Introduction to Spatial Data

Dipankar Bandyopadhyay

Div. of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, U.S.A.

January 17, 2012
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.

⇒ motivates hierarchical modeling and data analysis for complex spatial (and spatiotemporal) data sets.
Example: In an epidemiological investigation, we might wish to analyze lung, breast, colorectal, and cervical cancer rates

- by county and year in a particular state
- with smoking, mammography, and other important screening and staging information also available at some level.
Public health professionals who collect such data are charged not only with surveillance, but also statistical inference tasks, such as

- *modeling* of trends and correlation structures
- *estimation* of underlying model parameters
- *hypothesis testing* (or comparison of competing models)
- *prediction* of observations at unobserved times or locations.
Cressie (1990, 1993): the legendary ‘bible’ of spatial statistics, but
- rather high mathematical level
- lacks modern hierarchical modeling/computing

Wackernagel (1998): terse; only geostatistics

Chiles and Delfiner (1999): only geostatistics

Stein (1999a): theoretical treatise on Kriging.

Our primary focus is on the issues of modeling, computing, and data analysis.
- **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).
Map of PM2.5 sampling sites; plotting color indicates range of average 2001 level.
ArcView poverty map, regional survey units in Hennepin County, MN.
The previous figure is an example of a *choropleth map*, which uses shades of color (or greyscale) to classify values into a few broad classes, like a histogram.

From the choropleth map we know which regions are adjacent to (touch) which other regions.

Thus the “sites” $s \in D$ in this case are actually the regions (or *blocks*) themselves, which we will denote not by $s_i$ but by $B_i$, $i = 1, \ldots, n$.

It may be helpful to think of the county *centroids* as forming the vertices of an irregular lattice, with two lattice points being connected if and only if the counties are “neighbors” in the spatial map.
8-hour max. ozone levels (ppm) at 10 sites, July 15, ’95: Atlanta
Map of observed Scallops Sites
Introduction

Deterministic surface interpolation

- Spatial surface observed at finite set of locations
  \[ \mathcal{I} = \{s_1, s_2, \ldots, s_n\} \]

- Tessellate the spatial domain (usually with data locations as vertices)

- Fit an interpolating polynomial:
  \[ f(s) = \sum_i w_i(\mathcal{I}; s) f(s_i) \]

- “Interpolate” by reading off \( f(s_0) \).

Issues:
- Sensitivity to tessellations
- Choices of multivariate interpolators
- Numerical error analysis
Scallop data: Contour lines super-imposed
Scallop data: Drop-line scatter plot
Stockton real estate data: Surface plot
Stockton real estate data: Image contour plot
Locations form patterns
Surface features
Exemplified by residences of persons suffering from a particular disease, or by locations of a certain species of tree in a forest.

The response $Y$ is often fixed (occurrence of the event), and only the locations $s_i$ are thought of as random.

Such data are often of interest in studies of event clustering, where the goal is to determine whether points tend to be spatially close to other points, or result merely from a random process operating independently and homogeneously over space.

In contrast to areal data, here (and with point-referenced data as well) precise locations are known, and so must often be protected to protect the privacy of the persons in the set.
‘No clustering’ is often described through a homogeneous Poisson process:

\[ E[\text{number of occurrences in region } A] = \lambda |A|, \]

where \( \lambda \) is the \textit{intensity} parameter, and \( |A| \) is \text{area}(\( A \)).

Visual tests can be unreliable (tendency of the human eye to see clustering), so instead we might rely on \textit{Ripley’s K function},

\[ K(d) = \frac{1}{\lambda} E[\text{number of points within } d \text{ of an arbitrary point}], \]

where again \( \lambda \) is the intensity of the process, i.e., the mean number of points per unit area.
For a point process with no spatial dependence (a non-spatial process), \( K(d) = \pi d^2 = (1/\lambda) \ast (\lambda \pi d^2) \)

The usual estimator for \( K \) over region \( A \) is

\[
\hat{K}(d) = n^{-2} |A| \sum \sum_{i \neq j} p_{ij}^{-1} I_d(d_{ij}) ,
\]

where \( n \) is the number of points in \( A \), \( d_{ij} \) is the distance between points \( i \) and \( j \), \( p_{ij} \) is the proportion of the circle with center \( i \) and passing through \( j \) that lies within \( A \), and \( I_d(d_{ij}) \) equals 1 if \( d_{ij} < d \), and 0 otherwise.

Compare this to, say, \( K(d) = \pi d^2 \), the theoretical value for nonspatial processes.

Clustered data would have larger \( K \); uniformly spaced data would have a smaller \( K \).
The earth is round! So (longitude, latitude) \( \neq (x, y) \)!

A map projection is a systematic representation of all or part of the surface of the earth on a plane.

**Theorem:** The sphere cannot be flattened onto a plane without distortion

Instead, use an intermediate surface that can be flattened. The sphere is first projected onto the this developable surface, which is then laid out as a plane.

The three most commonly used surfaces are the cylinder, the cone, and the plane itself. Using different orientations of these surfaces lead to different classes of map projections...
Geometric constructions of projections

Regular Cylindrical

Transverse Cylindrical

Regular Conic

Polar Azimuthal (plane)

Oblique Cylindrical

Oblique Azimuthal (plane)
Map projections seek functions $f(\cdot)$ and $g(\cdot)$: Writing (longitude, latitude) as $(\lambda, \phi)$, projections are

\[
x = f(\lambda, \phi), \quad y = g(\lambda, \phi),
\]

where $f$ and $g$ are appropriate functions to be determined, based upon the properties we want our map to possess.

Compare infinitesimal patches on the sphere and the plane to derive a set of pde’s for $f$ and $g$.

Equal area projections must satisfy

\[
\left( \frac{\partial f}{\partial \lambda} \frac{\partial g}{\partial \phi} - \frac{\partial f}{\partial \phi} \frac{\partial g}{\partial \lambda} \right) = R^2 \cos \phi.
\]

Conformal (equal-angle) projections must satisfy

\[
\frac{\partial f}{\partial \lambda} \frac{\partial f}{\partial \phi} + \frac{\partial g}{\partial \lambda} \frac{\partial g}{\partial \phi} = 0.
\]
This *sinusoidal* projection obtained by specifying $\partial g / \partial \phi = R$, which yields equally-spaced straight lines for the parallels, and results in (with the 0 degree meridian as the central meridian),

$$f(\lambda, \phi) = R\lambda \cos \phi; \quad g(\lambda, \phi) = R\phi.$$
The Mercator projection is a conformal projection that distorts areas (badly at the poles):

\[ f(\lambda, \phi) = R\lambda; \quad g(\lambda, \phi) = R \ln \tan \left( \frac{\pi}{4} + \frac{\phi}{2} \right). \]
The basic geometry behind calculating geodesic distances

Consider two points on the surface of the earth, $P_1 = (\theta_1, \lambda_1)$ and $P_2 = (\theta_2, \lambda_2)$, where $\theta =$ latitude and $\lambda =$ longitude.

The geodesic distance we seek is $D = R\phi$, where

- $R$ is the radius of the earth
- $\phi$ is the angle subtended by the arc connecting $P_1$ and $P_2$ at the center
From elementary trigonometry, the coords on a sphere are

\[ x = R \cos \theta \cos \lambda, \quad y = R \cos \theta \sin \lambda, \quad \text{and} \quad z = R \sin \theta \]

Assume a unit sphere (i.e. \( R = 1 \)). Letting \( u_1 = (x_1, y_1, z_1) \) and \( u_2 = (x_2, y_2, z_2) \), we know

\[ \cos \phi = \frac{\langle u_1, u_2 \rangle}{||u_1|| \cdot ||u_2||} = \langle u_1, u_2 \rangle. \]

We now compute

\[
\begin{align*}
\langle u_1, u_2 \rangle &= \cos \theta_1 \cos \lambda_1 \cos \theta_2 \cos \lambda_2 + \cos \theta_1 \sin \lambda_1 \cos \theta_2 \sin \lambda_2 \\
&\quad + \sin \theta_1 \sin \theta_2 \\
&= \cos \theta_1 \cos \theta_2 \cos (\lambda_1 - \lambda_2) + \sin \theta_1 \sin \theta_2
\end{align*}
\]

For a sphere of radius \( R \), our final answer is

\[ D = R\phi = R \arccos[\cos \theta_1 \cos \theta_2 \cos(\lambda_1 - \lambda_2) + \sin \theta_1 \sin \theta_2]. \]