

1 GIS and Spatial Statistics

1.1 Loose coupling

- Data files are exchanged between GIS software and statistical software packages.
- GIS is used to retrieve spatial and attribute information, to perform spatial data manipulations and to visualize the spatial information and statistics.
- Specialized routines in statistical packages are employed to perform the spatial analysis and simulation experiments.
- Examples: SPDEP in the R-package, connection between SPSS and Maptitude in today’s lab.
- Teaching advantage: Students have control over all steps of the analysis and can inspect the exchanged data structures. Additional operation can be applied (e.g., transform distribution of variable towards symmetry).
- Teaching disadvantage: Time consuming and potentially distractive from substantive interpretations. Students must be knowledgeable in the statistical and the GIS environments.

1.2 Tight coupling

- Automated information exchange between a GIS and a statistical program and automated data processing through macros and COM.
- Example: S-Plus with the spatial statistics package and ArcView.
- Teaching advantage: Students do not need to be fluent in two software environments. Efficient setup of examples and labs.
- Teaching disadvantage: Data exchange and processing very much a black box. Both software packages must still be installed in a teaching lab.

1.3 Integrated

- All spatial statistical analysis can be performed within a GIS through integrated menu commands.
- Examples: ArcGIS’s “Geo-statistical Analyst”, IDRISI or GeoDa.
- Teaching advantages: Seamless data analysis and subsequent data visualization. Just one software package.
• Teaching disadvantages: Black box. Bound to the functionality implemented in the spatial statistical module.

2 Systematic Development of Spatial Pattern Analysis

• **Point Pattern Analysis**: The locations of points are the random component. No other attributes of the points are relevant. A varying background intensity (first order effect) of the point distribution can be incorporated into the statistical model. Distances or densities are used to model point pattern processes. Point pattern analysis is implemented in the SPLANCE package available for R and S-plus and CrimeStat

• **Spatial Statistically Analysis**: The attribute of an areal object is a random variable. Measurements on other attributes of the areal objects are possible to incorporate for the expected level of the attribute under investigation (first order effects). Spatial relationships among the areal units are used to model spatial dependencies (stochastic spatial pattern). Spatial statistical methods are implemented for instance in Spdep for R, S-plus, MatLab, or GeoDa as well as in software tools for SPSS or SAS

• **Geo-statistics**: Interpolation among sampling points. The observed value at the sampling point can be fixed or random (with nugget effect), its location is given. The impact of an underlying spatial field can incorporated into the statistical model. Distance among the sampling locations are used to model spatial covariation. Implicitly, positive spatial autocorrelation is assumed. A comprehensive package is available in the Geostatistical Analyst for ArcGIS.

2.1 First and second order effects

• First order effects are associated with external factors that influence an observed pattern.
• These external influences can be independent variables in a regression context, or the uneven distribution of the population at risk underlying an observed disease pattern.

• Second order effects are based on mutual endogenous interrelationships among the spatial objects that either tie them together or repel them.

• These internal forces can be exchange mechanisms between spatial objects such as the sharing of information or of common values or some dominance-subordinate relationships.

• *Positive* autocorrelation is usually observed for spatial objects or their attributes that are attracting each other mutually

• *Negative* autocorrelation can be conceptualized as mutual repulsion or competition.

• It is difficult to distinguish between first and second order effects in empirical settings:

![Figure 4.1 Difficulty of distinguishing first- and second-order effects.](image)

The underlying intensity in the left pattern exhibits an external trend (access point from the beach), the middle pattern exhibits internal clustering (family groups on the beach), whereas the right pattern could be generated by either effect or their mixture.

• A spatial pattern can be decomposed into a first order component (systematic signal) a second order component (stochastic signal) and some purely random component (white noise):
• It is critical to convey the concepts of first and second order effects to the students.

2.2 Spatial Dependence/Independence

• What is the first law of geography? Everything is related to everything else, but near things are more related than distant things (Waldo Tobler, 1970)
  ⇒ Only the presence of spatial autocorrelation allows us to do local spatial interpolations and predictions.
• The **zero hypothesis** in spatial analysis usually denotes the absence of second order effects with respect to an underlying relationship structure among the spatial objects. This zero hypothesis leads
  [a] in point pattern analysis to complete spatial randomness
  [b] in spatial statistics to the stochastic independence among the observations and
  [c] in geo-statistics to a constant variogram over the inter-point distances.
• This zero hypothesis serves usually as reference level for spatial tests.

![Conceptual framework for the statistical approach to spatial analysis](image)
3 Sample Lecture 1: Spatial Autocorrelation In Choropleth maps

Assumptions: Students were already exposed to scatterplots, correlation coefficient and basic statistical test theory. Students were expected to have read a relevant chapter on spatial analysis in a standard geographical statistical textbook.

3.1 Recap: Pearson's Correlation Coefficient:

\[ r_{Y,X} = \frac{\sum_{i=1}^{n} (y_i - \bar{y}) \cdot (x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2} \cdot \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \]

Notes: Measures the linear relationship between two random variables \( Y \) and \( X \)

- Centers \( Y \) and \( X \) around their means, that is, \( x_i - \bar{x} < 0 \) implies that the observation \( x_i \) is below average and, vice versa, \( x_i - \bar{x} > 0 \) implies that the observation \( x_i \) is above average

- Later we will learn that the mean is actually a regression estimate and the variations around the mean are regression residual.

- This concept of variation around a reference value easily generalizes to regression residuals

- Mapping this variation leads to a bipolar map theme with zero as neutral value
3.2. Concept of Autocorrelation

What is autocorrelation and how does it differ from bivariate correlation?

• Bivariate correlation measures the relationship between two variables $Y$ and $X$

• Autocorrelation measures the internal correlation among one set of observations $Y$ with itself. Clearly it is not the trivial $r_{Y,Y} = 1$

• We need to operationalize the concept of internal correlation. This implies that our observations internally structured (i.e., arrangement) and that the observations are dependent with respect to this internal structure.

• Note: All statistical techniques you have encountered so far assume that the observations are stochastically independent (i.e., knowledge of particular observations does not allow us to make predictions of new observations)

**Example: Serial (Temporal) Autocorrelation**

• Observed average daily temperatures at South Point (Ohio)

![Graph of observed average daily temperatures at South Point (Ohio)]

• Question: Is there a relationship between the observed daily temperature and the temperature at the previous day?

• To address this question we need to generate a second variable that shifts the observations by one day.
Note that the time series \( \text{temp} \) has an internal order and that the observations are sorted according to this order.

The lagged variable \( \text{lagtemp} \) displays the temperature at the previous day. For the first day we do not have lagged information.

Now we have a variable and an internally related variable, for which and we can calculate a correlation coefficient: \( r_{\text{temp,lagtemp}} = 0.697 \)

Example: Spatial Autocorrelation

The concept of an internal spatial relationship (order) is more tricky than in the time-series situation, where the order comes naturally (the future depends on the past)
In spatial analysis the relationships among the observations (spatial objects) are *multidirectional, multilateral* and *not equally spaced*. There are also more observations at the *edge of the study area* and factors like *spatial extend of the regions, spatial aggregation* and *underlying densities* or *differential population counts* (i.e., spatial heterogeneity) become important.

- How do we represent spatial objects: simple distances between points or representative points of areas
- Geographical theories provide many *concepts of spatial relationships* within a single variable. For instance:
  - neighborhood relationships between areas (rook's or queen's specification in square tessellations, higher order neighborhood relationship of regions several neighbors apart)
  - traffic flows or migration patterns
  - spatial hierarchies (hub and spokes)
  - diffusion processes etc.
- We will focus here on a simple neighborhood relationship for the Columbus crime dataset

- The variable *lag.v.c* is the neighborhood sum of *v.crime* surrounding each census tract
• The variable \textit{nofneig} is the number of neighbors for each census tract.

• The variable \textit{nave.v.c} is the average centered crime rate around each census tract (i.e., $\frac{\text{lag.v.c}}{\text{nofneig}}$).

• Again we have now a variable and its internally related counterpart, the spatially lagged variable.

• Plotting \textit{nave.v.c} against \textit{v.crime} gives the so-called \textit{Moran's scatterplot}, which is frequently used in exploratory spatial data analysis:

![Moran's Scatterplot](image)

• If a census tract has an above average crime rate then it is most likely that it is surrounded by census tracts with above average crime rates. Vice versa for below average crime rates. This positive relationship indicates \textit{positive spatial autocorrelation} among the observation of one spatially distributed variable, $r_{v.crime,nave.v.c} = 0.691$.

• Note: there are also two potential \textit{outliers} that do not fit into the general spatial relationship. Potential outliers can be identified by \textit{local spatial statistical techniques}.

• In spatial statistics the correlation between a spatially lagged variable and its reference values is not used, because the distribution of this spatial autocorrelation correlation coefficient is difficult to evaluate in order to \textit{test for the lack of spatial independence}.

\textbf{3.3 The Spatial Link Matrix}

• The spatial matrix operationalizes the underlying structure of the potential spatial relationships among the observations.
• For potential distance relationships we have the distance matrix (known from road atlases, perhaps using spherical distances)

• For potential neighborhood relationships we must use a binary spatial connectivity matrix

Example: Encoding a spatial tessellation as a binary connectivity matrix

\[
\begin{bmatrix}
0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0
\end{bmatrix}
\]

Spatial tessellation of 9 hexagonal cells with underlying connectivity structure

Binary 9 × 9 spatial connectivity matrix \( B \)

• An element \( b_{ij} = 1 \) denotes that the tiles \( i \) and \( j \) are adjacent and an element \( b_{ij} = 0 \) signifies that the tiles \( i \) and \( j \) are not common neighbors

• The spatial connectivity matrix \( B \) is symmetric

• A tile is not connected to itself. Thus all diagonal elements are zero

• For study areas with a large number \( n \) of individual regions the generation of the connectivity matrix \( B \) (or distance matrix) must be left to a GIS program. The connectivity matrix has \( n \times n \) elements

• See for instance the free GIS/spatial statistics program \textit{GeoDa} at \url{http://sal.agecon.uiuc.edu/csiss/index.html#geoda}

• Problems occur if we have island and holes in our study area. Usually machine generated connectivity matrices must be polished manually.

• Most of the elements are zero. There are efficient storage modes for sparse matrices. For empirical map patterns an area in the interior has on average 6 neighbors.
3.4 Join Count Spatial Autocorrelation Test

- The join count test is defined for **dichotomous** variables that follow a Bernoulli distribution:

\[
\Pr(y_i) = \begin{cases} 
\pi & \text{for } y_i = 1 \text{ or black or present} \\
1 - \pi & \text{for } y_i = 0 \text{ or white or absent}
\end{cases}
\]

- Higher order measurement scales can be **downgraded** to the nominal dichotomous scale by implementing a **cut-off** point (for instance, is below and above average).

- The join count statistic counts the number of **pairwise links** between **adjacent** regions for which both regions are Black-Black, Black-White, or White-White.

- These counts are tested whether they follow a **random spatial pattern**, or a more **spatially disperse** or a more **spatially clustered** pattern (see McGrew & Monroe pp 228-232 for the test statistic and test procedure).

- If we assume the total numbers of Black and White regions are **externally given** a **nonfree sampling** test procedure follows (sampling without replacement).

- If we assume the total number of Black and White regions is a **random** variable that follows the Bernoulli distribution the **free sampling** test procedure follows (sampling with replacement).

**Example: Different Black-White Pattern (Cholera death London 1832)**

• A negative $z(BW) = [BW - E(BW)]/\sqrt{Var(BW)} \sim N(0,1)$ indicates spatial clustering ([a] disease spreads from few sources, [b] large areas favoring similar disease levels, or [c] gradient in disease intensity across the map)

• A positive $z(BW)$ indicates a disperse pattern ([a] heterogenous study area, [b] pockets of immune and susceptible populations, or [c] small clusters of similar diseases)

3.5 Global Moran’s I Spatial Autocorrelation Test

• Moran’s I: $I = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} b_{ij} \cdot (y_{i} - \overline{y}) \cdot (y_{j} - \overline{y})}{\sum_{i=1}^{n} \sum_{j=1}^{n} b_{ij} \cdot (y_{i} - \overline{y})^2/n}$

• The $z$-transformed $z(I) = \frac{I - E(I)}{\sqrt{Var(I)}}$ is again approximately normal distributed for large $n$, with $z(I) < -1.96 \leftrightarrow z(I) > 1.96$ at the 5% error probability level

Examples of extreme map patterns for different tessellations:
• Another test statistic for global spatial autocorrelation is called Geary's $c = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} b_{ij} \cdot (y_i - y_j)^2}{2 \cdot \sum_{i=1}^{n} \sum_{j=1}^{n} b_{ij} \cdot (y_i - \bar{y})^2 / (n - 1)}$.

• It ranges approximately from $0 \leftrightarrow 1 \leftrightarrow 2$.

3.6 Local Moran's $I_i$ Spatial Autocorrelation Test

• Local Moran's $I_i$ measures the clustering or dispersal tendency around a reference cell $i$. We get $n$ local Moran's $I_i$ measures for a map.

• Local Moran's $I_i$ is defined by components of global Moran's $I$

\[
I_i = \frac{n \cdot \sum_{i=1}^{n} \sum_{j=1}^{n} b_{ij} \cdot (y_i - \bar{y}) \cdot \sum_{j=1}^{n} b_{ij} (y_j - \bar{y})}{\sum_{i=1}^{n} (y_i - \bar{y})^2 / n}
\]

• The average $I = \frac{1}{n} \cdot \sum_{i=1}^{n} I_i$ of the local Moran's $I_i$s gives the global Moran's $I$. Consequently, the local Moran's $I_i$ can be interpreted as variation around the global autocorrelation level.

Example: Forms of Local Spatial Autocorrelation

<table>
<thead>
<tr>
<th>Positive</th>
<th>local spatial Autocorrelation (spatial cluster)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><img src="image" alt="Positive Local Spatial Autocorrelation" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Negative</th>
<th>local spatial Autocorrelation (spatial outlier, hot spot)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><img src="image" alt="Negative Local Spatial Autocorrelation" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Local Spatial Independence</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Local Spatial Independence" /></td>
</tr>
</tbody>
</table>

• Another popular local test statistic is Getis and Ord's local $G_i$ measure. It tests for clusters of above or below average values.
Example: Revisit Columbus Crime Data

- Global Moran's $I$ for the variable Crime is $I = 0.5206$ and its $z$-value is $z(I = 0.5206) = 6.2557$.

- Thus the crime in Columbus has the tendency to cluster spatially (first map).

- The probability $\Pr(I_i < I_{i,\text{observed}})$ of local Moran's $I_i$ are given in the second map:

- Census tract 4 (OSU campus in North-West corner) is clearly a spatial outlier.

- Discussion of statistical problems:
  - The distribution of a single local Moran's $I_i$ is difficult to evaluate and depends on exogenous variable used to calculate the regression residuals or deviations around the mean. Thus simulations or randomizations are frequently used to approximate the distribution.
  - The local Moran's $I_i$ statistics are influenced by edge effects
  - The local Moran's $I_i$ statistics are correlated among each other
  - Testing a set of local Moran's $I_i$ statistics leads to the multiple testing problem and thus $\alpha$-error is inaccurate (for 100 tests we expect 5 tests to reject the zero hypothesis of local spatial independence).
4 Sample Lecture 2: Point pattern methods (without first order effects)
Assumptions: Students are familiar with basic inferential statistics and the Poisson distribution.


- There are two basic approaches point toward point pattern analysis:
  [a] use the density of points (see Quadrat Analysis below)
  [b] use the distances separating the points

4.1 Quadrat Analysis (density based):
- Point pattern: Agglomerative advantages (clustering) versus spatial competition/repulsion (regular pattern maximizes the distance among points)
- Assumption: Constant first order effect, that is, [a] the underlying areas have a homogeneous distribution or [b] the underlying intensity of the hypothetical point process is constant
- Problem of location definition: Do we record where a crime happens or where the victims or offenders live? Do we record the home or the work address of a person with a disease?
- Problems with reference grid generation: [a] boundary cells in irregular shaped study areas, [b] cell size and [c] shifts in the overlay grid pattern (spatial scale and aggregation problem)
- Empirical distribution: Number of points within each grid cell (Use overlay function to count the number of points within each grid).

4.2 Demonstration
- A 9 by 10 grid overlay over the study area (left map) and the count of the number of points in each grid cell (middle map). In comparison the right map pattern displays a simulated pattern under complete spatial randomness. It is more disperse than the observed pattern:
4.3 Reference distribution:

- Simulation of a random point pattern under complete spatial randomness:
  Select the latitude and the longitude of a point independently from a random uniform distribution (Note: the extend should match the bounding box of the study area)

- Under complete spatial randomness one can show that the number of points within a grid should theoretically follow a Poisson distribution:
  \[ \Pr(X = x) = \frac{e^{-\lambda} \cdot \lambda^x}{x!} \]
  with the intensity \( \lambda = \frac{\# \text{ of points in study area}}{\# \text{ of grid cells}} \)
  that we expect for each grid cell.

- Test with SPSS for the Poisson distribution by the Kolmogorov-Smirnov test

  **One-Sample Kolmogorov-Smirnov Test**

<table>
<thead>
<tr>
<th></th>
<th>( n )</th>
<th>( \text{deaths} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td></td>
<td>90</td>
</tr>
<tr>
<td>Poisson Parameter(^a,b)</td>
<td>Mean</td>
<td>6.41</td>
</tr>
<tr>
<td>Most Extreme</td>
<td>Absolute</td>
<td>.532</td>
</tr>
<tr>
<td>Differences</td>
<td>Positive</td>
<td>.532</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>-.175</td>
</tr>
<tr>
<td>Kolmogorov-Smirnov Z</td>
<td></td>
<td>5.045</td>
</tr>
<tr>
<td>Asymp. Sig. (2-tailed)</td>
<td></td>
<td>.000</td>
</tr>
</tbody>
</table>

  \(^a\) Test distribution is Poisson.

  \(^b\) Calculated from data.

- Note: the expectation and variance of the Poisson distribution are both equal to the intensity \( \lambda \).
This property leads to the variance-mean ratio test statistic \( VMR = \frac{s^2}{\bar{x}} \)
where the mean is \( \bar{x} = \hat{\lambda} \) and the observed variance is 
\[
\sum_{i=1}^{m} \frac{(x_i - \bar{x})^2}{m - 1}
\]
with \( x_i \) the number of points in the grid cell \( i \) and \( m \) the number of grid cells.

Interpretation: [a] if the points are more clustered, \( s^2 \) increases and [b] if the points are regular spaced \( s^2 \) decreases.

Thus \( VMR > 1 \) indicates spatial clustering and \( VMR < 1 \) points to regular spaced observations.

4.4 Approximations:

The statistics \((m - 1) \cdot VMR\) follows approximately a \( \chi^2 \)-distribution with \( df = m - 1 \).

The \( \chi^2 \)-distribution has a mean of \( df \) and a variance of \( 2 \cdot df \).

Thus for \( df > 30 \) the normal approximation can be used:

\[
z(VMR) = \frac{(m - 1) \cdot VMR - (m - 1)}{\sqrt{2 \cdot (m - 1)}} \sim N(0,1)
\]
5 Discussion:

- How can we improve the sample lectures?
  - What prerequisites are needed? Statistical and GIS foundation
  - How can we incorporate a textbook into the lecture?
  - What level of math is appropriate?
  - How can we time the lecture?
  - Shall we include brief demonstrations during the lecture?
  - What needs to be considered if spatial statistics is introduced in a GIS course or a general course introduction to statistics for spatial analysts?
  - How could we design a tutorial/lab reinforcing the concepts of the lecture?
  - Are there any classroom exercises that the students can conduct? Example: location decision of two ice-cream vendors.
  - What software environment is appropriate?
  - Many more questions …

6 Suggested exercise:

Assumption: SPACE participants are interested in GeoDa

- Comment on datasets and associated reading material
- Columbus crime analysis:
  - Show the dataset in SPSS; discuss spatial link matrix and calculate local Moran's $I_s$.
  - Show problem of generating a link matrix in GeoDa with sloppy digitized ArcGIS shape files
  - Import shape file into GeoDa and select NEIG as key
  - Generate choropleth map of crime rate and interpret the pattern
  - Attach local Moran's dBase file to the table.
  - Map variances of local Moran's $I_s$ to show the presence of edge effects
  - In GeoDa generate spatial link matrix
o In GeoDa calculate local Moran's $I_s$ for crime and interpret the observed pattern.

o Calculate residuals from a regression model crime $\sim$ discbd. Interpretation: Inclusion of a first order effect (distance decay of crime from CBD) eliminates the spatial dependencies in the regression residuals.