Introduction to Spatial Data

Sudipto Banerjee\(^1\) and Alan E. Gelfand\(^2\)

\(^1\) Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, U.S.A.

\(^2\) Statistics and Decision Sciences, Duke University, Durham, North Carolina, U.S.A.

May 16, 2010
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.
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,
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

This motivates hierarchical modeling and data analysis for complex spatial (and spatiotemporal) data sets.
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.
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.
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;
- **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-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.
Misaligned areal and point data: Atlanta zip codes and 8-hour maximum ozone levels (ppm) at 10 sites, July 15, 1995.
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.
Scallops Sites
Spatial surface observed at finite set of locations
\( I = \{s_1, s_2, \ldots, s_n\} \)

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

Fit an interpolating polynomial:

\[
  f(s) = \sum_i w_i(I; s) f(s_i)
\]

“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

ENAR 2009 Hierarchical Modeling and Analysis
Drop-line scatter plot

![Drop-line scatter plot image]
Surface plot

logSP

Latitude

Longitude

-121.36
-121.34
-121.32
-121.3
-121.28
Introduction to spatial data and models

Image contour plot

Scallops data: image and contour plots
Locations form patterns
Surface features
Interesting plot arrangements
The earth is round! So \((\text{longitude}, \text{latitude}) \neq (x, y)\)!

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
- Oblique Cylindrical
- Regular Conic
- Polar Azimuthal (plane)
- Oblique Azimuthal (plane)
Map projections seek functions $f(\cdot)$ and $g(\cdot)$: Writing (longitude, latitude) as $(\lambda, \theta)$, 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 $\frac{\partial g}{\partial \phi} = R$, 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 $\theta = \text{latitude}$ and $\lambda = \text{longitude}$.

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 \( \mathbf{u}_1 = (x_1, y_1, z_1) \) and \( \mathbf{u}_2 = (x_2, y_2, z_2) \), we know
  \[ \cos \phi = \frac{\langle \mathbf{u}_1, \mathbf{u}_2 \rangle}{||\mathbf{u}_1|| \ ||\mathbf{u}_2||} = \langle \mathbf{u}_1, \mathbf{u}_2 \rangle. \]

- We now compute
  \[ \langle \mathbf{u}_1, \mathbf{u}_2 \rangle = \cos \theta_1 \cos \lambda_1 \cos \theta_2 \cos \lambda_2 + \cos \theta_1 \sin \lambda_1 \cos \theta_2 \sin \lambda_2 \]
  \[ + \sin \theta_1 \sin \theta_2 \]
  \[ = \cos \theta_1 \cos \theta_2 \cos (\lambda_1 - \lambda_2) + \sin \theta_1 \sin \theta_2 \]

- For a sphere of radius \( R \), our final answer is
  \[ D = R \phi = R \arccos[\cos \theta_1 \cos \theta_2 \cos(\lambda_1 - \lambda_2) + \sin \theta_1 \sin \theta_2]. \]