Spatial Data Mining

Regression and Classification Techniques
Discrete class labels (left) vs. continues quantities (right) measured at locations (2D for geographic applications)

Build a model for predicting the measured quantity at any location

Additional (to spatial) attributes may exist
Geostatistics

- Analysis and inference of continuously-distributed variables
  - Analysis: Describing the spatial variability of the phenomenon under study
  - Inference: Estimating the unknown values
- Questions on measurements:
- How are they distributed? How are they related to each other? How can I infer a distribution from one sample?
Spatial continuity and stationarity

Why prediction is possible?

- **Continuity**: Spatial close measurements are more similar than distant ones

What does it mean?

- Model the underlying phenomenon with the model $f(x,w)$, $x$ the location vector and $w$ the measurement
- If not just noise, then continuity creates “smoothness” of $w$ values that can be modeled by $f(x,w)$

Can all locations be modeled by a single $f(x,w)$?

- **Stationarity**: Measurements generated by a single distribution at all locations
Spatial Autocorrelation

- Continuity produces autocorrelation: correlation of a variable with itself through space
  - First law of geography: “everything is related to everything else, but near things are more related than distant things” – Waldo Tobler

- 3 possible cases:
  - If nearby or neighboring areas are more alike, this is positive spatial autocorrelation
  - Negative autocorrelation describes patterns in which neighboring areas are unlike
  - Random patterns exhibit no spatial autocorrelation
Why to bother about spatial autocorrelation?

- Most statistics/data mining methods are based on the assumption that the values of observations in each sample are independent of one another.
- Positive spatial autocorrelation may violate this, if the samples were taken from nearby areas.
  - Spatial Autocorrelation is a kind of redundancy: the measurement at a location constrains, or makes more probable, the measurement in a neighboring location.
  - Models will be biased, since measurements tend to be concentrated and there are actually fewer number of independent observations than are being assumed.
Measures of autocorrelation

Objectives:
- Measure the strength of spatial autocorrelation
- Test the assumption of independence or randomness

Measures
- Moran’s I
- Variograms
- other (Geary’s C, Ripley’s K)
Moran’s I: A measure of spatial autocorrelation

- Compares the value of the variable at any one location with the value at all other locations

\[
I = \frac{N \sum_{i} \sum_{j} W_{i,j} (X_i - \bar{X})(X_j - \bar{X})}{(\sum_{i} \sum_{j} W_{i,j}) \sum_{i} (X_i - \bar{X})^2}
\]

- Similar to correlation coefficient, it varies between \(-1.0\) and \(+1.0\)
  - When autocorrelation is high, the coefficient is high
  - A high \(I\) value indicates positive autocorrelation
Symbols and Contiguity matrix

- $N$ is the number of cases
- $X_i$ is the variable value at location $i$
- $X_j$ is the variable value at location $j$
- $\bar{X}$ is the mean of the variable
- $W_{ij}$ is a weight applied to the comparison between location $i$ and location $j$

\[ I = \frac{N \sum_i \sum_j W_{i,j} (X_i - \bar{X})(X_j - \bar{X})}{(\sum_i \sum_j W_{i,j}) \sum_i (X_i - \bar{X})^2} \]

- $W_{ij}$ is a contiguity matrix
  - If location $j$ is adjacent to zone $i$, the interaction receives a weight of 1
  - Another option is to make $W_{ij}$ a distance-based weight which is the inverse distance between locations $i$ and $j$ ($1/d_{ij}$)
Example: Per Capita Income in Monroe County

Actual values: Moran’s I: 0.66

Random values: Moran’s I: 0.01
Local Moran’s I

Following Anselin’s (1995) definition, a local Moran’s $I_i$ may be defined as:

$$I_i = \frac{Z_i}{s^2} \sum_j w_{ij} Z_j, \ i \neq j$$

$Z$s are the deviations from the mean of $y$s

![Table of numbers](image)

$$I_{75} = \frac{75 - 55.82}{675.32} [71 + 85 + 61 + 63 - 4 \times 55.82] = 1.61$$
Global vs. Local Moran’s I: example

- Spatial pattern detection in China’s provincial development
- The variable used: per capita GDP
- Dynamic patterns – global Moran’s I
- Specific local spatial process – local Moran’s I and the Moran’s scatterplot
China: per capita GDP in 1978
China: per capita GDP in 2000

Eastern Region
Central Region
Western Region
Eastern Region

Yuan
869 - 1913
1914 - 3162
3163 - 4532
4533 - 8411
8412 - 15593

0 250 500 1,000 Miles
0 500 1,000 2,000 Kilometers
Global vs. Local Moran’s I: example

There is a clustering trend in China’s provincial level development (represented by per capita GDP).

But the global Moran’s I can’t tell on which side does the clustering trend take place.
Local Moran’s I in 1978

- Eastern Region
- Central Region
- Western Region
- Eastern Region

Local Moran’s I

- < - 0.3
- - 0.3 - 0
- 0 - 0.3
- 0.3 - 1.0
- > 1.0
Local Moran’s $I$ in 2000

- Eastern Region
- Central Region
- Western Region

Local Moran’s $I$
- $-0.3 - 0$
- $0 - 0.3$
- $0.3 - 1.0$
- $>1.0$
More details to the Chine GDP example

- First, China’s coast-interior divide persisted
  - Interior provinces exhibit great geographical similarity in economic development and spatial contributions to the global Moran’s $I$
- Second, the municipalities (Beijing, Tianjin, Shanghai) always contribute the most
  - Shanghai’s position is worth noting, its development changed the spatial pattern the most
- Third, Guangdong’s contribution to the global index corresponds with its changing spatial behavior depicted in the Moran scatterplot
- Fourth, while most of the interior provinces have similar patterns, coastal provinces vary greatly
- Fifth, Shandong fell into the low-low quadrant, and contributed very little to the global index
- Sixth, Guizhou and Yunnan, two provinces in southwest China, contributed relatively highly to the global index in 2000
  - The poorest ones tend to form a poor cluster
Variograms

- Analyse the observed variation in data values by distance bands using a spatial autocorrelation-like measure, $\gamma$:
  - Semivariance measure is most often used:

$$
\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{d_{ij}=h+\Delta/2} (z_i - z_j)^2 - \sum_{d_{ij}=h-\Delta/2} (z_i - z_j)^2
$$

- Bands have width $\Delta$. $N(h)$ is the number of pairs in the band with mid-point distance $h$

- After building an experimental variogram, we need to fit a theoretical function in order to model the spatial variation
Variograms

- Smallest observed separation
- Fitted curve
- Average semivariance for band 4
- \( C_1 = C_0 + C \) (structural variance)
- Range, \( A_0 \)
- Lag (distance band)

Exponential model: \( C_0 = 21400.0000, C_0 + C = 132900.0000, A_0 = 352.00, \gamma = 0.994 \).
## Variograms

<table>
<thead>
<tr>
<th>Model</th>
<th>Formula (Theoretical Fit)</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nugget effect</td>
<td>$\gamma(0) = C_0$</td>
<td>Simple constant. May be added to all models. Models with a nugget will not be exact</td>
</tr>
<tr>
<td>Linear</td>
<td>$\gamma(h) = C_1(h)$</td>
<td>No sill. Often used in combination with other functions. May be used as a ramp, with a constant sill value set at a range, a</td>
</tr>
<tr>
<td>Exponential</td>
<td>$\gamma(h) = C_1\left(1 - e^{-kh}\right)$</td>
<td>$k$ is a constant, often $k=1$ or $k=3$. Useful when there is a larger nugget and slow rise to the sill</td>
</tr>
<tr>
<td>Spherical</td>
<td>$\gamma(h) = C_1\left(\frac{3h}{2} - \frac{1}{2} h^3\right), h &lt; 1$ $\gamma(h) = C_1, h \geq 1$</td>
<td>Useful when the nugget effect is important but small. Given as the default model in some packages.</td>
</tr>
</tbody>
</table>
Approaches to spatial prediction

Value of the variable is predicted from “nearby” samples
- Example: concentrations of soil constituents (e.g. salts, pollutants)
- Example: vegetation density

Each interpolator has its own assumptions:
- Nearest neighbor and variations:
  - Average within a radius
  - Average of \( n \) nearest neighbors
  - Distance-weighted average within a radius
  - Distance-weighted average of \( n \) nearest neighbours

“Optimal” weighting -> Kriging
Nearest Neighbor Methods

- **k-NN Classification**: assign the class label of the majority of the k-NN

- **k-NN Regression**: assign the mean value of the k-NN

\[
\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i
\]

A common weighting scheme is to give each neighbor a weight of \(1/d\), where \(d\) is the distance to the neighbor.

1-NN: Voronoi Diagram
**Nearest Neighbor Methods**

**Pros:**
- Simple, no training (lazy)
- Benchmark:
  - $E_{1-NN} \leq 2 E_B$
- Often as good as more sophisticated methods
- Per-se considerations of autocorrelation

**Cons:**
- Slow classification (lazy)
- Prone to noise
- High-variance
- Need to determine $k$
  - Cross validation
- Need to determine weights (for variations)
Nearest Neighbor Methods

- When no spatial autocorrelation (random data):

\[
Z = f(X) + \varepsilon(X)
\]

\[
Z \approx \varepsilon(X)
\]

- CV (LOO) error is maximized for 1-NN:

\[
E = \frac{1}{N} \sum_{j=1}^{N} \left(Z_j - Z_{j,1NN}\right)^2 \approx \frac{1}{N} \sum_{j=1}^{N} \left(\varepsilon_j - \varepsilon_{j,1NN}\right)^2 \propto 2\text{Var}(\varepsilon) \approx 2\text{Var}(Z)
\]
Nearest Neighbor Methods

Random data
- No minimum occurs

Spatial autocorrelation
- Minimum occurs
**Nearest Neighbor Methods**

**Bias-Variance decomposition:**

\[
Err(x_0) = \sigma^2 + \left( f(x_0) - \frac{1}{k} \sum_{n=1}^{k} f(x_n) \right)^2 + \frac{\sigma^2}{k}
\]

- Original
- k=3
- k=1 (hi var)
- k=30 (hi bias)
“Optimal Weighting”: Kriging

Characteristics of “optimality”:

- Prediction is made as a **linear** combination of known data values (a **weighted average**)
  - Points closer to the point to be predicted have larger weight
- Prediction is **unbiased** and **exact at known points**
- Error estimate is based only on the sample configuration, not the data values
- **Prediction error should be as small as possible**

Why “optimal” and not optimal?

“**optimal**” with respect to the chosen model!
Overview of Kriging

1. **Sample**, preferably at different resolutions
2. **Calculate** the experimental variogram
3. **Model** the variogram with one or more authorized functions
4. **Apply** the kriging system, with the variogram model of spatial dependence, at each point to be predicted
   - Predictions are often at each point on a regular grid (e.g. a raster map)
5. Calculate the error of each prediction; this is based only on the sample point locations, *not* their data values.
**Ordinary Kriging (OK)**

- In OK, we model the value of variable \( z \) at location \( s_i \) as the sum of a **regional mean** \( m \) and a **spatially-correlated random component** \( e(s_i) \):

  \[
  Z(s_i) = m + e(s_i)
  \]

- The regional mean \( m \) is estimated from the sample, but not as the simple average, because there is spatial dependence.

  - It is **implicit** in the OK system.
Ordinary Kriging: Solution

\[ C_{ij} = C(0) - \gamma(h) = C_0 + C_1 - \gamma(h) \]

- Covariance matrix elements

- Substituting the values we find the weights

- Kriging estimator:
  \[ Z_{x_0}^* = \sum_{i=1}^{n} \lambda_i Z(x_i) \]

- Variance
  \[ \sigma_{k0}^2 = (C_0 + C_1 - \lambda^T k) \]
**Kriging usage**

- Supported by many GIS

- But be aware of polemics between classic statistics vs. geostatistics
  - Spatial dependence may be assumed or be verified?

- Kriging in scandal: Spatial dependence between borehole grades or blasthole grades was assumed at Bre-X's Busang property

- More details: