### Issues in Analyzing Spatial Data

**Chapter 2 Haining**

SIE 512 Lecture 6  
September 16, 2010

#### Data Sources: quality and quantity

- Spatial data previously much less available
- Earlier, spatial dimensions in data sets did not exist or were too inaccurate for much analysis
- New technologies driving collection of spatial data: satellite remote sensing, GPS, DOQ's
- Wireless sensor networks are rapidly appearing and generating data streams with high spatial and temporal resolution

*"Where data recording methods are automated the risk of undetected error increases."

#### Data Sources: quality and quantity

- More spatial libraries - data banks – potentially increase problems of data reliability.
- How carefully are the data checked?
- How well are they documented?
### Errors in spatial data

**Attribute error**
Census survey biases, remote sensing distortion, field sampling biases

**Positional error**
generalization, scale effects, straight line approximations-discretization effects, conversion: digitizer operator error, map errors, projection distortion, GIS operation errors

![Positional error diagram](image)

Length of shared boundary

Centroid to centroid distance

### Incompleteness
Sensor failure, missed readings, resolution constraints, data suppression

### Data collected over different, incompatible areal units

![Data collection diagram](image)

### Length of shared boundary

**School District**

**Census Units**

### Large data sets issues

**Particular problems - efficient exploratory processing and analysis**
15 variables, 10+locations, 3+depths, every 15 minutes, 7 years

![Large data sets issues](image)

### Modeling issues: hypothesis tests - more likely to reject the null hypothesis with very large samples.

In F tests, small values of $R^2$ and large n can lead to very large F hypothesis more frequently rejected with large sample size

\[
F = \frac{R^2/(1 - R^2)}{(n - k - 1/k)}[n - k - 1/k]
\]

Null hypothesis rejected when $F > F_{(k, n-k-1)}$

Modify significance levels as decreasing function of sample size
Small data set issues

- Analysis not meaningful
- Less likely to meet assumptions of normality
- Unable to discriminate between models
- More sensitive to data errors

Form and attributes - points and areas

Two main forms of spatial representation: points, areas

**Points**
- may be all occurrences of a phenomena within a region
- may be samples (discrete and continuous case)

**Areas:**
- areal framework for long term multipurpose analysis
- aim is a flexible system for aggregation of units,
  - be able to merge data from different sources in changing circumstances
- areal framework often designed for specific application - spatial scale, study objective, specific study attributes.

Areas

- area referenced data often formed from aggregation of primary units
  - can be regular grid or irregular partition
- regular tessellations more durable over time and comparative studies
- data from satellites, integrations over the surface gridded at various resolutions

**Area dependencies**
- attributes of primary units averaged over the areal unit
  - dependent on the area primary units
  - not dependent on primary units - population density
- Value assigned to an area ought to be representative of any sub area

Spatial Data Representation

Point
- (X_i, Y_i, Z_i)

Vector Representation
- (X_i, Y_i, Z_i)

Raster – Planar tessellation
- Cells partition plane in infinitely repeatable pattern

Spatial relations need to be made explicit
### Spatial Data Representation

Tessellation (raster) model most appropriate where data are continuous and grid resolution is small compared to scale of surface variation.

- Often a preferred structure for statistical analysis
- Easier to aggregate
- Easier to process

### Sources of spatial structure

Spatial structure in a measured attribute can arise from:

- Measurement error
- Continuity effects - spatial heterogeneity
- Operation of a spatial process where spatial relations explicitly effect the way the process behaves

Important - helps determine the type of spatial analysis - when to anticipate certain spatial patterns and what models will describe them best

---

### Measurement error

- Operator differences over a large area

<table>
<thead>
<tr>
<th>O₁</th>
<th>O₂</th>
<th>O₃</th>
<th>O₄</th>
<th>O₅</th>
<th>O₆</th>
</tr>
</thead>
</table>

- Instrument induced dependencies
  - spatially induced errors in remote sensors - greatest along scan lines
  - induced spatial correlation

### Continuity effects and spatial heterogeneity

Toblers Law – units closer together are more similar than those farther apart.

- Continuity of events in space generate form
- Form can be smoothly continuous or have relatively sharp boundaries

- Representation of spatial variation by regions
  - Regions as patches demarcated in terms of attribute levels
  - How regions or partitions are determined effects spatial structure
Regional Representations

Within a region individual primary units are assumed to be identical members of same population
Relations between regions unspecified
Assumed independent

Sharp delineation
Assumed internal uniformity

If areal units applied are smaller than the scale of surface variation then attribute values for these units are likely to be spatially correlated

Regional differentiation not a priori a property of space
May be more useful concept in the social sciences

Spatial Processes

Structure arises from operation of processes in which spatial relations enter explicitly in the way the process behaves
spatial process - one in which changes in state are due to spatial properties of an attribute
a process which generates spatial structure in an attribute
structure as a function of relationships in space and time

\[ y_{i,t+1} = f\left(y_{j,t}^{j \in N(i)}, y_{j,t-1}^{j \in N(i)}\right) \]

N(i) areas adjacent to i

Dispersal or Spread

Process in which population disperses and resulting spatial structure depends on dispersed

Week of December 1 1996
Week of December 9 1996
Types of analytical problems:

- problems in spatial data sampling
- problems in providing numerical summaries and characterizing spatial properties of data
- problems in the analysis of multivariate data sets

Problems in Spatial Sampling

Sampling: cost, effort, accuracy - how to account for spatial dependencies

Need to Determine:

- Number of sample sites required
- Their spatial location and spacing

Example – sampling for image classification and verification

Need to consider strength of correlation

Numerical Summaries

Summary as distribution of values: histograms, frequency plots, Box plots

Distribution characterized by mean (centrality), dispersion, skewness, and compared to standard models

Goal is to devise summary measures that characterize the spatial arrangement of values

Roughly in terms of clustered, random, alternating

Interested in describing patterns at different scales

Histogram says nothing about the spatial distribution
Correlation Measures

When two or more variables are measured we can compute correlation coefficients:

A numerical coefficient that describes the direction and character of the relationship between two variables

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

Positive relationship

No relationship

Negative relationship

Spearman's rank, Kendall's tau, Pearson's product moment are all aspatial measures of correlation.

May indicate no or no significant association

Need measures of spatial correlation- degree to which similar or dissimilar values of two variables are close to one another in space (and or time).

Spatial autocorrelation measures
Regression Models
Express functional relationship between response variable and explanatory variables

\[ y = X\beta + \epsilon \]
\[ y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i \quad \text{for } i = 1, \ldots, n \]

(k explanatory variables measured at sites i)

Can have dependencies in explanatory variable - spatially lagged explanatory

\[ y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \tau \sum_{j \in N(i)} x_{ij} + \epsilon_i \quad \text{for } i = 1, \ldots, n \]

N(i) denotes sites adjacent to i

Multiple Regression
Written in matrix notation

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

X is an \((n \times (k+1))\) matrix of observations on explanatory variables

Y is an \((n \times 1)\) vector of response variable

\( \epsilon \) is an \((n \times 1)\) vector of residuals

\( \beta \) is an \((k \times 1)\) vector of parameters to be estimated

Matrix representations

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

\[
\begin{bmatrix}
100 \\
95 \\
96 \\
92 \\
93 \\
94 \\
95 \\
\end{bmatrix}
\begin{bmatrix}
\beta_0 \\
\beta_1 \\
\beta_2 \\
\epsilon_1 \\
\epsilon_2 \\
\epsilon_3 \\
\epsilon_4 \\
\end{bmatrix}
= \begin{bmatrix}
1 \\
2 \\
3 \\
4 \\
5 \\
6 \\
7 \\
\end{bmatrix}
\]

\[
\epsilon^T = \begin{bmatrix}
\epsilon_1 \\
\epsilon_2 \\
\epsilon_3 \\
\epsilon_4 \\
\epsilon_5 \\
\epsilon_6 \\
\epsilon_7 \\
\end{bmatrix}
\]

\[
\epsilon \epsilon^T = \epsilon_1^2 + \epsilon_2^2 + \ldots + \epsilon_7^2
\]

\[
Y^TY = y_1^2 + y_2^2 + \cdots + y_7^2
\]

\[
1^TY = y_1 + y_2 + \cdots + y_7
\]
Regression Models

Can have dependencies in response variable - spatially lagged response

\[ y_j = \beta_0 + \beta_1 x_{j1} + \ldots + \beta_k x_{jk} + \rho \sum_{i \neq j} y_i + \epsilon_j \quad \text{for} \quad i, j \in \{1, 2, \ldots, n\} \]

When there are potentially numerous difficult to specify spatially correlated effects these effects can be addressed by covariance matrix \( V \)

\[ E(ee^T) = \sigma^2 I \]
\[ E(ee^T) = \sigma^2 V \]

\( V \) is a non-diagonal matrix that describes spatial dependence in the errors

\[ Y_{(s)} = X_{(s)}^T \beta + V_{(s)} \]

Can model spatial variation in the parameters \( \beta \) by allowing them to vary with space

For the constant or intercept coefficient:

\[ y_i = \alpha + \beta x_i + \gamma D_i + \epsilon_i \]

\( D_i \) is a dummy variable equal to 1 if site \( i \) is in area 1 and zero otherwise

When \( D=1 \), constant coefficient is \( (\alpha + \gamma) \)

When \( D=0 \) constant coefficient is \( \alpha \)

Model spatial variation in the slope coefficient

\[ y_i = \alpha + (\beta + \gamma D_i) x_i + \epsilon_i \]
\[ y_i = \alpha + \beta x_i + \gamma (D_i x_i) + \epsilon_i \]

\( D_i \) is a dummy variable equal to 1 if site \( i \) is in region 1 and zero otherwise

For region 1 slope coefficient is \( (\beta + \gamma) \)

For region 2 slope coefficient is \( \beta \)