Models for Continuous Spatial Processes

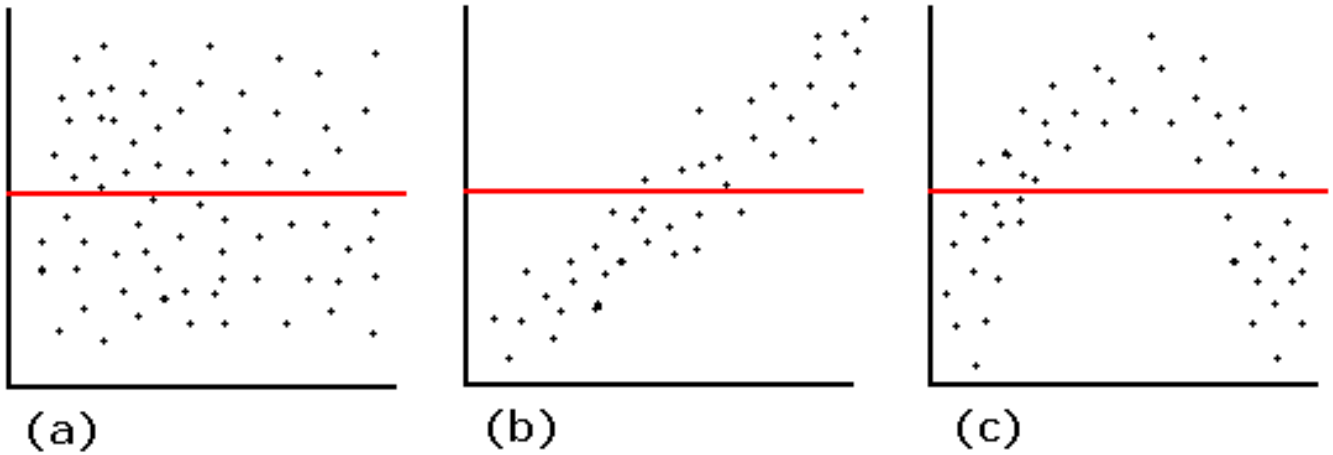
Models for Socio-Environmental Data

Chris Che-Castaldo, Mary B. Collins, N. Thompson Hobbs

August 1, 2019

$$\varepsilon_i = y_i - g(\boldsymbol{\theta}, \mathbf{x_i})$$

ε_i are iid



What goes wrong if we fail to account for autocorrelation?

The global issue is model checking. More specific issues include:

- Inference is excessively optimistic .
- Model selection favors over-parameterized models.
- Your paper will not be published if it goes to a savvy reviewer.

Roadmap: Modeling structure in data

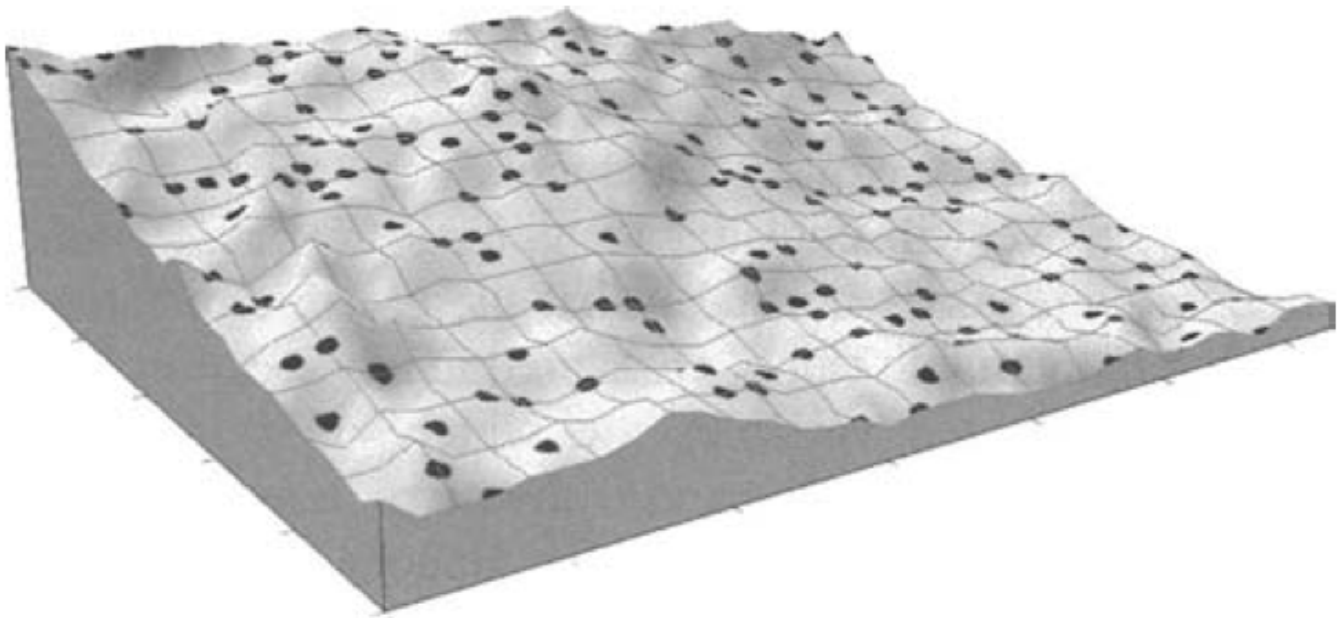
- Continuous spatial processes
  - Detecting spatial dependence
  - Distance matrices
  - Semi-variograms
  - Modeling spatial dependence
- Areal spatial processes (briefly)
  - Detecting spatial dependence
  - Modeling spatial dependence

Learning objectives

- Understand methods for identifying spatial dependence in residuals.
- Appreciate that modeling spatially structured data requires only minor modifications to the models you have developed earlier in the course.
- Write and code a simple model for spatially structured point data.

Most ecological data are spatial

Continuous spatial processes



Data for continuous spatial processes

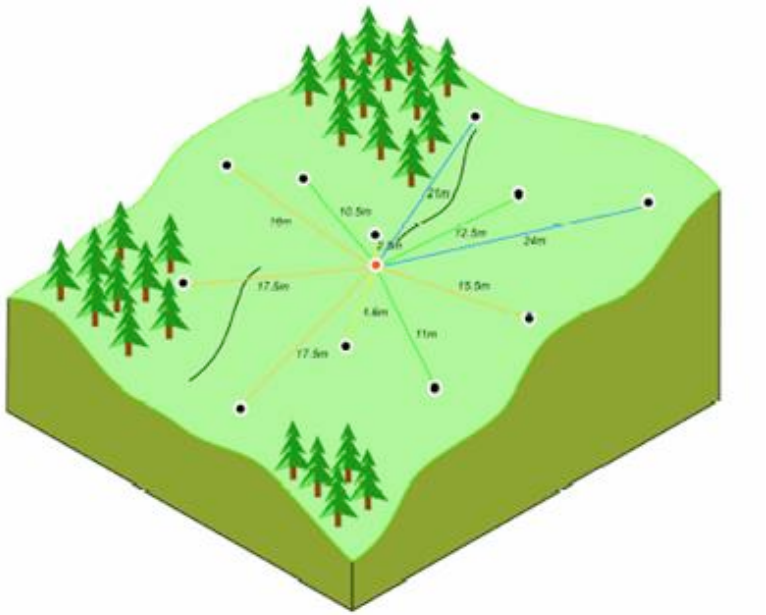
All data points include a spatial reference.

$$\begin{aligned} \text{aspatial data point} &: y_i \\ \text{spatially referenced data point} &: y(\mathbf{s}_i) \end{aligned}$$

Where $\mathbf{s}_i$ is a vector of spatial coordinates of length 1, 2, or 3. The data are said to be continuous because they can occur at any point $(\mathbf{s}_i)$ in one, two, or three dimensional space. This does not mean that the value at that point $(y(s_i))$ can not be discrete.

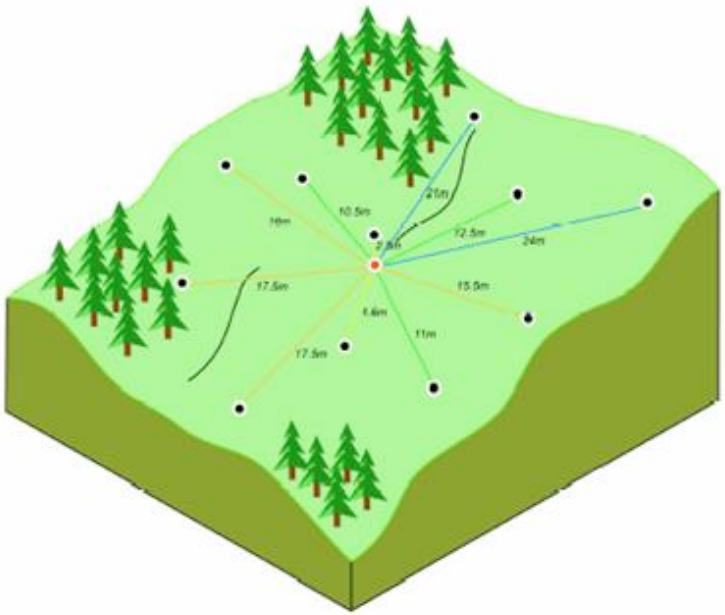
Distance matrices

$n \times n$ matrix, i indexes rows, j indexes columns



$$\begin{pmatrix} 0 & d_{1,2} & d_{1,3} & . & . & d_{1,n} \\ d_{2,1} & 0 & d_{2,3} & . & . & d_{2,n} \\ d_{3,1} & d_{3,2} & 0 & . & . & d_{3,n} \\ . & . & . & 0 & . & . \\ . & . & . & . & 0 & . \\ d_{n,1} & d_{n,2} & . & . & d_{n,n-1} & 0 \end{pmatrix}$$

http://planet.botany.uwc.ac.za/nisl/GIS/spatial/chap_1_38.htm



Often each pair of locations has a unique distance and there are often many pairs of points. To plot all pairs becomes tedious. Instead of plotting each pair, the pairs are often aggregated into *lag bins*. For example, we compute the average semivariance for all pairs of points that are greater than 40 meters apart but less than 50 meters. Thus, $N(d)$is the number of pairs at lag distance d.
(http://planet.botany.uwc.ac.za/nisl/GIS/spatial/chap_1_38.htm)

Assessing spatial correlation

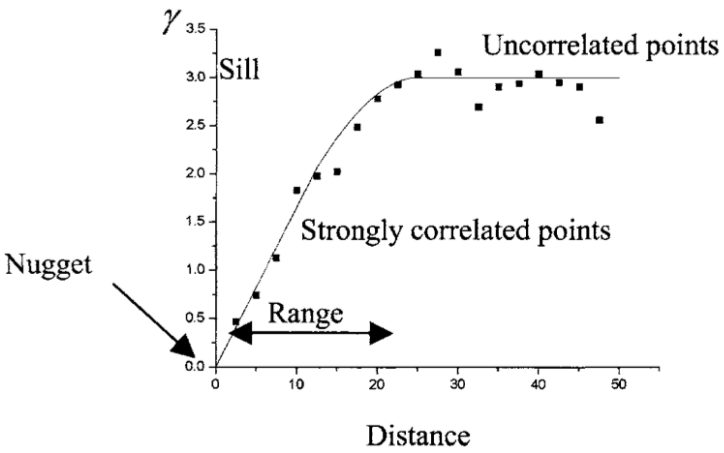
Let $\mu_i = g(\boldsymbol{\theta}, \mathbf{x}_i)$

1. Assume $\mathbf{y}$ is measured at n spatial locations.
2. Compute the residuals: $\mathbf{e} = \mathbf{y} - \boldsymbol{\mu}$.
3. Examine the residuals $\mathbf{e}$ for spatial correlation (i.e., autocorrelation).

Assessing spatial correlation

Empirical semi-variogram

$$\hat{\gamma}(d) = \frac{\sum_{i,j \in N(d)} (e_i - e_j)^2}{2N(d)}$$



$N(d)$ is the number of pairs of residuals at distance d.The y axis is the average squared difference between pairs of residuals at a given distance d divided by two. The x axis is the distance between pairs. Distances are often binned into categories.

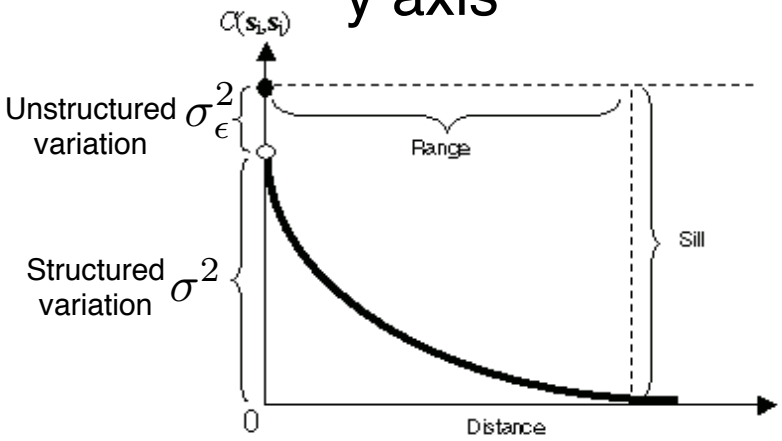
In MCMC, compute residuals at each iteration, compute $\gamma(d)^{(k)}$ and plot variogram using posterior mean of $\gamma(d)$. Or, better, sample MCMC output for residuals in R, use R functions (geoR package) to find the variogram.

Modeling spatial structure with two sources of variation

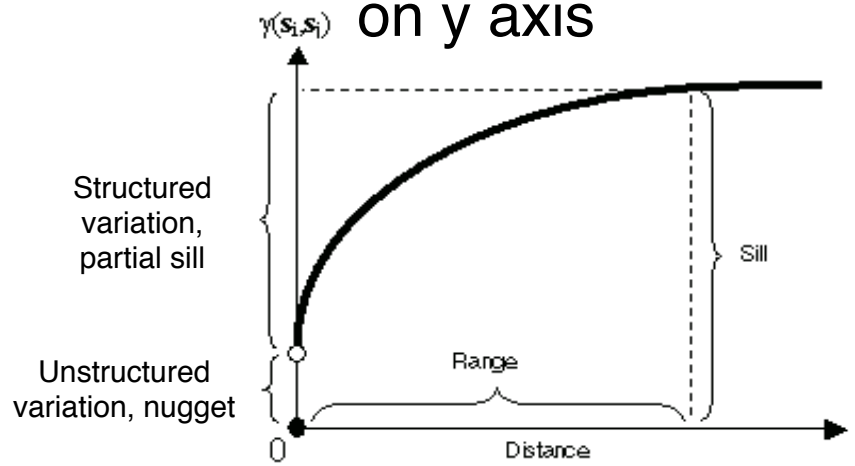
1. **Correlated error:** The structured, process component. Varies with distance between points. Process variance here.
2. **Uncorrelated error:** The unstructured, site specific component. It includes effects of fine scale heterogeneity and measurement error.

Modeling spatial structure with two sources of variation

Covariogram, covariance (cov) on y axis



Semi-variogram, semivariance γ(*d*) on y axis



Distance between observation *i* and *j*

$$\mathrm{cov}(d) = \mathrm{cov}(0) - \gamma(d)$$

Figures modified from ESRI ArcGIS Desktop online manual

Remember the covariance matrix $\mathbf{\Sigma}$

Imagine a vector of 3 random variables, $(z_i, z_2, z_3)'$ The covariance between any two of these random variables is simply an unstandardized version of the correlation between them– it is correlation measured in the units of the random variables. The covariance matrix (aka variance covariance matrix) of the random variable is:

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \text{Cov}_{1,2} & \text{Cov}_{1,3} \\ \text{Cov}_{2,1} & \sigma_2^2 & \text{Cov}_{2,3} \\ \text{Cov}_{3,1} & \text{Cov}_{3,2} & \sigma_3^2 \end{pmatrix}$$

Generalizing, a $m \times m$ covariance matrix has the variances of the random variable on the diagonal and the covariance on the off diagonal. The covariance between random variable i and j is $\text{Cov}_{ij} = \rho \sigma_i \sigma_j$ where ρ is the correlation coefficient, which takes on values between -1 and 1. Covariance can take on values between $-\infty$ and $+\infty$.

Remember the identity matrix $\mathbf{I}$

Using a 3 x 3 matrix to illustrate:

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$\sigma_\varepsilon^2\mathbf{I} = \begin{pmatrix} \sigma_\varepsilon^2 & 0 & 0 \\ 0 & \sigma_\varepsilon^2 & 0 \\ 0 & 0 & \sigma_\varepsilon^2 \end{pmatrix}$$

Modeling spatial structure with two sources of error

$\mu_i = g(\boldsymbol{\theta}, x_i)$, a model of a socio-ecological process that can take on real values (for now).
$\mu = g(\boldsymbol{\theta}, \mathbf{X})$, note that $\boldsymbol{\mu}$ is a vector with length = number of observations (n) and $\mathbf{X}$ is a data matrix with number of rows $= n$ and number of columns = number of predictor variables.

$$\mathbf{y} \sim \text{multivariate normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma} + \sigma_\varepsilon^2 \mathbf{I})$$

$\boldsymbol{\Sigma}$ is an $n \times n$ matrix with structured variance (σ^2) at distance 0 on the diagonal and the covariance between observation i and observation j on the off diagonals $(i \neq j)$. $\mathbf{I}$ an $n \times n$ matrix with ones on the diagonal and zeros elsewhere. σ_ε^2 is unstructured (uncorrelated) variance.

Alternative notation: random effects approach

$$\mathbf{y} = g(\boldsymbol{\theta}, X) + \boldsymbol{\eta} + \boldsymbol{\varepsilon}$$

1. Correlated Error: $\boldsymbol{\eta} \sim$ multivariate normal$(\mathbf{0}, \boldsymbol{\Sigma})$
2. Uncorrelated Error: $\boldsymbol{\varepsilon} \sim$ multivariate normal$(\mathbf{0}, \sigma_{\varepsilon}^2 \mathbf{I})$

Alternative notation: hierarchical approach

$$\begin{aligned} \mathbf{y} &\sim \text{multivariate normal}(g(\boldsymbol{\theta}, \mathbf{X}) + \boldsymbol{\eta}, \sigma_{\varepsilon}^2 \mathbf{I}) \\ \boldsymbol{\eta} &\sim \text{normal}(0, \boldsymbol{\Sigma}) \end{aligned}$$

1. Correlated Error: $\boldsymbol{\eta}$
2. Uncorrelated Error: σ_e^2

Modeling spatial structure with two sources of error

These both imply:

$$\mathbf{y} \sim \text{multivariate normal}(g(\boldsymbol{\theta}, \mathbf{X}), \boldsymbol{\Sigma} + \sigma_{\varepsilon}^2 \mathbf{I})$$

Modeling spatial structure with two sources of error

Do we really need to predict $\frac{1}{2}(n^2 - n)$ covariances? No. Instead, we model[1] them as a function of distance using parametric covariance functions. [2]:

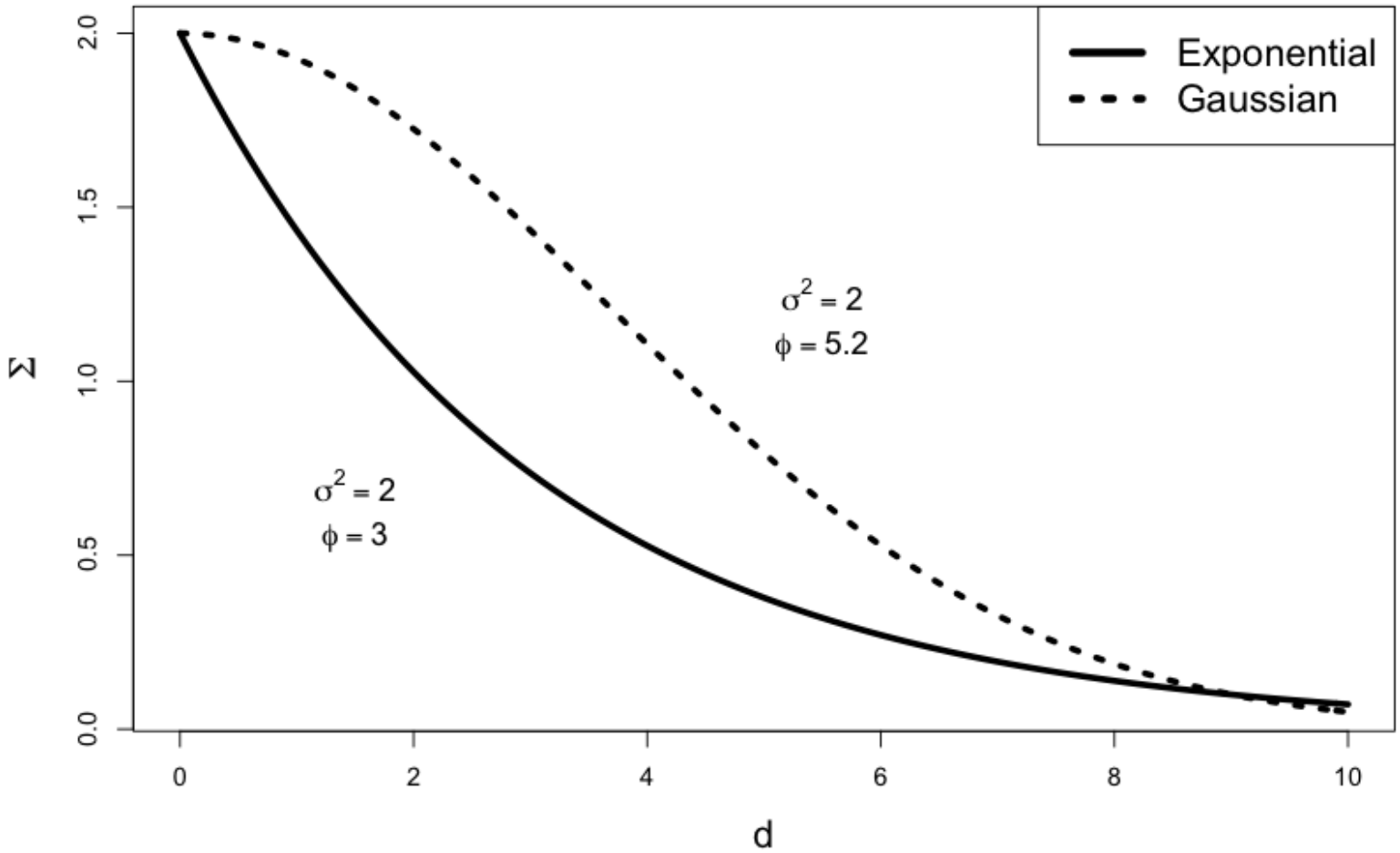
- Exponential: $\Sigma_{i,j} = \sigma^2 \exp\left(-\frac{d_{i,j}}{\phi}\right)$
- Gaussian: $\Sigma_{i,j} = \sigma^2 \exp\left(-\frac{d_{i,j}^2}{\phi^2}\right)$

where $d_{i,j}$ = distance between locations i and j. Note that an aspatial model would require approximating the posterior distribution of a single variance parameter σ^2. The spatial equivalent requires three: σ^2, ϕ and σ_ε^2. Also note that when $i = j$ such that we are "at" a location, $d_{i,j} = 0$ and $\Sigma_{i,j} = \sigma^2$.

[1] This is a great illustration of the main purpose of science: dimension reduction.

[2] There are many others, but these are used most frequently.

Modeling spatial structure with two sources of error



Important assumptions

- **Stationarity**: spatial structure does not vary with location, which means that the spatial correlation does not change within the area being analyzed.
- **Isotropy**: spatial structure does not vary with direction, which means the spatial correlation does not change with direction.

Toy illustration for 3 data points and simple linear regression

$$
\begin{aligned}
\mathbf{y} &= (y(\mathbf{s}_1), y(\mathbf{s}_2), y(\mathbf{s}_3)) \\
\boldsymbol{\beta} &= (\beta_0, \beta_1) \\
\mathbf{X} &= \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{pmatrix} \\
\mathbf{D} &= \begin{pmatrix} d_{1,1} & d_{1,2} & d_{1,3} \\ d_{2,1} & d_{2,2} & d_{2,3} \\ d_{3,1} & d_{3,2} & d_{3,3} \end{pmatrix}
\end{aligned}
\qquad
\begin{aligned}
\boldsymbol{\Sigma} &= \begin{pmatrix} \sigma^2 & \sigma^2 e^{-\frac{d_{1,2}}{\phi}} & \sigma^2 e^{-\frac{d_{1,3}}{\phi}} \\ \sigma^2 e^{-\frac{d_{2,1}}{\phi}} & \sigma^2 & \sigma^2 e^{-\frac{d_{2,3}}{\phi}} \\ \sigma^2 e^{-\frac{d_{3,1}}{\phi}} & \sigma^2 e^{-\frac{d_{3,,2}}{\phi}} & \sigma^2 \end{pmatrix} \\
\sigma_\varepsilon^2 \mathbf{I} &= \begin{pmatrix} \sigma_\varepsilon^2 & 0 & 0 \\ 0 & \sigma_\varepsilon^2 & 0 \\ 0 & 0 & \sigma_\varepsilon^2 \end{pmatrix} \\
\boldsymbol{\Sigma} + \sigma_\varepsilon^2 \mathbf{I} &= \begin{pmatrix} \sigma^2 + \sigma_\varepsilon^2 & \sigma^2 e^{-\frac{d_{1,2}}{\phi}} & \sigma^2 e^{-\frac{d_{1,3}}{\phi}} \\ \sigma^2 e^{-\frac{d_{2,1}}{\phi}} & \sigma^2 + \sigma_\varepsilon^2 & \sigma^2 e^{-\frac{d_{2,3}}{\phi}} \\ \sigma^2 e^{-\frac{d_{3,1}}{\phi}} & \sigma^2 e^{-\frac{d_{3,,2}}{\phi}} & \sigma^2 + \sigma_\varepsilon^2 \end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{y} &\sim \text{multivariate normal}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma} + \sigma_\varepsilon^2 \mathbf{I}) \\
[\boldsymbol{\beta}, \sigma^2, \sigma_\varepsilon^2, \phi \mid \mathbf{y}] &\propto [\mathbf{y} \mid \mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma} + \sigma_\varepsilon^2 \mathbf{I}] \\
&\times [\boldsymbol{\beta}][\sigma^2][\sigma_\varepsilon^2][\phi]
\end{aligned}
$$

Priors on ϕ

Choices for range parameter ϕ:

- $\phi \sim \text{gamma}(\gamma_1, \gamma_2)$
- $\log(\phi) \sim \text{normal}(\mu_\phi, \sigma^2_\phi)$
- $\phi \sim \text{half-Cauchy}(\gamma)$
- $\phi \sim \text{uniform}(0, \gamma)$

General spatial models

- Real valued, non-negative

$$\begin{aligned} g(\boldsymbol{\beta}, \mathbf{X}) &= \exp(\mathbf{X}\boldsymbol{\beta}) \\ \log(\mathbf{y}) &\sim \text{multivariate normal}(\log(g(\boldsymbol{\beta}, \mathbf{X}), \boldsymbol{\Sigma} + \sigma_{\varepsilon}^2 \mathbf{I}) \end{aligned}$$

- Counts

$$\begin{aligned} g(\boldsymbol{\beta}, \mathbf{X}) &= \exp(\mathbf{X}\boldsymbol{\beta}) \\ \log(\boldsymbol{\lambda}) &\sim \text{multivariate normal}(\log(g(\boldsymbol{\beta}, \mathbf{X}), \boldsymbol{\Sigma} + \sigma_{\varepsilon}^2 \mathbf{I})) \\ y_i &\sim \text{Poisson}(\lambda_i) \end{aligned}$$

- Binary

$$\begin{aligned} g(\boldsymbol{\beta}, \mathbf{X}) &= \text{logit}^{-1}(\mathbf{X}\boldsymbol{\beta}) \\ \text{logit}(\mathbf{p}) &\sim \text{multivariate normal}(\text{logit}(g(\boldsymbol{\beta}, \mathbf{X}), \boldsymbol{\Sigma} + \sigma_{\varepsilon}^2 \mathbf{I})) \\ y_i &\sim \text{Bernoulli}(p_i) \end{aligned}$$

Simulating data for a continuous spatial process

1. Choose locations $\mathbf{s}_i$ for $i = 1, \ldots, n$.
2. Choose the mean $\boldsymbol{\mu}$. This could be a scalar or it could vary spatially. It could be the output of a model with parameter values that you choose and $\mathbf{x}$ data.
3. Choose the unstructured variance σ^2_ε.
4. Choose range parameter ϕ and structured variance component σ^2.
5. Compute distance matrix $\mathbf{D}$ between all n locations of interest.
6. Calculate covariance matrix $\boldsymbol{\Sigma} = \sigma^2 \exp\left(-\frac{\mathbf{D}}{\phi}\right)$.
7. Sample the n-dimensional vector $\mathbf{y} \sim$ multivariate normal$(\boldsymbol{\mu}, \boldsymbol{\Sigma} + \sigma^2_\varepsilon \mathbf{I})$.