A Spatial Prediction Map of Selenium Concentrations

Andrea Mack

Department of Mathematical Sciences

Montana State University

May 3, 2017

A writing project submitted in partial fulfillment
of the requirements for the degree

Master of Science in Statistics

APPROVAL

of a writing project submitted by

Andrea Mack

This writing project has been read by the writing project advisor and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Statistics Faculty.

5/8/2017

Date

Andy Hoegh, Ph.D.
Writing Project Advisor

5/5/2017

Date

Mark C. Greenwood
Writing Projects Coordinator

ABSTRACT

Data were provided by the United States Geological Survey on selenium concentrations collected over Benton Lake National Wildlife Refuge near Great Falls, Montana. High selenium concentrations are known to decrease reproduction rates in lake waterfowl (USFWS, 2012, p. 204). Previous work in sediment selenium data included mean concentration estimates and comparisons using several variance estimation methods. The previous work found sediment selenium concentrations to be highest in Units 1 and 2, and in areas of the lake that are flooded. Current work confirms previous findings and provides a spatial prediction map of selenium concentrations across the entire lake. Novel to the current work is incorporation of spatial correlation and Bayesian methods into predicting sediment selenium concentrations. Work is extended by implementing the minimum average posterior prediction variance criterion developed by Diggle and Lophaven (2006) to provide suggestions for a reduced location sampling plan.

Contents

1 Introduction

Benton Lake National Wildlife Refuge is a waterfowl refuge located in north central Montana. High selenium concentrations at Benton Lake have been of concern in recent years as they are known to decrease reproduction rate in waterfowl (USFWS, 2012, p. 204). Reproductive failure results from embryonic deformities and death (Friend and Franson, 1999). Selenium is a necessary nutrient, but can be toxic at high levels. Selenium accumulates in sediment and can spread to roots and invertebrates. Selenium concentration samples were taken in water, sediment, roots, micro-invertebrate, and bird eggs. Previous analyses involved data manipulation, exploratory plots, and linear regression modeling in a classical framework. Further analysis of the data in this paper includes selenium concentration prediction maps across the entire lake using both spatial and non-spatial methods in a Bayesian framework.

Following the introduction is an overview to spatial modeling (2), including types of spatial data (2.1), modeling spatial correlation (2.2), and prediction (2.3). Bayesian samplers, cross validation and Bayes' Factor model comparisons are explained (3 & 4), and the data (5.1), sampling design criterion (5.2), model (6), and variables are then formally written along with computation methods (7). Prediction results from spatial and non-spatial models of selenium concentrations are provided (8). Section 9 includes a brief conclusion. The appendix (10) discusses possible violations to the isotropic assumption.

2 Spatial Background

2.1 Spatial Data

Spatial data have been classified into three categories: point process, point reference, and areal. Observational units for point process and point reference data are at exact point locations in the R^d space. The selenium data are in R^2. Observational units for areal data include a specified shape or area encompassing the point locations. Point process data retain only event locations, making the domain randomly determined by where events are randomly observed. For example, egg counts collected at nest locations may be point process data because nest locations, which are the sampling locations, may occur randomly. In point reference data, also known as geostatistical data, the domain is fixed and interest lies in sampling at both event and non-event locations. The selenium data are considered geostatistical because all locations on Benton Lake, a fixed domain, could be observational units.

2.2 Model Assumptions

Stationarity and isotropy are two assumptions when analyzing geostatistical data. Strong stationarity is met when the response distribution is constant over the entire surface. Second order stationarity is met when the first and second moments of the response distribution are constant over the entire surface. Intrinsic stationarity assumes a constant mean over the entire surface and assumes the variogram (defined below) only depends on the distance between two locations. The subtle difference between second order and intrinsic stationarity is that second order stationarity makes assumptions about the variation in observations and intrinsic stationarity makes assumptions about the variation in the difference of observations. A process is isotropic when the covariance only depends on the distance between two locations, and not, for example, direction. Isotropy is violated in the selenium data and is discussed in more detail in the appendix. A homogeneous process is both stationary and isotropic. Spatial models separate the residuals into two pieces, the spatially correlated and the uncorrelated portions. For the selenium data both pieces are assumed to follow normal distributions.

2.3 Spatial Correlation

Spatial models account for spatial correlation between observations. The choice of correlation function informs the smoothness of the process. It is chosen based on theoretical foundations, criteria penalizing complexity and rewarding parsimony, or by comparing the theoretical correlation structure to the empirical. The exponential and Gaussian (Table 1) are two common correlation structures.

Exponential	$R(h\|\phi) = exp[-(\frac{h}{\phi})]$ if h >0, 1 otherwise
Gaussian	$R(h\|\phi) = exp[-(\frac{h}{\phi})^2]$ if h >0, 1 otherwise

Table 1: Correlation Functions

The variogram (γ) is defined to be the expected squared difference between all pairwise combinations of responses and is used to assess the pairwise correlation of observations as a function of distance. The variogram is stated notationally in Equation 1.

$$\gamma_{i,j} = E[(Z(s_i) - Z(s_j))^2] = 2[Var(Z) + Cov(Z_i, Z_j)] = 2[\tau^2 + \sigma^2 R(h|\theta)] \tag{1}$$

for all i $\neq$ j.

The partial sill (σ^2) represents the spatial variation and the correlation function ($R(h|\theta)$) depends on some θ vector of parameters, here a single parameter, ϕ. The practical range (ϕ) is the the threshold separating distances for which the correlation function holds from the distances that are not spatially correlated. The distance between observations is denoted h. The second source of variability enters through the nugget (τ^2). The nugget can be viewed as the non-spatial variation, measurement error, and error from repeated sampling at that location and time. As Banerjee, Carlin, and Gelfand (2004) mention, the nugget does not separate the repeated sampling variability from the microscale variability. The sill represents the total variability by summing the spatial and non-spatial variation components.

The variogram is plotted versus pairwise distances between points to visualize how the variation in squared response differences changes by distance. Figure 1 was taken from PNNL's Visual Sample Plan website and nicely visualizes the sill, nugget, and range on a variogram model that is the typical shape of a variogram from spatially correlated data.

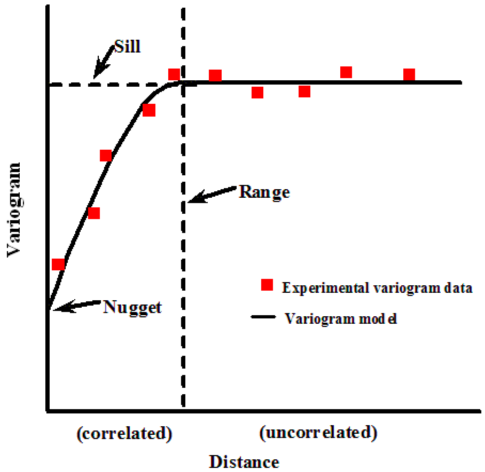


Figure 1: Semi-variogram Properties from PNNL

Variograms are not resistant to outliers. Response outliers shift the entire variogram up at all corresponding distances. A single distance outlier may show inconsistent or abnormal variogram behavior at extreme distances, leaving the center portion of the variogram intact. Clusters of distance outliers possibly affected by some anisotropic process may result in a misleading variogram in its entirety. When processes are anisotropic, variograms can be broken into components to visualize the directional aspects of their correlation (see the appendix).

Figure 2 includes data simulated under both the exponential and Gaussian correlation functions using the selenium data coordinates. These two simulated cases illustrate what would be expected under theoretical situations with the selenium data coordinates. The two variograms with simulated data are then compared to the empirical variogram for the selenium data. Also included for comparison is a variogram simulated from a pure nugget model.

The simulated variograms in Figure 2 were all found by fixing τ^2, σ^2, and ϕ. The model assumed a zero mean and the data were simulated from a multivariate Gaussian response model by using locations from the selenium data set. The exponential is slightly rougher than the Gaussian at distances less than 0.1. Overly smoothed correlation structures are not helpful in accounting for spatial variability, whereas rough correlation structures have a small squared prediction error for interpolated data, but may not do well predicting responses at new locations. Under the chosen parameters, there appears to be very little spatial correlation for distances less than 0.1, and high variability in the spatial correlation at distances at and above 0.1. The pure nugget model shows the effect of the number of pairs at each distance on the variogram. There is no spatial correlation between observations, and still the variogram has changing variability at distances greater than 0.09. This pattern is reflective of the fewer pairs of observations at those distances. The variogram plots show that with this set of parameters and locations, either there is not much difference in the covariance pattern between exponential and Gaussian spatially correlated data and uncorrelated data, or that the count at each set of distances considered does not lend an informative pattern. The exponential correlation function is the most common, and was used in creating the spatial prediction maps. The exponential correlation function limits to only allowing positive associations. Methods exist to allow negative associations but are not explored here.

2.4 Kriging

Methods for spatial prediction include linear methods, such as kriging, along with splines and other non-linear methods. Kriging is the optimal linear prediction method for spatially correlated data under the minimum mean squared error of prediction (MSEP) criterion. This means that kriging predictions are the

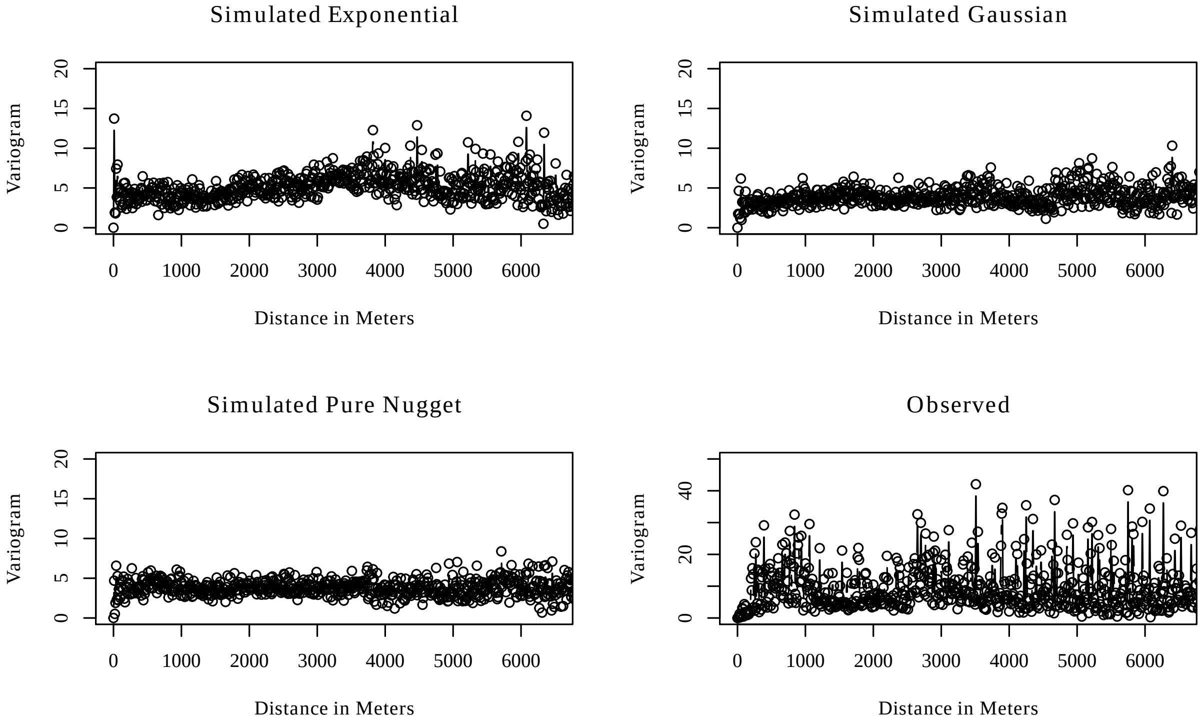


Figure 2: Variograms

best linear unbiased predictions. Kriging is an exact interpolator; it predicts values at observed locations as the values observed at those locations in a no nugget model. Under a MSEP criterion, we would expect an exact interpolator to perform better than a smooth prediction method (Papritz & Stein, p.86) such as splines. Kriging is an umbrella prediction method with many subsets. The most basic types of kriging are presented in Table 2. All kriging methods assume the response $Z(S_0)$ at new location S_0 comes from an infinite dimensional Gaussian distribution known as a Gaussian Random Process. The table below compares simple, ordinary, and universal kriging. Note that without covariates, universal kriging is ordinary kriging. Classical methods assume Σ, the covariance matrix of the residuals, is exactly known, which rarely holds.

Simple Kriging	$Z(S) = \mu + \epsilon(S)$	$\epsilon(S) \sim (0, \Sigma)$	μ, Σ known
Ordinary Kriging	$Z(S) = \mu + \epsilon(S)$	$\epsilon(S) \sim (0, \Sigma)$	μ unknown and constant, Σ known
Universal Kriging	$Z(S) = X(S)\beta + \epsilon(S)$	$\epsilon(S) \sim (0, \Sigma)$	μ unknown and varies, Σ known

Table 2: Kriging Models

Often the spatially varying mean is unknown, as in ordinary and universal kriging. Because ordinary kriging is a subset of universal kriging, for brevity, only the equations for universal kriging are provided. Using notation from Schabenberger and Gotway (2005), $p_{uk}(Z(S_0))$ denotes the kriging prediction at location S_0 and $\sigma^2_{uk}(S_0)$ denotes the prediction variance estimate (Equations 2 and 3).

$$p_{uk}(Z(S_0)) = X(S_0)^T\hat{\beta}_{GLS} + \sigma^T\Sigma^{-1}[Z(S) - X(S)^T\hat{\beta}_{GLS}] \tag{2}$$

$$\hat{\beta}_{GLS} = [1^T\Sigma^{-1}1]^{-1}1^T\Sigma^{-1}Z(s)$$

$$\sigma^2_{uk}(S_0) = \sigma_0 - \sigma^T\Sigma^{-1}\sigma + [X(S_0)^T - \sigma^T\Sigma^{-1}X(S)][X(S)^T\Sigma^{-1}X(S)]^{-1}[X(S_0)^T - \sigma^T\Sigma^{-1}X(S)]^T \tag{3}$$

Parameters in Equations 2 and 3 are defined as: σ_0, the variation at prediction location(s); σ, the covariance between new prediction(s) and observed data; Σ, the covariance between all responses such that if Σ_{11} is the covariance matrix between observed responses, Σ_{00} is the covariance of responses at predicted locations; and $\Sigma_{01} = (\Sigma_{10})^T$ is the covariance matrix between responses at new and observed locations, then $\Sigma = [\begin{smallmatrix}\Sigma_{11} & \Sigma_{10}\\ \Sigma_{01} & \Sigma_{00}\end{smallmatrix}]$.

The prediction is the least squares prediction as when assuming uncorrelated errors plus a correction that incorporates the magnitude of prediction error at observed locations. The variance includes both direct and cross covariances between observed and new locations.

The covariance matrix Σ is assumed known, but is often estimated using ordinary, weighted, and generalized least squares, as well as maximum and restricted maximum likelihood methods. However, the classical kriging methods do not incorporate the variability of the estimated covariance matrix into the prediction variance formulas. Le and Zidek proposed a Bayesian alternative to kriging in their 1992 paper. Bayesian kriging incorporates uncertainty in the posterior model into the posterior predictions and their associated errors. By ignoring uncertainty in parameter estimates, variation in predictions from classical methods may be larger than calculated. On the other hand, Bayesian methods account for the added uncertainty, making predictions “somewhat robust” to model mis-specification (Le and Zidek, 1992). Bayesian methods are also favored because they allow for continual updating of prior information, which is particularly helpful when data are collected repeatedly over time.

Partial versus fully Bayesian methods can be specified by treating some parameters as estimated and placing prior distributions on others. However, partial Bayesian methods lead to the same underestimation of variation problem as classical methods (Lee and Zedak, 1992). A sensitivity analysis can be used in deciding an appropriate method, but is not explored here. Other classical and Bayesian kriging extensions exist, but are saved for future research. Selenium concentrations are kriged using a fully Bayesian approach in alignment with the belief that uncertainty should be accounted for where it exists.

3 Bayesian Methods

Bayesian analyses begin by specification of prior distributions, p($\boldsymbol{\theta}$), for unknown parameters ($\boldsymbol{\theta}$). The data are assumed to follow some distribution $p(Z|\boldsymbol{\theta})$, with likelihood $p(\boldsymbol{\theta}|\boldsymbol{Z})$. Bayes' rule and reduction of the denominator to a constant yield the joint posterior distribution of $\boldsymbol{\theta}$ given as the proportionality below.

$$p(\boldsymbol{\theta}|\boldsymbol{Z}) = \frac{p(\boldsymbol{Z}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\theta} p(\boldsymbol{Z}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \propto p(\boldsymbol{Z}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{4}$$

In modeling the selenium data, $\boldsymbol{\theta}$ = $[\boldsymbol{\beta}, \tau^2, \omega]$ where ω is a random effect dependent on hyperparameters, σ^2 and ϕ (defined formally following Equation 1). It follows that the joint posterior distribution of interest is $p(\boldsymbol{\beta}, \tau^2, \omega, \sigma^2, \phi|\boldsymbol{Z})$. The posterior predictive distribution $p(Z_0|\boldsymbol{\theta}, \boldsymbol{Z})$ will be used to make predictions by an approximation to the integral in Equation 5.

$$p(\boldsymbol{Z_0}|\boldsymbol{\theta}, \mathbf{Z}) \propto \int_{\boldsymbol{\theta}} \mathbf{p}(\mathbf{Z_0}|\boldsymbol{\theta}, \mathbf{Z})\mathbf{p}(\boldsymbol{\theta}|\mathbf{Z})\mathbf{d}\boldsymbol{\theta} \tag{5}$$

To explicate the theoretical superiority of Bayesian prediction over classical prediction when model parameters are unknown, prior information in $p(\boldsymbol{\theta})$ and information in the data are both used to inform the posterior distribution of $\boldsymbol{\theta}$. Distribution specification allows for uncertainty in the posterior to carry through to the predictions, which are based on the information in the posterior and the information in the data.

The proportionality of $p(\boldsymbol{\theta}|\boldsymbol{Z})$ can be written down, however, it is not a valid probability density function as it does not integrate to one. Without the normalizing constants, it only provides relative frequencies over the domain. When $\int_{\boldsymbol{\theta}} p(\boldsymbol{Z}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$ does not have a closed form solution, $p(\boldsymbol{\theta}|\boldsymbol{Z})$ must be approximated so that $p(\boldsymbol{Z}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ can be scaled to a proper probability density function. Markov Chain Monte Carlo sampling (MCMC) methods are tools to approximate this integral and will be used in analyzing the selenium data. MCMC methods include Gibbs sampling and the Metropolis Hastings algorithm (MH). MCMC methods generate posterior draws of $\boldsymbol{\theta^{(g+1)}}$ that are only dependent on the previous draw, with $\boldsymbol{\theta^{(g)}}$ being independent for all g in independent samplers. Notationally, this means $\boldsymbol{\theta}^{(g+1)}|\boldsymbol{\theta}^{(g)} \perp [\boldsymbol{\theta}^{(g-1)}, \boldsymbol{\theta}^{(g-2)}, ...]$ and $\boldsymbol{\theta^{(g)}} \perp \boldsymbol{\theta^{(g^{'})}}$ for all $g \neq g^{'}$, respectively. It has been shown that generating samples that are all independent or independent except for the previous draw converge to the true posterior distributions when enough iterations are run, enough being relative. For completeness, these two methods are outlined below. Notation corresponds to that of the posterior predictive distribution for the selenium data.

GIBBS SAMPLING

Gibbs is useful for approximating joint posterior distributions when parameters are dependent. Consider the two parameter model where we wish to approximate $p(\boldsymbol{\beta}, \tau^2|\mathbf{Z})$ as in a Bayesian regression model with Gaussian, uncorrelated errors (selenium model assuming no spatial correlation). Joint posterior samples of $\boldsymbol{\theta}^{(g)} = [\boldsymbol{\beta}^{(g)}, \tau^{2(g)}]$ are desired. Samples from the full conditionals, $p(\boldsymbol{\beta}^{(g)}|\tau^{2(g-1)}, \mathbf{Z})$ and $p(\tau^{2(g)}|\boldsymbol{\beta}^{(g)}, \mathbf{Z})$ taken with the full conditionals iteratively updated to form the set of joint posterior samples.

METROPOLIS HASTINGS ALGORITHM

Gibbs sampling is useful when the full conditional distributions take the form of a named probability distribution so that $p(\boldsymbol{Z}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ can be sampled from. When the full conditional distribution does not take take the form of a named distribution, the MH algorithm is useful for sampling from the posterior distribution. MH works by specifying a proposal distribution which controls how the parameters in $\boldsymbol{\theta}$ can move about the parameter space through a random walk. A draw is taken from the proposal distribution and the ratio of the likelihood and proposal using $\boldsymbol{\theta}^{(g+1)}$ and $\boldsymbol{\theta}^{(g)}$ is compared to a draw from the Uniform(0,1) distribution. The new draw $\boldsymbol{\theta}^{(g+1)}$ is accepted ($\boldsymbol{\theta}^{(g+1)} = \boldsymbol{\theta}^{(g+1)}$) if the likelihood ratio is more than the uniform draw and rejected ($\boldsymbol{\theta}^{(g+1)} = \boldsymbol{\theta}^{(g)}$) if the ratio is less. Roughly half the time the ratio should be rejected and the other half accepted, which indicates efficient sampling that enables proper mixing. Refer to Hoff (2009) for a complete layout of the algorithm.

These two methods provide ways of looking at posterior distributions, but are approximations. Hoff (2009) briefly describes why theoretically the distributions resulting from these methods approximate the true posterior distributions. Of primary concern is how many iterations are necessary for approximations to converge to the true posterior distributions. Common diagnostics include sensitivity analyses, where the results are compared from a variety of priors, trace plots to ensure convergence, correlation plots to ensure parameters assumed to be independent are in their posterior form, and posterior distribution plots to ensure the shape resembles that expected.

4 Model Comparisons

Cross validation and Bayes' Factor (BF) are methods to compare posterior models. Minimum MSEP will be used as a cross-validation criterion in the selenium data analysis, however both are explained here. MSEP is used because the classical predictions found in kriging are the best, unbiased, linear predictors, meaning the predictions are unbiased and theoretically have the minimum MSEP of all other predictors. Notationally, MSEP=$E[(Z_i - \hat{Z}_i)^2]$. Cross validation can be implemented by predicting a single left out observation (LOO) or predicting when leaving out k observations (k-fold cross validation). According to Gelman (2004), LOO cross-validation should be used in all Bayesian prediction problems. LOO cross-validation operates by removing an observation, fitting the model, and predicting the observation that was removed. This is done for each observation, meaning that a posterior predictive model is fit n times, if there are n observations in the data set. Each model fit contains n-1 observations to create a posterior predictive distribution for the left out observation. Observations with large prediction errors have high influence in the model. Because the posterior predictive model must be fit n times. LOO cross-validation can be computationally demanding and therefore in this paper, 10-fold cross-validation was arbitrarily used. The idea is similar to LOO. The data were randomly separated into ten subsets. The algorithm is then run ten times, each time leaving out one of the subsets and using the information in the "left in" observations to predict the left out subset. MSEP is computed with the mean posterior predictions using the left out data.

BF is a ratio of the probability of the data given two models specified. Model parameters are reparameterized to include indicator variables that control model l, Λ_l, where l spans the number of models considered. The model considered for predicting selenium concentrations is given in Equation 6.

$$Z(s) = \mu + x(s)\beta + W(s, \phi, \sigma^2) + \epsilon(s) \tag{6}$$

It is then reparameterized as:

$$Z(s) = \mu + \lambda_1 x(s)\beta + \lambda_2 W(s, \phi, \sigma^2) + \epsilon(s) \tag{7}$$

The three models compared using the Λ notation, where $\mathbf{\Lambda} = [\lambda_1, \lambda_2]$, are $\mathbf{\Lambda}_{\boldsymbol{A}} = [1, 1]$; $\mathbf{\Lambda}_{\boldsymbol{B}} = [1, 0]$; $\mathbf{\Lambda}_{\boldsymbol{C}} = [0, 1]$.

$\mathbf{\Lambda}_{\boldsymbol{A}}$ represents the model with spatial covariance structure and zone main effects, $\mathbf{\Lambda}_{\boldsymbol{B}}$ represents the nugget only variance model with zone as fixed effects, and $\mathbf{\Lambda}_{\boldsymbol{C}}$ represents the mean only model with spatial covariance structure. Formally, the Bayes' Factor is the likelihood of model $\mathbf{\Lambda}_{\boldsymbol{l}}$ relative to model $\mathbf{\Lambda}_{\boldsymbol{l'}}$.

$$BF = \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}, \mathbf{\Lambda}_{\mathbf{l}})}{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}, \mathbf{\Lambda}_{\mathbf{l'}})} \tag{8}$$

These methods do not indicate where models break down (Gelman, 2004), but can be useful for model checking during the exploratory stage.

5 Selenium Data

A map of Benton Lake is shown in Figure 3. The lake is divided into nine units: Unit 1, Unit 2, Unit 3, Unit 4a, Unit 4b, Unit 4c, Unit 5, Unit 6, and the Interunit Canal. Within each unit are zones representing the type of water. Water zones considered for this analysis include: flooded, saturated, and intermittent. Waterlevels vary throughout the seasons as well as throughout the lake. Flooded zones have pumped and irrigation water flow, saturated have standing water, and intermittent zones have combinations of pumped, irrigated, and standing water. Generally mid-July through August Units 3-6 are dry. Water throughout the lake comes from combinations of pumped water from Muddy Creek, standing water, and natural run off.

Figure 4 is a plot of the observed distribution of selenium concentrations, which were highly skewed towards larger values. Selenium concentrations were recorded in micrograms per Liter (mg/L). A transformation of the response was not done as the assumption of normality in the residuals was reasonably met.

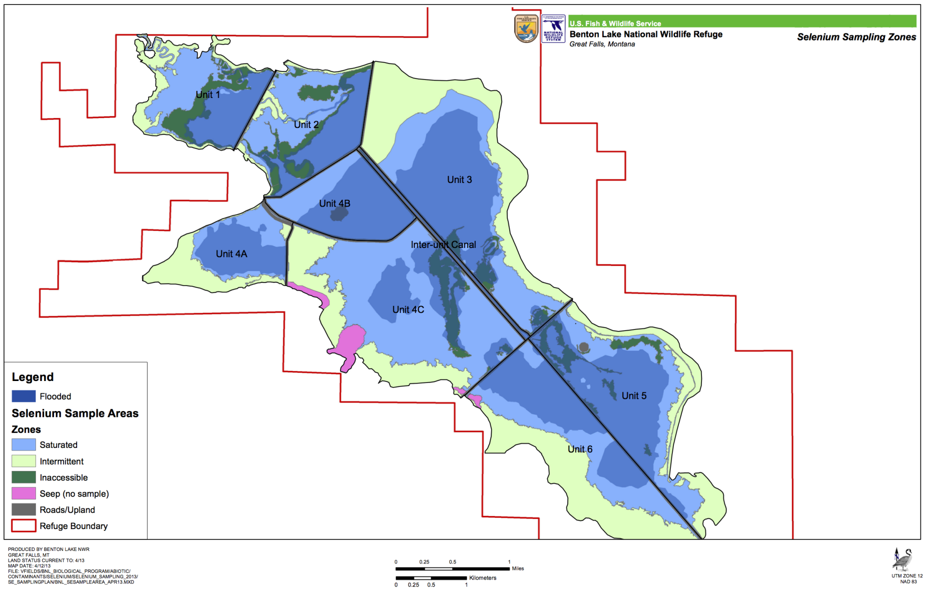


Figure 3: Map of Benton Lake

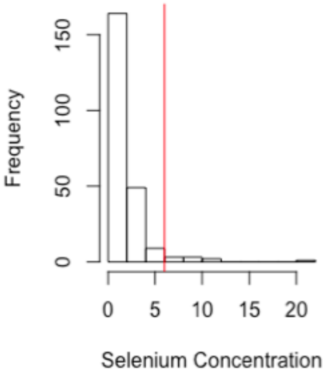


Figure 4: Histogram of Observed Selenium Concentrations in mg/L

5.1 Sampling at Benton Lake

Data were collected during summer months of 2013-2014 using two separate stratified Generalized Random Tessellation Sampling (GRTS) algorithms. In both cases, units were the strata. Input into the GRTS algorithm for sediment were proportionality constants set to over-sample zones with more standing water. Over-sampling was done to reduce the variability in those zones. Sediment was sampled in all three zones. Sampling was done in units 2, 4C, 5, 6, and the Interunit Canal were sampled in 2013 and in units 1, 4A, and 4B in 2014. After data collection, adjustments such as weighting based on area and realized sample size were done to the final data set.

Selenium concentrations were recorded in water, sediment, roots, invertebrate, and birds. The focus in this analysis is predicting sediment selenium concentrations in saturated, intermittent, and flooded water zones in all units after accounting for spatial correlations. Sampling month and year were also recorded but are confounded with unit information.

5.2 Sampling Criterion

Diggle et al. (2003) state that placing sampling locations in a regular grid is efficient for prediction when all model parameters are known. When parameters must be estimated, models with precise parameter estimates do not necessarily predict efficiently in a spatial setting. As described previously, classical kriging methods compute predictions assuming parameters are known and therefore do not incorporate uncertainty in parameter estimates in to the prediction efficiency. Diggle and Lophaven (2006) suggested the spatially averaged posterior prediction variance as a Bayesian design criteria to the geostatistical prediction problem. The design criteria are stated formally below for both the retrospective and prospective spatial design scenarios. Note the difference lies in the prospective design, where $Z(s_0)$ has not been observed and thus must be approximated through the expected value shown.

Retrospective Criterion:

$$\bar{\nu} = \int_A Var[Z(s_0)|Z]ds \tag{9}$$

Prospective Criterion:

$$E[\bar{\nu}] = \int_A E_{Z|\boldsymbol{\theta}_0}[Var(Z(s_0)|Z)]ds \tag{10}$$

The retrospective criterion in Equation 9 was applied in predicting a regular grid of 81 locations over the lake with the goal of removing 18 sampling locations. The grid contained locations in all three zones and in all units except the inter-unit canal. The criterion was applied to the top spatial and non-spatial models for comparison. Note that although the criteria were developed for the spatial setting, they can easily be adapted for the non-spatial setting.

6 Spatial Model

Two parameterizations of the selenium data model could be considered. Incorporating the random effect ω into the analysis allowed for simplification in the posterior computations.

$$\boldsymbol{Z}(s) = \boldsymbol{X}(s)\beta + \boldsymbol{\epsilon}(s) \tag{11}$$

where $\boldsymbol{\epsilon}(s) \sim \text{MVN}(\tau^2 I + \sigma^2 R(h|\phi))$

$$\boldsymbol{Z}(s) = \boldsymbol{X}(s)\boldsymbol{\beta} + \boldsymbol{\omega}(s, \phi, \sigma^2) + \boldsymbol{\epsilon}(s) \tag{12}$$

where $\boldsymbol{\epsilon}(s) \sim MVN(\mathbf{0}, \tau^2 I)$ and $\boldsymbol{\omega}(s, \phi, \sigma^2) \sim MVN(\mathbf{0}, \sigma^2 R(h|\phi))$

Notation is defined as follows: $\boldsymbol{Z}(s)$ is the sediment selenium concentration at location s; $\boldsymbol{X}_{231\times 3}$ is the design matrix with binary indicators for whether each location is in a saturated, flooded, or intermittent zone; $\beta_{3\times 1}$ is the covariate matrix for the effects of zone on mean selenium concentration; $\boldsymbol{\omega}(s, \phi, \sigma^2)$ ~N(0,$\sigma^2 R(h|\phi)$) is the random spatial variation assuming an exponential covariance function; $R(h|\phi) = exp[-\frac{h}{\phi}]$ is the exponential correlation function for distance h; and $\boldsymbol{\epsilon}(s)$ is the random variation not accounted for by the model.

7 Computation

Priors used were suggested by Banerjee, Carlin, and Gelfand (2004). The nugget and partial sill have inverse gamma priors, while the fixed and random effects were assumed to have multivariate normal prior distributions. The range prior was assumed to be uniform over the observed range of distances in the data set. All priors were uninformative and marginally conjugate.

Full conditional posterior distributions were approximated for $\boldsymbol{\beta}$, τ^2, and ω through Gibbs sampling methods and the MH algorithm was used for approximating the posterior of ϕ. ω is treated as a random effect, with hyperparameters σ^2 and ϕ in the covariance structure. Hyperparameters are parameters that indirectly influence predictions through the direct parameters.

For each set of posterior parameter estimates a sample of ω is obtained. This is known as composition sampling. Prediction for z_0 also occurs via composition sampling. The algorithm to approximate the joint posterior follows.

1. Sample from $p(\tau^{2(S)}|\omega^{(S-1)}, \boldsymbol{\beta}^{(S-1)}, \boldsymbol{Z}) \sim InvGam(\frac{n+\nu_o}{2}, \frac{\Sigma[y-x\boldsymbol{\beta}-\omega]^2+\nu_o\tau_o^2}{2})$
2. Sample from $p(\boldsymbol{\beta}^{(S)}|\omega^{(S-1)}, \tau^{2(S-1)}, \boldsymbol{Z}) \sim MVN([\boldsymbol{x}^T\tau^{-2}\boldsymbol{x}+\Sigma_o^{-1}][\boldsymbol{x}^T\tau^{-2}\boldsymbol{Z}+\boldsymbol{x}^T\tau^{-2}\omega+\boldsymbol{\beta}_o\Sigma_o^{-1}], [\boldsymbol{x}^T\tau^{-2}\boldsymbol{x}+\Sigma_o]^{-1})$
3. Sample from $p(\omega^{(S)}|\boldsymbol{\beta}^{(S)}, \tau^{2(S)}, \sigma^{2(S-1)}, \phi^{(S-1)}, \boldsymbol{X}, \boldsymbol{Z}) \sim MVN([\tau^{-2}+(\sigma^2 R(h|\phi))^{-1}]^{-1}[\tau^{-2}\boldsymbol{Z}-\boldsymbol{x}^T\tau^{-2}\boldsymbol{\beta}])$
    (a) Sample from $p(\sigma^{(S)}|\phi^{(S-1)}, \omega^{(S)}, \boldsymbol{Z}) \sim InvGam(\frac{n+\nu_{oo}}{2}, \frac{\omega^T R(h|\phi)^{-1}\omega+\nu_{oo}\sigma_o^2}{2})$
    (b) Sample from $p(\phi^{(S)}|\sigma^{2(S)}, \omega^{(S)}, \boldsymbol{Z})$ using MH

Iterate until convergence. Posterior predictions are found by independently sampling $Z(S_0)$ from Equation 13.

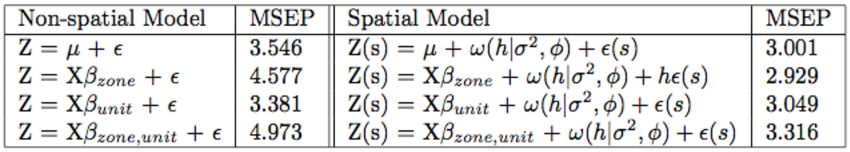
$$p(Z(S_0)|\boldsymbol{\beta}, \tau^2, \omega, Z) \sim MVN(\boldsymbol{X}(S_0)^T\hat{\boldsymbol{\beta}} + \sigma^T\Sigma^{-1}[Z(S) - X(S)\hat{\boldsymbol{\beta}}], \sigma_0 - \sigma^T\Sigma^{-1}\sigma) \tag{13}$$

Predictions from (13) are made for each posterior set of parameters past the burn-in period. The burn-in period is necessary because it may take a number of iterations through the algorithm before posterior parameters are in the most likely areas of the parameter space. Posterior parameters in the burn-in period are discarded, meaning they are not used for posterior parameter inference nor for posterior prediction.

8 Results

8.1 Cross-Validation Results

As previously described, the loss function used in cross-validation was the minimum MSEP. The tables below show the four spatial and non-spatial models that were considered with different fixed effects: the intercept only model, models with only zone and only unit, and models with both zone and unit. The top spatial model included zone as a predictor, and the top non-spatial model included unit as a predictor. Note that all the spatial models out perform the non-spatial models.



Non-spatial Model	MSEP	Spatial Model	MSEP
$Z = \mu + \epsilon$	3.546	$Z(s) = \mu + \omega(h\|\sigma^2, \phi) + \epsilon(s)$	3.001
$Z = X\beta_{zone} + \epsilon$	4.577	$Z(s) = X\beta_{zone} + \omega(h\|\sigma^2, \phi) + h\epsilon(s)$	2.929
$Z = X\beta_{unit} + \epsilon$	3.381	$Z(s) = X\beta_{unit} + \omega(h\|\sigma^2, \phi) + \epsilon(s)$	3.049
$Z = X\beta_{zone,unit} + \epsilon$	4.973	$Z(s) = X\beta_{zone,unit} + \omega(h\|\sigma^2, \phi) + \epsilon(s)$	3.316

8.2 Posterior Prediction Results

Posterior prediction maps are shown in Figure 5 for both the top spatial and the non-spatial model, as well as a map for the observed predictions. The spatial map more closely reflects the observed selenium concentrations and provides a more informative map by capturing the hot spots of selenium contamination. The non-spatial map predicts the same concentration for each unit and zone combination and does not show any locations contain extremely high selenium concentrations as were observed in the data.

8.3 Retrospective Design Criteria Results

This section will provide a response to the question concerned with which 18 sampling locations should be removed if future sampling efforts must be reduced. Diggle and Lophaven's (2006) retrospective criteria was applied to the top spatial model and the top non-spatial model. Recall the top spatial model had zone as a predictor while the top non-spatial model had unit as a predictor.

Eighteen locations were removed from the data set to compare predictive accuracy. Observed location were removed one by one from the data set and 1000 selenium concentration predictions were made at each

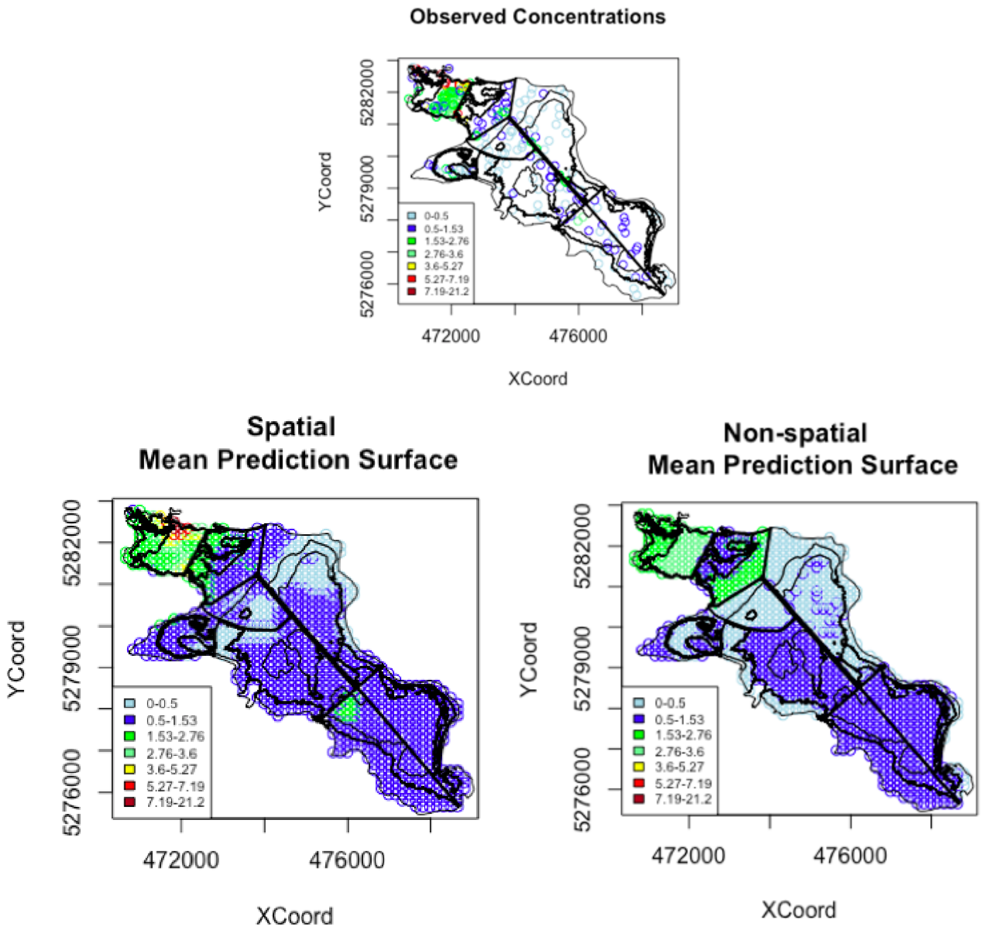


Figure 5: Concentration Maps

point in a spatial grid of 81 locations over the lake. Predictions were made on the grid when each observed location was discarded. The location that when left out resulted in the maximum posterior prediction variance was removed, and this process was repeated while sequentially removing points. The spatial model used was that in Section 6 and the kriging algorithm was explained in Section 7.

Figure 6 shows in red the 18 locations that were removed based on the retrospective criteria in both the spatial and non-spatial models. Patterns in locations removed were similar between the two models.

Both spatial and non-spatial maps suggest removal of locations in Units 1 and 2, where observed concentrations were the highest and where locations were over-sampled. In the non-spatial case, locations to be removed are concentrated in the flooded zone of Unit 1. The spatial case has the majority of locations to be removed in Unit 1. The maps show a balance of close observations and spaced locations to be removed, which reflects the criteria properties. An initial concern with the criterion is that it may favor smooth sur-

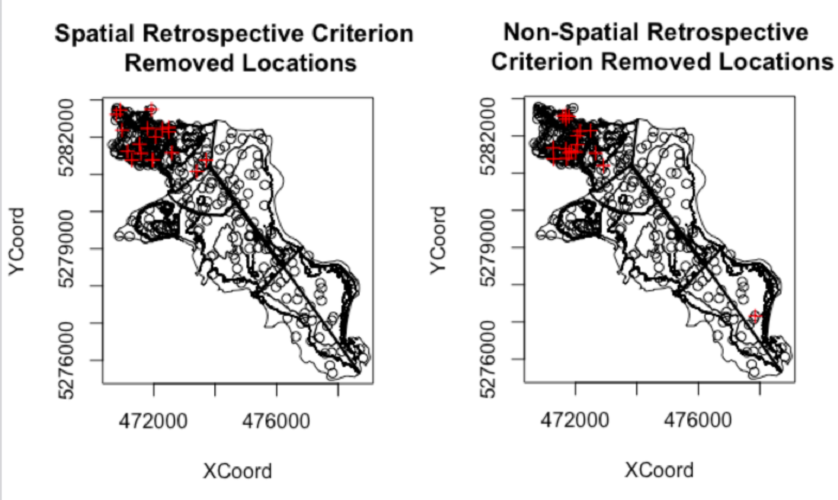

Spatial Retrospective Criterion Removed Locations
Non-Spatial Retrospective Criterion Removed Locations
YCoord
5276000
5279000
5282000
472000
476000
XCoord

Figure 6: Sampling Locations Based on Retrospective Criterion

faces with little variability. However, if the true distribution is volatile, the spatial prediction map should resemble that. Future work may involve examining variations of this criterion. Removing more locations, or adding locations, and comparing the resulting prediction maps are also potential extensions. One caveat concerning the criterion is the code run time is quite expensive, particularly in the spatial case.

9 Conclusions

Continued monitoring of selenium concentrations over Benton Lake is important to maintain the health and presence of lake waterfowl. Given that the selenium concentrations were collected over Benton Lake and may be spatially dependent, the accuracy of selenium concentration predictions could be improved by incorporating the spatial aspects of the lake into the analysis. By kriging in a Bayesian framework, predictions incorporate the uncertainty in the covariance parameters. The range and partial sill parameters were highly correlated, making convergence an issue individually for these parameters. Individual convergence was not a problem, however, because the parameters converged when combined. Recall that these were hyperparameters and so only the convergence of their combination in ω is of concern. The top MSEP spatial model had zone as a predictor and the top MSEP non-spatial model had unit as predictor. This indicates that the spatial covariance structure and unit share information. Predictions were made using both spatial and non-spatial models. The spatial model was able to predict higher selenium concentrations where the observed concentrations were highest. The spatial predictions were more reflective of the observed data. Because management is most concerned with the locations of high selenium concentrations, basing conclusions on non-spatial predictions would misrepresent the health of the lake.

Lake management plan to sample selenium data in the future as a part of continued monitoring efforts. The retrospective design criteria suggested removal of sampling locations in Units 1 and 2, which were purposefully over-sampled using the stratified GRTS algorithm previously. Further examination of the stratified GRTS algorithm is needed to make sampling design recommendations. The stratified GRTS sampling design over-sampled Units 1 and 2 to minimize the variance in those units, which would be counteracted by removal of some of those locations as suggested by the retrospective design criteria. Incorporation of the prospective design criteria to expand the sampling frame may also be of interest to explore in future sampling efforts.

This paper lends itself to extensions in sampling design, among others. All models were fit using the exponential correlation function because it is common. Further exploration of variograms and predictions under different correlation functions may be insightful. Recall that selenium concentrations were collected in water, sediment, roots, invertebrate, and birds. Co-kriging is a multivariate version of kriging and is most useful in the heterotopic case. Heterotopy occurs when at least one variable is over-sampled and predictions can be interpolated for the under-sampled variable. While preliminary examination of co-kriging in the selenium data suggested response variables were only weakly correlated, co-kriging results may be more accurate. Modeling the accumulation of selenium through the trophic levels would be very informative. Finally, the isotropic assumption may not be a reasonable, which is illustrated in the appendix.

10 Appendix: Anisotropy

The variogram definition assumes the spatial process is isotropic, or that the spatial correlation is only a function of the Euclidean distance (h_{ij}) between two arbitrary points *i* and *j*. Benton Lake has water flow and pumping systems which may contribute to isotropy violations. The directional semi-variogram indicated that the spatial process does exhibit anisotropic behaviors, particularly in the 135 degree angle direction. Different types of anisotropy can occur. A spatial process that attains the same sill in all directions but with practical ranges depending on direction is said to exhibit geometric (range) anisotropy. Zonal anisotropy occurs when different sills are seen in different directions (Schabenberger and Gotway, 151-2).

A linear transformation of coordinates is a solution to geometrical anisotropy. Allowing semi-variograms nested within "angle classes" is one proposed solution to zonal anisotropy. Anisotropy can sometimes be eliminated by using distance metrics that are transformably Euclidean through methods such as Multi-Dimensional Scaling. Suppose an island existed in the middle of Benton Lake and we have one sampling location (k) directly below another sampling location (l), but on opposite sides of the island. The shortest distance in terms of waterflow between the two points would wrap around the island. By using distances measures that more accurately reflect the geography of the lake, we may correct for anisotropy. Figure 7 is the empirical directional semi-variogram. Note that it does not show variability in the estimation of the directional semi-variograms.

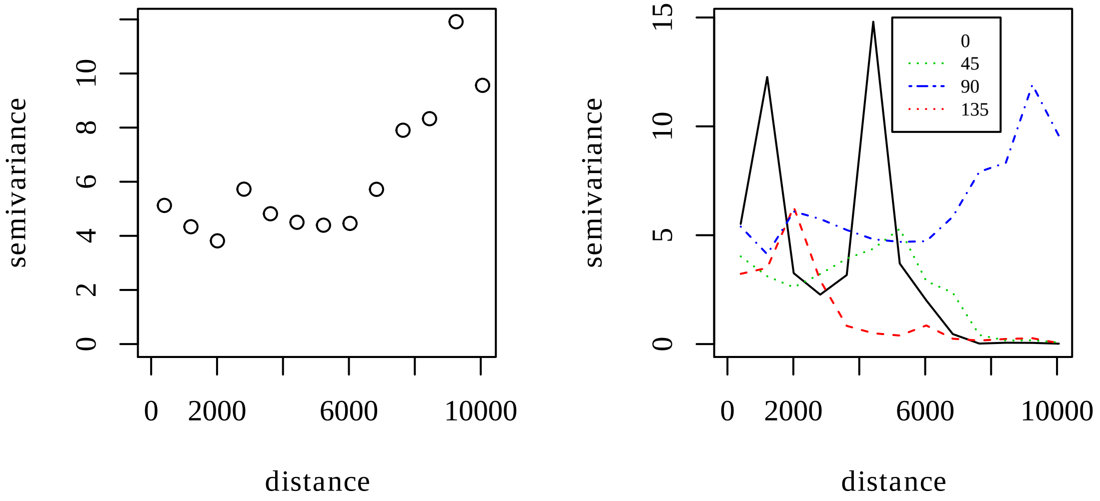


Figure 7: Empirical Directional Variograms Using geoR

Zonal anisotropy is suggested by the directional semi-variogram. However, nesting semi-variograms within angle classes will not necessarily result in positive definite covariance matrices. Addressing the anisotrophic nature of the data is a possible route for future work.

11 References

Banerjee, S., Carline, B.P., & Gelfand, A.E. (2004). *Hierarchical Modeling and Analysis for Spatial Data.* Boca Raton, FL: CRC Press LLC.

Diggle, P. & Lophaven, S. (2006). Bayesian geostatistical design. *Scandinavian Journal of Statistics, 33*(1):53-64.

Diggle, P., Ribeiro Jr, P.J., & O.F.C. (2003). An introduction to model-based geostatistics. *Spatial statistics and computational methods.*

Friend, M.& Franson, J.C. (1999). Field manual of wildlife diseases: general field procedures and diseases of birds. *U.S. Department of the Interior. U.S. Geological Survey.* Chapter 44. Retrieved from `https://www.nwhc.usgs.gov/publications/field_manual/chapter_44.pdf`.

Gelfand, A.E. & Banerjee S. (2009). *Multivariate Spatial Process Models.* Chapter 28.

Gelman A. (2004, Nov 8). Cross-validation for Bayesian multilevel modeling. Retrieved from `http://andrewgelman.com/2004/11/08/crossvalidation/`

Hoff, P.D. (2009). *A First Course in Bayesian Statistical Methods.* New York: Springer.

Irvine, K. & Fields, V. (2014). *Selenium Sampling Protocol for Baseline Assessment at Benton Lakes NWR.*

Isaaks, E.H. & Srivastava, R.M. (1989). *An Introduction to Applied Geostatistics.* New York: Oxford University Press.

Le, N.D. & Zidek, J.V. (1992). Interpolation with uncertain spatial covariances: a Bayesian alternative to kriging. *Journal of Multivariate Analysis 43*. 43:351-374.

Pilz and Spock (2007). Why do we need and how should we implement Bayesian kriging methods. *Stochastic Environmental Research and Risk Assessment, 22*:621-632.

Rosencrantz, J. & Irvine, K.M. (2015). *Data analysis report: baseline assessment of selenium concentrations at Benton Lake NWR (2013-2014).*

Schabenberger, O. & Gotway, C.A. (2005). *Statistical Methods for Spatial Data Analysis*. Boca Raton, FL: Taylor & Francis Group, LLC.

Shortridge, A. (2013). Retrieved from `https://msu.edu/~ashton/classes/866/notes/lect17/ni_cokrig.R`

U.S. Fish & Wildlife Service (March 2012). *Draft Comprehensive Conservation Plan and Environmental Assessment: Benton Lake National Wildlife Refuge Complex.* Lakewood, Colorado: U.S. Department of the Interior, Fish and Wildlife Service, Mountain-Prairie Region. 366 p.

Wackernagel, H. (1995). *Mutlivariate Geostatistics*. New York: Springer.