

---

# Spatial Multivariate Processes with Skewed Marginals

Hao Zhang

Department of Statistics  
Washington State University  
USA

# Multivariate Geospatial Data

---

In many studies, several variables, say  $Y_1, \dots, Y_p$ , are observed at spatial locations. Consider  $p = 2$  for simplicity.

- Each variable may represent an underlying spatial process  $Y_i(\mathbf{s})$ . Each process is spatially autocorrelated
- Two variables  $Y_1(\mathbf{s})$  and  $Y_2(\mathbf{s})$  are correlated, as seen from a scatter plot.
- If the two variables are analyzed individually, the aforementioned correlation is ignored.
- If the correlation  $\text{Corr}(Y_1(\mathbf{s}), Y_2(\mathbf{s}))$  is modeled and believed to be positive, then  $\text{Corr}(Y_1(\mathbf{s}_1), Y_2(\mathbf{s}_2)) > 0$  (for sufficiently close locations  $\mathbf{s}_1$  and  $\mathbf{s}_2$  if not for all different locations).

Some references: Wackernagel (2003), Chapter 5.

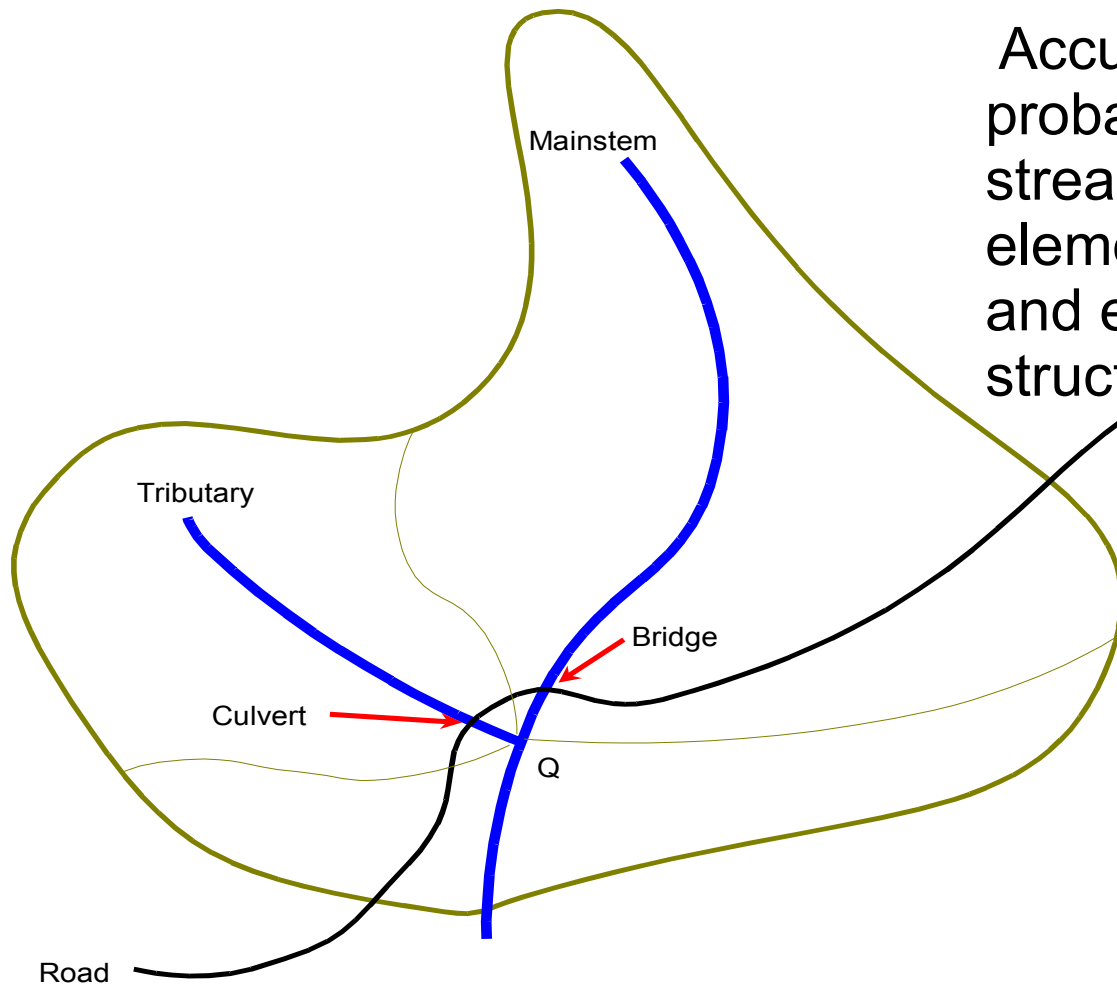
Chapters in Chilés and Delfiner (1999), Banerjee et al. (2003) and Le and Zidek (2006).

# Non-Gaussian Data

---

When the observed spatial variables are non-Gaussian, the shape of the distributions is of interest.

- Better prediction (e.g., Lognormal kriging).
- Calculation of probability of exceedance  $P(X(\mathbf{s}) > a)$ .



Accurate estimates of the joint probability of design flows at stream confluences are a crucial element in the design of efficient and effective highway drainage structures.

We need to calculate the probability of coincident exceedance  $P(FM+FT>Q)$ , or the conditional probabilities  $P(FM>Q1 | FT>Q2)$  or  $P(FT>Q2 | FM>Q1)$

Table 1. Existing criteria for selecting design return periods for concurrent flooding analysis  
(1999 AASHTO Model Drainage Manual & Montana DOT Drainage Manual)

| Area Ratio  | Frequency for Joint Occurrence |           |                |           |                 |           |
|-------------|--------------------------------|-----------|----------------|-----------|-----------------|-----------|
|             | 10-year design                 |           | 50-year design |           | 100-year design |           |
|             | main stream                    | tributary | main stream    | tributary | Main stream     | tributary |
| 10,000 to 1 | 1                              | 10        | 2              | 50        | 2               | 100       |
|             | 10                             | 1         | 50             | 2         | 100             | 2         |
| 1,000 to 1  | 2                              | 10        | 5              | 50        | 10              | 100       |
|             | 10                             | 2         | 50             | 5         | 100             | 10        |
| 100 to 1    | 5                              | 10        | 10             | 50        | 25              | 100       |
|             | 10                             | 5         | 50             | 20        | 100             | 25        |
| 10 to 1     | 10                             | 10        | 25             | 50        | 50              | 100       |
|             | 10                             | 10        | 50             | 25        | 100             | 50        |
| 1 to 1      | 10                             | 10        | 50             | 50        | 100             | 100       |
|             | 10                             | 10        | 50             | 50        | 100             | 100       |

For example, if a tributary stream has a drainage area of 20 acres and discharges into a main stream waterway with an area of 2000 acres, the Area Ratio is 100 to 1 and the 50-year design discharge on the tributary is assumed to occur when a 10-year design discharge is occurring on the main stream.

# Non-Gaussian Geospatial Processes

---

- Trans-Gaussian processes (lognormal, Box-Cox, etc)
- Spatial generalized linear mixed models (as in model-based geostatistics, Diggle et al., 1998)
- Scale-mixing of Gaussian processes (Gaussian-log-Gaussian, Palacios and Steel, 2006)

In this talk, our primary interest is to directly model stationary processes with continuous but skewed marginal distributions.

# Multivariate Distributions and Multivariate Processes

---

Consider only one variable that is observed at locations  $\mathbf{s}_1, \dots, \mathbf{s}_n$ . We know the marginals are skewed based on physical background or an exploratory data analysis.

One might attempt to use a skewed multivariate distribution to model  $\mathbf{Y} = (Y(\mathbf{s}_1), \dots, Y(\mathbf{s}_n))'$ . However, some multivariate distributions may not be appropriate for modeling the skewness of spatial data.

# Multivariate Skew-Normal Distribution

---

Azzalini and Dalla Valle (1996): Consider an  $n$ -dimensional normal variable  $\mathbf{X} = (X_1, \dots, X_n)'$  with standardized marginals, independent of  $X_0 \sim N(0, 1)$ . For constants  $m_j, \delta_j \in R$ ,  $\sigma_j \geq 0$ ,  $j = 1, \dots, n$ , define

$$Z_j = m_j + \delta_j |X_0| + \sigma_j X_j. \quad (1)$$

Then the joint distribution of  $\mathbf{Z} = (Z_1, \dots, Z_n)'$  is called a multivariate skew-normal distribution and each marginal distribution is skew-normal.

If  $Z_1, \dots, Z_n$  represent a partial realization of a *stationary* spatial process, then  $m_j = m$ ,  $\delta_j = \delta$ ,  $\sigma_j = \sigma$ . Hence

$$Z_j = m + \delta |X_0| + \sigma X_j, \quad j = 1, \dots, n.$$

Given any realization,  $\mathbf{Z}$  behaves like a multivariate normal. For example, the histogram of  $Z_1, \dots, Z_n$  will be bell-shaped.

Therefore, for the model, multiple realizations of the process have to be observed for inferences. For spatial data, only one realization is available.

---

Model the process, not the finite  
distribution.

# Trans-Gaussian Processes

---

Transform the observed  $Y(\mathbf{s})$ ,

$$Z(\mathbf{s}) = g(Y(\mathbf{s}))$$

for some function  $g$ , and  $Z(\mathbf{s})$  is stationary Gaussian. For example,  $g$  may be the Box-Cox transformation (De Oliveira, et al. 1997).

# Gamma Processes

---

Gamma process  $Y(\mathbf{s})$ ,  $\mathbf{s} \in R^d$ :

- Each  $Y(\mathbf{s})$  has a gamma marginal distribution.
- The correlation function is parametric and manageable.
- The full distribution is characterized by a few parameters.

# Gamma Processes

---

Let  $Z_i(\mathbf{s}), \mathbf{s} \in R^d$  ( $i = 1, \dots, q$ ) be  $q$  Gaussian stationary processes with standardized marginals and common correlation function  $\rho(\mathbf{h})$ . Assume the  $q$  processes are independent of each other. Define

$$Y(\mathbf{s}) = \sum_{i=1}^q Z_i(\mathbf{s})^2 / q.$$

1.  $Y(\mathbf{s})$  is stationary with  $Ga(q/2, q/2)$  marginals.
2.  $Y(\mathbf{s})$  has a correlation function  $\rho(\mathbf{h})^2$ .
3. The distribution  $(Y(\mathbf{s}_1), Y(\mathbf{s}_2), \dots, Y(\mathbf{s}_n))'$  is called multivariate gamma (Krishnamoorthy and Parthasarathy, 1951).

Extension to non-integer  $q$  is possible (Henderson and Shimakura, 2003)

# A New Class

---

Zhang and El-Shaarawi (2006)

$$Y(\mathbf{s}) = m(\mathbf{s}) + \sigma_1 |X_1(\mathbf{s})|^q + \sigma_2 X_2(\mathbf{s}) + \sigma_0 X_0(\mathbf{s}), \quad (2)$$

where

- $q > 0, \sigma_2 > 0, \sigma_0 > 0, \sigma_1 \in R$ .
- The three processes  $X_i(\mathbf{s})$  are independent of each other, each having standard margins.
- The processes  $X_1(\mathbf{s})$  and  $X_2(\mathbf{s})$  are spatially autocorrelated with correlation functions  $\rho_i(\mathbf{h}, \psi_i)$ ,  $i = 1, 2$ .
- $X_0(\mathbf{s})$  is a white noise.
- $m(\mathbf{s}) = \beta_0 + \sum_{j=1}^p g_j(\mathbf{s})\beta_j$ , for some observable explanatory variables  $g_j(\mathbf{s})$ .

$Y(\mathbf{s})$  is right-skewed if  $\sigma_1 > 0$ , left-skewed if  $\sigma_1 < 0$  or Gaussian if  $\sigma_1 = 0$ .

# A Skew-Gaussian Process

---

A special case,  $q = 1$ :

$$Y(\mathbf{s}) = m(\mathbf{s}) + \sigma_1 |X_1(\mathbf{s})| + \sigma_2 X_2(\mathbf{s}) + \sigma_0 X_0(\mathbf{s}).$$

- Marginally,  $Y(\mathbf{s})$  has a skew-normal distribution. However, the distribution  $(Y(\mathbf{s}_1), \dots, Y(\mathbf{s}_n))$  is not multivariate skew-normal.
- $Y(\mathbf{s})$  has a covariogram

$$\begin{aligned} C(\mathbf{h}) &= \frac{2\sigma_1^2}{\pi} \left( \sqrt{1 - \rho_1(\mathbf{h})^2} + \rho_1(\mathbf{h}) \arcsin(\rho_1(\mathbf{h})) - 1 \right) \\ &+ \sigma_2^2 \rho_2(\mathbf{h}) + \sigma_0^2 1_{\{\mathbf{h}=0\}}. \end{aligned} \tag{3}$$

# Maximum Likelihood Estimation

---

Observe  $\mathbf{Y} = (Y(\mathbf{s}_1), \dots, Y(\mathbf{s}_n))'$  from the skewed model.

The likelihood can't be calculated in closed form.

The EM algorithm seems straightforward.

Assume right-skewed marginals. Write the model

$$\mathbf{Y} = G\boldsymbol{\beta} + |\mathbf{X}|^q + \mathbf{W},$$

where  $\mathbf{X} \sim N(0, V_1)$ ,  $\mathbf{W} \sim N(0, V_2)$ . Then the covariance matrices  $V_i$  can be written as

$$V_1 = \tau_1 R_1, \quad V_2 = \tau_2 R_2 + \tau_0 I,$$

where  $R_k = R_k(\psi_k) = (\rho(\|\mathbf{s}_i - \mathbf{s}_j\|, \psi_k))_{i,j=1}^n$ ,  $k = 1, 2$ .

Note  $\mathbf{Y}|\mathbf{X}$  is  $N(G\boldsymbol{\beta} + |\mathbf{X}|^q, V_2)$ .

# EM Algorithm

---

Treat  $\mathbf{X}$  as unobservable latent variables. The complete-data log likelihood function for  $(\mathbf{X}, \mathbf{Y})$  is

$$\log L_c(\theta) = \log f(\mathbf{X}, \sigma_1, \tau_1) + \log f(\mathbf{Y}|\mathbf{X}, \beta, \tau_0, \tau_2, \psi_2).$$

The EM algorithm runs iteratively. Given estimate  $\theta^{(m)}$ , the new estimate  $\theta^{(m+1)}$  maximizes

$$\begin{aligned} Q(\theta|\theta^{(m)}) &= E_{\theta^{(m)}}(\log L_c(\theta, \mathbf{Y}, \mathbf{X})|\mathbf{Y}) \\ &= E_{\theta^{(m)}}(\log f(\mathbf{X}, \sigma_1, \psi_1)|\mathbf{Y}) + E_{\theta^{(m)}}(\log f(\mathbf{Y}|\mathbf{X}, \beta, \sigma_0, \sigma_2, \psi_2)|\mathbf{Y}) \end{aligned}$$

The conditional expectations can be approximated by a MCMC method.

# Slice Sampling

---

Recall the model  $\mathbf{Y} = G\boldsymbol{\beta} + |\mathbf{X}|^q + \mathbf{W}$

- Want to generate  $\mathbf{X}$  from conditional distribution  $f(\mathbf{X}|\mathbf{Y})$ .
- Metropolis-Hastings algorithm does not work well.
- Slicing sampling: Use auxiliary variables and sample from uniform distributions (Neal, 2003, Agarwal and Gelfand, 2005).

Let  $U|(\mathbf{X}, \mathbf{Y}) \sim U[0, f(\mathbf{Y}|\mathbf{X})]$ , then  $\mathbf{X}|(U, \mathbf{Y}) \sim f(\mathbf{X})1_{\{U < f(\mathbf{Y}|\mathbf{X})\}}$ .

Slicing sampling (Gibbs sampler): Start at some  $\mathbf{X}^{(0)}$ .

- (1) Given  $\mathbf{X}^{(t)}$ , generate  $U^{(t+1)} \sim U[0, f(\mathbf{Y}|\mathbf{X}^{(t)})]$ .
- (2) Generate  $\mathbf{X}^{(t+1)} \sim f(\mathbf{X})1_{\{U^{(t+1)} < f(\mathbf{Y}|\mathbf{X})\}}$ .
- (3) Repeat.

# Slice Sampling

---

Given  $\mathbf{X}^{(t)}$ , generate  $\tilde{U} \sim U[0, f(\mathbf{X}^{(t)})]$ , and generate  $\mathbf{X}^{(t+1)}$  uniformly on the set

$$\{\mathbf{x} : f(\mathbf{x}) 1_{\{U^{(t+1)} < f(\mathbf{Y}|\mathbf{x})\}} > \tilde{U}\} = \{\mathbf{x} : f(\mathbf{x}) > \tilde{U}\} \cap \{\mathbf{x} : f(\mathbf{Y}|\mathbf{x}) > U^{(t+1)}\}. \quad (4)$$

Note that  $f(\mathbf{x}) > \tilde{U}$  if and only if  $\log f(\mathbf{x}) > \log(\tilde{U}/f(\mathbf{X}^{(t)})) + \log(f(\mathbf{X}^{(t)}))$ , and if and only if

$$\mathbf{x}'V_1^{-1}\mathbf{x} < -2\log(\tilde{U}/f(\mathbf{X}^{(t)})) + \mathbf{X}^{(t)'}V_1^{-1}\mathbf{X}^{(t)} = r,$$

which is an  $n$ -dimensional oval.

Similarly  $f(\mathbf{Y}|\mathbf{x}) > U^{(t+1)}$  if and only if

$$(\mathbf{Y}-|\mathbf{x}|^q)'V_2^{-1}(\mathbf{Y}-|\mathbf{x}|^q) < -2\log(U^{(t+1)}/f(\mathbf{Y}|\mathbf{X}^{(t)})) + (\mathbf{Y}-|\mathbf{X}^{(t)}|^q)'V_2^{-1}(\mathbf{Y}-|\mathbf{X}^{(t)}|^q) = a.$$

# Slice Sampling

---

Given  $\mathbf{X}^{(t)}$ ,

- Generate  $\eta, \xi$  i.i.d  $\text{Exp}(1)$ , and let

$$r = 2\eta + \mathbf{X}^{(t)'} V_1^{-1} \mathbf{X}^{(t)}, \quad a = 2\xi + (\mathbf{Y} - |\mathbf{X}^{(t)}|^q)' V_2^{-1} (\mathbf{Y} - |\mathbf{X}^{(t)}|^q),$$

- Generate  $\mathbf{X}^{(t+1)}$  from the uniform distribution on

$$\{\mathbf{x} : \mathbf{x}' V_1^{-1} \mathbf{x} < r\} \cap \{\mathbf{x} : (\mathbf{Y} - |\mathbf{x}|^q)' V_2^{-1} (\mathbf{Y} - |\mathbf{x}|^q) < a\}.$$

# Bivariate Model of Skewed Processes

---

If two variables are observed and both have skewed distributions, we can extend the model to the bivariate or multivariate case:

$$\begin{aligned} Y_1(\mathbf{s}) &= m_1(\mathbf{s}) + |X_1(\mathbf{s})|^{q_1} + W_1(\mathbf{s}) \\ Y_2(\mathbf{s}) &= m_2(\mathbf{s}) + |X_2(\mathbf{s})|^{q_2} + W_2(\mathbf{s}) \end{aligned}$$

where  $\mathbf{X}(\mathbf{s}) = (X_1(\mathbf{s}), X_2(\mathbf{s}))'$  is bivariate stationary Gaussian with covariogram matrix  $\mathbf{C}(\mathbf{h}) = (C_{ij}(\mathbf{h}))$ ;  $\mathbf{W}(\mathbf{s}) = (W_1(\mathbf{s}), W_2(\mathbf{s}))$  is bivariate stationary Gaussian with covariogram  $\mathbf{K}(\mathbf{h}) = (K_{ij}(\mathbf{h}))$ .  $\mathbf{X}(\mathbf{s})$  is independent of  $\mathbf{W}(\mathbf{s})$ .

Multivariate covariogram models (Wackernagel, 1998, Chilés and Delfiner, 1999).

The EM algorithm and the slicing sampling can be applied similarly.

# Bivariate Matérn Covariogram

---

Matérn covariograms have been increasingly used recently. If both processes each has a Matérn covariogram, what would be the cross covariogram?

Consider the Matérn covariograms

$$\text{Cov}(X_i(\mathbf{s}), X_j(\mathbf{s} + \mathbf{h})) = C(\mathbf{h}, \sigma_{ij}, \alpha_{ij}) = \frac{\sigma_{ij}(\alpha_{ij}|\mathbf{h}|)^\nu}{2^{\nu-1}\Gamma(\nu)} \mathcal{K}_\nu(\alpha_{ij}|\mathbf{h}|), i, j = 1, 2$$

where  $\alpha_{ij} > 0$  and  $\Sigma = (\sigma_{ij}) > 0$ . Then  $C(\mathbf{h}, \sigma_{ij}, \alpha_{ij})$  has a spectral density

$$f_{ij}(u) = c_d \frac{\sigma_{ij} \alpha_{ij}^{2\nu}}{(\alpha_{ij}^2 + u^2)^{\nu+d/2}}.$$

It is a valid bivariate covariogram if and only if  $(f_{ij}(u))$  is non-negative definite for almost all  $u > 0$ .

# Bivariate Matern Covariogram

---

Necessary and sufficient conditions for the bivariate Matérn covariogram

$$\alpha_{12} \geq \max(\alpha_{11}, \alpha_{22}), \text{ and } \frac{\sigma_{12}}{(\sigma_{11}\sigma_{22})^{0.5}} \leq \left( \frac{\sqrt{\alpha_{11}\alpha_{22}}}{\alpha_{12}} \right)^{2\nu}.$$

Let  $\tau_{ij} = \sigma_{ij}\alpha_{ij}^{2\nu}$ . Then

$$\alpha_{12} \geq \max(\alpha_{11}, \alpha_{22}), \text{ and } (\tau_{ij}) \geq 0.$$

I believe that  $\tau_{ij}$  are microergodic parameters.

# An Application

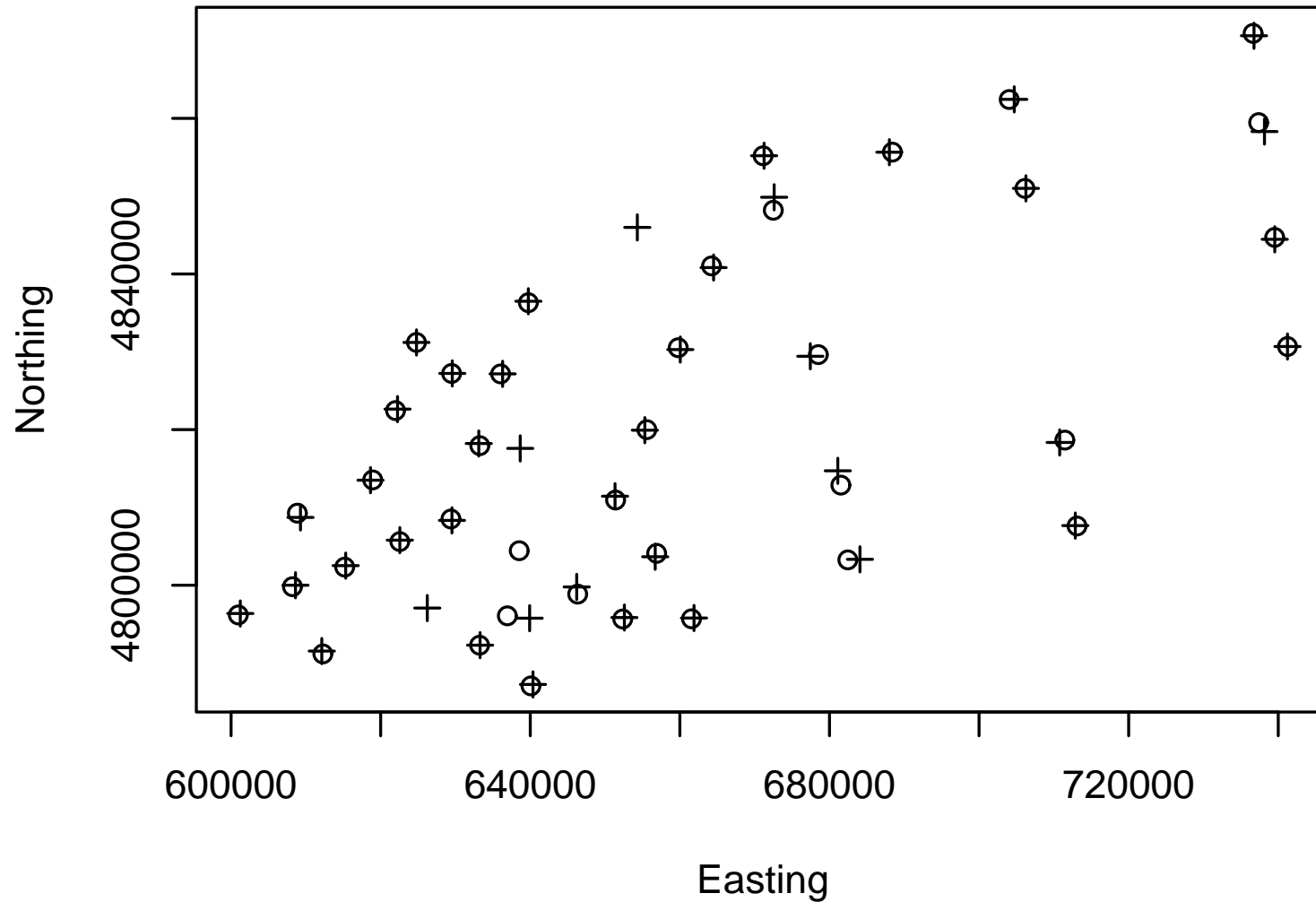
---

Lake Ontario Surveillance Program to monitor spatial and temporal changes in water quality of the Great Lakes.

- $Y(s)$ : Chlorophyll concentration at location  $s$ . A measure of phytoplankton biomass and thus lake productivity.
- One covariate is the sounding depth: shallow regions are high in nutrients and productivity.
- Dataset 1 was collected on April 29, 30 and May 1 of 1974 at 40 locations; Dataset 2 on Sep 30, Oct. 2, 4, 5 of 1974 on 42 locations.

# Locations

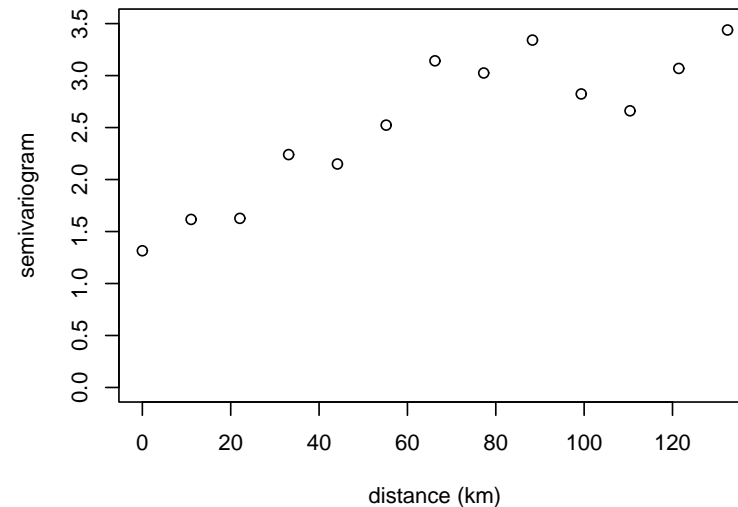
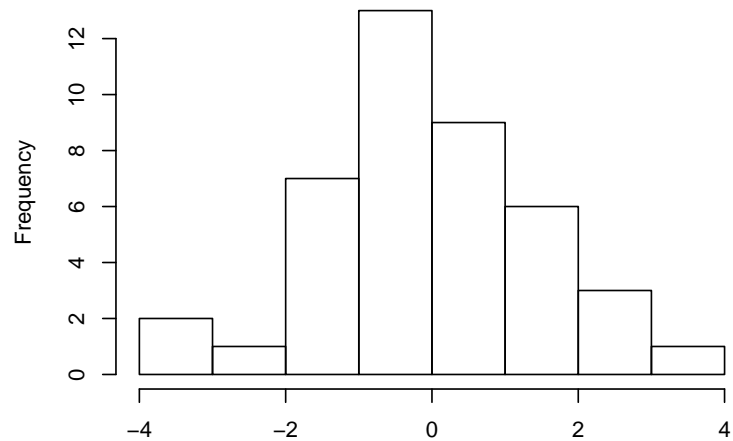
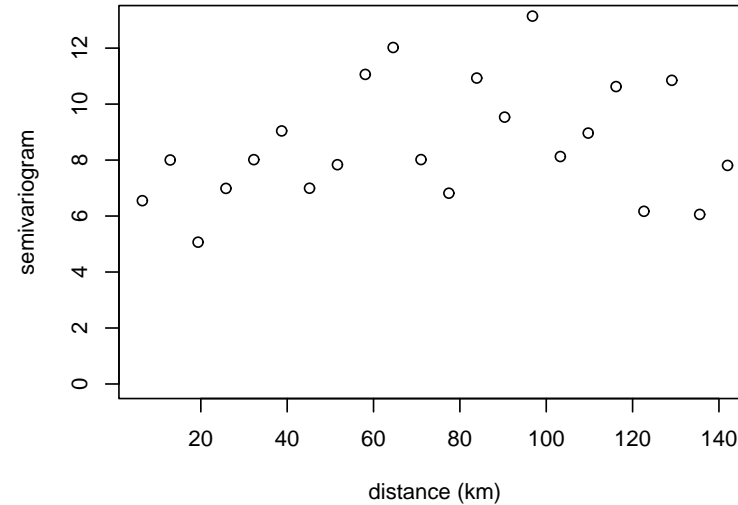
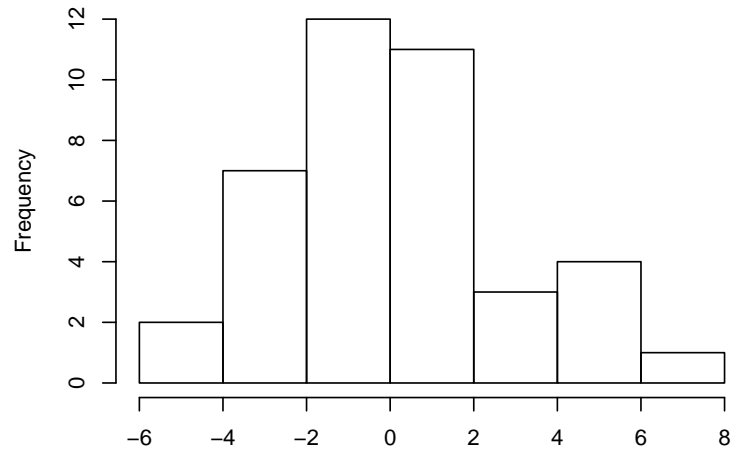
---



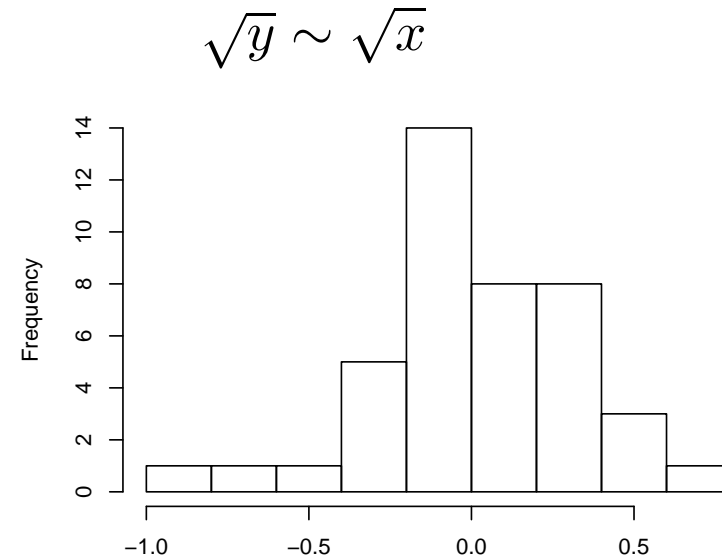
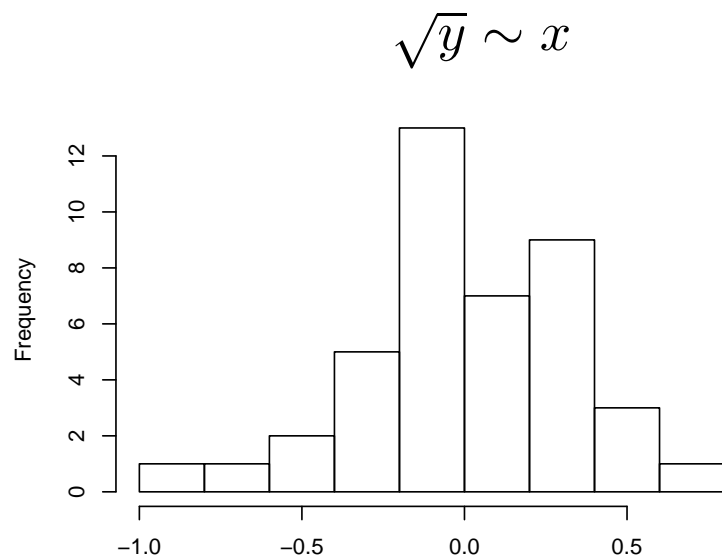
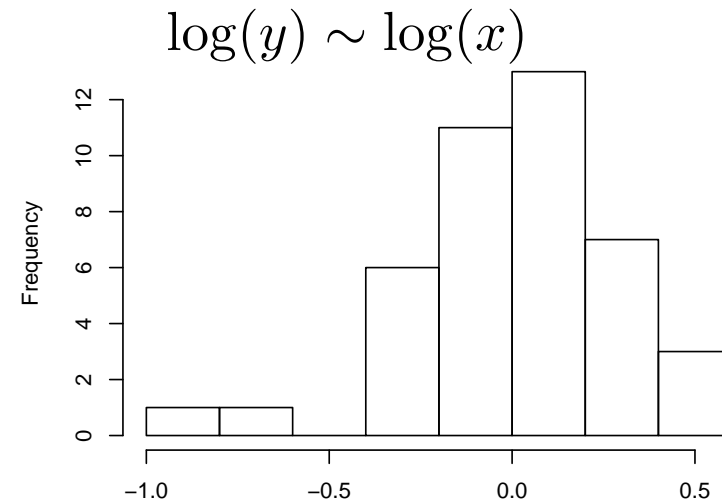
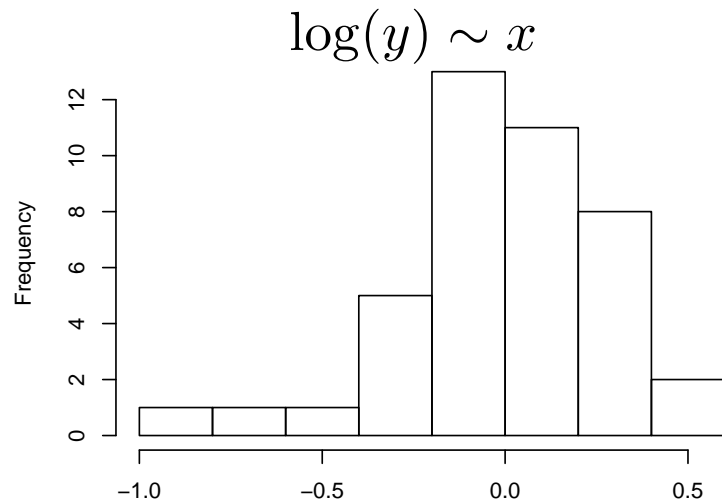
Observed locations: circle (○) for dataset 1 and plus (+) for dataset 2.

# Exploratory Data Analysis

Regress chlorophyll on sounding depth ignoring spatial correlation.



# Transformations



# Spatial Gaussian Linear Model

---

$$Y(\mathbf{s}_j) = \beta_0 + g(\mathbf{s}_j)\beta_1 + \epsilon(\mathbf{s}_j), \quad j = 1, \dots, n$$

where  $g(\mathbf{s}_j)$  is the sounding depth and  $(\epsilon(\mathbf{s}_1), \dots, \epsilon(\mathbf{s}_n))'$  is stationary Gaussian process with mean 0 and a Matérn covariogram that has a nugget effect.

Estimates :

Dataset 1:  $(\beta_0, \beta_1) = (12.267, -0.0805)$ , and  
 $(\sigma_0^2, \sigma_1^2, \phi, \nu) = (4.4505, 3.5724, 0.011, 0.25)$ .

Dataset 2:  $(\beta_0, \beta_1) = (7.0226, -0.0189)$ , and  
 $(\sigma_0^2, \sigma_1^2, \phi, \nu) = (1.9294, 0.3006, 0.3954, 0.5)$

Mean squared error of drop-one prediction

$$\begin{aligned} \sum_{i=1}^n (Y(\mathbf{s}_i) - \hat{Y}(\mathbf{s}_i))^2 / n &= 8.023 \text{ for Dataset 1} \\ &= 2.179 \text{ for Dataset 2.} \end{aligned}$$

# Skew-Gaussian Model

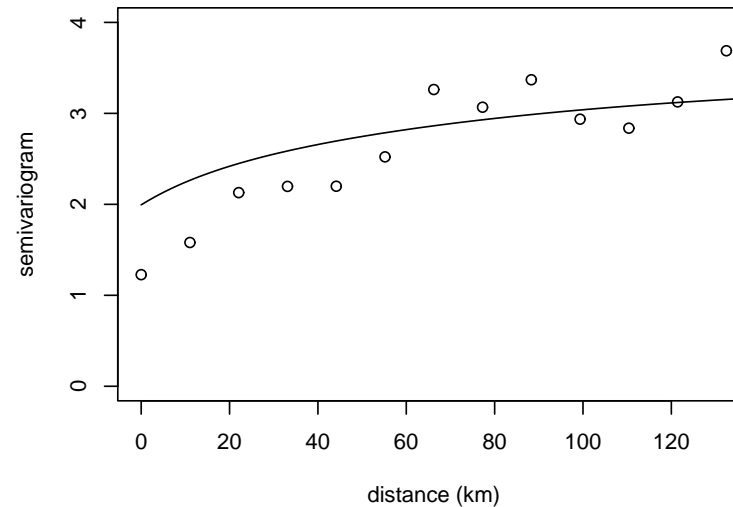
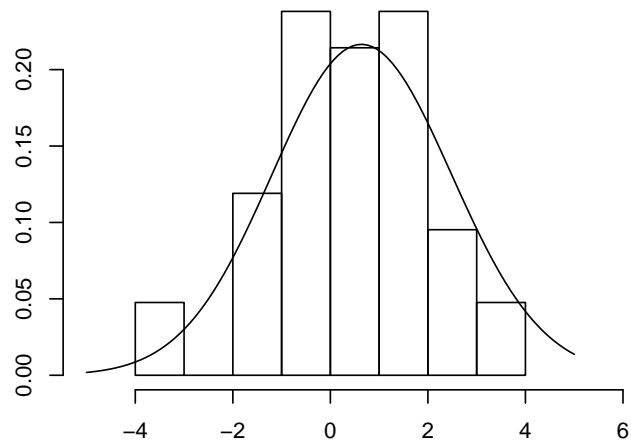
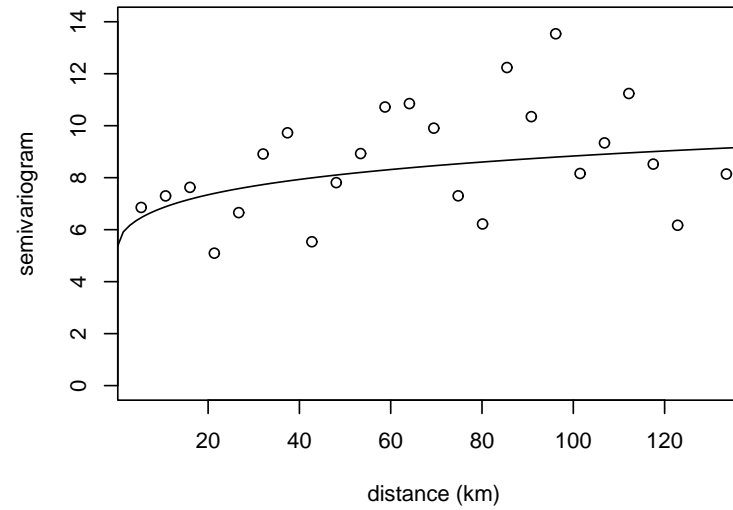
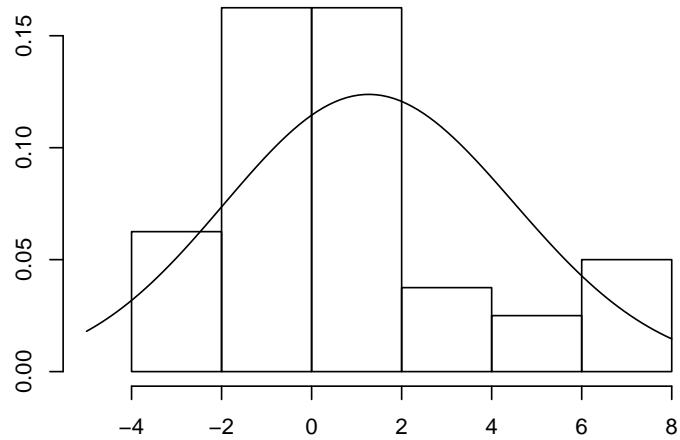
---

Employ slicing sampling to generate 250,000 Monte Carlo samples and keep every 50th one.

Dataset 1:  $(\hat{\beta}_0, \hat{\beta}_1) = (11.30, -0.0801)$ , and  
 $(\hat{\tau}_1, \hat{\phi}_1, \hat{\tau}_2, \hat{\phi}_2, \hat{\tau}_0, \hat{\nu}) = (2.7233, 0.3997, 4.1475, 1.9461, 5.2661, 0.25)$

Dataset 2:  $(\hat{\beta}_0, \hat{\beta}_1) = (6.6029, -0.0203)$ , and  
 $(\hat{\tau}_1, \hat{\phi}_1, \hat{\tau}_2, \hat{\phi}_2, \hat{\tau}_0, \hat{\nu}) = (0.6765, 0.4048, 1.1579, 1.1631, 1.9944, 0.5)$

# Residuals and Semivariograms



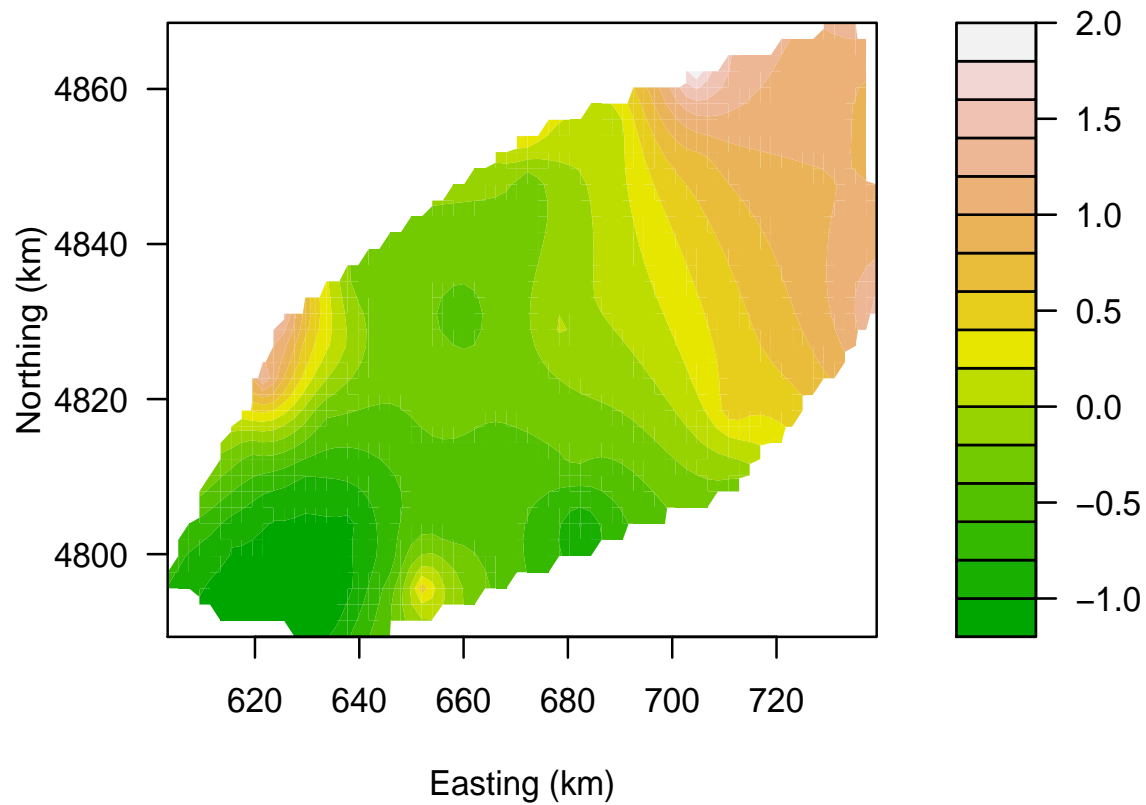
---

## Mean squared error of drop-one prediction

$$\begin{aligned} \sum_{i=1}^n (Y(\mathbf{s}_i) - \hat{Y}(\mathbf{s}_i))^2 / n &= 6.764 \text{ for Dataset 1 (v.s. 8.023)} \\ &= 1.873 \text{ for Dataset 2 (v.s. 2.179)} \end{aligned}$$

# Interpolation

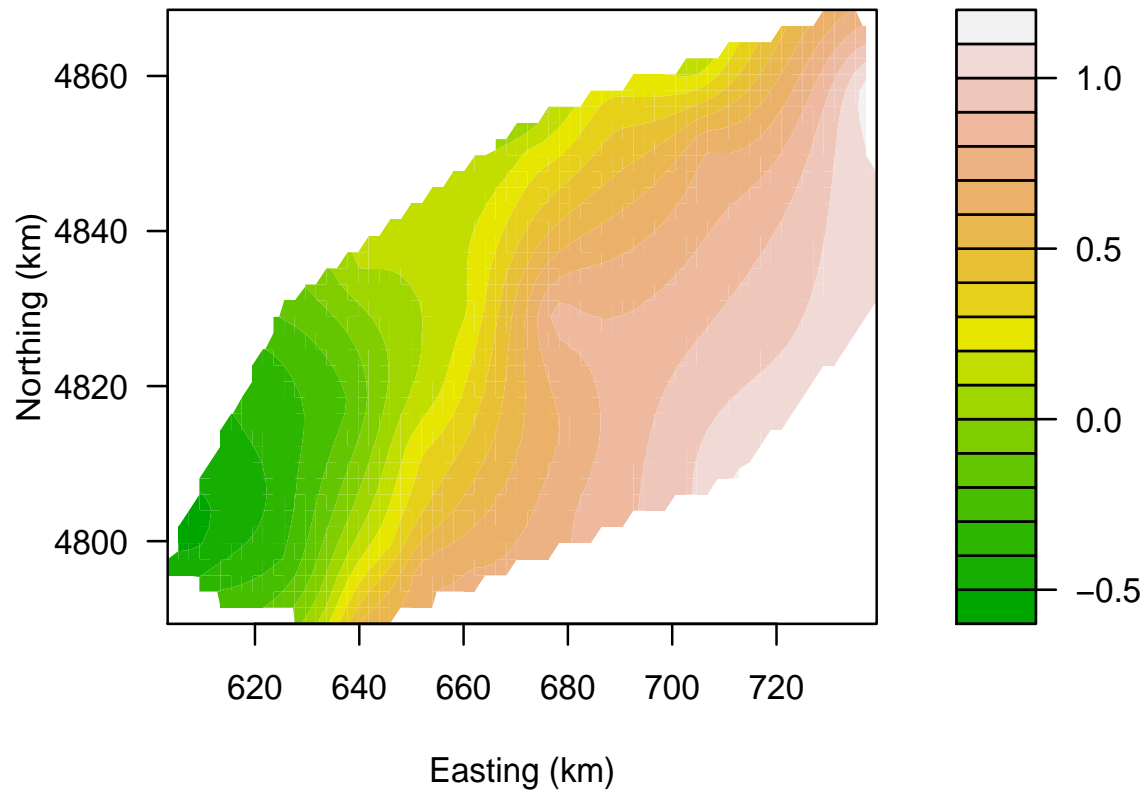
---



Dataset 1

# Interpolation

---



Dataset 2

## REFERENCES

- Azzalini, A. and Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika* **83**, 715–726.
- Banerjee, S., Carlin, B. and Gelfand, A. (2003). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC, New York.
- Chilés, J. P. and Delfiner, P. (1999). *Geostatistics: modeling spatial uncertainty*. Wiley, New York.
- De Oliveira, V., Kedm, B. and Short, D. A. (1997). Bayesian prediction of transformed gaussian random fields. *Journal of the American Statistical Association* **92**, 1422–1433.
- Diggle, P., Tawn, J. A. and Moyeed, R. A. (1998). Model-based geostatistics (with discussion). *Journal of the Royal Statistical Society. Series C. Applied Statistics* **47**, 299–350.
- Henderson, R. and Shimakura, S. (2003). A serially correlated gamma frailty model for longitudinal count data. *Biometrika* **90**, 355–366.
- Krishnamoorthy, A. S. and Parthasarathy, M. (1951). A multivariate gamma-type distribution (Corr: V31 p229). *The Annals of Mathematical Statistics* **22**, 549–557.
- Le, N. D. and Zidek, J. V. (2006). *Statistical Analysis of Environmental Space-Time Processes*. Springer, New York.
- Palacios, M. B. and Steel, M. F. J. (2006). Non-gaussian bayesian geostatistical modeling. *Journal of the American Statistical Association* **101**, 604–618.
- Wackernagel, H. (1998). *Multivariate Geostatistics: an Introduction with Applications*. Springer-Verlag Inc.
- Zhang, H. and El-Shaarawi, A. (2006). Stationary processes with skewed continuous marginals. (*Working paper*)