Measuring Spatial Patterns

- Determine whether or not a spatial pattern exists by DISTANCE
- Determine whether or not a spatial pattern exists by VALUES
- Tobler's 1st Law of Geography
  - Features that are closer together are more similar (in value/characteristics) than features that are further apart

author: Todd J. Schuble,
University of Chicago

Spatial Neighborhood

- Scale effects most spatial statistics
- Weights associated with features are usually dependent on how their "neighborhood" is defined
- Index the relative location and influence of all features
  - Define by adjacency (usually 1 or 0)
  - Define by distance (usually Euclidean)

author: Todd J. Schuble,
University of Chicago

Spatial Neighborhood

- A matrix is defined by the user or the GIS to determine the weight of influence
- Points are rarely adjacent to one another
- Polygons and lines only use adjacent "neighborhoods"
  - What about polygons that are adjacent on a corner?
  - Rook/Queen contiguity matrix
- Adjacent = 1, not-adjacent = 0

author: Todd J. Schuble,
University of Chicago

Spatial Neighborhood

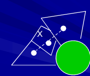
For Regular Polygons

rook case  or queen case 

For Irregular polygons

- All polygons that share a common border
- All polygons that share a common border or have a centroid within the circle defined by the average distance to (or the "convex hull" for) centroids of polygons that share a common border

For points



- The closest point (nearest neighbor)
  - Select the contiguity criteria
  - Construct weights matrix with 1 if contiguous, 0 otherwise

author: Todd J. Schuble,
University of Chicago

Spatial Neighborhood

- Distance neighborhoods can be defined in a variety of ways
- Points are easy to measure to-and-from
  - A to B
  - Points fall in or out of distance thresholds
- Polygons complicate matters
  - Distance from centroid or edge? Which edge?
  - Polygons fall completely within thresholds or partially within thresholds

author: Todd J. Schuble,
University of Chicago

Spatial Neighborhood

- Distance weight definitions
  - Direct measurement
  - Inverse distance
    - Weight = 1 / distance from A to B
  - Inverse distance squared
    - Weight = 1 / (distance from A to B)^2
- Proportional weights
  - Row-standardized weights
    - All areas given some influence

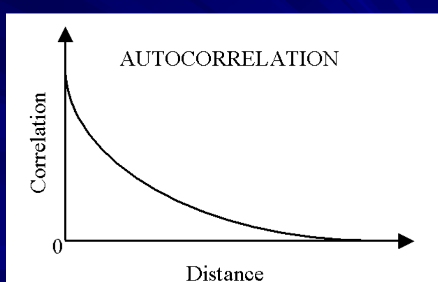
author: Todd J. Schuble,
University of Chicago

Spatial Autocorrelation

- Spatial autocorrelation
  - Is the distribution of values dependent on the distribution of features?
- Positive spatial autocorrelation: features surrounding each other have similar values
- Negative spatial autocorrelation: features surrounding each other have dissimilar values
- No spatial autocorrelation: feature values are randomly distributed
- Spatial autocorrelation can cause undetectable analysis errors
- Spatial autocorrelation is extremely useful and/or necessary in geographic methodology
- Most spatial data has positive spatial autocorrelation

author: Todd J. Schuble,
University of Chicago

Spatial Autocorrelation



author: Todd J. Schuble,
University of Chicago

Why Is Spatial Autocorrelation Bad?

- Ordinary Least Squares (OLS) regression assumes the independence and randomness of all observations
- Spatially autocorrelated observations violate these assumptions
  - Errors in estimation of regression coefficients
  - R^2 values not accurate

author: Todd J. Schuble,
University of Chicago

Why Is Spatial Autocorrelation Good?

- Sample sites are finite
  - Impossible to collect universal samples
- The assumption of spatial autocorrelation is used to interpolate areas to create continuous data samples
  - Methods: Inverse Distance Weighted (IDW), Kriging, etc.

author: Todd J. Schuble,
University of Chicago

Determining Spatial Autocorrelation

- Geary's ratio c statistic
  - Relates the difference between contiguous (or nearby) features to the differences in the entire set of features which comprises the study area
- A contiguity matrix is calculated
- The contiguity weight in the equation is only 0 or 1
  - 0 if the features are not within a certain distance or adjacent
  - 1 if the features are within a certain distance or adjacent

author: Todd J. Schuble,
University of Chicago

Geary's c

- Calculate the difference between feature values
- Calculate the how far feature values are from mean feature values (variance)
- Produce Geary's c ratio
  - If c is close to 0, features are clustered (small numerator, larger denominator)
  - If c is close to or > 1, features are more distributed (larger numerator, closer to the mean value)
- Test significance of Geary's c
  - Calculate a Z-score to see if features are clustered or dispersed (+ = dispersed, - = clustered)
  - $Z_c = c_o - c_e / SD_{ce}$

author: Todd J. Schuble,
University of Chicago

Moran's I

- Moran's I statistic
  - Relates the difference between feature values and their neighbors and the mean of all feature values ON AN INDIVIDUAL BASIS
- A contiguity matrix is calculated
- The contiguity weight in the equation is only 0 or 1
  - 0 if the features are not within a certain distance or adjacent
  - 1 if the features are within a certain distance or adjacent

author: Todd J. Schuble,
University of Chicago

Moran's I

- Calculate the difference between a feature's value and the mean AND the difference between it's neighbor's value and the mean
- Calculate the how far feature values are from mean values (variance)
- Produce Moran's I (index)
  - If -1 > I < 0, features' values are more distributed (pairs have less similar values)
  - If I = 0, features' values are distributed randomly
  - If 1 > I > 0, features' values are clustered (pairs have more similar values)
- Test significance of Moran's I
  - Calculate a Z-score to see if features are clustered or dispersed (- = dispersed, + = clustered)
  - $Z_I = I_o – I_e / SD_{ce}$

author: Todd J. Schuble,
University of Chicago

Dealing with Spatial Autocorrelation

- Descriptive spatial statistics
- Different types of data, lattice, point, geostatistical
- How to detect autocorrelation
  - Moran's I
  - Detect clusters
- How to deal with autocorrelation if it exists…which it always will

author: Todd J. Schuble,
University of Chicago

Dealing with Spatial Autocorrelation

- Spatial autocorrelation is useful to interpolate records for area that have not or cannot be sampled
- Spatial autocorrelation also causes data redundancy which produces misleading statistical results
- Resampling and filtering data are two of the most popular ways to combat the adverse effects of spatial autocorrelation

author: Todd J. Schuble,
University of Chicago

Dealing with Spatial Autocorrelation

- Data can be resampled using various means
  - 1) Group data according to preset data collection boundaries (counties, zip codes, tracts, etc.)
  - 2) Eliminate biased variance by restructuring boundaries and recategorizing data
    - Similar to Quadrat Analysis method
    - Use quadrat formula: side of quadrat = √2*(area/# of features)

author: Todd J. Schuble,
University of Chicago

Dealing with Spatial Autocorrelation

- Guarantee elimination of spatial autocorrelation
  - Resample data at a distance where autocorrelation is no longer significant
  - Use K-function (weighted) and calculate significance at different distances
  - Results will yield threshold distance

author: Todd J. Schuble,
University of Chicago

Local Spatial Statistics

- Patterns occur at different scales
- Determine where certain patterns occur across an entire study area
- Global measures identify if ALL features are (positively/negatively) spatially autocorrelated
- Local measures which features may be specifically contributing to spatial autocorrelation

author: Todd J. Schuble,
University of Chicago

Nearest Neighbor Hierarchical Clustering

- Finding cluster of clusters
- Calculate the threshold distance range
  - Choose a probability that features are not near each other simply due to chance….ex. 95%
  - At what threshold distance are features less likely to occur simply by chance?
    - Mean random distance = $0.5\sqrt{(\text{area of study area/number of features})}$
  - Standard error measures how much the mean random distance varies around its average
    - $SE = 0.26136/\sqrt{(n^2/A)}$

author: Todd J. Schuble,
University of Chicago

Nearest Neighbor Hierarchical Clustering

- A student t distribution is used based on your probability and the number of features (degrees of freedom) to give you a critical value
- The critical value is multiplied by the standard error
  - t * SE
- The product above is added or subtracted to the mean random distance
  - The THRESHOLD DISTANCE is the lower value of the range
  - $[t * (0.26136/\sqrt{(n^2/A)})] +/- (0.5\sqrt{(\text{area of study area/number of features})})$

author: Todd J. Schuble,
University of Chicago

Nearest Neighbor Hierarchical Clustering

- Measure distances between features
  - All features pairs separated by a distance < the threshold is assigned to a cluster
- 2nd order clusters are calculated using the centers of 1st order clusters
  - Confidence probabilities and threshold distances are calculated again
  - And so forth….

author: Todd J. Schuble, University of Chicago

Nearest Neighbor Hierarchical Clustering

- Why clusters occur
  - Analyze if the root is clustered
    - Ex. More illnesses are reported where there is higher population…more gas stations are located where there are more vehicles

author: Todd J. Schuble, University of Chicago

Local Moran's I

- Local Moran's I statistic is a modification of Moran's I
  - Developed by Luc Anselin (Arizona State) in the early 1990's
  - Different than Moran's index, the local variation identifies clusters of values within a study area
- Local Moran's I produces a statistic for each feature identifying if its "neighbors" are more alike or more different in value

author: Todd J. Schuble, University of Chicago

Local Moran's I

- Local Moran's I formula
  - z_i = difference between feature value and mean values
  - z_j = difference between neighbor's feature value and mean values
  - S = variance of value from the mean
  - w = weight of distance between features
  - $I_i = (z_i / s^2)\Sigma w_{ij}(z_j)$
- Z-score for local Moran's I is calculated to determine the likelihood that similarities or differences in values are not occurring by chance (p.170, Mitchell)

author: Todd J. Schuble,
University of Chicago

Local Moran's I

- Interpreting local Moran's I
  - Output is an index value and a Z-score for each feature
  - If index value is +, the feature has values similar to neighboring values
  - If index value is -, the feature is quite different from neighboring values
- Interpreting Z-score for local Moran's I
  - A high + Z-score for feature indicates the surrounding values are similar values
    - A group of adjacent features having high z scores indicates a cluster of similar values
  - A – Z-score for feature indicates the surrounding values are dissimilar values
  - Normal Z-score confidence levels are used to determine statistical significance of values

author: Todd J. Schuble,
University of Chicago

Local Moran's I

- Influencing local Moran's I results
  - Features with few neighbors, usually near edges of study area
  - Small numbers of features, usually under 30, can skew results if there are outliers
  - Z-scores can be misleading if many of the same neighbors are used for different features
    - No independence

author: Todd J. Schuble,
University of Chicago

Local G-statistic

- Related to the global statistic, general G
- Two versions of G-statistic, Gi and Gi*
  - Gi does not include target feature value
  - Gi* includes target feature value
- Both developed by Getis and Ord
- Gi* is more commonly used
  - Used for hot/cold spot analysis
  - Identifies clusters of high values and clusters of low values

author: Todd J. Schuble,
University of Chicago

Local G-statistic

- Related to the global statistic, general G
- Two versions of G-statistic, Gi and Gi*
  - Gi does not include target feature value
  - Gi* includes target feature value
- Both developed by Getis and Ord
- Gi* is more commonly used
  - Used for hot/cold spot analysis
  - Identifies clusters of high values and clusters of low values

author: Todd J. Schuble,
University of Chicago

Local G-statistic

- Local Moran's I and Gi* allows the user more "control" over the analysis
  - Distance measure must be inputted
  - Distance usually based on behavior of data
  - Helpful hint: run spatial autocorrelation test at certain distances, distance with highest Z-score should be used in test

author: Todd J. Schuble,
University of Chicago

Local G-statistic

- Interpreting Gi*
  - High Gi* values indicate a cluster of features with high values
  - Low Gi* values indicate a cluster of features with low values
  - Gi* values near zero indicate no concentration
- Interpreting Z-score for Gi*
  - High Z-scores indicate a cluster of features with high values
  - Low Z-scores indicate a cluster of features with low values
  - Z-scores near zero indicate no concentration

author: Todd J. Schuble,
University of Chicago