Lecture 4: Theory of spatial models

April 19, 2015

Why deal with space?

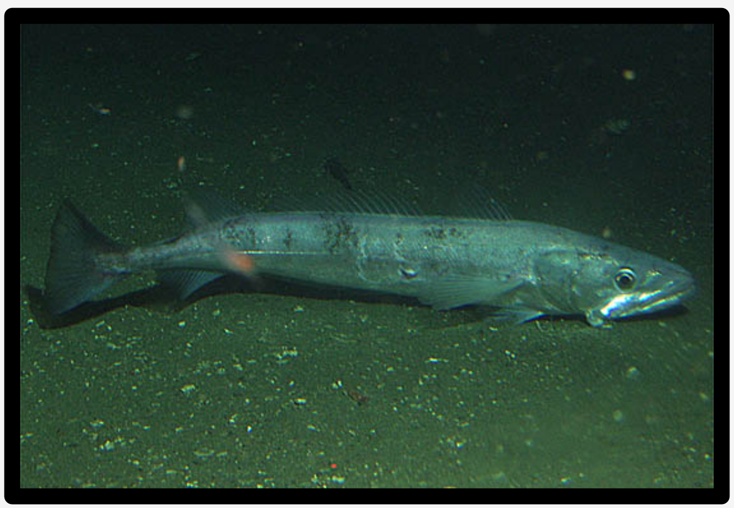
1. All processes have an underlying spatial domain to them.
2. Many of our observations contain spatial data (whether we chose to use it or not)
3. Gain a richer understanding of the dynamics by explicitly incorporating spatial dynamics and spatial autocorrelation

Why should we fit models that incorporate spatial autocorrelation?

- Failure to incorporate the spatial autocorrelation may lead to incorrect inference on the factors that describe the mean portion of the model.
- For example – what if have an animal that likes to group?
  - Some occupied habitat may be due to high quality habitat and some marginal habitat may be occupied to be close to other conspecifics. Failure to take the spatial correlation into account will lead to assuming that the marginal habitat is high quality.
- How might this play out for a territorial species?

Case Study

- Agostini et al. 2008. Climate–ocean variability and Pacific hake: A geostatistical modeling approach. *J of Marine Systems*, 71:237-248.

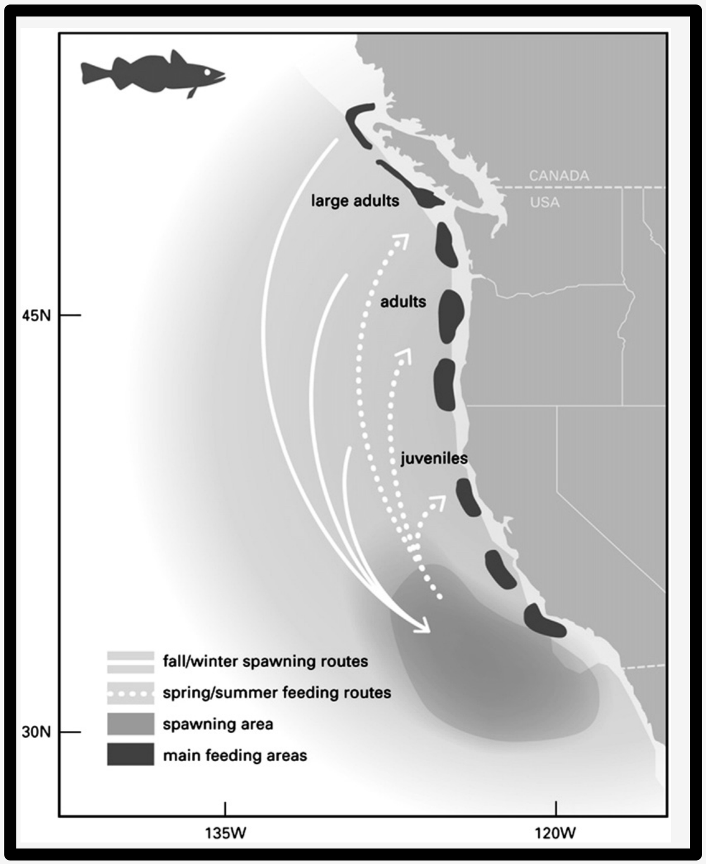


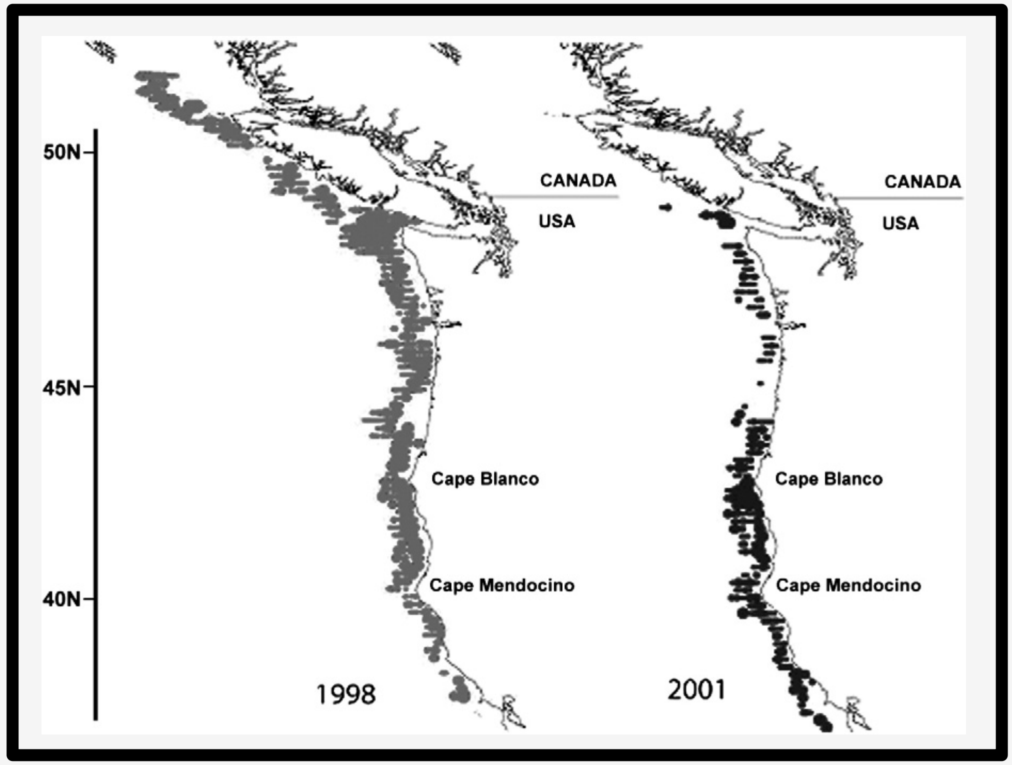
credit: sanc1717

credit: Rodney Johnson

Objective

- Determine environmental factors describing hake habitat while accounting for spatial autocorrelation (hake are a schooling fish so may expect spatial autocorrelation)





Approach

- Fit Generalized Linear Models with spatial autocorrelation
  (notation of Agostini et al. 2008)

$$Y_i = \mu_i + S(\mathbf{x}_i) + \varepsilon_i$$

where

- μ_i is a mean process affecting the role of habitat and is a function of covariates $\boldsymbol{C}_i$ (depth and current) via μ_i=$\boldsymbol{C}_i\boldsymbol{\beta}$
- $\mathbf{x}_\mathrm{i}$ is the observation location
- $S(\mathbf{x}_\mathrm{i})$ is a stationary Gaussian process with expected value $\mathrm{E}[S(\mathbf{x})] = 0$ and $\mathrm{cov}[S(\mathbf{x}_\mathrm{i}),S(\mathbf{x}_\mathrm{j})] = \sigma^2\rho(\mathbf{x}_\mathrm{i} - \mathbf{x}_\mathrm{j})$ where σ^2=variance and ρ = correlation coefficient)
- ε_i are mutually independent with variance = v^2

Modeling spatial autocorrelation

- Used an exponential variogram to model the spatial autocorrelation $S(\mathbf{x})$ (more on this later in the lecture)

$$\gamma(\mathbf{x}-\mathbf{x}') = \tau^2 + \sigma^2\left\{1 - \exp\left[-\frac{(\mathbf{x}-\mathbf{x}')}{\phi}\right]\right\}$$

where τ^2 is the small scale variability at small distances (nugget), ϕ is the distance over which samples are autocorrelated (range), and σ^2 is the background variability (sill)

  - Model $S(\mathbf{x})$ simultaneously with estimating $\boldsymbol{\beta}$ coefficients
  - Use GeoR package in R

Results

Without spatial autocorrelation AIC = 7057

With spatial autocorrelation AIC = 6014

Table 2

Model parameter estimates for the model structure with the lowest AIC value (autocorrelated model) and model parameter estimates for the same model without spatial autocorrelation

Parameter	Mean value	Standard deviation
Model without autocorrelation term: ${}^{nac}\beta_0 + {}^{nac}\beta_1 current^2 + {}^{nac}\beta_2 current + {}^{nac}\beta_3 depth$		
β_0^{nac}	12.68	0.49
β_1^{nac}	−5.73	13.5
β_2^{nac}	2.90	2.92
β_3^{nac}	−0.006	0.0004
Model with autocorrelation term $\beta_0 + \beta_1 current^2 + \beta_2 current + \beta_3 depth + S(\mathbf{x})$		
β_0	8.20	0.992
β_1	20.28	12.04
β_2	−3.75	3.009
β_3	−0.0025	0.00087
τ^2 (nugget)	2.81	
ϕ (range)	3.67 km	
σ^2 (sill)	39.28	

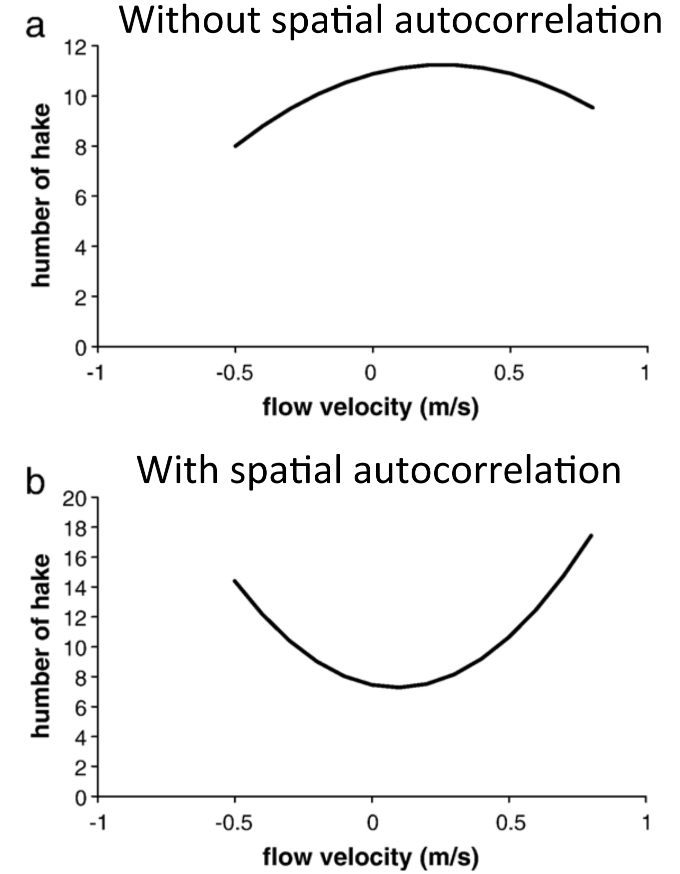


Fig. 5. Relationship between hake abundance and current velocity as predicted by the model without autocorrelation (upper panel) and the model with autocorrelation (lower panel).

Topics

- Geostatistical processes
- Generalized covariance matrix for state process
- Variogram functions
- Stationarity, isotropy, anisotropy
- Test of spatial independence – Moran's I
- Geostatistical modeling

Definitions -

- Spatial series-
  - Z_s is a noisy observation of a process at location s with expected value μ_s

$$Z_s \sim g(\mu_s)$$

  - μ_s is a function of covariates $\mathbf{X}_s$ and random effects $\varepsilon_{s,}$ the random effects may contain spatial autocorrelation and we want to model this appropriately

$$\mu_s = \mathbf{X}_s\mathbf{b} + \varepsilon_s$$

Stationarity

- Strong stationarity: spatial model for C_s is stationary if all statistics remain unchanged after spatial shifts, i.e. remain unchanged for all C_{s+h} for all possible h (all moments are invariant)

- Weak stationarity: spatial model for C_s is stationary if the mean function is constant and auto-covariance is a function of distance (only mean and variance are invariant)

Spatial covariance functions

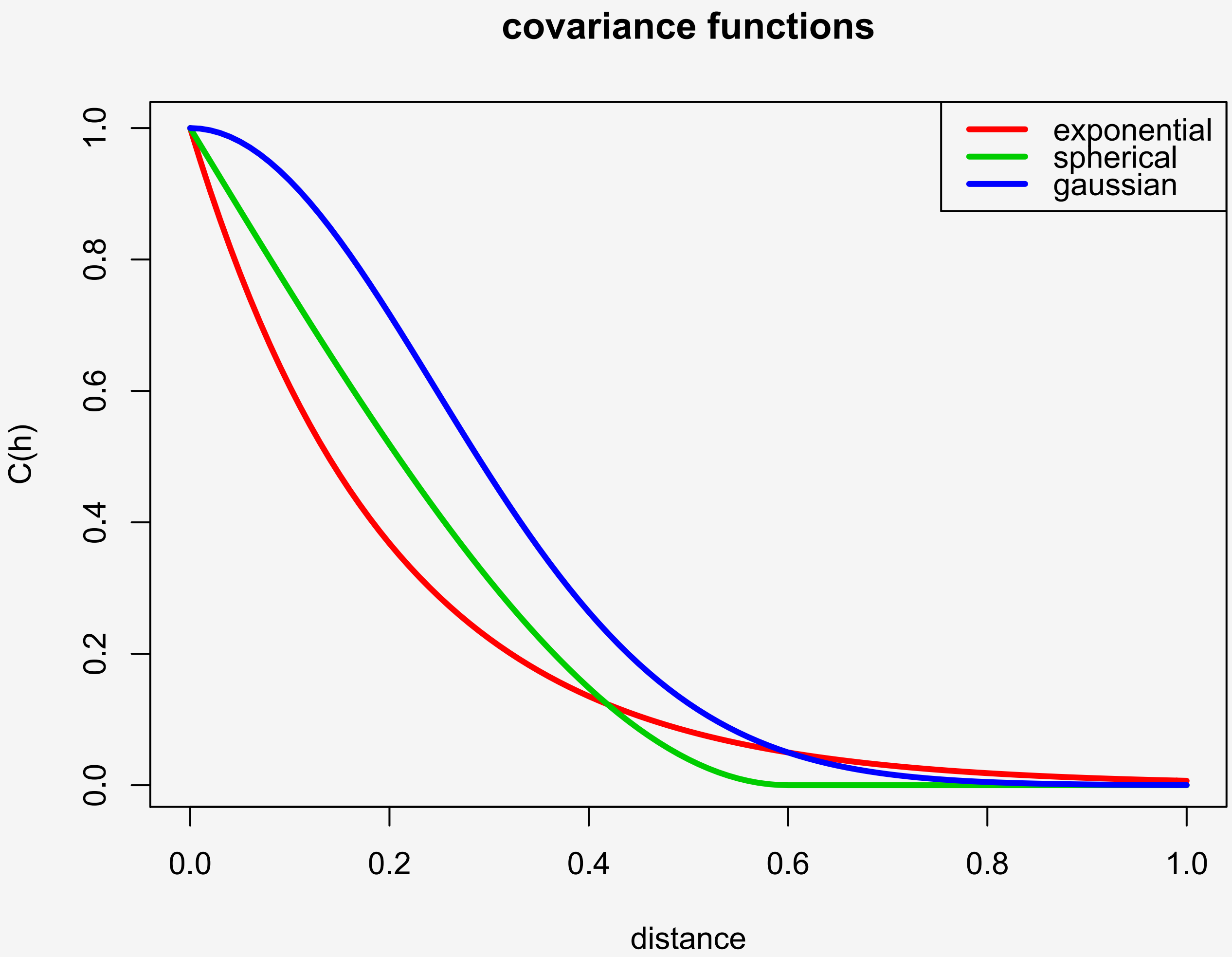
- Distance between sites in 2 (or more) dimensions can be calculated with the Euclidean norm

$$\left\|\mathbf{h}\right\| = \sqrt{h_1^2 + h_2^2 + \cdots + h_n^2} = \sqrt{\mathbf{h}\cdot\mathbf{h}}$$

- Covariance C_Y is defined as a function of the distance $||\mathbf{h}||$ between sites, for example the exponential covariance function, where $I()$ is the indicator function that equals 1 if the condition is TRUE

$$C_Y(\mathbf{h}|\sigma_0^2, \sigma_1^2, \theta_1) = \sigma_0^2 I(||\mathbf{h}|| = 0) + \sigma_1^2 \exp-(||\mathbf{h}||/\theta_1)$$

Examples of covariance functions



- If instead, we define a variance function (the covariogram) that is the variance of the differences in observations at 2 locations, Z_i and Z_j

$$\mathrm{var}(Z_i - Z_j) = \mathrm{var}(Z_i) + \mathrm{var}(Z_j) - 2\mathrm{cov}(Z_i, Z_j)$$

- If the process is second order stationary, then then $\mathrm{E}(\mathbf{Z})$ = 0, and $\mathrm{var}(Z_i)$ = $\mathrm{var}(Z_j)$ = $\mathrm{C}(0)$, where $\mathrm{C}(0)$ is the covariance at distance 0

$$\mathrm{var}(Z_i - Z_j) = 2(C(0) - C(x_i - x_j)$$

$$\mathrm{var}(Z_i - Z_j) = 2(C(0) - C(h_{i,j})$$

Variogram

- The covariogram or variogram is

$$2\gamma(\mathbf{h}) = 2(C(0) - C(\mathbf{h}))$$

- Half this quantity is the semivariogram

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

  - Nugget

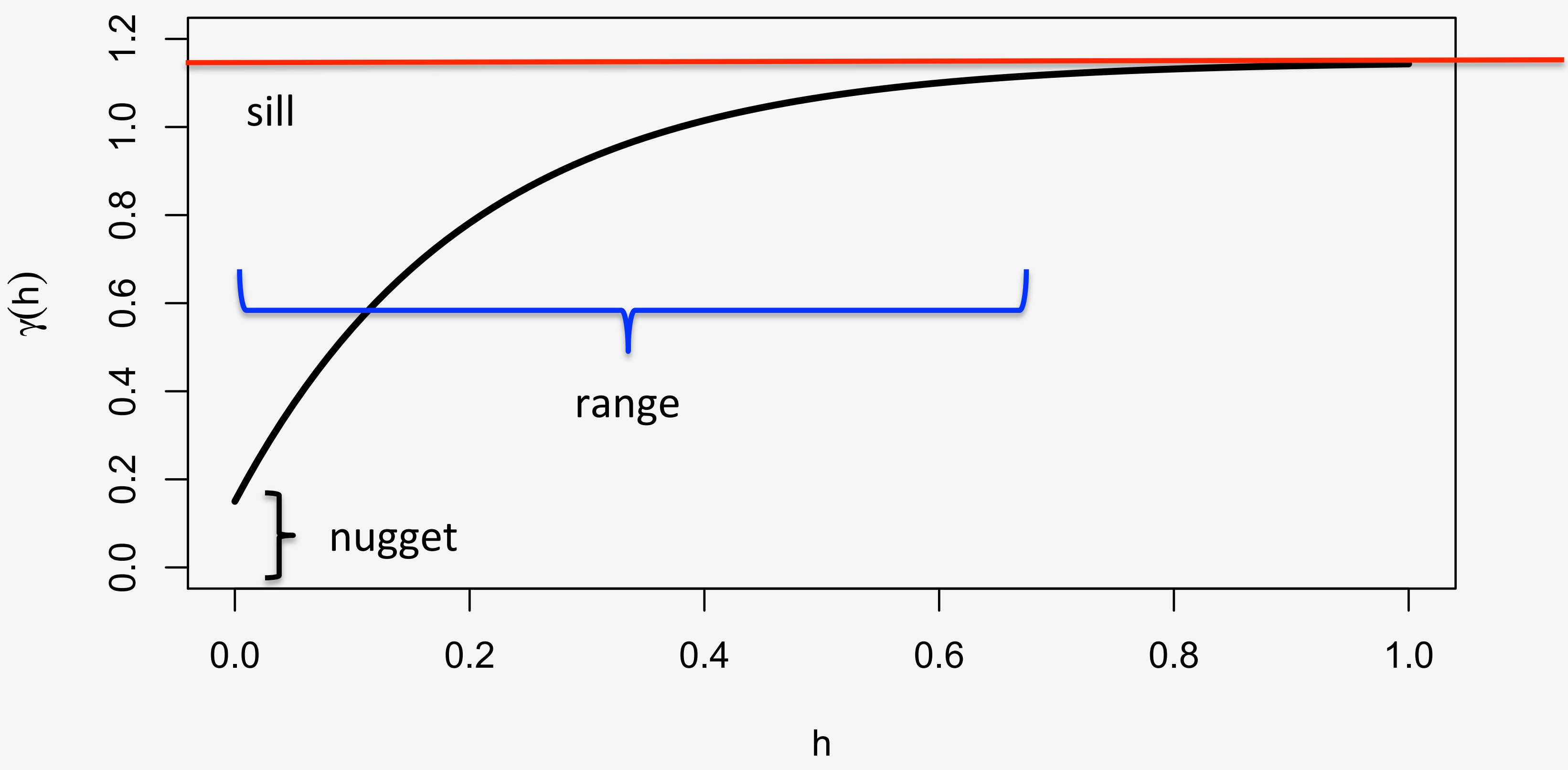
$$\gamma(h) \rightarrow c_0 \text{ as h} \rightarrow 0$$

$$\text{if } c_0 > 0, \text{ then 'nugget effect'}$$

$$\mathrm{c}_0 = c_{MS} + c_{ME}$$

  - Sill is the variance at large h, sill = $C(0)$, partial sill = $C(0)$ – c_0
  - Range = min r where $\gamma(\mathrm{r}(1+\mathrm{e}))$=$C(0)$ = sill

- Variogram components

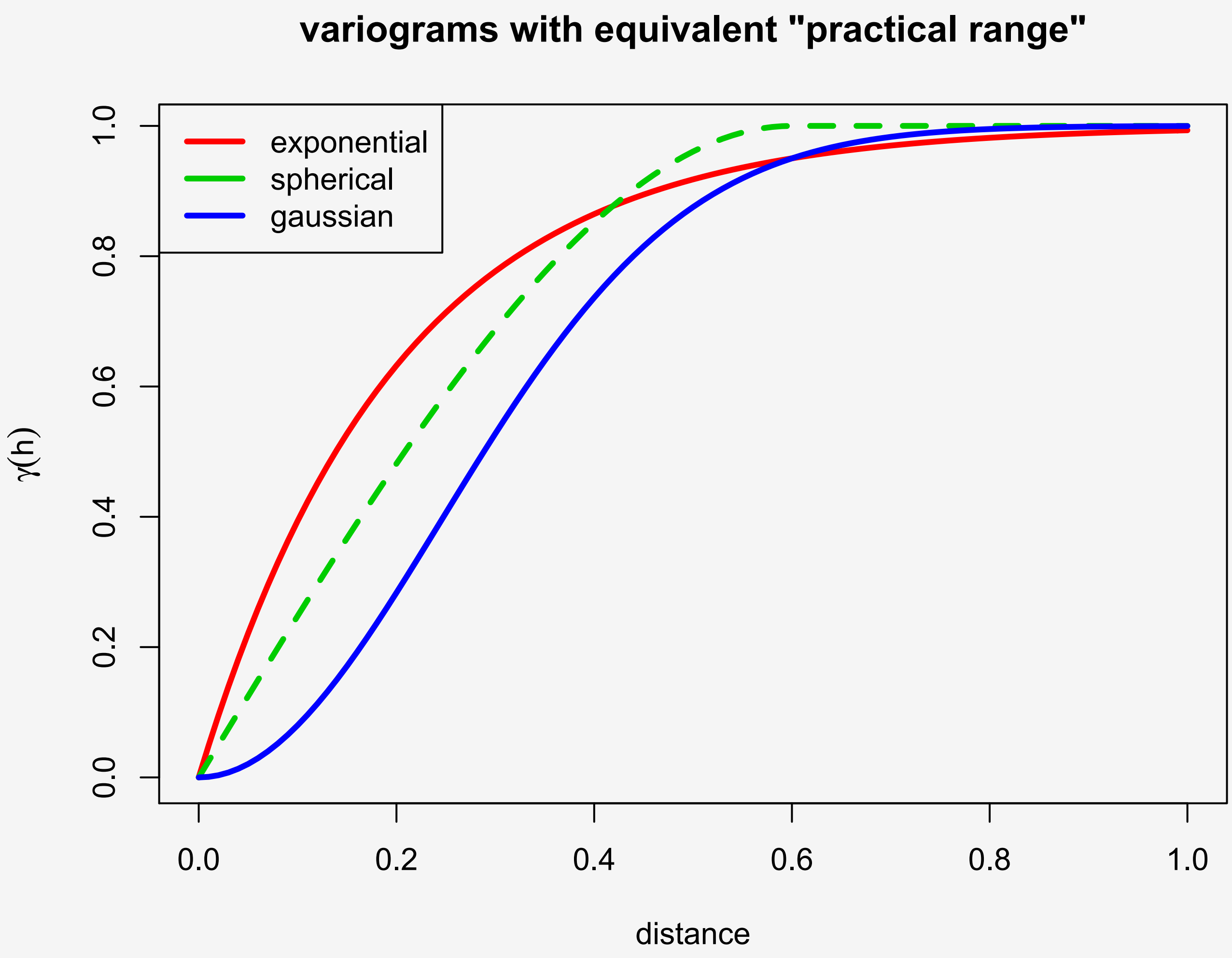


- The functional forms of the semivariogram $\gamma(\mathbf{h})$ are related to the covariance functions, for example the exponential semivariogram is:

$$\gamma_Y(\mathbf{h}) = \sigma_0^2 I(\|\mathbf{h}\| \neq 0) + (\sigma^2{}_1)(1 - \exp(-\|\mathbf{h}\|/\theta_1)$$

- where $I()$ is the indicator function that equals 1 if the condition is TRUE

Examples of variogram functions



Isotropy and Anisotropy

- If the spatial covariance structure is equivalent in all directions, it can be described as a function of distance only and is said to be *isotropic*

- If instead, the covariance structure varies in different directions, then it is a function of the distance and direction and is said to be *anisotropic*

- We must deal with anisotropy either by transforming or by constructing variograms for different directions

Estimating spatial autocorrelation

Two-step process:

1. Remove all trend so that have a mean 0 process
2. Evaluate spatial autocorrelation in residuals with Moran's I
3. Calculate empirical semivariogram
   - Classic estimator
   - Robust estimator
4. Fit theoretical variogram to empirical semivariogram
   - Obtain estimates of nugget, range, and sill

One-step process:

1. Estimate trend and spatial covariance parameters simultaneously

Test of spatial autocorrelation

- Need a test of whether spatial autocorrelation is likely

- Moran's I

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2}$$

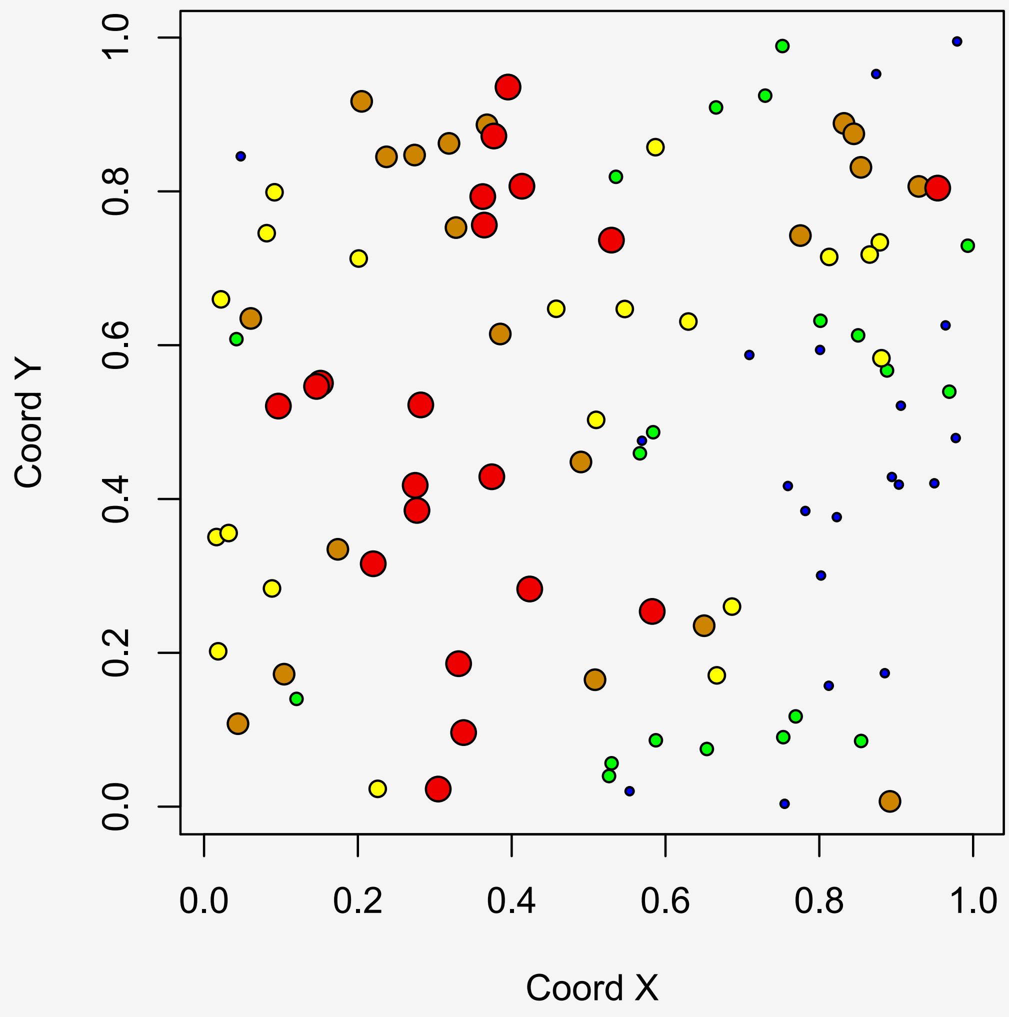
where *N* is the number of spatial units indexed by i and j and X is the variable of interest; and *w* is an element of a matrix of spatial weights.

The expected value of Moran's *I* under the null hypothesis of no spatial autocorrelation is

$$E(I) = \frac{-1}{N-1}$$

- Implemented in R packages **spdep** and **ape**

- Spatial data – we will start with a simulated, 2nd order stationary data series



Empirical variogram

- Classic variogram

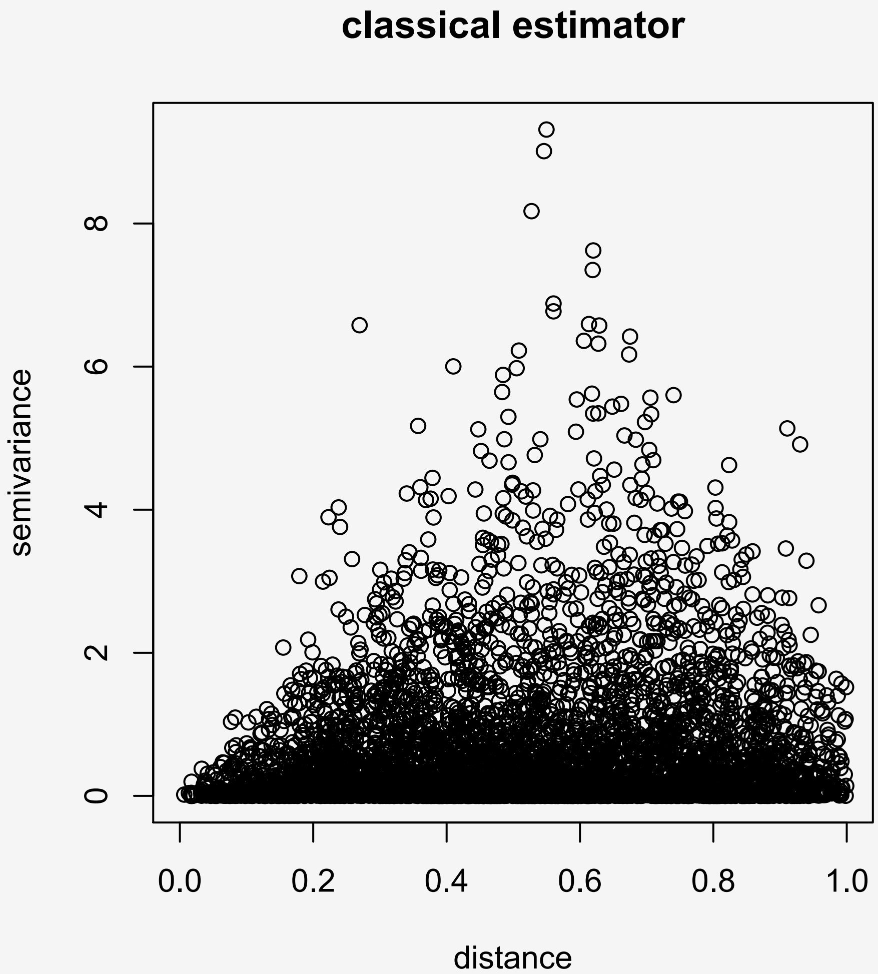
$$\hat{\gamma}(h)=\frac{1}{2}\cdot\frac{1}{n(h)}\sum_{i=1}^{n(h)}(Z(x_i+h)-Z(x_i))^2$$

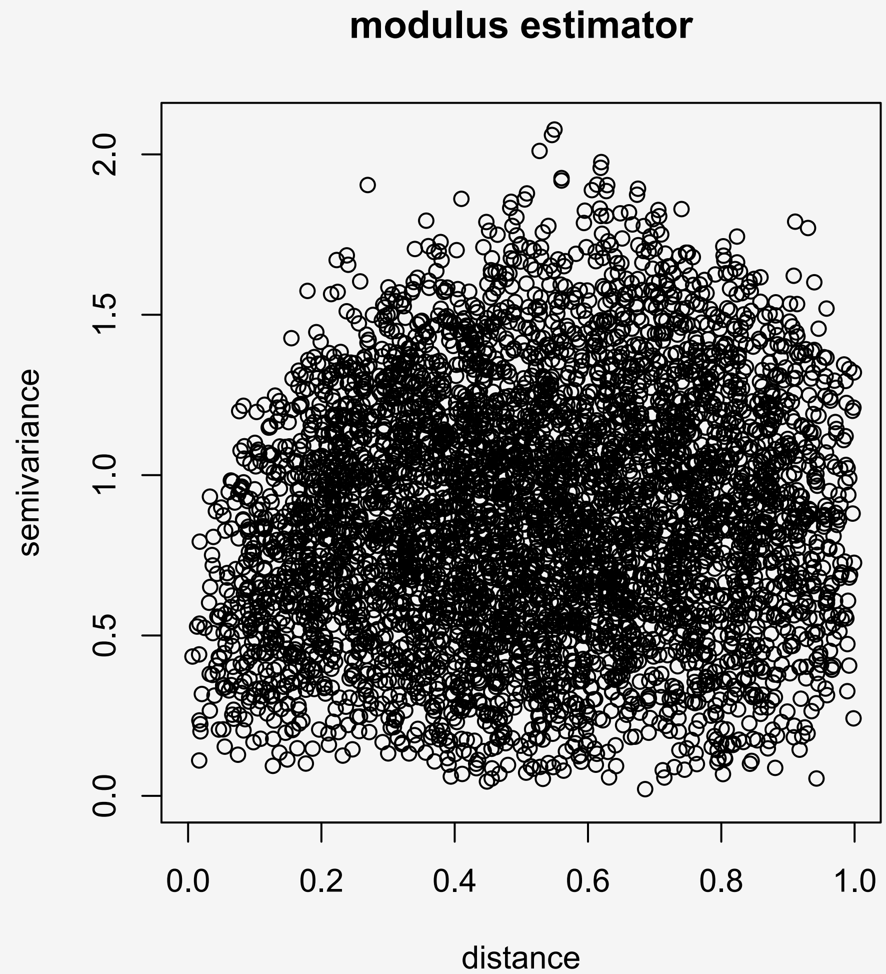
where z is a datum at a particular location, h is the distance between ordered data, and $n(h)$ is the number of paired data at a distance of h

- Modulus or robust variogram (Cressie 1993)

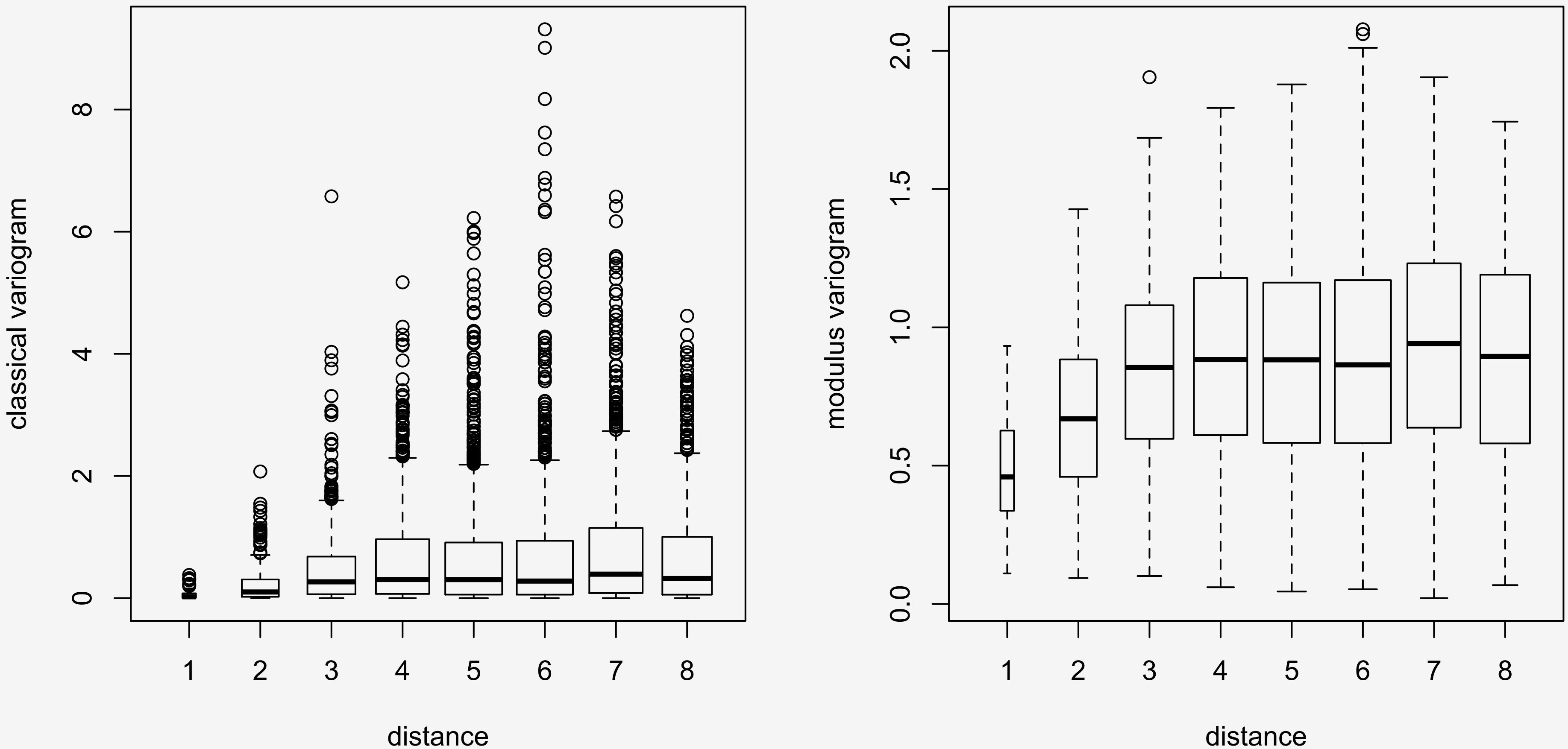
$$\hat{\gamma}(h)=\frac{1}{N(h)}\sum_{i=1}^{N_h}|Z(x_{i+h})-Z(x_i)|^{(1/2)}]^4)/(0.914+(0.988/N_h))$$

Variogram clouds

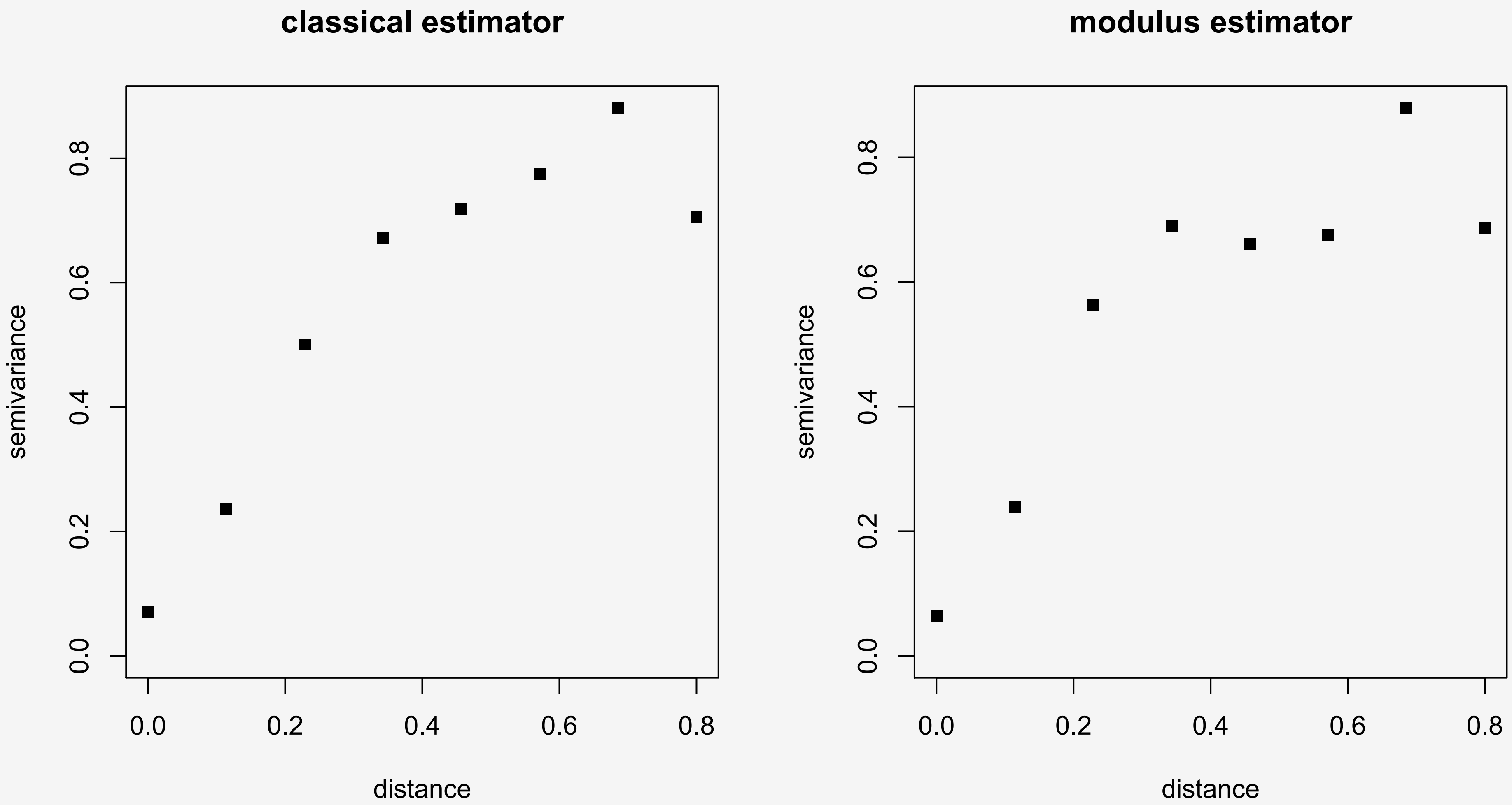




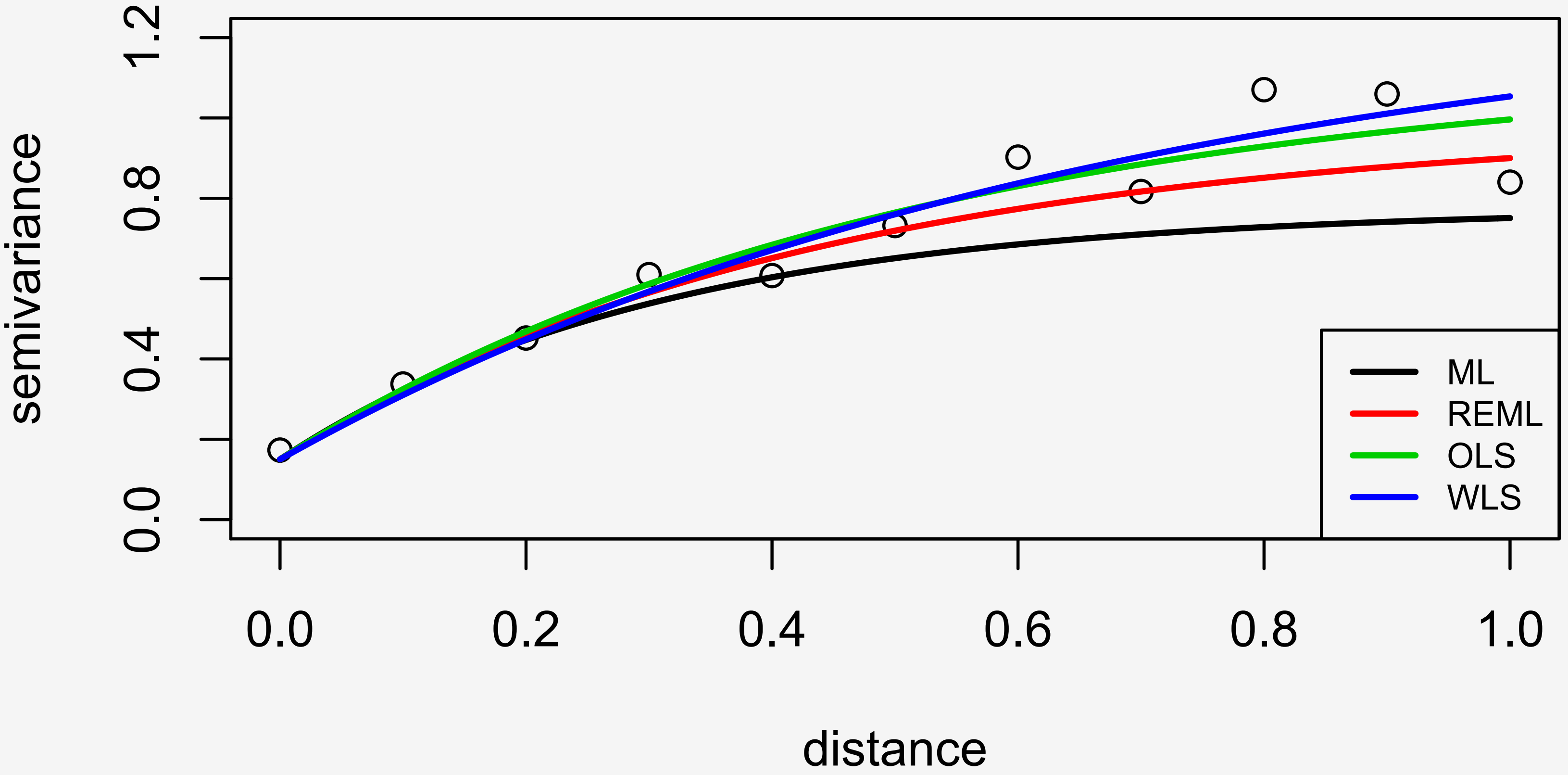
Binned variogram clouds



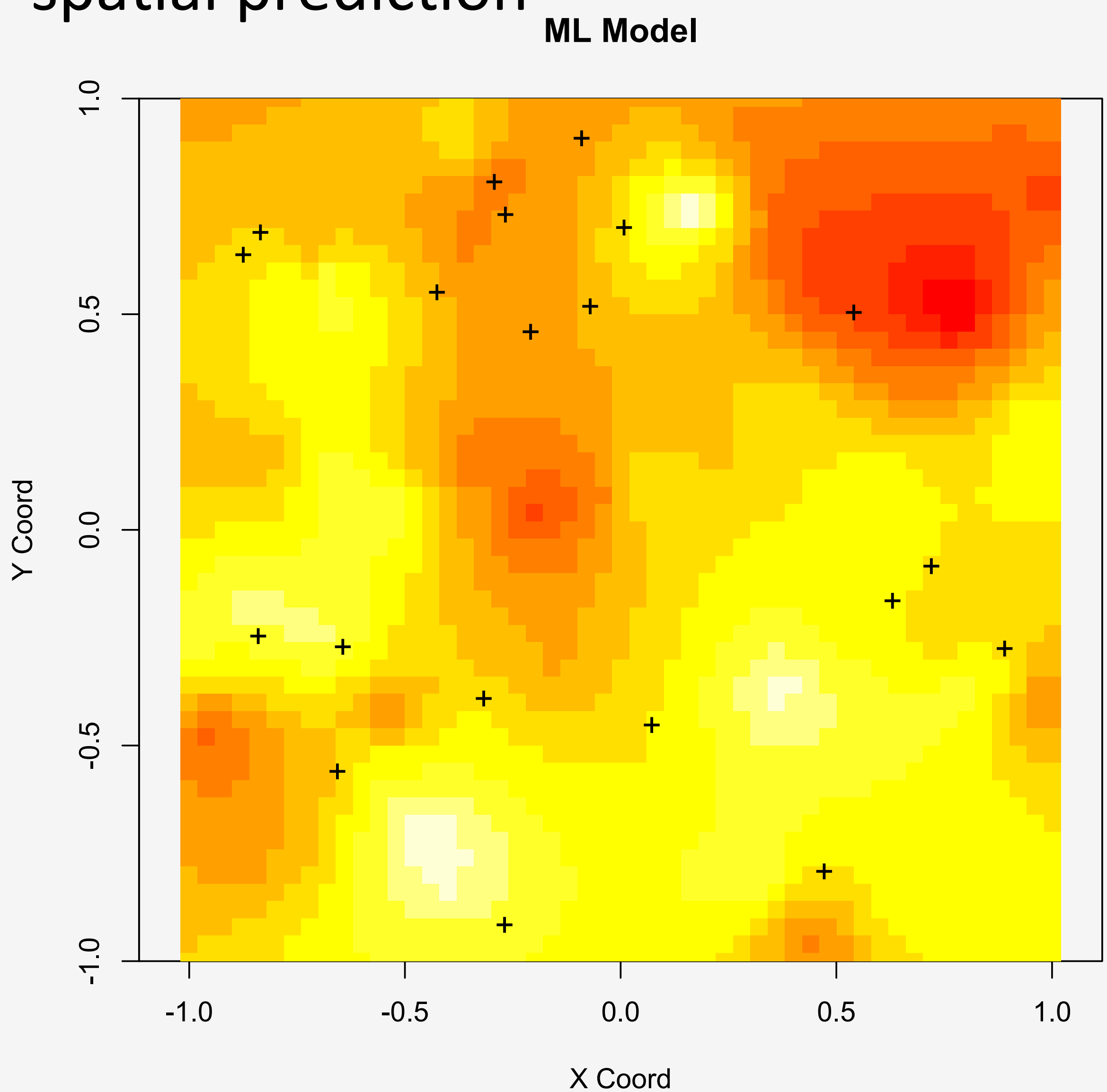
Empirical variogram



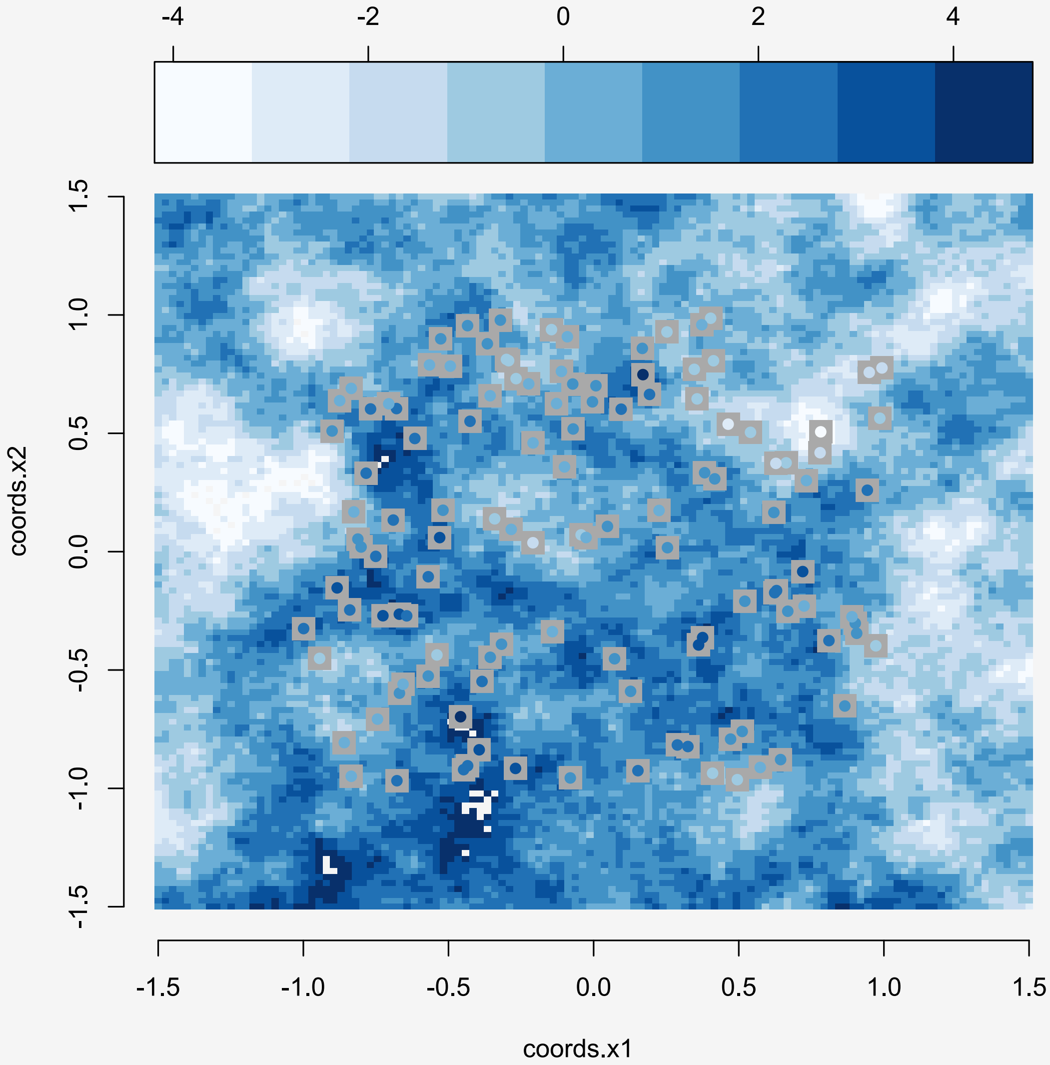
- Fitting theoretical to empirical variogram



- Kriging - spatial prediction



Simulating spatially autocorrelated data



Simulation methods

$$E(\mathbf{Z}) = (\mu(\mathbf{s}_1),\dots,\mu(\mathbf{s}_n)) = \mu$$

$$\mathrm{var}(\mathbf{Z}) = C(\mathbf{s}_i,\mathbf{s}_j) = \Sigma$$

where Σ is n x n positive definite

with i, jth elements $C(s_i,s_j)$

- Choleski Decomposition - decompose Σ using two triangular matrices L and L', use L to create correlated random effects from iid random variables $\mathbf{e}$

$$\Sigma = LL'$$

$$\mathbf{e} \sim N(0,1)$$

$$\mathbf{Z} = \mu + L\mathbf{e}$$

- **RandomFields** - simulation of stationary and isotropic random fields including circulant embedding, turning bands, and decomposition methods

Extensions:

- Anisotropy – variograms in multiple directions
- Ordinary Kriging
- Simple Kriging
- Universal Kriging - fitting a model for the trend and a model for the variogram or covariance for the spatial autocorrelation
- Cokriging – fitting models for multivariate data

Some R packages for geostatistical modeling

- fields
- geoR
- geoRGLM
- gstat
- Rgeostats
- Spatial
- spdep
- RandomFields