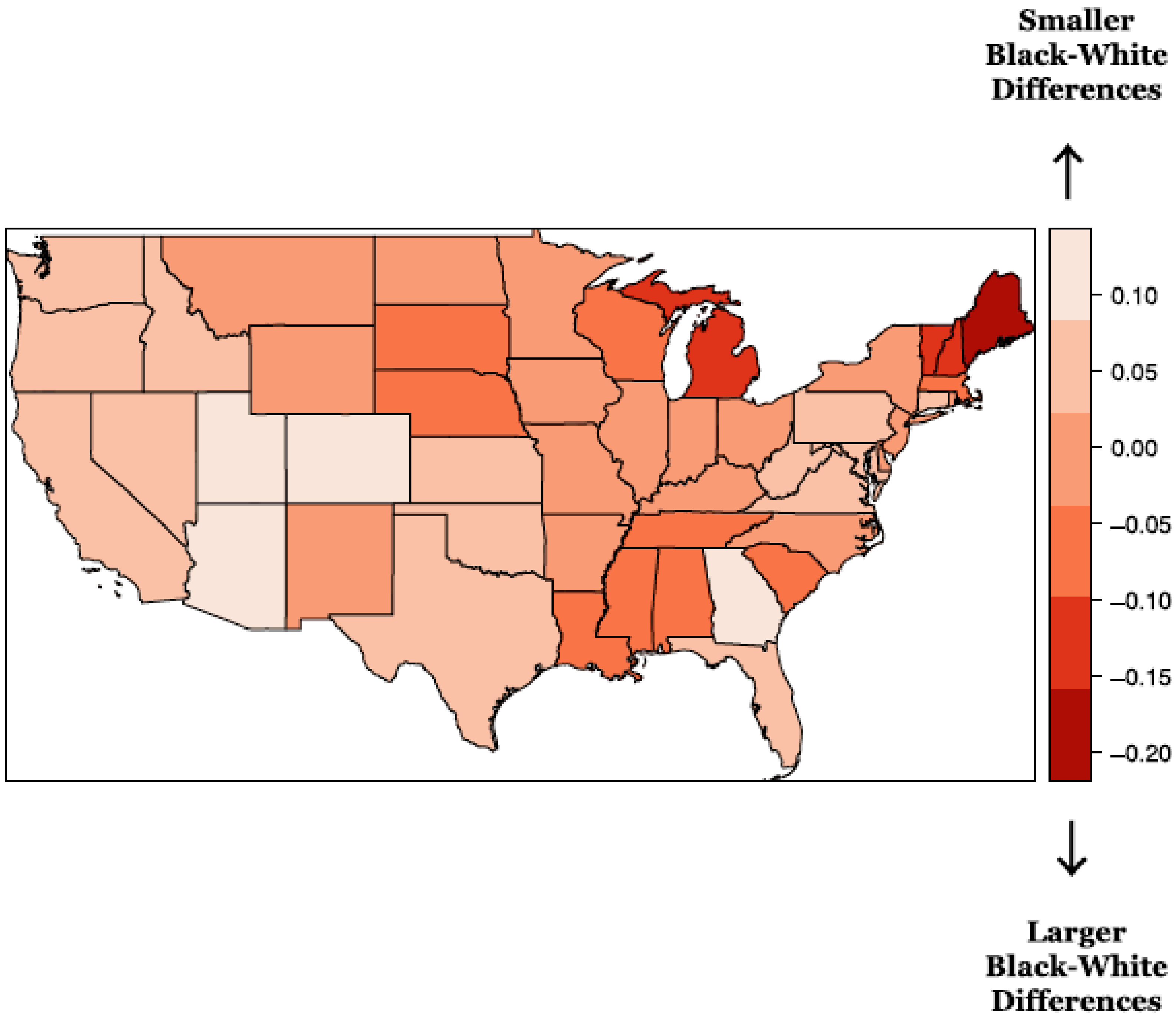


Fig. 3. Map of state level random effects from the fully adjusted spatially varying random coefficient (race) model for total CVH scores for all REGARDS participants, United States (US).* * REGARDS, Reasons for Geographic and Racial Differences in Stroke; CVH, cardiovascular health. Model includes intercept + race + age + sex + income + education + marital status + physical activity resources + walkability + favorable food stores + social engagement + living in the Stroke Belt + index of concentration at the extremes + race-varying state-level random effects.

Closing Thoughts…

History -> Now

Why Neighborhoods?

Exploratory Methods

Inferential Methods

What Next?

References

Provided as a separate document that accompanies these slides.

Questions?