Spatial Statistics
Areal Data Unit 2

PM569 Spatial Statistics

Lecture 6: October 21, 2016

Areal Data

- Spatial similarity
- Global indexes of spatial autocorrelation
- Local indexes of spatial autocorrelation (LISA)
- Spatial autoregressive models (CAR, SAR)

Areal Data: Global Indexes of Spatial Autocorrelation

- The goal of global indexes of spatial autocorrelation is to summarize the degree to which similar observations tend to occur near each other
- Global indexes are summaries over the entire study area, akin to testing clustering rather than a test to detect individual clusters
- Indexes share a common structure: calculate the similarity of values at locations i and j then weight the similarity by the proximity of locations i and j
- High similarities with high weight indicate similar values that are close together; low similarities with high weight indicate dissimilar values that are close together

Indexes of spatial autocorrelation

- We want to summarize similarity between nearby areal units
- Spatial autocorrelation is the the correlation of the same measurement taken at different areal units
- The similarity of values at locations B_i and B_j are weighted by the proximity of i and j
- The weight w_{ij} defines proximity
- In general the extent of similarity is represented by the weighted average of similarity between areal units: indexes of spatial autocorrelation are built on this basic form:

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

Moran's I

- Moran's I (1950) follows the basic form for global indexes of spatial autocorrelation with similarity between areal units i and j defined as the product of the respective difference between y_i and y_j with the overall mean
- Similarity $sim_{ij} = (y_i - \bar{y})(y_j - \bar{y})$
- Where $\bar{y} = \sum_{i=1}^{n} y_i / n$
- Divide the basic form by the sample variance to get the Moran's I statistic:
- $I = \frac{1}{s^2} \frac{\sum_i \sum_j (y_i - \bar{y})(y_j - \bar{y})}{\sum_i \sum_j w_{ij}}$
- Where $s^2 = \frac{\sum_i (y_i - \bar{y})^2}{n}$

Moran's I

- I is a random variable having a distribution defined by the distributions of and interactions between the y_i
- When neighbouring regions have similar values (pattern is clustered), I will be positive
- When neighbouring regions have different values (pattern is regular), I will be negative
- When there is no correlation between neighbouring values: $E(I) = -\frac{1}{n-1}$
- When $n \to \infty$, $E(I) \to 0$
- I is asymptotically normally distributed where $\frac{I+\frac{1}{n-1}}{\sqrt{Var(I)}} \sim N(0,1)$

Moran's I

- Moran's I is similar to Pearson's correlation but it is not bounded on [-1,1] because of the spatial weights
- Null hypothesis: NO spatial association, i.e. y_i iid
- Compare the z-score to a standard normal distribution
- The z-score that we compare to the standard normal is $z = \frac{I - E(I)}{\sqrt{Var(I)}}$ where $E(I) = -\frac{1}{n-1}$ and V(I) is a little complicated (shown later)

Moran's I in R using North Carolina SIDS data

```
# Define neighbours.  Choose k=2 NN
IDs<-row.names(as(nc, "data.frame"))
sids.kn2<-knn2nb(knearneigh(coordinates(nc), k=2,
RANN=FALSE), row.names=IDs)
# Convert to weight matrix (row standardized)
sids.kn2.w<-nb2listw(sids.kn2, style="W")
# Use moran.test for Moran's I
moranSIDS<-moran.test(sids79.rate,sids.kn2.w)
```

Moran's I in R using North Carolina SIDS data

Result:

```
data:  sids79.rate
weights:  sids.kn2.w

Moran I statistic standard deviate = 2.4465, p-value
= 0.007213
alternative hypothesis:  greater
sample estimates:
Moran I statistic Expectation Variance
0.214682382 -0.010101010 0.008442014
```

The null hypothesis of no spatial correlation is rejected.

Geary's c

- Geary (1954) devides the contiguity ratio or Geary's c
- Similarity $sim_{ij} = (y_i - y_j)^2$
- If regions i and j have similar values, sim$_{ij}$ wil be small
- $c = \frac{n-1}{2\sum_i (y_i - \bar{y})^2} \frac{\sum_i \sum_j w_{ij}(y_i - y_j)^2}{\sum_i \sum_j w_{ij}}$
- Like Moran's it is a weighted average, but here it is scaled by a measure of the overall variation around the mean, $\bar{y}$

Geary's c

- c ranges from 0 to 2 with 0 indicating perfect positive spatial correlation and 2 indicating perfect negative spatial correlation
- c is not a Pearson correlation (related to the Durbin-Watson statistic)
- Low values of Geary's c denote positive autocorrelation and high values indicate negative correlation
- Expected value, $E(c) = 1$ under spatial independence

Geary's c in R using North Carolina SIDS data

```
# Define neighbours.  Choose k=2 NN
IDs<-row.names(as(nc, "data.frame"))
sids.kn2<-knn2nb(knearneigh(coordinates(nc), k=2,
RANN=FALSE), row.names=IDs)
# Convert to weight matrix (row standardized)
sids.kn2.w<-nb2listw(sids.kn2, style="W")
# Use geary.test for Geary's c
gearySIDS<-geary.test(sids79.rate,sids.kn2.w)
```

Geary's c in R using North Carolina SIDS data

Result:

```
data:  sids79.rate
weights:  sids.kn2.w

Geary C statistic standard deviate = 1.6166, p-value
= 0.05299
alternative hypothesis:  Expectation greater than
statistic
sample estimates:
Geary C statistic Expectation Variance
0.83688116 1.00000000 0.01018165
```

We have a marginal p-value for negative spatial autocorrelation.

Moran's I

- Inference is performed on I under the randomization assumption or by Monte Carlo tests
- Null hypothesis is always that there is no spatial association
- We have a normality assumption of spatial independence such that all observations follow iid gaussian distribution
- Random permutations of the null distribution are computed

Moran's I

- Randomization means the observations are assigned at random in the B_i areal units
- Test statistic is z= (observed-expected)/s.d expected
- $E(I) = \frac{-1}{(n-1)}$ under null hypothesis of no autocorrelation
- $V(I)$ is more complicated and is dependent upon the weight matrix. The normal approximation of the variance under randomization is (Cliff and Ord, 1981):

$$V(I) = \frac{ns_1 - s_2 s_3}{(n-1)(n-2)(n-3)(\sum_i \sum_j w_{ij})^2}$$

$$s_1 = (n^2 - 3n + 3)(0.5 \sum_i \sum_j (w_{ij} + w_{ji})^2) - n(\sum_i (\sum_j w_{ij} + \sum_j w_{ji})^2) + 3(\sum_i \sum_j w_{ij})^2$$

$$s_2 = \frac{n^{-1} \sum_i (y_i - \bar{y})^4}{(n^{-1} \sum_i (y_i - \bar{y})^2)^2}$$

$$s_3 = 0.5 \sum_i \sum_j (w_{ij} + w_{ji})^2 - 2n(0.5 \sum_i \sum_j (w_{ij} + w_{ji})^2) + 6(\sum_i \sum_j w_{ij})^2$$

Moran's I

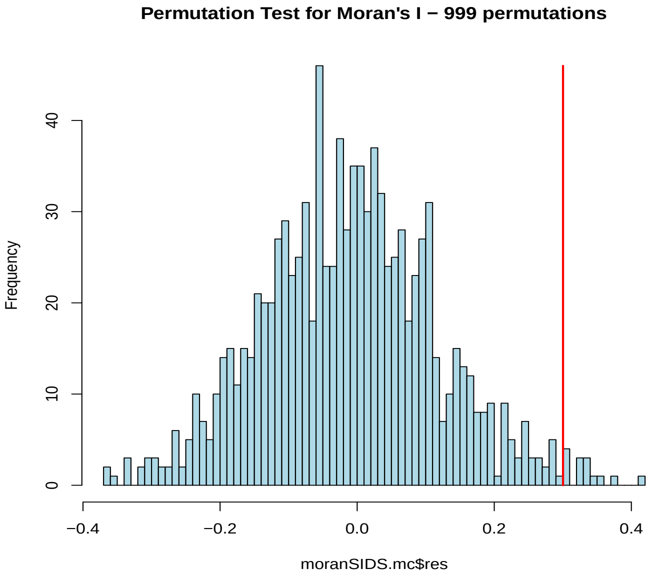
- Monte Carlo approach repeats randomization of the observations into the areal units a large number of times (e.g. $N_{sim} \sim 999$)
- For each randomization the Moran's I statistic is calculated
- Compare the observed Moran's I to the random set
- If the actual I falls at the 5th/95th percentile (or smaller/greater) then it is significant at $\alpha = 0.05$

- R output for monte carlo estimate of global Moran's I
- `moran.mc(sids79.rate,sids.kn1.w,nsim=999)`

```
Monte-Carlo simulation of Moran's I
data:  sids79.rate
weights:  sids.kn1.w
number of simulations + 1:  1000
statistic = 0.3003, observed rank = 987, p-value
= 0.013
alternative hypothesis:  greater
```

- The null hypothesis of no spatial correlation is rejected.

Areal Data: Global Indexes of Spatial Autocorrelation



Areal Data: Global Indexes of Spatial Autocorrelation

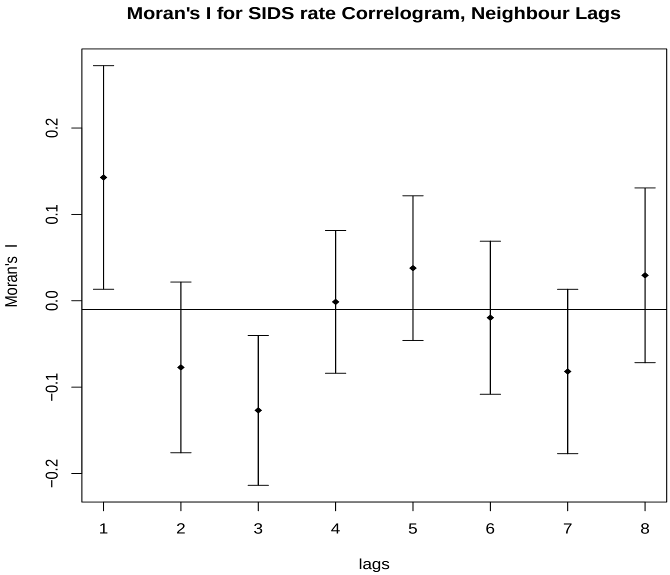
Issues with spatial autocorrelation tests:

- they assume that the mean trend has been removed, for example elevation effect when examining spatial pattern in precipitation
- this assumption is in centering the mean $y_i - \bar{y}$ as it's equivalent to saying the correct model has constant mean and the spatial pattern is represented in the spatial weights
- removing trend may be impossible if you don't have covariate data
- spatial weights may be misspecified for testing autocorrelation, for instance too few neighbour weights when spatial pattern is based on larger distances or vice versa
- they require $min(N) \sim 20$ to provide asymptotically accurate results

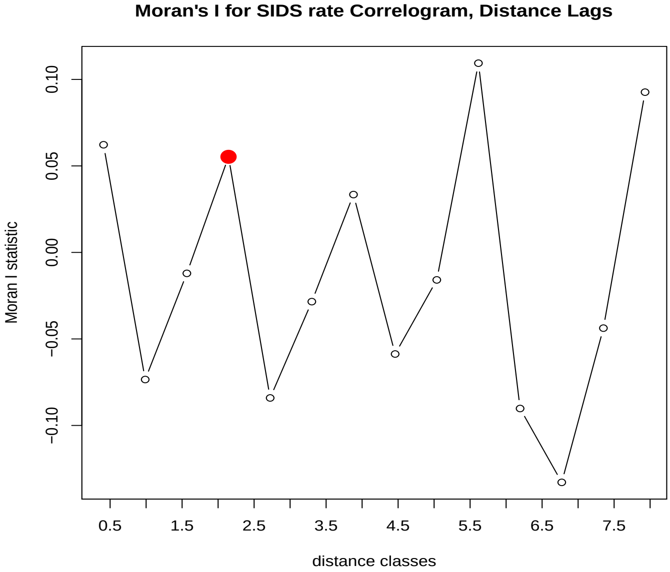
Areal Data: Global Indexes of Spatial Autocorrelation

- We can look at correlograms of the Moran's **I** statistic to determine appropriate number of neighbours or distance
- Calculate I based on knn for a range of k (e.g. 1,...,8) or number of borders shared (e.g. queen, rook)
- Calculate I based on d_{ij} for a range of distances

Areal Data: Global Indexes of Spatial Autocorrelation



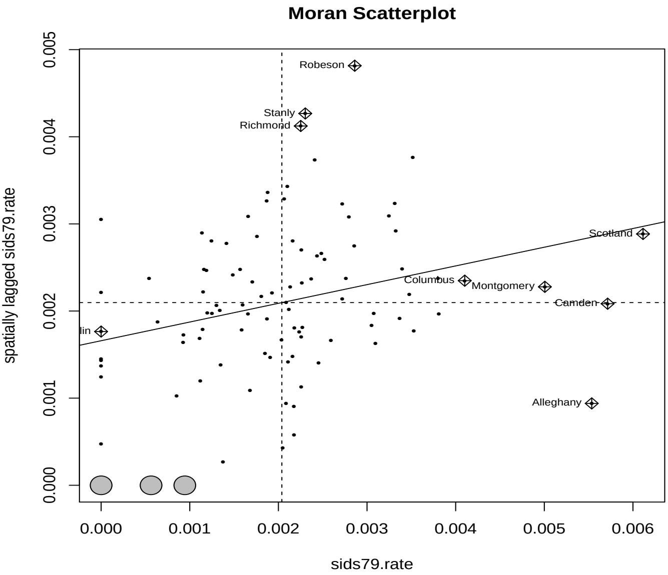
Areal Data: Global Indexes of Spatial Autocorrelation



Areal Data: Quasi Local Indexes of Spatial Autocorrelation

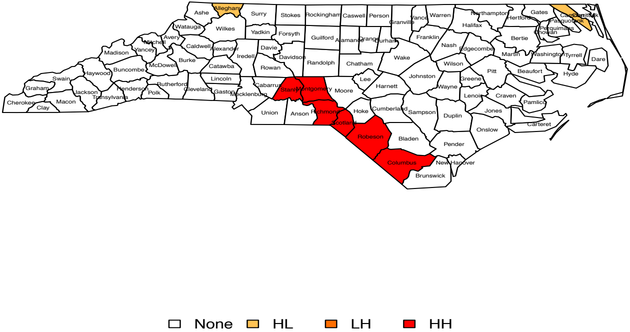
- Global tests of spatial autocorrelation can be broken down into components
- We can construct local tests that identify clusters
- One preliminary step is to construct a Moran's I scatterplot, where variable of interest (e.g. SIDS rate) is on the x-axis and their spatially lagged values on the y-axis

Areal Data: Quasi Local Indexes of Spatial Autocorrelation



Areal Data: Quasi Local Indexes of Spatial Autocorrelation

- This plot can be divided into quadrants of low-low, low-high, high-low and high-high
- The global Moran's I is the slope of this plot lm(wx x) where wx is the spatially lagged value of the values on the x-axis (e.g. SIDS rates)
- Can also detect outliers by testing for influence measures


Ashe
Alleghany
Surry
Stokes
Rockingham
Caswell
Person
Vance
Warren
Granville
Northampton
Hertford
Gates
Pasquotank
Perquimans
Chowan
Halifax
Watauga
Wilkes
Yadkin
Forsyth
Guilford
Alamance
Orange
Durham
Franklin
Nash
Edgecombe
Bertie
Avery
Mitchell
Yancey
Madison
Caldwell
Alexander
Davie
Iredell
Davidson
Randolph
Chatham
Wake
Wilson
Martin
Washington
Tyrrell
Dare
Burke
McDowell
Buncombe
Haywood
Swain
Catawba
Rowan
Lee
Johnston
Greene
Pitt
Beaufort
Hyde
Graham
Cherokee
Clay
Macon
Jackson
Transylvania
Henderson
Polk
Rutherford
Cleveland
Lincoln
Gaston
Mecklenburg
Cabarrus
Stanly
Montgomery
Moore
Harnett
Wayne
Lenoir
Craven
Pamlico
Union
Anson
Richmond
Scotland
Hoke
Cumberland
Sampson
Duplin
Jones
Onslow
Carteret
Robeson
Bladen
Pender
Columbus
New Hanover
Brunswick
None
HL
LH
HH

Areal Data: Local Indexes of Spatial Autocorrelation

- Global measures (Moran's I or Geary's c) are a single value that apply to the entire study area
- The same pattern or process occurs over the entire geographic area
- Global statistic suggests that there is clustering but does not identify areas of particular clusters
- Global test is often used first to determine if there is evidence of spatial association
- Want to detect local areas of similar values, need a local statistic
- LISAs are decompositions of global indicators into the contribution of each individual observation (i.e. $B_i \in D$)
- As a result the sum of LISAs is proportional to the equivalent global indicator
- Local Moran's I, Getis-Ord G*

Areal Data: Local Indexes of Spatial Autocorrelation

- With LISAs, each observation gives an indication of the extent of significant spatial clustering of similar values located around that observation
- The locations "around" one particular observation is defined as a neighbourhood and is formalized with the spatial adjacency weights matrix, W
- Recall, W can be based on sharing a border (full or partial) or distance
- Row standardization of W helps with interpretation of the statistic

Areal Data: Local Indexes of Spatial Autocorrelation

LISAs can be used to detect

- Clusters (areal units with similar neighbours): Local Moran's I
- Hotspots (areal units with dissimilar neighbours): Getis-Ord G*

Moran's I vs Local Moran's I

$$I = \frac{1}{s^2} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} (y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}}$$

$$s^2 = \frac{\sum_{i}^{n} (y_i - \bar{y})^2}{n}$$

$$I_i = \frac{y_i - \bar{y}}{s} \sum_{j}^{n} w_{ij} \frac{(y_j - \bar{y})}{s}$$

- I_i is calculated for each areal unit B_i

Areal Data: Local Indexes of Spatial Autocorrelation

Local Moran's I

$$I_i = \frac{y_i - \bar{y}}{s} \sum_{j}^{n} w_{ij} \frac{(y_j - \bar{y})}{s}$$

- Have a value of I for each B_i
- Sometimes the I_i are mapped to indicate units with high values indicating stronger local autocorrelation
- More often, z-score and significance of z-score is plotted
- As before, test statistics are generated under randomization
- Since in a local setting, there are multiple comparisons being made (neighbours sharing observations when calculating I_I) we need a Bonferroni adjustment

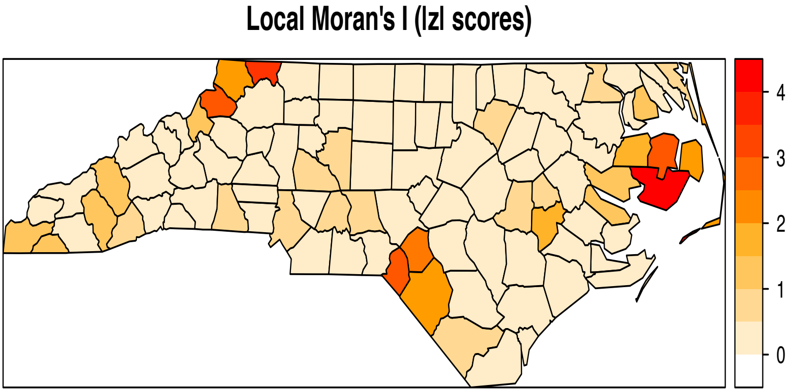
Areal Data: Local Indexes of Spatial Autocorrelation

Local Moran's I

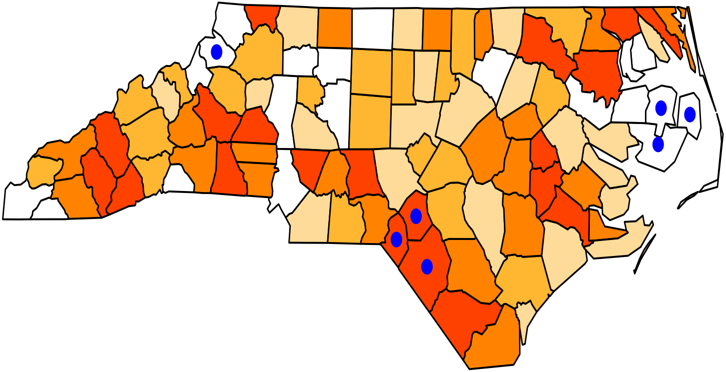
$$I_i = \frac{y_i - \bar{y}}{s} \sum_{j}^{n} w_{ij} \frac{(y_j - \bar{y})}{s}$$

- Have a value of I for each B_i
- Sometimes the I_i are mapped to indicate units with high values indicating stronger local autocorrelation
- More often, z-score and significance of z-score is plotted
- As before, test statistics are generated under randomization
- Since in a local setting, there are multiple comparisons being made (neighbours sharing observations when calculating I_I) we need a Bonferroni adjustment

Areal Data: Local Indexes of Spatial Autocorrelation



Statistically significant Local Moran's I as blue dots



Getis-Ord G vs Local Getis-Ord $G*$

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} y_i y_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j}$$

$$G_i* = \frac{\sum_{j=1}^{n} w_{ij} y_j}{\sum_{j=1}^{n} y_j}$$

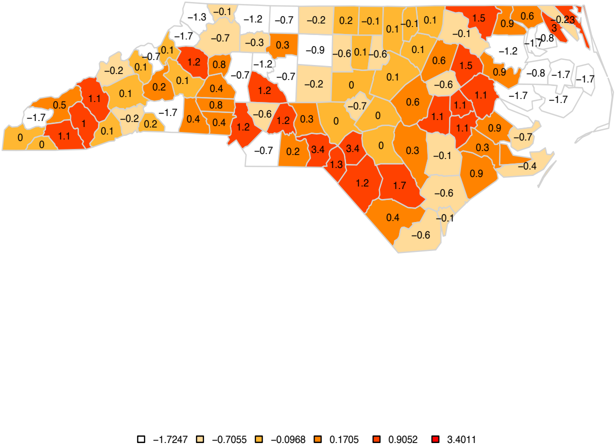
Getis-Ord G

- Again, we compute the z-score for spatially randomized G to determine if it is significantly different than our observed
- z=observed-expected/s.d. observed
- $E(G) = \frac{\sum_i \sum_j w_{ij}}{n(n-1)}$
- V(G) is complicated
- the sign of the z-score is important; positive z means high values cluster together, negative values means low values cluster together
- the p-value must be computed to determine significance of G

Local Getis-Ord G^*

- Similar to local Moran's I_i, G_i^* is calculated for each areal unit
- A group of areal units with high G_i^* indicates a "hotspot" where as low G_i^* means a "coldspot"

Getis-Ord G* for SIDS79 rates



Areal Data

Issues with spatial autocorrelation tests:

- They assume that the mean trend has been removed, for example household income effect when examining spatial pattern in SIDS rate
- One solution is to run a linear model and then test for spatial association on residuals
- Use autoregressive models

Areal Data: Models

Simultaneous Autoregressive models

- Similar to universal kriging or regression kriging
- Use regression on values from neighbouring areal units to account for spatial dependence
- Autocorrelation reflects self regression where you use observations of the outcome at other locations as additional covariates in the model
- $Y(s) \sim MVN(X\beta, \Sigma)$
- In universal kriging we modeled Σ as a parametric function of distance
- In areal modeling, we restrict distances to those between our areal units

Areal Data: Models

Simultaneous Autoregressive models

- We represent Σ as our residual errors $\epsilon(s_i) = \sum_j b_{ij}\epsilon(s_j) + \nu(s_i)$ and apply spatial correlation to these residuals

$$Y(s_i) = x(s_i)\beta + \sum_j b_{ij}\epsilon(s_j) + \nu(s_i)$$

$$Y(s_i) = x(s_i)\beta + \sum_j b_{ij}[Y(s_j) - x(s_j)\beta] + \nu(s_i)$$

The degree of spatial dependence is through the term $\sum_j b_{ij}[Y(s_j) - x(s_j)\beta]$

Areal Data: Models

Simultaneous Autoregressive models

- SAR models are often represented in matrix form
- From the equation on the previous slide,

$$Y = X^T\beta + B(Y - X^T\beta) + \nu$$