Hierarchical Modeling and Analysis for Spatial Data

Bradley P. Carlin, Sudipto Banerjee , and Alan E. Gelfand

brad@biostat.umn.edu, sudiptob@biostat.umn.edu, and alan@stat.duke.edu

University of Minnesota and Duke University

Introduction to spatial data and models

- 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.

$\Rightarrow$ motivates *hierarchical* modeling and data analysis for complex spatial (and spatiotemporal) data sets.

Introduction (cont'd)

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.

Introduction (cont'd)

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.

$\Longrightarrow$ all naturally accomplished through hierarchical modeling implemented via Markov chain Monte Carlo (MCMC) methods!

Existing spatial statistics books

- 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

More descriptive presentations: Bailey and Gattrell (1995), Fotheringham and Rogerson (1994), or Haining (1990).

Our primary focus is on the issues of modeling, computing, and data analysis.

Type of spatial data

- *point-referenced data*, where $Y(\mathbf{s})$ is a random vector at a location $\mathbf{s} \in \Re^r$, where $\mathbf{s}$ varies *continuously* over D, a fixed subset of $\Re^r$ 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. $Y(\mathbf{s})$ itself can simply equal 1 for all $\mathbf{s} \in D$ (indicating occurrence of the event), or possibly give some additional covariate information (producing a *marked point pattern process*).

Point-level (geostatistical) data

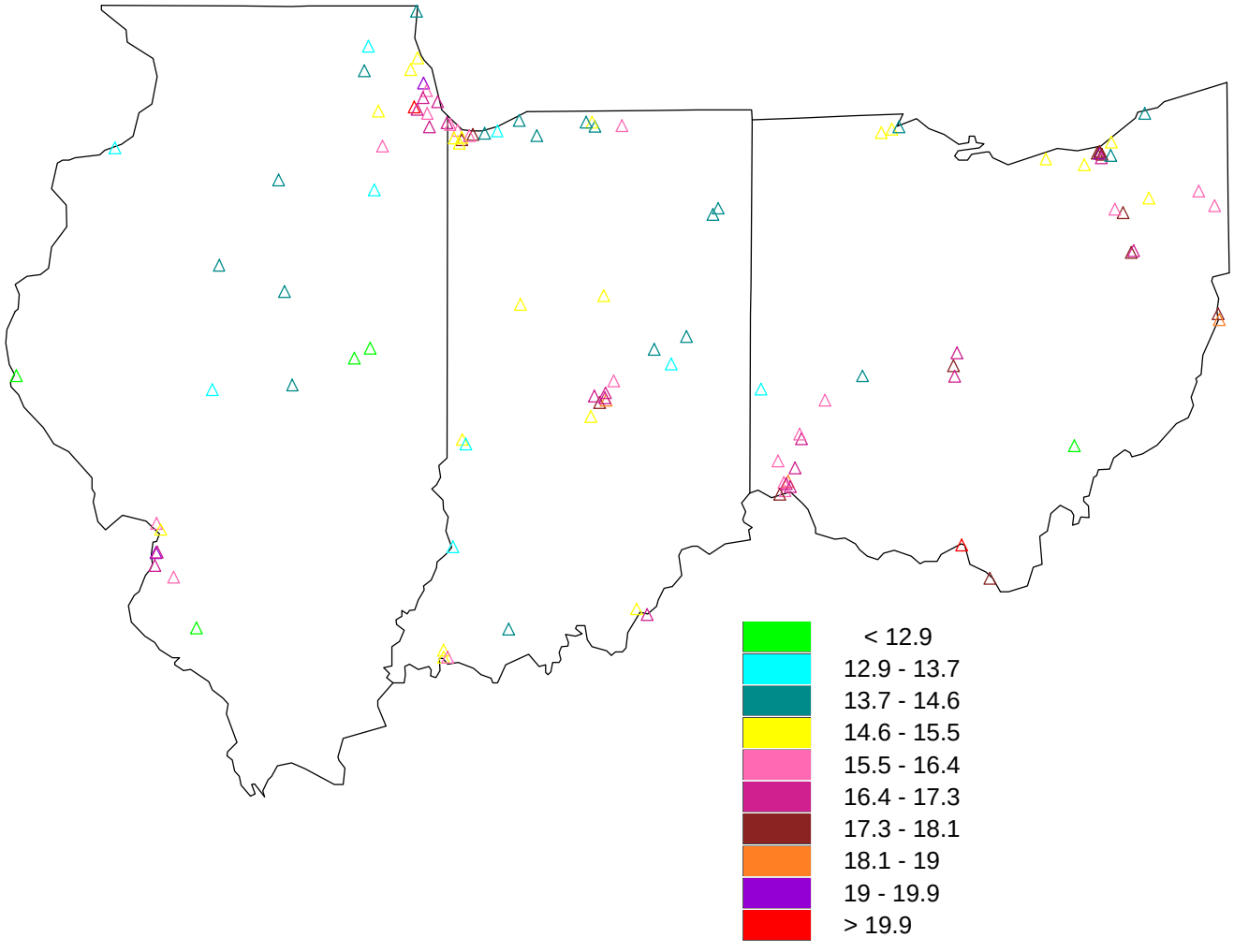


Figure 1: Map of PM2.5 sampling sites; plotting color indicates range of average 2001 level

Areal (lattice) data

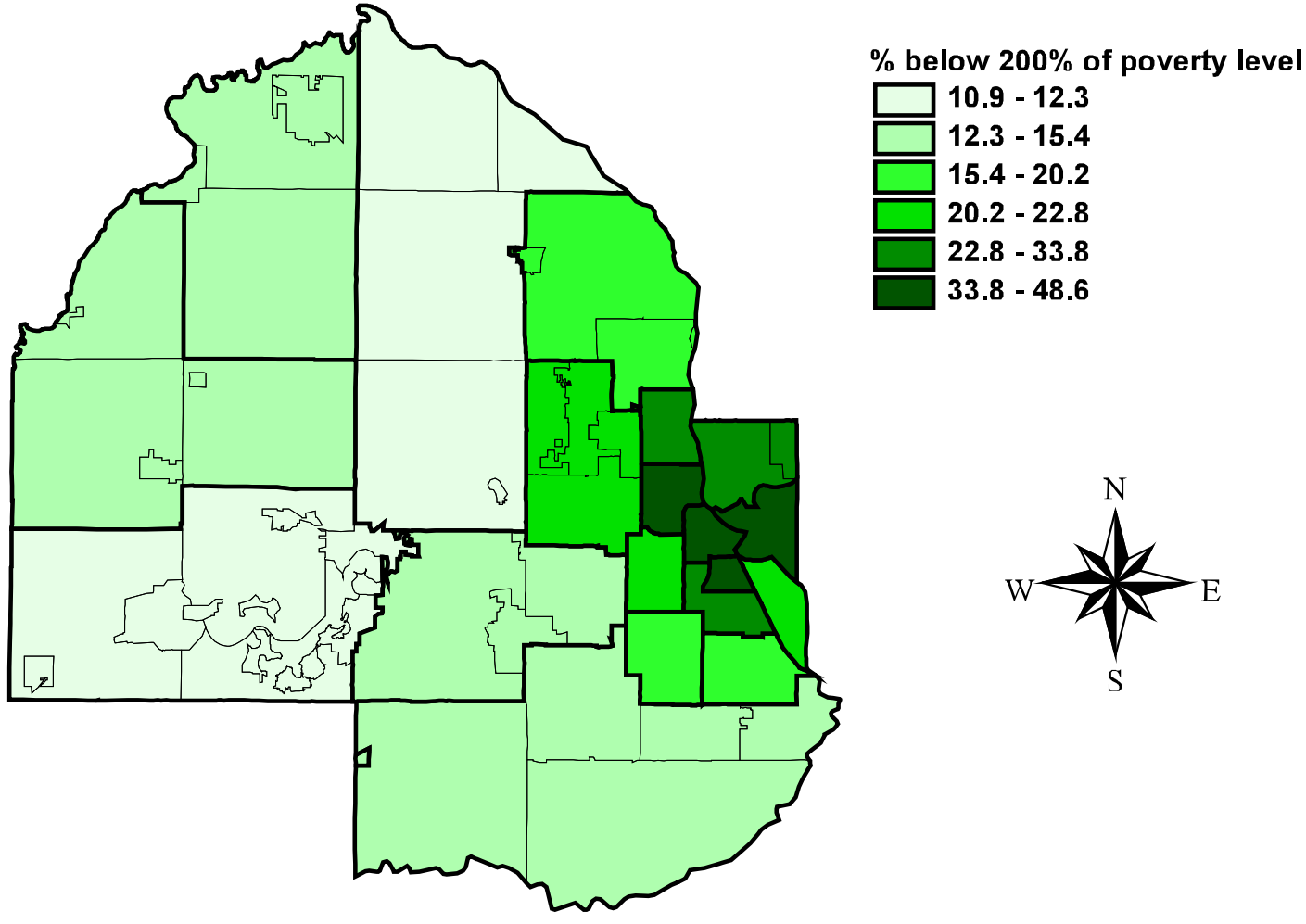


Figure 2: `ArcView` poverty map, regional survey units in Hennepin County, MN.

Notes on areal data

- Figure 2 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" $\mathbf{s} \in D$ in this case are actually the regions (or *blocks*) themselves, which we will denote not by $\mathbf{s}_i$ 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 (point and areal) data

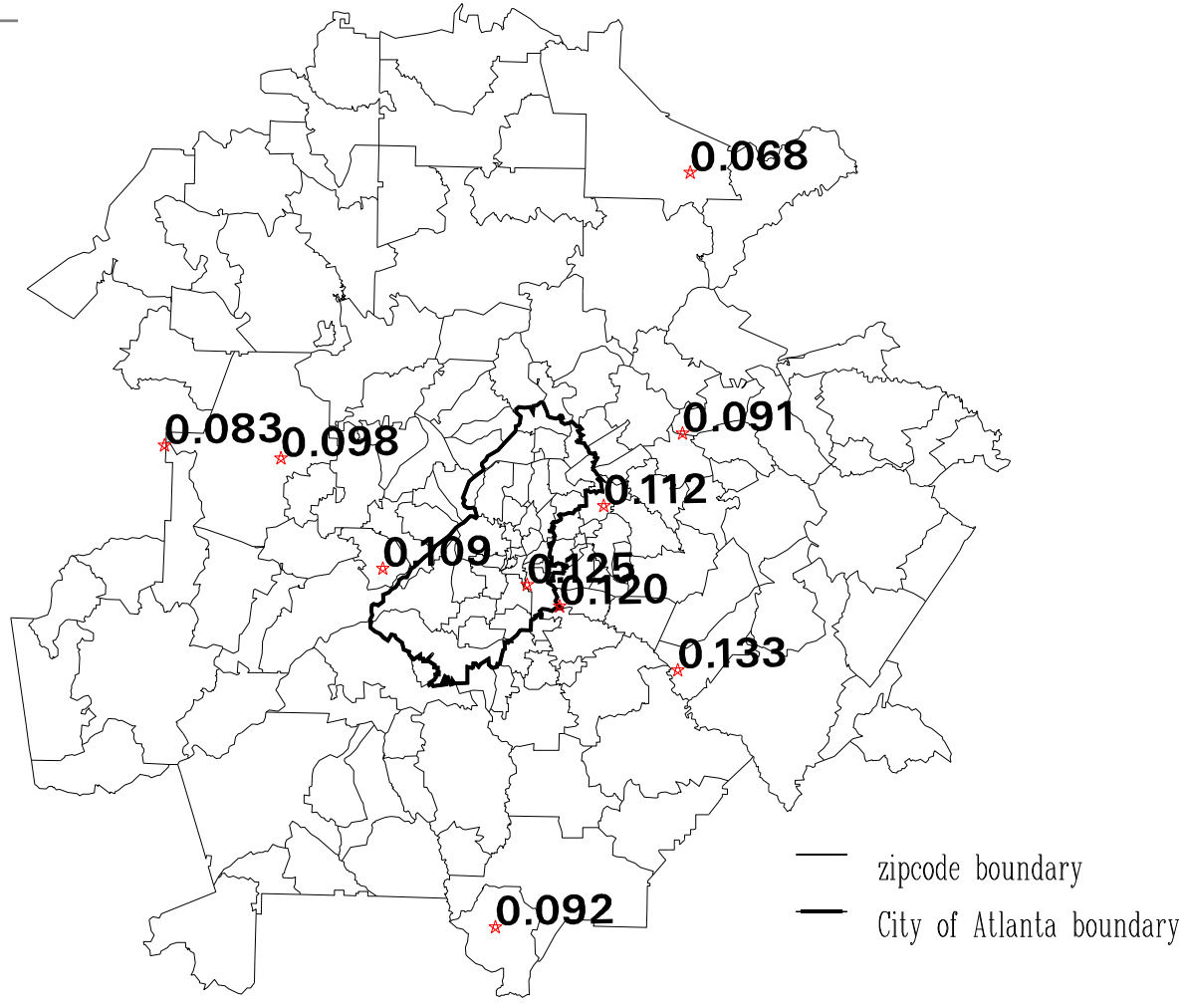


Figure 3: Atlanta zip codes and 8-hour maximum ozone levels (ppm) at 10 sites, July 15, 1995.

Spatial point process data

- 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 $\mathbf{s}_i$ 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.

Spatial point process data (cont'd)

- "No clustering" is often described through a homogeneous Poisson process:

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

where λ is the *intensity* parameter, and $|A|$ is area(A).

- Visual tests can be unreliable (tendency of the human eye to see clustering), so instead we might rely on *Ripley's K function*,

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

where again λ is the intensity of the process, i.e., the mean number of points per unit area.

Spatial point process data (cont'd)

- The usual estimator for K is

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

  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

Spatial point process summary

- A popular spatial add-on to the S+ package, `S+SpatialStats`, allows computation of K for any data set, as well as approximate 95% intervals for it
- Full inference likely requires use of the `Splancs` software, or perhaps a fully Bayesian approach along the lines of Wakefield and Morris (2001).

- We consider only a fixed index set D, i.e., random observations at either points $\mathbf{s}_i$ or areal units B_i; see
  - Diggle (2003)
  - Lawson and Denison (2002)
  - Møller and Waagepetersen (2004)

  for recent treatments of spatial point processes and spatial cluster detection and modeling.

Fundamentals of Cartography

- 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...

Developable surfaces

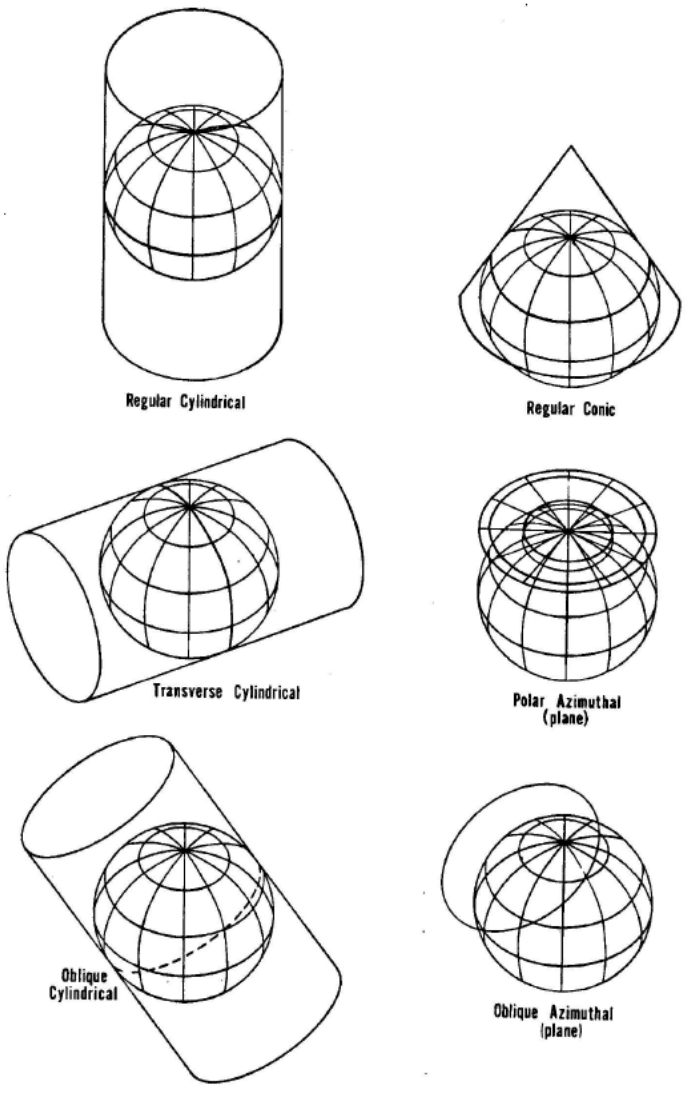
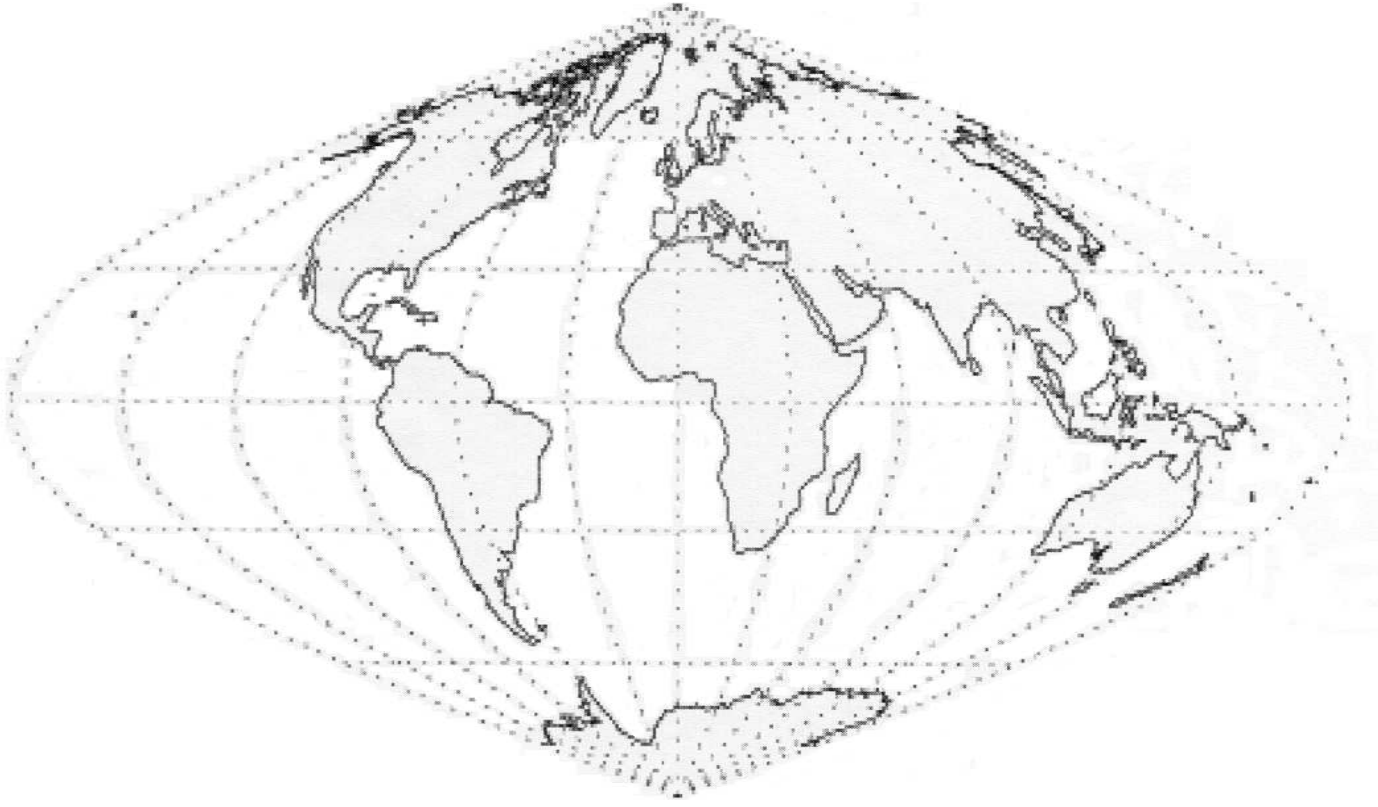


Figure 4: Geometric constructions of projections

Sinusoidal projection

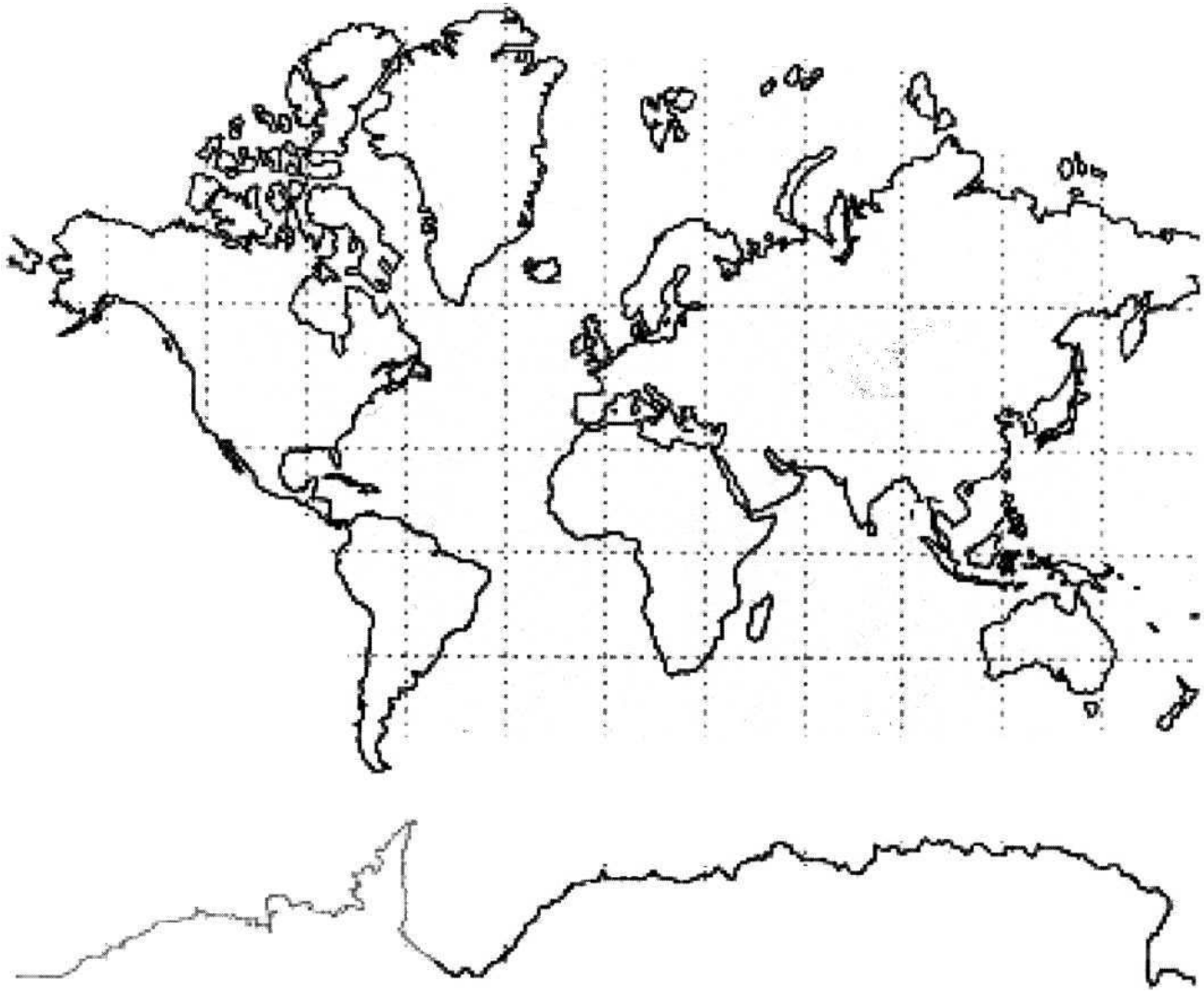


Writing (longitude, latitude) as (λ, θ), projections are

$$x = f(\lambda, \phi),\ y = g(\lambda, \phi)\ ,$$

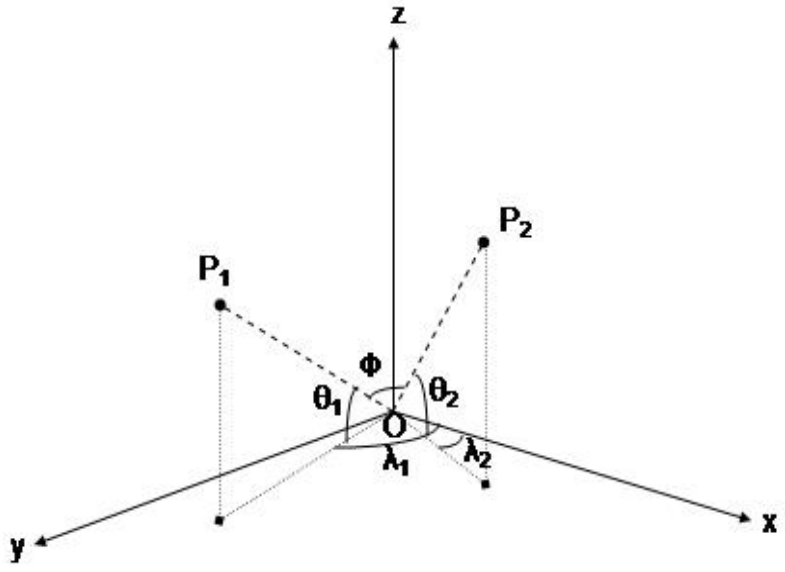
where f and g are chosen based upon properties our map must possess. This sinusoidal projection preserves area.

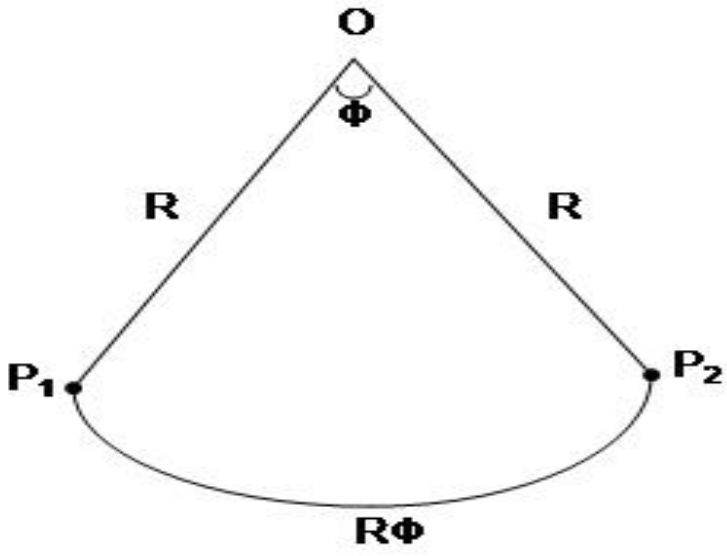
Mercator projection



While no projection preserves distance (Gauss' Theorema Eggregium in differential geometry), this famous conformal (angle-preserving) projection distorts badly near the poles.

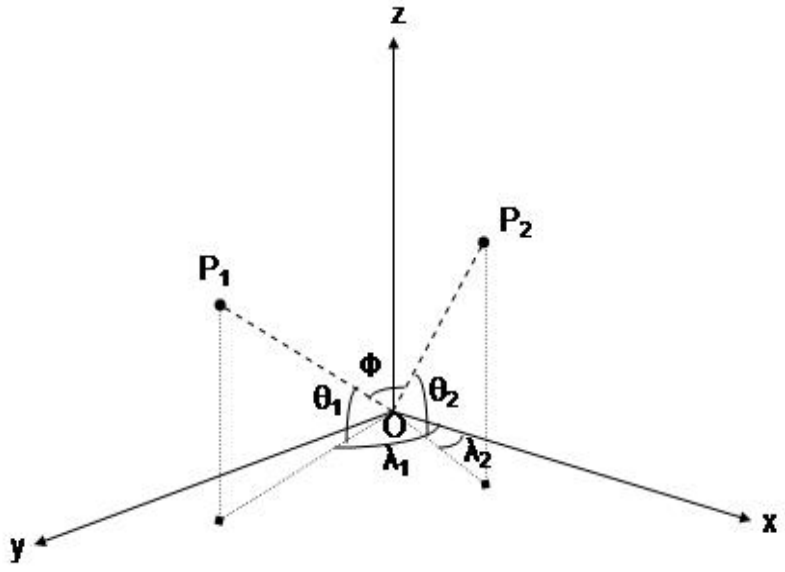
Calculation of geodesic distance

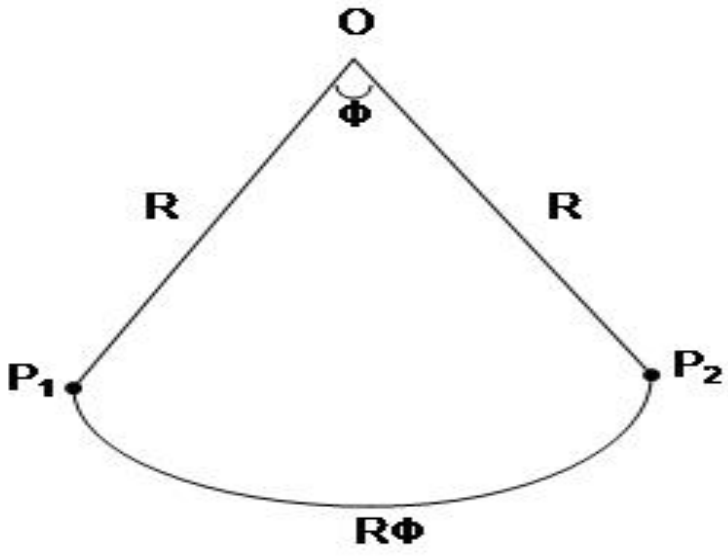




- Consider two points on the surface of the earth, $P_1 = (\theta_1, \lambda_1)$ and $P_2 = (\theta_2, \lambda_2)$, where $\theta =$ latitude and $\lambda =$ longitude.
- The geodesic distance we seek is $D = R\phi$, where
  - R is the radius of the earth
  - ϕ is the angle subtended by the arc connecting P_1 and P_2 at the center

Calculation of geodesic distance (cont'd)





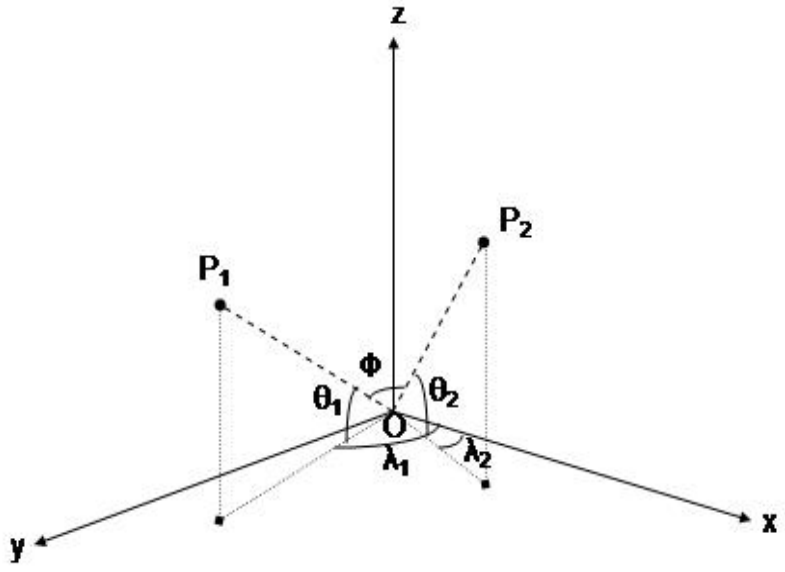
- From elementary trig, we have

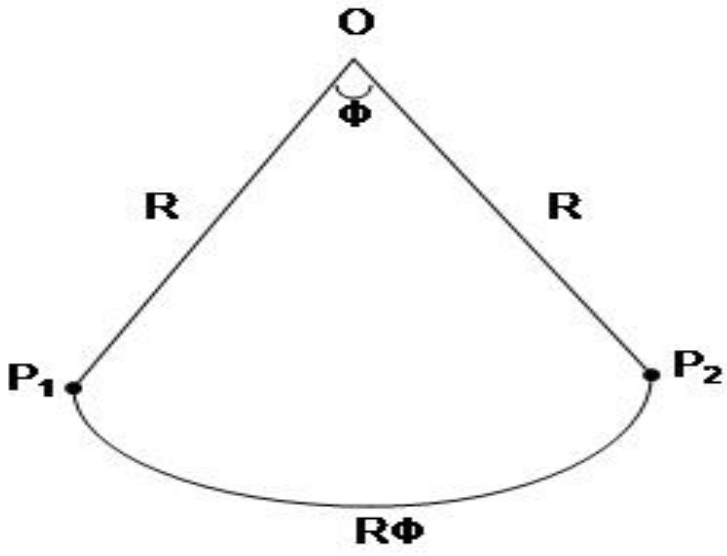
$$x = R\cos\theta\cos\lambda,\ y = R\cos\theta\sin\lambda, \quad \text{and} \quad z = R\sin\theta$$

- Letting $\mathbf{u}_1 = (x_1, y_1, z_1)$ and $\mathbf{u}_2 = (x_2, y_2, z_2)$, we know

$$\cos\phi = \frac{\langle \mathbf{u}_1, \mathbf{u}_2 \rangle}{||\mathbf{u}_1||\,||\mathbf{u}_2||}$$

Calculation of geodesic distance (cont'd)





- We then compute $\langle \mathbf{u}_1, \mathbf{u}_2 \rangle$ as

$$R^2 \left[\cos\theta_1 \cos\lambda_1 \cos\theta_2 \cos\lambda_2 + \cos\theta_1 \sin\lambda_1 \cos\theta_2 \sin\lambda_2 + \sin\theta_1 \sin\theta_2\right]$$
$$= R^2 \left[\cos\theta_1 \cos\theta_2 \cos\left(\lambda_1 - \lambda_2\right) + \sin\theta_1 \sin\theta_2\right] \ .$$

- But $||\mathbf{u}_1|| = ||\mathbf{u}_2|| = R$, so our final answer is

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