Analysis Methods for Area Data

Values are associated with a fixed set of areal units covering the study region.

We assume a value has been observed for all areas.

The areal units may take the form of a regular lattice or irregular units.

---

Analysis Methods for Area Data

Objectives

- Not prediction - there are typically no unobserved values, the attribute is exhaustively measured.
- Model spatial patterns in the values associated with fixed areas and determine possible explanations for such patterns.

Examples

Relationship of disease rates and socio-economic variables.
**Analysis Methods for Area Data**

\[ \{ Y(s), s \in R \} \]  
Random variable \( Y \) indexed by locations

\[ \{ Y(A_i), A_i \in R \} \]  
Random variable \( Y \) indexed by a fixed set of areal units

\[ A_1 \cup \cdots \cup A_n = R \]  
The set of areal units cover the study region \( R \)

Conceive of this sample as a sample from a super population – all realizations of the process over these areas that might ever occur.

The sample observations are one possible realization

---

**Visualizing Area Data**

Can use proportional symbols superimposed over the areal units. Symbols are proportionate to the attribute value of the area.

Can use choropleth maps – each area is shaded according the attribute value associated with the areal unit

- Attribute of interest is scaled to a set of discrete ranges or classes
- Each zone is shaded or colored according to its attribute value

---

**Analysis Methods for Area Data**

Explore these attribute values in the context of global trend or first order variation and second order variation – the spatial arrangement of the set of areas

- First order variation as variation in the of mean, \( \mu_i \) of \( Y_i \)
- Second order variation as variation in the \( COV(Y_i, Y_j) \)

---

**Visualizing Area Data**

Class intervals compress attribute ranges into a relatively small number of discrete values.

**Number of classes**

Should be a function of the range of data variability

Limited by human perception to \(~8\) classes

Rule of thumb: \# classes = \( 1 + 3.3 \log n \)

- 5 observations = 3 classes
- 20 observations = 6 classes
- 40 observations = 7 classes
- 200 observations = 8 classes
- 1500 observations = 11 classes
Visualizing Area Data

**Visualizing Area Data**

Same class intervals as applied to continuous data

- Equal Intervals - for fairly uniformly distributed data
- Trimmed equal intervals - handles a few outliers from a uniform distribution
- Percentiles - rank data and get $x$ evenly distributed classes of width $1/x$
- Quartile map - 4 classes, lowest quartile, second quartile, third, and highest
- Standard Deviates - divide data into units of standard deviation around the mean

**Visualizing Area Data**

Visual outcome depends heavily on class interval choice, color and shading

Use of discrete classes to assign colors can give false impressions of both uniformity (within units) and discontinuity (between units)

**Visualizing Area Data**

Problems with choropleth maps

Arbitrariness of the areal units

What was the origin of areal unit boundaries?

Designed for convenience or efficiency rather than reflection of the underlying spatial pattern – most enumeration units

Typically referred to as modifiable area unit problem - MAUP
**Modifiable Area Unit Problem**

Different zones will produce virtually any numbers from the same underlying distribution.

<table>
<thead>
<tr>
<th>Original data (individuals living in households)</th>
<th>mean 3.75 var 0.30</th>
<th>mean 3.75 var 0.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 4 6 1</td>
<td>3.75 3.75</td>
<td>3.75 3.75</td>
</tr>
<tr>
<td>3 6 3 5</td>
<td>3.75 3.75</td>
<td>3.75 3.75</td>
</tr>
<tr>
<td>1 5 4 2</td>
<td>3.75 3.75</td>
<td>3.75 3.75</td>
</tr>
<tr>
<td>5 4 5 4</td>
<td>3.75 3.75</td>
<td>3.75 3.75</td>
</tr>
</tbody>
</table>


---

**Visualizing Area Data**

**Problems with choropleth maps**

- Large areas tend to dominate the information content of the map.
- Area shading of attribute values means two visual variables apply to the data – color and area.
- Should not map absolute numbers with a choropleth map.
- With absolute counts there is no relation of the attribute variable to the absolute and/or relative size of the spatial units.
- Choropleth maps are not appropriate for counts unless the data are corrected for area, population or some other measure that factors out the size of the enumeration districts.

---

**Map of murder rates per 100,000 US and Canada**

- The problem of class limits
  - Use of continuous shading
- Area dependence problem
  - Correct for population or area
  - Transform areas to be proportionate to the attribute value
Corrects for population

There are two distinct and conflicting goals in the construction of cartograms:

- adjusting region sizes
- retaining region shapes

Cartograms
Cartograms are a non-continuous population cartogram based on the population values for the census zones of Leicestershire created using Dorling’s ‘circle growing’ algorithm. Both the circle areas and shades are proportionate to the zone populations.

Cartograms represent areas in relation to their population size. Patterns are displayed in relation to the number of people involved instead of the size of the area involved.

Comparison of traditional choropleth map with cartogram showing percent of male population of working age (1891)

The traditional map gives the impression that the majority of areas have approximately 56% (yellow, brown) of the male population of working age.

The cartogram gives a different overall view - the majority of areas have a male population of working age of over 60% (red and purple on the map).
Exploring Area Data

- Approaches for examining mean values over the areal units
- Techniques for exploring spatial dependence

What was essential information for exploration of these with point and continuous data?

Need measures of proximity for irregular areal units

Proximity Measures

- \( w_{ij} \) centroid of \( A_j \) is one of \( k \) nearest centroids to that of \( A_i \)
- \( w_{ij} \) centroid of \( A_j \) is within distance \( d \) of centroids of \( A_i \)
- \( w_{ij} \) \( A_j \) shares a boundary with \( A_i \)
- \( d_{ij} \) if intercentroid distance \( d_{ij} < \delta \)
- \( W_i \) otherwise

The spatial proximity matrix \( W \) with elements \( w_{ij} \) represents a measure of proximity of \( A_i \) to \( A_j \)

Adjacency

Not symmetric
**Proximity Measures**

Proximity measures can be specified as measures of different orders – spatial lags.

These can be represented as different proximity matrices for different lags.

For example $W_1$ represents first spatial lag, $W_2$ the second spatial lag, etc.

**Spatial Moving Averages**

One way to estimate the mean is by the average of the values in “neighboring” areas.

The proximity matrix $W$ represents the neighbors.

$$\tilde{\mu}_j = \frac{\sum_{j=1}^n W_{ij} y_j}{\sum_{j=1}^n W_{ij}}$$

**Median Polish**

If areal units are a regular grid can use median polish to estimate the trend. Median is more resistant to outliers in the data than the mean.

Each value is decomposed into:

$$y_{ij} = \mu + \mu_i + \mu_j + \epsilon_{ij}$$

where $\mu$ is a fixed overall effect.

$\mu_i$ and $\mu_j$ represent fixed row and column effects.

$\epsilon_{ij}$ is a random error term.

The overall mean is $\bar{\mu} = \mu_i + \mu_j$.
Median Polish algorithm

1. Take the median of each row and record the value to the side of the row – subtract the row median from each value in that row.

2. Compute the median of the row medians, and record the value as the overall effect. Subtract the overall effect from each of the row medians.

3. Take the median of each column and record the value beneath the column. Subtract the column median from each value in that particular column.

4. Compute the median of the column medians, and add the values to the current overall effect. Subtract this addition to the overall effect from each of the column medians.

5. Repeat steps 1-4 until no changes occur with the row or column medians.

Median Polish algorithm

Represents a global trend in the data.

Resulting cell values: $\hat{\mu}_{ij} = \hat{\mu} + \hat{\mu}_i + \hat{\mu}_j$

Also provides a method to remove trend by leaving a set of residuals to analyze.

Can produce banding effects since it decomposes trend according to the directions of the grid.

These directions may not correspond to the direction of trend in the data.

No control over the degree of smoothing.

Median Polish example

Source:
http://www.rem.sfu.ca/gis/Projects/Elu/Nzbirds/geostats/polish.htm
**Kernel Estimation**

Kernel estimation was employed to explore intensity variations for point pattern data and to describe changes in first order trend in continuous data.

It may be used in the area case, as well.

Kernel methods are really only applicable to distance, and therefore do not use the $W$ matrix.

$$\hat{\mu}_k(s) = \frac{1}{\sum_{i=1}^n k\left(\frac{s - s_i}{\tau}\right)} \sum_{i=1}^n k\left(\frac{s - s_i}{\tau}\right) y_i$$

use simple distance from the zone centroids to a set of points (small areas) $s$.

When observations are counts the previous approach is not appropriate - need an estimate of density

$$\hat{\lambda}_k(s) = \sum_{i=1}^n \frac{1}{\tau^2} k\left(\frac{s - s_i}{\tau}\right) y_i$$

Represents the total count per unit area

Count estimate derived from density estimate times area pf region or integrating density estimate over the region

One approach to cross areal interpolation

**Cross Areal Interpolation**

With areal data there is often a need to interpolate values from one set of areas to another.

Kernel estimation has been used to convert count data from set of irregular units to set of finer grid units.

$$\hat{\lambda}_k(s) = \sum_{i=1}^n \frac{1}{\tau^2} k\left(\frac{s - s_i}{\tau}\right) y_i$$

The data can then be re-aggregated to other sets of areas and used with data from an alternate set of areas.
Cross Areal Interpolation

Example

\[ + \]

\[ + \]

\[ + \]

\[ + \]