Areal Data Spatial Analysis

D G Rossiter

Cornell University

March 11, 2020
Outline

1. Areal data
   - Definition and examples
   - Characteristics
   - The "ecological fallacy"
   - Neighbours

2. Spatial autocorrelation
   - Global Moran’s I
   - Autocorrelation of categorical variables
   - Local Moran’s I
   - Hot-spot analysis

3. GeoDa and LISA
   - Exploratory graphics
   - Clustering
   - Weights and neighbours
   - Spatial correlation
   - Spatial regression

4. Spatially-explicit linear models

5. References
1 Areal data
   - Definition and examples
   - Characteristics
   - The "ecological fallacy"
   - Neighbours

2 Spatial autocorrelation
   - Global Moran’s I
   - Autocorrelation of categorical variables
   - Local Moran’s I
   - Hot-spot analysis

3 GeoDa and LISA
   - Exploratory graphics
   - Clustering
   - Weights and neighbours
   - Spatial correlation
   - Spatial regression

4 Spatially-explicit linear models

5 References
Data are presented as attributes of **fixed polygonal areas**
- generally irregularly-shaped, and/or not all same shape
- examples: census blocks, voting districts, forest parcels . . .
- but methods can apply to regular grids

Attributes can be analyzed in feature space (distribution, correlation, regression . . .) but:

Q: Is the data-generating **process**:
- non-spatial (all in feature space),
- spatial (all in geographic space), or
- mixed?

Q: If mixed, how does the **spatial structure** affect the **feature-space structure**?
Typical applications

- spatial econometrics [2]
- epidemiology [6, §11] [8]
- sociology / demographics [10]
- political science [18]
- natural resources, if data are presented as areal aggregates
  - forest management blocks, farms, ...
Example: Syracuse (NY) census and health

Syracuse city, relative Leukemia incidence

Syracuse city, % over 65

source: Bivand et al. [6, §9]

Q: Is leukemia incidence in a census tract correlated with mean age in the tract? Or are there local “hot spots” that might have a point-source cause?
This by rank, not relative incidence
Real world

Google Earth view of census tract boundaries (KML file); can zoom and pan
Characteristics of Syracuse leukemia data

Typical of most areal data:

- **Aggregated by reporting unit**
  - here, US census tracts; within City boundary
- Units were not designed for our purpose
  - here, study of the causes of leukemia
  - size, geographic and feature-space characteristics
- **Uneven size and shape** of units
- Different numbers of neighbours, lengths of common borders
- Units on edges have **unobserved neighbours**
- Uneven feature-space “size”
  - e.g., population, proportion residential vs. commercial
- “Points” (e.g., industry) assigned to the whole polygon
Example: Favourite NFL team by county (2012)


Question: What factor(s) determine this in feature & geographic spaces?
See next slide.
What factors control which team is the favourite?

- Team’s success (over what period?)
- Team’s games shown on local TV?
- Team featured often on national TV?
- Team plays in county’s state?
- If no team in state, team plays in neighbouring state?
- Team plays in migrants’ home state?
- Proximity of county to team’s stadium?
- Demographic factors (occupation, ethnicity)?
- Popular players on team from locality/local college?
- **Is there residual spatial correlation** after accounting for these factors? “Spillover effect”. 
The attributes relate to the whole area of the polygon, and can not be further localized.

- Various methods of dis-aggregation using covariates with finer spatial resolution
- e.g., satellite imagery to separate industrial and residential areas within one polygon

Often the attributes are aggregate measures
- e.g., population count, proportions

The attributes may already be normalized to the area of the polygon
- e.g., population density

Metadata is vital for proper processing and interpretation
- especially aggregation method
tessellation = division of the study area

- Can not be dis-aggregated by analyst

- The tessellation may have been done for a purpose not directly relevant to the analysis
  - E.g., crop yield statistics may be aggregated by political division, but the crop yield may be better modelled by agro-ecological zone changes to boundaries → longitudinal analyses
  - e.g., British county / authority boundaries; area code zones

- Metadata is vital for proper processing and interpretation
  - especially tessellation decisions, consistency
Characteristics of areal data – choice of scale

- The **scale** of the tessellation affects the analysis
  - a variation of the **bandwidth** problem for spatial fields
  - e.g., voting patterns by state vs. congressional district vs. county vs. ward; relation between e.g., family income and political preference
  - e.g., crop statistics by county may show strong spatial autocorrelation, which becomes much weaker at district or state level, although the underlying process is the same.

- Technical term: **modifiable areal unit problem** [12]
Example: Mexican electoral regions

Source: [18, Fig. 1]
Two levels of aggregation: state, region
Question: what socio-economic factors determine voting patterns?
Example: New Jersey housing

Source: [8, Fig. 1]
Percentage of homes built before 1950 (risk factor for Pb poisoning)
Aggregation level: USA census block group, ZIP code, county.
ZIP code

“hot spots” of 50–75% older houses

County

all 25–50% older houses
no hot spots
The “ecological fallacy” – non-spatial (1)

- “Ecological” = context of observations: “In an ecological correlation the statistical object is a group of persons”
- Fallacy: inferences about a fine-scale grouping can be deduced from inferences for a coarse-scale grouping
  - E.g.: regression/correlation of voting preferences based on socio-economic factors at state/province level vs. same relations at county level.
  - The aggregate relation (at states level) can not be obtained by aggregating fine-scale regressions (50 per-state relations)

- References: [16, 17];
  
  *Ecological Fallacy In: Encyclopedia of Survey Research Methods*

  https://dx.doi.org/10.4135/9781412963947.n151
Fallacy: inferences about **individuals** ("individual-level correlations/regressions") can be deduced from inference for their group.

- E.g., Strong empirical-statistical relation between age of schoolchildren and height does *not* imply that a randomly-selected 5th grader is taller than a randomly-selected 4th grader.
The “ecological fallacy” – spatial

- Fallacy: inferences about aggregate data at small area can be deduced from inferences about aggregate data for an enclosing larger area or from inferences from all individual observations.
  - E.g.: strong empirical-statistical relation between crime and size of police force (both normalized for population) at state level; does not imply that there is a strong relation at city level within a single state or overall.

- Message: analyze at the level that you want to understand / make policy.
The “ecological fallacy” and the MAUP

- Correlations at more general levels are generally **stronger** (higher $|r|$) than at finer levels.
- Regressions at more general levels are generally **stronger** (higher $R^2$) than at finer levels.
- This is because much noise has been averaged out.
- Factor for correlations: \[\frac{1-\sigma_{XA}\sigma_{YA}}{\sqrt{1-\sigma_{XA}^2} \sqrt{1-\sigma_{YA}^2}}\]
  - $\sigma_{XA}$, $\sigma_{YA}$: variation of the two variables $X$ and $Y$ between strata;
  - minimum possible value $= 1$ when there is no variation between strata.
Q: how do we quantify “nearby”? 

A1: **distance** between **centroids** of polygons 
   - as with spatial fields; represents polygons as points 
   - can use inverse distance, $ID^2 \ldots$

A2: common borders: **neighbours** (1\textsuperscript{st} order) 
   - “rook” (common line) vs. “queen” (common point) neighbours 
   - terminology from legal chess moves

A3: number of **steps** to reach a common border 
   - 1\textsuperscript{st}, 2\textsuperscript{nd}, 3\textsuperscript{rd}… **order** neighbours

Distance or steps? depends on purpose of analysis 
   - what is supposed to drive the **spatial process**?
R packages for areal data

- `sp` “Classes and Methods for Spatial Data” (Edzer Pebesma\(^1\), Roger Bivand\(^2\))
- `spdep` “Spatial Dependence: Weighting Schemes, Statistics and Models” (Roger Bivand)
- `splm` “Econometric Models for Spatial Panel Data” (Giovanni Millo\(^3\))

\(^1\)University of Münster (D)  
\(^2\)NHH: Norwegian school of economics  
\(^3\)Generali insurance
Syracuse city census tracts, queen and rook neighbours

Neighbours example
Finding neighbours

spdep functions:

- **knearneigh** find $k$ nearest neighbours for each polygon (class knn)
- **knn2nb** convert these to weights (class nb “neighbour list”)
- **dnearneigh** identify neighbours within a given distance band (class nb)
- **nbdist** Distances along each link of a neighbour list.
Nearest neighbours within a distance

Syracuse city census tracts, 1.2 km centroid neighbours
$k$ nearest neighbours

Syracuse city census tracts, nearest neighbour

Syracuse city census tracts, 2 nearest neighbours

Syracuse city census tracts, 3 nearest neighbours

Syracuse city census tracts, 4 nearest neighbours
In models of spatial processes (see below) neighbours are presumed to have influence on a target polygon.

- See examples in spatial autocorrelation and spatial modelling (below).

A neighbour can be more or less influential to a target polygon, depending on the spatial process.

So, assign a weight to each link in the graph → (a)symmetric weights matrix.

Weights style depends on presumed process (see next) – there is no “correct” weighting.
Weighting styles

- **Style B (binary):** weights of adjacent polygons affecting a target polygon are either 0 (not a neighbour of the target) or 1 (is a neighbour)
  - Implies process depends on the number of neighbours
  - Can also use with weighting based on distances between centroids: multiply 1’s by some distance measure

- **Style W:** weights of adjacent polygons affecting a target polygon must sum to 1 (row-standardized)
  - All $n$ neighbours equally influential $\rightarrow$ all weights $1/n$.
  - i.e., total influence to a target area is constant, influence from neighbours divided among them
  - Links originating at areas with few neighbours $\rightarrow$ larger weights (edge effect).
Assigning weights

spdep functions:

- **nb2listw** spatial weights for neighbours lists (class `listw`, `nb`); styles `W`, `B`, `C`, `U`, `S`
  
  `W` row-standardized
  
  `B` binary
  
  `C` globally-standardized: sum over all links to `n`
  
  `U` `C` divided by number of neighbours
  
  `S` variance-stabilizing

- **glist** argument to `nb2listw`: pass a list of vectors of weights corresponding to the neighbour relationships
  
  - example: pre-computed inverse-distance, ID$^2$W with `nbdist`s; use style `B`, will modify “binary” weights

- **listw2mat** show weights matrix
Example weights matrix – style ‘B’

1 = is a neighbour; 0 = not; by definition **symmetric**
Example weights matrix – ‘B’ with Inverse-distance weighting

<table>
<thead>
<tr>
<th></th>
<th>109</th>
<th>110</th>
<th>111</th>
<th>112</th>
<th>113</th>
<th>114</th>
<th>115</th>
<th>116</th>
<th>117</th>
</tr>
</thead>
<tbody>
<tr>
<td>109</td>
<td>0.0000</td>
<td>0.6035</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.6602</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>110</td>
<td>0.6035</td>
<td>0.0000</td>
<td>0.9265</td>
<td>0.5963</td>
<td>1.0111</td>
<td>1.4139</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>111</td>
<td>0.0000</td>
<td>0.9265</td>
<td>0.0000</td>
<td>0.9858</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>112</td>
<td>0.0000</td>
<td>0.5963</td>
<td>0.9858</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.7191</td>
<td>1.0020</td>
<td>1.1118</td>
<td>0.7829</td>
</tr>
<tr>
<td>113</td>
<td>0.6602</td>
<td>1.0111</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.3676</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>114</td>
<td>0.0000</td>
<td>1.4139</td>
<td>0.0000</td>
<td>0.7191</td>
<td>1.3676</td>
<td>0.0000</td>
<td>1.7476</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>115</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0020</td>
<td>0.0000</td>
<td>1.7476</td>
<td>0.0000</td>
<td>1.7162</td>
<td>0.0000</td>
</tr>
<tr>
<td>116</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.1118</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.7162</td>
<td>0.0000</td>
<td>1.0592</td>
</tr>
<tr>
<td>117</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.7829</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0592</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Neighbours weighted by inverse distance to centroids; e.g., \((110, 111)\) closer pair then \((110, 109)\), so 109 will have less influence on 110 than will 111.
Example weights matrix – style ‘W’

\[
\begin{array}{cccccccccc}
109 & 110 & 111 & 112 & 113 & 114 & 115 & 116 & 117 \\
109 & 0.0000 & 0.2000 & 0.0000 & 0.0000 & 0.2 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\
110 & 0.2000 & 0.0000 & 0.2000 & 0.2000 & 0.2 & 0.2000 & 0.0000 & 0.0000 & 0.0000 \\
111 & 0.0000 & 0.5000 & 0.0000 & 0.5000 & 0.0 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\
112 & 0.0000 & 0.1429 & 0.1429 & 0.0000 & 0.0 & 0.1429 & 0.1429 & 0.1429 & 0.1429 \\
113 & 0.1429 & 0.1429 & 0.0000 & 0.0000 & 0.0 & 0.1429 & 0.0000 & 0.0000 & 0.0000 \\
114 & 0.0000 & 0.2000 & 0.0000 & 0.2000 & 0.2 & 0.0000 & 0.2000 & 0.0000 & 0.0000 \\
115 & 0.0000 & 0.0000 & 0.0000 & 0.2500 & 0.0 & 0.2500 & 0.0000 & 0.2500 & 0.0000 \\
116 & 0.0000 & 0.0000 & 0.0000 & 0.2000 & 0.0 & 0.0000 & 0.2000 & 0.0000 & 0.2000 \\
117 & 0.0000 & 0.0000 & 0.0000 & 0.1429 & 0.0 & 0.0000 & 0.0000 & 0.1429 & 0.0000 \\
\end{array}
\]

\[0.2 = \frac{1}{5} \text{ equal weight to the 5 neighbours of target polygon 109;}\]
\[0.14286 = \frac{1}{7} \text{ equal weight to the 7 neighbours of target polygon 112 . . .}\]

Rows sum to unity; \(W\) is not necessarily symmetric.
1. Areal data
   - Definition and examples
   - Characteristics
   - The "ecological fallacy"
   - Neighbours

2. Spatial autocorrelation
   - Global Moran’s I
   - Autocorrelation of categorical variables
   - Local Moran’s I
   - Hot-spot analysis

3. GeoDa and LISA
   - Exploratory graphics
   - Clustering
   - Weights and neighbours
   - Spatial correlation
   - Spatial regression

4. Spatially-explicit linear models

5. References
Tobler’s first law of geography (1970): “Everything is related to everything else, but near things are more related than distant things”
- not always true!! It depends on the process that generated the spatial distribution of “everything”

“Auto” = the same feature-space attribute

Question 1: finding if this is true for a given attribute; quantifying the range and degree of autocorrelation.

Question 2: finding out why – really “auto” or due to some other spatially-distributed attribute?
Moran’s $I$ – motivation

- Q: are attribute values in neighbouring polygons (suitably weighted) more similar than is expected by chance?
  - A: using centroids and inverse distance as weights: variograms or correlograms
  - A: considering (weighted) neighbours: Moran’s $I$

Assumption: no spatial patterning due to some underlying spatial factor
  - i.e., apparent spatial correlation in this variable is *not* due to actual spatial correlation of another variable
  - This can be tested in simultaneous autoregressive model (SAR), see below.

Assumption: the assigned neighbour weights are appropriate to the process
Moran’s $I$ – formula

$$ I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2} \tag{1} $$

- Variables:
  - $y_i$ is the value of the variable in the $i$th of $n$ polygons
  - $\bar{y}$ is the global mean of the variable
  - $w_{ij}$ is the spatial weight of the link between polygons $i$ and $j$

- The second term numerator is the weighted covariance; the denominator normalizes by the variance

- The first term normalizes by the sum of all weights $\rightarrow$ the test is comparable among datasets with different numbers of polygons and using different weightings.
Global Moran’s $I$ test

- Compute for **all pairs** of polygons $(i, j)$
  - Test is about correlation across the **whole map** – is there any patterning anywhere?

- Assign weights according to hypothesis
  - 1$^{\text{st}}$ order: only immediate neighbours (rook? queen?) have non-zero weights
  - 2$^{\text{nd}}$, 3$^{\text{rd}}$... order: zero weights for immediate, 2$^{\text{nd}}$... neighbours, then non-zero weights for the next “ring” (boundary crossing)

- Expected value if random placement of response variable $-1/(n-1)$; complicated formula for variance

- Transform observed $I$ to a normal $Z$ score, compute probability it is by chance that different from the value expected if random allocation of the attribute value to polygons
Example: Syracuse leukemia

Equally-weighted first (rook) neighbours:

Moran’s I Expectation Variance
0.2075836 -0.0161290 0.0050781

Moran’s I test under randomisation
alternative hypothesis: greater
Moran I standard deviate = 3.1394
p-value = 0.000846

Conclusion: reject null hypothesis, there is **positive** spatial autocorrelation of leukemia incidence *across the map*. **Note** we have made no attempt (yet) to explain *why*.
Effect of weights

These represent different **hypotheses** about the relative importance of neighbours in the spatial process.

- **W** inversely proportional to the number of neighbours;
  - more weight to areas with few neighbours
- **B** binary: 1 for a neighbour, 0 otherwise;
- **C** globally standardized: inversely proportional to the total number of links;
- **IDW** inverse-distance to centroids

Syracuse leukemia Moran’s I with different weights:

<table>
<thead>
<tr>
<th>Style</th>
<th>Moran’s I</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>0.207</td>
<td>0.0008</td>
</tr>
<tr>
<td>B</td>
<td>0.224</td>
<td>0.0002</td>
</tr>
<tr>
<td>IDW</td>
<td>0.195</td>
<td>0.0018</td>
</tr>
</tbody>
</table>
Autocorrelation of categorical variables

- “BB join count”
- Analogous to Moran’s I for continuous attributes
- “BB” = “black/black” vs. “BW” = “black/white” etc., but can have more “colours” (categories)
- Similar to a contingency table for non-spatial attributes
- Tests whether same “colour” joins (mergers) occur more frequently than would be expected by chance (i.e., if the colours were randomly assigned to areas)
- Sensitive to the definition of neighbours and weights
- Sensitive to MAUP and aggregation method
  - mode (most common), nearest (centre)
- R package spdep, function joincount.test etc.
Example “BW” patterns

global Moran’s I with binary (0/1) weighting:

Separated
\[ I = +1 \]

Even
\[ I = -1 \]

Random
\[ I = -1/(36 - 1) = -0.028 \]

Local Moran’s I

- Compute Moran’s I for each polygon separately:

$$I_i = \frac{(y_i - \bar{y}) \cdot \sum_j (y_j - \bar{y})}{1/n \cdot \sum_i (y_i - \bar{y})^2}$$ (2)

- The denominator ensures that $\sum_i I_i = I$

- Show these on a scatterplot as Moran’s I (x-axis) vs. the average Moran’s I of all neighbours of the polygon

- The slope of the regression between these is global Moran’s I!

- Identifies “hot” and “cold” spots of spatial correlation that contribute most to the global Moran’s I
Example: Syracuse leukemia (1)

Moran scatterplot, Syracuse leukemia incidence

Slope of regression is global Moran’s $I$
Point numbers are polygon IDs.

**x-axis:** Leukemia in a district
**y-axis:** Leukemia weighted-averaged in neighbour districts
Marked points have high leverage (influence on global Moran’s $I$)
Example: Syracuse leukemia (2)

HH: the tract has high incidence, so do its neighbours; etc.
Topic: Hot-spot analysis

Question: are there portions of the study area with consistently higher (“hot”) or lower (“cold”) attribute values than average?

- **Point** data: interpolate from point values over a ‘fine” grid
  - kriging is a smoothing interpolator and will by construction show clusters
- **Area** data: compare areas to average
  - local Moran’s I
  - Getis-Ord local $G$
Getis-Ord local $G$ statistics

- Symbolized as $G_i$ and $G_i^*$; the subscript $i$ emphasizes that they are computed separately for each area.
- No attempt to characterize overall spatial dependency.
- They identify local areas where there may be dependency.
  
  "These statistics are especially useful in cases where global statistics may fail to alert the researcher to significant pockets of clustering." – Ord and Getis [15]

- Two variants: $G_i$ and $G_i^*$, where the ‘starred’ variant includes the self-weights $w_{ii}$ of each target polygon
  
  - $G_i$ shows whether an area is within a surrounding hot spot
  - $G_i^*$ shows whether the area itself is part of such a hot spot.
Getis-Ord local $G$ – original formulation

A simple concept [9]:

$$G_i(d) = \frac{\sum_j w_{ij}(d) \cdot x_j}{\sum_j x_j}$$

(3)

- the values of the target attribute
- $i$ index of the local area
- $j$ index running over all local areas, not including area $i$
- $d$ buffer distance, selected by analyst
- $w$ symmetric 0/1 matrix: 1 $\rightarrow$ area $j$ is within distance $d$ of area $i$; but $w_{ii} = 0$

$G_i(d)$ is the proportion of the total of an attribute within distance $d$ of target area $i$.

$G_i^*(d)$ includes the target area in the index $j$. 
Getis-Ord local $G$ – revised formulation

- Generalize [15] to any weighting, not just 0/1 and not just based on distances
- So it can use the same weighting styles as for Moran’s $I$
- Define as a standard (normal) variate
  - original $G_i$ less its expectation $W_i = \sum_{j \neq i} w_{ij} / (n - 1)$ . . .
  - . . . divided by the square root of the variance:

$$\text{Var}(G_i) = \frac{W_i(n - 1 - W_i)}{(n - 1)^2(n - 2)} \cdot \left[ \frac{s(i)}{\bar{x}(i)} \right]^2$$
Getis-Ord local $G$ – revised formula

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j}x_j - \bar{x} \sum_{j=1}^{n} w_{i,j}}{\text{Var}(G_i^*)^{1/2}}$$ (4)

- $\bar{x}, s$ are sample mean and standard deviation of the target variable; $n$ areas
- the $w_{i,j}$ are the neighbour weights
- the numerator shows the difference between area $j$’s weighted average of the target and the overall weighted average
- the denominator standardizes the index
- interpret as Z-score: $+$ $\rightarrow$ clustering of high values, $-$ $\rightarrow$ clustering of low values
Getis–Ord Gi, Syracuse leukemia incidence

Getis–Ord Gi*, Syracuse leukemia incidence
Outline

1. Areal data
   - Definition and examples
   - Characteristics
   - The "ecological fallacy"
   - Neighbours

2. Spatial autocorrelation
   - Global Moran’s I
   - Autocorrelation of categorical variables
   - Local Moran’s I
   - Hot-spot analysis

3. GeoDa and LISA
   - Exploratory graphics
   - Clustering
   - Weights and neighbours
   - Spatial correlation
   - Spatial regression

4. Spatially-explicit linear models

5. References
LISA

- Expanded and implemented in the GeoDa computer program⁴ “Exploratory Spatial Data Analysis & spatial regression”
- Attractive interface to these techniques
- The GeoDa program, documentation and sample data are freely available for download from the Geodata Center’s GitHub⁵.

---

⁴http://spatial.uchicago.edu/geoda
⁵http://geodacenter.github.io
Univariate exploratory graphics quantile plot

≈ equal numbers of observations in each quantile
Algorithm to minimize the within-class/between-class variance – equivalent to univariate k-means
2nd and 3rd quartiles (half the observations); hinges = 1.5 × Interquartile range; outside this are “boxplot outliers”
Box plot map with two hinge limits

Hinge = 3.0 only shows the most extreme.
Bivariate exploratory graphics: cartogram

Shows one variable by size, the other by colour, space by the centroids
Clustering

- **Objective**: group spatial units (e.g., census tracts) into “homogeneous” groups, according to their **feature-space** attributes
  - can also include **coordinates of centroids** as attributes to force **geographic compactness**

- **Method**: k-means
  - **one-step**: minimize within-class variance, maximize between-class variance; analyst fixes number of classes (k)

- **Method**: hierarchical clustering
  - **bottom-up** grouping to form increasingly-larger groups
  - each grouping has a “distance” between its members
  - can “cut” the dendrogram (graph) at any level to form any number of groups.
Clustering: univariate k-means (1)

Algorithm to minimize the within-class/between-class variance
Algorithm to minimize the within-class/between-class variance
Algorithm to minimize the within-class/between-class variance, while forcing clusters to be *spatially-contiguous*
Clustering: multivariate geographic k-means (2)

Algorithm to minimize the within-class/between-class variance, while forcing clusters to be spatially-contiguous
Clustering: multivariate hierarchical: specification and dendrogram

Group at any level of detail; see “distance” between groups in multivariate attribute space
Clustering: cluster statistics

Number of clusters: 6
Transformation: Standardize (Z)
Method: Ward's-linkage
Distance function: Euclidean

Cluster centers:

<table>
<thead>
<tr>
<th></th>
<th>POP8</th>
<th>PCTOWNHOME</th>
<th>PCTAGE65P</th>
<th>Z</th>
<th>PEXPOSURE</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>4.49078</td>
<td>-1.46223</td>
<td>-0.430876</td>
<td>-0.637778</td>
<td>-0.309463</td>
</tr>
<tr>
<td>C2</td>
<td>-0.205116</td>
<td>-0.867063</td>
<td>-0.709461</td>
<td>-0.2613</td>
<td>0.0199737</td>
</tr>
<tr>
<td>C3</td>
<td>0.345393</td>
<td>0.56543</td>
<td>0.0892871</td>
<td>-0.13715</td>
<td>-1.08593</td>
</tr>
<tr>
<td>C4</td>
<td>-0.173746</td>
<td>0.847444</td>
<td>-0.137512</td>
<td>-0.0483164</td>
<td>1.28624</td>
</tr>
<tr>
<td>C5</td>
<td>-1.75555</td>
<td>-0.2365</td>
<td>2.04196</td>
<td>3.16833</td>
<td>0.411726</td>
</tr>
<tr>
<td>C6</td>
<td>0.281756</td>
<td>-0.768694</td>
<td>2.39247</td>
<td>-0.012324</td>
<td>-0.00152647</td>
</tr>
</tbody>
</table>

The total sum of squares: 310

Within-cluster sum of squares:

<table>
<thead>
<tr>
<th></th>
<th>Within cluster S.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>0</td>
</tr>
<tr>
<td>C2</td>
<td>36.8377</td>
</tr>
<tr>
<td>C3</td>
<td>33.667</td>
</tr>
<tr>
<td>C4</td>
<td>27.6518</td>
</tr>
<tr>
<td>C5</td>
<td>9.74216</td>
</tr>
<tr>
<td>C6</td>
<td>4.68514</td>
</tr>
</tbody>
</table>

The total within-cluster sum of squares: 112.584
The between-cluster sum of squares: 197.416
The ratio of between to total sum of squares: 0.636827
Clustering: multivariate hierarchical: maps
Must operationalize the concept of "neighbour" and give each one a weight for tests of spatial correlation, and to use in spatial regression models.
Click on point in local Moran graph, highlights the polygon on the map; slope of line is global Moran’s I
Influential and clustered polygons for Moran’s I
Spatial regression model

Note “diagnostics for spatial dependence”, this is the next topic here
Our aim is to understand some spatial process – what explains the spatial distribution of a target variable?

We feel we’ve understood it if we can build a successful model.

A model can be used for prediction or policy decisions.

Special problems in spatial models:

- How much of the process is local (endogenous to an area)?
- How much of the process controlled by other spatially-distributed attributes (exogenous)?
- Is there a spillover effect by which exogenous factors in neighbouring areas affect the outcome?
- What is the proper representation of space? (distance, neighbours, weighting . . .)
Finding a correct model

- How do we know a model is correct, even if it fits well?
- Problem is **model mis-specification**
- Typical case: **apparent** spatial autocorrelation, caused by an **underlying factor** that is itself spatially-correlated
  - e.g., spatially-correlated productivity of forest blocks; related to spatially-correlated soil conditions.
  - Should analyze according to a **hypothesis** and **assumptions** based on **theory**.
- Method: compare models by their **likelihood** (see below)

Reference: Bivand et al. [6, §9]
A non-spatial analysis (in feature space) assumes independence of model residuals.

“Nearby” (in geographic space) areas may be similar because of some spatially-correlated underlying factor in geographic or feature space.

- e.g., house prices in adjacent city wards all affected by similar proximity to city centre (geographic space)
- e.g., crop yields in adjacent reporting districts all affected by the same climate and similar soils (feature space).

Feature-space attributes of “nearby” areas may affect the target attribute ("spillover effect")

- e.g., attractiveness of a ward for housing may depend not only on its own proportion of green space, but on the proportion in “nearby” neighbours
Question: **does the non-spatial** (feature-space) **model remove all the apparent spatial correlation?**

- If the **residuals** are spatially-correlated, the actual amount of information (roughly, “degrees of freedom”) is reduced.
  - Spatial autocorrelation usually **reduces** the amount of information supplied by each observations
  - This is because once we know surrounding areas we know something about a target area

- The feature-space model may have incorrect **coefficients**
Zero-mean models

- **Definition**: model where the expected deviance in each polygon from the global mean of a variable is zero.
- There may be spatial correlation but it is an attribute of the *spatial process* of the target variable only.
  - Example: diffusion of a pollutant from point sources through a homogeneous soil.
- Equivalent to *first-order stationarity* in random field theory (geostatistics).
- This is *not valid* if there is another spatially-distributed variable that, in feature space, (partially) determines the value of the target variable.
Combining feature and geographic space

1. Build a **feature-space** model (e.g., linear model)
2. Check **residuals** for spatial autocorrelation
   - for areal data, use global Moran’s I; for point data can also use variograms
3. If no autocorrelation, we are done, feature space explains everything
4. If present, build a model accounting for spatial autocorrelation. Various forms (see below):
   - Simultaneous Autoregressive Models (SAR)
     - spatial error SAR
     - spatial lag SAR
     - spatial Durbin SAR
     - spatial Durbin error SAR
   - Conditional Autoregressive Models (CAR)
5. Verify that the spatial model is more correct than the non-spatial model (e.g., Likelihood Ratio test)
Linear model with independent residuals

This is the **non-spatial** formulation; response is explained by predictors **in** attribute space only:

\[ Y = X\beta + \varepsilon \]  

- \( X \): design matrix of predictor values
- \( \varepsilon \): independent and identically-distributed \( \mathcal{N} \sim (0, 1) \) errors
- To estimate: \( \beta \), the linear regression coefficients
- Solve by minimization of \( \varepsilon^2 = (Y - X\beta)^2 \)
- **BLUE** is Ordinary Least Squares (OLS):

\[ \beta = \left( X^T X \right)^{-1} X^T Y \]
Example: Central NY leukemia

Leukemia incidence based on likely feature-space predictors:

**PEXPOSURE** exposure to TCE (trichloroethylene)
- toxic chemical linked to cancer

**PCTAGE65** % of residents > 65 years old
- cancer incidence may increase with age

**PCTOWNHOME** % of homes owned
- wealthier =? better health care? less likely to have worked in a chemical plant?

281 census tracts in 8 Central NY counties: Cayuga, Onondaga (includes Syracuse city), Madison, Chenango, Broome, Tioga, Tompkins, Cortland

---

Linear model results

Build an additive linear model using these predictors.

\[
\text{lm(formula = Z ~ PEXPOSURE + PCTAGE65P + PCTOWNHOME, data = NY8@data)}
\]

Coefficients:

|                      | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------------|----------|------------|---------|----------|
| (Intercept)          | -0.5173  | 0.1586     | -3.26   | 0.0012 **|
| PEXPOSURE            | 0.0488   | 0.0351     | 1.39    | 0.1648   |
| PCTAGE65P            | 3.9509   | 0.6055     | 6.53    | 3.2e-10 ***|
| PCTOWNHOME           | -0.5600  | 0.1703     | -3.29   | 0.0011 **|

Adjusted R-squared: 0.184

%> 65 years (+), % homeowners (-) significant, TCE (+) not
But does the model satisfy linear modelling assumptions?
Global Moran I for regression residuals
model: lm(formula = Z ~ PEXPOSURE + PCTAGE65P + PCTOWNHOME,
data = NY8@data)
weights: NY8listwB

Moran I statistic standard deviate = 2.64, p-value = 0.0042
alternative hypothesis: greater
sample estimates:
Observed Moran I  Expectation    Variance
  0.0830903   -0.0098913   0.0012423

The residuals are (positively) spatially-correlated among neighbours, i.e.,
similar residuals are clustered; so the OLS solution to the linear model is
not correct.
Simultaneous autoregressive models (SAR)

The solution is to use models that simultaneously solve for:

1. the regression coefficients, i.e., the effects of the predictors on the response;
2. the autoregressive error structure, i.e., the strength and nature of the spatial autocorrelation

Several forms of this, depending the cause of spatial autocorrelation:

- as a result of accounting for the predictors; a spatially-correlated residual effect: ‘induced spatial dependence’ ("spatial error model")
- as a result of a process within the target variable itself: ‘inherent spatial autocorrelation’ ("spatial lag model")
- both ("mixed model")
SAR model selection I

What process do we think is producing the spatial correlation?

- Spatially-correlated residual effect due to a spatially-correlated feature-space cause not included in the model: **SAR error model**
  - maybe we don’t suspect that it is a cause
  - maybe it has not been measured
  - leukemia example: occupation
  - ecology example: soil type (if not known or in model)
  - crime example: gun laws, sentencing guidelines

- A diffusion effect: **SAR lag model**
  - leukemia example: infection (e.g., feline leukemia, not known to occur in human leukemia)
  - ecology example: spread of an invasive species
  - social example: spread of a rumour by word-of-mouth

- A spillover effect: **SAR Durbin model**
  - this must also account for possible diffusion effects
SAR model selection II

- leukemia example: exposure to TCE in neighbouring areas, because residents in one area tend to shop or visit in neighbouring areas (??) and so the neighbours add to exposure
- ecology example: habitat quality in neighbouring forest patches affects bird population in a patch
- social example: amenities in neighbouring wards affecting desirability of living in a ward

- Compare models with the **Likelihood Ratio** or **Lagrange Multiplier** [4] tests
  - Likelihood Ratio: both models are fit with maximum likelihood, so the two likelihoods are known
  - “likelihood” = probability of the observed data being produced by the model with the given parameters
“spatial error”: the autoregressive process is found only in the error term, i.e., not accounted for by the linear model

- formula: \( Y = X\beta + \lambda Wu + \varepsilon \)
- \( W \) is a matrix representing the spatial structure (e.g., neighbour weights)
- \( u = (Y - X\beta) \) are the spatially-correlated residuals
- \( \lambda \) is the strength of autoregression of the errors
- \( \varepsilon \) is the independent error (not autoregressive)

The concept here is that there is some spatially-structured error, which cause we can not identify, but which we must account for to have a correct model.
“spatial lag”: the autoregressive process occurs only in the response variable

- formula:  \( Y = \rho W Y + X \beta + \varepsilon \)
- also can write \((I - \rho W) Y = X \beta + \varepsilon\)
- \(\rho\) is the strength of autoregression of the response variable
- Notice how the autocorrelation is applied to the response variable, not to the residuals, as in the SAR error model

The concept here is that the target variable in neighbouring areas affects the target in a given area.
“Durbin” or “mixed”: spatial autocorrelation affects both response (‘inherent spatial autocorrelation’) and explanatory (‘induced spatial dependence’) variables

- formula: $Y = \rho W Y + X \beta + WX \gamma + \epsilon$
- $\rho$ is the strength of autoregression of the response
- $\gamma$ is the strength of autoregression of the errors
SAR models in R

- package spdep
- SAR error model: functions errorsarlm
  - also spautlom; this can also compute Conditional Autoregressive (CAR) models
- SAR lag model: function lagsarlm with argument type="lag"
- SAR Durbin model: function lagsarlm with argument type="mixed"
Derivation of the SAR spatial error model

- Accounts for spatial autocorrelation of the residuals by a regression on the residuals from adjacent areas
- Residuals are partially the function of some (unobserved) ‘hot’ (or ‘cold’) spot of a spatially-distributed covariable
- Each area’s residual is modelled as a **linear function** of all the others (depending on neighbours and weights)

\[
e_i = \sum_{j=1}^{m} b_{ij} e_j + \varepsilon_i
\]  

(7)

- \(b_{ij}\) values: spatial dependence of \(e_i\) (residual in one area) on \(e_j\) (residual in neighbour area); set \(b_{ii} = 0\) (don’t self-regress)
SAR error model formulation

\[ Y = X\beta + B(Y - X\beta) + \varepsilon \]  \hspace{1cm} (8)

\[(I - B)(Y - X\beta) = \varepsilon \]  \hspace{1cm} (9)

To estimate: \(B\) (spatial dependence), \(\beta\) (regression)

This residual error \(\varepsilon\) is to be minimized; from the variance:

\[ \text{Var}[Y] = \sigma^2(I - B)^{-1}(I - B^T)^{-1} \]  \hspace{1cm} (10)

Reparameterize with explicit spatial autocorrelation parameter \(\lambda\) and spatial dependence matrix \(W\) (list of weights):

\[ \text{Var}[Y] = \sigma^2(I - \lambda W)^{-1}(I - \lambda W^T)^{-1} \]  \hspace{1cm} (11)

and solve for \(\lambda\) by maximum likelihood.
SAR error model example

```r
spautolm(formula = Z ~ PEXPOSURE + PCTAGE65P + PCTOWNHOME, 
          data = NY8, listw = NY8listwB)

            Estimate Std. Error   z value     Pr(>|z|)
(Intercept)   -0.618193  0.176784  -3.49690  0.0004707
PEXPOSURE      0.071014  0.042051   1.68877  0.0912635
PCTAGE65P      3.754200  0.624722   6.00942 1.862e-09
PCTOWNHOME    -0.419890  0.191329  -2.19464  0.0281930
```

Lambda: 0.040487 LR test value: 5.2438 p-value: 0.022026
Asymptotic standard error: 0.016214

- LR test value compares the models with and without spatial autocorrelation.
- p-value: probability that rejecting the null hypothesis (the two models are equally likely) would be a Type I error.
- If p-value is low $\rightarrow$ residuals of non-spatial model are autocorrelated.
These coefficients give the influence of **feature-space** predictors, **after** accounting for spatial correlation of residuals

- i.e., any spatial process is removed (not modelled)
- computing Moran’s I on the SAR residuals should confirm this

**λ** gives the **relative strength of the spatial process** vs. the feature-space process

- can visualize this with trend vs. stochastic residuals fits, see next page

The **form** of the spatial correlation is modelled by the **form of weights**

- depends on **neighbour list** and **weighting style**
- weighting style is set by modeller based on hypotheses of how the spatially-correlated error occurs; can compare several for robustness
Comparing regression coefficients: OLS vs. SAR/e

|                | Estimate | Pr(>|t|) |
|----------------|----------|---------|
| (Intercept)    | -0.5173  | 0.0012  |
| PEXPOSURE      | 0.0488   | 0.1648  |
| PCTAGE65P      | 3.9509   | 0.0000  |
| PCTOWNHOME     | -0.5600  | 0.0011  |

# SAR error model

|                | Estimate | Pr(>|t|) |
|----------------|----------|---------|
| (Intercept)    | -0.6182  | 0.0005  |
| PEXPOSURE      | 0.0710   | 0.0913  |
| PCTAGE65P      | 3.7542   | 0.0000  |
| PCTOWNHOME     | -0.4199  | 0.0282  |

**Substantial change in coefficients:** home ownership less important; exposure to TCE more important and now significant at $\alpha < 0.1$. 
Contributions to model fit

Trend

Stochastic

feature space

spatially-correlated residuals
SAR error model residuals

Leukemia incidence, linear model residuals

Leukemia incidence, SAR model residuals

points are TCE sites

SAR model has \textit{not} removed spatial correlation of residuals, just \textit{changed} it
The above analysis is with the **SAR error** model:

- ‘induced spatial dependence’
- process is **exogenous** to the response variable: local hotspots of some unmeasured factor
- leukemia example: could be local hotspots of carcinogens not included in the PEXPOSURE (exposure to TCE) term; could be local hotspots of an unknown or unaccounted for risk factor

An alternate formulation is the **SAR lag** model:

- ‘inherent spatial autocorrelation’
- process is **endogenous** to the response variable: diffusion or repulsion effects
- leukemia example: could be contagious (this is the case for feline leukemia – seems unlikely for humans)
### Spatial lag model

```r
lagsarlm(formula = Z ~ PEXPOSURE + PCTAGE65P + PCTOWNHOME,
         data = NY8, listw = NY8listwB, type = "lag")

| Term          | Estimate | Std. Error | z value | Pr(>|z|) |
|---------------|----------|------------|---------|----------|
| (Intercept)   | -0.514495| 0.156154   | -3.2948 | 0.000985 |
| PEXPOSURE     | 0.047627 | 0.034509   | 1.3801  | 0.167542 |
| PCTAGE65P     | 3.648198 | 0.599046   | 6.0900  | 1.129e-09|
| PCTOWNHOME    | -0.414601| 0.169554   | -2.4453 | 0.014475 |

Rho: 0.038893, LR test value: 6.9683, p-value: 0.0082967
Asymptotic standard error: 0.015053

This is more “likely” an explanation than the spatial error model: compare LR test values 6.97 (lag) vs. 5.24 (error).
Also, standard error 0.015 (SAR lag) is lower than 0.016 (SAR error)
Comparing regression coefficients: SAR_error vs. SAR_lag

# SAR_error

(Intercept)  -0.6182  0.0005
PEXPOSURE    0.0710  0.0913
PCTAGE65P    3.7542  0.0000
PCTOWNHOME   -0.4199  0.0282

# SAR_lag

   Estimate   Pr(>|z|)
(Intercept)  -0.5145  0.0010
PEXPOSURE    0.0476  0.1675
PCTAGE65P    3.6482  0.0000
PCTOWNHOME   -0.4146  0.0145

Less effect of all predictors after accounting for endogenous spatial autocorrelation in the leukemia incidence.
Both incorporate spatial correlation structure of the model residuals in a mixed model
  - GLS can include many other kinds of correlation structures

Spatial correlation in GLS depends on an authorized covariance function of separation between point-pairs
  - If polygons are reduced to their centroids, GLS can be used on area data

SAR uses weighted adjacency matrices to model the linear dependence of residuals on each other
  - so can work with polygons of any shape and size
To remember:

1. **areal data**: aggregated over (usually irregular) polygons

2. **apparent** spatial autocorrelation may depend on:
   - a spatial process of that variable;
   - spatially-structured covariable(s);
   - both.

3. Moran’s *I* measures strength of spatial autocorrelation

4. spatial structure depends on assumed process $\rightarrow$ weights matrix; based on distance, common boundary count or length . . .

5. paradigm: (1) formulate hypotheses; (2) build model to match hypotheses; (2) test model to see if there is evidence for/against hypotheses
Outline

1 Areal data
   - Definition and examples
   - Characteristics
   - The "ecological fallacy"
   - Neighbours

2 Spatial autocorrelation
   - Global Moran’s I
   - Autocorrelation of categorical variables
   - Local Moran’s I
   - Hot-spot analysis

3 GeoDa and LISA
   - Exploratory graphics
   - Clustering
   - Weights and neighbours
   - Spatial correlation
   - Spatial regression

4 Spatially-explicit linear models

5 References
Further reading

Theory: Anselin [3, 4], Openshaw [14]

Use in ecology, difference between SAR model types: Kissling and Carl [13]

Applications: see slide 5

In R: Bivand et al. [7, §9–10]

8-county leukemia study: Ahrens et al. [1]; original data Iwano [11]
Web pages

- Roger Bivand (Bergen): http://www.bias-project.org.uk/ASDARcourse/Unit 6 “Worked example: Spatial Weights, Autocorrelation”

- Lance Waller (Emory): analysis of public health from Waller and Gotway [19]: http://web1.sph.emory.edu/users/lwaller/ch7index.htm

- **GeoDa** Center for Spatial Data Science (Univ. Chicago, Luc Anselin); GeoDa computer program for exploratory ADSA http://spatial.uchicago.edu/geoda/ Text: Anselin and Rey [5]


