spatial randomness

positive and negative spatial autocorrelation

spatial autocorrelation statistics

spatial weights
Spatial Randomness
• The Null Hypothesis

  spatial randomness is absence of any pattern

  spatial randomness is not very interesting

  if rejected, then there is evidence of spatial structure
Interpreting Spatial Randomness

observed spatial pattern of values is equally likely as any other spatial pattern
value at one location does not depend on values at other (neighboring) locations
• Operationalizing Spatial Randomness

under spatial randomness, the location of values may be altered without affecting the information content of the data

random permutation or reshuffling of values
• Tobler’s First Law of Geography

  everything depends on everything else, but
  closer things more so

  structures spatial dependence

  importance of distance decay
Positive and Negative Spatial Autocorrelation
• Rejecting the Null Hypothesis

rejecting spatial randomness (s.r.)

like values in neighboring locations occur more frequently than for s.r.
= positive spatial autocorrelation

dissimilar (e.g., high vs low) in neighboring locations occur more frequently than for s.r.
= negative spatial autocorrelation
• Positive Spatial Autocorrelation

impression of clustering

clumps of like values

like values can be either high (hot spots) or low (cold spots)

difficult to rely on human perception
Quantile: ZAR09
-5.711 - 1.931 (16)
-1.802 - 1.099 (16)
-1.048 - 0.2079 (17)
0.2648 - 1.375 (17)
1.393 - 2.497 (17)
2.605 - 4.92 (17)

< positive s.a.

random >
• Negative Spatial Autocorrelation

  checkerboard pattern

  hard to distinguish from spatial randomness
Quantile: ZARN09

< negative s.a.

random >
Spatial Autocorrelation
Statistics
• What is a Test Statistic?

  a statistic is any value that summarizes characteristics of a distribution calculated from the data

  test statistic: calculated from the data and compared to a reference distribution

  how likely is the value if it had occurred under the null hypothesis (spatial randomness)

  when unlikely (low p value) the null hypothesis is rejected
compare value of test statistic to its distribution under the null hypothesis of spatial randomness
• **Spatial Autocorrelation Statistic**

  captures both attribute similarity and locational similarity

  how to construct an index from the data that captures both attribute similarity and locational similarity (i.e., neighbors are alike)
• Attribute Similarity

summary of the similarity (or dissimilarity) of observations for a variable at different locations

variable y

locations i, j

how to construct $f(y_i, y_j)$
• **Similarity Measure**

• cross product: \( y_i \cdot y_j \)

under randomness, cross product is not systematically large or small

when large values are systematically together, product will be larger, and vice versa
Dissimilarity Measure

- squared difference: \((y_i - y_j)^2\)
- absolute difference: \(|y_i - y_j|\)

Under randomness, difference measure will not be systematically large or small.

When small values or large values are systematically together, difference measures will be smaller.
• **Locational Similarity**

formalizing the notion of neighbors
= spatial weights \((w_{ij})\)

when are two spatial units \(i\) and \(j\) a priori likely to interact

not necessarily a geographical notion, can be based on social network concepts or general distance concepts (distance in multivariate space)
• General Spatial Autocorrelation Statistic

general form

sum over all observations of an attribute similarity measure with the neighbors

\[ \text{statistic} = \sum_{ij} f(x_i x_j)w_{ij} \]

\( f(x_i x_j) \) is attribute similarity between \( i \) and \( j \) for \( x \)

\( w_{ij} \) is a spatial weight between \( i \) and \( j \)
Spatial Weights
Basic Concepts
• Why Spatial Weights

formal expression of locational similarity

spatial autocorrelation is about interaction

n x (n - 1)/2 pairwise interactions but only n observations in a cross-section

insufficient information to extract pattern of interaction from cross-section

example: North Carolina has 100 counties 5,000 pairwise interactions, 100 observations
Solution

- impose structure

limit the number of parameters to be estimated

incidental parameter problem = number of parameters grows with sample size

for spatial interaction, number of parameters grows with $n^2$
• **Spatial Weights**

  - exclude some interactions
  - constrain the number of neighbors, e.g., only those locations that share a border
  - single parameter = spatial autocorrelation coefficient

• **strength of interaction = combined effect of coefficient and weights**

  - small coefficient with large weights
  - large coefficient with small weights
six polygons - neighbors share common border
neighbor structure as a graph
• Spatial Weights Matrix Definition

N by N positive matrix $W$ with elements $w_{ij}$

$w_{ij}$ non-zero for neighbors

$w_{ij} = 0$, i and j are not neighbors

$w_{ii} = 0$, no self-similarity
spatial weights matrix $W$ with elements $w_{ij}$
Geography-Based Spatial Weights
• Binary Contiguity Weights

contiguity = common border

i and j share a border, then $w_{ij} = 1$

i and j are not neighbors, then $w_{ij} = 0$

weights are 0 or 1, hence binary
The binary contiguity weights matrix for the six-region example is given by:

\[
W = \begin{bmatrix}
0 & 1 & 0 & 1 & 1 & 0 \\
1 & 0 & 0 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 \\
1 & 1 & 0 & 0 & 1 & 0 \\
1 & 1 & 1 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0
\end{bmatrix}.
\]
contiguity on a regular grid - different definitions
rook contiguity - edges only
2, 4, 6, 8 are neighbors of 5
bishop contiguity - corners only
1, 3, 7, 9 are neighbors of 5
Rook Contiguity

Neighbors for 5: 2, 4, 6, 8

Common Border

Queen Contiguity

8 Neighbors for 5

Both Border and Vertex

queen contiguity - edges and corners
5 has eight neighbors
Solano county, CA contiguity

rook

obs 1448 has 3 neighbors: 1263, 1282, 1302

queen

obs 1448 has 4 neighbors: 1263, 1282, 1283, 1302
• Distance-Based Weights

distance between points

distance between polygon centroids or central points

in general, can be any function of distance that implies distance decay, e.g., inverse distance

in practice, mostly based on a notion of contiguity defined by distance
• **Distance-Band Weights**

\[ w_{ij} \text{ nonzero for } d_{ij} < d \]

less than a critical distance \( d \)

potential problem: isolates = no neighbors

make sure critical distance is max-min, i.e., the largest of the nearest neighbor distance for each observation
distance-band weights

Solano county, CA, distance-band neighbors
\[ d = 90 \text{ mi} \]
distance-band weights

Elko county, NV, distance band neighbors
• k-Nearest Neighbor Weights

k nearest observations, irrespective of distance

fixes isolates problem for distance bands

same number of neighbors for all observations

in practice, potential problem with ties

needs tie-breaking rule (random, include all)
k-nearest neighbor weights

k = 4

k = 6

Elko county, NV, k-nearest neighbors
Table 1. Spatial weights formats supported by PySAL.

<table>
<thead>
<tr>
<th>Type</th>
<th>File extension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sparse contiguity (SpaceStat, GeoDa, R spdep, etc.)</td>
<td>GAL</td>
</tr>
<tr>
<td>Sparse general weights (SpaceStat, GeoDa, R spdep, etc.)</td>
<td>GWT</td>
</tr>
<tr>
<td>ArcGIS text weights</td>
<td>TXT</td>
</tr>
<tr>
<td>ArcGIS dbf weights</td>
<td>DBF</td>
</tr>
<tr>
<td>ArcGIS swm weights</td>
<td>SWM</td>
</tr>
<tr>
<td>Matlab spatial weights (old version)</td>
<td>DAT</td>
</tr>
<tr>
<td>Matlab spatial weights (new version)</td>
<td>MAT</td>
</tr>
<tr>
<td>Lotus weights</td>
<td>WK1</td>
</tr>
<tr>
<td>GeoBUGS weights</td>
<td>TXT</td>
</tr>
<tr>
<td>Stata weights</td>
<td>TXT</td>
</tr>
<tr>
<td>MatrixMarket weights</td>
<td>MTX</td>
</tr>
</tbody>
</table>

many spatial weights file formats
Spatial Weights Transformations
Row-Standardized Weights

rescale weights such that \( \sum_j w_{ij} = 1 \)

\[ w_{ij}^* = w_{ij} / \sum_j w_{ij} \]

constrains parameter space

makes analyses comparable

spatial lag = average of the neighbors
\[ W^* = \begin{bmatrix}
0 & 1/3 & 0 & 1/3 & 1/3 & 0 \\
1/3 & 0 & 0 & 1/3 & 1/3 & 0 \\
0 & 0 & 0 & 0 & 1/2 & 1/2 \\
1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\
1/4 & 1/4 & 1/4 & 1/4 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
\end{bmatrix}. \]

row-standardized weights matrix
• **Stochastic Weights**

  double standardization

  \[ w_{ij}^* = \frac{w_{ij}}{\sum_i \sum_j w_{ij}} \]

  rescaled such that \( \sum_i \sum_j w_{ij} = 1 \)

  similar to probability
\[ W^* = \begin{bmatrix}
0 & 1/16 & 0 & 1/16 & 1/16 & 0 \\
1/16 & 0 & 0 & 1/16 & 1/16 & 0 \\
0 & 0 & 0 & 0 & 1/16 & 1/16 \\
1/16 & 1/16 & 0 & 0 & 1/16 & 0 \\
1/16 & 1/16 & 1/16 & 1/16 & 0 & 0 \\
0 & 0 & 1/16 & 0 & 0 & 0 
\end{bmatrix} . \]

stochastic weights matrix
second order contiguity: neighbor of neighbor
redundancy in higher order contiguity
paths of length 2 between 1 and other cells
• Higher Order Weights

  recursive definition

• k-th order neighbor is first order neighbor of (k-1)th order neighbor

  avoid duplication, only unique neighbors of a given order (not both first and second order)

  pure contiguity or cumulative contiguity, i.e., lower order neighbors included in weights
Solano county, CA, second order contiguity

exclusive of first order

obs 1448 has 9 neighbors: 1006, 1102, 1142, 1169, 1173, 123

inclusive of first order

obs 1448 has 13 neighbors: 1006, 1102, 1142, 1169, 1173, 1234, 1
Properties of Weights
• Connectivity Histogram

histogram of number of neighbors

neighbor cardinality

diagnostic for “isolates” or neighborless units

assess characteristics of the distribution
Things to Watch for

isolates

need to be removed for proper spatial analysis

do not need to be removed for standard analysis

very large number of neighbors

bimodal distribution
connectivity histogram - contiguity weights
U.S. counties
Connectivity histogram - contiguity weights
U.S. counties

default distance 90 mi
critical distance 80 mi
contiguity histogram for k-nearest neighbors
or, what did you expect?