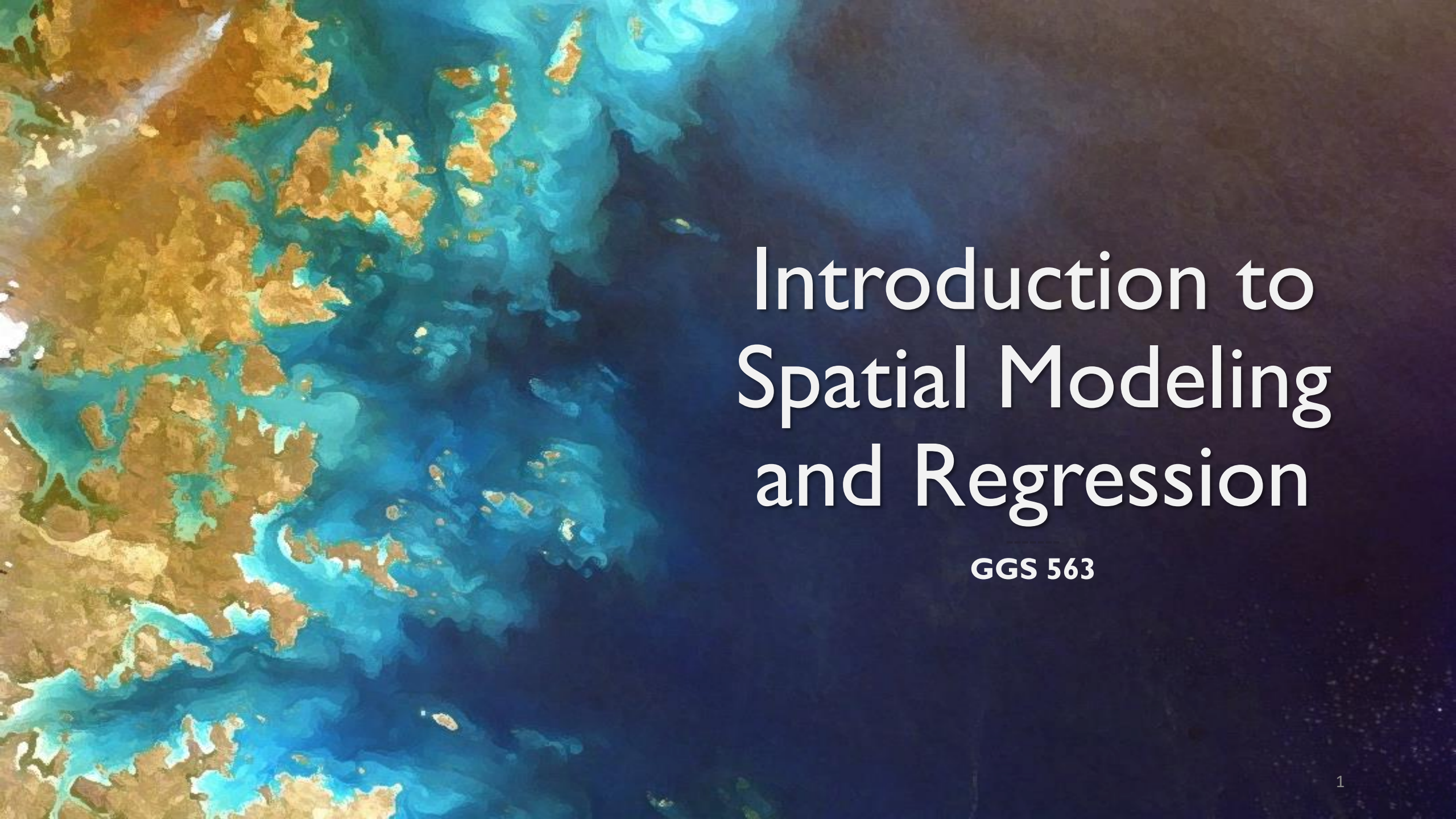


Introduction to Spatial Modeling and Regression

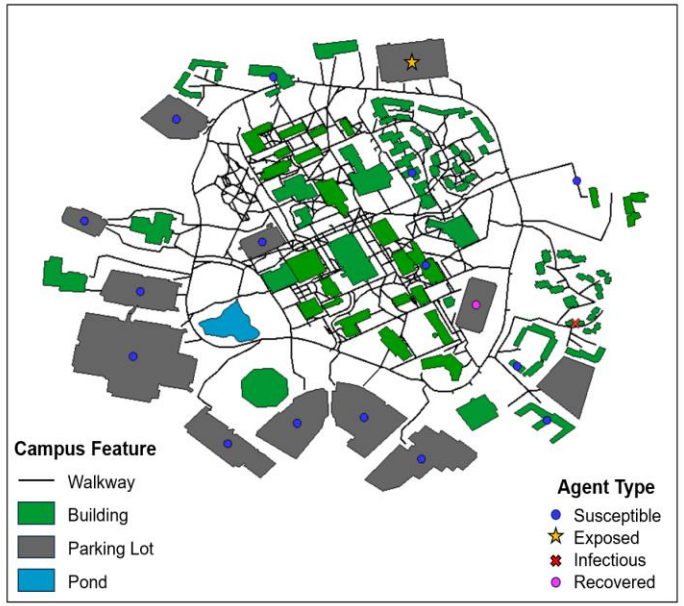
GGS 563

What is Spatial Modeling?

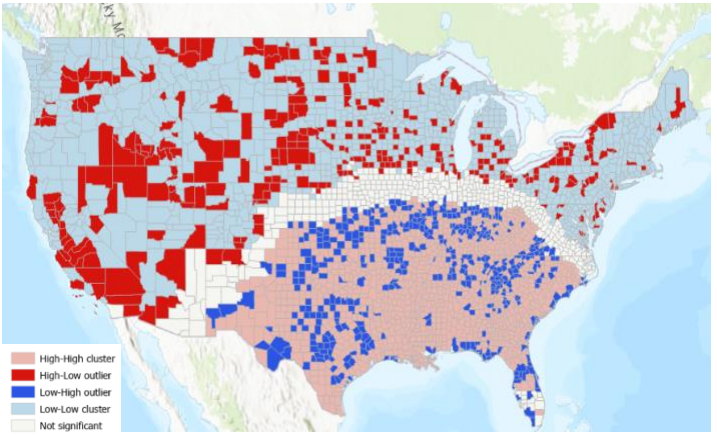
- **Spatial modeling** refers to a broad set of techniques used to represent, analyze, and simulate spatial phenomena
    - Spatial models capture spatial *patterns*, *relationships*, and *processes*

Types of Spatial Models:

- *Descriptive models:* Identifies spatial patterns
    - Clustering of TikTok users
- *Simulation models:* Represents processes over space and time
    - Agent-based modeling (ABMs) of COVID-19 spread
- *Predictive models:* Forecasts outcomes based on spatial variables
    - Identifying where flooding might occur
- *Geostatistical models:* Uses spatial interpolation
    - Kriging to estimate air pollution



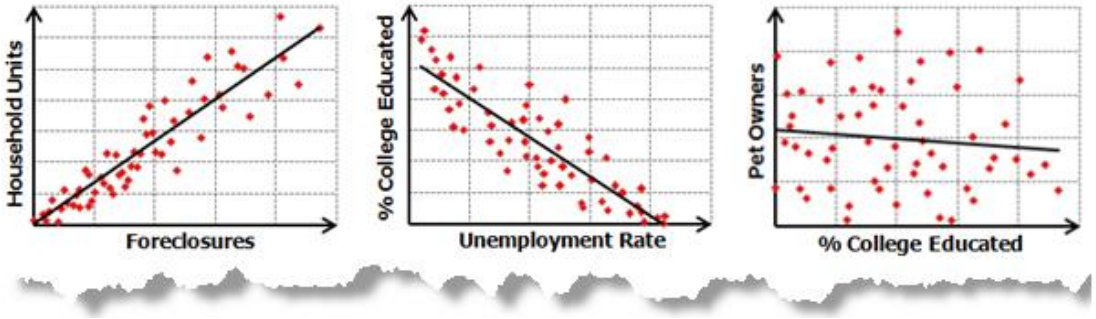
Spatial ABM of COVID-19 spread on GMU's campus.



Clusters of TikTok users across U.S. counties.

Traditional Regression vs. Spatial Regression

- **Traditional regression:** a statistical method used to describe the global relationship between a dependent variable and one or more independent variables
  - ***Ordinary Least Squares (OLS)***
  - What is the relationship between bachelors degree attainment and unemployment rate?



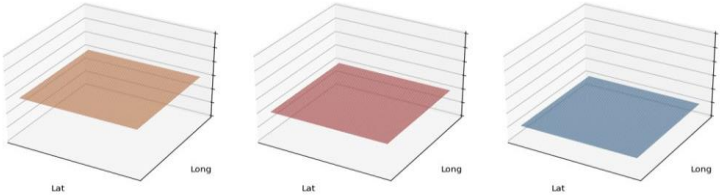
Global models explaining the relationship between certain variables, not accounting for spatial variation.

- Spatial regression considers **spatial non-stationarity**: the variation in relationships between certain variables over geographic space
  - How does the relationship between bachelors degree attainment and unemployment rate vary across US counties?

Varying coefficient surfaces for:

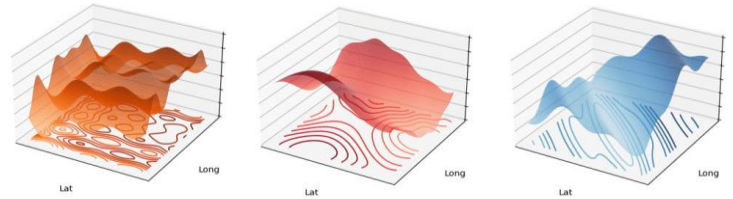
OLS

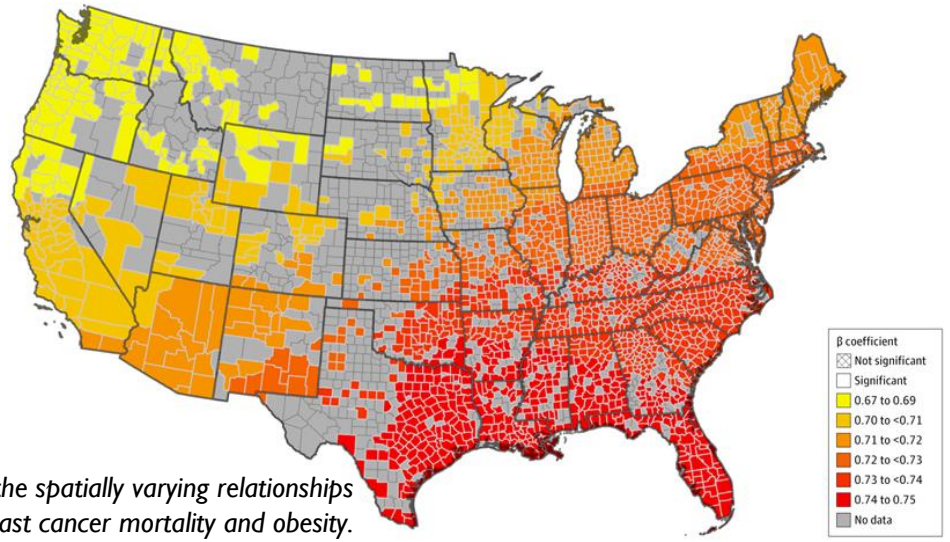
coefficients do not vary



GWR

coefficients change across space





Coefficients explaining the spatially varying relationships between breast cancer mortality and obesity.

Spatial Regression

- **Spatial regression** is a spatial modeling technique that helps us understand how different factors influence outcomes across space and time

- Takes into account that nearby places often influence each other
  - Tobler's first law of geography: *"Everything is related to everything else, but near things are more related than distant things."*

- Types of spatial regression:
  - *Spatial Lag Model (SLM)* – nearby outcomes influence the result
  - *Spatial Error Model (SEM)* – nearby errors affect the result
  - *Geographically Weighted Regression (GWR)* – relationships vary by location
  - *Multiscale Geographically Weighted Regression (MGWR)* – relationships vary by location and scale

Spatial Regression Applications

Public Health

- How do the relationships between county-level breast cancer mortality and population characteristics vary geographically in the US? [1]

Environment and Climate

- How do the associations between PM2. 5 concentrations and socioeconomic factors vary across China's major cities? [2]

Urban Planning

- What are the spatially varying relationships between landcover types and the urban climate (land surface temperature and moisture) in Southeast China? [3]

Transportation

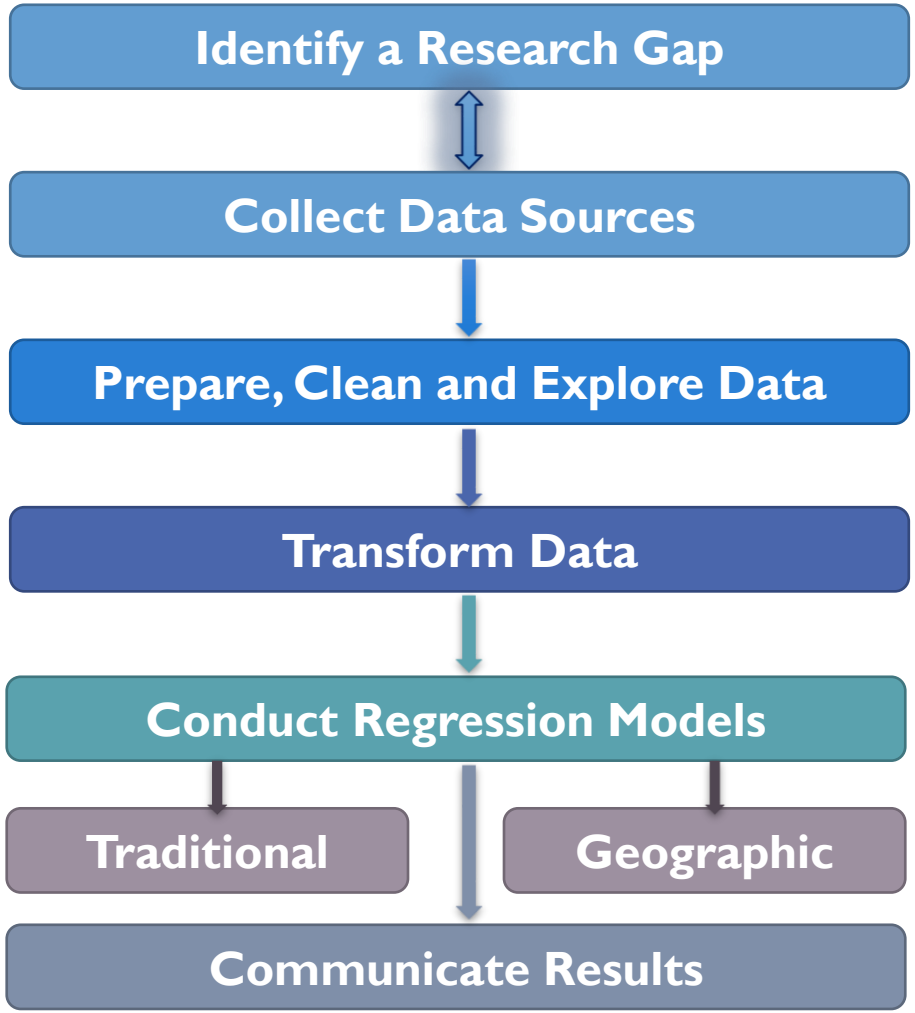
- What is the spatiotemporal influence of the built environment (e.g., sidewalk density, traffic volume) on transit ridership? [4]

Agriculture

- What is the spatially varying effects of mechanical and biochemical technologies, government payments, product diversity and farm size on farming viability in Ohio? [5]

1. *Anderson, T., Herrera, D., Mireku, F., Barner, K., Kokkinakis, A., Dao, H., ... & Pierobon, M. (2023). Geographical variation in social determinants of female breast cancer mortality across US counties. JAMA network open, 6(9), e2333618-e2333618.*
2. *Zhao, X., Zhou, W., Han, L., & Locke, D. (2019). Spatiotemporal variation in PM2. 5 concentrations and their relationship with socioeconomic factors in China's major cities. Environment international, 133, 105145.*
3. *Hu, X., & Xu, H. (2019). Spatial variability of urban climate in response to quantitative trait of land cover based on GWR model. Environmental Monitoring and Assessment, 191(3), 194.*
4. *Ma, X., Zhang, J., Ding, C., & Wang, Y. (2018). A geographically and temporally weighted regression model to explore the spatiotemporal influence of built environment on transit ridership. Computers, Environment and Urban Systems, 70, 113-124.*
5. *Jiang, Z., & Xu, B. (2014). Geographically weighted regression analysis of the spatially varying relationship between farming viability and contributing factors in Ohio. Regional Science Policy & Practice, 6(1), 69-84.*

Spatial Regression Project Workflow



Identify a Research Gap and Formulating a Research Plan

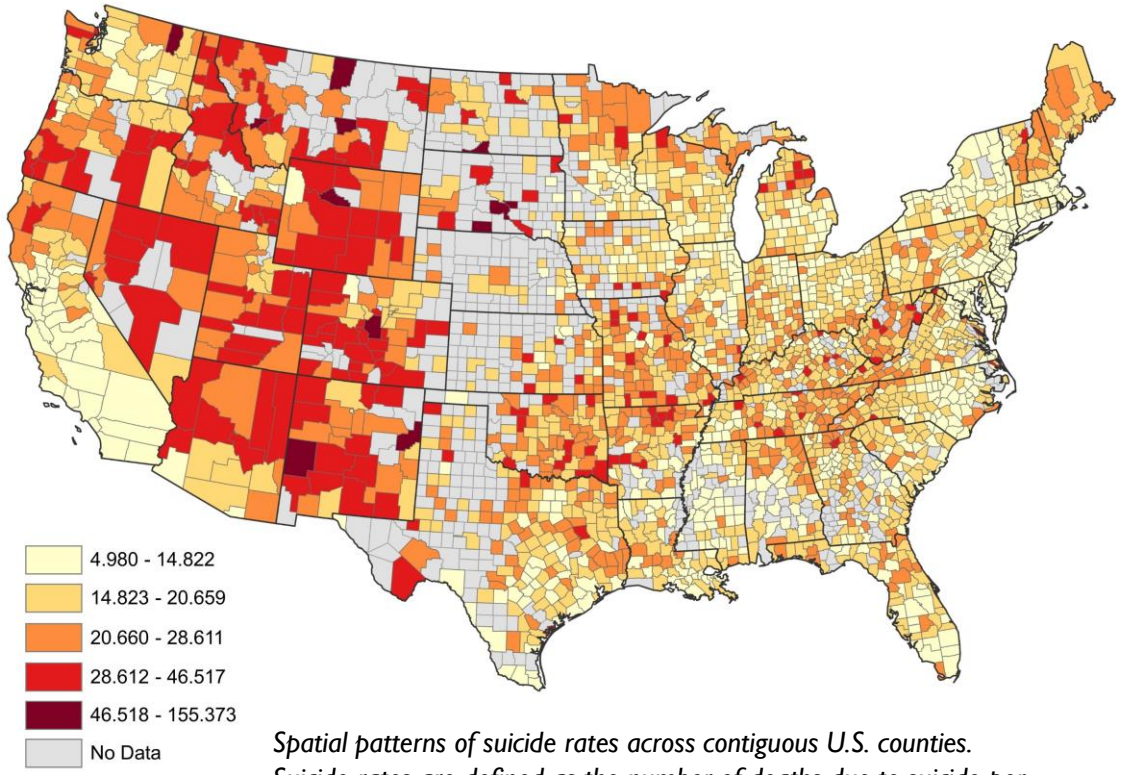
1. **Choose a domain of interest**
2. **Conduct a literature review**
   - Identify existing studies and common methods used
   - Look for areas where spatial variation has been ignored or oversimplified
   - Ask key questions:
     - *Where might spatial non-stationarity influence relationships between variables?*
     - *Have existing studies overlooked spatial effects or assumed global relationships?*
3. **Choose a study area and appropriate spatial scale**
   - What locations? Local, regional or national? Country, county or census tract scale?
4. **Develop a clear and focused research question**
   - Ensure the question is specific and measurable
5. **Align methods with the research gap**
   - Select appropriate spatial regression techniques (e.g., GWR)
   - Justify your choice (statement of need) based on the research gap



Think about available data throughout this process!

Identify a Research Gap and Formulating a Research Plan: *A Case Study*

1. **Choose a domain of interest:** Public Health
2. **Conduct a literature review**
   - Traditional regressions are typically used to assess health outcomes
   - Suicide mortality is a critical concern where spatial variation is often ignored
3. **Choose a study area and appropriate spatial scale**
   - United States, across all contiguous counties
4. **Develop a clear and focused research question**
   - How do the relationships between social determinants and suicide vary spatially across U.S. counties?
5. **Align methods with the research gap**
   - Spatial regression



Spatial patterns of suicide rates across contiguous U.S. counties. Suicide rates are defined as the number of deaths due to suicide per 100,000 population (age-adjusted) from 2016-2020.

Collect Data Sources

1. Identify variables based on your research question

- Dependent variable (e.g., disease rate, housing price)
- Independent variables (e.g., income, pollution, distance to services)
  - *Should be correlated to dependent variable in some aspect → determined with domain knowledge or literature review*
  - *Independent variables should NOT be correlated to each other → important to check before analysis!!!*

2. Explore common and reliable spatial data sources:

- Government agencies (e.g., U.S. Census Bureau, EPA, CDC)
- Open data portals (e.g., data.gov, local GIS portals, CHR, ESRI)
- Remote sensing & satellite imagery (e.g., Landsat, MODIS)
- Academic & NGO repositories (e.g., Harvard Dataverse, WHO)

3. Gather spatially referenced data

- Ensure all datasets include geographic identifiers (e.g., census tracts)
- Use consistent spatial and temporal resolution across variables

Recognize this is an iterative process!!!

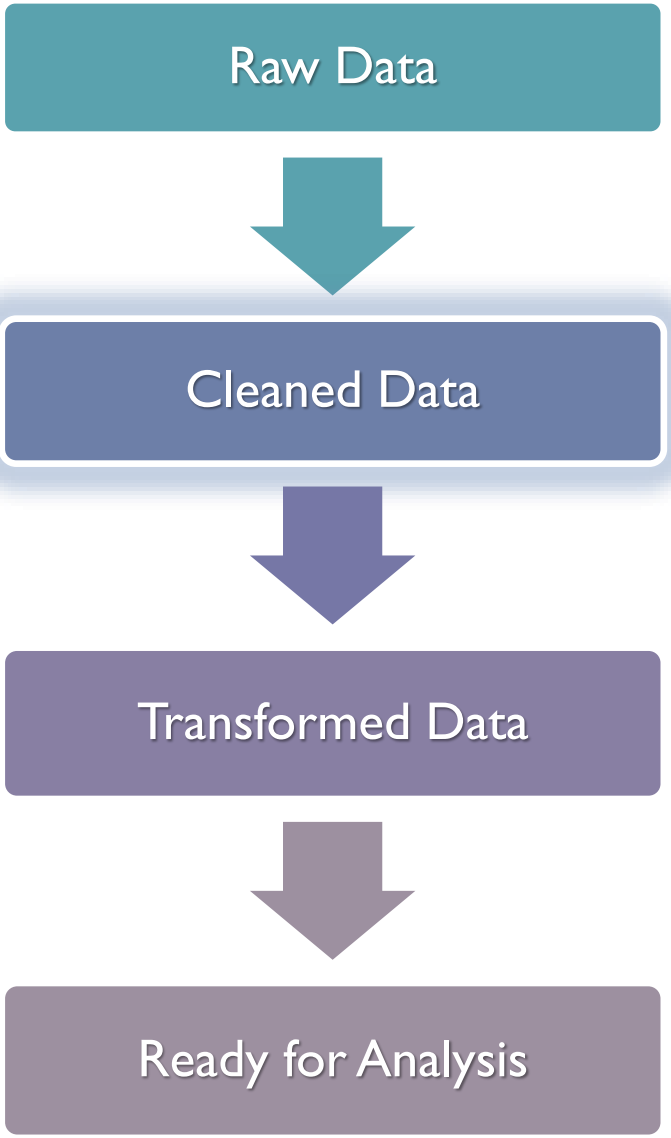
- **Available** data may refine your research question
- May require revisiting earlier steps (e.g., redefining variables, scale)

Data Description Table: *County-aggregated variables analyzed in a spatial regression project analyzing the spatial non-stationarity in suicide determinants. Data is from County Health Rankings & Roadmaps (CHR) and U.S. Census Bureau's American Community Survey (ACS).*

Variable	Description	Years	Source
Independent Variables			
Living Alone	Percent of population living alone	2017-2021	ACS
Smart Phone Access	Percent of households without access to a smart phone	2017-2021	ACS
Veteran	Percent of veteran civilian population aged 18+	2017-2021	ACS
Unemployed	Percent of population unemployed	2017-2021	ACS
Bachelor's Degree	Percent of population aged 25+ who completed at least a bachelor's degree	2017-2021	ACS
Food Insecurity	Percent of population who did not have access to a reliable source of food	2020	CHR
Mental Care Providers	Number of mental health providers per 100,000 population	2022	CHR
No Health Insurance	Percent of population with no health insurance	2017-2021	ACS
Poor Mental Health Days	Average number of mentally unhealthy days reported in past 30 days (age-adjusted)	2020	CHR
Primary Care Physicians	Number of primary care physicians per 100,000 population	2020	CHR
Rural	Percent of land considered rural area	2017-2021	ACS
American Indian/Alaskan Native	Percent of American Indian and Alaskan Native alone population	2017-2021	ACS
Black	Percent of Black population alone	2017-2021	ACS
Dependent	Percent of population within dependent age groups: under 18 or over 65	2017-2021	ACS
Divorced	Percent of divorced population	2017-2021	ACS
Female	Percent of Female population	2017-2021	ACS
White	Percent of White population alone	2017-2021	ACS
Dependent Variable			
Suicide Rate	Number of deaths due to suicide per 100,000 population (age-adjusted).	2016-2020	CHR

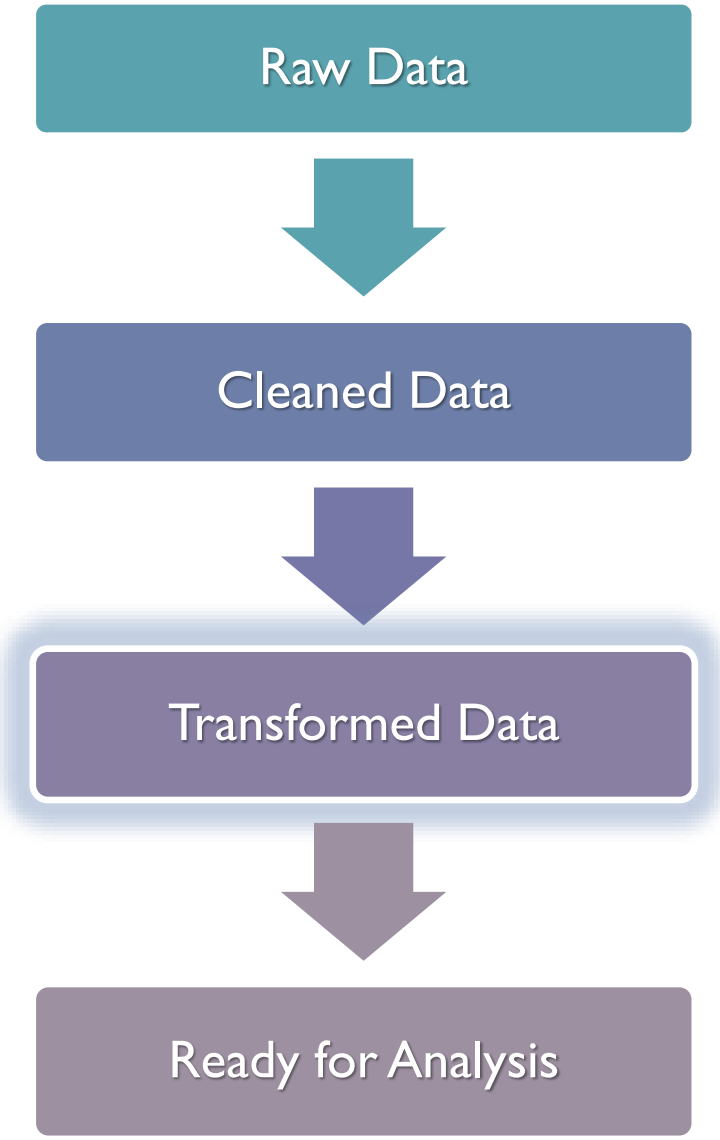
Prepare, Clean and Explore Data

1. **Combine and organize your datasets**
   - Append variables into a single spatial dataset through **joins**
2. **Check for missing or inconsistent values**
   - Identify and address missing data (e.g., imputation, exclusion)
3. **Create an analysis-ready dataset**
   - Final dataset should include all necessary variables
   - Clean, consistent, and ready for data transformation and spatial analyses
4. **Examine descriptive statistics**
   - Calculate measures (*mean, median, standard deviation*) explaining data
5. **Visualize spatial patterns**
   - Create choropleth maps of independent and dependent variables



Data Transformation

1. **Transform skewed variables**
   - Use log, square-root, or Box-Cox transformation
   - Helps meet regression assumptions of normality and linearity
2. **Standardize or normalize your finalized dataset**
   - Ensures variables are on comparable scales
   - Important for spatial regression methods where scale affects interpretation
   - Common methods: Z-score, Min-Max scaling



Fitting a Global Regression Model

- **Fit an initial global regression model**
  - Use cleaned and transformed dataset
  - Specify dependent and independent variables
- **Examine predictor relationships**
  - Look at the **significance (p-value),** as well as **magnitude** and **direction** of each coefficient
  - Identify potential issues with multicollinearity between variables using **Variance Inflation Factor (VIF)**
- **Interpret model performance**
  - Use measures like **R-squared** and **adjusted R-squared**
  - Consider overall model significance and explanatory power

OLS model results table *describing the relationship between county-level determinants and suicide rates, including coefficient, p-value and VIF for each standardized independent variable. P-values with an asterisk (*) indicate a statistically significant association.*

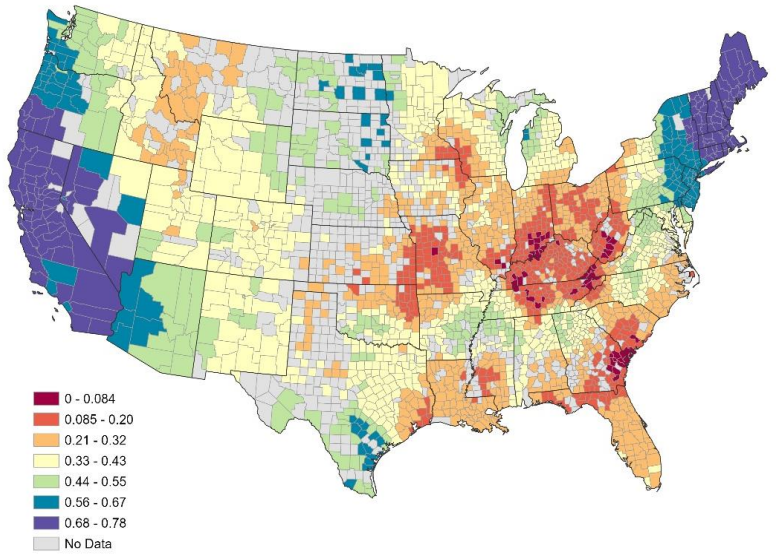
Standardized Variable	Coefficient	P-Value	VIF
Smart Phone Access	-0.1064	0.00*	2.85
Veteran	0.2628	0.00*	1.45
Social Connectedness	0.3925	0.00*	3.174
Unemployed	0.0216	0.26	1.81
Bachelor's Degree	0.0723	0.01*	4.10
Food Insecurity	0.0354	0.18	3.36
Poor Mental Health Days	0.1084	0.00*	1.97
Primary Care Physicians	0.0120	0.55	1.90
Mental Care Providers	0.0427	0.03*	1.90
No Health Insurance	0.1457	0.00*	2.06
Rural	-0.0448	0.09	3.37
Dependent Population	0.0672	0.00*	1.98
Divorced	0.1243	0.00*	1.77
White Population	0.0645	0.02*	3.59
Black Population	-0.2664	0.00*	2.45
American Indian/Alaskan Native Population	0.1229	0.00*	1.40
Female Population	-0.0533	0.00*	1.60
R-squared: 0.5051 , Adjusted R-squared: 0.5016			

Fitting a Spatial Regression Model

- **Preparing data**
  - Data must have a spatial component
  - Use cleaned, transformed and standardized dataset
  - Use the same variables as the global model for comparison
- **Fitting a spatial regression model**
  - Use GWR or MGWR
  - Addresses spatial non-stationarity
    - *The model fits a separate regression at each location, weighting nearby observations more heavily*
    - *Relationships between independent variables and the dependent variable vary across space*
    - *Produces results for each location in the analysis (e.g., each county)*
- **Interpret localized results**
  - Examine the **significance (p-value), direction** and **magnitude** for each spatial feature's coefficient
  - Look at the **R-squared** across spatial features
  - Results should be visualized using **maps** to show how relationships vary spatially

MGWR model results table *describing a subset of the spatially varying relationships between county-level determinants and suicides, including the mean coefficient, standard deviation, minimum coefficient, median coefficient, maximum coefficient, number and percentage of neighbors, and number and percentage of significant features for each standardized independent variable.*

Standardized Variable	Mean	Standard Deviation	Minimum	Maximum	Neighbors (% of Features)	Significance (% of Features)
Smart Phone Access	-0.087	0.0071	-0.1087	-0.072	2195 (97.08)	1153 (47.84)
Veteran	0.1757	0.0509	0.0903	0.2708	1286 (53.36)	2390 (99.17)
Social Connectedness	0.2021	0.1228	-0.0627	0.4593	264 (10.95)	1061 (44.02)
Unemployed	0.0187	0.0511	-0.1787	0.1275	409 (16.97)	0 (0.00)
Bachelor's Degree	-0.079	0.0033	-0.0823	-0.0688	2410 (100.00)	0 (0.00)
Food Insecurity	0.1099	0.0414	0.0572	0.1883	1363 (56.76)	1070 (44.40)
Poor Mental Health Days	0.1068	0.033	0.0688	0.172	1501 (62.28)	1396 (57.93)
Primary Care Physicians	0.0047	0.0052	-0.0014	0.0194	2410 (100.00)	0 (0.00)
Mental Care Providers	0.0095	0.0041	0.0053	0.022	2410 (100.00)	0 (0.00)
No Health Insurance	0.0184	0.0462	-0.0734	0.1094	888 (36.85)	0 (0.00)



Spatially varying R-squared values across US counties in the analysis assessing the spatial non-stationarity in suicide determinants.

Communicate Results

- **Summarize model performance**
  - Compare key metrics from traditional and spatial regression models using tables
  - Discuss and interpret: *We want to answer what do all these results really mean?*
- **Use visuals to tell the story**
  - Maps of local coefficients and R^2 to show spatial variation
  - Use clear legends and scales for accessibility
- **Highlight key spatial patterns**
  - Point out where and why relationships differ across space
  - Connect findings to real-world context
- **Acknowledge limitations**
  - Note data quality, model assumptions, and exploratory nature of GWR
  - Emphasize that spatial models complement — not replace — traditional ones
- **Tailor your message to the audience (e.g., stakeholders)**

Spatial Regression Project Workflow

Now its your turn to do all of this with project 3! ☺

