Introduction to Spatial Data and Models

Sudipto Banerjee¹ and Andrew O. Finley²

¹ Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, U.S.A.
² Department of Forestry & Department of Geography, Michigan State University, Lansing Michigan, U.S.A.

July 19, 2009

Researchers in diverse areas such as climatology, ecology, environmental health, and real estate marketing are increasingly faced with the task of analyzing data that are:
- highly multivariate, with many important predictors and response variables,
- geographically referenced, and often presented as maps, and
- temporally correlated, as in longitudinal or other time series structures.

This motivates hierarchical modeling and data analysis for complex spatial (and spatiotemporal) data sets.

Type of spatial data

- **point-referenced data**, where $Y(s)$ is a random vector at a location $s \in \mathbb{R}^r$, where $s$ varies continuously over $D$, a fixed subset of $\mathbb{R}^r$ that contains an $r$-dimensional rectangle of positive volume;

- **areal data**, where $D$ is again a fixed subset (of regular or irregular shape), but now partitioned into a finite number of areal units with well-defined boundaries;

- **point pattern data**, where now $D$ is itself random; its index set gives the locations of random events that are the spatial point pattern. $Y(s)$ itself can simply equal 1 for all $s \in D$ (indicating occurrence of the event), or possibly give some additional covariate information (producing a marked point pattern process).

---

Exploration of spatial data

- First step in analyzing data
  - **First Law of Geography**: Mean + Error
  - Mean: first-order behavior
  - Error: second-order behavior (covariance function)

- EDA tools examine both first and second order behavior
- Preliminary displays: Simple locations to surface displays

---

Scallops Sites
Spatial surface observed at finite set of locations
\[ \mathcal{X} = \{ s_1, s_2, \ldots, s_n \} \]
Tessellate the spatial domain (usually with data locations as vertices)
Fit an interpolating polynomial:
\[ f(s) = \sum_i w_i(\mathcal{X} ; s) f(s_i) \]
“Interpolate” by reading off \( f(s_0) \).
Issues:
- Sensitivity to tessellations
- Choices of multivariate interpolators
- Numerical error analysis

Introduction to spatial data and models
Scallops data: image and contour plots
Drop-line scatter plot
Surface plot
Image contour plot
Locations form patterns
Introduction to spatial data and models

Stationary Gaussian processes

Variograms

Suppose we assume $E[Y(s) - Y(s)] = 0$ and define

$$E[Y(s + h) - Y(s)]^2 = Var(Y(s + h) - Y(s)) = 2\gamma(h).$$

This is sensible if the left hand side depends only upon $h$. Then we say the process is intrinsically stationary.

$\gamma(h)$ is called the semivariogram and $2\gamma(h)$ is called the variogram.

Note that intrinsic stationarity defines only the first and second moments of the differences $Y(s + h) - Y(s)$. It says nothing about the joint distribution of a collection of variables $Y(s_1), \ldots, Y(s_n)$, and thus provides no likelihood.

Interesting plot arrangements

Suppose our spatial process has a mean, $\mu(s) = E(Y(s))$, and that the variance of $Y(s)$ exists for all $s \in D$.

- Strong stationarity: If for any given set of sites, and any displacement $h$, the distribution of $(Y(s_1), \ldots, Y(s_n))$ is the same as $(Y(s_1 + h), \ldots, Y(s_n + h))$.
- Weak stationarity: Constant mean $\mu(s) = \mu$, and $Cov(Y(s), Y(s + h)) = C(h)$: the covariance depends only upon the displacement (or separation) vector.
- Strong stationarity implies weak stationarity
- The process is Gaussian if $Y = (Y(s_1), \ldots, Y(s_n))$ has a multivariate normal distribution.
- For Gaussian processes, strong and weak stationarity are equivalent.

Intrinsic Stationarity and Ergodicity

- Relationship between $\gamma(h)$ and $C(h)$:

$$2\gamma(h) = Var(Y(s + h)) + Var(Y(s)) - 2Cov(Y(s + h), Y(s))$$

$$= C(0) + C(0) - 2C(h)$$

$$= 2[C(0) - C(h)].$$

- Easy to recover $\gamma$ from $C$. The converse needs the additional assumption of ergodicity: $\lim_{|u| \to \infty} C(u) = 0$.
- So $\lim_{|u| \to \infty} \gamma(u) = C(0)$, and we can recover $C$ from $\gamma$ as long as this limit exists.

$$C(h) = \lim_{|u| \to \infty} \gamma(u) - \gamma(h).$$
Introduction to spatial data and models

Isotropy

Examples: Spherical Variogram

\[ \gamma(t) = \begin{cases} 
\tau^2 + \sigma^2 & \text{if } t \geq 1/\phi \\
\tau^2 + \sigma^2 \left( \frac{2}{\phi} t - \frac{1}{2} (\phi t)^2 \right) & \text{if } 0 < t \leq 1/\phi \\
0 & \text{if } t = 0.
\end{cases} \]

- While \( \gamma(0) = 0 \) by definition, \( \gamma(0^+) \equiv \lim_{t \to 0^+} \gamma(t) = \tau^2 \); this quantity is the *nugget*.
- \( \lim_{t \to \infty} \gamma(t) = \tau^2 + \sigma^2 \); this asymptotic value of the semivariogram is called the *sill*. (The sill minus the nugget, \( \sigma^2 \) in this case, is called the *partial sill*.)
- Finally, the value \( t = 1/\phi \) at which \( \gamma(t) \) first reaches its ultimate level (the sill) is called the *range*, \( R \equiv 1/\phi \).

Some common isotropic covariance functions:

<table>
<thead>
<tr>
<th>Model</th>
<th>Covariance function, ( C(t) )</th>
</tr>
</thead>
</table>
| Linear      | \( C(t) = \begin{cases} 
0 & \text{if } t \geq 1/\phi \\
\tau^2 \left[ 1 - \frac{1}{2} \phi t + \frac{1}{2} (\phi t)^2 \right] & \text{if } 0 < t \leq 1/\phi \\
\tau^2 + \sigma^2 & \text{otherwise}
\end{cases} \)                        |
| Spherical   | \( C(t) = \begin{cases} 
\tau^2 \exp(-\phi t) & \text{if } t > 0 \\
\tau^2 + \sigma^2 & \text{otherwise}
\end{cases} \)                            |
| Exponential | \( C(t) = \begin{cases} 
\tau^2 \exp(-\phi t) & \text{if } t > 0 \\
\tau^2 + \sigma^2 & \text{otherwise}
\end{cases} \)                            |
| Matérn     | \( C(t) = \begin{cases} 
\tau^2 (1 + \phi t) \exp(-\phi t) & \text{if } t > 0 \\
\tau^2 + \sigma^2 & \text{otherwise}
\end{cases} \)                        |

Notes on exponential model

\[ C(t) = \begin{cases} 
\tau^2 + \sigma^2 & \text{if } t = 0 \\
\sigma^2 \exp(-\phi t) & \text{if } t > 0.
\end{cases} \]

- We define the *effective range*, \( t_0 \), as the distance at which this correlation has dropped to only 0.05. Setting \( \exp(-\phi t_0) = 0.05 \) equal to this value we obtain \( t_0 \approx 3/\phi \), since \( \log(0.05) \approx -3 \).
- Finally, the form of \( C(t) \) shows why the nugget \( \tau^2 \) is often viewed as a “non-spatial effect variance,” and the partial sill \( \sigma^2 \) is viewed as a “spatial effect variance.”
The Matérn Correlation Function

- Much of statistical modelling is carried out through correlation functions rather than variograms
- The Matérn is a very versatile family:
  \[
  C(t) = \begin{cases} 
  \frac{2^{\nu-1}}{\Gamma(\nu)} (2\sqrt{\nu}t)^{\nu} K_{\nu}(2\sqrt{\nu}t\phi) & \text{if } t > 0 \\
  \sigma^2 \left( t^2 + \nu \right) & \text{if } t = 0
  \end{cases}
  \]

  where \( K_{\nu} \) is the modified Bessel function of order \( \nu \) (computationally tractable)
  \( \nu \) is a smoothness parameter (a fractal) controlling process smoothness.

How do we select a variogram? Can the data really distinguish between variograms?

Empirical Variogram:

\[
\gamma(t) = \frac{1}{2|N(t)|} \sum_{s_i, s_j \in N(t)} (Y(s_i) - Y(s_j))^2
\]

where \( N(t) \) is the number of points such that \( \|s_i - s_j\| = t \) and \( |N(t)| \) is the number of points in \( N(t) \).

Grid up the \( t \) space into intervals \( I_1 = (0, t_1), I_2 = (t_1, t_2), \) and so forth, up to \( I_K = (t_{K-1}, t_K) \). Representing \( t \) values in each interval by its midpoint, we define:

\[
N(t_k) = \{(s_i, s_j) : \|s_i - s_j\| \in I_k\}, k = 1, \ldots, K.
\]