5-1-2010

Applying Multiple Imputation with Geostatistical Models to Account for Item Nonresponse in Environmental Data

Breda Munoz
RTI International, breda@rti.org

Virginia M. Lesser
Oregon State University, lesser@science.oregonstate.edu

Ruben A. Smith
Oregon State University, RASmith@cdc.gov

Follow this and additional works at: http://digitalcommons.wayne.edu/jmasm

Part of the Applied Statistics Commons, Social and Behavioral Sciences Commons, and the Statistical Theory Commons

Applying Multiple Imputation with Geostatistical Models to Account for Item Nonresponse in Environmental Data

Breda Munoz   Virginia M. Lesser   Ruben A. Smith
RTI International,   Oregon State University,
RTP, NC   Corvallis, OR

Methods proposed to solve the missing data problem in estimation procedures should consider the type of missing data, the missing data mechanism, the sampling design and the availability of auxiliary variables correlated with the process of interest. This article explores the use of geostatistical models with multiple imputation to deal with missing data in environmental surveys. The method is applied to the analysis of data generated from a probability survey to estimate Coho salmon abundance in streams located in western Oregon watersheds.

Key words: Environmental surveys; missing data; nonresponse.

Introduction

Environmental surveys are often subject to missing data. An entire observational unit, such as a sampling site, may be missing; conversely, one or a few variables for an observational unit may be missing. These types of missing data are referred to in the survey literature as either unit or item nonresponse, respectively (Lessler & Kalsbeek, 1992). Causes for missing data in environmental studies include failure of the measuring instruments (resulting in unit and/or item nonresponse), inaccessibility of the site (unit nonresponse), and data lost or damaged (unit and/or item nonresponse). A multiple

Breda Munoz is a Senior Research Statistician at RTI International. Email her at: breda@rti.org. Virginia M. Lesser is a Professor of Statistics and Director of the Survey Research Center, Oregon State University. Email her at: lesser@science.oregonstate.edu.   Ruben   A. Smith currently serves as a Mathematical Statistician for the Applied Sciences Branch, Division of Reproductive Health, National Center for Chronic Disease Prevention and Health Promotion, Centers for Disease Control and   Prevention.   Email   him   at: RASmith@cdc.gov.

imputation approach is proposed for handling missing item nonresponse data that occurs at one sample point in time data in environmental surveys.

Further study of the magnitude and factors resulting in missing data is necessary to interpret the data that has been collected. The impact of missing data in the estimation stage depends on the missing data mechanism or random process leading to it and also on whether the observed missingness is related to any variables in the dataset (Little & Rubin, 2002). Specifically, the impact of nonresponse on survey error depends on how the missing data occurred, the percent of nonresponse, and the parameters to be estimated (Lessler & Kalsbeek, 1992; Little & Rubin, 2002).

Let $\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_{obs} \\ \mathbf{Y}_{miss} \end{bmatrix}$ denote the matrix of

complete data corresponding to observations of a random process, where $\mathbf{Y}_{miss}$ and $\mathbf{Y}_{obs}$ denote the missing and observed components of $\mathbf{Y}$, respectively. Missing data can be classified as missing completely at random (MCAR), missing at random (MAR), and nonignorable or informative nonresponse (Little & Rubin, 2002). Data is called MCAR if the observed data ($\mathbf{Y}_{obs}$) can be considered a representative sample of the population, that is, the missingness does not depend on the response ($\mathbf{Y}$) or other variables

measured at the site or regional level. Under this assumption, valid results are obtained when analysis techniques developed for complete data sets are performed on the observed data ($\mathbf{Y}_{obs}$) (Little & Rubin, 2002; Lessler & Kalsbeek, 1992; Lohr 2001).

When the missingness does not depend on the unobserved response but depends only on observed values of auxiliary variables, then the missing data mechanism is known as MAR. This is also referred to as ignorable nonresponse. A model for this nonresponse mechanism can be formulated and incorporated into either design-based or model-based analysis techniques to explain and account for the nonresponse. For example, among the design based approaches, weighting methods - such as a weighting class adjustment - can be used to produce estimates to adjust for the nonresponse (Lohr, 2001).

Finally, if the probability of nonresponse depends on the response and cannot be completely explained by the values of the auxiliary variables, then the nonresponse is nonignorable (Little & Rubin, 2002). Models for the nonignorable missing mechanism are usually more complicated than models for ignorable nonresponse because they depend on the unobserved values.

Recognized approaches to handle missing data problems include deletion of the records, hot or cold deck imputation (Chen & Shao, 1999), substitution, parametric and semi parametric modeling techniques (Rotnitzky, et al., 1998; Robins, 1995), and multiple imputation (Little & Rubin, 2002). More innovative techniques include neural networks (Gupta & Lam, 1996), Bayesian models (Sebastiani & Ramoni, 2000; Kleinman, et al., 1998), maximum likelihood estimation approaches (Little & Schluchter, 1985; Schneider, 2001; Little 1982), and linear and generalized linear model imputation assuming nonignorable missing data (Greenless, et al., 1982; Baker & Laird, 1988; Ibrahim, 1990).

Most of these approaches result in a single imputation of the missing data, generating one complete data set. Analyses are then applied to the complete data set. The results of data analysis on single imputation data neither reflect the missing-data uncertainty nor on the consequence of imputation. Furthermore, analyses based on a single imputation may result in under-estimated standard errors, incorrect p-values, and high Type I error rates. This problem increases as the rate of missing information and the number of model parameters increases (Schafer & Olsen, 1998).

Another method to deal with nonresponse is the well-known multiple imputation (MI) methodology. This method incorporates the uncertainty of the missing data into the inference (Rubin, 1987). MI replaces each missing item with m values from a distribution of likely values. This process generates m complete data sets on which the same analysis procedure is performed. The final inferences combine the individual estimates obtained from the m complete data sets, thus allowing a researcher to account for the variability due to imputation and to analyze the data using standard techniques and software available for complete datasets (Schafer & Olsen, 1998; Schafer, 1997).

To account for the spatial variability inherent in environmental monitoring programs, a geostatistical model is considered as the imputation model. Kriging and other stochastic predictors for spatial data are referred to as geostatistical models in the spatial statistics literature (Diggle, et al., 1998). Kriging is a well-known technique for spatial interpolation that generates predictions for the unobserved values of the spatial random process at the unvisited sites. The kriging estimator is a minimum error weighted linear predictor that assumes a Gaussian distribution for the random process and a model for the variance-covariance matrix (see Cressie, 1993 for more details). Diggle, et al. (1998) extended the concept of geostatistical models to non-Gaussian situations within the framework of generalized linear models (see McCullagh & Nelder, 1989 for more details on generalized linear models).

In this study MI is explored using geostatistical models for handling missing data in environmental surveys for item nonresponse. An advantage of using geostatistical models in MI is the possibility of imputing missing values for both continuous and discrete environmental variables.

Multiple Imputation

Multiple imputation (MI) is a simulation-based approach analyzing missing data that incorporates the uncertainty of missing data into the inference (Rubin, 1987; Rubin, 2002, Harrel & Zhou, 2007). In MI, each missing datum is replaced by a set of $m > 1$ simulated plausible values from their predictive distribution creating m complete data sets. Each complete data set is analyzed separately. The final estimator is the average of the estimators obtained in the individual analyses. The variability introduced by the m analyses is combined with an estimate of the sample variance to provide a single variability measure for the parameters of interest (Schafer, 1997).

Following Rubin (1996) and Schafer (1997), $\hat{Q}_i$ is denoted as a point estimate (e.g., an estimate of salmon abundance in the State of Oregon) of the parameter of interest, Q (e.g., salmon abundance in the State of Oregon), where $i = 1,\ldots,m$. Let $\hat{U}_i$ denote the estimated variance of $\hat{Q}_i$ obtained from the ith individual analysis, $i = 1,\ldots,m$. The overall point estimate is obtained as

$$\bar{Q}_m = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i$$

and the overall within imputation variance estimate is given by

$$\bar{U}_m = \frac{1}{m}\sum_{i=1}^{m}\hat{U}_i.$$

The between imputation variance estimate, defined as

$$B_m = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{Q}_i - \bar{Q}_m)^2,$$

reflects the extra inferential uncertainty due to the imputation of the missing data. The total variance of $\bar{Q}_m$, is calculated as

$$T_m = \bar{U}_m + (1+m^{-1})B_m.$$

A confidence interval for the parameter of interest, Q, can be obtained as: $\bar{Q}_m \pm t_{df}\sqrt{T_m}$, where t_{df} is the df-quantile of the t-Student distribution, and

$$df = (m-1)\left(1+\frac{m\bar{U}_m}{(m+1)B_m}\right)^2$$

denotes the corresponding degrees of freedom (Barnard & Rubin, 1999).

To ensure valid inferences when using MI, researchers must assume a mechanism of missingness, a model for the complete data $f(\mathbf{Y}_{\text{miss}}, \mathbf{Y}_{\text{obs}})$, and a prior distribution for the parameters of the model. A MAR mechanism for the missing data was assumed and imputations for $\mathbf{Y}_{\text{miss}}(\mathbf{s})$ from the posterior predictive distribution of the missing data $f(\mathbf{Y}_{\text{miss}} \mid \mathbf{Y}_{\text{obs}})$ were generated. The posterior predictive distribution of $\mathbf{Y}_{\text{miss}}$ can be obtained by Bayes's Theorem as

$$f(\mathbf{Y}_{\text{miss}} \mid \mathbf{Y}_{\text{obs}}) = \int_{\Theta} f(\mathbf{Y}_{\text{miss}} \mid \mathbf{Y}_{\text{obs}}, \underline{\boldsymbol{\theta}}) f(\underline{\boldsymbol{\theta}} \mid \mathbf{Y}_{\text{obs}}) \mathrm{d}\underline{\boldsymbol{\theta}}$$

(1)

where $\underline{\boldsymbol{\theta}}$ represents the vector of parameters of the imputation model for the complete data (e.g., $f(\mathbf{Y}_{\text{miss}}, \mathbf{Y}_{\text{obs}})$), $f(\mathbf{Y}_{\text{miss}} \mid \mathbf{Y}_{\text{obs}}, \underline{\boldsymbol{\theta}})$ is the posterior predictive distribution of $\mathbf{Y}_{\text{miss}}$ given $\underline{\boldsymbol{\theta}}$ and the observed data (e.g., $\mathbf{Y}_{\text{obs}}$), $f(\underline{\boldsymbol{\theta}} \mid \mathbf{Y}_{\text{obs}})$ is the posterior distribution of $\underline{\boldsymbol{\theta}}$ given the observed data (e.g., $\mathbf{Y}_{\text{obs}}$), and Θ denotes the parameter space (Schafer, 1997; Little & Rubin, 2002). It can be shown that $f(\underline{\boldsymbol{\theta}} \mid \mathbf{Y}_{\text{obs}}) \propto L(\underline{\boldsymbol{\theta}} \mid \mathbf{Y}_{\text{obs}})\pi(\underline{\boldsymbol{\theta}})$, where $L(\underline{\boldsymbol{\theta}} \mid \mathbf{Y}_{\text{obs}})$ is the observed data likelihood, and $\pi(\underline{\boldsymbol{\theta}})$ is an assumed prior for $\underline{\boldsymbol{\theta}}$.

The resulting posterior predictive density of $\mathbf{Y}_{\text{miss}}(\mathbf{s})$, $f(\mathbf{Y}_{\text{miss}} \mid \mathbf{Y}_{\text{obs}})$, may not be a recognizable distribution. Whether the

distribution is recognizable depends on the assumptions adopted for the conditional distributions and the priors. In some cases $f(\mathbf{Y}_{\text{miss}} \mid \mathbf{Y}_{\text{obs}})$ can be written as the product of conditional and marginal known densities.

In other cases, only an approximation can be obtained by means of computational analyses such as the Markov Chain Monte Carlo (MCMC) methods, which consist of a collection of techniques for drawing pseudo random values from approximate or exact predictive distributions (Schafer, 1997; Gelman, et al., 1995). These methods include the Gibbs sampling algorithm, data augmentation methods, the Metropolis-Hasting algorithm and a series of hybrid algorithms.

MCMC is one of the primary methods for generating MI's in nontrivial problems. MCMC is discussed in the literature for parameter simulation by creating a dependent sequence of random draws of parameters from Bayesian posterior distributions under complicated parametric models (Gilks, et al., 1996). However, in MI-related applications MCMC is used to create a small number of independent draws of the missing data from a predictive distribution; these draws are then used for multiple-imputation inference (Schaffer, 1997; Rubin, 2003).

The MCMC methods generate sequential realizations of the posterior predictive density of $\mathbf{Y}_{\text{miss}}(\mathbf{s})$, $\{\mathbf{Y}^{(t)}_{\text{miss}}(\mathbf{s}) : t = 1, 2, ...\}$. Each term in the sequence (e.g., $\mathbf{Y}^{(t)}_{\text{miss}}(\mathbf{s})$) depends on the preceding one, and the limiting distribution of the sequence converges to the posterior predictive density of $\mathbf{Y}_{\text{miss}}(\mathbf{s})$. These methods are attractive because the convergence of the MCMC algorithms does not require that the starting values for the distribution of $\mathbf{Y}_{\text{miss}}(\mathbf{s})$ to be actual realizations of the posterior predictive density of $\mathbf{Y}_{\text{miss}}(\mathbf{s})$. Close starting values are recommended, however, to assure faster convergence (Gelman & Rubin, 1992; Shafer, 1997). Finally, the posterior predictive mean is defined as the expected value of the posterior predictive distribution of $\mathbf{Y}_{\text{miss}}$, $E(\mathbf{Y}_{\text{miss}} \mid \mathbf{Y}_{\text{obs}}, \boldsymbol{\theta})$. Diagnostic assessment of the convergence of the MCMC chains can be made using the convergence diagnostics of Geweke (1992) and Heidelberger and Welch (1983). Both convergence diagnostics assess the stationary distribution assumption of the chain.

Geostatistical Models

In environmental science, researchers use geostatistical techniques to model environmental processes that evolve in space and time. Geostatistical models are proposed (Handcock & Stein, 1993; Le & Zidek, 1992; Diggle, et al., 1998; Diggle & Ribeiro, 2002; Christensen & Waagepetersen, 2002) in conjunction with MI (Schafer, 1997; Rubin, 1996; Little & Rubin, 2002) to handle missing data in environmental surveys.

An environmental process of interest is generated by an unobserved spatial random field, Y, defined over a continuous region of interest, $D \subset R^2$. $Y(\mathbf{s})$ denotes the outcome of the process of interest at location $\mathbf{s}$, and $\mathbf{s}$ be the coordinates of a site or point in D, $\mathbf{s} \in D$. The observed data is collected from a finite number of sites, $S = \{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n\}$. The sites can be selected either from a probability or a non-probability sampling design. Missing data occurrs in n_1 of the n sites, with $n_1 < n$.

For each point $\mathbf{s}$ in D, the random process of interest, Y, has a distribution with mean $\mu(\mathbf{s})$, $E[Y(\mathbf{s})] = \mu(\mathbf{s})$. A continuous differentiable function g of μ exists, such that $g[\mu(\mathbf{s})] = \mathbf{X}\beta + Z(\mathbf{s}) + \varepsilon(\mathbf{s})$, where $\mathbf{X}$ is a vector of covariates, correlated with the random process Y, that is available at the site level, and β is a vector of unknown parameters. Z denotes a spatial random effect with mean 0 and its variance-covariance matrix $\sigma_Z^2 \mathbf{R}(\boldsymbol{\theta})$. $\mathbf{R}(\boldsymbol{\theta})$ is a correlation matrix. This correlation matrix is a function of the distance between two sites and $\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is a vector of unknown correlation parameters and σ_Z^2 is the unknown structural parameter or constant variance. In addition, ε denotes an independent non-spatial random effect with mean 0 and variance-covariance matrix $\sigma_\varepsilon^2 \mathbf{I}$. In this case, σ_ε^2 represents the classical nugget effect and captures

measurement error or a combined effect of measurement error and any small scale spatial variation (Diggle & Ribeiro, 2002).

The posterior predictive density $\mathbf{Y}_{\text{miss}}(\mathbf{s})$ is obtained by integrating the following expression with respect to the parameters $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, σ_{ε}^2 and σ_Z^2 (see Equation 1) is:

$$f(\mathbf{Y}_{\text{miss}} \mid \mathbf{Y}_{\text{obs}}, \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma_{\varepsilon}^2, \sigma_Z^2, \mathbf{Z})$$
$$f(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma_{\varepsilon}^2, \sigma_Z^2, \mathbf{Z} \mid \mathbf{Y}_{\text{obs}})$$
$$\propto f(\mathbf{Y}_{\text{miss}} \mid \mathbf{Y}_{\text{obs}}, \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma_{\varepsilon}^2, \sigma_Z^2, \mathbf{Z})$$
$$f(\boldsymbol{\beta} \mid \mathbf{Y}_{\text{obs}}, \boldsymbol{\theta}, \sigma_{\varepsilon}^2, \sigma_Z^2, \mathbf{Z}) f(\mathbf{Z} \mid \boldsymbol{\theta}, \sigma_s^2)$$
$$\times f(\boldsymbol{\theta} \mid \mathbf{Y}_{\text{obs}}) f(\sigma_{\varepsilon}^2 \mid \mathbf{Y}_{\text{obs}})$$
$$f(\sigma_Z^2 \mid \mathbf{Y}_{\text{obs}}) \pi(\boldsymbol{\beta}) \pi(\boldsymbol{\theta}) \pi(\sigma_{\varepsilon}^2) \pi(\sigma_Z^2)$$

An exact expression for the integral will depend on the distribution (such as normal, Poisson, gamma, Bernoulli, binomial) assumed for the complete data, $f(\mathbf{Y}_{\text{miss}}, \mathbf{Y}_{\text{obs}})$, the distributions assumed for the two random components of the model, $f(\mathbf{Z} \mid \boldsymbol{\theta}, \sigma_Z^2)$ and $f(\boldsymbol{\varepsilon} \mid \sigma_{\varepsilon}^2)$, and the priors assumed for the parameters, $\pi(\boldsymbol{\beta})$, $\pi(\boldsymbol{\theta})$, $\pi(\sigma_{\varepsilon}^2)$ and $\pi(\sigma_Z^2)$. Diggle and Ribeiro (2002), Handcock and Stein (1993) and Omre and Halvorsen (1989) investigated the case assuming a Gaussian distribution for the data and a number of prior distributions for the parameters; their results are applied when selecting appropriate priors for the simulation and illustrative examples herein.

Methodology

The use of MI with a geostatistical model was assessed in a simulation. In addition, these procedures were applied to data collected from a 2002 probability survey of Coho salmon located in streams in western Oregon watersheds.

Simulation Example

One realization from a multivariate normal process with mean vector equal to $\mathbf{0}$, and a variance covariance matrix equal to $\sigma_z^2 \mathbf{R}(\boldsymbol{\theta}) + \sigma_{\varepsilon}^2 \mathbf{I}$ over a 21 by 21 regular grid was

generated and variances were chosen to be unequal and small. The variance, $\sigma_Z^2 = 0.8$ is the variance of the latent spatial random process and $\sigma_{\varepsilon}^2 = 0.2$ is the variance of the non-spatial random process. $\mathbf{R}(\theta)$ denotes the one-parameter 21 by 21 correlation matrix generated assuming an exponential correlation function, $e^{\|\mathbf{s}_i - \mathbf{s}_j\| / \theta}$, with s_i and s_j denoting two different sites, and $\theta = 2$ denoting the maximum distance where correlation between two sites is expected.

The parameter θ is known as the scale parameter and controls how fast the correlation decays with distance. Large values of θ correspond to a strong spatial correlation and small values to a weak spatial correlation. $\mathbf{I}$ is the 21 by 21 identity matrix. This simulated process accounts for spatial variation and measurement error. The collection of 441 observations defines the population values.

To induce a missing at random (MAR) mechanism on the response, stratification was imposed to the region of interest by dividing it into seven equal area vertical regions and then assigning a different response rate to each stratum; each stratum consists of 63 sites. Specification of the response rate range was based on the observed response rates from seven environmental surveys ranging from 0.69 to 0.90, as reported by Herger and Hayslip (2000) and Flitcroft, et al (2002). A range of response rates from 0.70 to 0.90 was assumed and randomly assigned to the seven strata. Within each stratum, 63 values of a uniform random variable P was assigned randomly to the 63 sites. A site, $\mathbf{s}$, if selected, would be missing if $P(\mathbf{s}) \leq 1 - \alpha$, where $P(\mathbf{s})$ denotes the value of the random variable P assigned to the site $\mathbf{s}$, and α denotes the stratum response rate.

Samples of size $n = 152$ were selected at random using equal allocation. Missing rates of 5%, 15%, 25%, 35% and 45% were assumed. For each missing rate, the number of missing sites in the sample was allocated proportional to the stratum response rates. Using the same sampling design, 2,000 samples of size $n = 152$ were generated. The Horvitz-Thompson (HT) mean and variance estimators for the continuous domain (Cordy, 1993) were calculated under the

following settings: (1) the observed data; (2) hot deck imputation; (3) a single imputation obtained from the geostatistical imputation model; (4) the predictive posterior mean imputation calculated as the mean of independent realizations from the predictive posterior distribution at each missing site; (5) hot deck multiple imputation using five and ten multiple imputations for the missing data and (6) multiple imputations for the predictive posterior mean imputation using five and ten multiple imputations for the missing data.

For the single and multiple imputation approaches, a multivariate mixed Gaussian model with constant mean β and variance covariance matrix $\sigma_Z^2 \mathbf{R}(\theta) + \sigma_\varepsilon^2 \mathbf{I}$ was assumed. $\mathbf{R}(\theta)$ is a correlation matrix that is a function of the distance between sites and an unknown parameter θ. The parameters of the posterior distribution were estimated by implementing MCMC techniques using a MATLAB program (Smith, 2004). An exponential correlation function and a uniform prior for β, an exponential prior for the correlation parameter with mean 1, and an inverse gamma distribution with parameters $\alpha = 0.1$ and $\beta = 10$ for the variance parameters σ_Z^2 and σ_ε^2 were assumed. As discussed by both Diggle and Ribeiro (2002) and Banerjee, et al. (2004), these prior selections lead to proper posterior distributions.

Imputation values for the missing data were obtained after verifying that the sample auto-correlations of the MCMC traces were less than 0.01 to ensure independence of the MCMC realizations. Values were randomly selected from the collection of independent realizations and used for the single and multiple imputations.

Salmon Example

This approach was illustrated with the 2002 winter Coho salmon spawning probability survey conducted by the Oregon Department of Fish and Wildlife (ODFW). This survey provides annual inventories of the Coho salmon abundance in streams located within western Oregon watersheds. These streams drain into the Pacific Ocean south of the Columbia River and are considered suitable habitat for salmon (Flitcroft, et al., 2002). The target population

consists of all streams located in a United States Geographical Survey (USGS) hydrography data layer of Oregon, except those streams located upstream of large dams that blocked anadromous fish passage (Flitcroft, et al., 2002).

The ODFW uses a generalized random tessellation stratified (GRTS) probability design (Stevens & Olsen, 1999) to select the sample site locations within the population of stream segments. The objective of these surveys is to estimate spawning Coho salmon abundance in both the entire area as well as within five monitoring areas (MA): North Coast, Mid Coast, Mid South Coast, Umpqua and South Coast.

Approximately 120 sites are selected per year within each MA, except in the South Coast MA where the sample size is about 60 sites per year. A total of 495 sites were surveyed in 2002. An additional 61 sites were originally selected in the sample but not visited because of time constraints or inaccessibility of the site location, resulting in 11% missing rate. It was assumed that these missing values resulted from a MAR mechanism. Figure 1 shows the location of the surveyed and missing sites corresponding to the year 2002. Stars represent surveyed sites, and open dots denote the missing sites in the same year. Each sampling site is approximately one-mile in length. At each selected site, counts of spawning Coho are obtained by visual observation. The population abundance of returning adult Coho in individual sites is estimated using area-under-the curve (AUC) techniques (Jacobs, et al., 2002).

Let Y_i denote the total number (abundance) of spawning Coho salmon observed at site $\mathbf{s}_i$ in 2002 and l_i be the length of the site $\mathbf{s}_i$ (in kilometers). Let λ_i be the density of spawning Coho salmon (counts per kilometer) at site $\mathbf{s}_i, i = 1, \ldots, n$, where n is the total number of surveyed sites. The total number of spawning Coho salmon at each site, Y_i, was assumed a noisy version of an unobserved spatial random process Z_i, and that conditional on Z_i, Y_i has a Poisson distribution with mean $l_i \lambda_i$. In other words, $Y_i \mid Z_i \sim \text{Poisson}(l_i \lambda_i)$, where $\log(\lambda_i) = \mu_i + Z_i + \varepsilon_i$, where μ_i denotes a

systematic component, Z_i denotes the spatial random component and ε_i the non-spatial random component, $i = 1,\ldots,n$.

The systematic component is assumed constant within each MA:

$$\mu_i = \beta_0 + \sum_{j=1}^{4} \beta_j x_{ij}$$

where $\beta_1, \beta_2, \beta_3$ and β_4 are the regression coefficients measuring the MA effects (North Coast, Mid-Coast, Mid-South and Umpqua, respectively, compared to the South Coast MA). The variable x_{ij}, is denoted by the value 1 if the ith site is located in MA j, and 0 otherwise, $i = 1,\ldots,n$, $j = 1,2,3,4$.

Figure 1: Site Locations for ODFW 2002 Spawning Locations

○ Visited sites
✳ Missing Sites
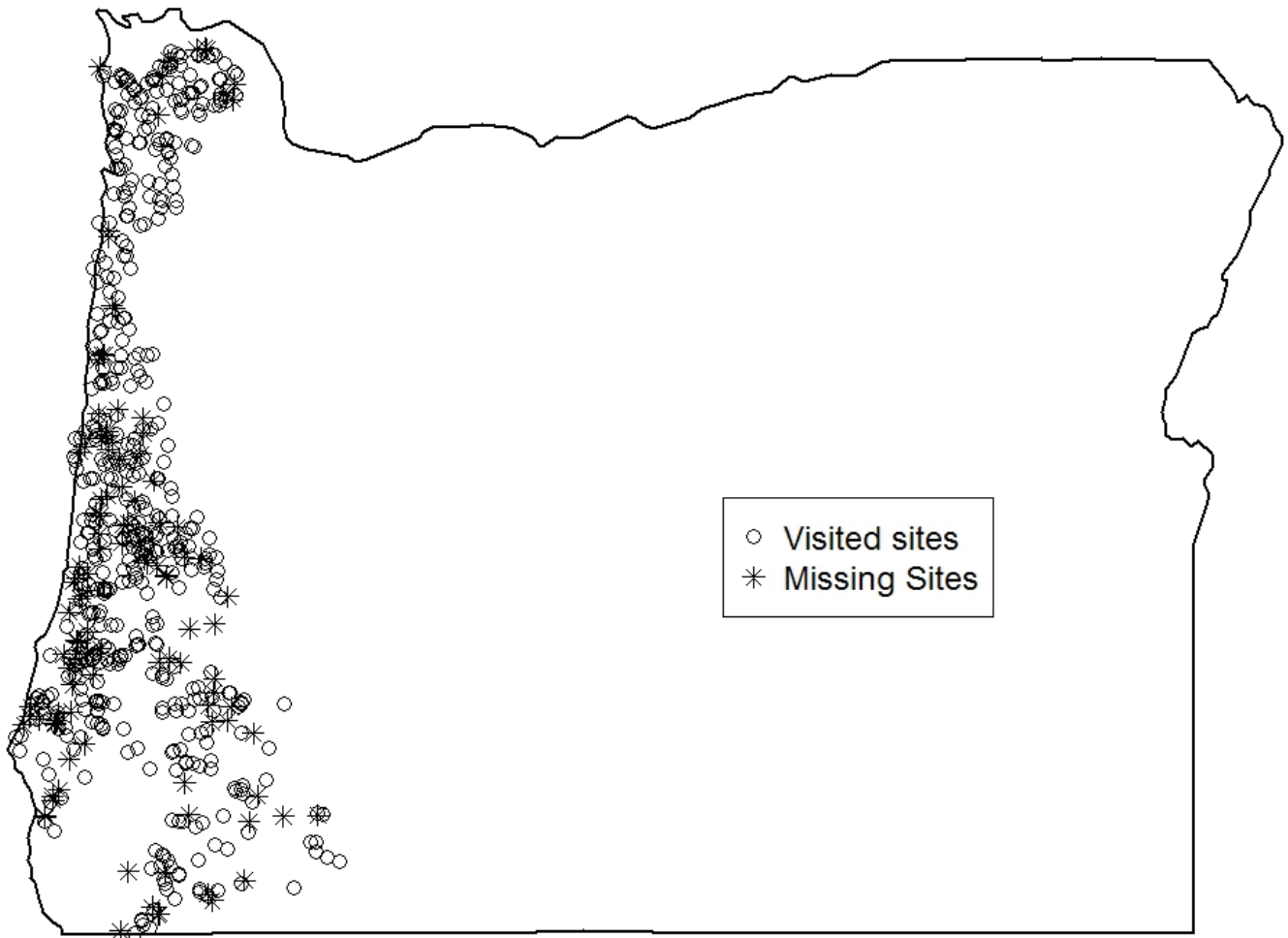
The spatial random process $\mathbf{Z}$ is assumed to have a multivariate normal distribution with $\underline{\mathbf{0}}$ mean vector and variance-covariance matrix given by $\sigma_Z^2 \mathbf{R}(\theta)$, where θ is the spatial correlation parameter, and $R_{ij}(\theta) = e^{\|\mathbf{s}_i - \mathbf{s}_j\|/\theta}$ denotes the exponential model. The non-spatial random effects, ε_i, are assumed to be independent and normally distributed with mean $\underline{\mathbf{0}}$ and variance σ_ε^2.

All parameters are assumed independent; vague prior distributions for the parameters were also assumed based on discussions from scientists experienced with these studies. An inverse-gamma $(\alpha = 0.1, \beta = 10)$ prior for σ_Z^2 and σ_ε^2, which has a wide distribution due to a long tail, and a proper prior $\pi(\theta) = 1/\theta^2$ for θ on the interval $[0.01, 50]$ was assumed. Selection of the upper limit of 50 kilometers was based on the assumption that it is unlikely to observe spatial correlation beyond this value. For the components of $\boldsymbol{\beta}$, independent improper uniform priors were used. Mathematical expressions for the marginal posterior distributions follow those presented in Christensen and Waagepetersen (2002).

A MATLAB program was used to obtain realizations from the posterior distributions of θ, σ_Z^2 and σ_ε^2, and each of the elements of $\mathbf{Z}$ and $\boldsymbol{\beta}$ (Smith, 2004). The MCMC simulation was run for 250,000 iterations after a 250,000 burn-in period. In order to reduce serial correlation in the simulated values, particularly in the chain for the parameter θ, each chain was re-sampled to obtain a final sample of 2,500 values of almost uncorrelated values (auto-correlation = 0.01) from the posterior for $\theta, \sigma_Z^2, \sigma_\varepsilon^2$ and each of the elements of $\boldsymbol{\beta}$, $\mathbf{Z}$, and $\log(\boldsymbol{\lambda})$.

Results
Simulation Example
The Geweke's statistics and two sided p-value for the model parameters $\beta, \theta, \sigma_Z^2$ and σ_ε^2 are 0.107 and 0.915; 0.875 and 0.382; 0.871 and 0.384; and 0.826 and 0.401, respectively, suggesting no evidence exists against convergence for each parameter. Similar results were achieved with the Heidelberger and Welch test for the model parameters, suggesting that chain convergence was achieved immediately after the 10,000 burn-in period for each model parameter (p-values for $\beta, \theta, \sigma_Z^2$ and σ_ε^2 are 0.552, 0.891, 0.926 and 0.784, respectively).

Table 1 shows the simulated root mean squared error (RSME), the average width of the 95% confidence interval, and the coverage rate of the simulated 95% confidence interval for each missing rate. A number of observations can be made from this simulation. As the percentage of missing data increases, the coverage rate decreases. As the missing rate increases, the imputation approaches all appear to be much closer to the 95% coverage as compared to the observed data. The multiple imputation approaches increase the RMSE slightly as compared to the simple and posterior mean imputation approach. In general, all multiple imputation methods (M = 20 not shown) performed similarly suggesting that there is no considerable gain in precision with more than 5 imputations.

Salmon Example
Sensitivity to selection of hyper-parameters was explored and no meaningful change was observed in the results. The convergence of the MCMC traces was assessed with the Geweke's statistic and the Heidelberger and Welch test. The Geweke's statistics and two sided p-values for the model parameters $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \theta, \sigma_Z^2$ and σ_ε^2 are −0.052 and 0.959, −1.081 and 0.230, 0.222 and 0.824, −0.154 and 0.878, -−0.240 and 0.810, −0.588 and 0.556, 0.910 and 0.363, and 0.551 and 0.5821, respectively, suggesting that no evidence exists against convergence for each parameter. Similar results were achieved with the Cramer-von-Mises statistics for the model parameters, suggesting that chain convergence was achieved for each model parameter (p-values: 0.886, 0.753, 0.921, 0.989, 0.667, 0.410, 0.944, and 0.366). As a result, the iterations

Table 1 Simulated Root Mean Squared Error (RMSE) of the Mean Estimate, Average Width and Coverage Rate of the 95% Confidence Interval for 5%, 15%, 25%, 35% and 45% Missing Rates

Missing Response Rate	Analysis Method	RMSE $\times$ 100	Width of Interval $\times$ 100	Coverage Rate(%) $\times$ 100
5% Missing	Observed Data	5.502	21.569	95.10
	Single Posterior Imputation	5.425	21.266	95.85
	Hot Deck Imputation	5.677	21.319	96.11
	Posterior Mean Imputation	5.423	21.259	96.00
	Multiple Imputation (M=5)	5.446	21.349	95.94
	Multiple Imputation (M=10)	5.446	21.351	96.00
	Hot Deck Multiple Imputation (M=5)	5.601	21.956	93.80
	Hot Deck Multiple Imputation (M=10)	5.553	21.768	94.10
15% Missing	Observed Data	5.480	21.482	92.05
	Hot Deck Imputation	5.509	20.693	93.81
	Single Imputation Data	5.264	20.636	94.55
	Predictive Posterior Mean Imputation	5.259	20.615	94.65
	Multiple Imputation (M=5)	5.280	20.700	94.83
	Multiple Imputation (M=10)	5.279	20.705	94.85
	Hot Deck Multiple Imputation (M=5)	5.432	21.293	93.20
	Hot Deck Multiple Imputation (M=10)	5.354	20.988	93.73
25% Missing	Observed Data	5.477	21.468	88.40
	Single Imputation Data	5.103	20.001	93.10
	Hot Deck Imputation	5.174	20.056	93.36
	Predictive Posterior Mean Imputation	5.093	19.964	92.90
	Multiple Imputation (M=5)	5.111	20.035	90.14
	Multiple Imputation (M=10)	5.110	20.051	93.35
	Hot Deck Multiple Imputation (M=5)	5.382	21.097	90.21
	Hot Deck Multiple Imputation (M=10)	5.313	20.827	93.23
35% Missing	Observed Data	5.490	21.519	82.20
	Single Imputation Data	4.944	19.381	91.45
	Hot Deck Imputation	5.174	19.434	91.70
	Predictive Posterior Mean Imputation	4.931	19.330	91.20
	Multiple Imputation (M=5)	4.952	19.414	92.00
	Multiple Imputation (M=10)	4.950	19.433	91.90
	Hot Deck Multiple Imputation (M=5)	5.264	20.634	89.23
	Hot Deck Multiple Imputation (M=10)	5.271	21.662	90.30
45% Missing	Observed Data	5.480	21.483	73.05
	Single Imputation Data	4.810	18.854	91.55
	Hot Deck Imputation	5.033	18.837	91.80
	Predictive Posterior Mean Imputation	4.792	18.785	90.85
	Multiple Imputation (M=5)	4.811	18.863	91.24
	Multiple Imputation (M=10)	4.809	18.887	91.45
	Hot Deck Multiple Imputation (M=5)	5.124	20.086	88.70
	Hot Deck Multiple Imputation (M=10)	5.212	20.431	89.23

$\boldsymbol{\beta}^{(t)}, \theta^{(t)}, \sigma_Z^{2(t)}, \ \sigma_{\varepsilon}^{2(t)}, Z^{(t)}$ and $\log\left(\boldsymbol{\lambda}^{(t)}\right)$ for $t = 1, \ldots, 2{,}500$ can be treated as a sample from the joint posterior distribution $p\left(\log(\boldsymbol{\lambda}), \mathbf{Z}, \boldsymbol{\beta}, \theta, \sigma_Z^2, \sigma_{\varepsilon}^2 \mid \mathbf{Y}\right).$

The posterior mean, median and the 95% Bayesian credible interval for each of the parameters in the model are shown in Table 2. The regression coefficients for the region covariates indicate that the MAs Mid-Coast, North Coast, Mid-South Coast and Umpqua tend to have a higher abundance of spawning Coho salmon than the MA South Coast. In addition, the posterior 95% Bayesian credible intervals for all region parameters except the Mid-Coast include zero, suggesting that all MAs except the Mid-Coast have a similar abundance of spawning Coho salmon.

The quantiles for σ_Z^2 (1.93; 4.73) (on the log scale) are above zero, indicating that after the inclusion of the five-level region covariates in the model there is substantial unexplained spatial variation left in the model. The 0.025 and 0.975 quantiles for the distance-scale parameter θ (8.50; 34.66) (in kilometers) indicate that there is significant spatial dependence after the inclusion of the five-level region covariate. The quantiles for σ_{ε}^2 (0.82; 1.95) (on the log scale) are above zero,

indicating that after the inclusion of the five-level region covariate and the spatial random effect, some additional variability may be attributed to observation error and other small-scale variation not accounted for in the model.

Using the 2,500 iterations of the posterior predictive parameters, the geostatistical imputation model is compared with hot deck imputation. The single imputation method was obtained by selecting one independent draw from the posterior predictive distribution. Multiple imputation was used to assess the impact of the error for this method using five and ten draws. This method was compared to the hot deck imputation, also employing both five and ten imputations.

Finally, the mean of the 2,500 values from the predictive posterior distribution of each missing site was used to estimate the predictive posterior mean for the missing site. These imputation methods are compared with the complete observed data ignoring the missing values. The predicted values were back transformed and the Horvitz-Thompson (HT) estimator for the total estimate for the abundance of spawning Coho salmon, the standard error using the local-variance estimator (Stevens & Olsen, 2003), and the 95% confidence intervals for the total were calculated.

Table 2: Mean, Median, and 95% Bayesian Credible Intervals for the Parameters of the Model

Parameter	Mean	Median	0.025 Quantile	0.975 Quantile
β_0 (South Coast)	0.17	0.16	-1.06	1.41
β_1 (North Coast)	1.64	1.67	-0.19	3.39
β_2 (Mid-Coast)	2.48	2.50	0.87	4.07
β_3 (Mid-South)	1.52	1.51	-0.03	3.11
β_4 (Umpqua)	1.28	1.28	-0.16	2.68
θ	17.49	16.10	8.50	34.66
σ_Z^2	3.07	2.98	1.93	4.73
σ_{ε}^2	1.39	1.39	0.82	1.95

Table 3 shows a summary of the results; the total estimate using only the observed data provides the lowest total counts estimate of all approaches. No adjustment for missing data was made for this estimate. Examination of the data reveals that the highest level of missing data was found in the Mid-Coast and the highest abundance values were located in this region. All imputation methods that made adjustments for this differential nonresponse across regions provided larger total estimates than the observed data.

The single posterior imputation obtains just one draw and may be more variable than an imputation based on multiple or the mean of multiple draws. The standard error for the MI method is larger than that obtained with the other methods: : this was expected because MI accounts for uncertainty due to the imputations (Schafer, 1997). As a result, the 95% confidence intervals using only the observed data (ignoring the missing values), single imputation and mean imputation, are less conservative than that which uses multiple imputation.

Conclusion

Statistical techniques that incorporate the spatial structure of the data in the random and/or systematic part of a model are currently used for modeling environmental phenomena, either discrete or continuous. Therefore, it seems natural to explore the efficiency of a multiple imputation approach that incorporates the spatial

structure of the latent process while accounting for missing data. The use of generalized mixed models to account for the missing data in environmental surveys was explored in this article. Generalized mixed models are recent techniques used for modeling environmental phenomena in an attempt to capture any spatial and/or temporal structure in the data. The possibility of implementing generalized linear models to different data distributions make them appealing for handling missing data in environmental surveys. Evaluations of the selection of the priors and the model specifications are performed before any imputation is conducted. This allows the researcher to explore different models for the covariance matrix and different priors that may better reflect the study data.

Simulation results from this study suggest that all imputation methods perform well at 5% and 15% missing rates. When the missing rate is 15% or higher, the performance of the statistics decays similarly for all techniques considered. However, the coverage rates for the 95% confidence intervals for all imputation methods are improved over no imputation. The performance of the statistics observed with 5 and 10 multiple imputations at all response rates, suggests that as in human populations (Schafer, 1997, Little & Rubin, 2002), little is gained when the number of imputations exceeds 5.

The method was illustrated by estimating the mean of an environmental

Table 3: Total, SE and 95% Confidence Intervals for the Abundance of Spawning Coho Salmon (Total counts of Spawning Coho Salmon) in the Oregon Coast

Imputation Technique	Total	SE	0.025 Quantile	0.975 Quantile
Observed Data (No Imputation)	227,885	16,648	195,255	260,514
Hot Deck Imputation	249,271	16,966	216,018	282,524
Single Posterior Imputation	238,185	16,919	205,023	271,346
Posterior Mean Imputation	250,921	16,519	218,543	283,298
MI Hot Deck (m=5)	257,931	18,193	222,274	293,589
MI Posterior (m=5)	250,213	21,689	206,302	294,127

variable, the abundance of spawning Coho salmon in the Oregon coastal streams. It is expected that multiple imputation methods which incorporate auxiliary information into the systematic part may render better results than the observed data. By incorporating auxiliary variables correlated with the process of interest into an imputation geostatistical model, the variances of the spatial component and the measurement error may be reduced resulting in narrowed posterior prediction intervals for the missing data. This implies that imputations may be closer to the unobserved true value, which will improve the imputation results. However, given the variability expected in natural environments, it is important to account for the imputation error through a multiple imputation approach.

References

Baker, S. G., & Laird, N. M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, *83*(*41*), 62-69.

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2004). *Hierarchical modeling and analysis for spatial data.* New York: Chapman & Hall.

Barnard, J., & Rubin, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika*, *86*, 949-955.

Chen, Y., & Shao, J. (1999). Inference with survey data imputed by hot deck when imputed values are nonidentifiable. *Statistica Sinica*, *9*(*2*), 361-384.

Christensen, O. F., & Waagepetersen, R. P. (2002). Bayesian prediction of spatial count data using generalized linear mixed models. *Biometrics*, *58*, 280-286.

Cordy, C. B. (1993). An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe. *Statistics & Probability Letters*, *18*, 353-362.

Cressie, N. (1993). *Statistics for spatial data.* New York: Wiley.

Diggle, P. J., Tawm, J. A., & Moyeed, R. A. (1998). Model-based geostatistics. *Applied Statistician*, *47*(*3*), 299-350.

Diggle, P. J., & Ribeiro, P. J. (2002). Bayesian inference in gaussian model-based geostatistics. *Geographical & Environmental Modeling*, *6*(*2*), 129-146.

Flitcroft, R. L., Jones, K. K., Reis, K. E. M., & Thom, B. A. (2002). *Year 2000 stream habitat conditions in western Oregon. Monitoring Program Report Number OPSW-ODFW-2001-05*, Oregon Department of Fish and Wildlife, Portland.

Gelman. A., Rubin, D. B., Carlin, J., & Stern, H. (1995). *Bayesian data analysis.* London: Chapman and Hall.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457-472.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain monte carlo in practice*. London: Chapman & Hall.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments, in *Bayesian Statistics*, J. M. Bernado, J. O. Berger, A. P. David, & A. F. M. Smith (Eds.). Oxford, U.K.: Clarendon Press.

Greenless, J. S., Reece, W. S., & Zieschang, K. D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, *77*(*378*), 251-261.

Gupta, A., & Lam, M. S. (1996). Estimating missing values using neural networks. *Journal of the Operational Research Society*, *47*, 229-238.

Handcock, M. S., & Stein, M. L. (1993). A Bayesian analysis of kriging. *Technometrics*, *35*(*4*), 403-410.

Harel1, O., & Zhou, X. (2007). Multiple imputation: Review of theory, implementation and software. *Statistics in Medicine*, *26*, 3057-3077.

Heidelberger, P., & Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. *Opinions Research*, *31*, 1109-1144.

Herger, L. G., & Hayslip, G. (2000). *Ecological condition of streams in the Coast Range ecoregion of Oregon and Washington*, EPA-910-R-00-002. U.S. Environmental Protection Agency, Region 10, Seattle, Washington.

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association*, *85*(*411*), 765-769.

Jacobs, S., et al. (2002). Status of Oregon Coastal Stocks of Anadromous Salmonids, 2000-2001 and 2001-2002, Oregon Plan for Salmon and Watersheds Monitoring, Report No. OPSW-ODFW-2002-3.

Kleinman, K. P., Ibrahim, J. G., & Laird, N. M. A. (1998). Bayesian framework for intent-to-treat analysis with missing data. *Biometrics*, *54*(*1*), 265-278.

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

Lessler, J. T., & Kalsbeek, W. D. (1992). *Nonsampling Error in Surveys*. New York: Wiley.

Little, R. J. A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, *77*(*378*), 237-250.

Little, R. J. A., & Schluchter, M. D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*, *72*(*3*), 497-512.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (*2nd Ed.*). New York: Wiley.

Lohr, S. L. (2001). *Sampling: design and analysis*. New York: Brooks/Cole Publishing Company.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. London: Chapman and Hall.

Omre, H., & Halvorsen, K. B. (1989). The Bayesian bridge between simple and universal kriging. *Mathematical Geology*, *21*, 767-786.

Plummer, M., Best, N., Cowless, K., & Vines, K. (2003). *The coda package. Output analysis and diagnostics for MCMC*. http://www-fis.iarc.fr/coda.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, *90*, 106-121.

Rotnitzky, A., Robins, J. M., & Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, *93*(*444*), 1321-1339.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*, 473-489.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Schafer, J. L. & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: a data analyst's perspective. *Multivariate Behavioral Research*, *33*, 545-571.

Schafer, J. L., & Schenker, N. S. (2000). Inference with imputed conditional means. *Journal of the American Statistical Association*, *95*(*449*), 144-154.

Schneider, T. (2001). Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. *American Meteorological Society*, *14*, 853-871.

Sebastiani, P., & Ramoni, M. (2000). Bayesian inference with missing data using bound and collapse. *Journal of Computational and Graphical Statistics*, *9*(*4*), 779-800.

Smith, R. A. (2004). A MATLAB package for geostatistic analysis. Unpublished manuscript.

Stevens, D. L., & Olsen, A. R. (1999). Spatially restricted surveys over time for aquatic resources. *Journal of Agricultural, Biological, and Environmental Statistics*, *4*, 415-428.

Stevens, D. L., & Olsen, A. R. (2003) Variance estimation for spatially balances samples of environmental resources. *Environmetrics*, *14*, 1-18.