Local statistics: local indicators of spatial association

GIS 5923 Spatial Statistics

COLLEGE OF ATMOSPHERIC AND GEOGRAPHIC SCIENCES
DEPARTMENT OF GEOGRAPHY
AND ENVIRONMENTAL SUSTAINABILITY
The UNIVERSITY of OKLAHOMA

Local Statistics

- A local statistic is any descriptive statistic associated with a spatial data set whose value varies from place to place.
- In the broadest sense, any spatial data set is a collection of local statistics, since the recorded attribute values are different at each location.
- But, a "local statistic" usually is one that is derived by considering a subset of the data local to or nearby the spatial location where it is being calculated. Example: the localized mean

Two topics in Local Statistics

- We will cover two general topics:

- In the last set of slides, we focused on geographically-weighted regression: regression models in which we allow the coefficients to vary spatially

- In this set of slides, we will cover Local Indicators of Spatial Association (LISAs): used to identify hotspots and outliers

Local statistics

- In principle, almost any spatial statistics can be turned into a "local statistic": instead of summarizing over a whole data set, we summarize over only the data in the locality of each point

- In practice, the term "local statistic" refers most commonly to three statistics: The Getis-Ord G_i and G_i^* statistics, and the local Moran's I statistic

COLLEGE OF ATMOSPHERIC AND GEOGRAPHIC SCIENCES
DEPARTMENT OF GEOGRAPHY AND ENVIRONMENTAL SUSTAINABILITY
The UNIVERSITY of OKLAHOMA

The weights matrix revisited

$$W = \begin{bmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & w_{nn} \end{bmatrix}$$

- Typical interpretation: the element w_{ij} represents the hypothesized spatial influence of location *j* on location *i*
- Therefore, we can say that a row matrix $W_i = [w_{i1} \quad \cdots \quad w_{in}]$ describes the local neighborhood of each location *i*
- We've seen that there are many different ways to construct the weights matrix: adjacency, distance, k-nearest neighbors, etc.

The Getis-Ord G_i statistic

The Getis-Ord G_i statistic is used to detect local concentrations of high or low values in a variable x. For a location i it is calculated as:

$$G_i = \frac{\sum_j w_{ij} x_j}{\sum_{j=1}^{n} x_j} for\ all\ i \neq j$$

If the weights matrix is based on adjacency, then G_i can be interpreted as the proportion of the sum of all x values in the study area accounted for just by the neighbors of i

The Getis-Ord G_i^* Statistic

- The Getis-Ord G_i^* statistic is calculated by including location *i* itself in the numerator and denominator

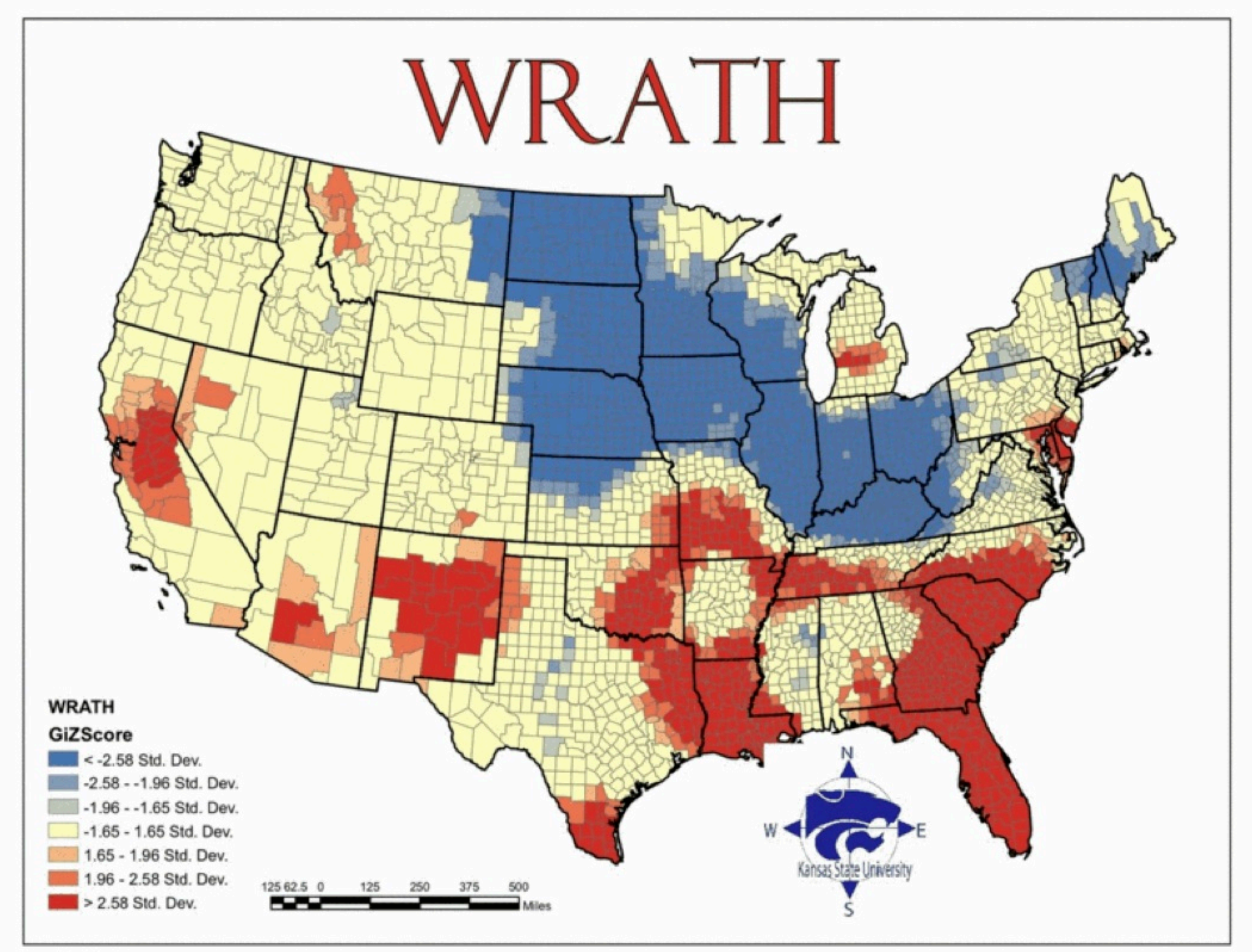
- Important: the attribute under consideration must be a ratio-scale variable with a natural origin
  - For example, you can see that the value of G_i will be different if you add a constant value to every location or if you transform the variable

Statistical significance of G and G^*

- Getis and Ord (1992) give the expected value and variance of G_i and G_i^* under a null hypothesis of CSR.

- As a result, we can assign a z-score to each location's G or G^* value

Wrath: Total number of violent crimes (murder, assault and rape) per capita



Kansas State University

COLLEGE OF ATMOSPHERIC AND GEOGRAPHIC SCIENCES
DEPARTMENT OF GEOGRAPHY
AND ENVIRONMENTAL SUSTAINABILITY
The UNIVERSITY of OKLAHOMA

Interpreting z-scores

- The typical approach would be to consider z scores outside the range -1.96 to +1.96 to be significant
- More care is required in making inferences from local statistics.
- Potential pitfalls:
  - 1) When events are rare, the statistic is being calculated on a small number of cases
  - 2) Edge effects
  - 3) Multiple testing problem: if we have 100 areal units in the study area, then on average 5 will be significant just by chance

Multiple test corrections

- The simplest approach to correct for multiple comparisons is to use a Bonferroni correction, and set a' = a / n.
  - Some authors find this too conservative. What is a' if a = 0.05 but n = 100?

- Alternatively, a Monte Carlo approach could be taken by holding the value at location *i* constant and shuffling the other attribute values among locations in the data set

Global Moran's I: an index of autocorrelation

Before discussing local Moran's I, let's review the formula for global Moran's I

Divide by the overall data set variance

Covariance term; subscripts i and j refer to different areal units

$$I = \left[\boxed{\frac{n}{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \right] \times \left[\frac{\boxed{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (y_i - \bar{y})(y_j - \bar{y})}}{\boxed{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}}} \right]$$

Normalize by the total spatial weights in the map

Local Moran's I

- Local Moran's I_i values are onents that are summed to calculate the Global Moran's I

This summation is omitted for local I

Global I

$$I = \left[\frac{n}{\sum_{i=1}^{n} (y_i - \bar{y})^2}\right] \times \left[\frac{\boxed{\sum_{i=1}^{n}}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\boxed{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}}\right]$$

Local I is not normalized by the sum of the spatial weights

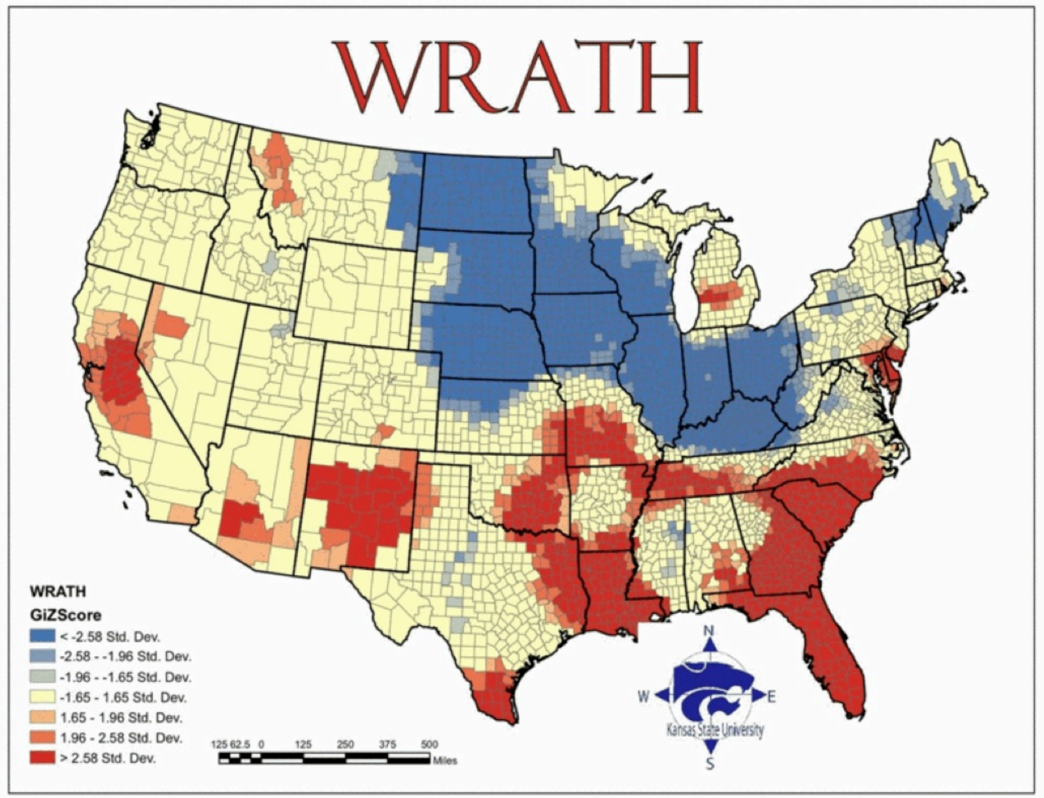
Local I

$$I_i = \left[\frac{n}{\sum_{i=1}^{n} (y_i - \bar{y})^2}\right] \times \left[\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})\right]$$

COLLEGE OF ATMOSPHERIC AND GEOGRAPHIC SCIENCES
DEPARTMENT OF GEOGRAPHY
AND ENVIRONMENTAL SUSTAINABILITY
The UNIVERSITY of OKLAHOMA

Which statistic to use?

- When exploring spatial data it is suggested that users apply both the G_i and the local Moran's I statistics.
- These statistics explore different but complementary processes underlying the observed spatial distribution of attribute values.



Kansas State University

COLLEGE OF ATMOSPHERIC AND GEOGRAPHIC SCIENCES
DEPARTMENT OF GEOGRAPHY
AND ENVIRONMENTAL SUSTAINABILITY
The UNIVERSITY of OKLAHOMA

R

- Let’s look at some examples in R…