Clustering I

Lecture #16 | GEOG 510
GIS & Spatial Analysis in Public Health

Varun Goel

Outline

- Terminology
- Spatial Autocorrelation
  - Global and Local
- Introducing Time

Clustering?

- What is clustering?
  - Clustering
    - Identifying whether events/values are clustered in space
    - Global, does not tell us "where"
      - Pattern is not random
    - For observations with values, spatial autocorrelation

Clustering?

- What is clustering?
  - Cluster detection
    - Identifying clusters of events/values in space (deviations from expected)
    - Local regions having…
      - High/Low values (e.g., incidence rate)
    - Spatial autocorrelation, scan statistics

Spatial Autocorrelation

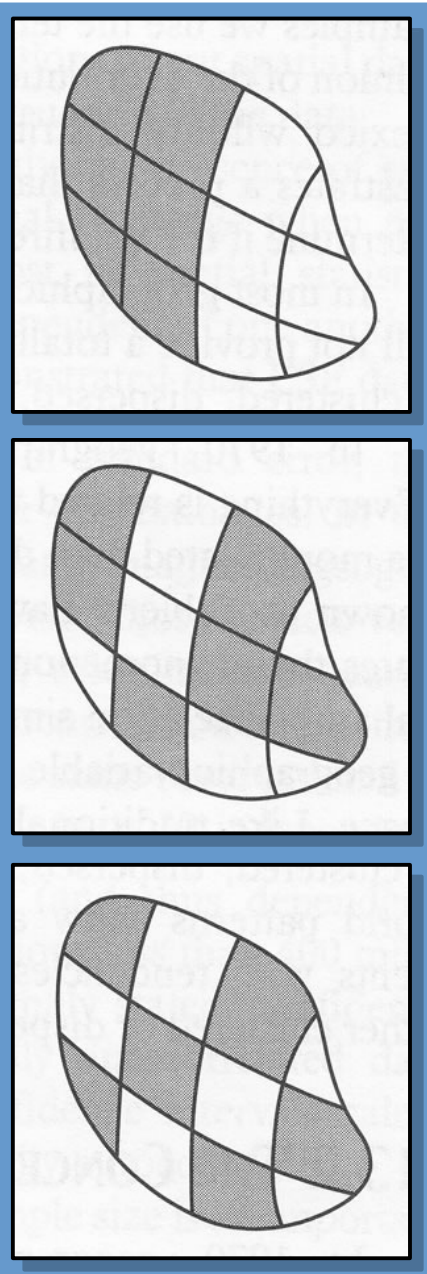
- Spatial Autocorrelation
  - The degree of similarity between objects that are located near each other
  - The arrangement or pattern of "values" within the landscape
    - Clustered, Random, Dispersed
  - Can be measured, quantitatively
    - For an entire region (global)
    - In a smaller area within the region (local)

Spatial Autocorrelation

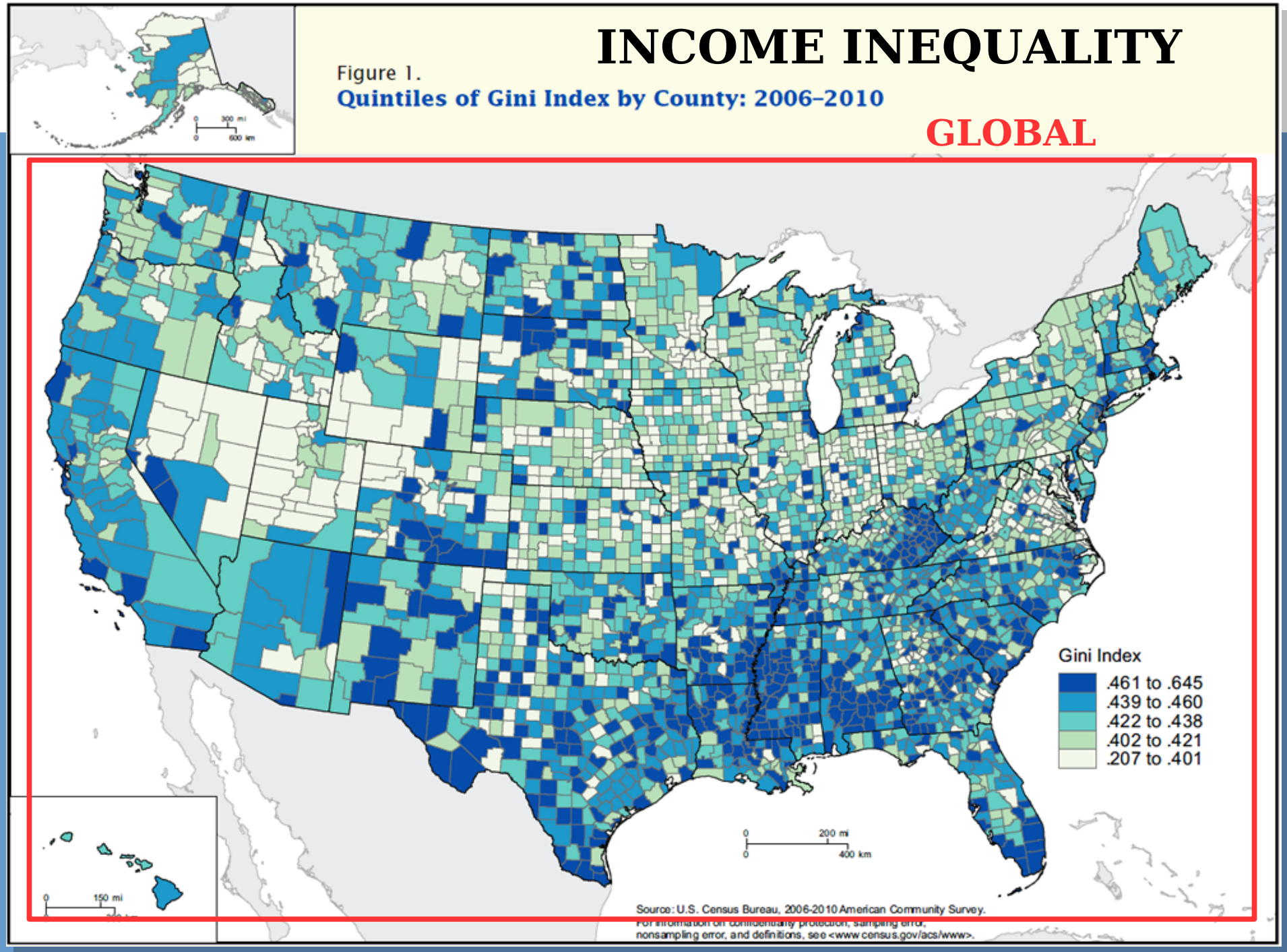
- For areal (polygon), point, or raster data, we measure how variable values are arranged
  - Not simply the locations of the objects, but the **attributes** associated with them
  - Be sure to control for variations in the number of people (at risk)!
    - Not recommended for count data unless population generating counts are exactly the same from place to place

Spatial Pattern

- Concepts
  - Clustered
    - Objects are configured or distributed near to one another
  - Random
    - Objects are configured or distributed such that there is no regular pattern
  - Ordered (dispersed)
    - Objects are configured or distributed in a regular or repeating fashion

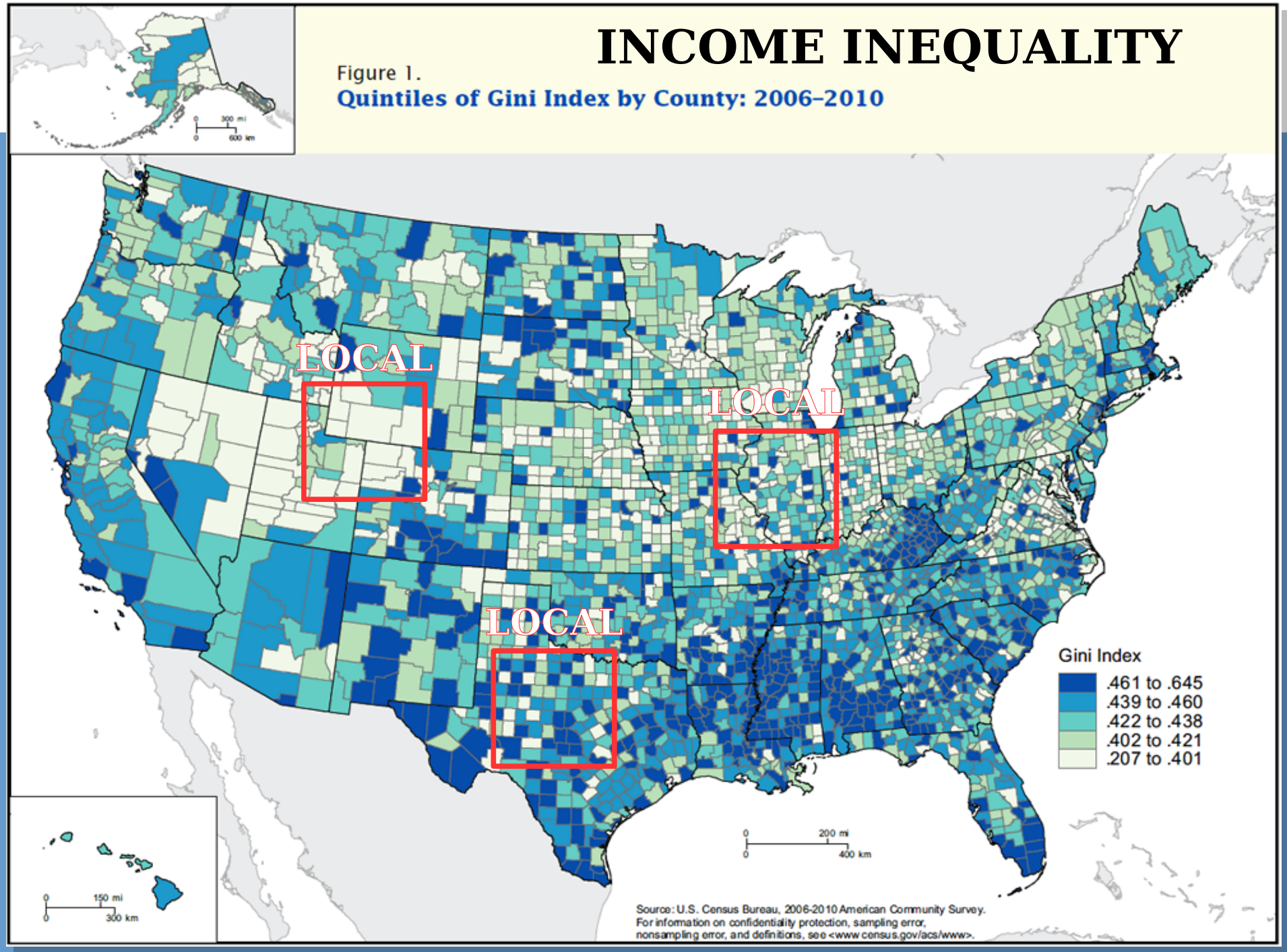


INCOME INEQUALITY



http://irjci.blogspot.com/2012/03/how-unequal-is-household-income-in-your.html

INCOME INEQUALITY



http://irjci.blogspot.com/2012/03/how-unequal-is-household-income-in-your.html

Spatial Autocorrelation

- Spatial Autocorrelation
  - The degree of **similarity** between objects that are located **near** each other
    - Requires the definition of neighbors to evaluate similarity
  - Note: this is "self" similarity
    - Correlation between variables will come a bit later

Spatial Autocorrelation

- Spatial Autocorrelation
  - We will start by only considering cross-sectional data
    - Values at a single point in time (or one value measured over a fixed time period)
      - Strictly spatial clustering, clusters

Moran's *I*

- Oft-used statistic for describing/testing the spatial autocorrelation within a region
  - Global (considers the whole region)
  - Measures the magnitude of spatial autocorrelation
    - Returns a single result (I)
    - In addition, it provides a *p*-value
      - The probability associated with I

Moran's *I*

- Global value
  - Moran's *I* ranges from -1 to 1 (continuous)
    - Perfectly dispersed: -1
    - Random: 0
    - Perfectly clustered: 1
  - Similar to Pearson's *R* (correlation)
  - Compares *I* of observed data to expected *I* (CSR)

Moran's *I*

- Formula

$$I = \frac{n}{S} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \overline{X})(x_j - \overline{X})}{\sum_{i=1}^{n}(x_i - \overline{X})^2}$$

n = number of areas
w_{ij} = the weight between area i and j
x_i = the value for area i
x_j = the value for area j
$\overline{X}$ = mean of all values
S = sum of all weights

$$S = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$$

Moran's *I*

- Formula

area *i*'s neighbors

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(x_i - \bar{X})(x_j - \bar{X})}{S \quad \text{area } i \quad \sum_{i=1}^{n}(x_i - \bar{X})^2}$$

For every *i*, compare deviation from mean of neighbors' value and self value

n = number of areas
w_{ij} = the weight between area i and j
x_i = the value for area i
x_j = the value for area j
$\bar{X}$ = mean of all values
S = sum of all weights

$$S = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}$$

Moran's *I*

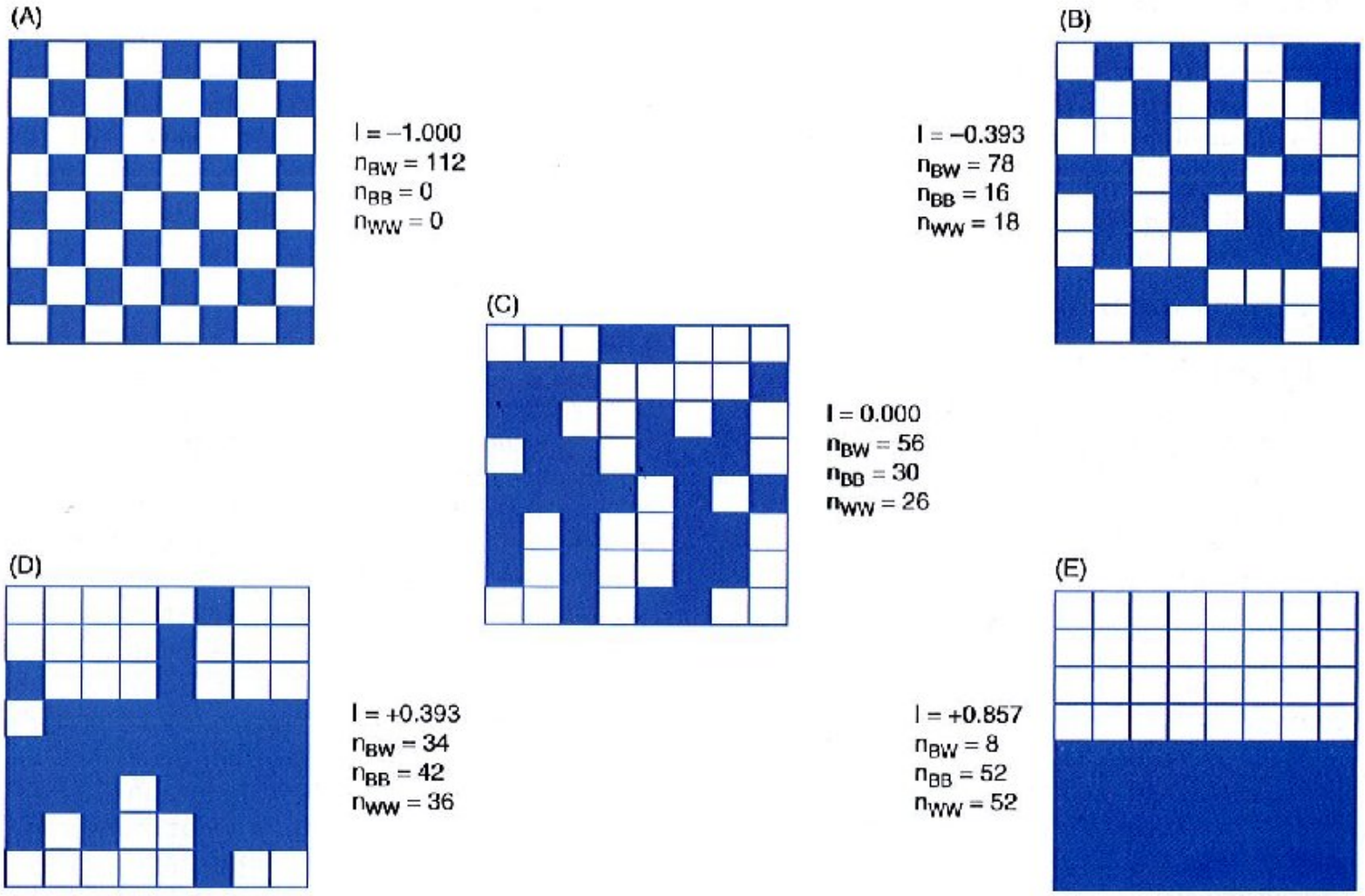


Figure 4.1 Field arrangements of blue and white cells exhibiting: (A) extreme negative spatial autocorrelation; (B) a dispersed arrangement; (C) spatial independence; (D) spatial clustering; and (E) extreme positive spatial autocorrelation. The values of the *I* statistic are calculated using the equation in Section 4.6 (*Source*: Goodchild 1986 CATMOG, GeoBooks, Norwich)

Moran's *I*

- Interpreting output
  - Magnitude
    - The closer to 1, the more clustered the values are
    - The closer to -1, the more dispersed the values are
  - Significance
    - Interpret *p*-value (e.g., < 0.05)

Moran's *I*

- Interpreting output
  - Importance
    - Beware significant, but unimportant deviation from random pattern
      - For example, $I = 0.04$, $p < 0.001$)
    - Like other inferential statistics, *p*-value is affected by number of observations
      - Personal interpretation system:
      - >0 to 0.1, barely clustered (pretty much random)
      - 0.1 to 0.3, slightly clustered
      - 0.3 to 0.5, moderately clustered
      - >0.5, highly clustered

Moran's *I*

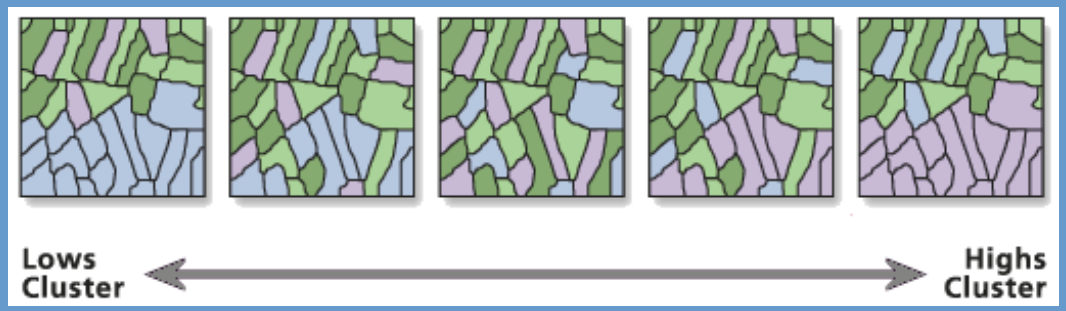
- Robustness test
  - Multiple neighborhood definitions

Table A2. Moran's *I* values for NME rate (%) for block group observations, under ten neighborhood definitions.

YEAR	ID(5)	ID(10)	ID(15)	ID(20)	KNN(5)	KNN(10)	KNN(15)	KNN(20)	CON(Q)	CON(R)
2000	0.073	0.076	0.078	0.079	0.106	0.093	0.083	0.079	0.119	0.104

Getis-Ord General *G*

- Alternate global measure of autocorrelation
  - Detects whether data driven by clusters of high values, low values (or CSR)



$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} x_i x_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j}$$

Getis-Ord General *G*

- Detects whether data has clusters of high values, low values, or CSR
  - Returns a single result (G) with corresponding *Z* score
  - In addition, it provides a *p*-value
    - The probability associated with G
  - We don't interpret the raw G value
    - Affected by magnitude of values
      - Not transferable
- Available in ArcPro

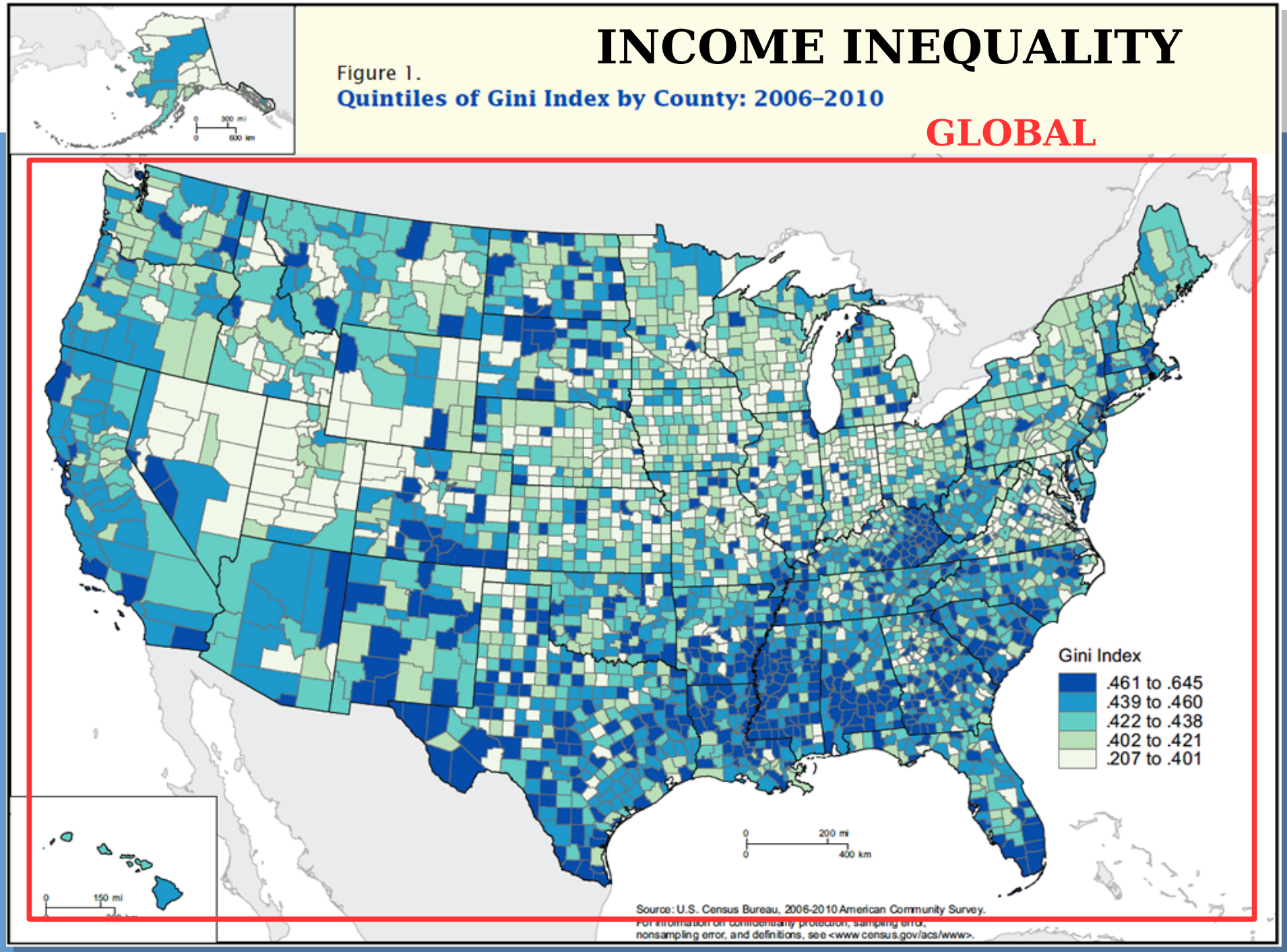
Global Autocorrelation

- For me, in practice...
  - Start with Moran's *I*
    - Clustered vs not Clustered
    - Ease of interpretation
  - Check Getis Ord *G*
    - Nature of Clustering (high/low values)

Stationary vs. Non...

- Autocorrelation
  - Global
    - Assumes that autocorrelation is stationary across space
      - Invariant from place to place
  - Local
    - Assumes that autocorrelation is non-stationary across space
      - Varies from place to place

INCOME INEQUALITY



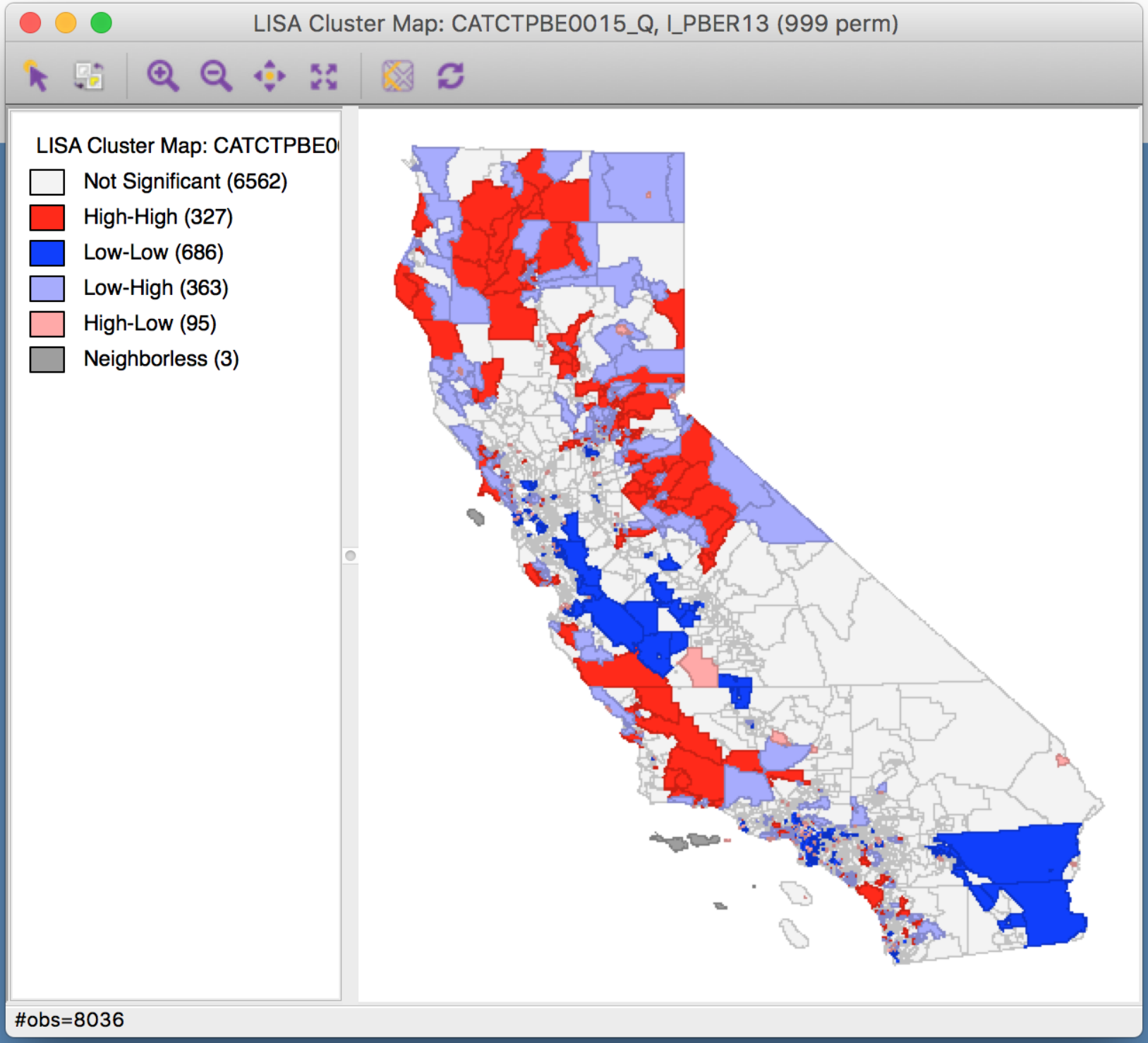
http://irjci.blogspot.com/2012/03/how-unequal-is-household-income-in-your.html

LISA

- Local Indicator of Spatial Association
  - Local version of Moran's *I*
    - Iterates through each observation and provides a measure of autocorrelation
      - And, associated *p*-value
    - Unlike global measures, results can be mapped
      - Reveals the nature of spatial autocorrelation throughout the study area

LISA

- Local Indicator of Spatial Association
  - Observations can be "hot" or "cold" spots (ugh), high or low outliers, or not significant
    - High-High (observation high, neighbors high)
    - Low-Low (observation low, neighbors low)
    - High outlier (observation high, neighbors low)
    - Low outlier (observation low, neighbors high)
  - Extremely useful for understanding "where" spatial autocorrelation is strong/weak


LISA Cluster Map: CATCTPBE0015_Q, I_PBER13 (999 perm)
LISA Cluster Map: CATCTPBE0
Not Significant (6562)
High-High (327)
Low-Low (686)
Low-High (363)
High-Low (95)
Neighborless (3)
#obs=8036

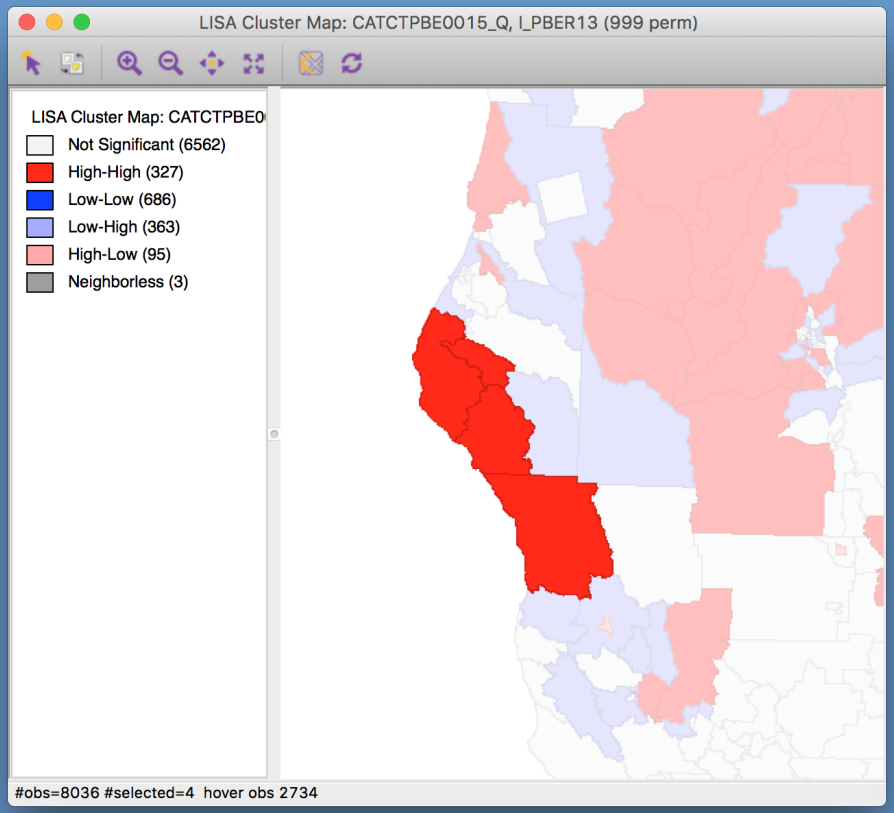


Table - CATCTPBE0015

	R12	PBER13	PBER14	PBER15	LISA_I	LISA_CL	LISA_P
498	15789	0.600000	0.290909	0.117647	27.7429107	1	0.0060000
2827	57692	0.049180	0.048387	0.084746	1.8241487	1	0.0010000
2833	21101	0.370787	0.406250	0.369369	24.6830557	1	0.0050000
3487	95238	0.195122	0.225806	0.066667	5.6619757	1	0.0110000
1	00000	0.059524	0.013889	0.013333	0.1447826	0	0.2190000
2	00000	0.008621	0.050000	0.010000	-0.0776767	0	0.1390000
3	18868	0.000000	0.008929	0.000000	0.1503457	0	0.1870000
4	00000	0.000000	0.000000	0.000000	0.1540371	0	0.1210000
5	20408	0.031746	0.000000	0.018692	-0.0700253	0	0.0800000
6	18018	0.009434	0.000000	0.000000	0.0949752	2	0.0270000
7	00000	0.045455	0.000000	0.000000	-0.1869132	4	0.0010000
8	00000	0.007692	0.008403	0.000000	-0.1011829	0	0.1330000
9	00000	0.000000	0.000000	0.000000	-0.1869093	0	0.1120000
10	00000	0.000000	0.000000	0.000000	0.1688026	2	0.0400000
11	09709	0.000000	0.000000	0.000000	0.1596978	2	0.0140000
12	00000	0.000000	0.000000	0.000000	0.1176653	0	0.1540000
13	00000	0.000000	0.583333	0.684211	-0.2310434	0	0.0740000
14	35714	0.019608	0.021277	0.019802	0.0038928	0	0.4210000
15	16667	0.035714	0.012346	0.013889	-0.0360511	0	0.4920000
16	08547	0.000000	0.000000	0.035211	0.1440382	0	0.1110000
17	99338	0.091837	0.056818	0.023585	0.2728499	0	0.2180000
18	00000	0.000000	0.000000	0.000000	-0.1822374	0	0.1030000
19	00000	0.000000	0.000000	0.000000	0.1470644	0	0.1300000
20	00000	0.000000	0.014286	0.000000	0.1577026	0	0.0930000
21	12658	0.007194	0.005618	0.000000	0.0603476	0	0.3650000
22	03774	0.000000	0.000000	0.002976	0.0792863	0	0.4070000

#obs=8036 #selected=4

Getis-Ord G_i^*

- Local version of Getis Ord *G*
  - Iterates through each observation and provides a measure of autocorrelation (*G*) and *Z* score
    - And, associated *p*-value
- Two versions
  - G_i^* , Considers observation and neighbors
  - G_i , Considers only neighbors

Getis-Ord G_i^*

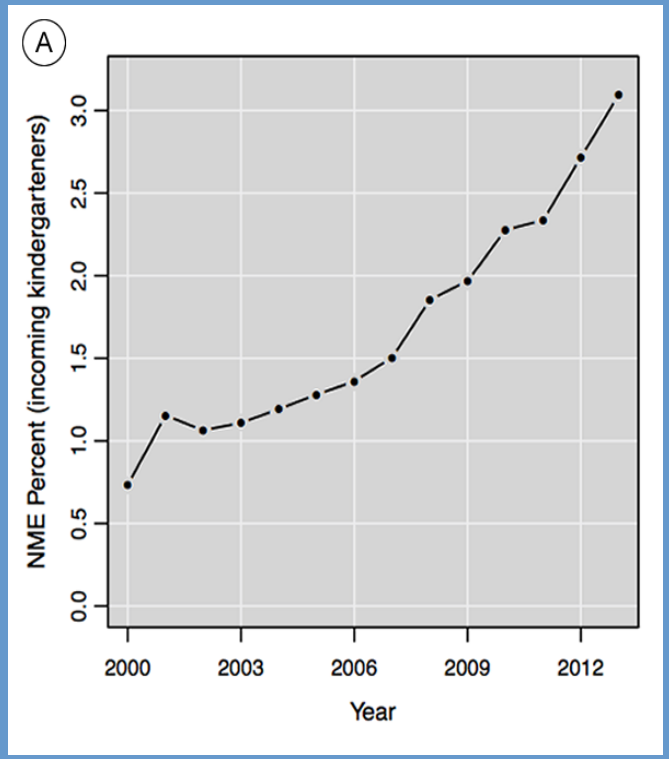
- Local version of Getis Ord *G*
  - Returns "hot" and "cold" spots
    - But, not outliers
    - Provides the hotness or coldness of each feature, via interpretation of the p-value
      - Basically, how confident we are that this is a true cluster of high/low values

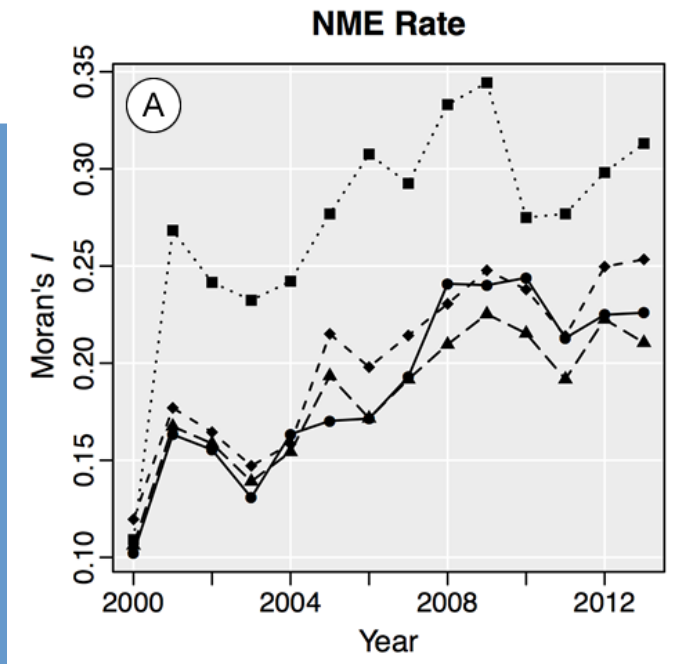
Introducing Time

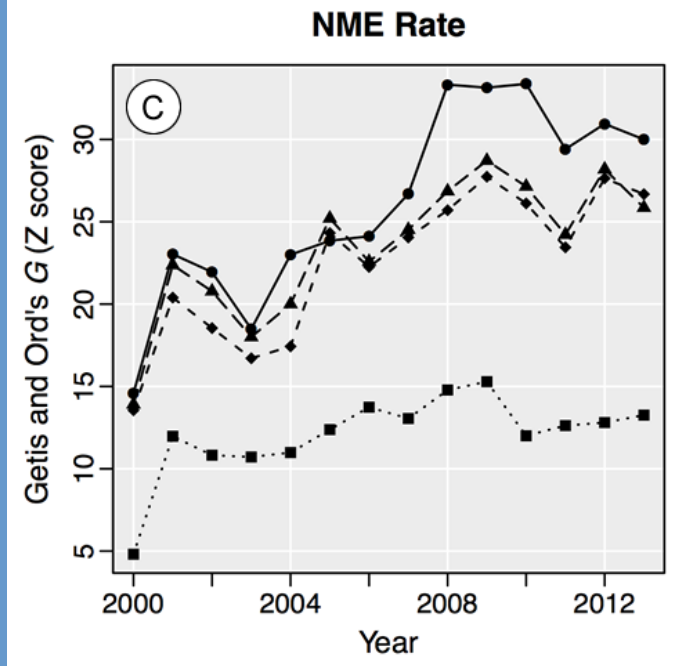
- Often, we are interested in whether things changed over time
  - Did the overall pattern (nature of clustering) change over time?
  - Did the location of local clusters change over time?
  - Were the changes clustered?
  - Where are local clusters of change?

Introducing Time

- Did the overall pattern (nature of clustering) change over time?
  - Evaluate global clustering method at different points in time, e.g., yearly

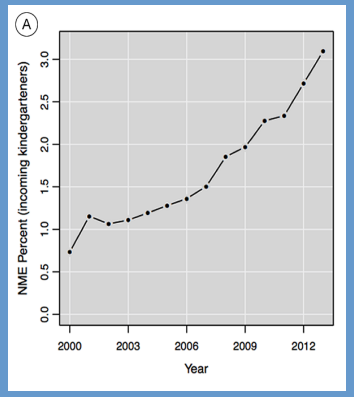


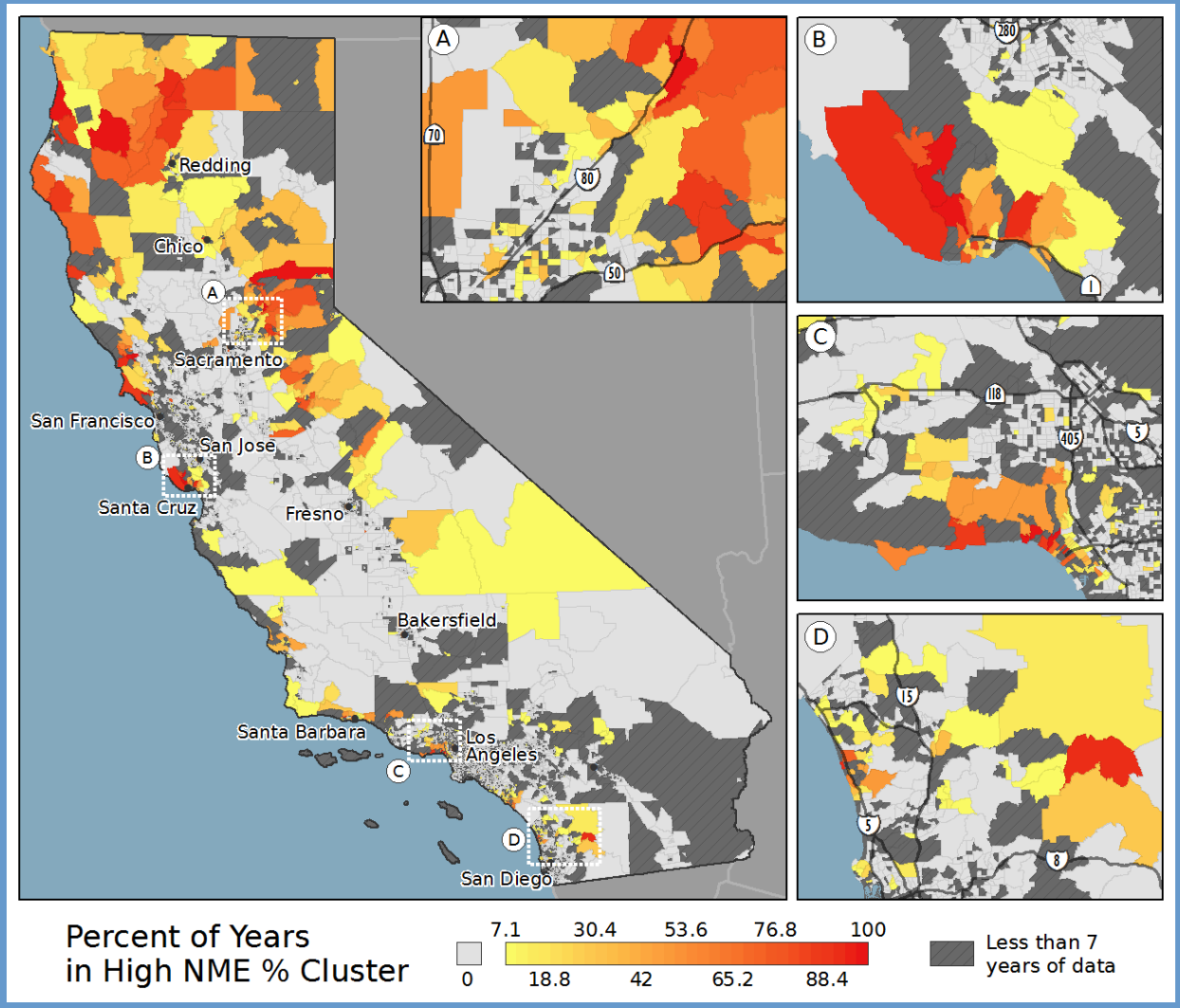




Introducing Time

- Did the location of local clusters change over time?
  - Evaluate local clustering method at different points in time, e.g., yearly, overlay results

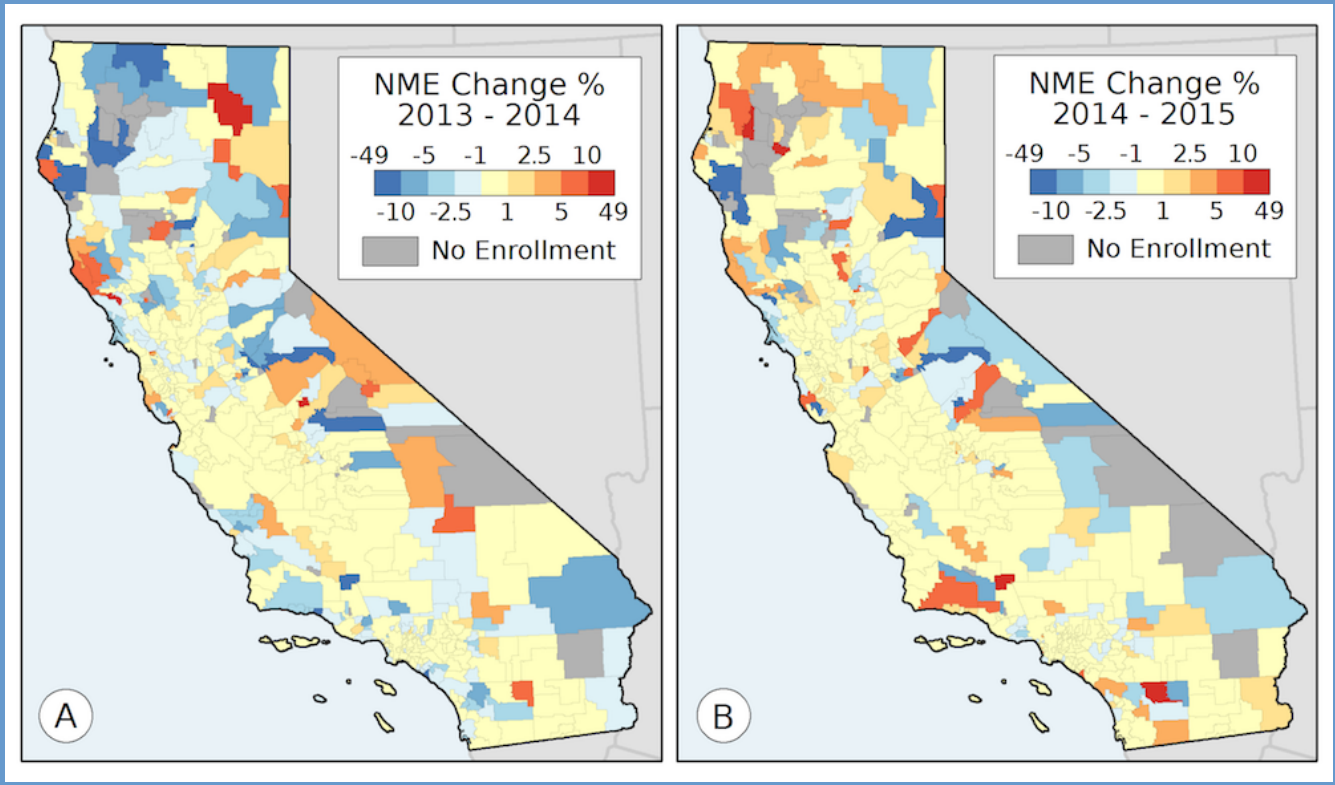


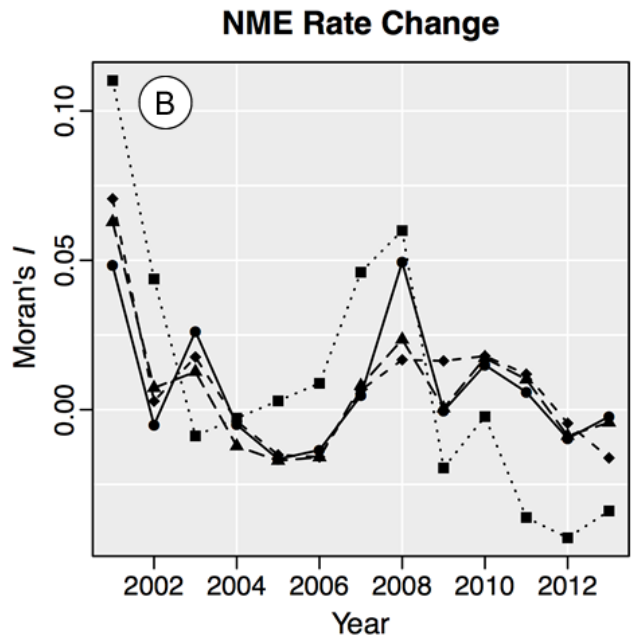


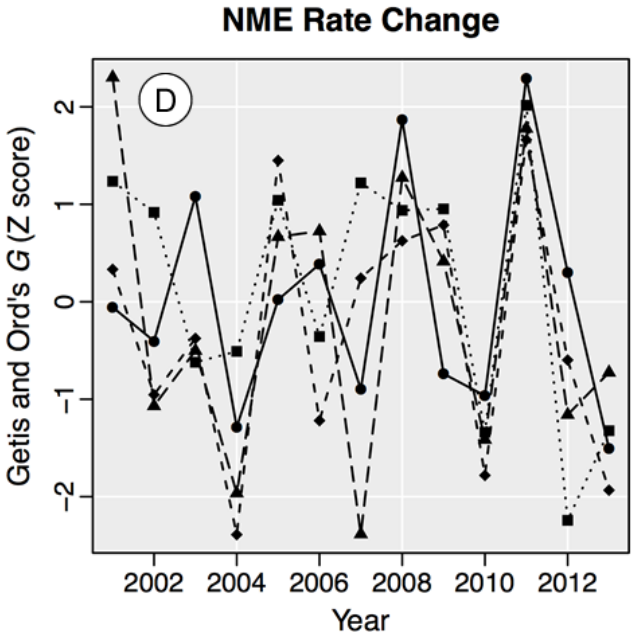
Introducing Time

- Were the changes clustered?
- Where are local clusters of change?

Differential Moran's I (and LISA) in GeoDa







Introducing Time

- Bivariate Moran's *I* and LISA in GeoDa
  - Evaluates observation's value at time *t* to neighbors' values at *t*-1
  - DO NOT USE FOR CORRELATION AMONG TWO DIFFERENT VARIABLES
    - Does not consider “in place” correlation
  - Limited usefulness, but possibly interesting

Spatiotemporal Clustering

- Local spacetime clusters (clusters in both space and time)
  - In ArcGIS Pro, not ArcMap
  - Extension of Getis Ord G_i^* to include temporal neighbors
    - Instead of a data layer, think of it as a 3D data cube (2D space + 1D time)
    - Considers spatial patterns and temporal trends
      - A lot of potential output values – adds temporal aspect to the hot/cold spatial output

Updates

- Final Project guidelines
  - Uploaded next monday
- Grading
  - Submit all outstanding assignments/exercises by Monday, March 10th.
  - "Midterm" grade released next Friday

Keywords

- Clustering, Cluster detection
- Spatial Autocorrelation
  - Global
    - Moran's *I*, Getis-Ord *G*
  - Local
    - LISA, Getis-Ord G_i^*, G_i
- Change over time