CBMS Lecture 1

Alan E. Gelfand
Duke University

Introduction to spatial data and models

- Researchers in diverse areas such as climatology, ecology, environmental exposure, public 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 risk factors, e.g., age, race, smoking, mammography, and other important screening and staging information also available at some level.

Introduction (cont'd)

Example: In a meteorological investigation, we might wish to analyze temperature and precipitation data

- with say hourly or daily measurements at a network of monitoring station
- with a mean surface that reflects say elevation, perhaps a trend in elevation

Introduction (cont'd)

Example: In an ecological setting, we may be interested in the point pattern of locations for say two different species, e.g., juniper trees and pine trees

- with geo-coded locations for each of the trees and a label indicating which species
- and environmental features to explain species distribution
- possibly collected over time in order to see change, evolution, diffusion of the patterns

Introduction (cont'd)

One may be interested in displaying the data collected but may also have interest in carrying out statistical *inference* tasks, such as

- modeling of trends and correlation structures
- estimation of underlying model parameters
- hypothesis testing (or comparison of competing models)
- prediction at unobserved times or locations
- regression specifications to explain spatial response.
- Conceptualize a process model specification:
  [data—process,parameters]
  [process—parameters][parameters]

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

Existing spatial statistics books

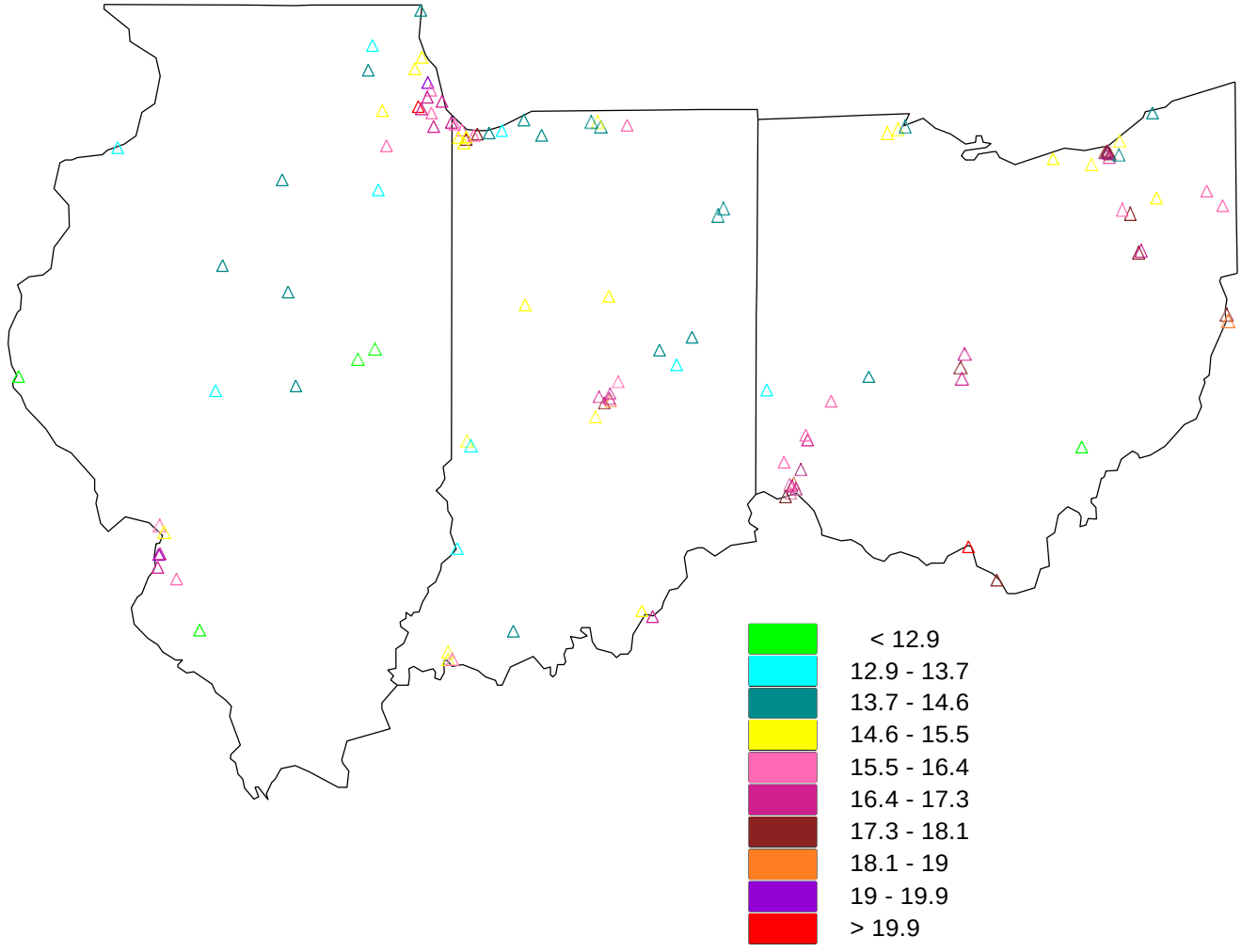
- Cressie (1990, 1993): the legendary "bible" of spatial statistics, but rather high mathematical level, lacks modern hierarchical modeling/computing
- Update is Cressie and Wikle (2011) Statistics for Spatio-temporal Data
- The Handbook of Spatial Statistics (Gelfand et al., 2010)
- Wackernagel (1998): terse; only geostatistics
- Chiles and Delfiner (1999): only geostatistics
- Stein (1999a): theoretical treatise on kriging
- So, of course Banerjee, Carlin, and Gelfand (2014)!
- More descriptive presentations: Bailey and Gattrell (1995), Fotheringham and Rogerson (1994), or Haining (1990).

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

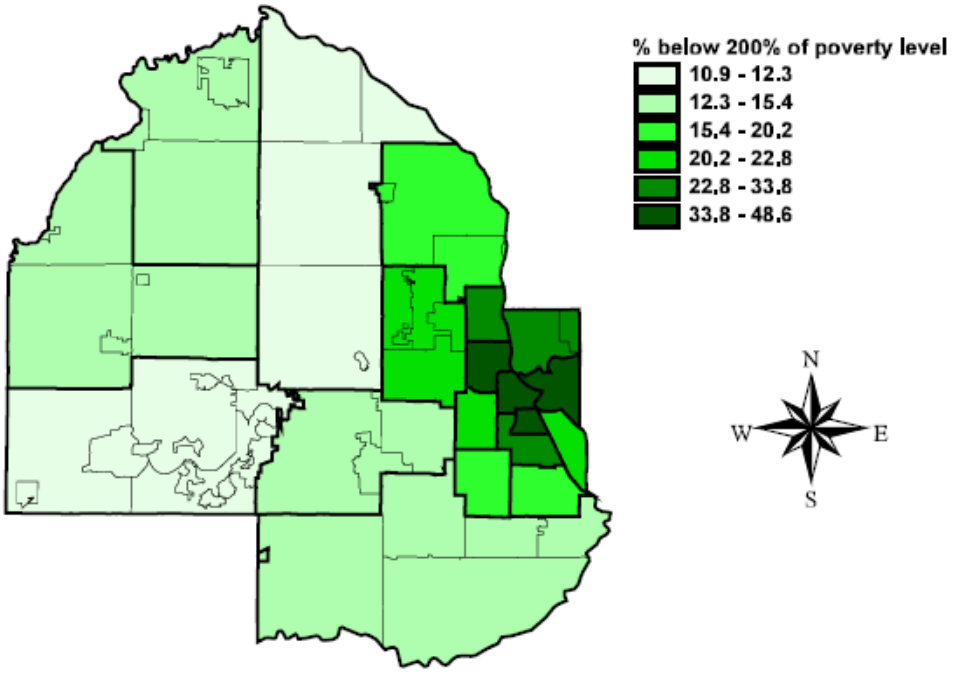
Types of spatial data

- point-referenced data, where $Y(\mathbf{s})$ is a random vector at a selected location $\mathbf{s} \in \Re^r$ and $\mathbf{s}$ varies continuously over D, a fixed subset of $\Re^r$;
- 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 and observations are associated with the areal units; discrete spatial data
- point pattern data, where now the set of locations in D is itself random; its index set gives the locations of random events that are the spatial point pattern. Can assign $Y(\mathbf{s}) = 1$ for all $\mathbf{s} \in D$ (indicating occurrence of the event), or possibly assign labels to the points (producing a marked point pattern process).

Point-level (geostatistical) data



Areal (lattice) data

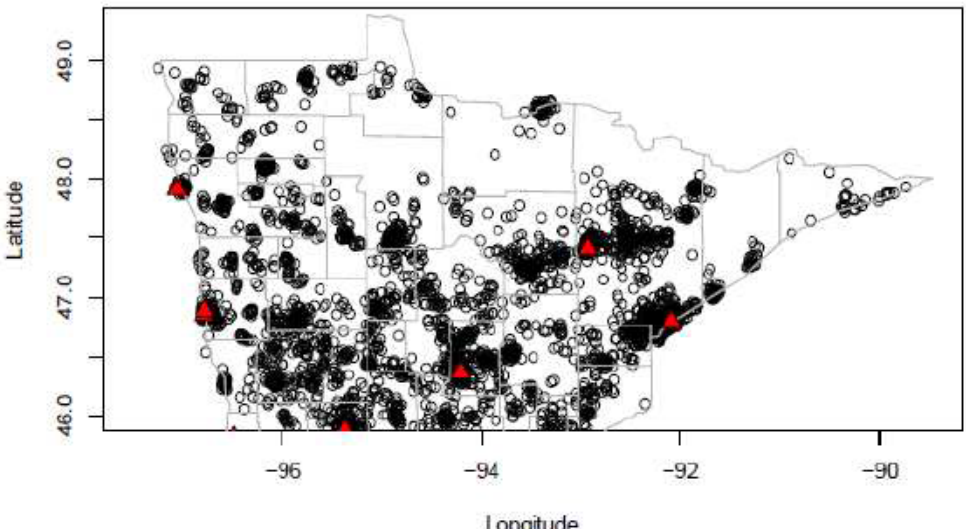


Notes on areal data

- This 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 (share a boundary or a corner) 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 $B_i,\ i = 1, \ldots, n$.

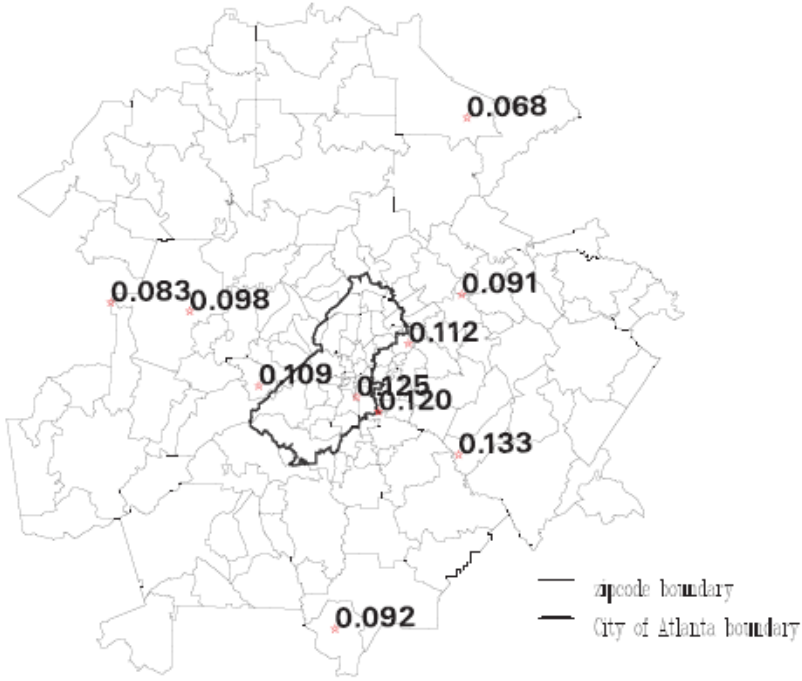
Point pattern data

N Minnesota Breast Cancer Data



- Jittered residential locations of cases, as well as radiation treatment facilities (RTFs; triangles), northern Minnesota, 1998–2002.

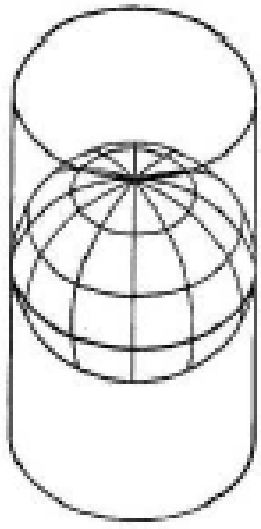
Misaligned (point and areal) data



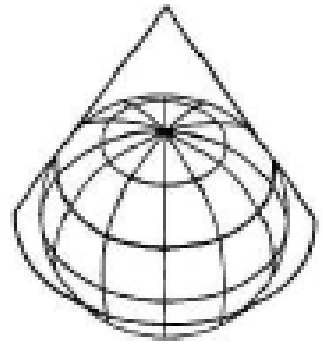
A few words on cartography

- 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 leads to different classes of map projections...

Developable surfaces

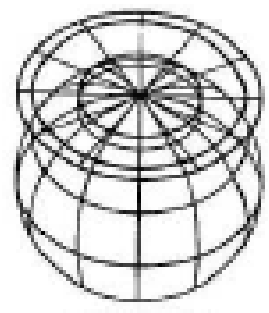


Regular Cylindrical

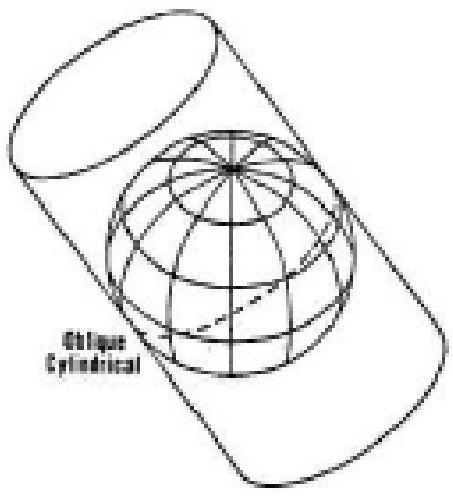


Regular Conic

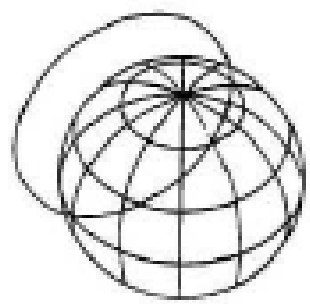
Transverse Cylindrical



Polar Azimuthal (plane)



Oblique Cylindrical



Oblique Azimuthal (plane)

Sinusoidal projection

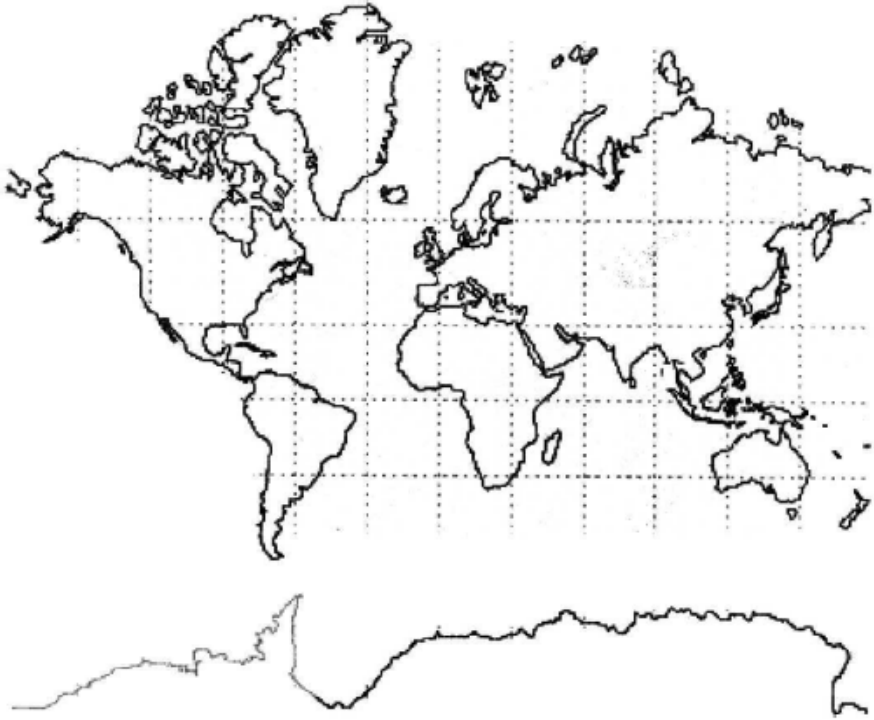
cont.

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



cont.

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

Basics of Point-Referenced Data Models

- Basic tool is a *spatial process*, $\{Y(\mathbf{s}), \mathbf{s} \in D\}$, where $D \subset \Re^r$
- Note that time series follows this approach with $r = 1$; we will usually have $r = 2$ or 3
- We begin with essentials of point-level data modeling, including stationarity, isotropy, and variograms – key elements of the "Matheron school"
- No formal inference, just least squares optimization
- We add the spatial (typically Gaussian) process modeling that enables likelihood (and Bayesian) inference in these settings.

Scallops catch sites, NY/NJ coast, USA

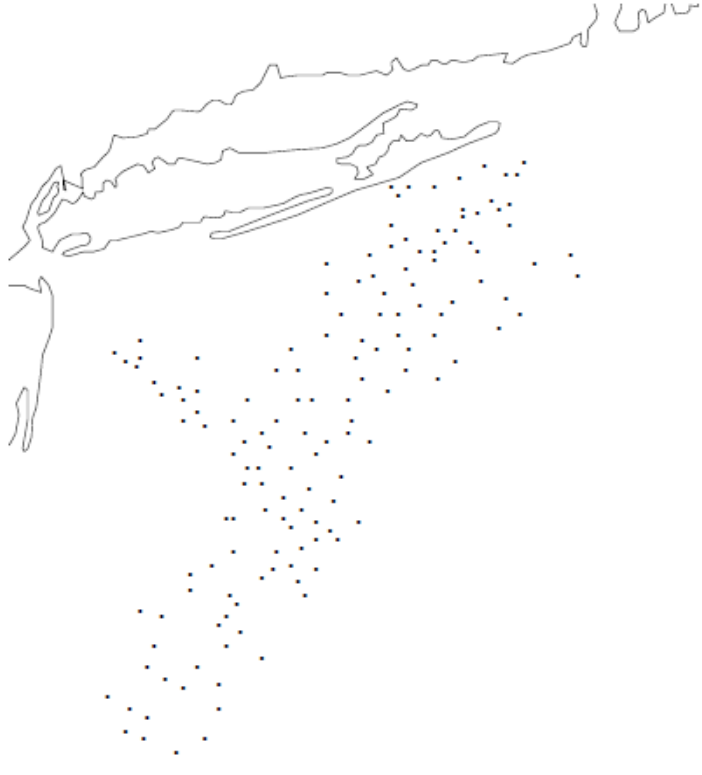
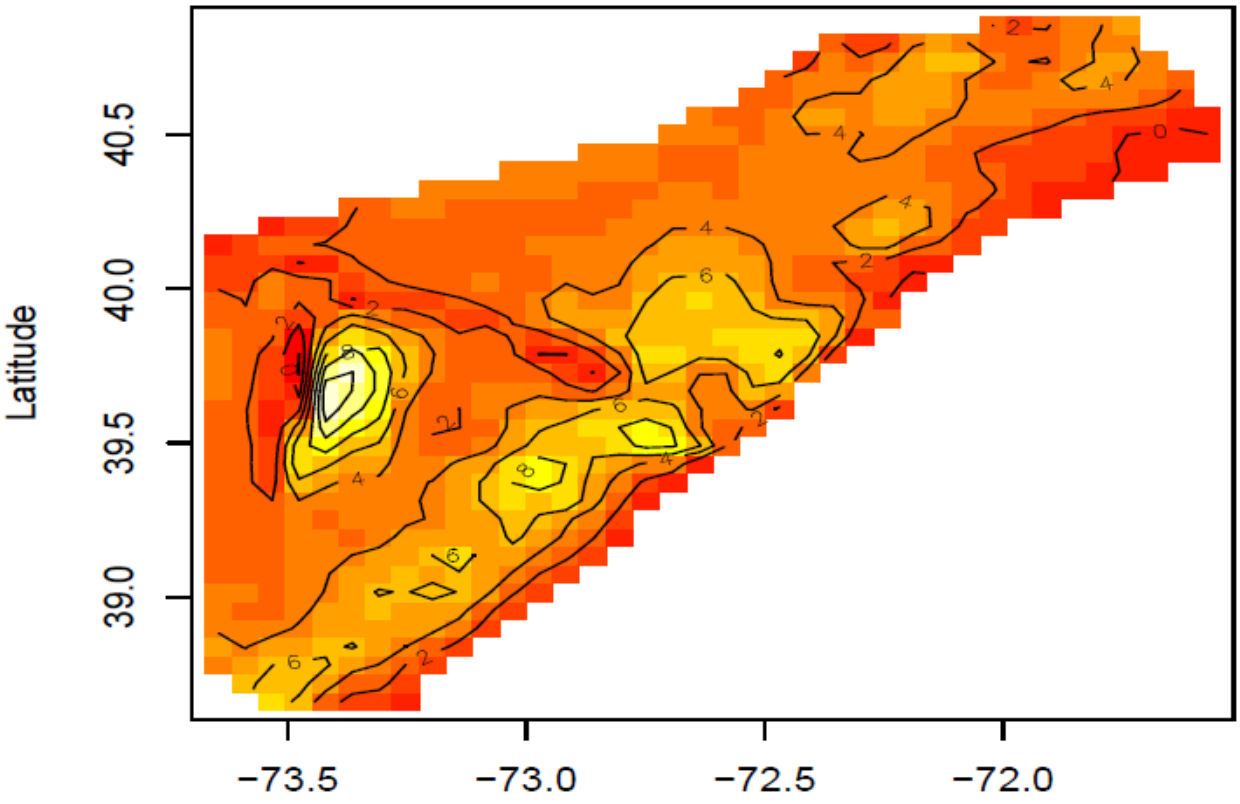


Image plot with log-catch contours



Stationarity

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

- The process is said to be strictly stationary (also called strongly stationary) if, for any given $n \geq 1$, any set of n sites $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ and any $\mathbf{h} \in \Re^r$, the distribution of $(Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))$ is the same as that of $(Y(\mathbf{s}_1 + \mathbf{h}), \ldots, Y(\mathbf{s}_n + \mathbf{h}))$.
- A less restrictive condition is given by weak stationarity (also called second-order stationarity): A process is weakly stationary if $\mu(\mathbf{s}) \equiv \mu$ and $Cov(Y(\mathbf{s}), Y(\mathbf{s} + \mathbf{h})) = C(\mathbf{h})$ for all $\mathbf{h} \in \Re^r$ such that $\mathbf{s}$ and $\mathbf{s} + \mathbf{h}$ both lie within D.

Notes on Stationarity

- Weak stationarity says that the covariance between the values of the process at any two locations $\mathbf{s}$ and $\mathbf{s} + \mathbf{h}$ can be summarized by a covariance function $C(\mathbf{h})$ (sometimes called a covariogram), and this function depends only on the separation vector $\mathbf{h}$.
- Note that with all variances assumed to exist, strong stationarity implies weak stationarity.
- The converse is not true in general, but it does hold for Gaussian processes

Gaussian processes

- The process $Y(\mathbf{s})$ is said to be Gaussian, i.e., a Gaussian process, a GP, if, for any $n \geq 1$ and any set of sites $\{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n\}$, $\mathbf{Y} = (Y(\mathbf{s}_1), Y(\mathbf{s}_2), ..., Y(\mathbf{s}_n))$ has a multivariate normal distribution.
- How do we create the multivariate normal distribution?
- We specify a mean function $\mu(\mathbf{s})$ and a "valid" covariance function $C(\mathbf{s}, \mathbf{s}') \equiv \mathrm{cov}(Y(\mathbf{s}), Y(\mathbf{s}'))$
- Then, $\mathbf{Y} \sim N(\boldsymbol{\mu}, \Sigma)$ where $\boldsymbol{\mu}_i = \mu(\mathbf{s}_i)$ and $\Sigma_{ij} = C(\mathbf{s}_i, \mathbf{s}_j)$.
- The mean function is usually some sort of regression specification
- The covariance function is specified through a few parameters say $\boldsymbol{\theta}$, so we have $\Sigma(\boldsymbol{\theta})$ providing *structured* dependence

Why do we love GPs?

- Restriction to Gaussian processes enables several advantages.
- Convenient specification: the mean function and the covariance function determine all distributions.
- Convenient distribution theory. Joint marginal and conditional distributions are all immediately obtained from standard theory given the mean and covariance structure.
- With hierarchical modeling, a Gaussian process assumption for spatial random effects at the second stage of the model aligns with the way independent random effects with variance components are customarily introduced in foregoing linear or generalized linear mixed models.
- Technically, with Gaussian processes and stationary models, strong stationarity is equivalent to weak stationarity
- It is difficult to criticize a Gaussian assumption. We have $\mathbf{Y} = (Y(\mathbf{s}_1), Y(\mathbf{s}_2), ..., Y(\mathbf{s}_n))$, a single realization from an n-dimensional distribution. With a sample size of one, how can we criticize any multivariate distributional specification?

Wait a minute!

- Strictly speaking this last assertion is not quite true with a Gaussian process model.
- That is, the joint distribution is a multivariate normal with mean, say, 0, and a covariance matrix that is a parametric function of the parameters in the covariance function.
- As n grows large enough, the effective sample size will also grow.
- By linear transformation we can obtain a set of approximately uncorrelated variables through which the adequacy of the normal assumption might be studied.

Variograms

- Suppose we assume $E[Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})] = 0$ and define

$$E[Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})]^2 = Var\left(Y\left(\mathbf{s}+\mathbf{h}\right) - Y\left(\mathbf{s}\right)\right) = 2\gamma\left(\mathbf{h}\right) .$$

- This expression only looks at the difference between variables. If the left hand side depends *only* on $\mathbf{h}$ and not the particular choice of $\mathbf{s}$, we say the process is intrinsically stationary.
- The function $2\gamma\left(\mathbf{h}\right)$ is then called the variogram, and $\gamma\left(\mathbf{h}\right)$ is called the semivariogram.

Intrinsic stationarity requires only the first and second moments of the differences $Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})$. It says nothing about the joint distribution of a collection of variables $Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n)$, and thus provides no likelihood.

Relationship between $C(\mathbf{h})$ and $\gamma(\mathbf{h})$

- We have

$$\begin{aligned} 2\gamma(\mathbf{h}) &= Var\left(Y\left(\mathbf{s}+\mathbf{h}\right)-Y\left(\mathbf{s}\right)\right) \\ &= Var(Y(\mathbf{s}+\mathbf{h}))+Var(Y(\mathbf{s}))-2Cov(Y(\mathbf{s}+\mathbf{h}),Y(\mathbf{s})) \\ &= C(\mathbf{0})+C(\mathbf{0})-2C\left(\mathbf{h}\right) \\ &= 2\left[C\left(\mathbf{0}\right)-C\left(\mathbf{h}\right)\right]. \end{aligned}$$

- Thus,

$$\gamma\left(\mathbf{h}\right)=C\left(\mathbf{0}\right)-C\left(\mathbf{h}\right).$$

- So given C, we are able to determine γ.
- But what about the converse: can we recover C from γ?...

Relationship between $C(\mathbf{h})$ and $\gamma(\mathbf{h})$

- In the relationship $\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h})$ we can $\pm$ a constant on the right side so $C(\mathbf{h})$ is not identified
- Usually, we want the spatial process to be ergodic. Otherwise, no good inference properties.
- This means $C(\mathbf{h}) \to 0$ as $||\mathbf{h}|| \to \infty$, where $||\mathbf{h}||$ is the length of $\mathbf{h}$.
- If so, then, as $||\mathbf{h}|| \to \infty$, $\gamma(\mathbf{h}) \to C(\mathbf{0})$
- Hence, $C(\mathbf{h}) = C(\mathbf{0}) - \gamma(\mathbf{h})$ and both terms on the right side depend on $\gamma(\cdot)$ So $C(\mathbf{h})$ is now well defined given $\gamma(\mathbf{h})$

So, previous slide showed that weak stationarity implies intrinsic stationarity. The converse is not true in general but is with the above condition on $\gamma(\mathbf{h})$

Isotropy

- If the semivariogram $\gamma(\mathbf{h})$ depends upon the separation vector only through its length $||\mathbf{h}||$, then we say that the process is *isotropic*.
- For an isotropic process, $\gamma(\mathbf{h})$ is a real-valued function of a univariate argument, and can be written as $\gamma(||\mathbf{h}||)$.
- If the process is intrinsically stationary and isotropic, it is also called homogeneous.

Isotropic processes are popular because of their simplicity, interpretability, and because a number of relatively simple parametric forms are available as candidates for C (and γ). Denoting $||\mathbf{h}||$ by t for notational simplicity, the next two tables provide a few examples...

Some common isotropic covariograms

Model	Covariance function, $C(t)$
Linear	$C(t)$ does not exist
Spherical	$C(t) = \begin{cases} 0 & \text{if } t \geq 1/\phi \\ \sigma^2\left[1 - \frac{3}{2}\phi t + \frac{1}{2}(\phi t)^3\right] & \text{if } 0 < t \leq 1/\phi \\ \tau^2 + \sigma^2 & \text{if } t = 0 \end{cases}$
Exponential	$C(t) = \begin{cases} \sigma^2 \exp(-\phi t) & \text{if } t > 0 \\ \tau^2 + \sigma^2 & \text{if } t = 0 \end{cases}$
Powered exponential	$C(t) = \begin{cases} \sigma^2 \exp(-\lvert\phi t\rvert^p) & \text{if } t > 0 \\ \tau^2 + \sigma^2 & \text{if } t = 0 \end{cases}$
Matérn at $\nu = 3/2$	$C(t) = \begin{cases} \sigma^2 (1 + \phi t) \exp(-\phi t) & \text{if } t > 0 \\ \tau^2 + \sigma^2 & \text{if } t = 0 \end{cases}$

Some common isotropic variograms

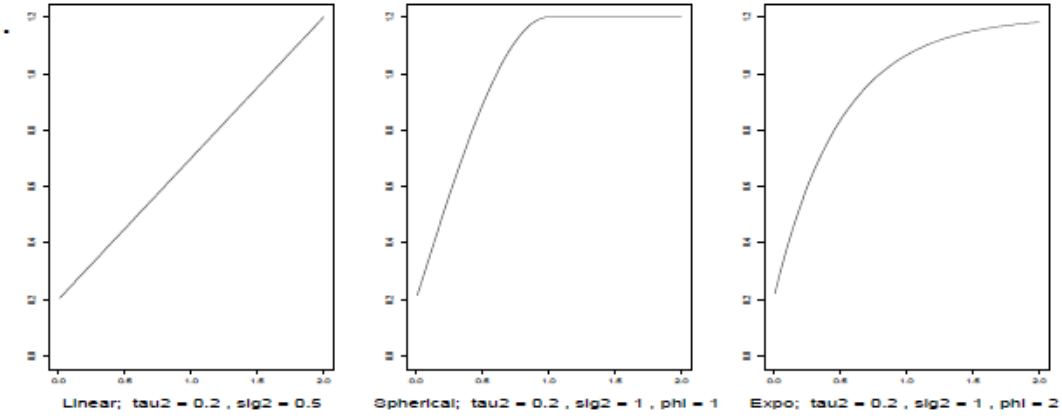
model	Variogram, $\gamma(t)$
Linear	$\gamma(t) = \begin{cases} \tau^2 + \sigma^2 t & \text{if } t > 0 \\ 0 & \text{if } t = 0 \end{cases}$
Spherical	$\gamma(t) = \begin{cases} \tau^2 + \sigma^2 & \text{if } t \geq 1/\phi \\ \tau^2 + \sigma^2 \left[\frac{3}{2}\phi t - \frac{1}{2}(\phi t)^3\right] & \text{if } 0 < t \leq 1/\phi \\ 0 & \text{if } t = 0 \end{cases}$
Exponential	$\gamma(t) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-\phi t)) & \text{if } t > 0 \\ 0 & \text{if } t = 0 \end{cases}$
Powered exponential	$\gamma(t) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-\lvert\phi t\rvert^p)) & \text{if } t > 0 \\ 0 & \text{if } t = 0 \end{cases}$
Matérn at $\nu = 3/2$	$\gamma(t) = \begin{cases} \tau^2 + \sigma^2 \left[1 - (1 + \phi t)\, e^{-\phi t}\right] & \text{if } t > 0 \\ 0 & \text{if } t = 0 \end{cases}$

Example: Spherical semivariogram

$$\gamma(t) = \begin{cases} \tau^2 + \sigma^2 & \text{if } t \geq 1/\phi \\ \tau^2 + \sigma^2 \left[\frac{3}{2}\phi t - \frac{1}{2}(\phi t)^3\right] & \text{if } 0 < t \leq 1/\phi \\ 0 & \text{otherwise} \end{cases}$$

- 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, σ^2 in this case, is called the *partial sill*.)
- The value $t = 1/\phi$ at which $\gamma(t)$ first reaches its ultimate level (the sill) is called the *range*, here $R \equiv 1/\phi$. (Both R and ϕ are sometimes referred to as the "range," but ϕ should be called the *decay* parameter.)

3 common semivariogram models



- For the linear model (left panel), $\gamma(t) \to \infty$ as $t \to \infty$, not to a constant (which would be $C(\mathbf{0})$). So, this semivariogram does not correspond to a weakly stationary process but it is intrinsically stationary.
- The nugget is τ^2, but the sill and range are both infinite.

The exponential model

- The sill is only reached asymptotically, meaning that strictly speaking, the range is infinite.
- To define an "effective range", for $t > 0$, we see that as $t \to \infty$, $\gamma(t) \to \tau^2 + \sigma^2$ which would become $C(0)$.
- Again,

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

- Then the correlation between two points distance t apart is $\exp(-\phi t)$;
- We define the *effective range*, t_0, as the distance at which this correlation $= 0.05$. Setting $\exp(-\phi t_0)$ equal to this value we obtain $t_0 \approx 3/\phi$, since $\log(0.05) \approx -3$.

cont.

- We introduce an *intentional* discontinuity at 0 for both the covariance function and the variogram.
- To clarify why, suppose we write the error at $\mathbf{s}$ in our spatial model as $w(\mathbf{s}) + \epsilon(\mathbf{s})$ where $w(\mathbf{s})$ is a mean 0 process with covariance function $\sigma^2\rho(t)$ and $\epsilon(\mathbf{s})$ is so-called "white noise", i.e., the $\epsilon(\mathbf{s})$ are i.i.d. $N(0, \tau^2)$
- Then, we can compute $\text{var}(w(\mathbf{s}) + \epsilon(\mathbf{s})) = \sigma^2 + \tau^2$
- And, we can compute $\text{Cov}(w(\mathbf{s}) + \epsilon(\mathbf{s}), w(\mathbf{s} + \mathbf{h}) + \epsilon(\mathbf{s} + \mathbf{h})) = \sigma^2\rho(||\mathbf{h}||)$
- So, the form of $C(t)$ shows why the nugget τ^2 is often viewed as a "nonspatial effect variance," and the partial sill (σ^2) is viewed as a "spatial effect variance."

The Matérn Correlation Function

- The Matèrn is a very versatile family:

$$C(t) = \begin{cases} \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(2\sqrt{\nu}t\phi)^{\nu}K_{\nu}(2\sqrt{(}\nu)t\phi) & \text{if } t > 0 \\ \tau^2 + \sigma^2 & \text{if } t = 0 \end{cases}$$

  K_ν is the modified Bessel function of order ν (computationally tractable in C/C++ or geoR)
- ν is a smoothness parameter:
  - $\nu = 1/2 \Rightarrow$ exponential; $\nu \to \infty \Rightarrow$ Gaussian; $\nu = 3/2 \Rightarrow$ convenient closed form for $C(t), \gamma(t)$
  - in two-dimensions, the greatest integer in ν indicates the number of times process realizations will be mean-square differentiable.

A bit more on covariance functions

- To be a valid covariance function the function must be positive definite
- Whether a function is positive definite or not can depend upon dimension
- c is a valid covariance functions if and only if it is the characteristic function of a symmetric about 0 random variable (Bochner's Theorem), i.e., $c(\mathbf{h}) = \int \cos(\mathbf{w}^T\mathbf{h})G(d\mathbf{w})$
- Fourier transform, spectral distribution, spectral density
- In principle, the inversion formula could be used to *check* if $c(\mathbf{h})$ is valid

Constructing valid covariance functions

Construct valid covariance functions by using properties of characteristic functions

- multiply valid covariance functions (corresponds to summing independent random variables)
- mixing covariance functions (corresponds to mixing distributions)
- convolving covariance functions (if c_1 and c_2 are valid then $c_{12}(\mathbf{s}) = \int c_1(\mathbf{s} - \mathbf{u}) c_2(\mathbf{u}) d\mathbf{u}$ is valid).
- There are conditions for valid variograms but difficult and not of interest for us.

Variogram model fitting

How does one choose a good parametric variogram model?

- First, one plots the *empirical semivariogram*,

$$\widehat{\gamma}(t) = \frac{1}{2N(t)} \sum_{(\mathbf{s}_i, \mathbf{s}_j) \in N(t)} [Y(\mathbf{s}_i) - Y(\mathbf{s}_j)]^2 \ ,$$

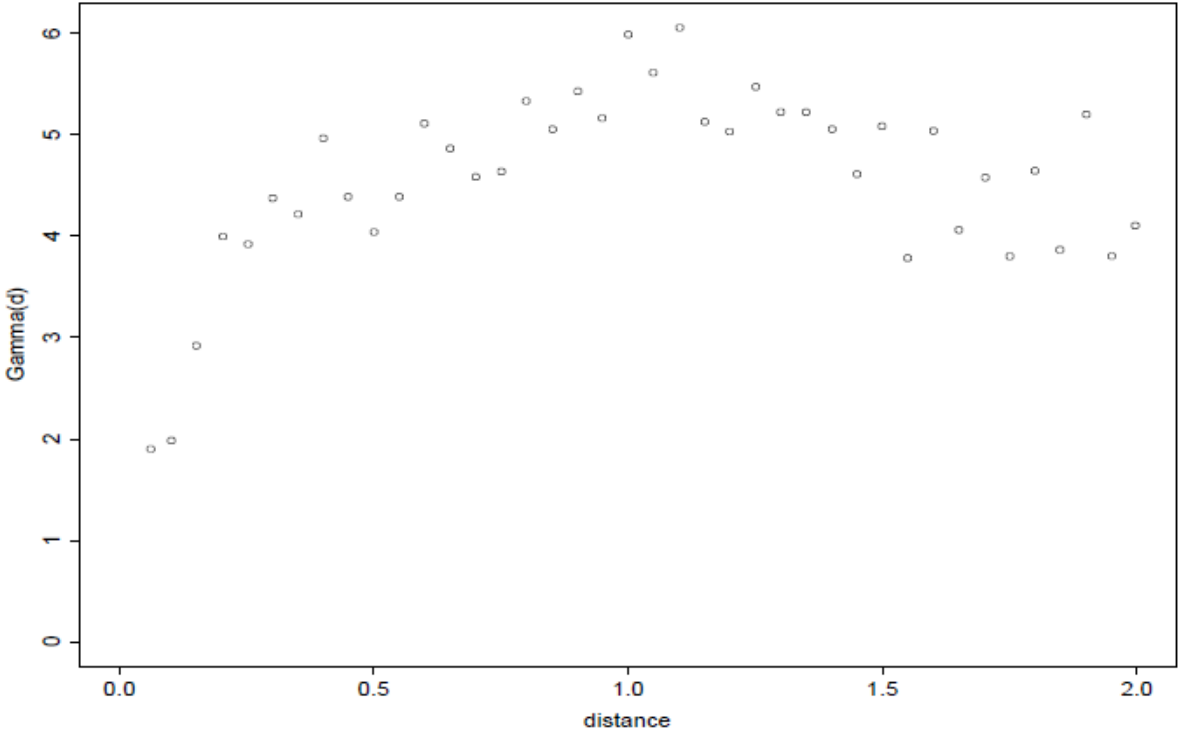
  where $N(t)$ is the set of pairs such that $||\mathbf{s}_i - \mathbf{s}_j|| = t$, and $|N(t)|$ is the number of pairs in this set.

- Usually need to "grid up" the t-space into bins $I_1 = (0, t_1), \ldots, I_K = (t_{K-1}, t_K)$ for $0 < t_1 < \cdots < t_K$. Represent each interval by its midpoint, and redefine

$$N(t_k) = \{(\mathbf{s}_i, \mathbf{s}_j) : ||\mathbf{s}_i - \mathbf{s}_j|| \in I_k\} \ , \ k = 1, \ldots, K \ .$$

- $\hat{\gamma}(t)$ will not be *valid*

Empirical variogram: scallops data



Variogram model fitting (cont'd)

- This method of moments estimator (analogue of the usual sample variance s^2) has problems:
- It will be sensitive to outliers
- A sample average of squared differences can be badly behaved.
- It uses data differences, rather than the data itself.
- The components of the sum will be dependent within and across bins, and $N(t_k)$ will vary across bins.
- Informally, one plots $\widehat{\gamma}(t)$, and then an appropriately shaped theoretical variogram is fit by eye or by trial and error to choose the nugget, sill, and range.
- Formal fitting using least squares, weighted least squares or generalized least squares

Anisotropy

- Isotropy implies circular contours in terms of decay in spatial dependence, i.e., association doesn't depend upon direction
- Stationarity is more general in that it allows association to depend upon the separation vector between locations (i.e., direction and distance).
- As special case is geometric anisotropy, where

$$c(\mathbf{s}-\mathbf{s}') = \sigma^2 \rho((\mathbf{s}-\mathbf{s}')^T B(\mathbf{s}-\mathbf{s}')) \; .$$

- B is positive definite with ρ a valid correlation function.
- Since the equation $(\mathbf{s}-\mathbf{s}')^T B(\mathbf{s}-\mathbf{s}') = k$ is an ellipse in 2-dim space, spatial dependence is constant on ellipses. This means dependence depends on direction. In particular, the contour corresponding to $\rho = .05$ provides the effective range in each spatial direction.

Anisotropy (cont'd)

- Both geometric anisotropy and product geometric anisotropy are special cases of range anisotropy (Zimmerman, 1993)
- Suggests we might also define sill anisotropy: Given a variogram $\gamma(\mathbf{h})$, what is the behavior of $\gamma(c\mathbf{h}/\left\|\mathbf{h}\right\|)$ as $c \to \infty$? Does it depend upon $\mathbf{h}$
- nugget anisotropy: Given a variogram $\gamma(\mathbf{h})$, what is the behavior of $\gamma(c\mathbf{h}/\left\|\mathbf{h}\right\|)$ as $c \to 0$? Does it depend upon $\mathbf{h}$.

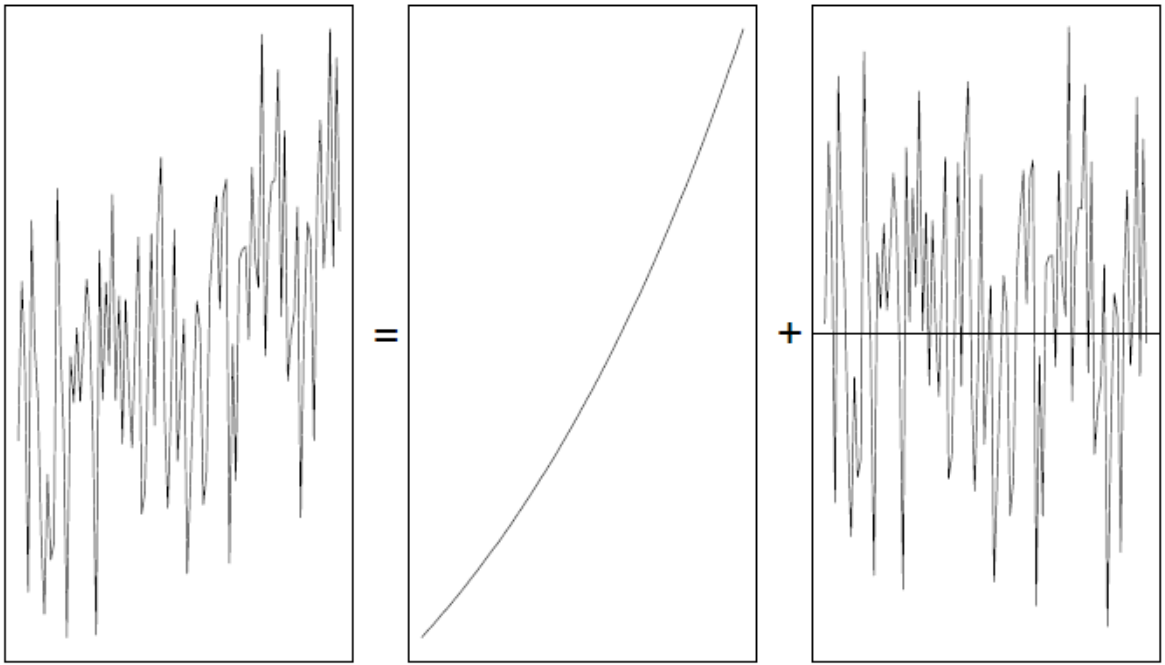
Exploration of Spatial Data

- First step in analyzing data
- First Law of Geostatistics: Mean + Error
- Mean: first-order behavior
- Error: second-order behavior (covariance function)
- Wide variety of EDA tools to examine both first and second order behavior (Cressie's book)
- Crucial point: the spatial structure you might see in the $Y(\mathbf{s})$ surface need not look anything like the spatial structure in the residual surface, after you have fit an explanatory model for the mean, say $\mu(\mathbf{s})$.

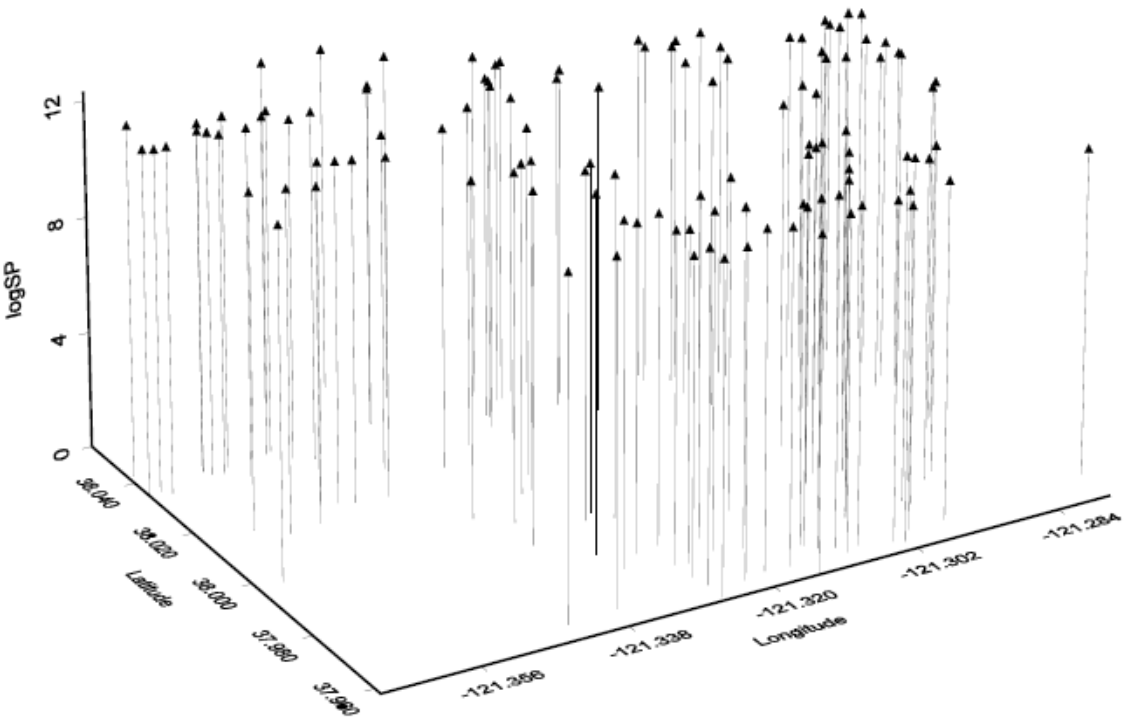
$$E((Y(\mathbf{s}) - \mu(\mathbf{s}))(Y(\mathbf{s}') - \mu(\mathbf{s}')) = E((Y(\mathbf{s}) - \mu)(Y(\mathbf{s}') - \mu)$$

$$+(\mu - \mu(\mathbf{s}))(\mu - \mu(\mathbf{s}'))$$

First Law of Geostatistics



Drop-line scatter plot



Surface plot

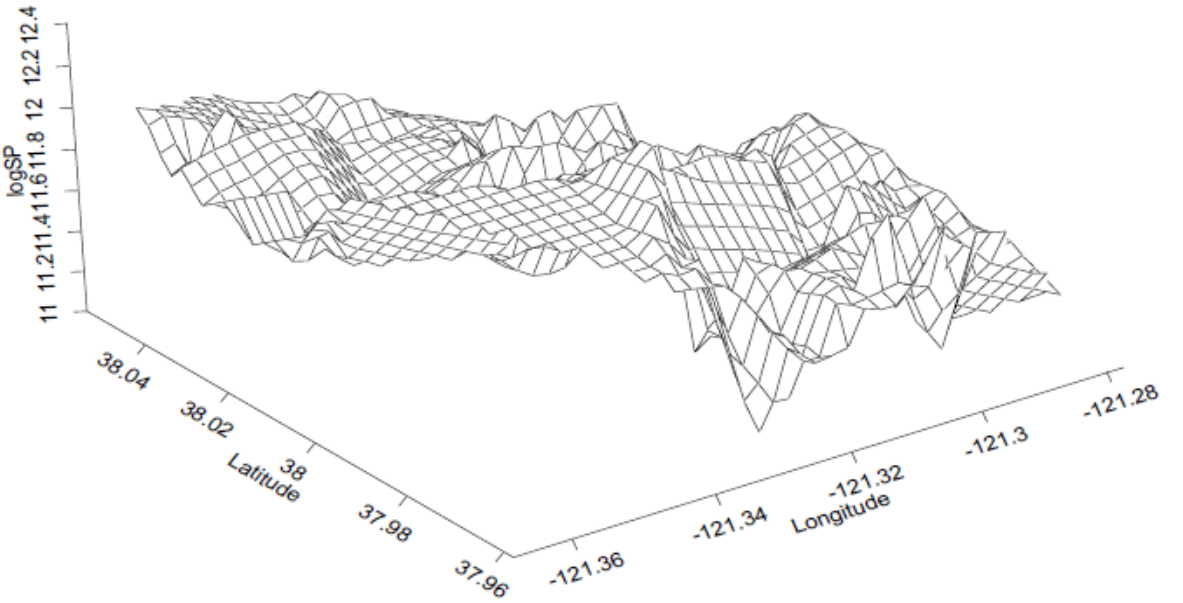
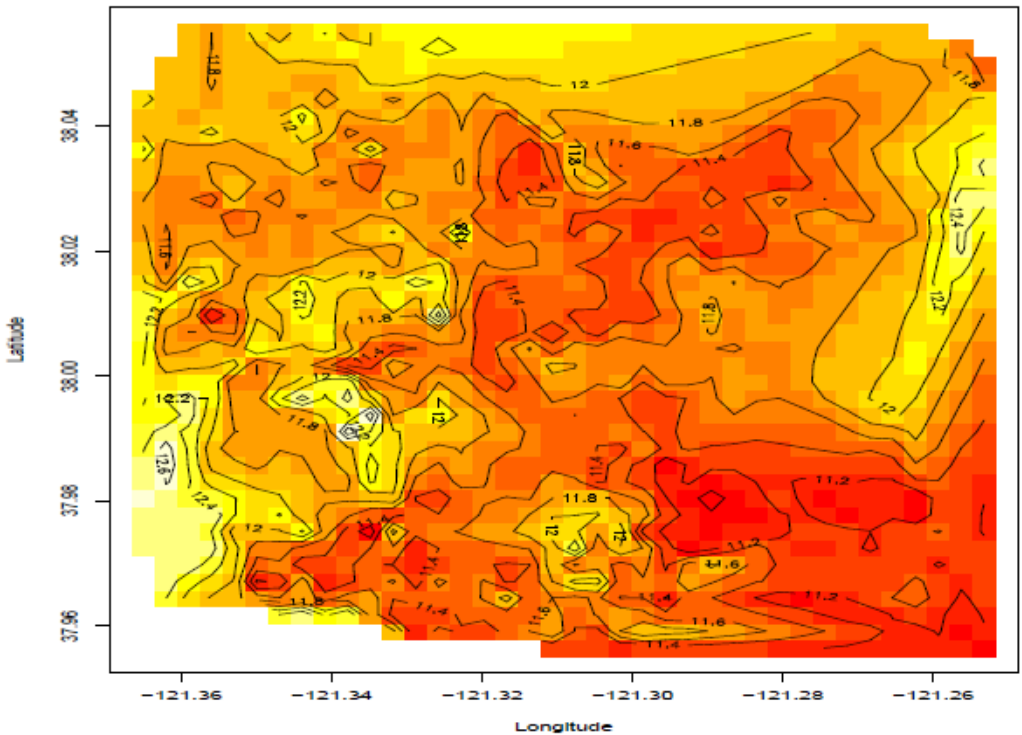


Image-contour plot



Classical spatial prediction (Kriging)

- Named in honor of D.G. Krige, a South African mining engineer whose seminal work on empirical methods for geostatistical data inspired the general approach
- Optimal spatial prediction: given observations of a random field $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))'$, predict the variable Y at a site $\mathbf{s}_0$ where it has not been observed
- Under squared error loss, the best linear prediction minimizes $E[Y(\mathbf{s}_0) - (\sum \ell_i Y(\mathbf{s}_i) + \delta_0)]^2$ over δ_0 and ℓ_i.
- Under intrinsic stationarity, adopting unbiasedness, δ_0 drops out. Obviously, $\bar{Y}$ is not best.
- With an estimate of γ, one immediately obtains the ordinary kriging estimate.
- No distributional assumptions are required for the $Y(\mathbf{s}_i)$.

Difficulties

- Limitation of a constant mean - so introduce a mean surface and then *universal* kriging
- mean surface unknown
- variogram unknown
- if we put estimates of both into the kriging equations we fail to take into account the uncertainty in these estimates
- so, we turn to Gaussian process, work with the covariance function and now have a likelihood
- we redo prediction in this setting

Kriging with Gaussian processes

- Given covariate values $\mathbf{x}(\mathbf{s}_i)$, $i = 0, 1, \ldots, n$, suppose

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \text{ where } \boldsymbol{\epsilon} \sim N(\mathbf{0}, \Sigma) \ .$$

- For a spatial covariance structure having no nugget effect, we specify Σ as

$$\Sigma = \sigma^2 H(\phi) \text{ where } (H(\phi))_{ij} = \rho(\phi; d_{ij}) \ ,$$

with $d_{ij} = ||\mathbf{s}_i - \mathbf{s}_j||$, the distance between $\mathbf{s}_i$ and $\mathbf{s}_j$, and ρ is a valid correlation function.

- For a model having a nugget effect, we instead set

$$\Sigma = \sigma^2 H(\phi) + \tau^2 I \ ,$$

where τ^2 is the nugget effect variance.

Kriging with Gaussian processes

- We seek the function $g(\mathbf{y})$ that minimizes the mean-squared prediction error, $E\left[\left(Y(\mathbf{s}_0) - g(\mathbf{y})\right)^2 \,\middle|\, \mathbf{y}\right]$, i.e., we work with the conditional distribution of $Y(\mathbf{s}_0)|\mathbf{y}$
- It is well known that the (posterior) mean minimizes expected squared error loss.
- So, it must be that the predictor $g(\mathbf{y})$ that minimizes the error is the conditional expectation, $E(Y(\mathbf{s}_0)|\mathbf{y})$.

Kriging with Gaussian processes

- Now consider estimation of this best predictor, first in the completely unrealistic situation in which all the population parameters ($\boldsymbol{\beta}, \sigma^2, \phi$, and τ^2) are known. Suppose

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix} \right),$$

  where $\Omega_{21} = \Omega_{12}^T$.
- Then $f(\mathbf{Y}_1|\mathbf{Y}_2)$ is normal with mean and variance

$$\begin{aligned} E[\mathbf{Y}_1|\mathbf{Y}_2] &= \boldsymbol{\mu}_1 + \Omega_{12}\Omega_{22}^{-1}(\mathbf{Y}_2 - \boldsymbol{\mu}_2) \\ \text{and } Var[\mathbf{Y}_1|\mathbf{Y}_2] &= \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21} \,. \end{aligned}$$

Kriging with Gaussian processes

- In our framework, $\mathbf{Y}_1 = Y(\mathbf{s}_0)$ and $\mathbf{Y}_2 = \mathbf{y}$, meaning that $\Omega_{11} = \sigma^2 + \tau^2$, $\Omega_{12} = \boldsymbol{\gamma}^T$, and $\Omega_{22} = \Sigma = \sigma^2 H(\phi) + \tau^2 I$, where $\boldsymbol{\gamma}^T = \left(\sigma^2 \rho(\phi; d_{01}), \ldots, \sigma^2 \rho(\phi; d_{0n})\right)$.
- Substituting these values into the mean and variance formulae above, we obtain

$$\begin{aligned} E[Y(\mathbf{s}_0)|\mathbf{y}] &= \mathbf{x}_0^T \boldsymbol{\beta} + \boldsymbol{\gamma}^T \Sigma^{-1} (\mathbf{y} - X\boldsymbol{\beta}) \ , \\ \text{and} \, Var[Y(\mathbf{s}_0)|\mathbf{y}] &= \sigma^2 + \tau^2 - \boldsymbol{\gamma}^T \Sigma^{-1} \boldsymbol{\gamma} \ . \end{aligned}$$

- Pretty forms but useless since we don't know any of the model parameters

Kriging with Gaussian processes

- So, consider how these answers are modified in the more realistic scenario where the model parameters are unknown. We modify $g(\mathbf{y})$ to

$$\hat{g}\left(\mathbf{y}\right) = \mathbf{x}_0^T \widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\gamma}}^T \widehat{\Sigma}^{-1} \left(\mathbf{y} - X\widehat{\boldsymbol{\beta}}\right) \ ,$$

where $\widehat{\boldsymbol{\gamma}} = \left(\hat{\sigma}^2 \rho(\hat{\phi}; d_{01}), \ldots, \hat{\sigma}^2 \rho(\hat{\phi}; d_{0n})\right)^T$, $\widehat{\Sigma} = \hat{\sigma}^2 H(\hat{\phi})$, and $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{WLS} = \left(X^T \widehat{\Sigma}^{-1} X\right)^{-1} X^T \widehat{\Sigma}^{-1} \mathbf{y}$.

- Thus $\hat{g}\left(\mathbf{y}\right)$ can be written as $\boldsymbol{\lambda}^T \mathbf{y}$, where

$$\boldsymbol{\lambda} = \widehat{\Sigma}^{-1} \widehat{\boldsymbol{\gamma}} + \widehat{\Sigma}^{-1} X \left(X^T \widehat{\Sigma}^{-1} X\right)^{-1} \left(\mathbf{x}_0 - X^T \widehat{\Sigma}^{-1} \widehat{\boldsymbol{\gamma}}\right) \ .$$

Kriging with Gaussian processes

- If $X(\mathbf{s}_0)$ is unobserved, we can still do the spatial prediction
- We estimate $X(\mathbf{s}_0)$ and $Y(\mathbf{s}_0)$ jointly by iterating between the formula for $\hat{g}(\mathbf{y})$ and a corresponding one for $\hat{\mathbf{x}}_0$, namely

$$\hat{\mathbf{x}}_0 = X^T \boldsymbol{\lambda} \, ,$$

  which arises simply by multiplying both sides of the previous equation by X^T and simplifying.
- This is essentially an EM (expectation-maximization) algorithm, with the calculation of $\hat{\mathbf{x}}_0$ being the E step and the updating of λ being the M step.
- In the classical framework, restricted maximum likelihood (REML) estimates are often plugged in above and have some optimal properties.