Data augmentation for the handling of censored spatial data

by

Brooke Leann Fridley

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Philip Dixon, Major Professor
Mark Kaiser
Kenneth Koehler
Stephen Vardeman
Thomas Glanville

Iowa State University

Ames, Iowa

2003

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

Graduate College
Iowa State University

This is to certify that the doctoral dissertation of

Brooke Leann Fridley

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Major Professor

Signature was redacted for privacy.

For the Major Program

TABLE OF CONTENTS

GENERAL INTRODUCTION 1

1 Introduction 1

2 Methods for censored independent observations 2

2.1 Deletion and substitution methods 2

2.2 Sample median, trimmed mean and Winsorized datasets 3

2.3 Likelihood approach 5

2.4 Probability plotting methods 6

2.5 System error approach 8

2.6 Probability of acceptance curves 9

3 MCMC estimation and inference 10

4 Kriging and Bayesian prediction 11

5 Dissertation organization 13

References 15

DATA AUGMENTATION FOR A BAYESIAN SPATIAL MODEL INVOLVING CENSORED OBSERVATIONS 17

Abstract 17

1 Introduction 18

2 Censored data and data augmentation 19

3 Spatial Bayesian model and prediction 21

4 Markov chain Monte Carlo for data augmentation 23

5 Illustrative example I: Missouri dioxin contamination 25
5.1 Description of data 25
5.2 Model specification 27
5.3 Results 28
6 Illustrative example II: site 15 38
6.1 Description of data 38
6.2 Model specification and results 39
7 Discussion and conclusions 51
Appendix 54
References 56

DATA AUGMENTATION FOR A CONDITIONALLY SPECIFIED GAUSSIAN SPATIAL MODEL INVOLVING CENSORED OBSERVATIONS 59

Abstract 59
1 Introduction 60
2 Censored data and data augmentation 62
3 Bayesian conditionally specified Gaussian spatial model 63
4 Markov chain Monte Carlo for data augmentation 65
5 Illustrative example: site 15 67
5.1 Description of data 67
5.2 Model specification and results 68
6 Illustrative example: Missouri dioxin contamination 80
6.1 Description of data 80
6.2 Model specification and results 81
7 Discussion and conclusions 87
Appendix 90

References 92

SIMULATION STUDY: DATA AUGMENTATION FOR THE HANDLING OF SPATIALLY CENSORED OBSERVATIONS 95

Abstract 95

1 Introduction 96

2 Data augmentation procedure 97

3 Data augmentation within a Bayesian spatial model 99

3.1 Model specification and data augmentation procedure 99

3.2 Simulation Study I 101

3.2.1 Estimation 101

3.2.2 Prediction 106

3.3 Simulation Study II 111

4 Conditionally specified Gaussian spatial model 122

4.1 Model specification and data augmentation procedure 122

4.2 Simulation Study III 125

4.2.1 Estimation 125

4.2.2 Prediction 129

4.3 Simulation Study IV 134

5 Conclusions 142

Appendix I 144

Appendix II 147

References 149

GENERAL CONCLUSIONS 151

ACKNOWLEDGMENTS 153

GENERAL INTRODUCTION

1 Introduction

Censored data (left, right or interval censored) occur in a variety of applications. In the case involving independent observations, numerous methods have been proposed to deal with the analysis of censored data (Helsel, 1990; Gibbons, 1995; Porter, Ward and Bell, 1988). In contrast, there are few adequate methods for the handling of censored observations involving spatial dependence. There are various statistical methods that allow for the analysis of spatially dependent data, but none of these statistical methods deal with the case involving censored data (Cressie, 1993; Ecker and Gelfand, 1997; Besag, 1974; Kaiser and Cressie, 2000).

In most spatial settings, if censoring has occurred, it usually results in left censored observations. Often, all the censored observations are set equal to some constant value, which results in single imputation for the censored observations. For example, in the case involving the measurement of environmental pollutants, some function of the level of detection (e.g. LOD, $LOD/2$) is commonly imputed for the censored observation. This single imputation method results in biased estimates of the mean, variability and spatial dependence.

This dissertation will present and illustrate a data augmentation approach, a method first proposed by Tanner and Wong (1987) and Li (1988), for the analysis of spatially correlated data, in which some of the observations are censored. Both a Bayesian geosta-

tistical model and a Bayesian conditionally specified Gaussian model will be presented within the data augmentation framework for the handling of censored observations. The method can be easily extended to the cases of interval censored and right censored spatial data. Comparison of the data augmentation method to the method of replacing the censored observations with LOD or $LOD/2$ will also be illustrated using two different studies involving soil contamination.

2 Methods for censored independent observations

Censored data is a type of missing data that is "non-ignorable" (Little and Rubin, 2002). If we were to throw out or ignore the censored observations, the resulting parameter estimates would be biased. Censoring, whether left or right, also results in the loss of information. The loss of information or censoring needs to be accounted for in the statistical analysis. There are various methods for the analysis of censored data in the case of independent observations. Some methods are more efficient than other methods. A few of the common methods to analyze censored data in the case of independence are outlined below.

2.1 Deletion and substitution methods

The easiest methods to handle censored data are the deletion and substitution methods. In the deletion procedure, observations reported below the detection limit are not used in the computation of the mean and the standard deviation. Hence, the mean is over-estimated while the standard deviation is under-estimated. Another method commonly used is the substitution method where one replaces the censored observations with either 0, $LOD/2$ or LOD. Then, based on this "imputed dataset", estimates of the mean and standard deviation are computed.

For example, let the truth be

0.5, 1, 4, 5, 5, 6

and the observed data be reported as

$$< 2, < 2, 4, 5, 5, 6.$$

If we were to replace the two censored values with their level of detection, the sample mean and standard deviation would be 4 and 1.67, respectively. If we were to replace the censored values with half their level of detection, the mean and standard deviation would be 3.67 and 2.16. Lastly, if we were to replace the censored values with 0, the resulting mean and standard deviation would be 3.33 and 2.66, respectively. In contrast, the true mean and standard deviation is 3.58 and 2.29, respectively.

Both the deletion and the substitution methods result in biased parameter estimates. In the case of large datasets with very few censored observations, the bias is not as extreme as in the case of small datasets or studies involving a large number of censored observations. In addition to biased estimates, there is no statistical justification for which constant to impute for the censored values. Due to the bias and arbitrary choice of the constant used in the imputation, these methods are not recommended (Newman, 1995; Gilbert, 1987; Helsel, 1990).

2.2 Sample median, trimmed mean and Winsorized datasets

The sample median and the trimmed mean are ways to produce a reasonable estimate of the mean or average. For example, instead of computing the sample mean as the measure of center, the sample median can be used. This approach is appropriate if not more than 50% of the observations are censored and if the underlying distribution is symmetric.

Another option is the use of the trimmed mean. A $100p\%$ trimmed mean, where $0 < p < 0.50$, is computed by finding the mean of the middle $100(1-2p)\%$ of the ordered observations. That is, the mean is computed on the middle $n(1-2p)$ observations,

where the largest np and smallest np observations are excluded from the computation. If the number of censored observations is no more than np, the trimmed mean can be computed. Thus, in the presence of a large proportion of censored observations, the trimmed mean can not be computed.

An idea similar to the trimming of datasets to computed a trimmed mean is the idea of Winsorizing. Winsorizing replaces the censored observations in a way that produces unbiased estimates of the mean and standard deviation. This method produces what is called the Winzorized mean and standard deviation. Assuming a symmetric distribution, the censored values are replaced by the smallest observation above the LOD. Then, the the same number of the largest observed values are replaced with the next smallest observation.

For example, if the dataset is

NA, NA, NA, 2, 3, 4, 4, 6, 7, 9, 10, 11,

the three NA values would be replaced with the value 2 and the values 9, 10 and 11 would be replaced with the value 7. Thus, the Winsorized dataset is

2, 2, 2, 2, 3, 4, 4, 6, 7, 7, 7, 7.

The Winsorized mean is the mean of the Winsorized dataset. The Winsorized standard deviation is $S_w(n-1)/(v-1)$, where S_w is the standard deviation of the Winsorized dataset, n is the number of observations and v is the number of unchanged observations. The mean and standard deviation computed using the Winsorized dataset, are unbiased estimates of the true mean and standard deviation. Thus, for our example, the Winsorized mean and standard deviation are 4.42 and 4.10. This method fails if there is more than one level of detection and if the number of censored observations is greater than or equal to $n/2$.

2.3 Likelihood approach

A more model oriented approach to analyze censored data is through the use of a likelihood function that accounts for the censored observations. In using this method, an assumption of the distributional form for the responses is required. This can be a drawback to the method. With the presence of censored observations, the assignment of a distribution form can be difficult. It is often hard or impossible to verify the distribution form in many cases involving censored data, leaving the distributional assumption as one's best guess.

The likelihood is composed of a piece representing the observed data and a piece representing the censored data. Let y_i have probability distribution function (or probability mass function) $f_y(y_i;\theta)$ for $i = 1, \ldots n$, where $F_y(\cdot)$ is the CDF of $f_y(\cdot)$. The following are the likelihoods involving the three types of censored data (i.e. left, right and interval censored).

- Left Censoring at a:

  $L(\theta;y) = \prod_{i=1}^{n} f_y(y_i;\theta)^{\delta_i} F_y(a)^{1-\delta_i}$, where δ_i is 1 if y_i is observed and 0 if censored.

- Right Censoring at b:

  $L(\theta;y) = \prod_{i=1}^{n} f_y(y_i;\theta)^{\delta_i} (1-F_y(b))^{1-\delta_i}$, where δ_i is 1 if y_i is observed and 0 if censored.

- Interval Censoring between a and b:

  $L(\theta;y) = \prod_{i=1}^{n} f_y(y_i;\theta)^{\delta_i} (F_y(b) - F_y(a))^{1-\delta_i}$, where δ_i is 1 if y_i is observed and 0 if censored.

The likelihood or log-likelihood function is then maximized in terms of θ, producing maximum likelihood parameter estimates.

2.4 Probability plotting methods

Probability plotting is a commonly used approach for the estimation of the mean and standard deviation in the presence of censored observations. The probability plotting method is outlined below for the case involving left censored observations.

1. Let $y \sim \text{NOR}(\mu, \sigma^2)$. Let the data (before censoring) be $y_1, y_2, ..., y_n$ where n_1 observations are censored (less than the LOD) and n_2 observations are observed (greater than the LOD).

2. Let $y_{i:n}$ represent the i^{th} order statistic for $i = n_1 + 1, ..., n$ (observed responses).

3. The cumulative percentage corresponding to each observation is then estimated and a plot of the cumulative percentage verses concentration is constructed.

4. A line that follows the data is then drawn on the plot.

5. The estimate of μ is taken to be the 50% cumulative percentage (P50).

6. The standard deviation is estimated by finding P16 (16%-tile) and P84 (84%-tile). The estimate of σ is taken to be (P84-P16)/2.

This method has the disadvantage of subjectivity in the fitting of the line. This problem can be overcome by using regression techniques to fit the line. In addition, if more than 16% of the data is censored, the method is not able to compute the estimate of the standard deviation.

A variation on the probability plotting method is the robust probability method (Helsel, 1990). This method is a modification of the probability plotting method that combines the observed data with extrapolated or imputed values for the censored observations to produce estimates of the mean and standard deviation. In doing so, a distributional form is assumed for the data. The method is as follows.

1. Assuming a distributional form, for each observation above the LOD, a z score is computed.

2. A plot of log(response) verses the z scores is constructed for which a regression line is then fit.

3. This regression line is then used to extrapolate/predict values for the censored observations.

4. A back transformation is then applied to return to values to the original units.

5. This process yields an "imputed dataset" from which estimates of the mean and standard deviation are computed.

Bias corrections for the use of back-transformation have been discussed in the literature and can be used to correct the bias due to transformation.

Lastly, the ad hoc quantile method is a combination of both the robust method and the probability plotting method (Cressie, 1998). Again, a drawback to this method is that estimates produced are not MLE's. The basic idea is to use the observed data to produce a regression line. From the regression line, prediction/extrapolation for the censored data is completed. Using the "imputed" dataset, estimates of the mean and standard deviation are then computed. This procedure of imputing and estimation is done until convergence. For the case of left censoring, the idea is as follows.

1. Let $y \sim \text{NOR}(\mu, \sigma^2)$. Let the data (before censoring) be $y_1, y_2, ..., y_n$ were n_1 observations are censored (less than the LOD) and n_2 observations are observed (greater than the LOD).

2. Let $y_{i:n}$ represent the i^{th} order statistic for $i = n_1 + 1, ..., n$ (observations observed).

3. Based on these order statistics, estimate μ and σ^2 by using the standard normal Q-Q plot. This is done by fitting the line $y = \mu + \sigma z$.

4. Based on the estimates of μ and σ^2, $\hat{\mu}$ and $\hat{\sigma}^2$, define $y_{i:n}$ to be $y_{i:n} = \hat{\mu} + \hat{\sigma}\Phi^{-1}(\frac{i-1/2}{n})$ for $i = 1, ..., n_1$ (i.e. imputation for the censored observations).

5. Using the imputed values for the censored values and the observed values, estimate μ and σ^2 to be $\tilde{\mu} = \sum_{i=1}^{n} \frac{y_{i:n}}{n}$ and $\tilde{\sigma}^2 = \sum_{i=1}^{n} \frac{(y_{i:n} - \tilde{\mu})^2}{n}$.

6. This procedure involving imputation and estimation (steps 4 and 5) is repeated until convergence.

Note, the estimate of σ^2 does not account for the variability involved in the imputation. Also, the method does not produce MLE's. The point estimates are adequate, but the standard errors are too small. The difference between the robust probability plotting method and the ad hoc quantile method is the iteration of the ad hoc quantile procedure until convergence. As with the likelihood method, an incorrect distributional assumption will lead to incorrect parameter estimates.

2.5 System error approach

Tackling the problem of censored data in a more philosophical approach is the idea of system error or measurement error approach. As stated by Porter, Ward and Bell (1988), "More information is gained when a numerical result and an estimate of measurement precision are reported for every measurement, as opposed to reporting "not detected" or "less than"". They further state that system error should be considered with the analysis of monitoring data.

Consider the following measurement error model,

$$X_m = X_p + e(X_p),$$

where X_m represents the measured amount, X_p represents the true amount, and $e(X_p)$ represents the measurement error. We wish to find out about the quantity X_p by using the observed data X_m. Thus, applying the idea of measurement error to censored data,

one does not report values as falling below a detection level. Instead, every value is reported with a numerical value and an estimate of the measurement precision. That is, measurements are reported as $X_m \pm$ *measurement error* for all values, including values falling below a detection level or non-detectable. A valid reporting definition of a non-detect or censored observation would be an interval that covers 0 (i.e. $0 \in X_m \pm$ *measurement error*) (Porter, Ward and Bell, 1988).

2.6 Probability of acceptance curves

Lastly, a method closely related to the system error approach is a method proposed by Lambert, Peterson and Terpenning (1991). The method introduces the use of a probability of acceptance curve, $p(m)$, which relates the probability of detection to the measured response. In doing so, the 'minimum reliably detected concentration' is defined as

$$\begin{aligned} \pi(C) &= \Pr(\text{acceptance} \mid \text{concentration} = C) \\ &= \int \Pr(\text{acceptance} \mid \text{measurement} = m)\ f(m|C)dm \\ &= \int p(m) f(m|C)dm. \end{aligned}$$

where C is a spiked concentration from a quality control sample, m is a measurement obtained from a field sample, and $f(m|C)$ is the density for the field samples with true concentration C. The 90^{th} percentile of $\pi(C)$ is referred to as the minimum reliably detected concentration and is the censoring limit. For example, assuming all non-detects fall below the smallest detected value is reasonable if the probability of acceptance curve rises sharply from 0 to 1.

The advantage to this approach is that it combines data from field samples and quality control samples. A disadvantage of this method of defining detection limits is that $p(m)$ requires an analyst to make the binary detection decision using their own detection criteria. Hence, "Lambert's method models the detection criterion of the

analyst but not the actual capabilities of the analytical method itself" (Gibbons, 1995). Therefore, different acceptance curves can be produced by difference analysts.

3 MCMC estimation and inference

Using Markov chain Monte Carlo (MCMC) in the form of a Gibbs sampler, a model can be fit with parameter estimation and inference based on the resulting simulated values from the chain. Let $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_p)$ represent a random vector with joint distribution $P(\boldsymbol{Y})$. The Gibbs sampler is a successive substitution sampler, in which draws are taken from the conditional distribution of each element in $\boldsymbol{Y}$ given all other elements. That is, at iteration t, $Y_1^{(t)}$ is generated from $p(Y_1|Y_2^{(t-1)}, ..., Y_p^{(t-1)})$, $Y_2^{(t)}$ is generated from $p(Y_2|Y_1^{(t)}, Y_3^{(t-1)}, ..., Y_p^{(t-1)})$ and $Y_p^{(t)}$ is generated from $p(Y_p|Y_1^{(t)}, ..., Y_{p-1}^{(t)})$ (Geman and Geman, 1984; Shafer, 1997; Gilks, Richardson and Spiegelhalter, 1996).

After the chain has converged, say at iteration k, $\boldsymbol{Y}^{(k*)}$, $\forall\ k^* \geq k$ can be considered as simulated values from the true joint posterior distribution, leading to an estimate of the joint posterior distribution, $P(\boldsymbol{Y})$, or any marginal posterior distributions that may be of interest. For example, an estimate for the random quantity Y_1 could be found by using a summary feature of marginal posterior distribution which can be estimated from the simulated values $Y_1^{(k^*)}$, $\forall k^* \geq k$ produced by the Gibbs sampler. In addition to a point estimate for Y_1, an approximate 95% Bayesian equal-tail credible interval for the random quantity Y_1 can be found by taking the 2.5 and the 97.5 percentiles of the simulated values $Y_1^{(k^*)}$, $\forall k^* \geq k$. If the posterior distribution is symmetric and unimodal, the equal-tail intervals correspond to the highest posterior density (HPD) credible set (Gelman, Carlin, Stern and Rubin, 1995; Carlin and Louis, 1996; de Oliveira and Ecker, 2002).

4 Kriging and Bayesian prediction

Along with estimating quantities of interest, the goal of many spatial analyses is to predict the value at unobserved locations. Let $Y_u(t_1)$ represent an ungauged (unobserved) value at location t_1. Let $\boldsymbol{Y}_g = (Y(s_1), Y(s_2), .., Y(s_g))$ represent a gauged (observed) vector for locations $\{s_1, s_2, .., s_g\}$. The goal is to predict $Y_u(t_1)$, where t_1 is the unobserved or ungauged location. One method to predict follows from the geostatistical literature called kriging. Kriging is done by considering only linear unbiased predictors of the form

$$\hat{Y}_u(t_1) = \sum_{i=1}^{g} \lambda_i Y(s_i).$$

Assuming stationarity, $E(Y(s_i)) = \mu$, this constraints $\sum_{i=1}^{n} \lambda_i = 1$. Hence, under square error loss we need to minimize

$$E\{Y(t_1) - \sum_{i=1}^{g} \lambda_i Y(s_i)\}^2 - 2m\{\sum_{i=1}^{g} \lambda_i - 1\}.$$

This minimization yields $\boldsymbol{\lambda} = \Gamma^{-1}\boldsymbol{\gamma}$ where $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_g, m)^T$, $\boldsymbol{\gamma} = (\gamma(t_1 - s_1), ..., \gamma(t_1 - s_g), 1)^T$ and

$$\Gamma = \begin{pmatrix} \gamma(s_1 - s_1) & \dots & \gamma(s_1 - s_g) & 1 \\ .. & .. & .. & .. \\ \gamma(s_g - s_1) & \dots & \gamma(s_g - s_g) & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}.$$

The kriging weights (λ_i) are written as a function of the semi-variogram, which is half the variogram. The variogram is defined as

$$2\gamma(s_i - s_j) = \text{Var}(Y(s_i) - Y(s_j)) = 2\text{Var}(Y(s_i)) - 2\text{Cov}(Y(s_i), Y(s_j))$$

with $\gamma(.)$ representing the semi-variogram. For an exponential parameterization of the spatial covariance matrix,

$$\gamma(Y(s_i) - Y(s_j)) = \tau^2 + \sigma^2(1 - \exp\{-d_{ij}/\phi\}).$$

The standard error for the prediction is $\sigma^2(t_1) = \boldsymbol{\lambda}^T\boldsymbol{\gamma}$. In place of $\gamma(\cdot)$, which is unknown, we use the estimate of the semi-variogram based on the estimated parameter values (Cressie, 1993; Matheron, 1963).

Hence, kriging at a given location results in the computation of a weighted mean, where the weights are based on the spatial dependence and variability parameters of the spatial model considered. In the case of independence, $\lambda_i = \frac{1}{g}$ for $\forall i = 1,...,g$ (i.e. equal weights). If censored data is present, by replacing the censored observations with a constant (e.g. $LOD/2$) not only are the subsequent parameter estimates biased, but also predictions. By applying data augmentation to spatial censored data, we hope to get more accurate parameters estimates along with better predictions.

An alternative to the traditional geostatistical kriging method is Bayesian prediction or Bayesian kriging. Again, let $\boldsymbol{Y}_u$ represent an ungauged (unobserved) vector $\boldsymbol{Y}_u = (Y(t_1), Y(t_2), .., Y(t_u))$ for locations $\{t_1, t_2, .., t_u\}$. Let $\boldsymbol{Y}_g = (Y(s_1), Y(s_2), .., Y(s_g))$ represent a gauged (observed) vector for locations $\{s_1, s_2, .., s_g\}$. Bayesian prediction uses the posterior predictive distribution

$$p(\boldsymbol{Y}_u|\boldsymbol{Y}_g) = \int p(\boldsymbol{Y}_u|\boldsymbol{Y}_g, \boldsymbol{\Theta})p(\boldsymbol{\Theta}|\boldsymbol{Y}_g)d\boldsymbol{\Theta}$$

for prediction purposes.

In the case involving censored data and data augmentation, let $\boldsymbol{Y}_u$,$\boldsymbol{Y}_g$, $\boldsymbol{Y}_{go}$, $\boldsymbol{Y}_{gc}$ represent the ungauged vector, gauged vector, gauged observed vector and the gauged censored vector, respectively. The joint distribution of $\boldsymbol{Y}_u$ and $\boldsymbol{Y}_g$ is then

$$\text{MVN}\left(\begin{pmatrix} \boldsymbol{\mu}_u \\ \boldsymbol{\mu}_g \end{pmatrix}, \begin{pmatrix} \Sigma_{uu} & \Sigma_{ug} \\ \Sigma_{gu} & \Sigma_{gg} \end{pmatrix}\right),$$

with mean vectors $\boldsymbol{\mu}_u$ and $\boldsymbol{\mu}_g$ of appropriate lengths, $\Sigma_{uu} = V(\sigma^2, \phi, D_{uu}) + \tau^2 I$, $\Sigma_{gg} = V(\sigma^2, \phi, D_{gg}) + \tau^2 I$, $\Sigma_{ug} = V(\sigma^2, \phi, D_{ug})$, and $\Sigma_{gu} = V(\sigma^2, \phi, D_{gu})$, where D_{uu}, D_{gg}, D_{ug}, D_{gu} are matrices containing distances between the ungauged and gauged sites.

Therefore, the conditional distribution for the ungauged sites given the gauged sites is multivariate normal.

Approximation of the posterior predictive distribution $p(\boldsymbol{Y}_u|\boldsymbol{Y}_g)$ can be accomplished using a Monte Carlo approach, where one simulates predictions from

$$\boldsymbol{Y}_u|\boldsymbol{Y}_{go}, \boldsymbol{Y}_{gc}^{(k)}, \boldsymbol{\Theta}^{(k)} \sim \text{MVN}(\boldsymbol{\mu}_{u.g}^{(k)}, \Sigma_{u.g}^{(k)}),$$

with $\boldsymbol{\mu}_{u.g}^{(k)} = \boldsymbol{\mu}_u^{(k)} + \Sigma_{ug}^{(k)}\Sigma_{gg}^{-1(k)}(\boldsymbol{Y}^{*(k)} - \boldsymbol{\mu}_g^{(k)})$, $\Sigma_{u.g}^{(k)} = \Sigma_{uu}^{(k)} - \Sigma_{ug}^{(k)}\Sigma_{gg}^{-1(k)}\Sigma_{gu}^{(k)}$, and $\boldsymbol{Y}^{*(k)} = (\boldsymbol{Y}_{go}^T, \boldsymbol{Y}_{gc}^{(k)T})^T$, for a large number MCMC iterations, k, (Carlin and Louis, 1996; de Oliveira and Ecker, 2002; Gelman, Carlin, Stern and Rubin, 1995). One advantage of the Bayesian prediction method is that the posterior predictive distribution reflects the variability in parameter estimation when predicting; kriging does not. Prediction standard errors produced via the kriging method are too small.

5 Dissertation organization

This dissertation provides a solution to the analysis of censored spatial data. Spatially dependent data occurs in a variety of applications in which observations are associated with a spatial location. Traditional methods to analyze spatial data are not appropriate when censored observations are present. In environmental studies, it is not uncommon for measurements of contaminants to fall below a level of detection (LOD). There are many statistical methods for the analysis of censored data when the observations are independent, but what does one do when spatial correlation is present? A solution presented in this dissertation is to use data augmentation for the analysis of censored spatial data.

The first paper will look at a geostatistical model (Cressie, 1993; Matheron, 1986). The model is set in the Bayesian framework, leading naturally to the data augmentation procedure. Prior distributions must be specified for all model parameters. So this will also be discussed (Ecker and Gelfand, 1997). In addition to parameter estimation and

inference, a main focus of many geostatistical analysis is spatial prediction. Censored observations cause some problems for traditional prediction methods. Spatial prediction will also be presented and illustrated involving censored data. Comparison of the data augmentation method to methods which replace any censored observations with half the level of detection and level of detection will be presented using data from two environmental studies; the first study investigating dioxin contamination in Missouri (Zirschky and Harris, 1986) and the second study looking at metal soil contamination at an old industrial site, called site 15.

The second paper explores the use of data augmentation for censored spatial data in the context of a Bayesian conditionally specified Gaussian or conditional auto-regressive model (Kaiser and Cressie, 2000; Besag, 1974; Daniels, Lee, and Kaiser, 2001). Once again, the use of a Bayesian model leads to data augmentation in a Gibbs sampler. Specification of prior distributions will be discussed. As opposed to the Bayesian geostatistical model, the focus of this paper is not on prediction, but on parameter estimation and subsequent inference. Comparison of the data augmentation method to the common method of replacing censored values with level of detection (LOD) and LOD/2 are illustrated using both the Missouri dioxin study and the site 15 metal contamination study.

In the third paper, results from an extensive simulation study, conducted to investigate the effect of different factors on the effectiveness of augmentation for the handling of censored spatial data, are presented and discussed. The simulation study will try and answer questions like, "Does the method work for high levels of censoring?," "Does the method work well for small samples?," "Does the method work better if there is large spatial dependence present in the data?" Two simulation studies were conducted, one for the geostatistical model and one for the conditionally specified Gaussian model, to answer these questions. In addition to simulation studies investigating factors that may impact the data augmentation procedure, two additional simulation studies were

conducted to look at the general adequacy of the augmentation procedure for both the geostatistical and conditionally specified Bayesian models.

These three papers are followed by a summary chapter giving the general conclusions for the entire dissertation. The summary discusses the superiority of the data augmentation method for the analysis of censored spatial data for both the geostatistical model and the conditionally specified model. The discussion concludes with general comments regarding the simulations studies presented in the fourth chapter of this dissertation.

References

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, B*, **36**, 192-236.

Carlin, B.P. and Louis, T.A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman & Hall, London.

Cressie, N.A.C. (1998). Environmental Statistics Course Notes, Iowa State University. January 1998 to May 1998.

Cressie, N.A.C. (1993). *Statistics for Spatial Data, Revised Edition.* John Wiley & Sons, Inc., New York.

Daniels, M.J., Lee, Y.D., and Kaiser, M.S. (2001). Assessing sources of variability in measurement of ambient particular matter. *Environmetrics*, **12**, 547-558.

de Oliveira, V., and Ecker, M.D. (2002). Bayesian hot spot detection in the presence of spatial trend: application to total nitrogen concentration in Chesapeake Bay. *Environmetrics*, **13**, 85-101.

Ecker, M.D., and Gelfand, A.E. (1997). Bayesian Variogram Modeling for an Isotropic Spatial Process. *Journal of Agricultural, Biological, and Environmental Statistics*, **2**, 347-369.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). *Bayesian Data Analysis.* Chapman & Hall, London.

Geman, S., and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Transactions on Patter Analysis and Machine Intelligence*, **6**, 721-741.

Gibbons, R. (1995). Some Statistical and Conceptual Issues in the Detection of Low-Level Environmental Pollutants. *Environmental & Ecological Statistics*, **2**, 125-167.

Gilbert, R.O. (1987) *Statistical Methods for Environmental Pollution Monitoring.* Van Nostrand Reinhold, New York.

Gilks, W.R., Richardson, S., and Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice.* Chapman & Hall, London.

Helsel, D.R. (1990). Less than Obvious. Statistical Treatment of Data Below the Detection Limit. *Environmental Science Technology*, **24**, 1766-1774.

Kaiser, M.S., and Cressie, N. (2000). The Construction of Multivariate Distributions form Markov Random Fields. *Journal of Multivariate Analysis*, **73**, 199-220.

Lambert, D., Peterson, B., and Terpenning, I. (1991). Nondetects, Detection Limits, and the Probability of Detection. *Journal of the American Statistical Association*, **86**, 266-277.

Li, K.H. (1988). Imputations Using Markov Chains. *Journal of Statistical Computation and Simulation*, **30**, 57-79.

Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data,* 2^{nd} *Ed.*. Wiley, New York.

Matheron, G. (1963). Principles of Geostatistics. *Economic Geology*, **58**, 1246-1266.

Newman, M.C. (1995). *Quantitative Methods in Aquatic Ecotoxicology.* Lewis Publishers, London.

Porter, P.S., Ward, R.C., Bell, H.F. (1988). The Detection Limit. Water Quality Monitoring Data Are Plagued with Levels of Chemicals That Are Too Low to Be Measured Precisely. *Environmental Science Technology*, **22**, 856-861.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data.* Chapman & Hall, London.

Tanner, M.A., and Wong, W.H. (1987). The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, **82**, 528-540.

Zirschky, J.H., and Harris, D.J. (1986). Geostatistical Analysis of Hazardous Waste Site Data. *Journal of Environmental Engineering*, **112**, 770-784.

DATA AUGMENTATION FOR A BAYESIAN SPATIAL MODEL INVOLVING CENSORED OBSERVATIONS

A paper to be submitted to Environmetrics

Brooke Fridley and Philip Dixon

Abstract

The analysis of spatially dependent data involving observations falling below a detection level occurs occasionally in environmental applications. With the increased interest in long term exposures to low level contaminants, better methods for handling small levels of a response variable, which lead to many censored observations, are needed. The most common practice for the handling of censored data in spatial settings is to replace the censored observations with some function of the level of detection (LOD), like $LOD/2$. The resulting parameter estimates and standard errors found using this single imputation method are biased. A data augmentation procedure using a Gibbs sampler for the analysis of censored spatial data in the context of a Bayesian spatial model is presented. Comparison of the data augmentation method to the LOD method and the LOD/2 method will be illustrated using data from a dioxin contaminated site and an old industrial area contaminated with heavy metals.

1 Introduction

Environmental studies, where contamination levels are measured at geographic locations, often result in some observations falling below a level of detection (LOD). Hence, some of the data are left censored. A method commonly used to handle censored spatial data is to assume independent observations and then use one of many methods available to handle censored data (Helsel, 1990; Gibbons, 1995; Porter, Ward and Bell, 1988). Another common approach that does not ignore the spatial dependency, is to replace the censored observation with some function of the level of detection (e.g. $LOD/2$, LOD). This ad hoc method of replacing all censored values with a constant results in biased estimates of the mean, variability, and spatial dependence.

Analysis of missing data has been an area of extensive research. The basic ideas and principles of missing data and the analysis of missing data have been outlined by Little and Rubin (2002). The idea of data augmentation to handle missing data using Markov chain Monte Carlo (MCMC) was first presented by Tanner and Wong (1987). Hopke, Liu and Rubin (2001) use a data augmentation procedure to provide k complete, augmented datasets which can then be analyzed using traditional statistical methods. Dempster, Laird and Rubin (1977) provide a general methodology using the EM (expectation/maximization) algorithm that can be used in a variety of missing data problems. The EM algorithm works by iteratively maximizing the data likelihood whereby setting the missing data or missing variable equal to its expectation until a convergence criteria has been satisfied. The EM algorithm and various hybrids of the algorithm have been used extensively for the handling of missing or censored data in mixed model (Hughes, 1999; Smith and Helms, 1995; Pettitt, 1986). Hybrids using both EM and MCMC ideas have also been used to handle missing data. Shafer (1997) further outlines the use of the EM algorithm and data augmentation to handle missing data and discusses similarities between the EM algorithm and data augmentation. The

EM algorithm has expectation and maximization steps, while data augmentation has an imputation step followed by a posterior step.

Much work has been done in spatial statistics, were observations are thought of as resulting from a stochastic process $\{Z(s) : s \in D\}$, where s represents a location and $D \in \Re^d$ (Cressie, 1993; Matheron, 1963). In addition to traditional geostatistical and likelihood approaches for data analysis, recently much work has been focused on applying Bayesian ideas to spatial data analysis. Prior specification for geostatistical spatial models is presented by Ecker and Gelfand (1997). Berger, de Oliveira and Sanso (2001) address non-informative prior specification resulting in the use of a special reference prior for the dependency parameter. The use of the reference prior ensures a proper joint posterior distribution.

A major goal in many spatial analysis is to identify areas of high contamination that may require clean-up. Detection of areas of extreme contamination using a Monto Carlo approximation to the Bayesian posterior predictive distribution for a set of predicted locations is discussed in de Oliveira and Ecker (2001). Ancona and Tawn (2002) discuss the use of conditional independence and integration to account for the censored observations in the data portion of the model analyzed via MCMC methods. Handling censored observations via a data augmentation procedure in a Bayesian spatial analysis has yet to be discussed. In this paper, we combine the ideas of data augmentation and a Bayesian spatial model to analyze left censored spatial data.

2 Censored data and data augmentation

Censoring is a type of missing data mechanism that is "non-ignorable" (Little and Rubin, 2002). If we were to throw out or ignore the censored observations, the resulting parameter estimates would be biased. One solution to the problem of censored data is to integrate the censored data out of the joint posterior distribution,

$\int p(\boldsymbol{\Theta}|\boldsymbol{Y}_c, \boldsymbol{Y}_o)p(\boldsymbol{Y}_c|\boldsymbol{Y}_o)d\boldsymbol{y}_c$. The problem with trying to implement this solution is the required integration, which may be very difficult. Another method for handling censored data is imputation, that is, to replace every censored observation with a real value. The question then becomes "What value to impute?" The easiest (but not always the best) method is to replace the censored values with a constant. Commonly, half the level of detection ($LOD/2$) is imputed. Another method would be to use some random imputation scheme to impute a value for the censored data (single imputation or multiple imputation). By imputing a constant (LOD or $LOD/2$), one is going to bias subsequent parameter estimation. Also, there is no sound justification for which value or constant to impute. The advantage of using multiple imputation over single imputation is that one is able to quantify the additional error in estimation due to the imputation.

The main issue in a single or multiple imputation scheme is which parameter values (based on the model) to use for the imputation? The answer to this question is to handle both the imputation for the censored data and the parameter estimation using Markov Chain Monte Carlo sampling. That is, it is possible to apply the idea of data augmentation as proposed by Tanner and Wong (1987) to the case of censored spatial data. The idea is as follows.

- Given the current value of the parameters $\boldsymbol{\Theta}^{(t)}$, draw a vector $\boldsymbol{Y}_c^{(t+1)}$ for the censored data from $p(\boldsymbol{Y}_c|\boldsymbol{Y}_o, \boldsymbol{\Theta}^{(t)})$.

- Then based on $\boldsymbol{Y}_c^{(t+1)}$, draw $\boldsymbol{\Theta}^{(t+1)}$ from $p(\boldsymbol{\Theta}|\boldsymbol{Y}_o, \boldsymbol{Y}_c^{(t+1)})$, the complete data posterior for $\boldsymbol{\Theta}$.

At every iteration of the simulation we are "augmenting" the data with imputed values for the censored observations. In doing so, we have eliminated the need to work with the observed data posterior $p(\boldsymbol{\Theta}|\boldsymbol{Y}_o)$, which in many cases is intractable or difficult to obtain. This process yields a stochastic sequence $\{\boldsymbol{\Theta}^{(t)}, \boldsymbol{Y}_c^{(t)} : t = 1, 2, ...\}$ whose stationary

distribution is $p(\boldsymbol{\Theta}, \boldsymbol{Y}_c|\boldsymbol{Y}_o)$ (Shafer, 1997; Gilks, Richardson and Spiegelhalter, 1996). Data augmentation can be thought of as using Markov chain Monte Carlo to perform imputation.

Data augmentation can also be looked at as a method that solves the problem of having to integrate out the censored observations from $p(\boldsymbol{\Theta}|\boldsymbol{Y}_c, \boldsymbol{Y}_o)$. As presented in Tanner and Wong (1987), let y=observed data, z=augmentated data (missing data), θ = parameters. If both y and z are observed, then $p(\theta|y, z)$ is easily calculated, whereas, if z is not observed, $p(\theta|y) = \int p(\theta|y, z)p(z|y)dz$ may be difficult to calculate. Thus, multiple values of z for augmentation are generated from the predictive distribution $p(z|y)$ in two steps. The first steps is to generate a value of θ, say ϕ, and based on this value ϕ, the second step is to generate z from $p(z|\phi, y)$. Then, $p(\theta|y)$ can be approximated by averaging $p(\theta|y, z)$ over the generated values of z (i.e. $\int p(\theta|y, z)p(z|y)dz$).

3 Spatial Bayesian model and prediction

Define $\{Y(s) : s \in D\}$ to be a spatial stochastic process, where s varies continuously over D, D in $\Re^2$. We specify a spatial isotropic model as

$$Y(s_i) = \mu + W(s_i) + \varepsilon(s_i), \tag{1}$$

where $Y(s_i)$ represents the observation at location s_i, μ is the overall mean, $\varepsilon(s_i)$ represents the random observational error at location s_i with $\varepsilon(s_i) \sim \text{NOR}(0, \tau^2)$, and $W(s_i)$ represents the random spatial effect at location s_i with $\boldsymbol{W}(\boldsymbol{s}) \sim \text{MVN}(\boldsymbol{0}, V(\sigma^2, \phi))$ where $V(\sigma^2, \phi)_{ij} = \sigma^2 \exp\{-d_{ij}/\phi\}$, $d_{ij} = \|s_i - s_j\|$ and $V^*(\phi)_{ij} = \exp\{-d_{ij}/\phi\}$. Note, there are various alternate ways to parameterize $V(\cdot)$.

Hence, we have $\boldsymbol{Y} \sim \text{MVN}(\boldsymbol{\mu}, V(\sigma^2, \phi) + \tau^2 I)$ and $\boldsymbol{Y}|\boldsymbol{W} \sim \text{MVN}(\boldsymbol{\mu} + \boldsymbol{W}, \tau^2 I)$. To complete the Bayesian model specification, prior distributions are put on all parameters in the model. There are various choices for the prior specifications, ranging from elicited

or conjugate proper priors to non-informative or improper priors. Whatever the choice, sensitivity analysis for the final inferences with respect to the prior distributions is recommended.

Prior specification involving non-informative, improper priors would be

$$p(\sigma^2) \propto (\sigma^2)^{-1},$$
$$p(\tau^2) \propto (\tau^2)^{-1},$$
$$p(\mu) \propto 1,$$
$$p(\phi) \propto \left(tr[W_\phi^2] - \tfrac{1}{(n-p)}(tr[W_\phi])^2\right)^{1/2},$$

where $W_\phi = \left(\left(\frac{\partial}{\partial\phi}\right)\Sigma_\phi\right)\Sigma_\phi^{-1}P_\phi^\Sigma$, $P_\phi^\Sigma = I - \mathbf{1}(\mathbf{1}^T\Sigma_\phi^{-1}\mathbf{1})^{-1}\mathbf{1}^T\Sigma_\phi^{-1}$, $\Sigma_{\phi,ij} = K_\phi(\|s_i - s_j\|)$, and $K_\phi(\|\boldsymbol{s} - \boldsymbol{u}\|) = \text{corr}\{Z(\boldsymbol{s}), Z(\boldsymbol{u})\}$ is an isotropic correlation function (Berger, de Oliveira, and Sanso, 2001). As discussed in Berger, de Oliveira and Sanso (2001), it is this reference prior for ϕ that ensures a proper joint posterior distribution.

An alternative spatial Bayesian model would be to place proper prior distributions on all the parameters in the following fashion:

$$\sigma^2 \sim \text{INGAM}(\alpha, \beta),$$
$$\tau^2 \sim \text{INGAM}(\gamma, \delta),$$
$$\mu \sim \text{NOR}(\lambda, \psi^2),$$
$$\phi \sim \text{GAM}(\eta, \theta).$$

One thing to note is that there is no conjugate prior for ϕ leading to easy computation of the full conditional distribution for ϕ.

In addition to fitting a model to produce parameter estimates, prediction is often a goal of spatial studies. Let $\boldsymbol{Y}_u$ represent an ungauged (unobserved) vector and $\boldsymbol{Y}_g$ represent a gauged (observed) vector. Bayesian prediction uses the posterior predictive distribution, $p(\boldsymbol{Y}_u|\boldsymbol{Y}_g)$, as the method for prediction. The joint distribution of $\boldsymbol{Y}_u$ and $\boldsymbol{Y}_g$ can be written as

$$\text{MVN}\left(\begin{pmatrix} \boldsymbol{\mu}_u \\ \boldsymbol{\mu}_g \end{pmatrix}, \begin{pmatrix} \Sigma_{uu} & \Sigma_{ug} \\ \Sigma_{gu} & \Sigma_{gg} \end{pmatrix}\right),$$

resulting in a multivariate normal distribution as the conditional distribution for the ungauged sites given the gauged sites. Approximation of the posterior predictive distribution can be accomplished by simulating predictions from this multivariate normal distribution.

For the case involving censored data and data augmentation, the approximation of the posterior predictive distribution can be modified to account for the censored observations. Let $\boldsymbol{Y}_u$,$\boldsymbol{Y}_g$, $\boldsymbol{Y}_{go}$, and $\boldsymbol{Y}_{gc}$ represent the ungauged vector, gauged vector, gauged observed vector and the gauged censored vector, respectively. Approximation of the posterior predictive distribution, $p(\boldsymbol{Y}_u|\boldsymbol{Y}_g)$, is accomplished by simulating predictions from

$$\boldsymbol{Y}_u|\boldsymbol{Y}_{go}, \boldsymbol{Y}_{gc}^{(k)}, \boldsymbol{\Theta}^{(k)} \sim \text{MVN}(\boldsymbol{\mu}_{u.g}^{(k)}, \Sigma_{u.g}^{(k)}),$$

with $\boldsymbol{\mu}_{u.g}^{(k)} = \boldsymbol{\mu}_u^{(k)} + \Sigma_{ug}^{(k)}\Sigma_{gg}^{-1(k)}(\boldsymbol{Y}^{*(k)} - \boldsymbol{\mu}_g^{(k)})$, $\Sigma_{u.g}^{(k)} = \Sigma_{uu}^{(k)} - \Sigma_{ug}^{(k)}\Sigma_{gg}^{-1(k)}\Sigma_{gu}^{(k)}$, and $\boldsymbol{Y}^{*(k)} = (\boldsymbol{Y}_{go}^T, \boldsymbol{Y}_{gc}^{(k)T})^T$, for various MCMC iterations k (Carlin and Louis, 1996; de Oliveira and Ecker, 2002; Gelman, Carlin, Stern and Rubin, 1995).

4 Markov chain Monte Carlo for data augmentation

For the analysis of censored spatial data modeled with a Bayesian spatial or geostatistical model with proper priors, data augmentation can be completed within a Gibbs sampler (Tanner and Wong, 1987; Geman and Geman, 1984). The Gibbs sampler is a special case of the data augmentation procedure outlined by Tanner and Wong in which t=1, where t is the number of augmentated datasets created at iteration to approximate the current posterior distribution.

With the Baysian spatial model involving censored data satisfying the assumptions for the Gibbs sampler, the data augmentation procedure can be completed as follows. At each iteration of the Gibbs sampler, the censored data are imputed by generating values from $\boldsymbol{Y}_c$'s full conditional distribution, $p(\boldsymbol{Y}_c \,|\boldsymbol{Y}_o, \mu, \tau^2, \sigma^2, \phi)$, using the auxillary information that $\boldsymbol{Y}_c < \boldsymbol{LOD}$. This results in an augmented complete dataset. Using this updated complete dataset, the parameters μ, τ^2, σ^2 and ϕ are generated from their corresponding full conditional distributions. This process yields a sequence $\{\boldsymbol{\Theta}^t, \boldsymbol{Y}_c^t : t = 1, 2, ...\}$ that is a stochastic process with stationary distribution $p(\boldsymbol{\Theta}, \boldsymbol{Y}_c|\boldsymbol{Y}_o)$, where $\boldsymbol{\Theta}$ contains μ, τ^2, σ^2 and ϕ. Derivations of the full conditional distributions required for the Gibbs sampler are located in the appendix. The MCMC data augmentation algorithm is as follows.

1. Set starting values for $\mu^{(0)}$, $\tau^{2(0)}$, $\sigma^{2(0)}$, $\boldsymbol{W}^{(0)}$, and $\phi^{(0)}$. Set $m = 0$.

2. Set censored values equal to their level of detection, $\boldsymbol{Y}_c^{(0)} = \boldsymbol{LOD}$. Let $\boldsymbol{Y}^{T(m)} = (\boldsymbol{Y}_c^{(m)}, \boldsymbol{Y}_o)^T$, where $\boldsymbol{Y_c}$ and $\boldsymbol{Y_o}$ represent the censored data and observed data, respectively.

3. Generate $\mu^{(m+1)}$ from $\text{NOR}(\mu_1^{(m+1)}, \sigma_1^{2(m+1)})$, with $\mu_1^{(m+1)} = (\frac{\psi^2\tau^{2(m)}}{\tau^{2(m)}+\psi^2})[\frac{1}{\psi^2}\lambda + \frac{1}{\tau^{2(m)}}(\bar{Y}^{(m)} - \bar{W}^{(m)})]$ and $\sigma_1^{2(m+1)} = (\frac{1}{n})(\frac{\psi^2\tau^{2(m)}}{\tau^{2(m)}+\psi^2})$.

4. Generate $\tau^{2(m+1)}$ from $\text{INGAM}(n/2+\gamma, (1/2)(\boldsymbol{Y}^{(m)} - (\boldsymbol{\mu}^{(m+1)} + \boldsymbol{W}^{(m)}))^T(\boldsymbol{Y}^{(m)} - (\boldsymbol{\mu}^{(m+1)} + \boldsymbol{W}^{(m)})) + \delta)$.

5. Generate $\sigma^{2(m+1)}$ from $\text{INGAM}(n/2+\alpha, (1/2)\boldsymbol{W}^{T(m)}V^*(\phi^{(m)})^{-1}\boldsymbol{W}^{(m)} + \beta)$.

6. Generate $\boldsymbol{W}^{(m+1)}$ from $\text{MVN}(\boldsymbol{\mu}_w^{(m+1)}, \Sigma_w^{(m+1)})$, where $\boldsymbol{\mu}_w^{(m+1)} = [V^{-1}(\sigma^{2(m+1)}, \phi^{(m)}) + \frac{1}{\tau^{2(m+1)}}I]^{-1}[\frac{1}{\tau^{2(m+1)}}(\boldsymbol{Y}^{(m)} - \boldsymbol{\mu}^{(m+1)})]$ and $\Sigma_w^{(m+1)} = [V^{-1}(\sigma^{2(m+1)}, \phi^{(m)}) + \frac{1}{\tau^{2(m+1)}}I]^{-1}$.

7. Using Metropolis-Hastings step(s), simulate $\phi^{(m+1)}$ from

   $p(\phi|\mu^{(m+1)},\tau^{2(m+1)},\sigma^{2(m+1)},\boldsymbol{W}^{(m+1)},\boldsymbol{Y}^{(m)})$

   $\propto \frac{\phi^{\eta-1}}{|V^*(\phi)|^{1/2}} \exp\{\frac{-1}{2\sigma^{2(m+1)}}\boldsymbol{W}^{T(m+1)}V^*(\phi)^{-1}\boldsymbol{W}^{(m+1)} - \theta\phi\}.$

8. Now have $\boldsymbol{\Theta}^{(m+1)} = (\mu^{(m+1)},\tau^{2(m+1)},\sigma^{2(m+1)},\phi^{(m+1)},\boldsymbol{W}^{(m+1)})$.

9. Using $\boldsymbol{\Theta}^{(m+1)}$ and $\boldsymbol{Y}^{(m)}$, impute values for $\boldsymbol{Y}_c$ to produce $\boldsymbol{Y}_c^{(m+1)}$. Let

   $\boldsymbol{Y}_c = (Y_{1c}, Y_{2c}, .., Y_{kc})$.

   (a) Generate $Y_{1c}^{(m+1)}$ from $\mathrm{N}(\mu^{(m+1)} + W_1^{(m+1)}, \tau^{2(m+1)})$, truncated at LOD_1.

   ...

   (b) Generate $Y_{kc}^{(m+1)}$ from $\mathrm{N}(\mu^{(m+1)} + W_k^{(m+1)}, \tau^{2(m+1)})$, truncated at LOD_k.

10. Complete prediction for a set of locations based on $\boldsymbol{Y}^{(m+1)}$ and $\boldsymbol{\Theta}^{(m+1)}$.

11. Set $m = m + 1$ and repeat algorithm a large number of times.

By introducing the spatial random variable (W) to the model, the imputation step of the algorithm simplifies to the generation of values from univariate truncated normal distributions.

5 Illustrative example I: Missouri dioxin contamination

5.1 Description of data

In 1971, dioxin (2,3,7,8-tetrachlorodibenzo-p-dioxin or TCDD) contaminated waste was dumped along sections of a country road in Missouri. Vehicles, animals and precipitation have since transported some of the dioxin away from the original contaminated areas. As a result of the pollution, a number of animals died. In November of 1983, the USEPA investigated the contaminated site to determine which areas required clean-up.

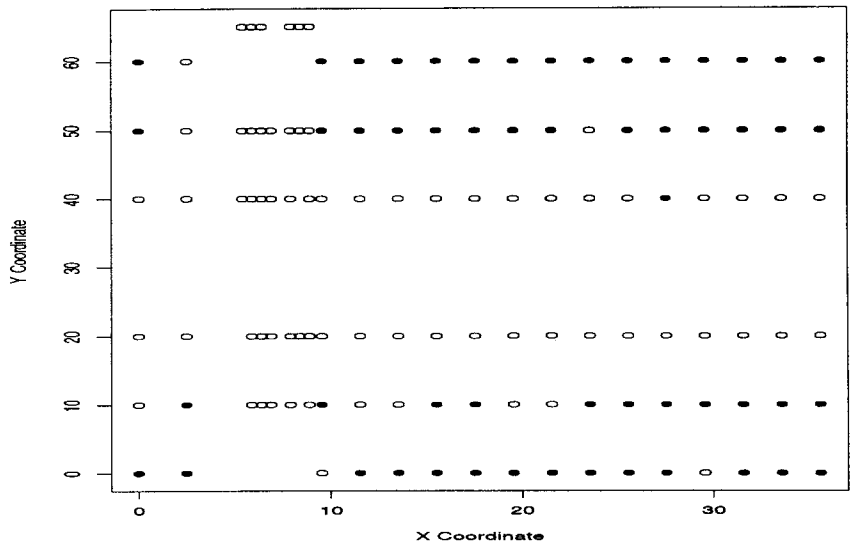

Figure 1 Missouri study locations, o represents an observed value and • represents a censored value

They sampled various areas, including the shoulder of the road, to determine their contamination levels. The data reported in Zirschky and Harris (1986), which only includes the sampled areas along the shoulder of the road, will be used to illustrate the use of data augmentation for spatial censored data. The goal of the analysis is to identify portions of the shoulder requiring clean-up.

The spatial directions are the X-direction (measured in $(\frac{1}{100})$feet), representing direction parallel to the road, and the Y-direction (measured in feet), representing the direction perpendicular to or away from the road. The road is located at the Y coordinate of 30. The shoulder of the road was divided into long transects in the X direction, most 200 feet, in which 8 samples were taken. The 8 samples were aggregated together to give one measurement per transect. For illustration purposes, we will treat the values reported as coming from one sampled location, with the X coordinate indicating the start of the transect.

From the samples taken, roughly 43% of the observations were censored, falling below some level of detection (LOD). The level of detections range from 0.10 μg/kg to 0.79 μg/kg. The level of detection is very "matrix" dependent; different amounts of soil, type of soil, moisture level, etc. may affect the limit of detection. All samples were analyzed according to USEPA approved procedures - USEPA, "Determination of 2,3,7,8-TCDD in Soil and Sediment", USEPA Region VII Laboratory, Kansas City, KS. The clean-up criteria for dioxin is 1 μg/kg. The goal is to perform spatial prediction that results in a map of predicted contamination levels for the entire area.

5.2 Model specification

The Bayesian spatial model assumes normality. A log transformation was applied to the original observations, resulting in a clean-up level of 0 ln(μg/kg). In addition to a log transformation of the data, a transformation of the original X coordinate by dividing by 100 was also required. This was due to a possible problem with the assumption of isotropy (no directional dependence). After initial investigation, there seemed to be a directional dependence in the data in the X direction. After transforming the X coordinate (i.e. defining a different distance measure), the isotropy assumption seemed reasonable.

The model for the analysis was the Bayesian spatial model outlined in Section 3 with prior distributions of $\mu \sim \text{NOR}(0, 50)$, $\sigma^2 \sim \text{INGAM}(2.1, 6.6)$, $\phi \sim \text{GAM}(2, 0.1)$, and $\tau^2 \sim \text{INGAM}(2.1, 0.55)$. These priors have large, but finite, variance with the distributions centered roughly around the parameter estimates found by replacing the censored values with their levels of detection in a non-Bayesian geostatistical analysis. For this analysis, the data were used to choose priors, but only to give a rough idea of appropriate prior means for the model parameters. Alternatively, a fully Bayesian analysis could be applied involving the specification of hyper-priors. Again, the question comes down to the specification of the hyper-prior parameters. The use of improper or

Table 1 Dioxin: Median and 95% credible intervals based on the simulated marginal posterior distributions

	DA		LOD/2		LOD	
	Median	Interval	Median	Interval	Median	Interval
μ	-0.701	(-1.744, 0.609)	-0.646	(-1.488, 0.338)	-0.441	(-1.305, 0.531)
τ^2	0.169	(0.076, 0.372)	0.193	(0.090, 0.383)	0.170	(0.083, 0.322)
σ^2	7.425	(3.85, 17.74)	4.122	(2.330, 9.178)	3.337	(1.783, 8.087)
ϕ	17.697	(8.93, 40.51)	15.760	(7.90, 36.51)	16.599	(7.96, 44.21)

flat priors for the hyper-parameters is an option, but care should be taken to insure a proper joint posterior distribution. As in the case of the first level priors, special consideration for the dependence parameter ϕ was needed in order to insure a proper joint distribution. In this case, a proper prior or a specific reference prior (Berger, de Olivieria and Sanso 2001) is required to insure a proper joint posterior distribution.

For the simulation of ϕ via Metropolis-Hastings step(s), the candidate generating distribution of GAM$(2X, 2)$ was used, where X represents the current value of ϕ. By choose GAM$(2X, 2)$, the mean of the candidate generating distribution for the current iteration of the chain is the current value for the random variable. At each iteration of the Gibbs sampler, 5 Metropolis-Hastings steps were completed for the simulation of ϕ. The chain was run for 10,000 iterations, excluding the first 500 iterations for burn-in. Convergence was checked via time-series plots constructed for each parameter.

5.3 Results

Summaries comparing the spatial analysis using data augmentation (DA) for censored observations to the method that replaces the censored observations with half the level of detection ($LOD/2$) or the level of detection (LOD) are presented in Table 1. Table 1 displays medians and 95% credible intervals for the parameters μ, τ^2, σ^2, and ϕ. In addition to numerical summaries, Figures 2 through 5 provide approximate marginal densities for the parameters μ, τ^2, σ^2 and ϕ using DA, LOD/2 and LOD methods for

Table 2 Dioxin: Point Estimates found using Weighted Least Squares

	LOD/2	LOD
μ	-0.871	-0.569
τ^2	0.502	0.435
σ^2	5.876	4.423
ϕ	20.672	21.430

handling censored data. From these results, one notices in addition to difference in posterior medians, the data augmentation procedure produced larger variability in the approximated marginal densities as compared to the LOD/2 and the LOD methods. The biggest difference between the three methods is in the estimation of the spatial variability parameter σ^2. The median of the posterior distribution for σ^2 is 7.425 using data augmentation, while half the level of detection and the level of detection methods produce medians of 4.122 and 3.337, respectively.

A comparison of the Bayesian method and the traditional method to find estimates using Weighted Least Squares to fit a variogram model was also investigated. The estimates found using Weighted Least Squares (WLS) are presented in Table 2. The table presents results based on replacing the censored observations with half the level of detection (LOD/2) and the level of detection (LOD). Comparing the results in Table 1 and Table 2, we see that the WLS method produced slightly larger point estimates, with the largest difference in regards to the estimation of τ^2. Overall, the two methods agree fairly well, with the possible difference between the two methods due to the prior specification involved in the Bayesian analysis.

Since the goal of this study is the identification of areas requiring clean-up based on a criteria of 0 $\ln(\mu/\text{kg})$, Bayesian prediction results are presented in Figures 6 and 7. Figure 6 displays the median of the approximated data augmentation Bayesian posterior predictive distribution. Figure 7 contain corresponding graphs of the posterior proba-

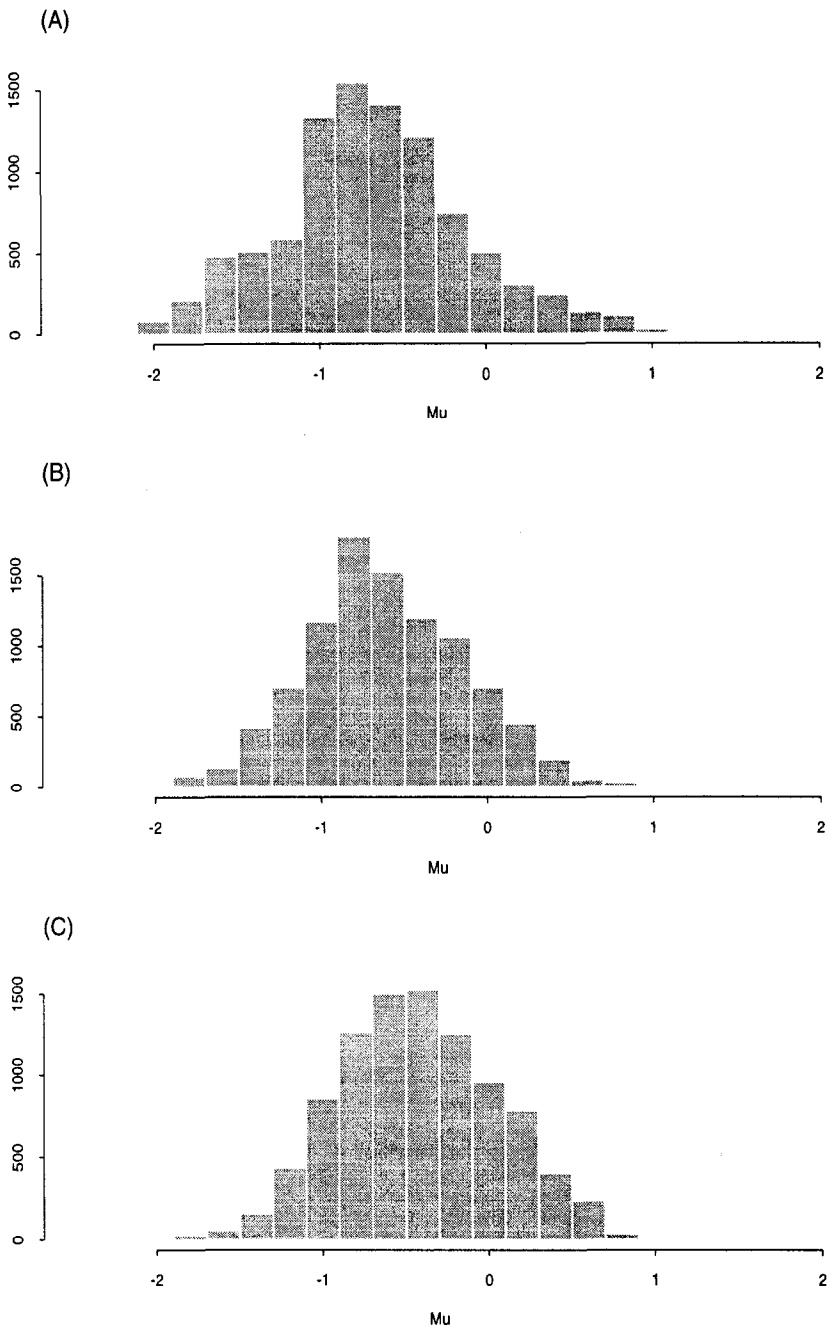

Figure 2 Dioxin: Simulated marginal posterior distributions for μ (A) data augmentation for censored values (B) censored values replaced by $LOD/2$ (C) censored values replaced by LOD

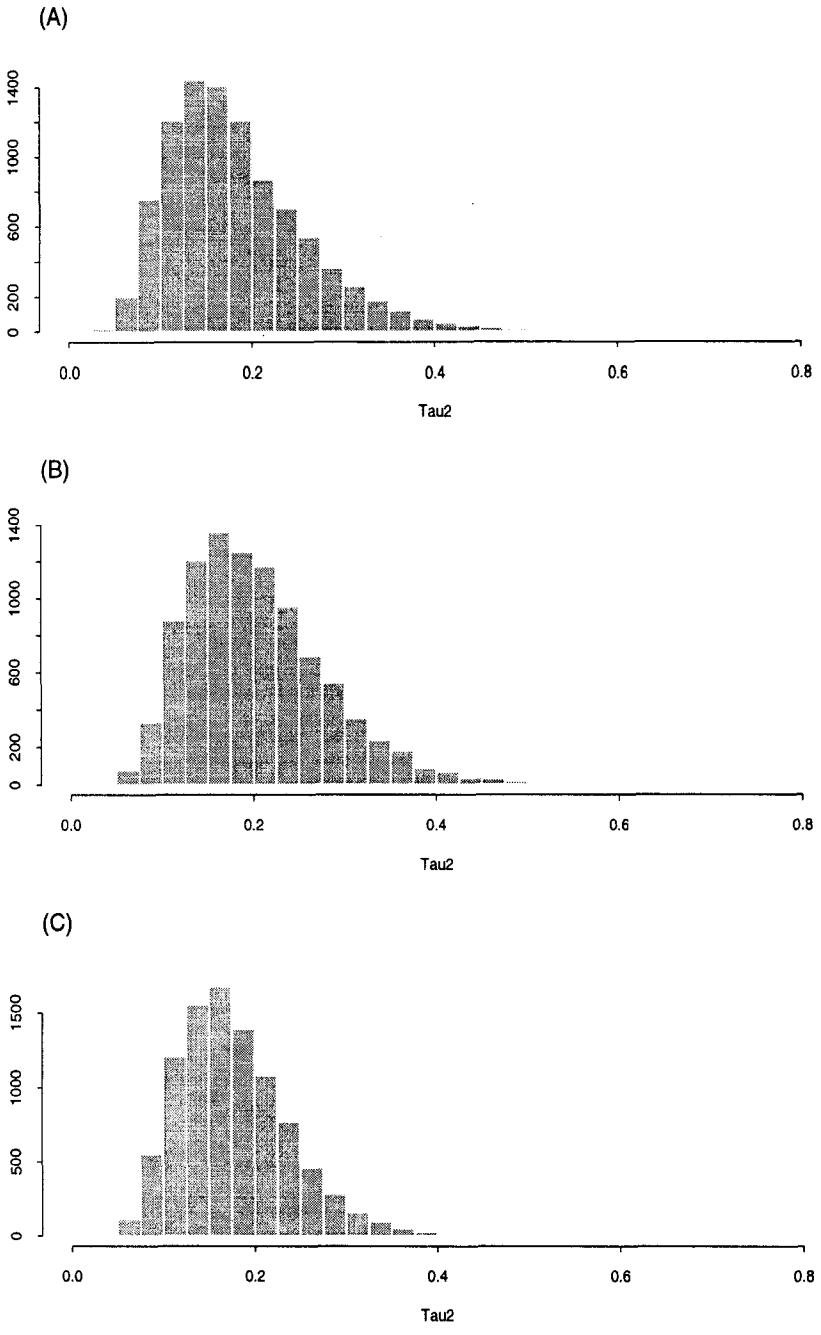

Figure 3 Dioxin: Simulated marginal posterior distributions for τ^2 (A) data augmentation for censored values (B) censored values replaced by $LOD/2$ (C) censored values replaced by LOD

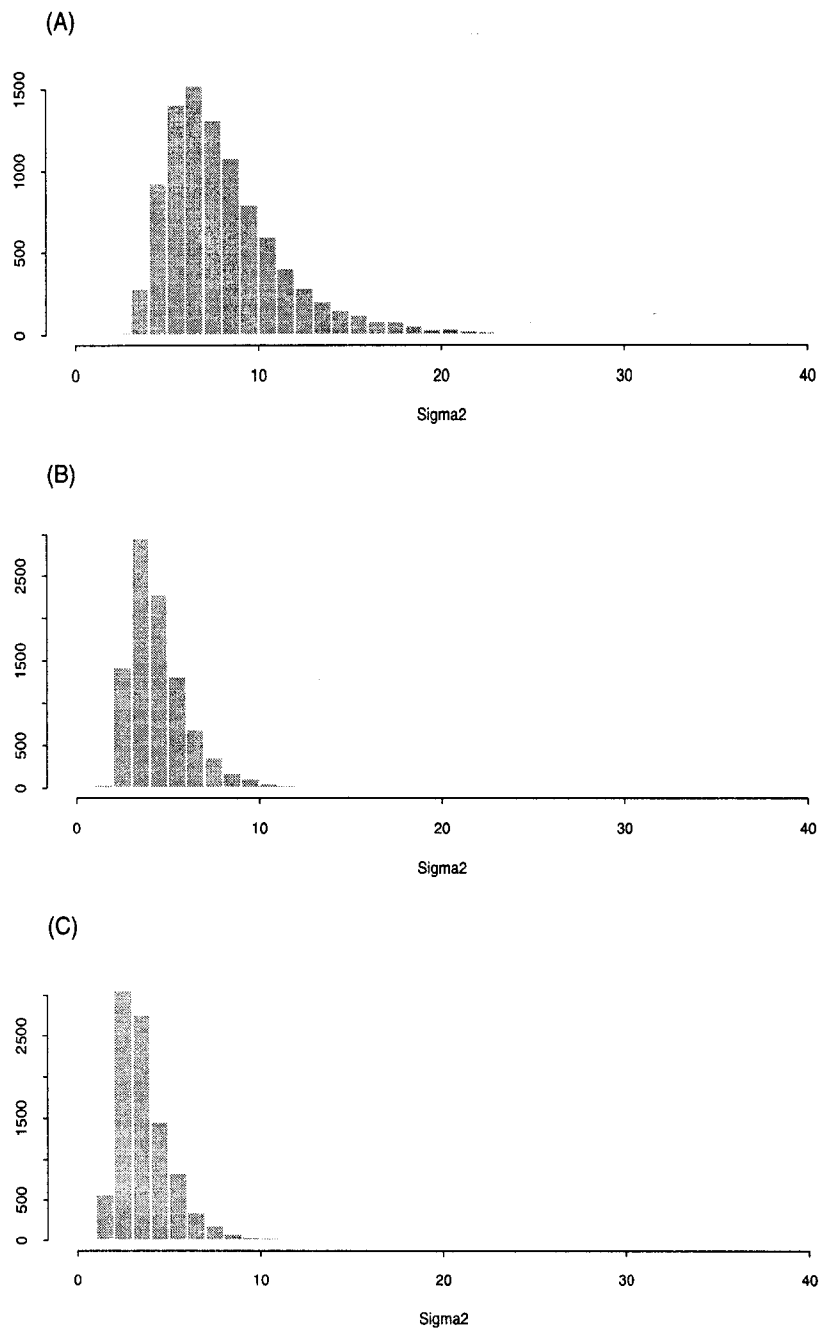

Figure 4 Dioxin: Simulated marginal posterior distributions for σ^2 (A) data augmentation for censored values (B) censored values replaced by $LOD/2$ (C) censored values replaced by LOD

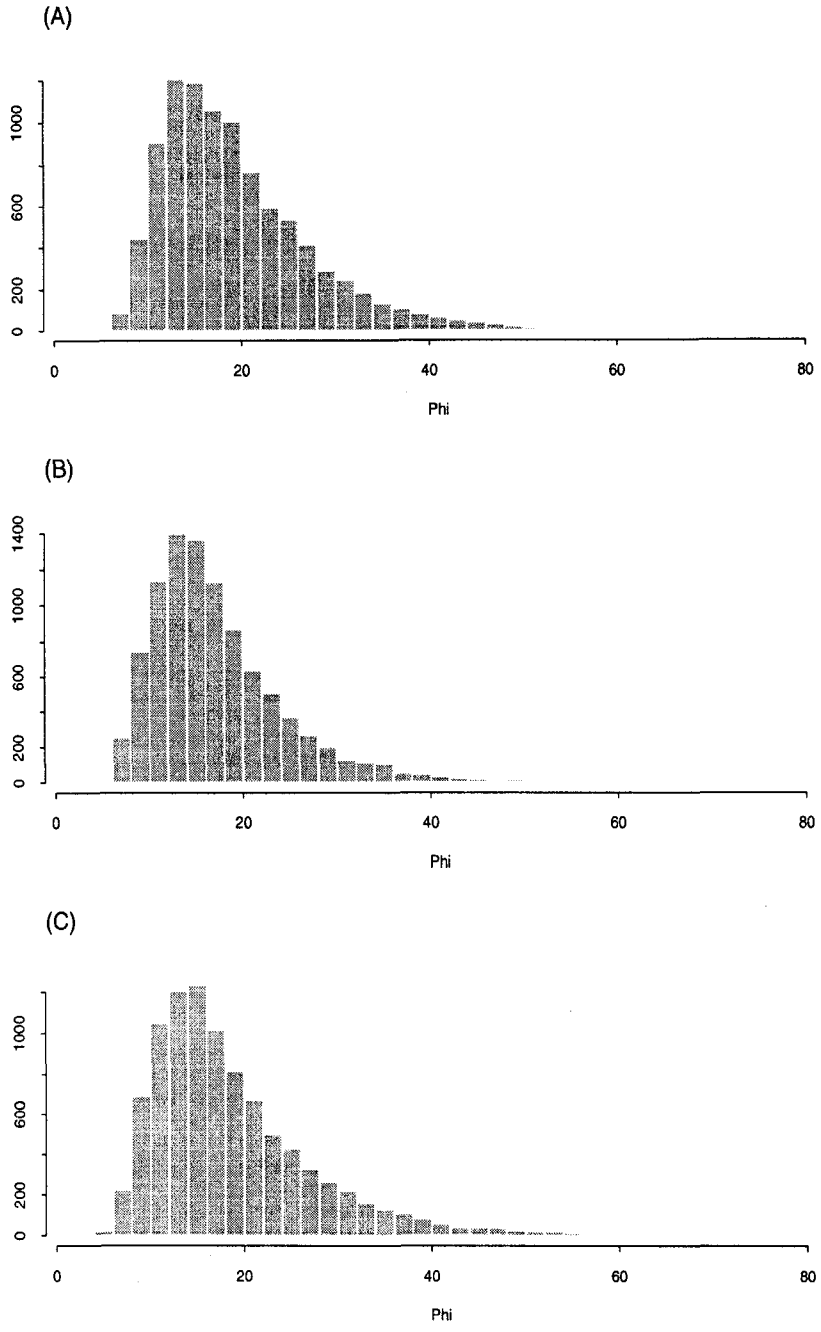

Figure 5 Dioxin: Simulated marginal posterior distributions for ϕ (A) data augmentation for censored values (B) censored values replaced by $LOD/2$ (C) censored values replaced by LOD

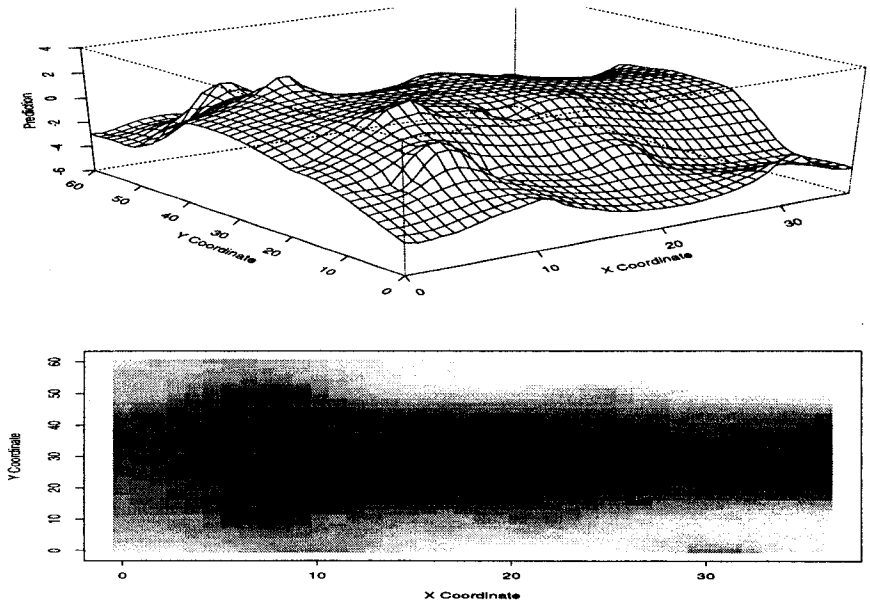

Figure 6 Dioxin: Posterior median of the Bayesian predictive distribution using data augmentation for censored values

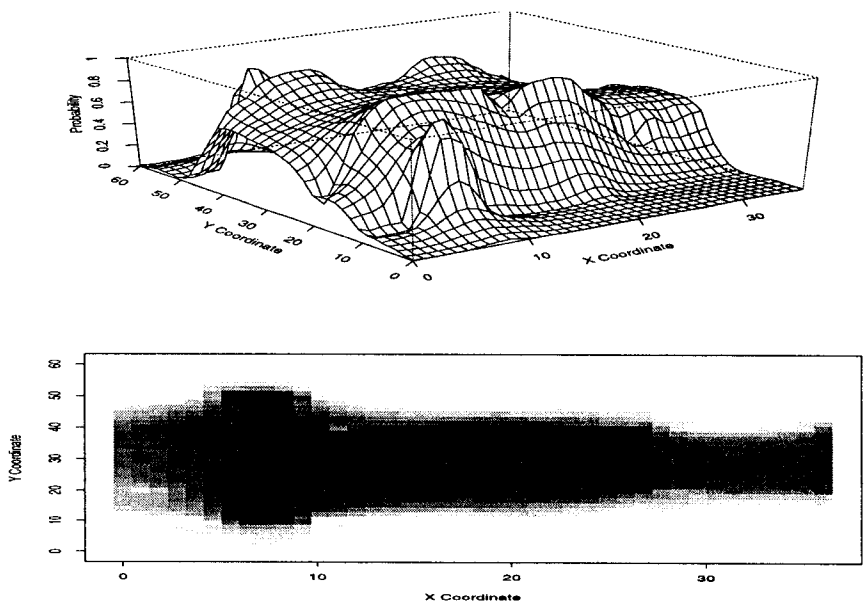

Figure 7 Dioxin: Posterior probability of prediction being greater than the clean-up criteria using data augmentation for censored values

bility of a prediction being greater than the clean-up criteria of 0 $\ln(\mu/\text{kg})$. Based on these plots or other summaries of the posterior predictive distribution, clean-up decisions can be made which better reflect the true contamination levels, by accounting for the censored observations adequately.

Figures 8 and 9 provide comparison of predictions produced by the DA and LOD/2 methods. These figures portray the difference in medians of posterior predictive distributions and posterior predictive probability being greater than clean-up criteria produced by using the DA and LOD/2 methods. The two figures show that setting censored observations equal to half the level of detection resulted in larger predictions in the areas far away from the road (Y direction), in particular for locations far down the road (in the positive X direction). With respect to the posterior probability of a location's contamination being greater than the clean-up criteria of 0 $\ln(\mu\text{g/kg})$, there are two major areas of discrepancy; along the shoulder (Y coordinates of 10 to 20 and 40 to 50) and at large values of X located on the road (Y coordinates around 30). Along the shoulder, replacing the censored values with half the level of detection resulted in larger posterior probabilities of contamination while data augmentation produced larger probabilities on and around the road at large values of the X coordinate.

To illustrate the difference in the clean-up regions determined by the DA and LOD/2 methods, Figure 10 contains contour plots for the probability being greater than the clean-up criteria were plotted for probabilities of 0.60, 0.70, 0.80, and 0.90. These probabilities of being greater than the clean-up criteria can be used to determine which areas needed to be cleaned up. The clean-up region is the area inside the plotted line, where a smaller clean-up region was found using the DA method as compared to the LOD/2 method. For this study, there was a moderate difference in the clean-up regions. Other studies may show larger difference in clean-up regions or no difference in clean-up regions; the DA method produces better parameter estimates and predictions which in some examples still will not result in any meaningful difference in clean-up regions

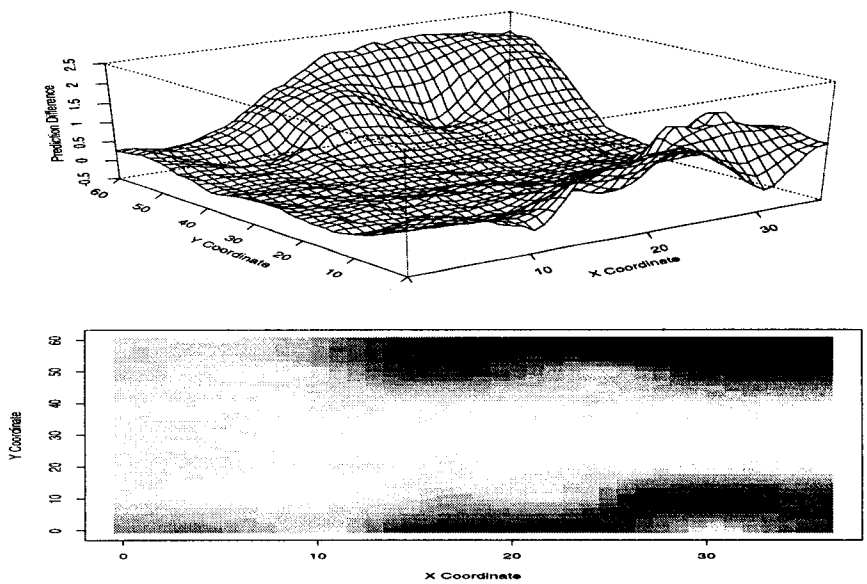

Figure 8 Dioxin: Difference in posterior medians for DA and LOD/2 methods for handling censored values (LOD/2 - DA)

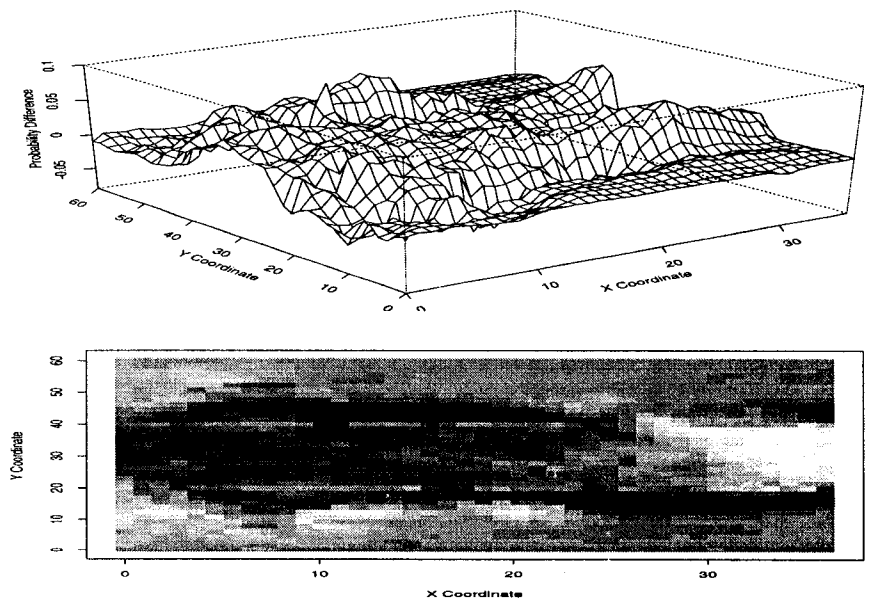

Figure 9 Dioxin: Difference in posterior predictive probability being greater than the clean-up criteria for DA and LOD/2 methods (LOD/2 - DA)

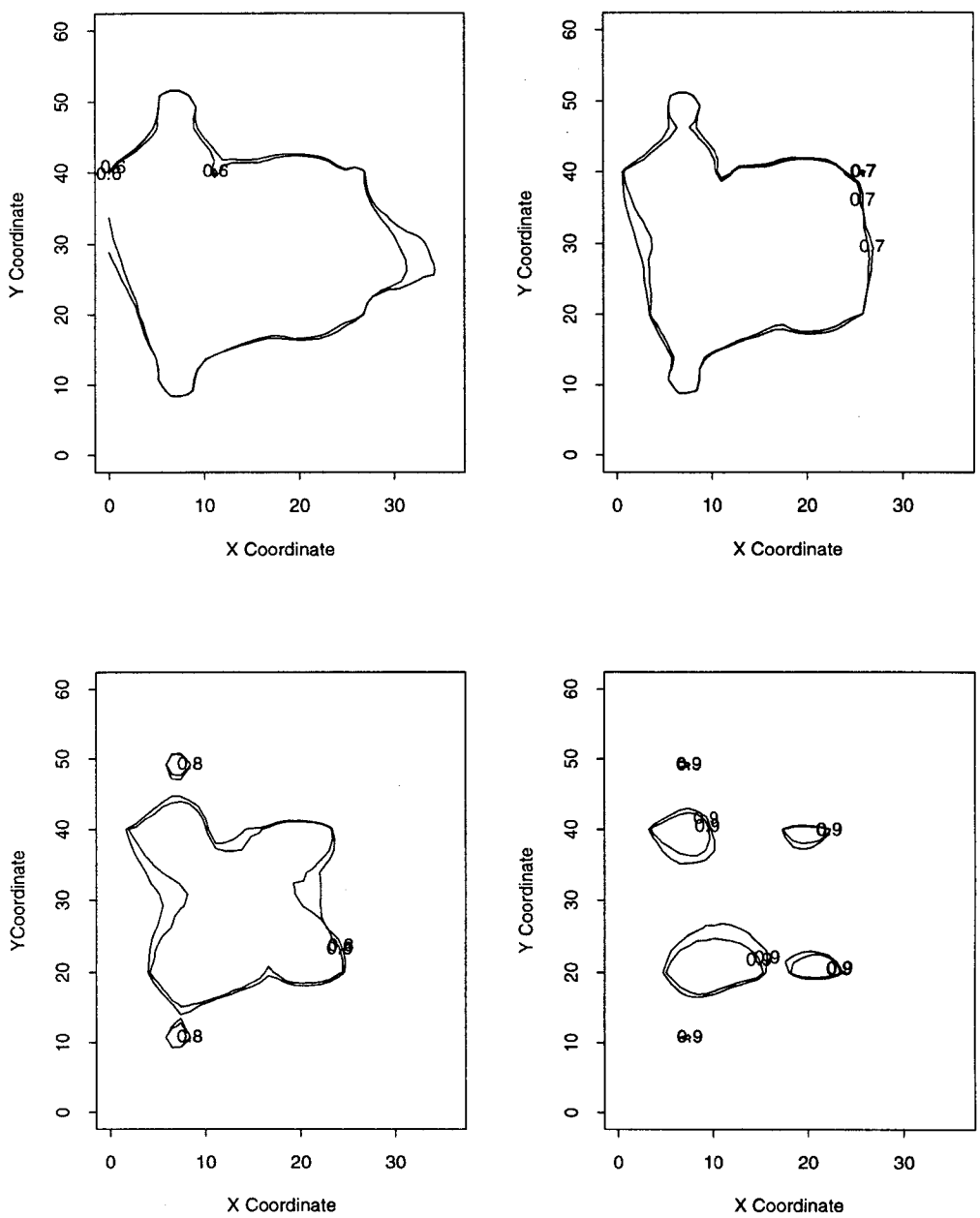

Figure 10 Dioxin: Difference in clean-up regions between the DA method (interior line) and the LOD/2 method (exterior line) based on different probabilities of being above clean-up cut-off values

Table 3 Dioxin: Median and 95% credible intervals based on the simulated marginal posterior distributions for three different prior specifications

	Primary Analysis		**Second Analysis**		**Third Analysis**	
	Median	Interval	Median	Interval	Median	Interval
μ	-0.701	(-1.744, 0.609)	-0.144	(-0.917, 0.519)	-0.681	(-1.901, 0.285)
τ^2	0.169	(0.076,0.372)	0.209	(0.105, 0.434)	0.246	(0.119, 0.509)
σ^2	7.425	(3.853,17.740)	7.951	(4.727,13.850)	7.394	(3.792, 17.087)
ϕ	17.697	(8.931,40.511)	19.373	(11.828, 31.263)	18.493	9.448, 41.927)

between the DA and LOD/2 methods.

Lastly, sensitivity analysis was performed to investigate the impact of the prior distributions on the parameter estimates. Two more analyses were completed using prior distributions of $\mu \sim \text{NOR}(0, 20)$, $\sigma^2 \sim \text{INGAM}(3, 12)$, $\phi \sim \text{GAM}(10, 0.5)$, $\tau^2 \sim \text{INGAM}(3, 1)$ and $\mu \sim \text{NOR}(0, 100)$, $\sigma^2 \sim \text{INGAM}(2.1, 4.4)$, $\phi \sim \text{GAM}(2.2, 0.1)$, $\tau^2 \sim \text{INGAM}(2.1, 1.1)$. Comparison of parameter estimates for the primary analysis and the two additional analyses can be seen in Table 3. As Table 3 presents, there are only small differences among the three analyses in terms of parameter estimation, with the largest differences for the estimation of μ. Overall, the priors used in the primary analysis seem appropriate.

6 Illustrative example II: site 15

6.1 Description of data

Site 15, an old industrial site converted into a park, is a site of metal contamination. A study was conducted to investigate the level of metal contamination at the site. The main purpose of the study was to determine the amount of Metal C present in the soil and whether clean-up was required. The second goal of the study was to determine if Metal C was associated with other metals of interest (e.g. Metal B).

A soil core was drilled at each sampled location and measurements taken at different depths. The depths intervals (which varied from location to location) were determined by the type of soil and soil characteristics, with no information available beyond the depth sampled. That is, no information was available on the type of soil, only the depth. To demonstrate the data augmentation procedure, we investigated the amount of Metal B present in the soil. The clean-up criteria for Metal B is 1 mg/kg for both residential and non-residential areas. Censored observations occurred for metal B due to detection limits or LOD (i.e. left censoring). The detection limits varied, due partially to the amount of sample analyzed and the amount of moisture in the sample.

To illustrate the data augmentation method and comparison to the LOD/2 and LOD methods for the handling of censored spatial data in the context of a Bayesian Spatial model, only the second depth measurements were analyzed (i.e. measurements right below the topsoil). The data augmentation procedure can be easily extended to the 3-dimensional setting. For the site 15 dataset, it was not clear how to handle the depth dimension, since the depths varied from location to location with no information recorded on the type of soil. Of the 82 observations, 32 (39%) were censored with the largest LOD being 1.5 mg/kg. Thus, the highest LOD is greater than the clean-up criteria for residential areas with a moderate amount of the data censored. Sampled locations for Metal B are displayed in Figure 11.

6.2 Model specification and results

The analysis of Metal B was completed using the spatial model outlined in Section 3. A log-transformation was applied to the original response variable to meet the normality assumption required for the model. This resulted in a clean-up criteria of 0 ln(mg/kg). After initial investigation and diagnostics, the assumption of isotropy seemed reasonable. To complete the Bayesian model, proper priors were placed on the parameters. The priors were chosen to have large, finite variances with the distributions centered around

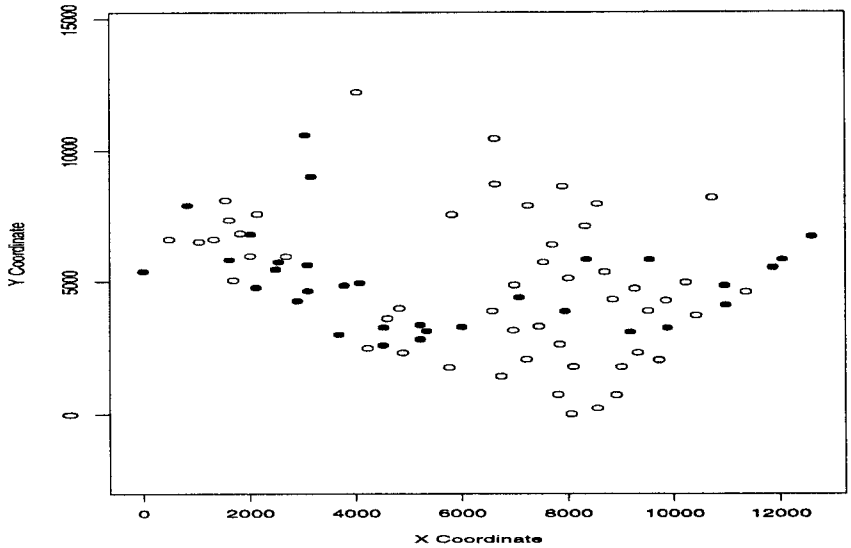

Figure 11 Sampled locations of Metal B, o represent observed values and • represent censored values

reasonable values found from initial analysis and investigation. The prior distributions used were $\mu \sim \text{NOR}(-1, 50)$, $\tau^2 \sim \text{INGAM}(2.1, 1.65)$, $\sigma^2 \sim \text{INGAM}(2.1, 2.97)$ and $\phi \sim \text{GAM}(50, 0.1)$. One thing to note is that the two closest points in the study are roughly 200 units apart, with the X coordinates and Y coordinates ranging roughly from 0 to 12,000. The prior distribution for ϕ results in a average range being 1500. That is, responses observed at locations more than 1500 units apart can be considered independent.

For the simulation of ϕ, 5 Metropolis-Hastings steps were completed at each iteration of the Gibbs sampler, where the candidate generating distribution was $\text{GAM}(2X, 2)$ with X representing the current value of ϕ. The value 2 is thought of as a "tuning parameter" that can be changed to increase "mixing" of the chain. The chain was run for 10,000 iterations, excluding the first 500 iterations for burn-in. Convergences was checked using time-series plots for all model parameters. The analysis was run three times, using either

Table 4 Metal B: Median and 95% credible intervals based on the simulated marginal posterior distributions

	DA		LOD/2		LOD	
	Median	Interval	Median	Interval	Median	Interval
μ	-1.807	(-2.263, -1.391)	-1.465	(-1.796, -1.124)	-1.217	(-1.534, -0.903)
τ^2	1.189	(0.329, 3.190)	0.875	(0.305, 2.074)	0.841	(0.296, 1.815)
σ^2	2.089	(0.676, 4.033)	1.532	(0.570, 2.611)	1.298	(0.545, 2.267)
ϕ	14.459	(11.43, 17.89)	14.434	(11.51, 17.81)	14.462	(11.50, 17.83)

the data augmentation (DA) method, the LOD/2 method or the LOD method as the means to handle the censored observations.

The results of the three analysis are displayed in Table 4 and Figures 12 to 15. As with the Missouri dioxin example, the level of spatial variability (σ^2) is vastly under-estimated using either the LOD/2 or the LOD method. In addition to the difference in the estimation of σ^2, there was also a difference in the estimation of μ and τ^2 between the three methods. Data augmentation produced an estimate (median of simulated posterior density) of τ^2 to be 1.189, while LOD/2 and LOD produced estimates of 0.875 and 0.841. Another interesting difference between the DA and the LOD/2 and LOD methods is the amount of variability in the posterior distributions. Like the Missouri dioxin example, data augmentation produced posterior distributions with more variability. In other words, the LOD/2 and LOD methods are under-estimating the true posterior variability. Thus, credible intervals produced from posterior distributions found using the LOD/2 and LOD methods tend to be too small. Lastly, there was little difference between the methods with regards to the estimation of ϕ (i.e. range parameter). The estimates of ϕ indicate no spatial dependence present in the data.

Results from analyses performed using Weighted Least Squares (WLS) are presented in Table 5. Comparison of the WLS method to the Bayesian method, in which the censored observations are replace with either the $LOD/2$ or LOD, show similar results

Table 5 Dioxin: Point Estimates found using Weighted Least Squares

	LOD/2	LOD
μ	-1.495	-1.224
τ^2	1.332	1.252
σ^2	1.506	1.206
ϕ	1709	1773

with the except of ϕ and τ^2. As with the Missouri example, WLS produced much larger estimates of τ^2 as compared to the Bayesian method. As for the large estimates of ϕ produced via WLS, this may be due to the fact that with very little spatial dependence present in the data. The standard errors estimates produced for the estimates of ϕ were on the same order of magnitude as the estimates, thus indicating little precision in estimation.

With a major goal of spatial analysis the prediction and identification of areas requiring clean-up, Bayesian prediction was also completed. Bayesian prediction in the setting of censored spatial data is outlined in Section 3. Displays of the predictions (i.e. medians of posterior predictive distributions) found via the DA method and the difference between the LOD/2 and DA methods are located in Figures 16 and 17. Since little spatial dependence is present in the data, predictions for the DA method are close to -1.8 ln(mg/kg), the estimate of μ. Likewise, the difference in predictions between the LOD/2 and the DA methods is close to 0.35 ln(mg/kg), the difference in the estimates of μ for the LOD/2 and DA methods. Based on the clean-up criteria of 0 ln(mg/kg), there does not seem to be excessive amount of Metal B present in the soil. For this example, both the LOD/2 and the DA methods produce similar conclusions in terms of clean-up. But, in terms of prediction and parameter estimation, the data augmentation procedure produced vastly different marginal densities for μ, τ^2 and σ^2 along with producing smaller predictions as compared to the LOD/2 method.

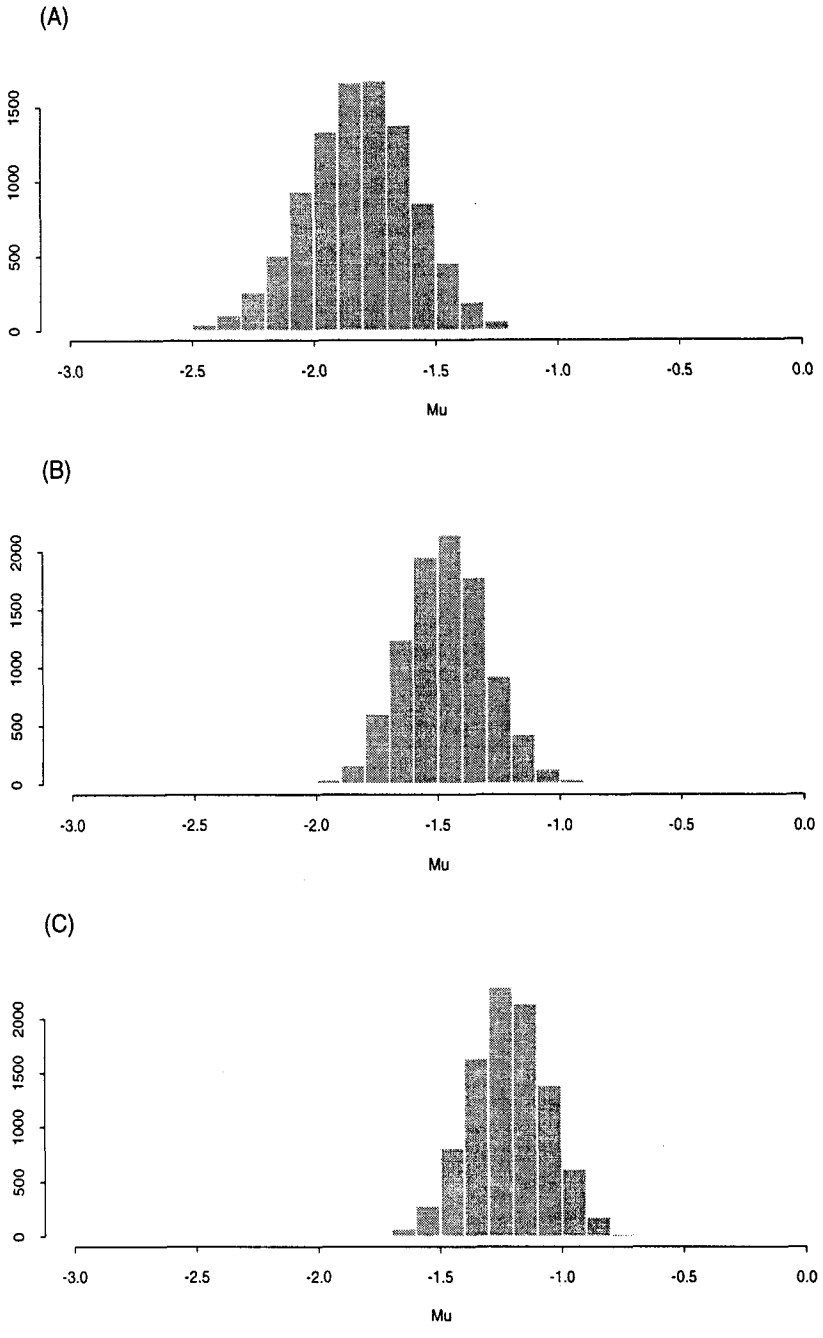

Figure 12 Metal B: Simulated marginal posterior distributions for μ (A) data augmentation for censored values (B) censored values replaced by $LOD/2$ (C) censored values replaced by LOD

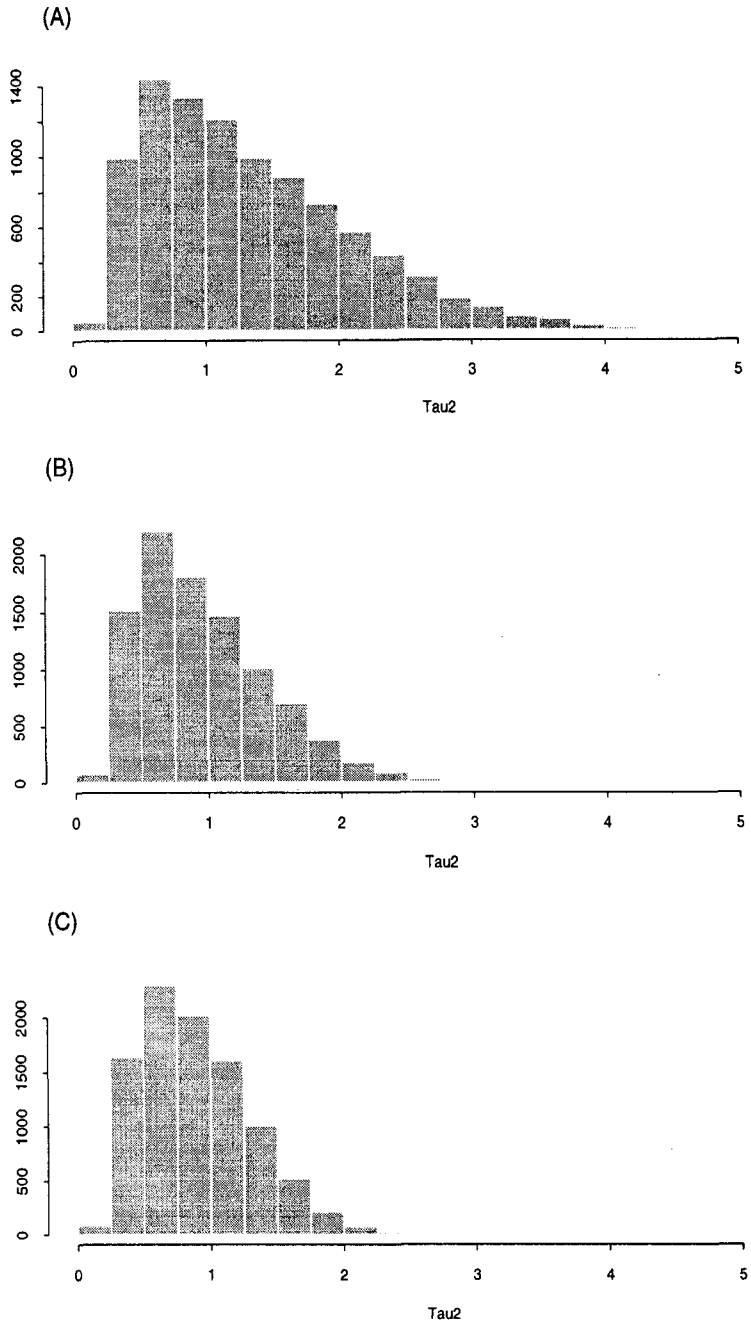

Figure 13 Metal B: Simulated marginal posterior distributions for τ^2 (A) data augmentation for censored values (B) censored values replaced by $LOD/2$ (C) censored values replaced by LOD

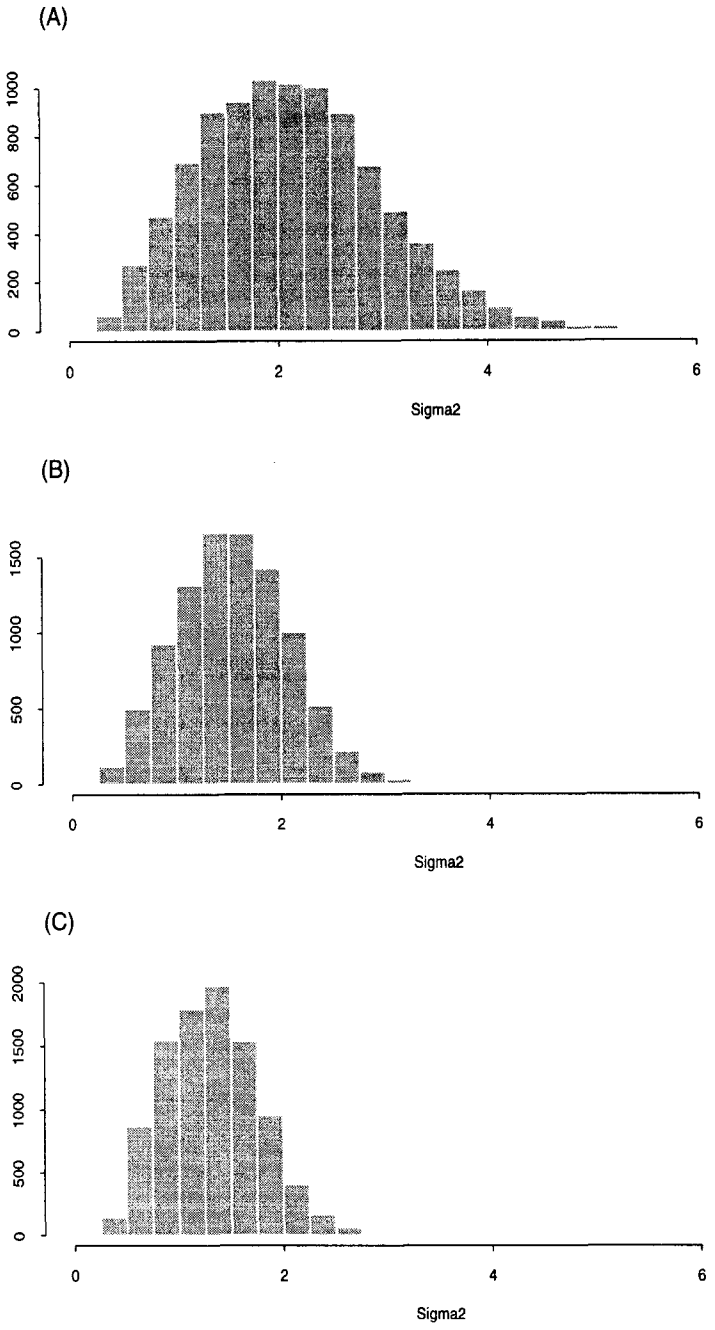

Figure 14 Metal B:Simulated marginal posterior. distributions for σ^2 (A) data augmentation for censored values (B) censored values replaced by $LOD/2$ (C) censored values replaced by LOD

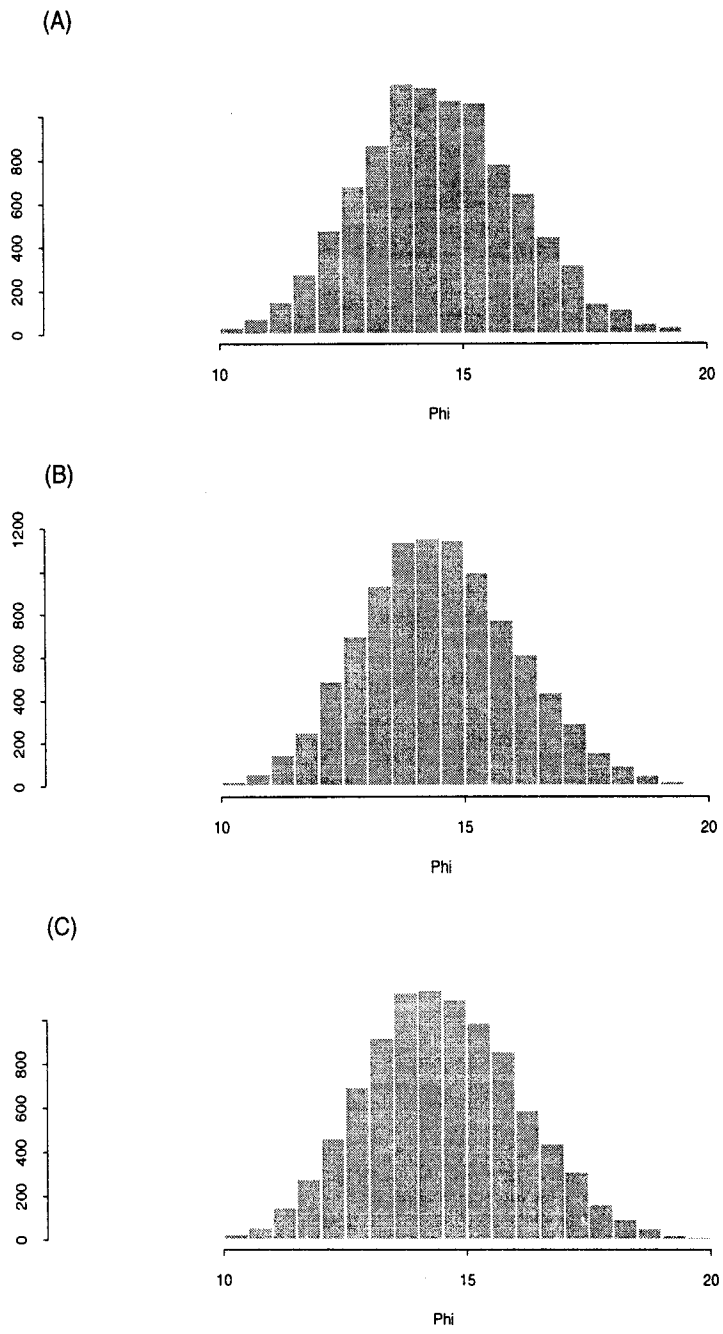

Figure 15 Metal B: Simulated marginal posterior distributions for ϕ (A) data augmentation for censored values (B) censored values replaced by $LOD/2$ (C) censored values replaced by LOD

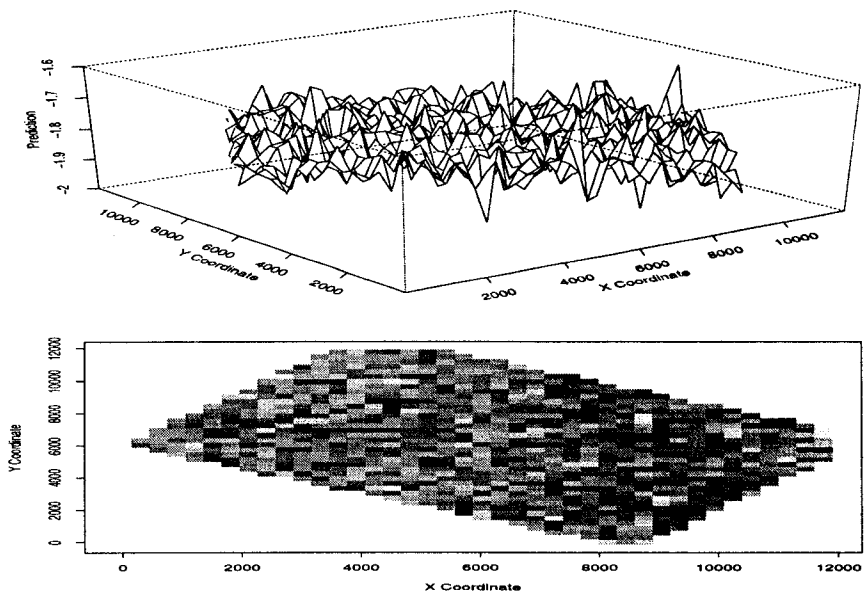

Figure 16 Metal B: Posterior median of the Bayesian predictive distribution using data augmentation for censored values

Along with constructing time-series plots for checking convergence, autocorrelations were computed for various lags. Plots of the autocorrelations for the DA analysis and the LOD/2 analysis are displayed in Figure 18 and 19. From these plots, we see that the DA method produces higher levels of autocorrelation as compared to the LOD/2 method. For example, independent iterates for the parameter μ occurs around a lag of 30 with data augmentation, while with the LOD/2 method independence of iterates occurs around a lag of 12. The data augmentation method produces lags that are roughly twice as large as the lags produces by the LOD/2 method. As stated on page 84 of *Analysis of Incomplete Multivariate Data* by Shafer (1997), "If the missing information is a large portion of the total information, the θ will depend heavily on Y_{mis} at each P-step, which will in turn depend on the value of θ used in the previous I-step; successive iterates of θ will tend to be highly correlated and convergence will be slow." In the case of censored spatial data, this is even more evident. Since the censored data is informative, as the

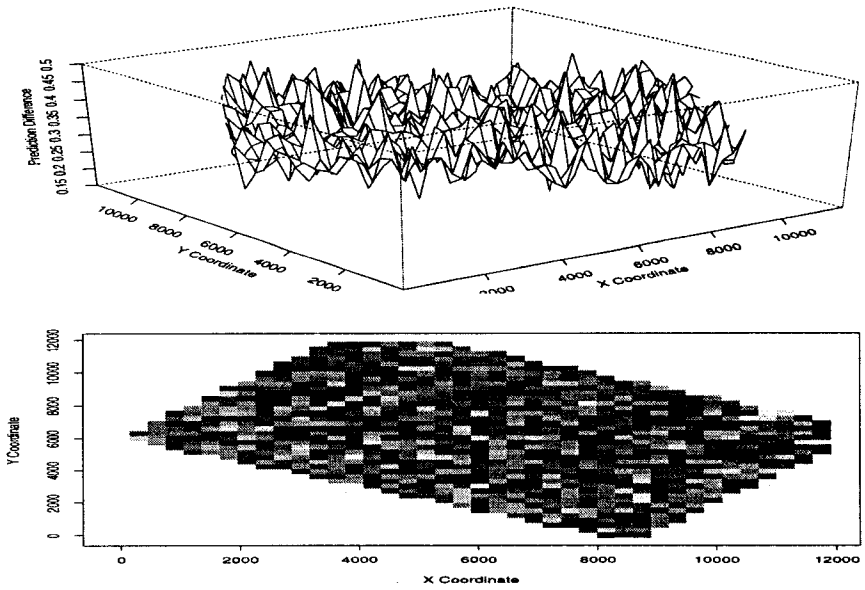

Figure 17 Metal B: Difference in posterior medians for DA and LOD/2 methods for handling censored values (LOD/2 - DA)

percent of censored observations increases, so does the amount of serial correlation when using the data augmentation procedure. Therefore, one may wish to use every k iterate when performing inferences, based on the amount of serial correlation.

Lastly, to investigate the effects of the prior distributions on estimation, sensitivity analysis was performed. Two more analyses were completed with prior distributions for the parameters being $\mu \sim \text{NOR}(0, 50)$, $\sigma^2 \sim \text{INGAM}(2.1, 3.3)$, $\phi \sim \text{GAM}(10, 0.1)$, $\tau^2 \sim \text{INGAM}(2.1, 2.2)$ and $\mu \sim \text{NOR}(-1, 100)$, $\sigma^2 \sim \text{INGAM}(2.1, 4.4)$, $\phi \sim \text{GAM}(5, 0.1)$, $\tau^2 \sim \text{INGAM}(2.1, 3.3)$. These two sets of prior specifications are very similar to the priors used in the first or primary analysis. The main difference is with regards to the means for the prior distributions for ϕ. Parameter estimates for the primary analysis and the two additional analyses are displayed in Table 6.

The results for μ, τ^2 and σ^2 are similar for the three sets of prior distributions. With regards to the parameter ϕ, the specified prior distribution seems to have an

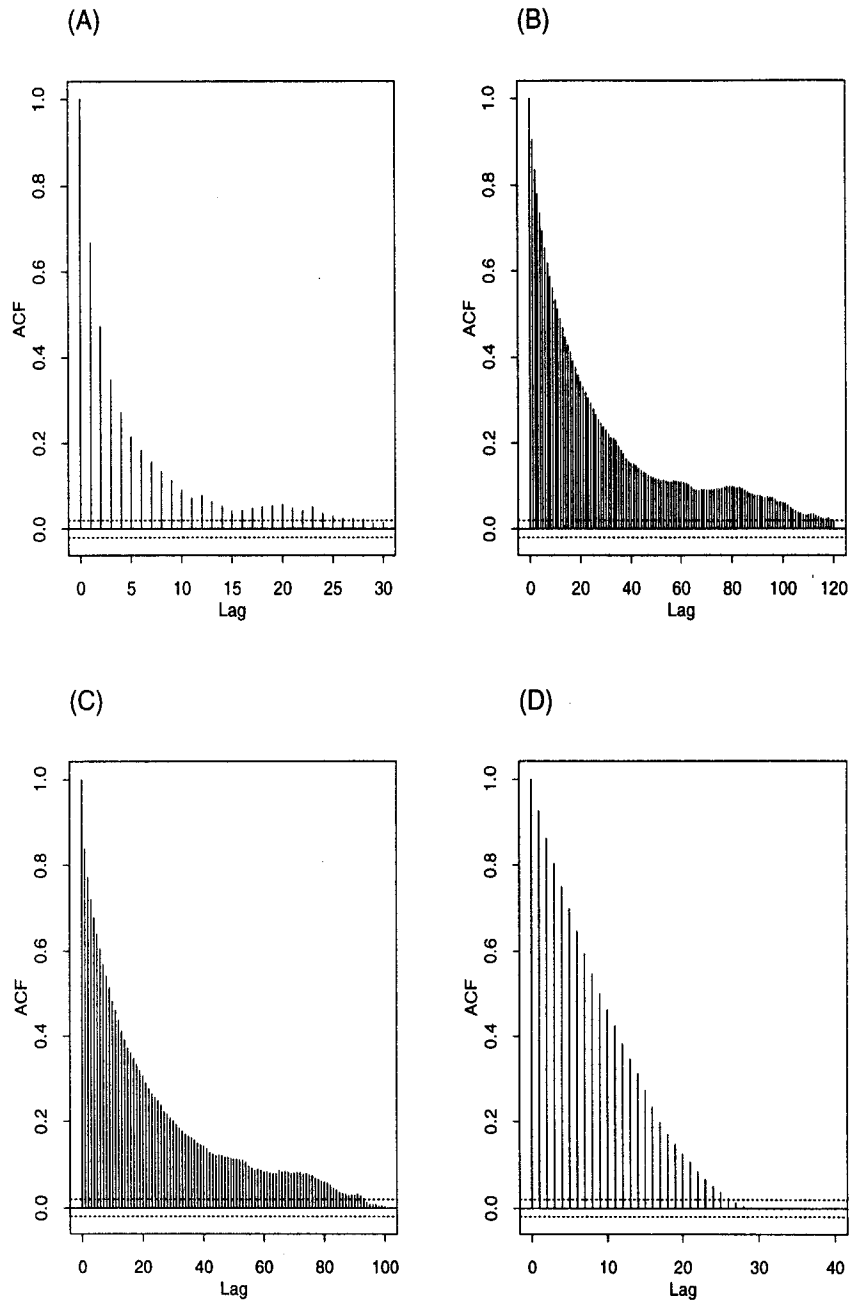

Figure 18 Metal B: Plot of autocorrelation function (ACF) for the parameters (A) μ (B) τ^2 (C) σ^2 and (D) ϕ using data augmentation

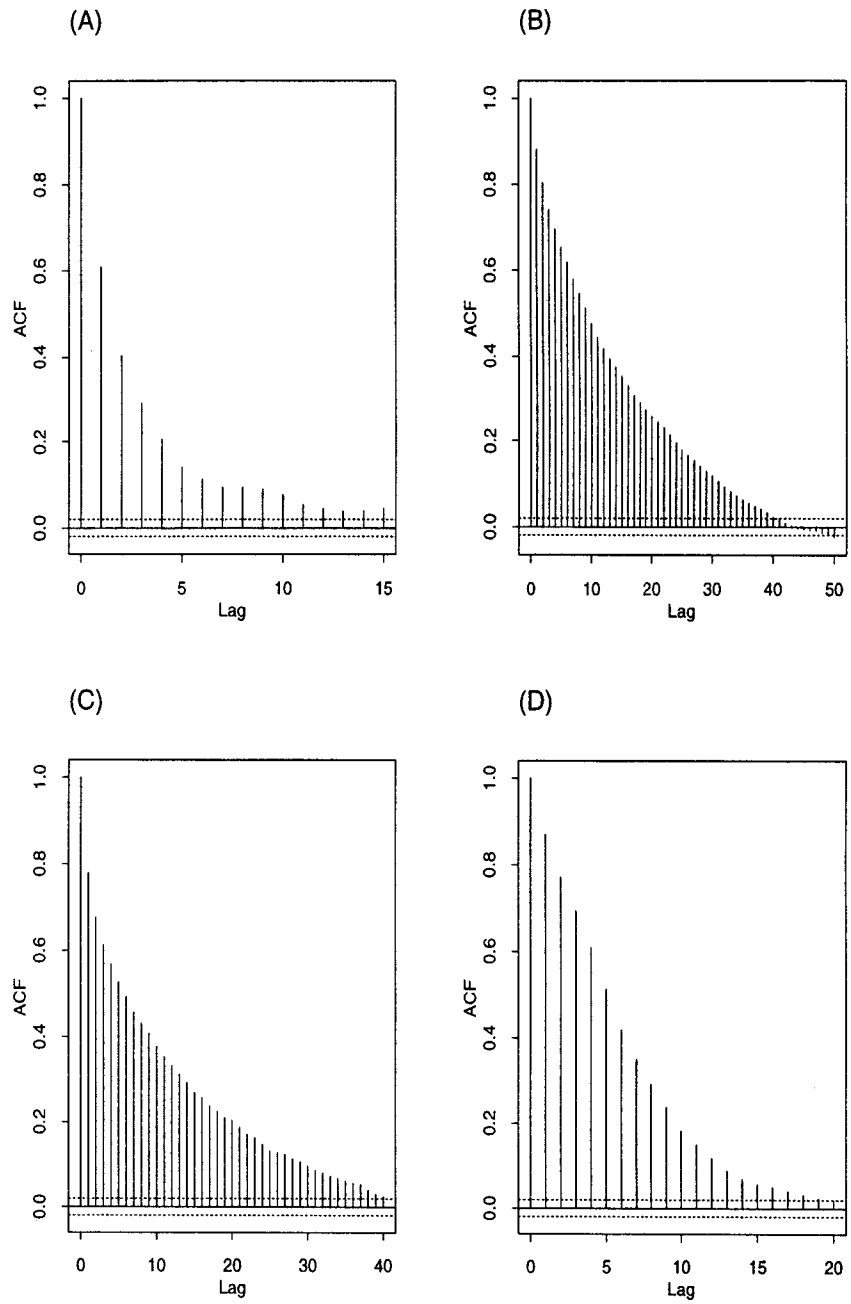

Figure 19 Metal B: Plot of autocorrelation function (ACF) for the parameters (A) μ (B) τ^2 (C) σ^2 and (D) ϕ using LOD/2 method

Table 6 Metal B: Median and 95% credible intervals based on the simulated marginal posterior distributions for three different prior specifications

	Primary Analysis		**Second Analysis**		**Third Analysis**	
	Median	Interval	Median	Interval	Median	Interval
μ	-1.807	(-2.263, -1.391)	-1.747	(-2.217, -1.328)	-1.871	(-2.348, -1.445)
τ^2	1.189	(0.329, 3.190)	1.418	(0.449, 3.314)	1.555	(0.560, 3.387)
σ^2	2.089	(0.676, 4.033)	1.906	(0.704, 3.757)	1.971	(0.796, 3.775)
ϕ	14.459	(11.43, 17.89)	5.916	(3.738, 8.810)	4.430	(2.321, 6.914)

effect on estimation. The primary analysis used a prior for ϕ that reflects a small amount of spatial dependence, with observations further than 1500 units apart being considered as independence observations. From this first analysis, we find the level of spatial dependence to be even smaller, with ϕ estimated to be 14.459. For all practical considerations for the site 15 study, an estimate of ϕ equal to 14 or 5 both result in little to no spatial dependence (i.e. range of around 45 or 15), leading to very similar predictions.

7 Discussion and conclusions

We have proposed a data augmentation method for the analysis of spatially correlated data in which some of the observation are censored. A Bayesian spatial or geostatistical model was used in which the spatial dependency was modeled using an exponential form. We also discussed the process of spatial prediction for unobserved locations using the augmented data and parameter estimates. The data augmentation procedure for censored spatial data was illustrated and compared to the LOD/2 and LOD methods using two environmental contamination sites; one involving dioxin and one involving a heavy metal.

The use of a model involving a spatial random effect allowed for imputation of the

censored observations to be completed using truncated univariate normal distributions. If not for the introduction of a spatial random effect to the model, the imputation step of the Gibbs sampler would have required the generation of the censored observations from a truncated multivariate normal distribution, $p(\boldsymbol{Y}_c|\boldsymbol{Y}_o, \boldsymbol{\Theta}, \boldsymbol{Y}_c \leq \boldsymbol{LOD})$, where $\boldsymbol{LOD}$ represents a vector containing the level of detections for the censored observations. One approach to generate values from a truncated multivariate normal distribution would be to implement the multivariate generation inside another Gibbs sampler, updating censored values one at a time. This method would be more computer intensive, requiring re-decomposition of the mean vector and the covariance matrix and subsequent calculation of the univariate conditional normal distribution for each censored observation at every iteration of the MCMC.

Likewise, the assumption of geometric anisotropy in the Missouri dataset lead to simplification of the analysis. Another option to handle anisotropy would be to model a trend in the X direction. A median polish procedure could also be performed and the resulting residuals used in the data augmentation procedure. But again, there is the question of how to deal with censored observations in a median polish procedure for the removal of a trend effect. For example, if all censored values were replaced with their level of detections, the median polish procedure would be removing the trend from the detection levels. For the Missouri data, by using a different distance measure, we were able to avoid the problems related to the median polish procedure when censored data are present. One thing to note, is that the method does not require isotropy. The procedure can be extended to cases involving directional dependence where simple techniques/solutions to handle directional dependence are not applicable.

The procedure could be easily extended to other Bayesian spatial models and other forms of censoring (e.g. right censoring, interval censoring). Instead of modeling the spatial dependence between observations with an exponential form, a spherical or Gaussian form could be applied. Since the data augmentation/imputation of the censored data is

based on the model, the results may be dependent upon this modeling choice. Future work is needed to investigate the robustness of the procedure to model misspecification and model diagnostics involving spatial data augmentation.

Convergence and serial correlation is another important issue with the analysis of censored spatial data using the data augmentation method. As seen with the site 15 example, the amount of serial correlation is larger when using the data augmentation method as opposed to a method that replaces the censored values with a constant, like $LOD/2$. As the percent of censoring increase, so does the amount of serial correlation and the number of iterations needed to reach convergence. Thus, in addition to model misspecification and diagnostics, work is needed to investigate the issues of convergence and serial correlation in cases involving moderate to large proportions of censored responses.

In addition to the extension of the method to different models, sensitivity analysis with respect to the prior distributions needs to be done. The data augmentation procedure for the analysis of censored spatial data can also be extended to a fully hierarchical Bayesian model using hyper-priors. Care must be taken when specifying prior distributions in the setting of spatial analysis to ensure proper joint distributions. As stated on page 81 of *Analysis of Incomplete Multivariate Data* by Shafer (1997), "Even when an improper prior is known to yield a proper posterior in the case of complete data, it may not necessarily do so when some data are missing."

In conclusion, this paper presents the use of data augmentation for the analysis of censored spatial data, which occurs often in environmental applications. Data augmentation produces more accurate parameter estimates as opposed to the common method of replacing the censored observations with half the level of detection. Along with producing biased parameter estimates, the common practice of replacing censored observations with a function of the level of detection under-estimates the variability in the approximated marginal densities. This under-estimation of the variability parameters

and variability in the marginal densities was also found when applying the data augmentation method in the context of a Bayesian conditionally specified Gaussian Model (Fridley and Dixon, 2003). Data augmentation can be easily applied to analyze censored spatial data, producing more accurate marginal posterior distributions and predictions. As seen in the two illustrative examples, the difference in predicted contamination levels between the ad hoc methods and the data augmentation can be extreme. These differences may lead to varying clean-up regions, which may have severe health, political and cost implications.

Appendix

This appendix presents the derivation of the full conditional distributions required for the Gibbs Sampler involving proper prior distributions.

Full conditional distribution for σ^2:

The full conditional distribution for σ^2 is

$$p(\sigma^2|\tau^2,\phi,\mu,\boldsymbol{W},\boldsymbol{X}) \propto p(\boldsymbol{W}|\sigma^2,\phi)p(\sigma^2)$$
$$\propto |\sigma^2 V^*(\phi)|^{-1/2}\exp\{\tfrac{-1}{2}\boldsymbol{W}^T(\sigma^2V^*(\phi))^{-1}\boldsymbol{W}\}(\sigma^2)^{-1(\alpha+1)}\exp\{-\beta/\sigma^2\},$$

where $V^*(\phi) = \exp\{-d/\phi\}$. Therefore, the full conditional distribution for σ^2 is

$$\sigma^2|\tau^2,\mu,\phi,\boldsymbol{W},\boldsymbol{X} \sim \text{INGAM}(n/2+\alpha,(1/2)\boldsymbol{W}^TV^*(\phi)^{-1}\boldsymbol{W}+\beta).$$

Full conditional distribution for τ^2:

The full conditional distribution for τ^2 is

$$p(\tau^2|\sigma^2,\mu,\phi,\boldsymbol{W},\boldsymbol{X}) \propto p(\boldsymbol{X}|\boldsymbol{W},\mu,\tau^2)p(\tau^2)$$
$$\propto \tfrac{1}{(\tau^2)^{n/2}(\tau^2)^{\gamma+1}}\exp\{\tfrac{-1}{2\tau^2}(\boldsymbol{X}-(\boldsymbol{\mu}+\boldsymbol{W}))^T(\boldsymbol{X}-(\boldsymbol{\mu}+\boldsymbol{W}))-\tfrac{\delta}{\tau^2}\}.$$

Therefore, the full conditional distribution for τ^2 is

$$\tau^2|\sigma^2,\mu,\phi,\boldsymbol{W},\boldsymbol{X} \sim \text{INGAM}(n/2+\gamma,(1/2)(\boldsymbol{X}-(\boldsymbol{\mu}+\boldsymbol{W}))^T(\boldsymbol{X}-(\boldsymbol{\mu}+\boldsymbol{W}))+\delta).$$

Full conditional distribution for μ:

The full conditional distribution for $\boldsymbol{\mu}$ is

$$p(\boldsymbol{\mu}|\tau^2,\sigma^2,\phi,\boldsymbol{W},\boldsymbol{X}) \propto p(\boldsymbol{X}|\boldsymbol{W},\boldsymbol{\mu},\tau^2)p(\boldsymbol{\mu}).$$

We will first find the full conditional distribution for $\boldsymbol{\mu}$ and then the full conditional distribution for μ. Thus,

$$p(\boldsymbol{\mu}|\tau^2,\sigma^2,\phi,\boldsymbol{W},\boldsymbol{X})$$
$$\propto \exp\{\tfrac{-1}{2}((\boldsymbol{X}-\boldsymbol{W})-\boldsymbol{\mu}))^T(\tau^2 I)^{-1}((\boldsymbol{X}-\boldsymbol{W})-\boldsymbol{\mu}))+\tfrac{-1}{2}(\boldsymbol{\mu}-\boldsymbol{\lambda})^T(\psi^2 I)^{-1}(\boldsymbol{\mu}-\boldsymbol{\lambda})\}.$$

By completing the square, we have

$$\boldsymbol{\mu}|\tau^2,\sigma^2,\phi,\boldsymbol{W},\boldsymbol{X} \sim \text{MVN}(\boldsymbol{\mu_o},\Sigma_o),$$

where $\boldsymbol{\mu_o} = (\frac{\psi^2\tau^2}{\tau^2+\psi^2})(\frac{1}{\psi^2}\boldsymbol{\lambda}+\frac{1}{\tau^2}(\boldsymbol{X}-\boldsymbol{W}))$ and $\Sigma_o = \frac{\psi^2\tau^2}{\tau^2+\psi^2}I$. Since $(\mathbf{1}^T/n)\boldsymbol{\mu} = \mu$, the full conditional distribution for μ is

$$\mu|\sigma^2,\tau^2,\phi,\boldsymbol{W},\boldsymbol{X} \sim \text{NOR}(\mu_1,\sigma_1^2),$$

where $\mu_1 = (\frac{\psi^2\tau^2}{\tau^2+\psi^2})[\frac{1}{\psi^2}\lambda+\frac{1}{\tau^2}(\bar{X}-\bar{W})]$ and $\sigma_1^2 = (\frac{1}{n})(\frac{\psi^2\tau^2}{\tau^2+\psi^2})$.

Full conditional distribution of $\boldsymbol{W}$:

The full conditional distribution for the spatial random effects, $\boldsymbol{W}$, is

$$p(\boldsymbol{W}|\boldsymbol{X},\mu,\tau^2,\sigma^2,\phi) \propto p(\boldsymbol{X}|\boldsymbol{W},\mu,\tau^2)p(\boldsymbol{W}|\sigma^2,\phi)$$
$$\propto \exp\{\tfrac{-1}{2}(\boldsymbol{X}-(\boldsymbol{\mu}+\boldsymbol{W}))^T(\tau^2 I)^{-1}(\boldsymbol{X}-(\boldsymbol{\mu}+\boldsymbol{W}))\} \times \exp\{\tfrac{-1}{2}\boldsymbol{W}^T V(\sigma^2,\phi)^{-1}\boldsymbol{W}\}$$
$$= \exp\{\tfrac{-1}{2}((\boldsymbol{X}-\boldsymbol{\mu})-\boldsymbol{W})^T(\tau^2 I)^{-1}((\boldsymbol{X}-\boldsymbol{\mu})-\boldsymbol{W})+\tfrac{-1}{2}\boldsymbol{W}^T V(\sigma^2,\phi)^{-1}\boldsymbol{W}\}.$$

By completing the square, we have the full conditional distribution for $\boldsymbol{W}$ to be

$$\boldsymbol{W}|\boldsymbol{X}, \mu, \tau^2, \sigma^2, \phi \sim \text{MVN}(\boldsymbol{\mu}_w, \Sigma_w),$$

where $\boldsymbol{\mu}_w = [V^{-1}(\sigma^2, \phi) + \frac{1}{\tau^2} I]^{-1}[\frac{1}{\tau^2}(\boldsymbol{X} - \boldsymbol{\mu})]$ and $\Sigma_w = [V^{-1}(\sigma^2, \phi) + \frac{1}{\tau^2} I]^{-1}$.

Full conditional distribution of ϕ:

The full conditional distribution for ϕ is

$$p(\phi|\mu, \tau^2, \sigma^2, \phi, \boldsymbol{W}, \boldsymbol{X}) \propto p(\boldsymbol{W}|\sigma^2, \phi)p(\phi)$$
$$\propto \frac{\phi^{\eta-1}}{|V^*(\phi)|^{1/2}} \exp\{\frac{-1}{2\sigma^2} \boldsymbol{W}^T V^*(\phi)^{-1} \boldsymbol{W} - \theta\phi\}.$$

Hence, there is no closed form (i.e. known distribution) for the full conditional for ϕ.

The full conditional distribution for ϕ is only known up to a proportional constant.

References

Ancona, M.A., Tawn, J.A. (2002). *Spatial Statistics Through Applications*, Chapter 8: Modelling extreme rainfall events. pg 177-201. Editors: J.Mateu and F. Montes. WIT Press, Boston.

Berger, J.O., de Oliveira, V., and Sanso, B. (2001). Objective Bayesian Analysis of Spatially Correlated Data. *Journal of the American Statistical Association*, **96**, 1361-1374.

Carlin, B.P. and Louis, T.A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman & Hall, London.

Cressie, N.A.C. (1993). *Statistics for Spatial Data, Revised Edition.* John Wiley & Sons, Inc., New York.

Dempster, A.P., Laird, N.M, and Rubin, D.B. (1977). Maximum likelihood estimation form incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B*, **39**, 1-38.

de Oliveira, V., and Ecker, M.D. (2002). Bayesian hot spot detection in the presence of spatial trend: application to total nitrogen concentration in Chesapeake Bay. *Environmetrics*, **13**, 85-101.

Ecker, M.D., and Gelfand, A.E. (1997). Bayesian Variogram Modeling for an Isotropic Spatial Process. *Journal of Agricultural, Biological, and Environmental Statistics*, **2**, 347-369.

Evans, M., and Swartz, T. (2000). *Approximating Integrals via Monte Carlo and Deterministic Methods.* Oxford University Press, Oxford.

Fridley, B.L. and Dixon, P. (2003). Data Augmentation for a Conditionally Specified Gaussian Spatial Model involving Censored Observations. In preparation.

Gaudard, M., Karson, M., Linder, E., and Sinha, D. (1999). Bayesian spatial prediction. *Environmental and Ecological Statistics*, **6**, 147-171.

Gelfand, A.E., and Smith, A.F.M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, **85**, 398-409.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). *Bayesian Data Analysis.* Chapman & Hall, London.

Geman, S., and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Transactions on Patter Analysis and Machine Intelligence*, **6**, 721-741.

Gibbons, R. (1995). Some Statistical and Conceptual Issues in the Detection of Low-Level Environmental Pollutants. *Environmental & Ecological Statistics*, **2**, 125-167.

Gilks, W.R., Richardson, S., and Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice.* Chapman & Hall, London.

Hastings, W.K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, **57**, 97-109.

Helsel, D.R. (1990). Less than Obvious. Statistical Treatment of Data Below the Detection Limit. *Environmental Science Technology*, **24**, 1766-1774.

Hopke, P.K., Liu, C., and Rubin, D.B. (2001). Multiple Imputation for Multivariate Data with Missing and Below-Threshold Measurements: Time-Series Concentrations of Pollutants in the Arctic. *Biometrics*, **57**, 22-33.

Hughes, J.P. (1999). Mixed Effects Models with censored Data with Application to HIV RNA Levels. *Biometrics*, **55**, 625-629.

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

Li, K.H. (1988). Imputations Using Markov Chains. *Journal of Statistical Computation and Simulation*, **30**, 57-79.

Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data,* 2^{nd} *Ed.*. Wiley, New York.

Matheron, G. (1963). Principles of Geostatistics. *Economic Geology*, **58**, 1246-1266.

Pettitt, A.N. (1986). Censored Observations, Repeated Measures and Mixed Effects Models: An Approach Using the EM Algorithm and Normal Errors. *Biometrika*, **73**, 635-643.

Porter, P.S., Ward, R.C., Bell, H.F. (1988). The Detection Limit. Water Quality Monitoring Data Are Plagued with Levels of Chemicals That Are Too Low to Be Measured Precisely. *Environmental Science Technology*, **22**, 856-861.

Robert, C.P. (1995). Simulation of truncated normal variables. *Statistics and Computing*, **5**, 121-125.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data.* Chapman & Hall, London.

Smith, F.B., Helms, R.W. (1995). EM Mixed Model Analysis of Data From Informatively Censored Normal Distributions. *Biometrics*, **51**, 425-436.

Tanner, M.A., and Wong, W.H. (1987). The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, **82**, 528-540.

Zirschky, J.H., and Harris, D.J. (1986). Geostatistical Analysis of Hazardous Waste Site Data. *Journal of Environmental Engineering*, **112**, 770-784.

DATA AUGMENTATION FOR A CONDITIONALLY SPECIFIED GAUSSIAN SPATIAL MODEL INVOLVING CENSORED OBSERVATIONS

A paper to be submitted to the Journal of the American Statistical Association: Case Studies and Applications

Brooke Fridley and Philip Dixon

Abstract

Censored data occur in numerous areas of application. When independence of observation can be assumed, various methods have been proposed to analyze censored data. When one adds the complexity of spatial dependency between observations, methods for handling censored observations are not as clear. There are various statistical methods that allow for the analysis of spatial data, but none of these standard methods deal with the occurrence of censored data. In many spatial analyses involving pollutants or contamination, censoring often occurs from contamination values falling below a level of detection (LOD). That is, often censored spatial data are left censored. A common practice is to set the censored observations equal to the LOD or some function of the level of detection, like $LOD/2$. This single imputation method results in biased parameter estimates. This paper will present and illustrate a data augmentation approach for the analysis of spatially correlated data using a Bayesian conditionally specified Gaussian model, in which some of the observations are left censored. Comparison of the data

augmentation method to the methods of replacing the censored observations with LOD and $LOD/2$ are also illustrated using a study looking at metal contamination in an old industrial area and a dioxin contaminated area in Missouri.

1 Introduction

Analysis of data involving spatial dependence arise in many applications, some of which include environmental sciences and epidemiology. In the case of lattice data, modeling can be done using Markov random fields. That is, one models the distribution of a random variable conditional on all other variables. Using the Markovian property, the conditional distribution of a random variable only depends on its neighbors. Recently, much research has been done using Markov random fields or conditional autoregressive models within a Bayesian framework. In particular, much of the work has been in regards to disease mapping applications (Stern and Cressie, 1999; Xia and Carlin, 1998; Bell and Broemeling, 2000; Gumpertz, Graham, and Ristaino, 1997). In addition to disease mapping, Daniels, Lee and Kaiser (2001) fit a hierarchical model for the analysis of particulate matter in Pittsburgh, where the random spatial effect is model with a conditionally specified Gaussian model.

In specifying a Bayesian conditionally specified model, one must specify the prior distributions along with the neighborhood structure for modeling the spatial dependence. The definition of the neighbors of a given location or neighborhood system varies. For example, one neighborhood structure could depend on distance from the given location, while another neighborhood structure would only include the nearest neighbors in the neighborhood structure (Besag, 1974; Kaiser and Cressie, 2000; Cressie, 1993). In addition to specifying the neighborhood structure and priors, one must also specify the distribution for the random variables. A common distributional assumption is that of Gaussian. Gelman and Meng (1991) discuss the use of Gaussian conditional distribu-

tions to model multivariate data. Exponential families are discussed in detail by Besag (1974). Kaiser, Cressie and Lee (2002) extend the use of exponential family conditional distributions to spatial mixture models. The use of Markov chain Monte Carlo and the Bayesian paradigm has also been discussed by Besag and Green (1993) and Besag, Green, Higdon and Mengersen (1995).

In environmental applications, it is not uncommon for some observations to fall below some detection level (LOD), resulting in left censored observations. A common technique to handle censored observations in the spatial setting is to replace the censored observation with some function of the level of detection. This method of replacing all censored values with a constant results in biased parameter estimates. Another common approach is to treat the data as independent observations and then use methods that can be applied in the case of independence (Helsel, 1990; Gibbons, 1995; Porter, Ward and Bell, 1988). The drawback with this approach is that one ignores the spatial information available.

Much research has been done and is currently being done in the area of missing data, where censored observations represent a form of non-ignorable missing data. Little and Rubin (2002) outline the basics for analyzing data involving missing observations. The EM algorithm (Dempster, Laird and Rubin, 1977) is a procedure that can be used to analyze missing data. The EM algorithm has been used extensively for the handling of missing or censored data involving mixed model (Hughes, 1999; Smith and Helms, 1995; Pettitt, 1986). The idea of data augmentation was first presented by Tanner and Wong (1987) in which analysis and augmentation of missing data is done within a Markov chain Monte Carlo. Hopke, Liu and Rubin (2001) use a data augmentation procedure to produce k sets of complete data, which can then be analyzed by traditional statistical methods. Hybrids using both EM and MCMC ideas have also been used to handle missing data. Shafer (1997) further outline the use of the EM algorithm and data augmentation to handle missing data. Implementation of data augmentation for

the handling of censored data in a spatial setting has yet to be addressed. In this paper, we combine the ideas of spatial analysis using a conditionally specified Gaussian model and data augmentation as a means for analyzing censored spatial data.

2 Censored data and data augmentation

Censored data is a form of missing data that if not accounted for will result in biased parameter estimates. In the terminology of Little and Rubin (2002), censored data is a type of "non-ignorable" missing data. One possible solution to the problem is to integrate out the censored data from the posterior distribution, $\int p(\boldsymbol{\Theta}|\boldsymbol{Y}_c, \boldsymbol{Y}_o)p(\boldsymbol{Y}_c|\boldsymbol{Y}_o)d\boldsymbol{y}_c$. A problem with this solution is that the integration may be difficult or intractable. To solve this problem, one may employ the idea of data augmentation within a Markov chain Monte Carlo. Data augmentation is a method that solves the problem of having to integrate out the censored observations from $p(\boldsymbol{\Theta}|\boldsymbol{Y}_c, \boldsymbol{Y}_o)$.

Let y represent the observed data, z represent the augmentated data (censored data), and θ represent the parameters. The posterior distribution $p(\theta|y, z)$ is easy to compute if both y and z are observed, whereas, $p(\theta|y) = \int p(\theta|y, z)p(z|y)dz$ may be cumbersome to calculate if z is not observed (Tanner and Wong, 1987). Hence, multiple realizations of z are generated from the predictive distribution $p(z|y)$. The generation of z from $p(z|y)$ can be decomposed into the following two steps: (1) a value of θ is generated, say ϕ, and (2) based on ϕ, generate z from $p(z|\phi, y)$. Averaging $p(\theta|y, z)$ over the simulated values of z results in an approximation for $p(\theta|y)$.

Data augmentation can be thought of as using Markov chain Monte Carlo to perform imputation. Data augmentation results in "augmenting" or imputing values for the censored observations at each iteration of the chain, followed by a posterior step that generates values of the parameters conditional on the augmented data. The idea is the following. Given the current value of the parameters $\boldsymbol{\Theta}^{(t)}$, draw a vector $\boldsymbol{Y}_c^{(t+1)}$

for the censored data from $p(\boldsymbol{Y}_c|\boldsymbol{Y}_o, \boldsymbol{\Theta}^{(t)})$. Then based on $\boldsymbol{Y}_c^{(t+1)}$, draw $\boldsymbol{\Theta}^{(t+1)}$ from the complete data posterior $p(\boldsymbol{\Theta}|\boldsymbol{Y}_o, \boldsymbol{Y}_c^{(t+1)})$. Repeating this process numerous times yields a stochastic sequence $\{\boldsymbol{\Theta}^{(t)}, \boldsymbol{Y}_c^{(t)} : t = 1, 2, ...\}$ whose stationary distribution is $p(\boldsymbol{\Theta}, \boldsymbol{Y}_c|\boldsymbol{Y}_o)$. In addition, $\{\boldsymbol{\Theta}^{(t)} : t = 1, 2, ...\}$ has stationary distribution $p(\boldsymbol{\Theta}|\boldsymbol{Y}_o)$. Hence, the sequence $\{\boldsymbol{\Theta}^{(t)} : t = 1, 2, ...\}$ can be used to estimation of the joint posterior distribution $p(\boldsymbol{\Theta}|\boldsymbol{Y}_o)$ (Shafer, 1997; Geman and Geman, 1984; Gilks, Richardson and Spiegelhalter, 1996).

3 Bayesian conditionally specified Gaussian spatial model

Fitting a conditionally specified Gaussian or conditional autoregressive model, let $\{Y(s_i) : i = 1, ..., n\}$ represent a set of random variables at locations $\{s_i : i = 1, ..., n\}$. Then, $Y(s_i)$ is model as

$$Y(s_i)|Y(N_i) \sim \text{NOR}(\mu_i, \tau^2),$$

where $Y(N_i)$ represent all observations that are neighbors to s_i. In addition, the parameterization of μ_i takes the form

$$\mu_i = \alpha_i + \sum_{j=1}^{n} c_{ij}(y(s_j) - \alpha_j).$$

We then define $\alpha_i = \alpha$, and $c_{ij} = \eta(d_{ij})^{-1}$ if $s_j \in N_i$. The joint distribution of $\boldsymbol{Y} = (Y(s_1), ..., Y(s_n))$ is then given by

$$\boldsymbol{Y} \sim GAU(\boldsymbol{\alpha}, (I - C)^{-1}M), \tag{1}$$

where C contains the elements c_{ij} which involve η, and M is a diagonal matrix containing τ^2 (Besag, 1974; Kaiser and Cressie, 2000). In this conditionally specified Gaussian model, we are modeling the inverse covariance matrix as opposed to the covariance matrix as in geostatistical models. In addition to modeling the inverse covariance matrix, α represents the large scale model and c_{ij} models the spatial dependence or small scale

model. If covariates are available, α_i can be modeled as $\alpha_i = X_i^T\beta$. Just as there are ways to model α_i with covariate information, there are other forms for the parameterization of C. For the remainder of this paper, we will not focus on the modeling aspect of the analysis, but instead on the data augmentation procedure for the handling of censored spatial data.

For the model specified in equation (1), $C = \eta H$, where H is a known symmetric matrix containing inverse distances. The covariance matrix $(I-C)^{-1}M$ is a non-negative definite matrix. This does not guarantee that H is non-negative or positive definite, since the eigenvalues of H can be positive or negative. If $h_1, h_2, ..., h_n$ represent the ordered eigenvalues of H, then $|(I - C)| = \prod_{i=1}^{n}(1 - \eta h_i)$, where η must be such that each term is positive (Kaiser and Cressie, 2000; Cressie, 1993). Thus, we have the following bounds for η.

- If $0 \leq h_1$, then $\hat{\eta} < 1/h_n$
- If $h_n \leq 0$, then $\hat{\eta} > 1/h_1$
- If $h_1 < 0 < h_n$, then $1/h_1 < \hat{\eta} < 1/h_n$ (most common case)

To finish the specification of the model, prior distributions are placed on all parameters in the model. A possible prior specification involving non-informative, improper priors would be

$$p(\alpha) \propto 1,$$
$$p(\tau^2) \propto (\tau^2)^{-1},$$
$$p(\eta) \propto 1 \text{ over the possible range of } \eta.$$

Another option would be to put proper prior distributions on all parameters. The following is a possible prior specifications involving proper priors and the priors used for the remainder of the paper:

$$\alpha \sim \text{NOR}(\mu, \sigma^2),$$

$$\tau^2 \sim \text{INGAM}(\gamma, \beta),$$

$$\eta \sim \text{Transformed BETA}(\psi, \phi),$$

where $\eta = y(\frac{h_1 - h_n}{h_1 h_n}) + \frac{1}{h_1}$, and $y \sim \text{BETA}(\psi, \phi)$. To ensure $(I - C(\eta))^{-1}$ is positive definite, we exclude $\frac{1}{h_1}$ and $\frac{1}{h_n}$ from the support set of the transformed beta distribution, of which both have measure 0. If both ψ and ϕ are set equal to 1, the transformed beta prior reduces to a uniform prior over the range $(\frac{1}{h_1}, \frac{1}{h_n})$. To place either a informative or non-informative prior distribution on η, we need to compute the largest and smallest eigenvalues of H. Based on the eigenvalues, the prior distribution will have support $(-\infty, 1/h_n)$, $(1/h_1, \infty)$, or $(1/h_1, 1/h_n)$.

4 Markov chain Monte Carlo for data augmentation

The handling of censored spatial data using a data augmentation procedure is done within a Markov chain Monte Carlo (MCMC). For the augmentation procedure, the Gibbs sampler will be utilized with an additional augmentation or imputation step. The Gibbs sampler is a special case of the data augmentation procedure presented by Tanner and Wong (1987), where only one augmentated dataset is generated at each iteration of the chain.

With the assumptions of the Gibbs sampler satisfied for the Bayesian conditionally specified Gaussian model involving censored data, the data augmentation procedure can be completed as follows. At each iteration of the Gibbs sampler, the censored data will be imputed by generating values from $p(\boldsymbol{Y}_c | \boldsymbol{Y}_o, \alpha, \tau^2, \eta)$. Using the augmented-complete dataset, the parameters α, τ^2, and η will be generated from their corresponding full conditional distributions. Repeating this process numerous times, yields a stochastic process with stationary distribution $p(\boldsymbol{\Theta}, \boldsymbol{Y}_c | \boldsymbol{Y}_o)$, where $\boldsymbol{\Theta} = (\alpha, \tau^2, \eta)$ (Geman and Geman, 1984). Derivation of the full conditional distributions required for the Gibbs

Sampler are located in the appendix. The MCMC data augmentation algorithm within the framework of a Bayesian conditionally specified Gaussian model involving censored observations is as follows.

1. Set starting values for $\alpha^{(0)}$, $\tau^{2(0)}$, and $\eta^{(0)}$. Set $m = 0$.

2. Set censored values equal to their level of detection, $\boldsymbol{Y}_c^{(0)} = \boldsymbol{LOD}$, where $\boldsymbol{Y}_c$ represent the vector of censored observations.

3. Let $\boldsymbol{Y}^{T(m)} = (\boldsymbol{Y}_c^{(m)}, \boldsymbol{Y}_o)^T$, were $\boldsymbol{Y}_o$ represent the observed values.

4. Generate $\alpha^{(m+1)}$ from $N(\mu_\alpha^{(m+1)}, \sigma_\alpha^{2(m+1)})$ with

   $\mu_\alpha^{(m+1)} = \frac{1}{n}\mathbf{1}^T(\frac{1}{\sigma^2}I + \frac{1}{\tau^{2(m)}}(I - C))^{-1}(\frac{\mu}{\sigma^2}\mathbf{1} + \frac{1}{\tau^{2(m)}}(I - C)\boldsymbol{Y}^{(m)})$ and

   $\sigma_\alpha^{2(m+1)} = \frac{1}{n^2}\mathbf{1}^T(\frac{1}{\sigma^2}I + \frac{1}{\tau^{2(m)}}(I - C))^{-1}\mathbf{1}$.

5. Generate $\tau^{2(m+1)}$ from $INGAM(\frac{n}{2} + \gamma, \frac{1}{2}(\boldsymbol{Y}^{(m)} - \boldsymbol{\alpha}^{(m+1)})^T(I - C)$

   $(\boldsymbol{Y}^{(m)} - \boldsymbol{\alpha}^{(m+1)}) + \beta)$.

6. Using Metropolis-Hastings step(s), simulate $\eta^{(m+1)}$ from

   $p(\eta|\boldsymbol{Y}^{(m)}, \tau^{2(m+1)}, \alpha^{(m+1)}) \propto [\prod_{i=1}^{n}(1 - \eta h_i)]^{1/2} \exp\{\frac{\eta}{2\tau^{2(m+1)}}(\boldsymbol{Y}^{(m)} - \boldsymbol{\alpha}^{(m+1)})^T$

   $H(\boldsymbol{Y}^{(m)} - \boldsymbol{\alpha}^{(m+1)})\}\ (\eta - \frac{1}{h_1})^{\psi-1}[1 - (\eta - \frac{1}{h_1})(\frac{h_n h_1}{h_1 - h_n})]^{\phi-1}$.

7. Now have $\boldsymbol{\Theta}^{(m+1)} = (\alpha^{(m+1)}, \tau^{2(m+1)}, \eta^{(m+1)})$.

8. Using $\boldsymbol{\Theta}^{(m+1)}$, impute values for the censored values $\boldsymbol{Y}_c$ and produce $\boldsymbol{Y}_c^{(m+1)}$. Let $\boldsymbol{Y}_c = (Y_{1c}, Y_{2c}, .., Y_{kc})$ represent k censored observations. Let $\mu_i = \alpha + \sum_{j=1}^{n} c_{ij}(Y(s_j) - \alpha)$, and $c_{ij} = \eta(d_{ij})^{-h}$ for $s_j \in N_i$.

   (a) Generate $Y_{1c}^{(m+1)}$ from $Y_{1c}|Y_{2c}^{(m)}, .., Y_{kc}^{(m)}, \boldsymbol{Y}_o, \Theta^{(m+1)}$ which is a univariate normal distribution $\mathrm{N}(\mu_1^{(m+1)}, \tau^{2(m+1)})$, truncated at LOD_1.

(b) Generate $Y_{2c}^{(m+1)}$ from $Y_{2c}|Y_{1c}^{(m+1)}, Y_{3c}^{(m)}, .., Y_{kc}^{(m)}, \boldsymbol{Y}_o, \Theta^{(m+1)}$ which is a univariate normal distribution $\mathrm{N}(\mu_2^{(m+1)}, \tau^{2(m+1)})$, truncated at LOD_2.

. .

(c) Generate $Y_{kc}^{(m+1)}$ from $Y_{kc}|Y_{1c}^{(m+1)}, .., Y_{(k-1)c}^{(m+1)}, \boldsymbol{Y}_o, \Theta^{(m+1)}$ which is a conditional univariate normal distribution $\mathrm{N}(\mu_k^{(m+1)}, \tau^{2(m+1)})$, truncated at LOD_k.

9. Set $m = m + 1$ and repeat the algorithm a large number of times.

The reason behind using a Metropolis-Hastings algorithm for the generation of η instead of a rejection algorithm is due to fact that a bound M for the function is required for a rejection algorithm. By using a Metropolis-Hastings algorithm, we were not required to find the bound M, only to specify a candidate generating distribution. If the chain converges slowly or does not mix well with respect to η, one may wish to use a different candidate generating distribution for the Metropolis-Hastings step(s).

5 Illustrative example: site 15

5.1 Description of data

Site 15 is an old abandoned industrial site that was later converted into a park. A study was conducted to look at the level of metal contamination at the site. The purpose of original study was to determine if the soil contained excessive amounts of Metal C and if clean-up was required. In addition, the study was designed to investigate possible association of Metal C with other metals found in the soil (e.g. Metal A and Metal B).

At each location sampled, a soil core was drilled. Measurements were taken at different depths based on soil characteristics. No information was available on the type of soil, only the depth. Censored observations occurred for various metals. For a given metal the detection limits varied, partly due to the amount of sample used in the analysis or the amount of moisture in a sample.

After initial investigation, the assumption of isotropy was adequate and a log transformation was required for the normality assumption. To illustrating the data augmentation method for censored spatial data, only the second depth measurements were analyzed (i.e. observations right below topsoil). The data augmentation procedure can be extended to the 3-dimension setting. For site 15, it was not clear how to handle the depth dimension, due to the fact that no information was available on the type of soil. The only information available was the depth of the samples taken, which differed from location to location.

For illustration purposes, we will only investigate two metals, Metal A and Metal B. Metal A was recorded in units mg/kg with a EPA clean-up criteria of 340 mg/kg for non-residential and 14 mg/kg for residential areas. Of the 82 observations, 52 (63%) were censored with varying levels of detection, the largest LOD being 35 mg/kg. Metal B was also recorded in mg/kg, but with a clean-up criteria of 1 mg/kg for both residential and non-residential areas. Of the 82 observations, 32 (39%) were censored with the largest LOD being 1.5 mg/kg. For both metals, their highest LOD is greater then the clean-up criteria for residential areas with moderate to large proportions of the data censored. Sampled locations for Metal A and Metal B are displayed in Figures 1 and 2. Metal A and Metal B will be used to illustrate the application of data augmentation for the handling of spatial censored data in the context of a Bayesian conditionally specified Gaussian model.

5.2 Model specification and results

For the analysis of both Metal A and Metal B, a Bayesian conditionally specified Gaussian model given in Section 3 was used with priors specifications of $\alpha \sim \text{NOR}(0, 50)$, $\tau^2 \sim \text{INGAM}(2.1, 2.2)$, and $\eta \sim$ transformed $\text{BETA}(1, 1)$, excluding $1/h_1$ and $1/h_n$ from the support set for η. This specification resulted in very diffuse priors, with finite variance, for all parameters. By using a transformed beta distribution as the prior for η,

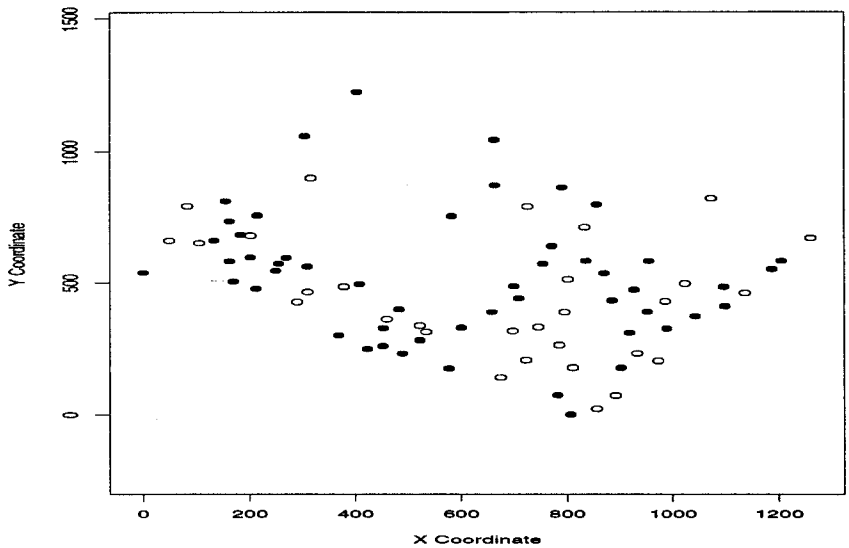

Figure 1 Sampled locations of Metal A, o represent observed values and • represent censored values

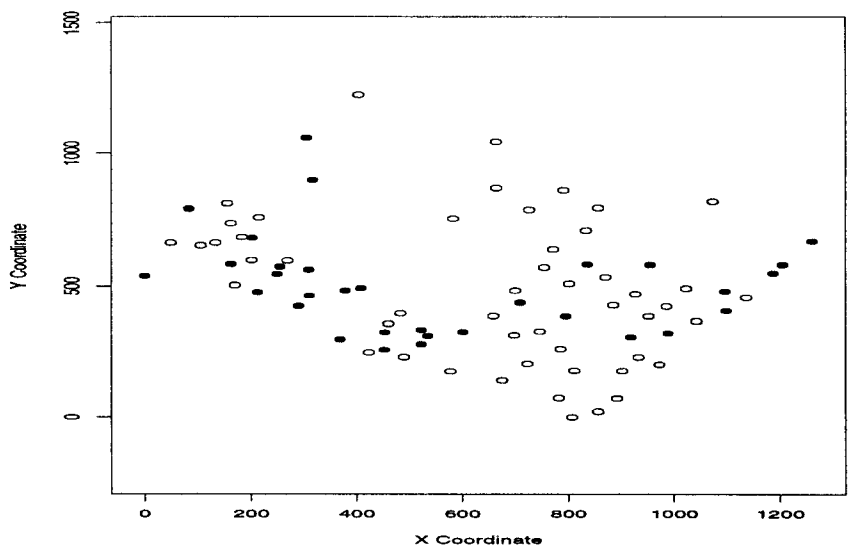

Figure 2 Sampled locations of Metal B, o represent observed values and • represent censored values

one can either specify a non-informative prior or an informative prior by changing the parameter values for the beta distribution. To look at the effects of using an informative prior on η, another model was fit with prior specification of $\eta \sim$ transformed BETA$(8, 2)$. Derivation of the transformed beta distribution can be found in the Appendix. The eigenvalues for the H matrix depend on the distances between sampled locations. The smallest and largest eigenvalues for the Site 15 locations are h_1 = -.051219 and h_n = 0.26088, respectively. Hence, for this model, $\eta \in$ (-19.524, 3.833).

For the Metropolis-Hastings steps used for the simulation of η within the Gibbs sampler, a transformed beta distribution with the support of (-19.524, 3.833) was used as the candidate generating distribution. The candidate generating distribution used was a TBETA$(\beta_1 X, \beta_1(1 - X))$, resulting in the mean of the generating distribution to be centered around the current value, X, for η. The value of β_1 was set to be 2 and can be thought of as a tuning parameter that can be changed to increasing "mixing" of the chain. The results presented are based on 10,000 iterations, excluding the first 500. Time-series plots were constructed to verify convergence. At each iteration of the Gibbs sampler, 5 Metropolis-Hastings steps were completed.

The analysis of Metal A using data augmentation (DA), half the level of detection (LOD/2) and the level of detection (LOD) for the handling of the censored data resulted in vastly different parameter estimates for α and τ^2. As presented in Table 1 and Figures 3 through 5, data augmentation produced a smaller estimate for α and a much larger estimate for τ^2 as compared to the LOD/2 and the LOD methods. For the estimation of η, the data augmentation procedure produced a negative estimate for η, along with the DA method producing more variation in the marginal posterior distribution as compared to the LOD/2 and LOD methods. All three methods produce results that indicate no spatial dependence, with zero contained in the credible intervals. The lack of precision in estimating η may be due to the fact that 52 out of the 82 observations are censored. Hence, the lack of information available resulted in low precision in the estimation of

Table 1 Median and 95% credible intervals based on the simulated marginal posterior distributions for Metal A and Metal B

	DA		**LOD/2**		**LOD**	
A	Median	Interval	Median	Interval	Median	Interval
α	0.520	(-0.093, 0.990)	1.248	(0.921, 1.568)	1.687	(1.356, 2.00)
τ^2	2.808	(1.723, 4.942)	1.475	(1.111, 2.018)	1.155	(0.865, 1.593)
η	-3.815	(-15.532, 2.823)	-0.331	(-9.295, 3.121)	0.827	(-6.346, 3.240)
B	Median	Interval	Median	Interval	Median	Interval
α	-1.886	(-2.623, -1.220)	-1.371	(-1.799, -0.683)	-1.225	(-1.688, -0.747)
τ^2	3.343	(2.269, 5.091)	2.373	(1.784, 3.238)	2.078	(1.553, 2.860)
η	1.895	(-3.161, 3.385)	1.904	(-3.778, 3.441)	1.418	(-4.841, 3.327)

the dependence parameter η.

Metal B analysis produced similar findings with regards to the differences in results between the three methods. Table 1 and Figures 6 to 8 display the estimated marginal posterior distributions, medians and credible intervals for the parameters α, τ^2 and η. Once again, a lower estimate of α and a larger estimate of τ^2 were produced by the data augmentation method. By replacing the censored values with $LOD/2$ or LOD, the estimate of the variability was underestimated and the estimate of the mean was over estimated. With regards to the estimation of η, the three methods produced similar results, with data augmentation producing a slightly larger estimate of η.

With the locations of the censored observations, the observed values at locations close to the censored locations and the varying level of detections all effecting the estimation of the dependence parameter η, it is hard to say that data augmentation will always produce lower estimates of spatial dependence as compared to the LOD/2 and the LOD methods. For the case were the level of detections vary, with some level of detections being very large, it becomes even more difficult to make general statements about how the DA, LOD/2 and LOD methods will compare for the estimation of the spatial dependence parameter.

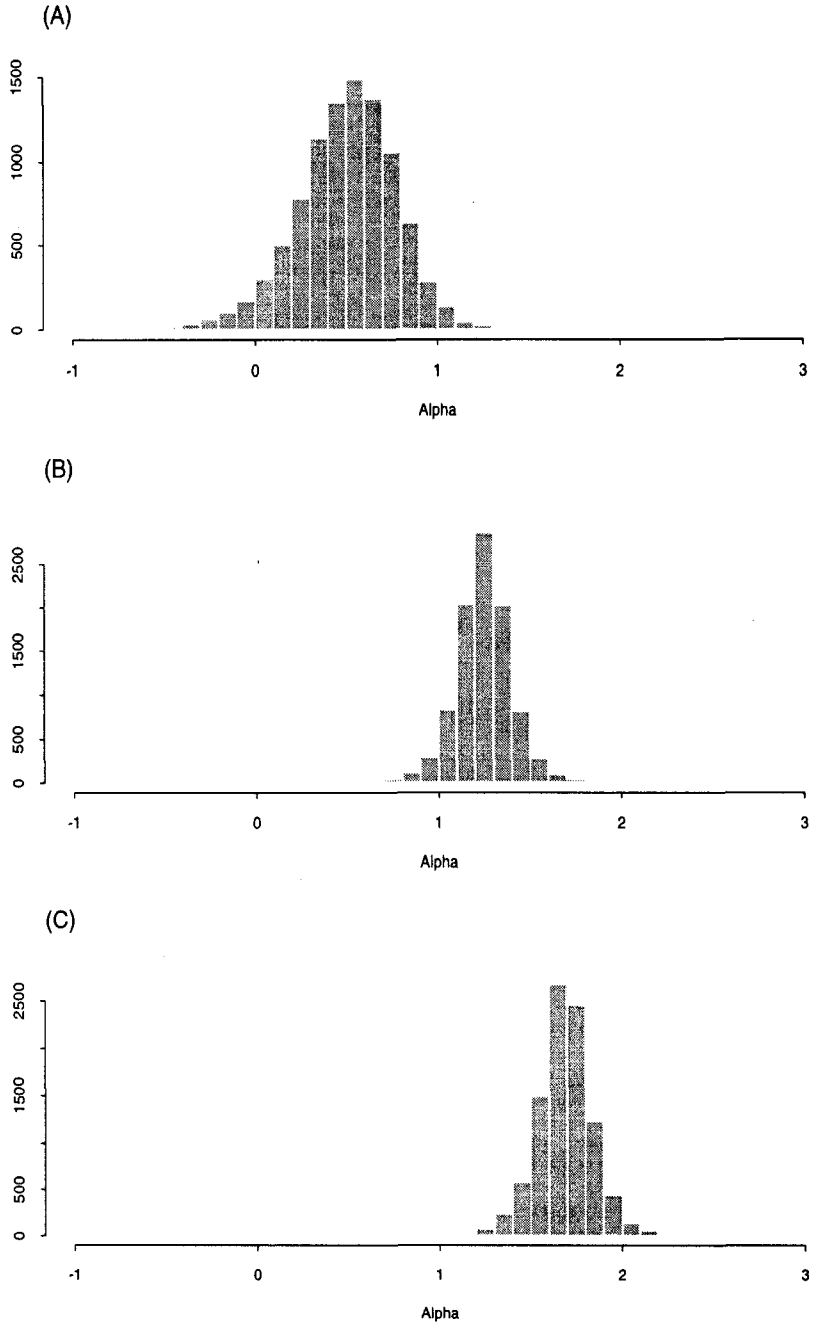

Figure 3 Metal A: Simulated marginal posterior distributions for α (A) data augmentation for censored values (B) censored values replaced by LOD/2 (C) censored values replaced by LOD

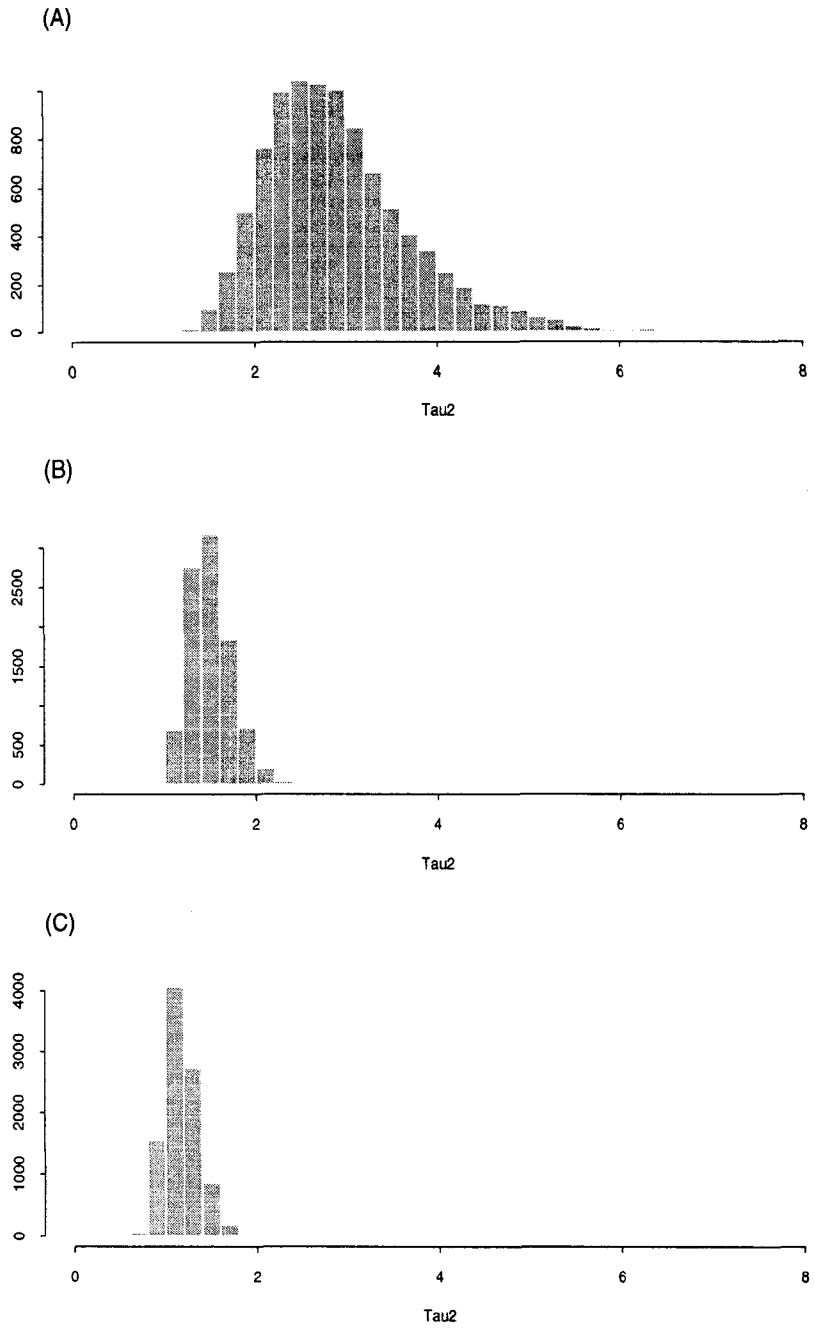

Figure 4 Metal A: Simulated marginal posterior distributions for τ^2 (A) data augmentation for censored values (B) censored values replaced by LOD/2 (C) censored values replaced by LOD

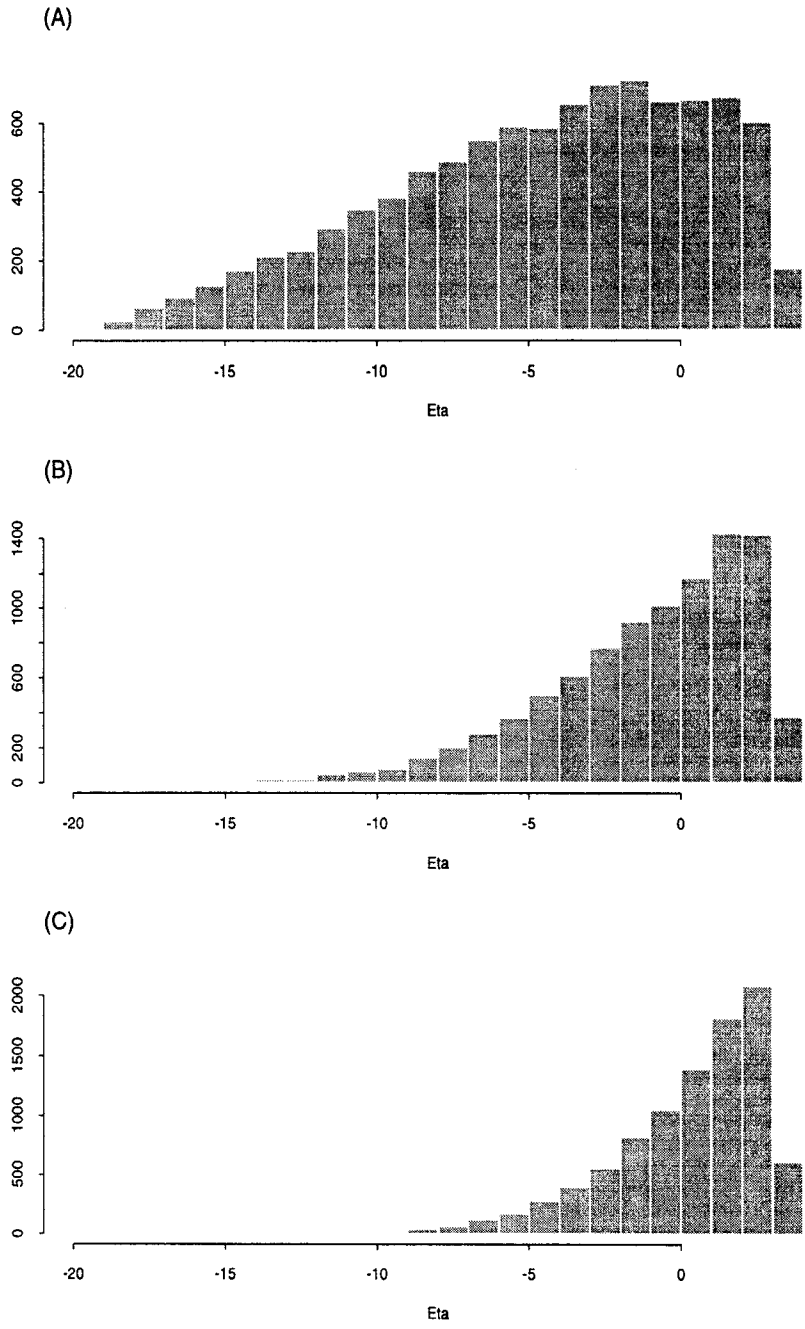

Figure 5 Metal A: Simulated marginal posterior distributions for η (A) data augmentation for censored values (B) censored values replaced by LOD/2 (C) censored values replaced by LOD

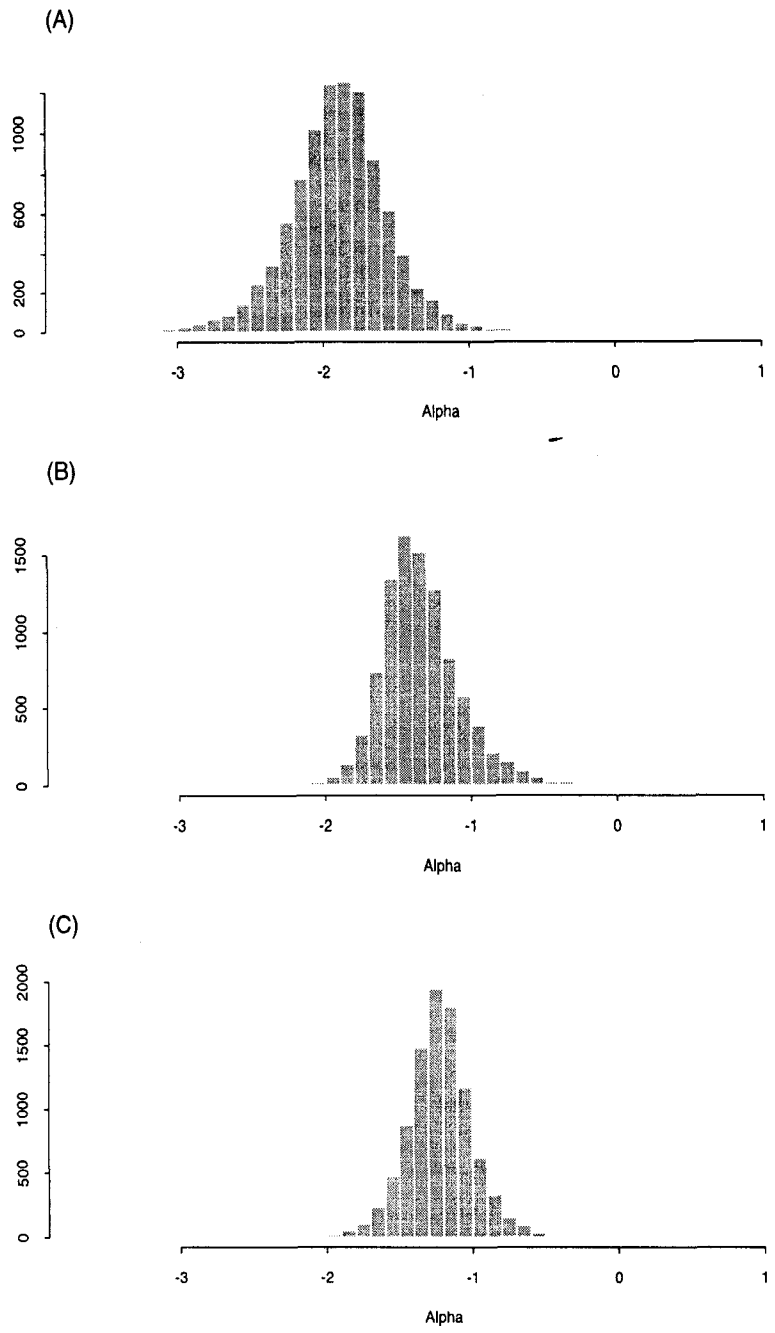

Figure 6 Metal B: Simulated marginal posterior distributions for α (A) data augmentation for censored values (B) censored values replaced by LOD/2 (C) censored values replaced by LOD

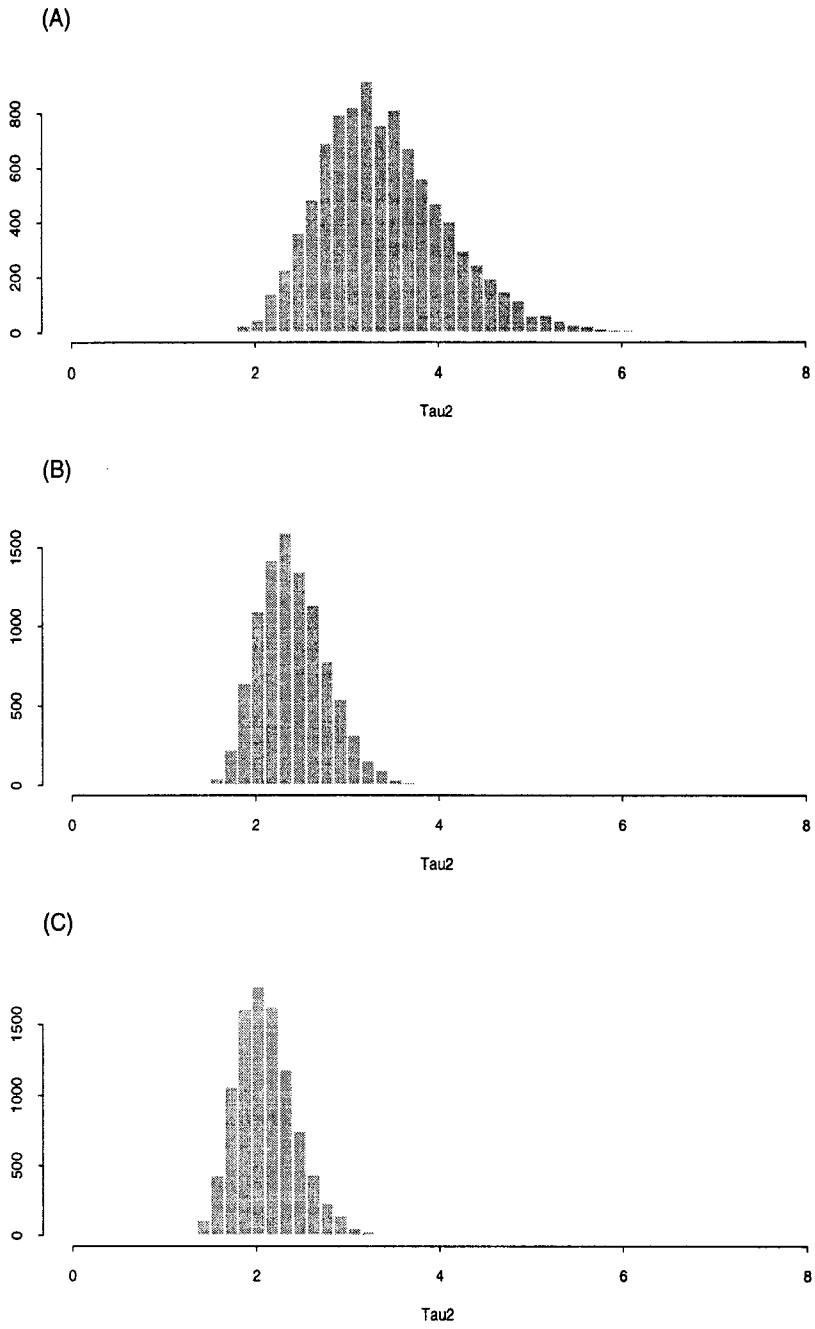

Figure 7 Metal B: Simulated marginal posterior distributions for τ^2 (A) data augmentation for censored values (B) censored values replaced by LOD/2 (C) censored values replaced by LOD

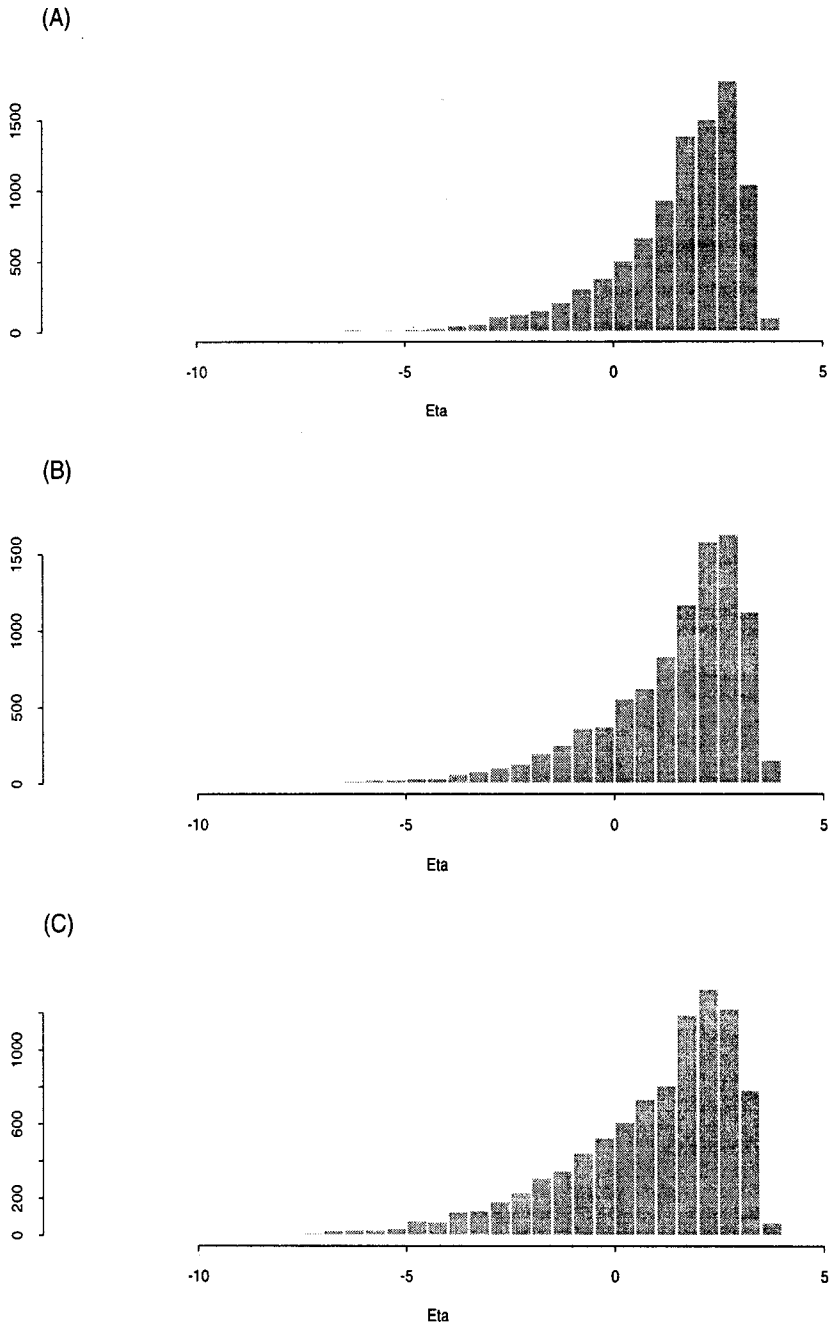

Figure 8 Metal B: Simulated marginal posterior distributions for η (A) data augmentation for censored values (B) censored values replaced by LOD/2 (C) censored values replaced by LOD

Table 2 Median and 95% credible intervals based on the simulated marginal posterior distributions for Metal A and Metal B using an informative prior for η

	Metal A		Metal B	
	Median	Interval	Median	Interval
α	0.546	(-0.100, 1.029)	-1.890	(-2.571,-1.263)
τ^2	2.781	(1.755, 4.909)	3.387	(2.322, 5.184)
η	-0.516	(-6.980,2.846)	1.708	(-2.473, 3.225)

Data augmentation produces estimates based on the observed values by using MCMC to integrate out the censored observations while the single imputation method treats the censored data as actual observed values. Thus, the variability in estimation with the LOD/2 or LOD method is under-estimated while data augmentation produces more accurate measures of estimation variability. Since variability in estimation is directly related to sample size and since the LOD/2 and LOD methods are treating all N observations as observed, these single imputation methods over-estimate the precision in estimation.

With specifying the prior for η as a transformed beta distribution, one has the flexibility of either using an informative prior or a non-informative prior. To look at the effects of using an informative prior on η, another model was fit with prior distribution for η being transformed BETA$(8,2)$. The range of possible values for η is -19.524 to 3.833. The use of this transformed beta distribution results in less probability given to large negative values of η and more probability given to values of η around 0.

Table 2 presents estimates and 95% credible intervals for Metal A and Metal B using an informative prior for η and data augmentation to handle the censored observations. Comparing the results in Table 2 to the results displayed in Table 1, we see no difference in parameters estimates for α and τ^2. With regards to the estimation of η, we see a large difference between the use of a non-informative and an informative prior. The use of an

Table 3 Median and 95% credible intervals based on the simulated marginal posterior distributions for Metal A and Metal B using three different non-informative prior specifications

	Primary Analysis		**Second Analysis**		**Third Analysis**	
A	Median	Interval	Median	Interval	Median	Interval
α	0.520	(-0.093, 0.990)	0.503	(-0.116, 0.970)	0.520	(-0.069, 0.977)
τ^2	2.808	(1.723, 4.942)	2.864	(1.771, 5.065)	2.803	(1.704, 4.909)
η	-3.815	(-15.53, 2.823)	-3.767	(-15.32 ,2.852)	-3.733	(-15.36, 2.866)
B	Median	Interval	Median	Interval	Median	Interval
α	-1.886	(-2.623, -1.220)	-1.753	(-2.376, -1.066)	-1.744	(-2.362, -0.907)
τ^2	3.343	(2.269, 5.091)	3.333	(2.318, 5.017)	3.392	(2.341, 5.082)
η	1.895	(-3.161, 3.385)	2.082	(-3.292, 3.425)	2.058	(-3.036, 3.451)

informative prior distribution resulted in both higher precision in estimation and larger point estimates for η. Thus, both the prior and the data are impacting the estimation of the dependence parameter. Care should be taken when using an informative prior for η. But, if there is prior knowledge with regards to the value of η, it could be incorporated into the prior distribution. In the case of the spatial dependence, more then likely there is no spatial dependence ($\eta = 0$) or positive spatial dependence ($\eta > 0$). Thus, it may seem reasonable to use a prior that puts less probability at large negative values of η.

Finally, sensitivity analysis was conducted to investigate the impact of the prior distributions. For both Metal A and Metal B, two additional analyses with varying prior distributions were completed. The prior specifications for Metal A were $\alpha \sim$ NOR$(0, 100)$, $\tau^2 \sim$ INGAM$(2.1, 3.3)$, $\eta \sim$ TBETA$(1, 1)$ and $\alpha \sim$ NOR$(1, 50)$, $\tau^2 \sim$ INGAM$(2.1, 2.2)$, $\eta \sim$ TBETA$(1, 1)$. The two additional analyses for Metal B used the prior distributions $\alpha \sim$ NOR$(-1, 50)$, $\tau^2 \sim$ INGAM$(2.1, 3.3)$, $\eta \sim$ TBETA$(1, 1)$ and $\alpha \sim$ NOR$(0, 100)$, $\tau^2 \sim$ INGAM$(2.1, 4.4)$, $\eta \sim$ TBETA$(1, 1)$. Results of the primary analyses and the two additional analyses for both Metal A and Metal B are presented in Table 3. The results show similar parameter estimates for the three different analyses using different prior distributions. Thus, we feel comfortable that the priors used in the

primary analysis of Metal A and Metal B are adequate.

6 Illustrative example: Missouri dioxin contamination

6.1 Description of data

In 1971, sections of a country road in Missouri were polluted with dioxin (2,3,7,8-tetrachlorodibenzo-p-dioxin or TCDD) contaminated waste. In November of 1983, investigation and determination of areas requiring clean-up was completed by the USEPA. Portions of this data, reported by Zirschky and Harris (1986), will be used to illustrate the data augmentation procedure for the analysis of censored spatial data. The data published by Zirschky and Harris only includes the sampled areas along the shoulder of the country road. The original study conducted by the USEPA was a much larger study that included areas beyond the shoulder of the road.

In the sampling of the locations, a regular sampling pattern was used with the X-direction representing direction parallel to the road and the Y-direction representing the direction perpendicular to or away from the road. The sampling was done by dividing the shoulder of the road into long transects in the X direction, in which 8 samples were taken. To get one measurement per transect, the 8 samples taken in a given transect were aggregated. Figure 9 displays the sampled locations, along with displaying which observations were censored. For our purposes, we will treat the values reported as coming from one sampled location, with the X coordinate indicating the start of the transect with the Y coordinate of 30 representing the road.

Of the 126 sampled locations, 43% of the observations fell below some level of detection (LOD). The detection levels varied, ranging from 0.10 μg/kg to 0.79 μg/kg. The clean-up criteria for dioxin is 1 μg/kg. Thus, none of the levels of detection were greater than the clean-up criteria. Varying levels of detections are due in part to the amount of soil, the type of soil, the moisture level, etc.

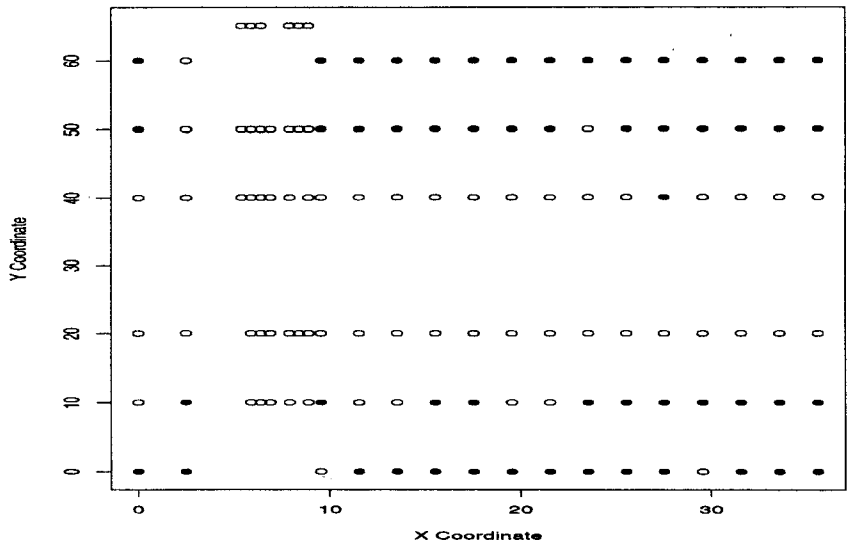

Figure 9 Missouri study locations, o represents an observed value and • represents a censored value

6.2 Model specification and results

Bayesian conditionally specified Gaussian model outlined in Section 3 was used to analyze the amount of dioxin present on the shoulder of the Missouri road. To satisfy the Gaussian assumption, a log-transformation was performed on the original observations. In addition to transforming the response variable, the X coordinate was transformed by a factor of 100 (i.e. X/100). In other words, the distance measure used was a variation of the commonly used Euclidean distance measure. Using Euclidean distance or the original X scale, there seems to be directional dependence, which was all but eliminated with the transformation or alternative distance measure. Another option would be to model the directional dependence (e.g. $c_{ij} = \eta_1 \sin^2(\theta_{ij})(\frac{m}{d_{ij}})^h + \eta_2 \cos^2(\theta_{ij})(\frac{m}{d_{ij}})^h$ if $s_j \in N_i$).

Priors specifications of $\alpha \sim \text{NOR}(0, 50)$, $\tau^2 \sim \text{INGAM}(2.1, 2.2)$, and $\eta \sim$ transformed BETA(1, 1), excluding $1/h_1$ and $1/h_n$ from the support set, were used in the analysis. These hyper-parameters result in diffuse prior distributions. We have chosen to place

Table 4 Median and 95% credible intervals based on the simulated marginal posterior distributions for the Missouri dataset

	DA		LOD/2		LOD	
	Median	Interval	Median	Interval	Median	Interval
α	-1.205	(-7.016, 3.969)	-0.852	(-5.249, 3.591)	-0.555	(-4.378, 3.321)
τ^2	4.401	(2.529, 13.083)	2.617	(1.487, 7.292)	1.917	(1.081, 5.404)
η	0.117	(0.100, 0.118)	0.117	(0.104, 0.118)	0.117	(0.105, 0.118)

a flat prior on η with the transformed BETA$(1,1)$ distribution resulting in an uniform distribution. An informative prior could be used for η by changing the hyper-parameter values in the specification of the transformed beta distribution. Care should be taken when using informative priors. For the parameter η, it may seem reasonable to use a prior which places less probability on large negative values, since η represents spatial dependence. For the sampled locations in the Missouri study, h_1 = -2.414 and h_n = 8.483, giving $\eta \in$ (-0.4143, 0.1179). Derivation of the transformed beta distribution can be found in the Appendix.

The Gibbs sampler with a data augmentation step was ran for 10,000 iterations. For the simulation of η, 5 Metropolis-Hastings steps were completed at each iteration of the Gibbs sampler. The candidate generating distribution used in the Metropolis-Hastings steps was a transformed BETA$(\beta_1 X, \beta_1(1-X))$ over the support (-0.4143, 0.1179), where X represents the current value for η. The value β_1, a "tuning" parameter, was set to 5 for the analysis. Time-series plots were used to verify convergence of the chain. Inferences were based on the last 9,500 iterations. Results are presented in Table 4 and Figures 10 to 12.

As with the site 15 example, the data augmentation method produced a much larger estimate for τ^2. Data augmentation produced an estimate of 4.401, while the LOD/2 and the LOD methods produced parameter estimates of 2.617 and 1.917, respectively. Along with producing a larger point estimate, Figures 10 through 12 illustrate the fact that the

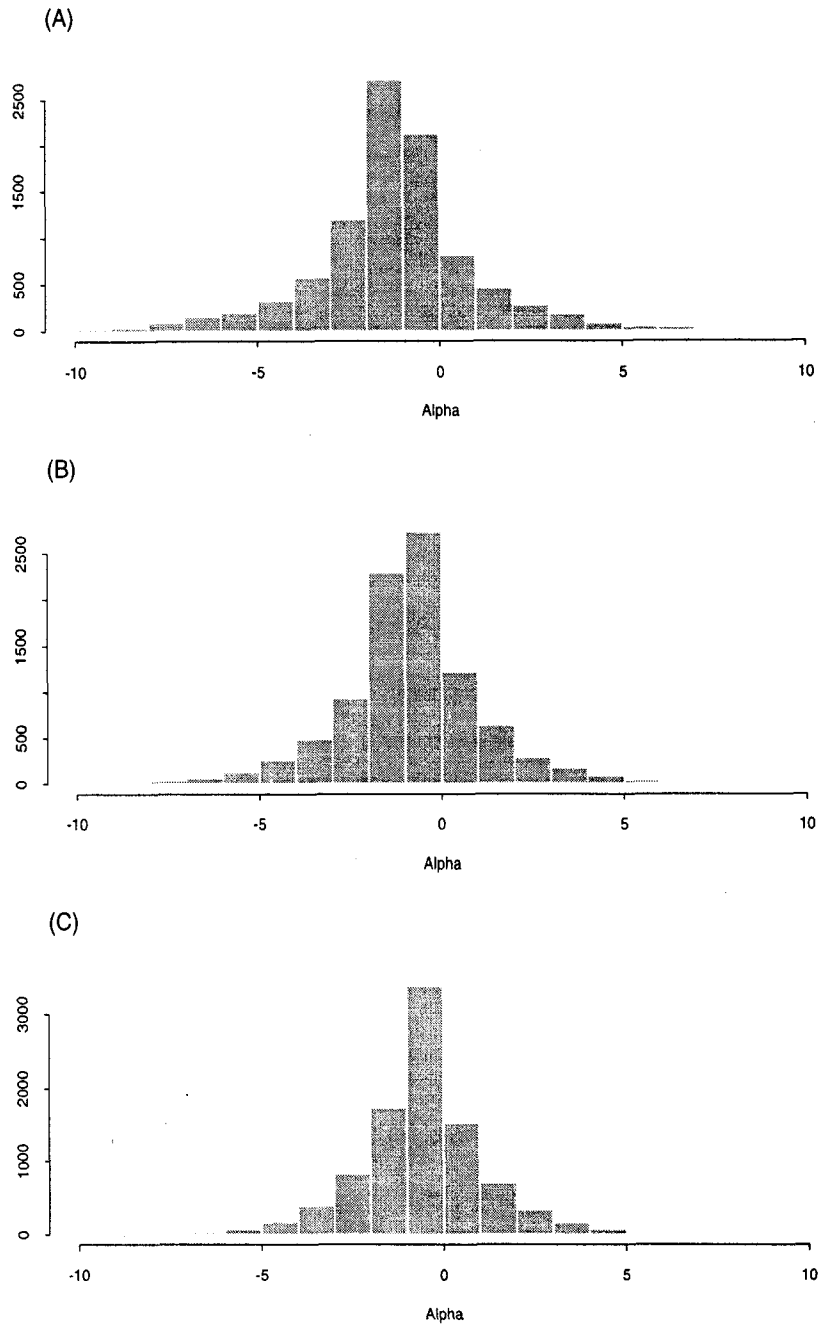

Figure 10 Missouri: Simulated marginal posterior distributions for α (A) data augmentation for censored values (B) censored values replaced by LOD/2 (C) censored values replaced by LOD

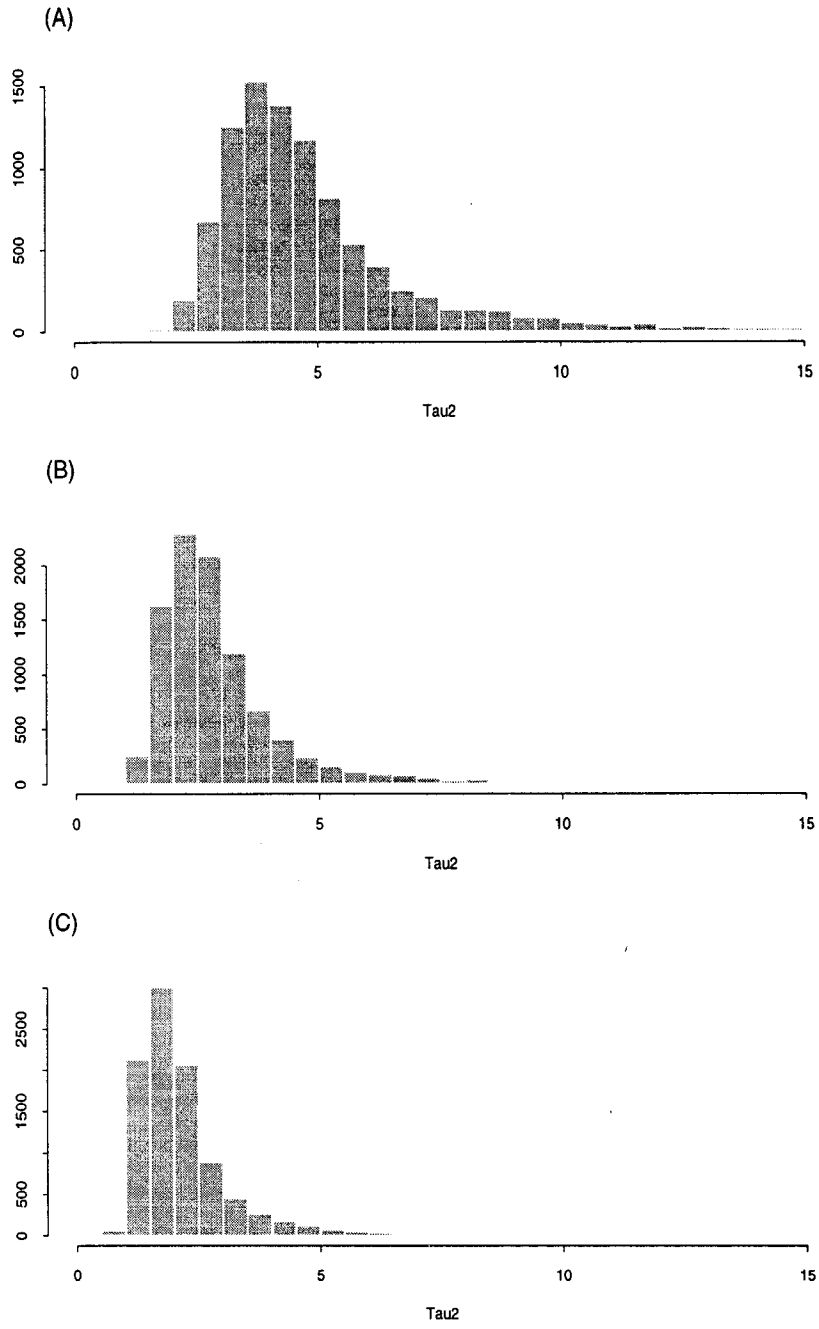

Figure 11 Missouri: Simulated marginal posterior distributions for τ^2 (A) data augmentation for censored values (B) censored values replaced by LOD/2 (C) censored values replaced by LOD

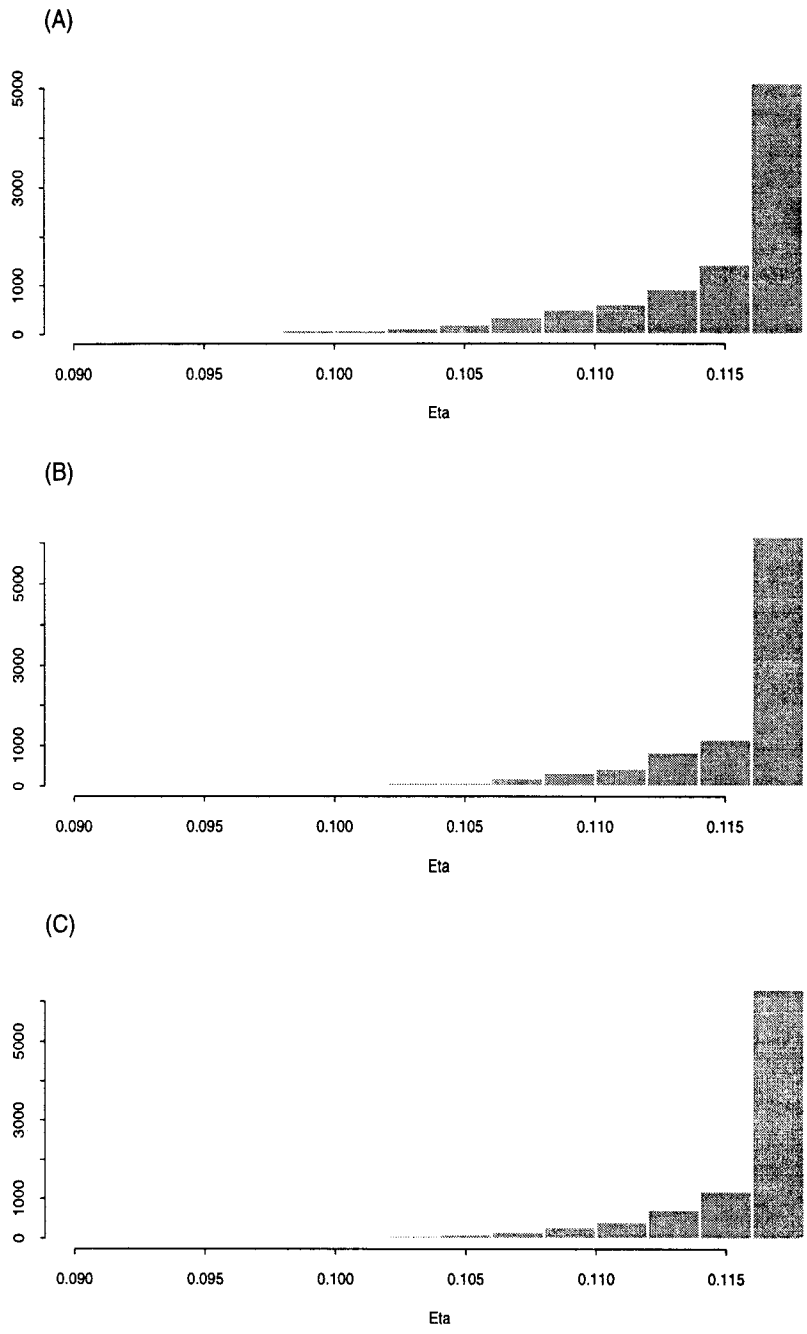

Figure 12 Missouri: Simulated marginal posterior distributions for η (A) data augmentation for censored values (B) censored values replaced by LOD/2 (C) censored values replaced by LOD

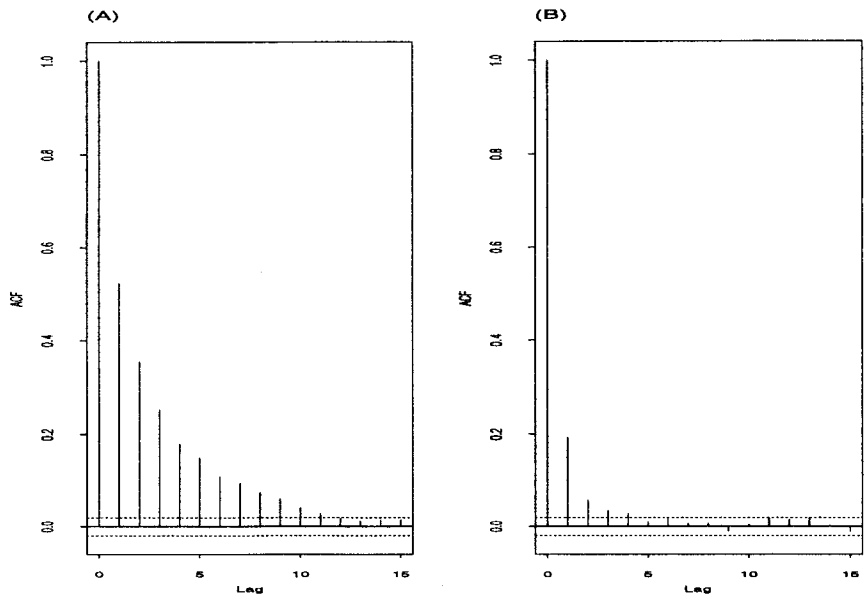

Figure 13 Missouri: Plot of autocorrelation function (ACF) for τ^2; (A) DA (B) LOD/2

data augmentation procedure produced more variability in the approximated marginal densities as compared to the LOD/2 and LOD methods. While data augmentation produced different parameter estimates for τ^2, the marginal densities for α and η did not differ greatly between the 3 methods, with spatial dependence indicated in all three results.

In addition to time-series plots for the verification of convergence, autocorrelation was computed for various lags. Figure 13 displays plots of the autocorrelation for the parameter τ^2. Figures 13 (A) and (B) represent the autocorrelations produced when data augmentation and the LOD/2 method are used to handle the censored observations, respectively. The autocorrelation for τ^2 is twice as large for the data augmentation method, as compared to the LOD/2 method. Censored spatial data model with a geostatistical model also produce larger autocorrelation with the data augmentation procedure (Fridley and Dixon, 2003). This occurrence of larger autocorrelations when data aug-

Table 5 Median and 95% credible intervals based on the simulated marginal posterior distributions for three different prior specifications for the Missouri dataset

	Primary Analysis		**Second Analysis**		**Third Analysis**	
	Median	Interval	Median	Interval	Median	Interval
α	-1.205	(-7.016, 3.969)	-1.203	(-3.676, 1.193)	-1.236	(-3.199, 0.681)
τ^2	4.401	(2.529, 13.083)	4.299	(2.810, 7.376)	4.333	(2.883, 7.041)
η	0.117	(0.100, 0.118)	0.111	(0.091, 0.117)	0.110	(0.090, 0.117)

mentation is employed to handle censored data is not unexpected. As Shafer states on page 84 of *Analysis of Incomplete Multivariate Data*, "If the missing information is a large portion of the total information, the θ will depend heavily on Y_{mis} at each P-step, which will in turn depend on the value of θ used in the previous I-step; successive iterates of θ will tend to be highly correlated and convergence will be slow." One may wish to use every k iterate for the estimation and inference if the autocorrelation is high, where k is set to the lag at which two iterates are uncorrelated.

Lastly, two additional analyses for the Missouri dataset was completed to investigate the impact of the prior specification. The two additional analyses used prior distributions $\alpha \sim \text{NOR}(1, 50)$, $\tau^2 \sim \text{INGAM}(2.1, 5.5)$, $\eta \sim \text{TBETA}(8, 2)$ and $\alpha \sim \text{NOR}(0, 50)$, $\tau^2 \sim \text{INGAM}(2.1, 3.3)$, $\eta \sim \text{TBETA}(1, 1)$. Results of the primary analysis and the two additional analyses for dioxin are presented in Table 5. The parameter estimates and intervals show no major differences in the results based on the three different prior specifications. Hence, we feel comfortable with the prior distributions used in the primary analysis and the subsequent results and inferences.

7 Discussion and conclusions

We have proposed a data augmentation approach for the handling of censored spatial observations model with a Bayesian conditionally specified Gaussian model. In doing

so, we discussed the use of a transformed beta distribution as the prior distribution for η. This allows for the specification of either a non-informative or an informative prior distribution that may reflect any prior knowledge about the spatial dependence parameter η.

The demonstration of the data augmentation method for the analysis of left censored observations was illustrated using an old industrial site contaminated with heavy metals and a site in Missouri contaminated with dioxin. Comparison of results for the site 15 and the Missouri site using data augmentation verses replacing the censored values with the level of detection and half the level of detection were also presented. These comparisons illustrated the differences in parameter estimation between the three methods. In the analysis Metal A, Metal B and dioxin, data augmentation produced larger estimates of variability as compared to estimates produced using the LOD/2 and the LOD methods. Data augmentation for the analysis of censored spatial data using a Bayesian geostatistical model produced similar results for the parameters representing variability (Fridley and Dixon, 2003).

The method can be easily extended to more complex models involving possible hyperpriors, hierarchical modeling, different parameterization of neighborhood structure and varying forms of censoring (i.e. interval censoring). Also, one should note that the imputation of the censored values at each iteration of the chain is conditional on the model. An incorrect model for the spatial process would lead to inaccurate augmentation or imputation for the censored data. Further work is needed to investigate the application of data augmentation to spatial settings and the robustness of the procedure to model misspecification, especially if the proportion of censored observations is large.

Along with robustness of the procedure, investigation into the issue of serial correlation in cases involving large proportions of censored data is needed. As seen with the dioxin example, the serial correlation for τ^2 was twice as high for the data augmentation method as compared to the LOD/2 method. The occurrence of large autocorrelations

using data augmentation for the analysis of censored data using a geostatistical model has also been illustrated and discussed by Fridley and Dixon (2003). The amount of serial correlation is directly related to the amount of censored observations. As the amount of censored observations increases, so does the level of autocorrelation. Further work is needed to investigate the issue of serial correlation when using the data augmentation method for the analysis of censored spatial data.

In addition to the investigation of model misspecification, sensitivity analysis is recommended with regards to the prior specifications. As an alternative to specifying values for the parameters in the prior distributions, a fully Bayesian analysis could be implemented. In doing so, care should be taken when using non-informative or improper priors in the setting of data augmentation, in that the resulting joint posterior distribution is proper. Shafer (1997) recommends using proper priors whenever in doubt due to the fact that even when a improper prior distribution is known to yield a proper joint posterior distribution in the complete data scenario, this is not always the case when it comes to data augmentation for missing/censored data.

In conclusion, this paper presents a data augmentation approach for the analysis of censored spatial data. Commonly, censored observations are set equal to some function of their level of detection. This ad hoc method of replacing the censored values with a constant results in biased parameter estimates. By imputing or augmenting values for the censored data at iteration of a Markov chain Monte Carlo, we more accurately estimate parameters in the setting involving censored data. As seen in the site 15 and Missouri dioxin examples, the level of variability was under-estimated with the LOD/2 and the LOD methods. Along with producing more accurate parameter estimates, data augmentation also produced more variability in the approximated marginal densities, particular in the case of estimating the variability parameter τ^2. Hence, data augmentation is a procedure that can be applied to analyze censored spatial data, which often occurs in environmental applications.

Appendix

This appendix presents the derivation of the full conditional distributions required for the Gibbs sampler using proper prior distributions and the derivation of the transformed beta distribution.

Full conditional distribution for τ^2:

The full conditional distribution for τ^2 is

$$p(\tau^2|\boldsymbol{y},\alpha,\eta) \propto p(\boldsymbol{y}|\tau^2,\alpha,\eta)p(\tau^2)$$
$$\propto (\tau^2)^{-(n/2+\gamma_o+1)} \exp\{\tfrac{-1}{\tau^2}(\tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{\alpha})^T(I-C)(\boldsymbol{y}-\boldsymbol{\alpha})+\beta_o)\}.$$

Hence, the full conditional distribution for τ^2 is

$$\tau^2|\boldsymbol{y},\alpha,\eta \sim INGAM(\tfrac{n}{2}+\gamma_o, \tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{\alpha})^T(I-C)(\boldsymbol{y}-\boldsymbol{\alpha})+\beta_o).$$

Full conditional distribution for α:

The full conditional distribution for α is

$$p(\alpha|\boldsymbol{y},\tau^2,\eta) \propto p(\boldsymbol{y}|\alpha,\tau^2,\eta)p(\alpha)$$
$$\propto \exp\{\tfrac{-1}{2}(\boldsymbol{y}-\boldsymbol{\alpha})^T M^{-1}(I-C)(\boldsymbol{y}-\boldsymbol{\alpha}) + \tfrac{-1}{2}(\boldsymbol{\alpha}-\mu_o\mathbf{1})^T(\sigma_o^2 I)^{-1}(\boldsymbol{\alpha}-\mu_o\mathbf{1})\}.$$

We will first find the full conditional distribution for $\boldsymbol{\alpha}$ and then the full conditional distribution for α, where $\boldsymbol{\alpha} = \mathbf{1}\alpha$. Completing the square, we have the full conditional distribution for $\boldsymbol{\alpha}$ to be

$$\boldsymbol{\alpha}|\boldsymbol{y},\tau^2,\eta \sim MVN(\boldsymbol{\mu}_\alpha, \Sigma_\alpha),$$

where $\boldsymbol{\mu}_\alpha = (\frac{1}{\sigma_o^2}I + \frac{1}{\tau^2}(I-C))^{-1}(\frac{\mu_o}{\sigma_o^2}\mathbf{1} + \frac{1}{\tau^2}(I-C)\boldsymbol{y})$ and $\Sigma_\alpha = (\frac{1}{\sigma_o^2}I + \frac{1}{\tau^2}(I-C))^{-1}$. Therefore, the full conditional distribution for α is

$$\alpha|\boldsymbol{y},\tau^2,\eta \sim N(\mu_\alpha, \sigma_\alpha^2),$$

where $\mu_\alpha = \frac{1}{n}\mathbf{1}^T\boldsymbol{\mu}_\alpha$ and $\sigma_\alpha^2 = \frac{1}{n^2}\mathbf{1}^T\Sigma_\alpha^2\mathbf{1}$.

Full conditional distribution for η:

The full conditional distribution for η is

$$p(\eta|\boldsymbol{y},\alpha,\tau^2) \propto p(\boldsymbol{y}|\alpha,\tau^2,\eta)p(\eta)$$
$$\propto |(I-C)^{-1}M|^{-1/2}\exp\{\tfrac{-1}{2}(\boldsymbol{y}-\boldsymbol{\alpha})^T M^{-1}(I-C)(\boldsymbol{y}-\boldsymbol{\alpha})\}$$
$$\times(\tfrac{h_n h_1}{h_1-h_n})^{\psi_o}\tfrac{1}{B(\psi_o,\phi_o)}(\eta-\tfrac{1}{h_1})^{\psi_o-1}[1-(\eta-\tfrac{1}{h_1})(\tfrac{h_n h_1}{h_1-h_n})]^{\phi_o-1}$$
$$\propto [\textstyle\prod_{i=1}^n(1-\eta h_i)]^{1/2}\exp\{\tfrac{\eta}{2\tau^2}(\boldsymbol{y}-\boldsymbol{\alpha})^T H(\boldsymbol{y}-\boldsymbol{\alpha})\}\times(\eta-\tfrac{1}{h_1})^{\psi_o-1}[1-(\eta-\tfrac{1}{h_1})(\tfrac{h_n h_1}{h_1-h_n})]^{\phi_o-1}.$$

There is no closed form for η's full conditional distribution (i.e. no known distribution). The full conditional distribution is only known up to a proportional constant. That is,

$$p(\eta|\boldsymbol{y},\alpha,\tau^2) \propto$$
$$[\textstyle\prod_{i=1}^n(1-\eta h_i)]^{1/2}\exp\{\tfrac{\eta}{2\tau^2}(\boldsymbol{y}-\boldsymbol{\alpha})^T H(\boldsymbol{y}-\boldsymbol{\alpha})\}(\eta-\tfrac{1}{h_1})^{\psi_o-1}[1-(\eta-\tfrac{1}{h_1})(\tfrac{h_n h_1}{h_1-h_n})]^{\phi_o-1}.$$

Derivation of transformed Beta distribution:

If the support of x is $1/h_1 \le x \le 1/h_n$ and $y = (x-\frac{1}{h_1})(\frac{h_n h_1}{h_1-h_n})$, we have $0 \le y \le 1$. Likewise, if the support of y is $0 \le y \le 1$ and $x = y(\frac{h_1-h_n}{h_n h_1}) + \frac{1}{h_1}$, we have $\frac{1}{h_1} \le x \le \frac{1}{h_n}$. Let $y \sim Beta(\alpha,\beta)$ and $x = g(y) = y(\frac{h_1-h_n}{h_n h_1}) + \frac{1}{h_1}$. Hence, we have $y = g^{-1}(x) = (x-\frac{1}{h_1})(\frac{h_n h_1}{h_1-h_n})$. By transformation, we have

$$f_x(x) = f_y((x-\tfrac{1}{h_1})(\tfrac{h_n h_1}{h_1-h_n})) \times |\tfrac{d}{dx}(x-\tfrac{1}{h_1})(\tfrac{h_n h_1}{h_1-h_n})|.$$

Now, $f_y((x-\frac{1}{h_1})(\frac{h_n h_1}{h_1-h_n})) = \frac{1}{B(\alpha,\beta)}[(x-\frac{1}{h_1})(\frac{h_n h_1}{h_1-h_n})]^{\alpha-1}[1-(x-\frac{1}{h_1})(\frac{h_n h_1}{h_1-h_n})]^{\beta-1}$ and $|\frac{d}{dx}(x-\frac{1}{h_1})(\frac{h_n h_1}{h_1-h_n})| = \frac{h_n h_1}{h_1-h_n}$. Thus, the distribution for the transformed beta random variable x is

$$f_x(x) = \tfrac{1}{B(\alpha,\beta)}(\tfrac{h_n h_1}{h_1-h_n})^{\alpha}[x-\tfrac{1}{h_1}]^{\alpha-1}[1-(x-\tfrac{1}{h_1})(\tfrac{h_n h_1}{h_1-h_n})]^{\beta-1},$$

with $1/h_1 \leq x \leq 1/h_n$.

References

Bell, B.S., and Broemeling, L.D. (2000). A Bayesian analysis for spatial processes with application to disease mapping. *Statistics in Medicine*, **19**, 957-974.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, B*, **36**, 192-236.

Besag, J., and Green, P.J. (1993). Spatial Statistics and Bayesian Computation. *Journal of the Royal Statistical Society, B*, **55**, 25-37.

Besag, Green, Higdon, and Mengersen (1995). Bayesian Computation and Stochastic Systems. *Statistical Science*, **10**, 3-66.

Carlin, B.P. and Louis, T.A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman & Hall, London.

Cressie, N.A.C. (1993). *Statistics for Spatial Data, Revised Edition.* John Wiley & Sons, Inc., New York.

Daniels, M.J., Lee, Y.D., and Kaiser, M.S. (2001). Assessing sources of variability in measurement of ambient particular matter. *Environmetrics*, **12**, 547-558.

Dempster, A.P., Laird, N.M, and Rubin, D.B. (1977). Maximum likelihood estimation form incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B*, **39**, 1-38.

Evans, M., and Swartz, T. (2000). *Approximating Integrals via Monte Carlo and Deterministic Methods.* Oxford University Press, Oxford.

Fridley, B.L., and Dixon, P. (2003). Data Augmentation for a Bayesian Spatial Model involving Censored Observations. In preparation.

Gelfand, A.E., and Smith, A.F.M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, **85**, 398-409.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). *Bayesian Data Analysis.* Chapman & Hall, London.

Gelman, A., and Meng, X. (1991). A note on bivariate distributions that are conditionally normal. *The American Statistican*, **45**, 125-126.

Geman, S., and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Transactions on Patter Analysis and Machine Intelligence*, **6**, 721-741.

Gibbons, R. (1995). Some Statistical and Conceptual Issues in the Detection of Low-Level Environmental Pollutants. *Environmental & Ecological Statistics*, **2**, 125-167.

Gilks, W.R., Richardson, S., and Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice.* Chapman & Hall, London.

Gumpertz, M.L., Graham, J.M., and Ristaino, J.B. (1997). Autologistic Model of Spatial Pattern of Phytophthora Epidemic in Bell Pepper: Effects of Soil Variables on Disease Presence. *Journal of Agricultural, Biological, and Environmental Statistics*, **2**, 131-156.

Hastings, W.K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, **57**, 97-109.

Helsel, D.R. (1990). Less than Obvious. Statistical Treatment of Data Below the Detection Limit. *Environmental Science Technology*, **24**, 1766-1774.

Hopke, P.K., Liu, C., and Rubin, D.B. (2001). Multiple Imputation for Multivariate Data with Missing and Below-Threshold Measurements: Time-Series Concentrations of Pollutants in the Arctic. *Biometrics*, **57**, 22-33.

Hughes, J.P. (1999). Mixed Effects Models with censored Data with Application to HIV RNA Levels. *Biometrics*, **55**, 625-629.

Kaiser, M.S., and Cressie, N. (2000). The Construction of Multivariate Distributions form Markov Random Fields. *Journal of Multivariate Analysis*, **73**, 199-220.

Kaiser, M.S., Cressie, N., and Lee, J. (2002). Spatial Mixture Models Based on Exponential Family Conditional Distributions. *Statistica Sinica*, **12**, 449-474.

Li, K.H. (1988). Imputations Using Markov Chains. *Journal of Statistical Computation and Simulation*, **30**, 57-79.

Little, R.J.A., and Rubin, D.B. (1987). *Statistical Analysis with Missing Data,* 2^{nd} *Ed.* Wiley, New York.

Pettitt, A.N. (1986). Censored Observations, Repeated Measures and Mixed Effects Models: An Approach Using the EM Algorithm and Normal Errors. *Biometrika*, **73**, 635-643.

Porter, P.S., Ward, R.C., Bell, H.F. (1988). The Detection Limit. Water Quality Monitoring Data Are Plagued with Levels of Chemicals That Are Too Low to Be Measured Precisely. *Environmental Science Technology*, **22**, 856-861.

Robert, C.P. (1995). Simulation of truncated normal variables. *Statistics and Computing*, **5**, 121-125.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data.* Chapman & Hall, London.

Smith, F.B., Helms, R.W. (1995). EM Mixed Model Analysis of Data From Informatively Censored Normal Distributions. *Biometrics*, **51**, 425-436.

Stern, H. and Cressie, N. (1999). *Disease Mapping and Risk Assessment for Public Health*, Chapter 5: Inference for Extremes in Disease Mapping. Wiley & Sons, Inc., New York.

Tanner, M.A., and Wong, W.H. (1987). The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, **82**, 528-540.

Xia, H., and Carlin, B.P. (1998). Spatio-Temporal Models with Errors in Covariates: Mapping Ohio Lung Cancer Mortality. *Statistics in Medicine*, **17**, 2025-2043.

Zirschky, J.H., and Harris, D.J. (1986). Geostatistical Analysis of Hazardous Waste Site Data. *Journal of Environmental Engineering*, **112**, 770-784.

SIMULATION STUDY: DATA AUGMENTATION FOR THE HANDLING OF SPATIALLY CENSORED OBSERVATIONS

A paper to be submitted to the Journal of Computational and Graphical Statistics

Brooke Fridley and Philip Dixon

This paper will present four simulation studies investigating the use of a data augmentation method for the analysis of censored spatial data. Censored spatial data occurs often in environmental applications. The two basic classes of models for the analysis of spatially correlated data, that of geostatistical and Markov random field models, are applied. Therefore, the simulation studies were completed for each type of model in a Bayesian framework. The goal of Simulation Studies I and III is to access the data augmentation procedure for the analysis of censored spatial data in the terms of parameter estimation and prediction. In addition to assessing the data augmentation procedure, comparison of the augmentation procedure to the common practice of replacing the censored values with half the level of detection will also be discussed. In contrast to Simulation Studies I and III, which investigates the general accuracy of the data augmentation method for only one combination of parameter values, Simulation Studies II and IV were designed to identify possible factors, like level of censoring and level of spatial dependence, that may impact the performance of the data augmentation procedure for the analysis of censored spatial data for each type of spatial model.

1 Introduction

In many environmental studies, observations may be censored for one reason or another. For example, in the measurement of wind speeds, it may be impossible to measure wind speeds accurately if they pass beyond some threshold value. The resulting observations are recorded as falling above the threshold value, or right censored. Conversely, left censoring occurs often when one is measuring trace amounts of pollutants in soil, air or water. In these cases, the observations are recorded as falling below some level of detection (LOD), which is often attributed to the analysis procedure or equipment. Due to the nature of the data, assuming independence is often invalid, leaving standard methods to handling censored observation inapplicable.

A common approach for the analysis of censored spatial data is to replace the censored values with some function of the level of detection (e.g. $LOD/2$, LOD). Once the censored values have been set equal to a constant, data analysis is completed on this "imputed" dataset as if all the values were observed. This approach has the disadvantage of producing biased parameter estimates, especially with the estimation of parameters representing sources of variability. In addition to producing biased point estimates, replacing the censored values with a constant, like $LOD/2$, will result in under-estimating the variability in the posterior distribution or standard errors.

One method that handles censored data in a spatial setting more adequately is the use of data augmentation. Data augmentation was first introduced by Tanner and Wong (1987). Fridley and Dixon (2003) have since applied the idea of augmentation to the analysis of censored spatial data in the context of both a Bayesian geostatistical model and a Bayesian conditionally specified or conditional auto-regressive (CAR) model. Fridley and Dixon found that replacing the censored observations with half their level of detection ($LOD/2$) or the level of detection (LOD) resulted in parameters measuring variability being underestimated. On the other hand, the data augmentation method

produced larger estimates of variability and smaller estimates of the mean. Likewise, Fridley and Dixon observed that predictions found in conjuncture with data augmentation for the handling of the censored observations resulted in different predictions as compared to the method of replacing the censored values with $LOD/2$.

The purpose of these four simulation studies is to investigate whether these results occur in many datasets. The analysis procedure used are Bayesian in nature, but we will use a frequentist approach for the evaluation of parameter estimates and predictions. Given parameter values, datasets are simulated for which analysis and posterior distribution are computed. Then for each dataset, the posterior median is computed as the point estimate from which characteristics of the estimates are examined using frequentist ideas of bias and mean square error.

The goal of this paper is to assess via simulation the data augmentation procedure presented by Fridley and Dixon (2003) for the analysis of censored spatial data. Simulation Studies I and III assess the data augmentation procedure in terms of parameter estimation and prediction in the context of a Bayesian geostatistical model and a Bayesian conditionally specified Gaussian model. In addition, comparison of the data augmentation method to the method of replacing the censored observations with half their level of detection ($LOD/2$) is also presented. In contrast to Simulation Studies I and III, Simulation Studies II and IV investigate factors, like percent censored and variability, that may impact the data augmentation procedure. For simplicity, we will refer to the methods to handle censored data as DA for data augmentation procedure and LOD/2 for the method that replaces the censored values with half the level of detections.

2 Data augmentation procedure

First introduced by Tanner and Wong (1987) and Li (1988), data augmentation is a procedure that can be use to handle missing data. In doing so, a Markov chain Monte

Carlo is used to "augment" or impute values for the missing or censored values at each iteration of the chain. Following the imputation for the censored or missing values, which Tanner and Wong called the I-step or imputation step, a posterior step or P-step is performed in which parameter values are simulated, conditional on the augmented data. That is, given the current value of the parameters $\boldsymbol{\Theta}^{(t)}$, augmentation is complete by drawing a vector $\boldsymbol{Y}_c^{(t+1)}$ from $p(\boldsymbol{Y}_c|\boldsymbol{Y}_o, \boldsymbol{\Theta}^{(t)})$, where $\boldsymbol{Y}_c$ and $\boldsymbol{Y}_o$ represent the censored and observed data. Then based on the current augmented data $\boldsymbol{Y}^{(t+1)} = (\boldsymbol{Y}_o^T, \boldsymbol{Y}_c^{(t+1)T})^T$, a posterior step is completed in which $\boldsymbol{\Theta}^{(t+1)}$ is generated from $p(\boldsymbol{\Theta}|\boldsymbol{Y}^{(t+1)})$. Once the chain has converged, say at iteration t^*, $\{\boldsymbol{\Theta}^{(t)} : t \geq t^*\}$ and $\{\boldsymbol{Y}_c^{(t)} : t \geq t^*\}$ can be thought of as draws from $p(\boldsymbol{\Theta}|\boldsymbol{Y}_o)$ and $p(\boldsymbol{Y}_c|\boldsymbol{Y}_o)$, respectively. That is, this process produces a stochastic sequence $\{\boldsymbol{\Theta}^{(t)}, \boldsymbol{Y}_c^{(t)} : t = 1, 2, ...\}$ whose stationary distribution is $p(\boldsymbol{\Theta}, \boldsymbol{Y}_c|\boldsymbol{Y}_o)$ (Shafer, 1997; Geman and Geman, 1984; Gilks, Richardson and Spiegelhalter, 1996). This procedure in essence integrates out the censored data from the posterior distribution, $p(\boldsymbol{\Theta}|\boldsymbol{Y}_o, \boldsymbol{Y}_c)$.

Therefore, data augmentation within a Gibbs sampler can be performed as a method to handle censored observations in a spatial setting, for both a Bayesian spatial model and the Bayesian conditionally specified Gaussian model. At each iteration of the Gibbs sampler, data will be imputed for the censored spatial data conditional on the current values of the parameters with subsequent generation of the parameters conditional on the complete, augmented dataset. In doing so, model and prior specification will be described along with the details of the Gibbs sampler algorithm used in the simulation studies (Fridley and Dixon, 2003).

3 Data augmentation within a Bayesian spatial model

3.1 Model specification and data augmentation procedure

Let a spatial stochastic process be represented by $\{Y(s) : s \in D\}$, for which s varies continuously over D, D in $\Re^2$. An isotropic spatial or geostatistical model is then,

$$Y(s_i) = \mu + W(s_i) + \varepsilon(s_i), \tag{1}$$

where $Y(s_i)$ represents the observation at location s_i, μ the overall mean, $\varepsilon(s_i)$ the random observational error at location s_i with $\varepsilon(s_i) \sim N(0, \tau^2)$, and $W(s_i)$ the random spatial effect at location s_i with $\boldsymbol{W}(\boldsymbol{s}) \sim MVN(\mathbf{0}, V(\sigma^2, \phi))$ (Cressie, 1993; Carlin and Louis, 1996). For the simulation study, the parameterization of the covariance matrix for the spatial dependence has an exponential form, with $V(\sigma^2, \phi)_{ij} = \sigma^2 \exp\{-d_{ij}/\phi\}$, $d_{ij} = \|s_i - s_j\|$ and $V^*(\phi) = \exp\{-d_{ij}/\phi\}$.

For the simulation study, we have chosen to use proper priors to insure that the joint posterior distribution is proper. The prior distributions placed on the parameters were

$$\begin{aligned}
\sigma^2 &\sim \text{INGAM}(\alpha, \beta), \\
\tau^2 &\sim \text{INGAM}(\gamma, \delta), \\
\mu &\sim \text{NOR}(\lambda, \psi^2), \\
\phi &\sim \text{GAM}(\eta, \theta).
\end{aligned}$$

For the Bayesian spatial model given in equation (1) with exponential parameterization of spatial covariance and proper priors, the following is the the Markov chain Monte Carlo with data augmentation step implemented within a Gibbs sampler as present by Fridley and Dixon (2003). Derivation of full conditional distributions can be found in Appendix I.

1. Set starting values for $\mu^{(0)}$, $\tau^{2(0)}$, $\sigma^{2(0)}$, $\boldsymbol{W}^{(0)}$, and $\phi^{(0)}$. Set censored values equal to their level of detection, $\boldsymbol{Y}_c^{(0)} = \boldsymbol{LOD}$ and $m = 0$.

2. Let $\boldsymbol{Y}^{T(m)} = (\boldsymbol{Y}_c^{(m)}, \boldsymbol{Y}_o)^T$, where $\boldsymbol{Y}_o$ represents the observed data and $\boldsymbol{Y}_c$ represents the censored data.

3. Generate $\mu^{(m+1)}$ from NOR$(\mu_1^{(m+1)}, \sigma_1^{2(m+1)})$, with

   $\mu_1^{(m+1)} = (\frac{\psi^2\tau^{2(m)}}{\tau^{2(m)}+\psi^2})[\frac{1}{\psi^2}\lambda + \frac{1}{\tau^{2(m)}}(\bar{Y}^{(m)} - \bar{W}^{(m)})]$ and $\sigma_1^{2(m+1)} = (\frac{1}{n})(\frac{\psi^2\tau^{2(m)}}{\tau^{2(m)}+\psi^2})$.

4. Generate $\tau^{2(m+1)}$ from INGAM$(n/2+\gamma, (1/2)(\boldsymbol{Y}^{(m)} - (\boldsymbol{\mu}^{(m+1)} + \boldsymbol{W}^{(m)}))^T(\boldsymbol{Y}^{(m)} - (\boldsymbol{\mu}^{(m+1)} + \boldsymbol{W}^{(m)})) + \delta)$.

5. Generate $\sigma^{2(m+1)}$ from INGAM$(n/2+\alpha, (1/2)\boldsymbol{W}^{T(m)}V^*(\phi^{(m)})^{-1}\boldsymbol{W}^{(m)} + \beta)$.

6. Generate $\boldsymbol{W}^{(m+1)}$ from MVN$\boldsymbol{\mu}_w^{(m+1)}, \Sigma_w^{(m+1)})$, where

   $\boldsymbol{\mu}_w^{(m+1)} = [V^{-1}(\sigma^{2(m+1)}, \phi^{(m)}) + \frac{1}{\tau^{2(m+1)}}I]^{-1}[\frac{1}{\tau^{2(m+1)}}(\boldsymbol{Y}^{(m)} - \boldsymbol{\mu}^{(m+1)})]$ and

   $\Sigma_w^{(m+1)} = [V^{-1}(\sigma^{2(m+1)}, \phi^{(m)}) + \frac{1}{\tau^{2(m+1)}}I]^{-1}$.

7. Using Metropolis-Hastings step(s), simulate $\phi^{(m+1)}$ from

   $p(\phi|\mu^{(m+1)}, \tau^{2(m+1)}, \sigma^{2(m+1)}, \boldsymbol{W}^{(m+1)}, \boldsymbol{Y}^{(m)})$

   $\propto \frac{\phi^{\eta-1}}{|V^*(\phi)|^{1/2}} \exp\{\frac{-1}{2\sigma^{2(m+1)}}\boldsymbol{W}^{T(m+1)}V^*(\phi)^{-1}\boldsymbol{W}^{(m+1)} - \theta\phi\}$

8. Have $\boldsymbol{\Theta}^{(m+1)} = (\mu^{(m+1)}, \tau^{2(m+1)}, \sigma^{2(m+1)}, \phi^{(m+1)}, \boldsymbol{W}^{(m+1)})$.

9. Using $\boldsymbol{\Theta}^{(m+1)}$, impute values for $\boldsymbol{Y}_c$ and get $\boldsymbol{Y}_c^{(m+1)}$. Let $\boldsymbol{Y}_c = (Y_{1c}, Y_{2c}, .., Y_{kc})$.

   (a) Generate $Y_{1c}^{(m+1)}$ from NOR$(\mu^{(m+1)} + W_1^{(m+1)}, \tau^{2(m+1)})$, truncated at LOD_1.

   ...

   (b) Generate $Y_{kc}^{(m+1)}$ from NOR$(\mu^{(m+1)} + W_k^{(m+1)}, \tau^{2(m+1)})$, truncated at LOD_k.

10. Set $m = m + 1$ and repeat algorithm a large number of times.

3.2 Simulation Study I

The first simulation study was conducted using data augmentation for the analysis of censored observations within a Bayesian geostatistical spatial model. The goal of the simulation study was to investigate properties of the estimates and predictions produced by imputing values for the censored observations within a Markov chain Monte Carlo. In addition to assessing the validity of the data augmentation procedure, Simulation Study I also compares the data augmentation method to the method of replacing the censored observations with half their level of detection.

3.2.1 Estimation

The first goal of Simulation Study I is to assess properties of the parameter estimates produced by the DA and LOD/2 methods. 1000 generated datasets were constructed containing 100 observations on a 10x10 regular grid or lattice. The data were simulated using the exponential parameterization of the spatial covariance matrix with parameter values of $\mu = 0$, $\tau^2 = 1$, σ^2 =5, ϕ=10 and % censored = 20%. To finish the specification of the Bayesian model, proper diffuse priors, centered at the truth, were specified. The priors used in the simulation study were

$$\mu \sim \text{NOR}(0, 50),$$
$$\tau^2 \sim \text{INGAM}(2.1, 1.1),$$
$$\sigma^2 \sim \text{INGAM}(2.1, 5.5),$$
$$\phi \sim \text{GAM}(1, 0.1).$$

For each simulated dataset, the Gibbs sampler outlined in Section 3.1 was run for 3,000 iterations, with a single Metropolis-step for the simulation of ϕ. The estimation was based on the last 2,000 iterations of the chain. In addition to the use of DA for the handling of the censored observations, an analysis replacing the censored observations with half their level of detection was completed in order to compare the two methods.

The results using the DA and the LOD/2 methods for the 1000 simulated datasets are displayed in Tables 1 through 3 and Figures 1 and 2.

Figure 1 present plots of the 1000 estimates produces via DA verses the 1000 estimates produced using the LOD/2 method. Estimates were taken to be the median of their marginal posterior density. Figure 1 (A) and (C) show the DA method systemically producing smaller estimates for μ and larger estimates of σ^2 as compared to the LOD/2 method. Estimates of $P(\hat{\mu}_{DA} < \hat{\mu}_{LOD/2})$, $P(\hat{\tau}^2_{DA} < \hat{\tau}^2_{LOD/2})$, $P(\hat{\sigma}^2_{DA} < \hat{\sigma}^2_{LOD/2})$ and $P(\hat{\phi}_{DA} < \hat{\phi}_{LOD/2})$ were found to be 0.993, 0.272, 0.00 and 0.491, respectively.

Summary and graphical displays of the estimates for μ, τ^2, σ^2 and ϕ across the 1000 simulated datasets are displayed in Table 1 and Figure 2. From these displays, one can observe that the DA method produced estimates of μ, τ^2 and σ^2 closer to the true values of 0, 1, and 5, with little difference in the estimation of ϕ between the two methods. Boxplots displaying $\hat{\mu} - \mu$, $\hat{\tau}^2 - \tau^2$, $\hat{\sigma}^2 - \sigma^2$, and $\hat{\phi} - \phi$ are presented in Figure 2. These boxplot illustrate the difference in estimation between the DA and LOD/2 methods. The largest discrepancy between the two methods is in regards to the estimation of the spatial variability, σ^2. With the LOD/2 method, the average estimate of σ^2 was 2.778, while data augmentation produced an average estimate of 4.897, almost twice as large. Furthermore, Figure 2 shows the data augmentation method producing more variability in the estimates for the parameters τ^2 and σ^2 in relation to the LOD/2 method.

To assess the estimation procedure more quantitatively, estimates of the mean square error (MSE), bias $E(\hat{\theta}) - \theta$, and variance $V(\hat{\theta})$ were computed. Estimates of MSE, bias and variance were found by computing the sample mean and sample variance of the 1000 estimates, producing an estimate of $E(\hat{\theta})$ and $Var(\hat{\theta})$. The estimates of MSE, bias and variance for the DA and the LOD/2 methods are displayed in Table 2. The estimated MSE for μ and σ^2 are much larger for the LOD/2 method. As for τ^2, the estimate of the MSE is larger for data augmentation, even though estimates produced by the LOD/2 method are more biased. This is due to the fact that there is more variability in the

Table 1 Summary of estimates for the 1000 simulated datasets

DA	Parameter	Min	Q1	Median	Mean	Q3	Max
	μ	-1.156	-0.228	0.011	0.011	0.254	1.205
	τ^2	0.458	0.644	0.752	0.833	0.928	3.719
	σ^2	2.388	4.047	4.768	4.897	5.628	9.016
	ϕ	1.738	7.598	9.539	9.673	11.646	21.202
LOD/2	Parameter	Min	Q1	Median	Mean	Q3	Max
	μ	-0.788	0.094	0.359	0.353	0.610	1.771
	τ^2	0.335	0.577	0.664	0.704	0.780	2.206
	σ^2	1.320	2.289	2.660	2.778	3.164	5.298
	ϕ	1.723	7.075	9.421	9.611	11.812	23.601

Table 2 Estimates of bias, variance and mean square error for estimation of μ, τ^2, σ^2 and ϕ using data augmentation and LOD/2 method.

DA	Parameter	Bias	Variance	MSE
	μ	0.011	0.134	0.134
	τ^2	-0.167	0.103	0.131
	σ^2	-0.103	1.305	1.315
	ϕ	-0.326	9.728	9.834
LOD/2	Parameter	Bias	Variance	MSE
	μ	0.353	0.138	0.262
	τ^2	-0.296	0.036	0.124
	σ^2	-2.222	0.459	5.394
	ϕ	-0.389	13.033	13.184

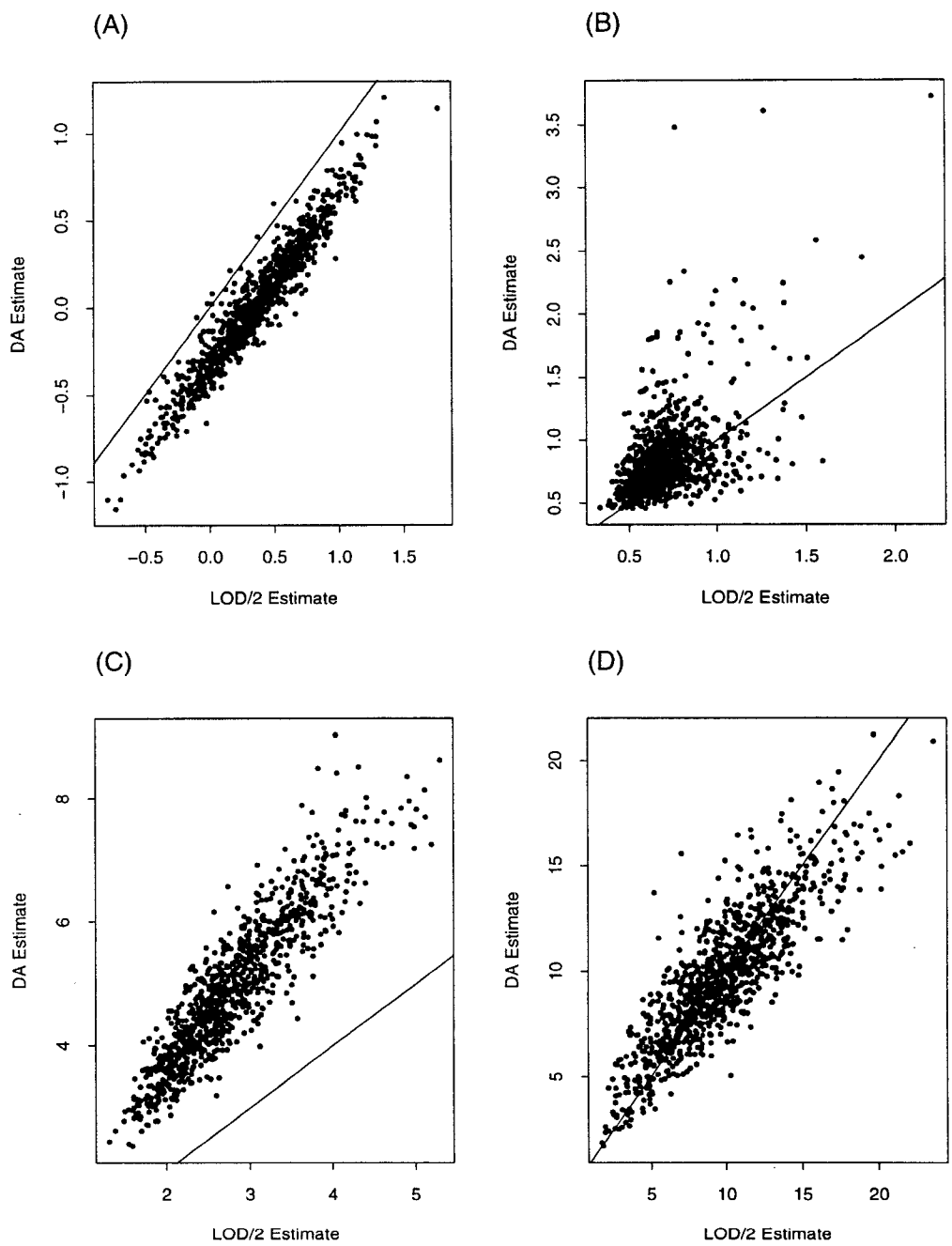

Figure 1 Scatterplot of estimates found via DA and LOD/2; (A) μ (B) τ^2 (C) σ^2 (D) ϕ

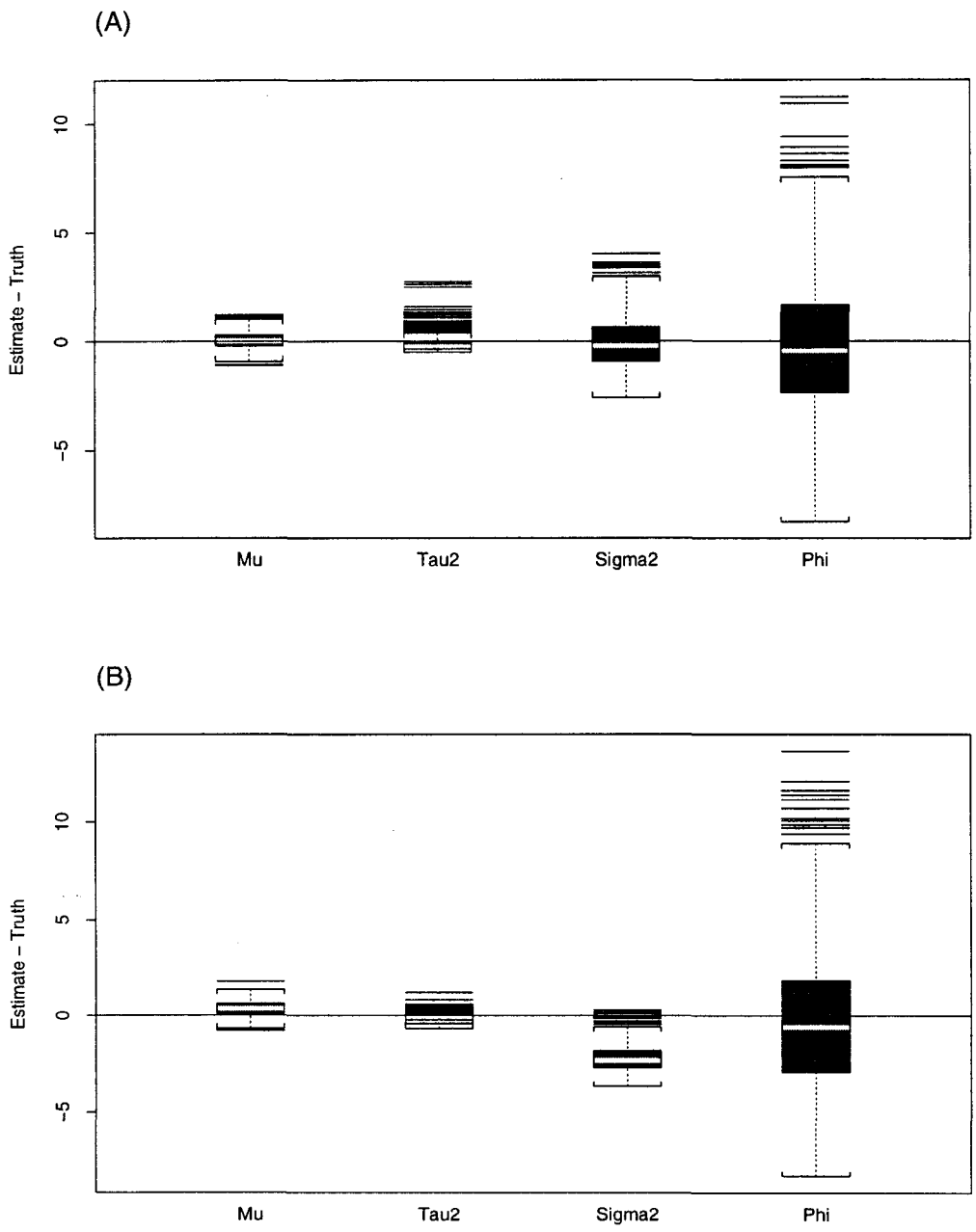

Figure 2 Boxplots of $\hat{\mu}$ - μ, $\hat{\tau}^2$ - τ^2, $\hat{\sigma}^2$ - σ^2 and $\hat{\phi}$ - ϕ using (A) data augmentation and (B) LOD/2 method

Table 3 Summary of lengths for 95% credible intervals for the 1000 simulated datasets

DA	Parameter	Min	Q1	Median	Mean	Q3	Max
	μ	0.744	1.381	1.578	1.589	1.788	2.488
	τ^2	1.092	2.268	2.845	2.912	3.434	7.025
	σ^2	2.514	4.872	5.747	5.932	6.703	18.257
	ϕ	5.900	13.968	17.341	19.106	21.780	64.779
LOD/2	Parameter	Min	Q1	Median	Mean	Q3	Max
	μ	0.555	1.130	1.339	1.348	1.574	2.364
	τ^2	0.634	1.432	1.724	1.759	2.024	4.221
	σ^2	1.227	2.545	3.099	3.223	3.725	7.551
	ϕ	5.564	14.869	19.248	21.367	26.254	57.415

estimation of τ^2 using the data augmentation procedure as compared to the LOD/2 method (0.103 vs. 0.036).

In addition to investigating point estimates, lengths of 95% equal-tail credible intervals were also computed. Summary results are presented in Table 3. As seen with point estimates, intervals for τ^2 and σ^2 tended to be larger with the use of data augmentation. Intervals for σ^2 and ϕ tended to be large, with a few intervals for ϕ being quite large. This lack of precision in estimating the spatial range parameter ϕ may be attributed to the sample size. With only 100 observations, in which 20% are censored, it maybe quite difficult to estimate the spatial range parameter with any precision.

3.2.2 Prediction

The second goal of Simulation Study I is to compare the error in prediction produced using the data augmentation method to the prediction error resulting from replacing the censored observations with half their level of detection ($LOD/2$). To investigate the aspect of prediction, 50 simulated datasets were constructed on a regular 15 x 15 lattice with 5 units between nearest neighbors. This resulted in 225 observations per dataset. The datasets were simulated with an exponential parameterization of the spatial

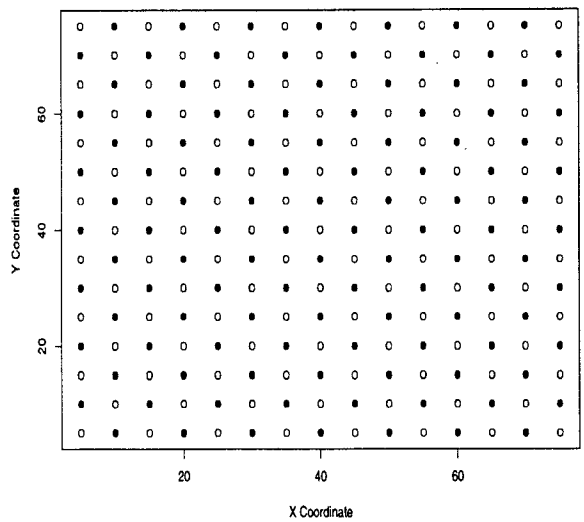

Figure 3 Locations for simulation study investigating prediction error, o represent locations used in parameter estimation and • represent locations used for prediction

covariance matrix using parameter values of $\mu = 0$, $\tau^2 = 1$, $\sigma^2 = 5$ and $\phi = 10$ with 20% of the observations censored. Half of the simulated dataset, 112 observations, was set aside for use in the prediction stage of the simulation study. This dataset would be use as the "truth" for which subsequent predictions would be compared. The remaining 113 sampled locations, constituting the observed data, were used in parameter estimation along with prediction. To illustrate further, Figure 3 displays the observed locations and the predicted locations. One thing to note is that the locations for prediction represent the best possible scenario for prediction, since most locations are surrounded by four observed locations.

The prediction stage of the analysis was completed using the Bayesian prediction method. In Bayesian prediction, the posterior predictive distribution, $p(\boldsymbol{Y}_u|\boldsymbol{Y}_g)$, is used as the means for prediction, where $\boldsymbol{Y}_g$ represent the gauged or observed locations and $\boldsymbol{Y}_u$ represent the ungauged or predicted locations. In the case of censored data

and augmentation, $\boldsymbol{Y}_g$ is partition into $\boldsymbol{Y}_{go}$ and $\boldsymbol{Y}_{gc}$ representing the gauged observed values and the gauged censored values, respectively. Since the joint distribution of $\boldsymbol{Y}_u$ and $\boldsymbol{Y}_g$ follows a multivariate normal distribution, the posterior predictive distribution can be approximated by simulating predictions from

$$\boldsymbol{Y}_u|\boldsymbol{Y}_{go}, \boldsymbol{Y}_{gc}^{(k)}, \boldsymbol{\Theta}^{(k)} \sim \text{MVN}(\boldsymbol{\mu}_{u.g}^{(k)}, \Sigma_{u.g}^{(k)}),$$

with $\boldsymbol{\mu}_{u.g}^{(k)} = \boldsymbol{\mu}_u^{(k)} + \Sigma_{ug}^{(k)}\Sigma_{gg}^{-1(k)}(\boldsymbol{Y}^{*(k)} - \boldsymbol{\mu}_g^{(k)})$, $\Sigma_{u.g}^{(k)} = \Sigma_{uu}^{(k)} - \Sigma_{ug}^{(k)}\Sigma_{gg}^{-1(k)}\Sigma_{gu}^{(k)}$, and $\boldsymbol{Y}^{*(k)} = (\boldsymbol{Y}_{go}^T, \boldsymbol{Y}_{gc}^{(k)T})^T$, for a large number of MCMC iterations, k (Carlin and Louis, 1996; de Oliveira and Ecker, 2002; Fridley and Dixon, 2003).

The same analysis procedure and priors outlined in Section 3.2.1 were used for the estimation of parameters within a Markov chain Monte Carlo. Approximation of the posterior predictive distribution was completed using every 5^{th} iteration from iteration 1000 to 3000. In other words, the posterior predictive distribution for each location was approximated via 400 simulated predictions. The prediction at a given location i, $\hat{y}_i$, was then taken to be the median of the simulated predicted distribution. Using these predictions and the truth, the mean prediction error (MPE) and mean squared prediction error (MSPE) were computed for each simulated dataset (i.e. $\sum_{i=1}^{n}(\hat{y}_i - y_i)/n$ and $\sum_{i=1}^{n}(\hat{y}_i - y_i)^2/n$). To compare the data augmentation method to the LOD/2 method, this procedure was completed for the 50 simulated datasets. Each simulated dataset was analyzed twice; once using data augmentation for the handling of the censored observations and once using the LOD/2 method. Results are displayed in Table 4 and Figures 4 through 6.

Table 4 and Figures 4 and 5 illustrate the fact that the data augmentation method not only produces better parameter estimates, but also better predictions. Across the 50 simulated datasets, data augmentation produced smaller MSPEs, with the except of one simulated dataset. The case when data augmentation out-performed the LOD/2 the most and vice verse are displayed in Figure 6. Figure 6 (A) represents the case when

Table 4 Summary of mean prediction error (MPE) and mean squared prediction error (MSPE) for the 50 simulated datasets using data augmentation and LOD/2 method for the handling of censored data

DA	Measure	Min	Q1	Median	Mean	Q3	Max
	MPE	-0.448	-0.191	-0.023	-0.027	0.119	0.356
	MSPE	2.197	2.925	3.186	3.203	3.526	4.308
LOD/2	Measure	Min	Q1	Median	Mean	Q3	Max
	MPE	0.047	0.288	0.465	0.443	0.583	0.875
	MSPE	2.752	3.255	3.698	3.778	3.975	5.798
LOD/2-DA	Measure	Min	Q1	Median	Mean	Q3	Max
	MSPE	-0.138	0.342	0.567	0.575	0.728	1.543

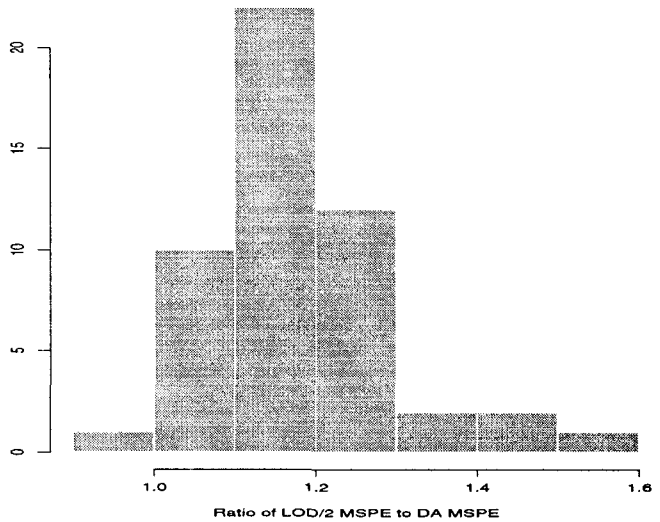

Figure 4 Histogram of the ratio of LOD/2 MSPE to DA MSPE

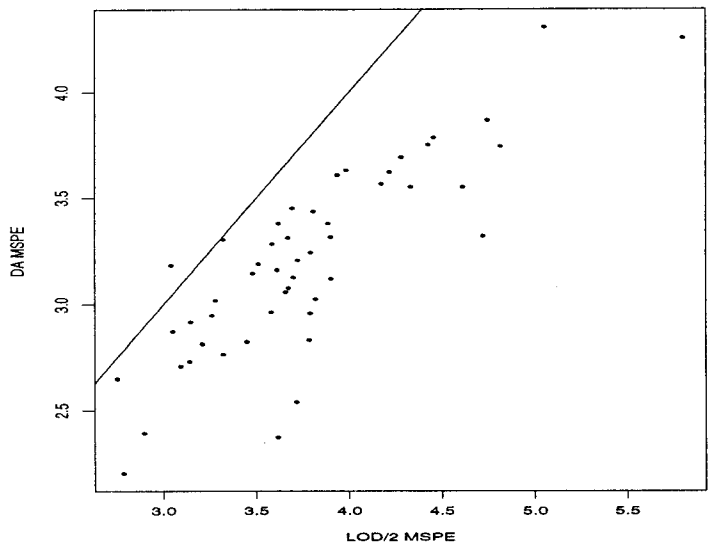

Figure 5 Scatterplot of DA MSPE and LOD/2 MSPE

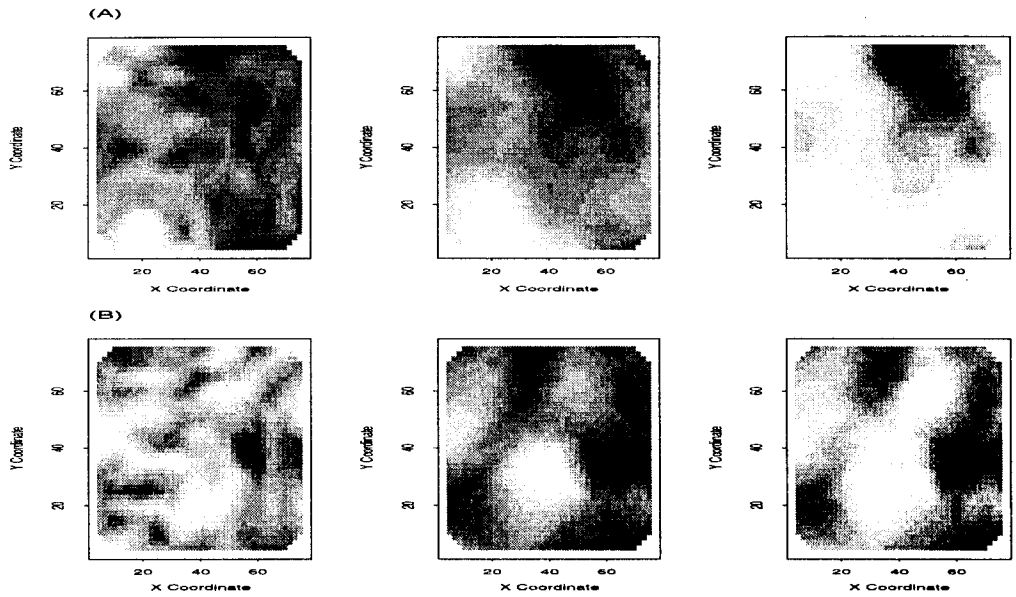

Figure 6 Map of the truth, predicted surface using DA and predicted surface using LOD/2 method; (A) Simulated dataset that resulted in largest superior performance with DA (B) Simulated dataset that resulted in the only superior performance via the LOD/2 method

data augmentation produced a MSPE of 4.256 while LOD/2 produced a value of 5.798. Conversely, Figure 6 (B) displays the simulation resulting in a MSPE equaling 3.044 for the LOD/2 method and 3.183 for data augmentation. In addition to the LOD/2 method producing larger MSPEs, with the largest MSPE being 5.798, each simulated dataset produced MPE greater than 0 (i.e. $\sum_{i=1}^{n}(\hat{y}_i - y_i)/n > 0$). Hence, the LOD/2 method is over-estimating when it comes to prediction.

Simulation Study I shows the data augmentation procedure for the handling of censored spatial data in the context of a Bayesian spatial model to be superior to the common method of replacing the censored values with $LOD/2$. Along with producing more accurate point estimates, the DA method produced larger credible intervals for the parameters (i.e. more variability in the approximated marginal densities for the parameters). Simulation Study I and all results were based on only one set of parameter values ($\mu = 0$, $\tau^2 = 1$, $\sigma^2 = 5$, $\phi = 10$, percent censored = 20%, N=100). The generalities of these results for other parameter combinations are investigated in Section 3.3: Simulation Study II. In addition, Simulation Study II is focused on determining which factors, if any, impact the performance of the data augmentation procedure for the analysis of censored spatial data.

3.3 Simulation Study II

Simulation Study II is a study to investigate factors that may impact the performance of the data augmentation procedure for spatially censored data, in terms of accuracy and precision in estimation. The factors investigated are sample size (N), percent censored, level of variability (τ^2), level of spatial variability (σ^2), and level of spatial dependence (ϕ). The factor levels for Simulation Study II can be found in Table 5. The standard parameter values, sample size and percent censored were set to be μ=0, $\tau^2 = 1$, $\sigma^2 = 5$, ϕ=10 and N=100 (10x10 regular lattice) with 20% of the data censored. For example, to investigate the effects of percent censored, the simulation of the datasets would be

Table 5 Factor levels for Simulation Study II

Factor	Level 1	Level 2	Level 3	Level 4
Sample Size (N)	7x7	10x10	15x15	—
% Censored	0	20	40	60
Variability (τ^2)	0.5	1.5	5.0	—
Spatial Variability (σ^2)	0.5	1.5	5.0	—
Spatial Dependence (ϕ)	5.0	10.0	15.0	—

completed using μ=0, τ^2 = 1, σ^2 =5, ϕ=10 and N=100 with percent censored levels of 0%, 20%, 40% and 60%. This gives a total of 16 scenarios. For each scenario, 50 simulated datasets were generated using the spatial model outlined in Section 3.1 with exponential parameterization of the spatial covariance matrix.

The sampled locations were on a square lattice with 10 units distance between nearest neighbors. To produce censored observations, level of detection values were determined based on the level of the percent censored factor. For instance, if 20% of the observations were to be censored, a LOD value would be found such that the proportion of the data below the LOD value was 20%. Any observation falling below the set level of detection would be coded as "$< LOD$". Therefore, the detection level did not vary within a simulated dataset.

To complete the analysis, proper priors were placed on all parameters. The hyper-parameters used in the prior specification resulted in distributions center around the truth with large, but finite, variances. Let σ^{2*}, τ^{2*} and ϕ^* represent the true value of σ^2, τ^2 and ϕ used in the generation of the simulated datasets. Based on the factor being investigated, the priors used in the analysis were

$$\mu \sim NOR(0, 50),$$
$$\tau^2 \sim INGAM(2.1, 1.1(\tau^{2*})),$$
$$\sigma^2 \sim INGAM(2.1, 1.1(\sigma^{2*})),$$
$$\phi \sim GAM(0.1(\phi^*), 0.1).$$

For example, analysis involving the percent censored factor level of 20% would use the prior distributions $\mu \sim \text{NOR}(0, 50)$, $\tau^2 \sim \text{INGAM}(2.1, 1.1(1))$, $\sigma^2 \sim \text{INGAM}(2.1, 1.1(5))$, and $\phi \sim \text{GAM}(0.1(10), 0.1)$.

The data augmentation procedure outlined in Section 3.1 was used for the analysis of the simulated spatial data involving censored observations. The Gibbs sampler was run for 4,000 iterations with the last 3,000 iterations used for estimation and inference. The simulation of ϕ, within the Gibbs sampler, was completed with one Metropolis-Hastings step using the candidate generating distribution GAM($2X$, 2), where X represents the current value of ϕ. Results for Simulation Study II are presented in Tables 6 and 7 and Figures 7 through 12.

Results in terms of estimation accuracy are displayed in Table 6 and Figures 7 to 9. Average parameter estimates for the parameters at the various factor levels are presented in Table 6. Figures 7, 8 and 9 graphically display estimated bias and 95% confidence intervals are plotted for each bias estimate. All factors seem to have some impact (small or moderate) on estimation, with the exception of the percent censored. A possible explanation for the level of percent censoring not impacting the estimates is the fact that the censored observations are handling in a reasonable fashion. That is, the censored observations were intergrated out of the joint posterior via MCMC. Thus, the percent censoring does not impact point estimation but instead the precision in estimation. The figures also show that higher sample sizes resulted in lower levels of bias in the estimation of σ^2 and ϕ, while at higher levels of variability (τ^2) the amount of bias in ϕ increased. The level of spatial variability (σ^2) seemed to impact the estimation bias of all parameters, with the exception of μ. The level of spatial variability impacted the estimation of both ϕ and τ^2; at the lowest level of σ^2, ϕ was estimated poorly while the estimation of τ^2 improved as the level of spatial variability decreased. Lastly, the level of dependence only effected the accuracy in estimating the spatial parameters (σ^2 and ϕ). At the lower level of ϕ, there was more difficulty in estimating σ^2 and

Table 6 Average parameter estimates for the various factor levels

Factor	Level 1	Level 2	Level 3	Level 4
N	49	100	225	—
μ	0.02	-0.06	-0.02	—
τ^2	0.81	0.86	0.81	—
σ^2	4.35	4.89	4.97	—
ϕ	8.78	9.49	9.748	—
% Censored	0	20	40	60
μ	0.03	0.02	0.00	0.05
τ^2	0.76	0.86	0.82	0.77
σ^2	5.20	5.32	4.88	5.07
ϕ	10.47	10.31	10.31	9.30
Variability, τ^2	0.5	1.5	5.0	—
μ	0.02	-0.05	0.03	—
τ^2	0.40	1.25	4.92	—
σ^2	5.08	5.01	4.90	—
ϕ	9.78	9.31	7.86	—
Spatial Variability, σ^2	0.5	1.5	5.0	—
μ	0.00	0.06	0.00	—
τ^2	1.00	0.95	0.84	—
σ^2	0.42	1.47	4.90	—
ϕ	8.38	9.94	9.52	—
Spatial Dependence, ϕ	5	10	15	—
μ	-0.06	-0.08	0.08	—
τ^2	0.87	0.79	0.78	—
σ^2	4.70	5.12	5.24	—
ϕ	3.17	9.35	14.91	—

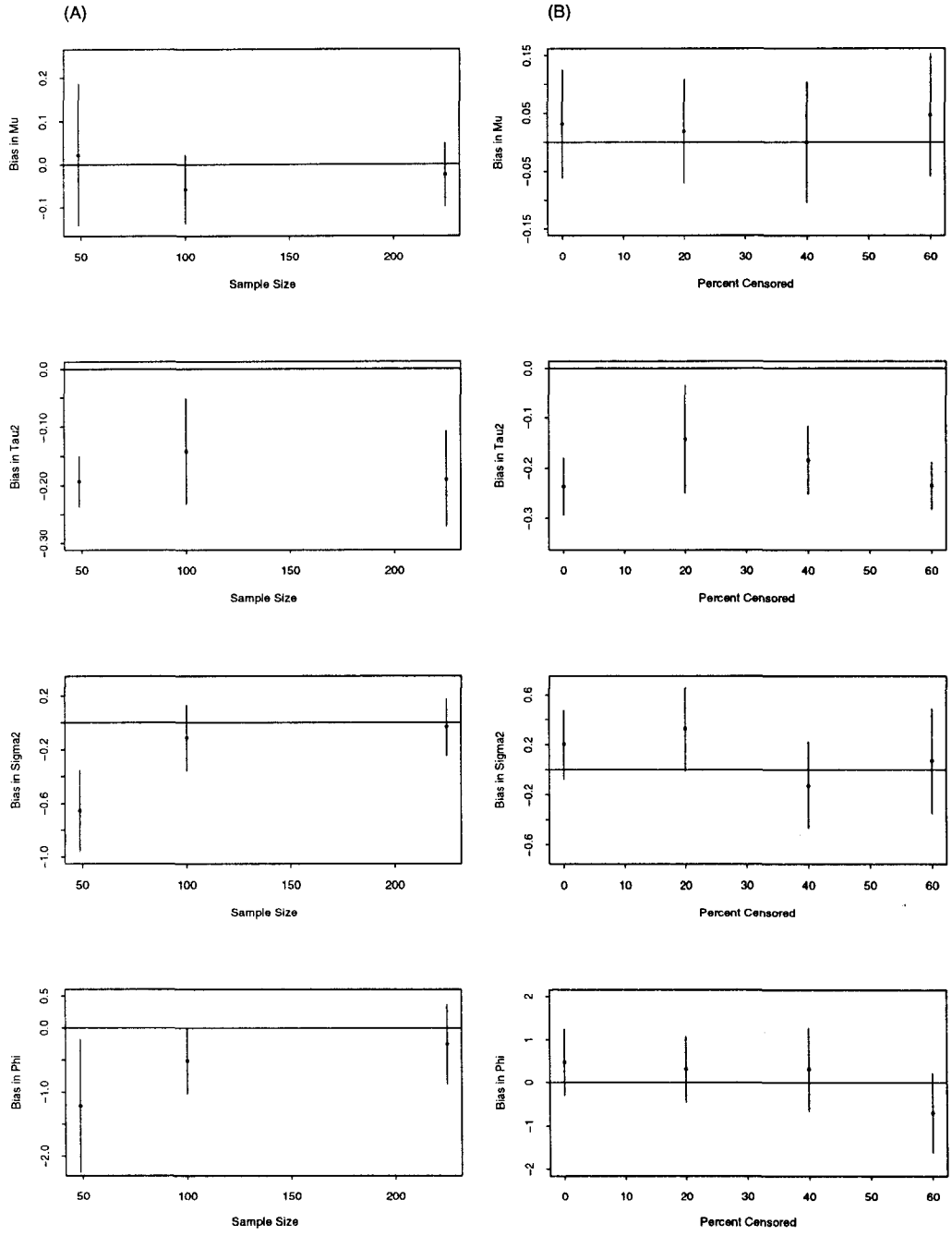

Figure 7 Plot of 95% confidence intervals for the average estimate for the various factor levels; (A) Sample Size (B) Percent Censored

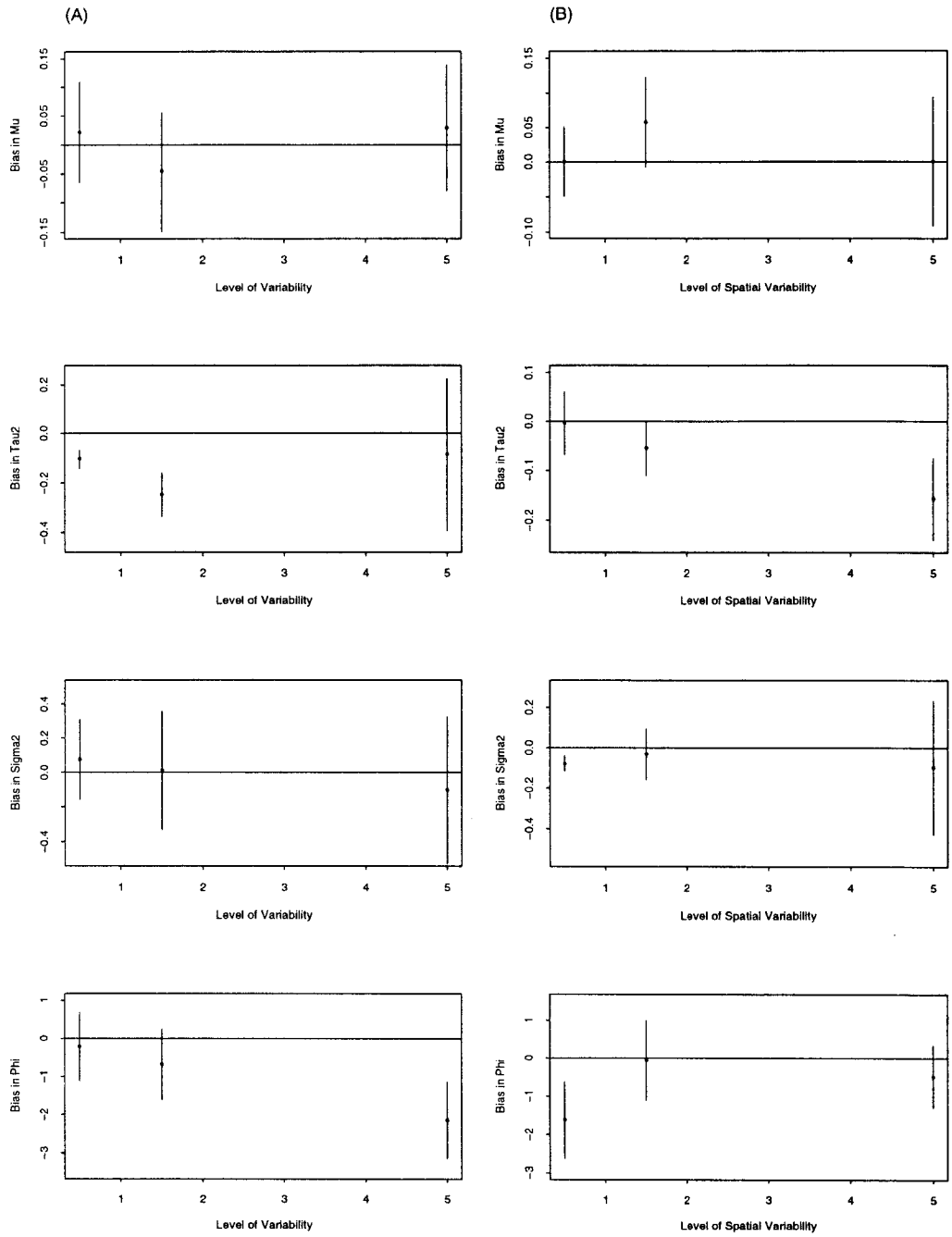

Figure 8 Plot of 95% confidence intervals for the average estimate for the various factor levels; (A) Variability (B) Spatial Variability

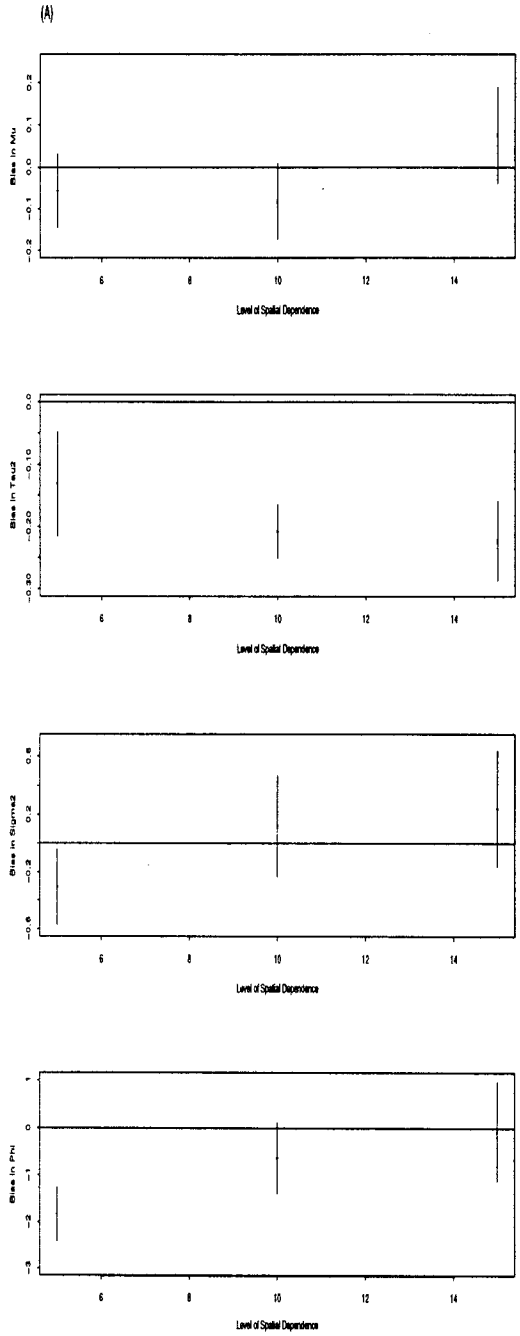

Figure 9 Plot of 95% confidence intervals for the average estimate for the various levels of spatial dependence

Table 7 Average length of credible intervals for the various factor levels

Factor	Level 1	Level 2	Level 3	Level 4
N	49	100	225	—
μ	2.10	1.58	1.08	—
τ^2	3.23	2.94	2.55	—
σ^2	6.86	5.88	4.24	—
ϕ	23.27	18.10	12.79	—
% Censored	0	20	40	60
μ	1.68	1.73	1.72	1.67
τ^2	2.63	2.92	2.99	3.26
σ^2	5.79	6.44	6.55	7.58
ϕ	17.21	18.87	20.83	19.21
Variability, τ^2	0.5	1.5	5.0	—
μ	1.60	1.62	1.62	—
τ^2	2.11	3.76	8.34	—
σ^2	5.56	6.54	9.29	—
ϕ	16.01	20.88	24.83	—
Spatial Variability, σ^2	0.5	1.5	5.0	—
μ	0.72	1.13	1.58	—
τ^2	1.28	1.82	3.02	—
σ^2	1.13	2.40	6.07	—
ϕ	29.97	27.96	19.44	—
Spatial Dependence, ϕ	5	10	15	—
μ	1.10	1.60	1.91	—
τ^2	3.97	3.03	2.31	—
σ^2	5.63	6.21	6.81	—
ϕ	10.81	17.42	27.30	—

ϕ accurrately. The effects of ϕ and σ^2 on one another maybe due to the connection between σ^2 and ϕ; as σ^2 tends to 0, ϕ is undefined.

In addition to accuracy in estimation, another goal of the simulation study was to investigate factors that may possibly impact precision in estimation. Length of 95% equal-tail Bayesian credible intervals were used as the measure of precision. Table 7 displays the average length of intervals for the various factor levels. As seen in Table 7, as sample size increased the mean length of credible intervals decreased, while most of the interval lengths increased as the amount of variability (τ^2 or σ^2) increased. Figures

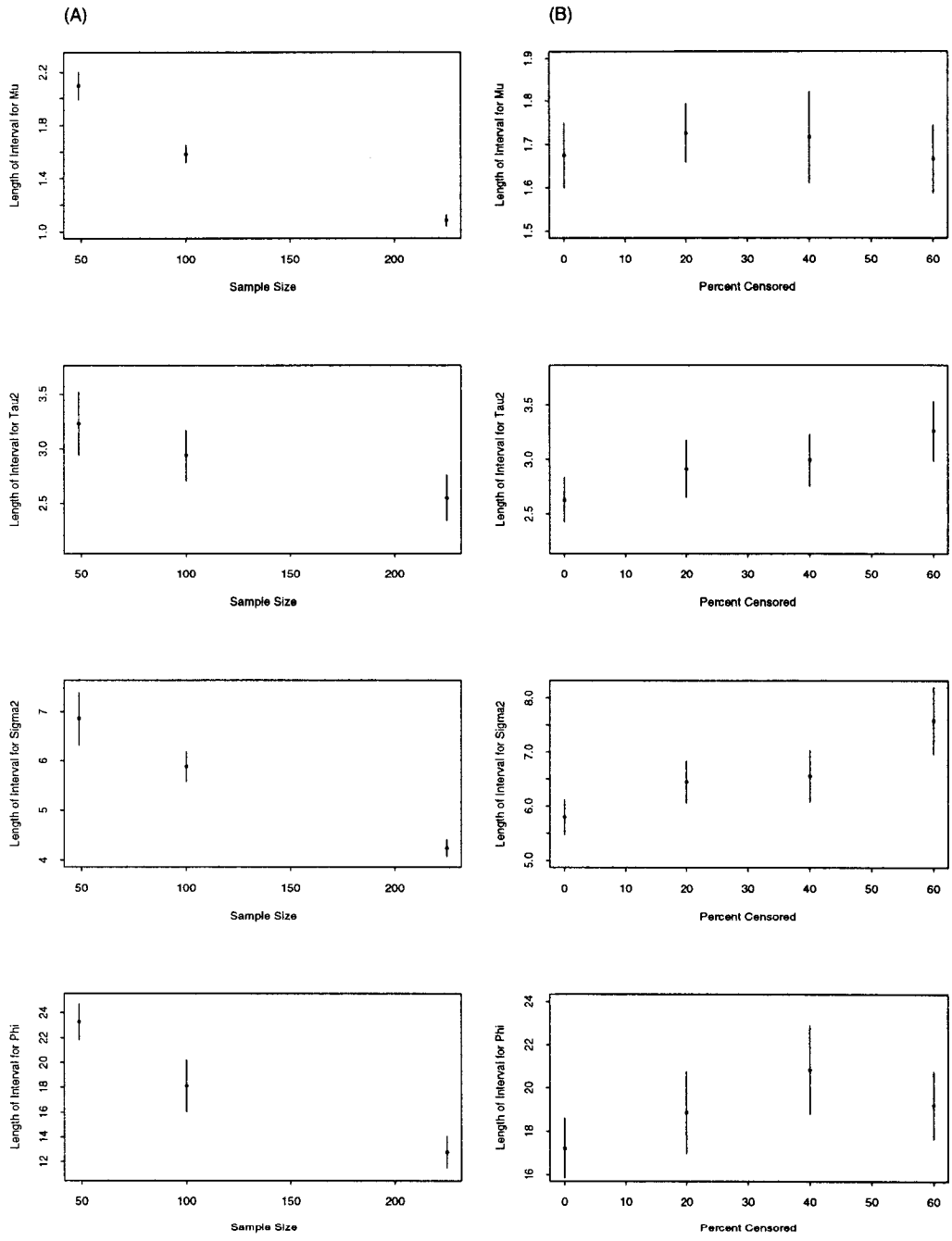

Figure 10 Plot of 95% confidence intervals for the average length of credible intervals for the various factor levels; (A) Sample Size (B) Percent Censored

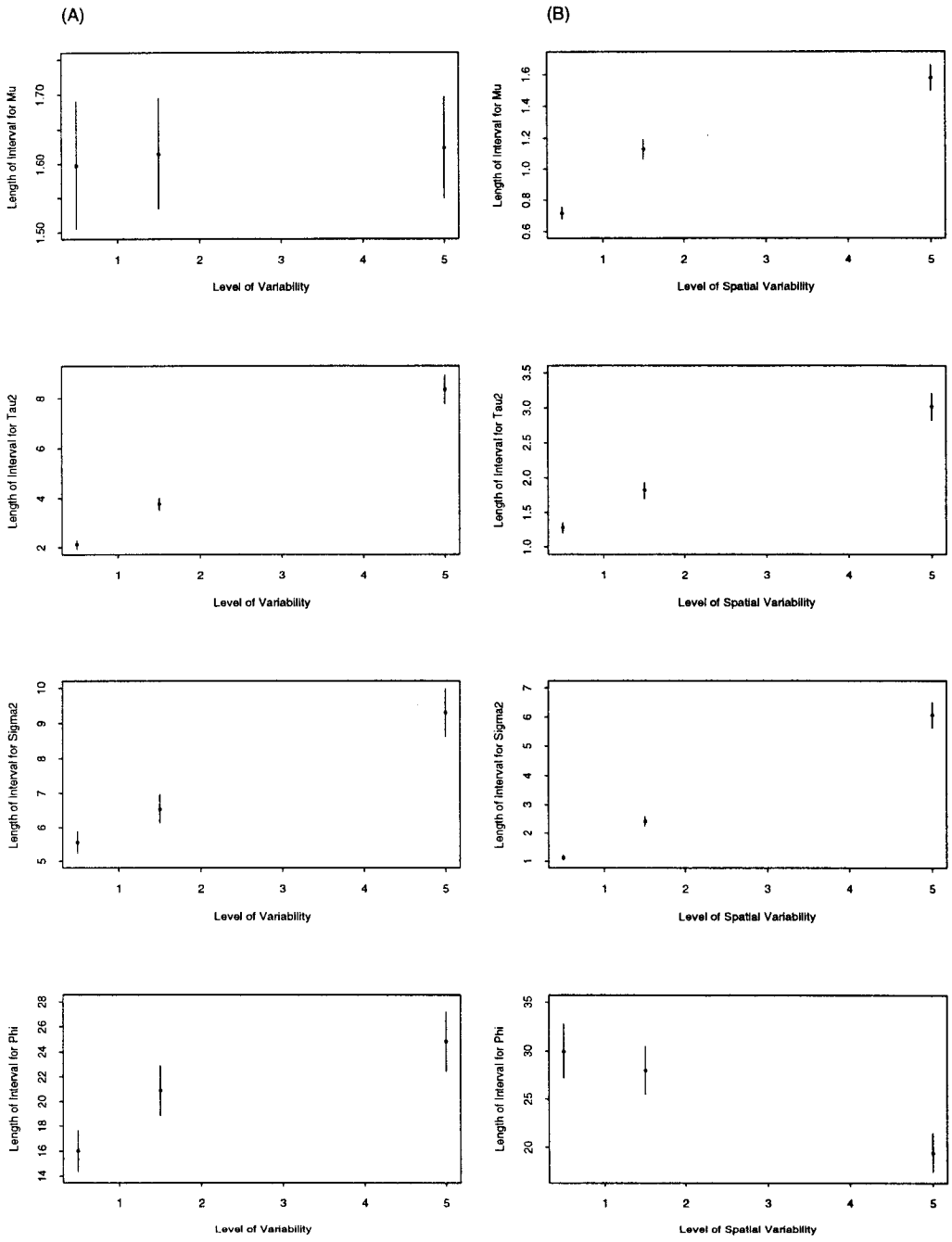

Figure 11 Plot of 95% confidence intervals for the average length of credible intervals for the various factor levels; (A) Variability (B) Spatial Variability

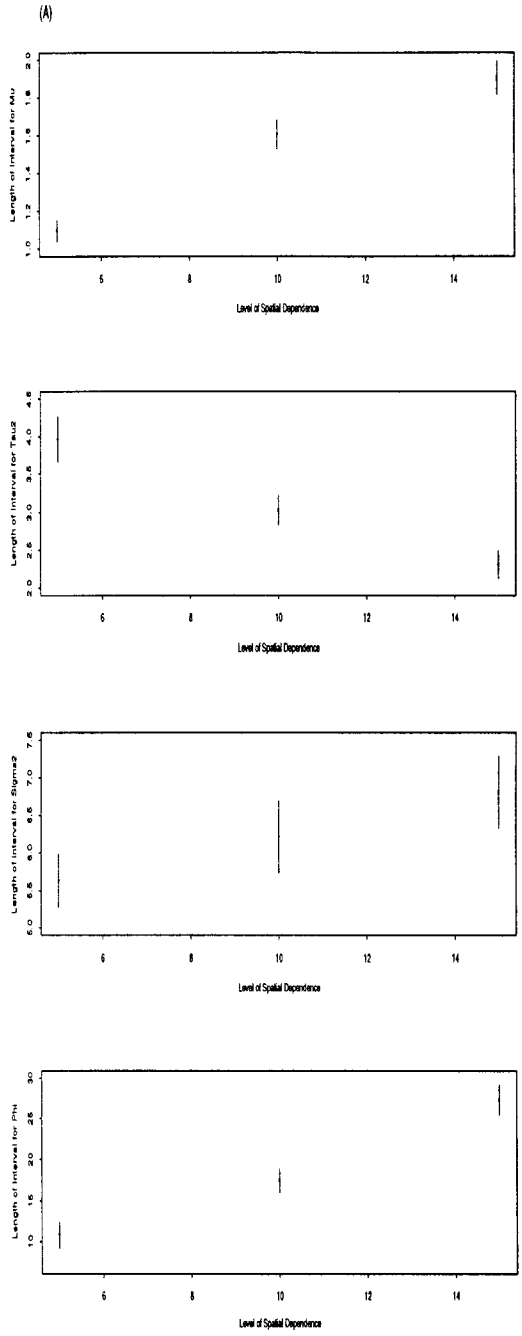

Figure 12 Plot of 95% confidence intervals for the average length of credible intervals for the various levels of spatial dependence

10, 11 and 12 are plots of 95% confidence intervals for the mean interval lengths. A few interesting things to note from these figures: (1) surprisingly, the percent censored has little impact on interval length, (2) intervals for the parameter ϕ reduce in length as the level of spatial variability (σ^2) increased due to the fact that σ^2 and ϕ are connected with ϕ being undefined as σ^2 tends to 0, (3) as the level of spatial dependence increased (ϕ), length of intervals increased since we are basically reducing our amount of total information (due to the dependence), with the exception of τ^2.

For an investigator, these results indicate that a sample size of 50 is too small to produce accurate estimates, in which 20% of the observations are censored. When using a grid sampling design with 10 units between adjacent locations, accuracy in estimation of ϕ is poor when the level of spatial dependence is low (i.e. $\phi = 5$). In addition to estimation, if an investigator wishes to have high percision in estimation, factors other than sample size come into play. If a high amount of variability or spatial dependence is present, a larger sample size will be needed to produce precise results. Surprisingly, the amount of censored data had only a mild impact on interval length or precision. Therefore, when designing a study, in addition to sample size, investigators need to take into account the amount of variability and spatial dependence thought to be present in any collected data.

4 Conditionally specified Gaussian spatial model

4.1 Model specification and data augmentation procedure

Let $\{Y(s_i) : i = 1, ..., n\}$ represent a set of random variables at location $\{s_i : i = 1, ..., n\}$. Then $Y(s_i)$, an observation at location s_i, is model as

$$Y(s_i)|Y(N_i) \sim \text{NOR}(\mu_i, \tau^2),$$

where $\mu_i = \alpha + \sum_{j=1}^{n} c_{ij}(y(s_j) - \alpha)$ and $c_{ij} = \eta(d_{ij})^{-1}$ if $s_j \in N_i$. This model results in the joint distribution for $Y(s_1), Y(s_2), ...Y(s_n)$ being

$$\boldsymbol{Y} \sim MVN(\boldsymbol{\alpha}, (I - C)^{-1}M), \tag{2}$$

where C contains the elements c_{ij}, with $C = \eta H$, where H is a known symmetric matrix containing inverse distances, and M is a diagonal matrix containing τ^2 (Besag, 1974; Kaiser and Cressie, 2000).

Proper prior distributions were placed on all parameters to insure a proper joint posterior distribution. The prior distributions used in the simulation study were

$$\alpha \sim \text{NOR}(\mu, \sigma^2)$$
$$\tau^2 \sim \text{INGAM}(\gamma, \beta)$$
$$\eta \sim \text{Transformed BETA}(\psi, \phi).$$

Based on the eigenvalues of H, the prior distribution for η will have support $(\frac{1}{h_1}, \frac{1}{h_n})$, where h_1 and h_n represent the smallest and largest eigenvalues of the matrix H, respectively. That is, $\eta = y(\frac{h_1 - h_n}{h_1 h_n}) + \frac{1}{h_1} \sim$ Transformed BETA(ψ, ϕ), where $y \sim$ BETA(ψ, ϕ) (Fridley and Dixon, 2003). Derivation of the transformed beta distribution can be found in Appendix II.

For the Bayesian conditionally specified Gaussian model represented in equation (2), the data augmentation algorithm using a Gibbs sampler as outlined by Fridley and Dixon (2003), is as follows. Derivation of full conditional distributions can be found in Appendix II.

1. Set starting values for $\alpha^{(0)}$, $\tau^{2(0)}$, and $\eta^{(0)}$. Set $m = 0$.

2. Set censored values equal to their level of detection, $\boldsymbol{Y}_c^{(0)} = \boldsymbol{LOD}$, where $\boldsymbol{Y}_c$ represents the vector of censored observations and $\boldsymbol{LOD}$ represents a vector containing the level of detection values.

3. Let $\boldsymbol{Y}^{T(m)} = (\boldsymbol{Y}_c^{(m)}, \boldsymbol{Y}_o)^T$, were $\boldsymbol{Y}_o$ represent the observed values.

4. Generate $\alpha^{(m+1)}$ from NOR$(\mu_\alpha, \sigma_\alpha^2)$ with $\sigma_\alpha^2 = \frac{1}{n^2}\mathbf{1}^T(\frac{1}{\sigma^{2(m)}}I + \frac{1}{\tau^{2(m)}}(I - C(\eta^{(m)})))^{-1}\mathbf{1}$ and $\mu_\alpha = \frac{1}{n}\mathbf{1}^T(\frac{1}{\sigma^{2(m)}}I + \frac{1}{\tau^{2(m)}}(I - C(\eta^{(m)})))^{-1}(\frac{\mu}{\sigma^{2(m)}}\mathbf{1} + \frac{1}{\tau^{2(m)}}(I - C(\eta^{(m)}))\boldsymbol{Y}^{(m)})$.

5. Generate $\tau^{2(m+1)}$ from INGAM$(\frac{n}{2} + \gamma, \frac{1}{2}(\boldsymbol{Y}^{(m)} - \boldsymbol{\alpha}^{(m+1)})^T(I - C(\eta^{(m)}))(\boldsymbol{Y}^{(m)} - \boldsymbol{\alpha}^{(m+1)}) + \beta)$.

6. Using Metropolis - Hastings step(s), simulate $\eta^{(m+1)}$ from $p(\eta|\boldsymbol{Y}^{(m)}, \tau^{2(m+1)}, \alpha^{(m+1)}) \propto [\prod_{i=1}^{n}(1-\eta h_i)]^{1/2} \exp\{\frac{\eta}{2\tau^{2(m+1)}}(\boldsymbol{Y}^{(m)} - \boldsymbol{\alpha}^{(m+1)})^T H(\boldsymbol{Y}^{(m)} - \boldsymbol{\alpha}^{(m+1)})\} (\eta - \frac{1}{h_1})^{\psi-1}[1 - (\eta - \frac{1}{h_1})(\frac{h_n h_1}{h_1 - h_n})]^{\phi-1}$.

7. Now have $\boldsymbol{\Theta}^{(m+1)} = (\alpha^{(m+1)}, \tau^{2(m+1)}, \eta^{(m+1)})$.

8. Using $\boldsymbol{\Theta}^{(m+1)}$ and $\boldsymbol{Y}^{T(m)}$, impute values for the censored values $\boldsymbol{Y}_c$ and get $\boldsymbol{Y}_c^{(m+1)}$. Let $\boldsymbol{Y}_c = (Y_{1c}, Y_{2c}, .., Y_{kc})$ represent k censored observations. Let $\mu_i^{(m+1)} = \alpha^{(m+1)} + \sum_{j=1}^{n} c_{ij}^{(m+1)}(Y(s_j) - \alpha^{(m+1)})$, and $c_{ij}^{(m+1)} = \eta^{(m+1)}(d_{ij})^{-h}$ for $s_j \in N_i$.

   (a) Generate $Y_{1c}^{(m+1)}$ from $Y_{1c}|Y_{2c}^{(m)}, .., Y_{kc}^{(m)}, \boldsymbol{Y}_o, \Theta^{(m+1)}$ which is a univariate normal distribution NOR$(\mu_1^{(m+1)}, \tau^{2(m+1)})$, truncated at LOD_1.

   (b) Generate $Y_{2c}^{(m+1)}$ from $Y_{2c}|Y_{1c}^{(m+1)}, Y_{3c}^{(m)}, .., Y_{kc}^{(m)}, \boldsymbol{Y}_o, \Theta^{(m+1)}$ which is a univariate normal distribution NOR$(\mu_2^{(m+1)}, \tau^{2(m+1)})$, truncated at LOD_2.

   . .

   (c) Generate $Y_{kc}^{(m+1)}$ from $Y_{kc}|Y_{1c}^{(m+1)}, .., Y_{(k-1)c}^{(m+1)}, \boldsymbol{Y_o}, \Theta^{(m+1)}$ which is a univariate normal distribution NOR$(\mu_k^{(m+1)}, \tau^{2(m+1)})$, truncated at LOD_k.

9. Set $m = m + 1$ and repeat the algorithm.

4.2 Simulation Study III

Simulation Study III was conducted using data augmentation for analysis of censored spatial data within a Bayesian conditionally specified Gaussian model. The aim of the study was to investigate the accuracy of the data augmentation procedure along with comparing the procedure to the method of replacing censored observations with $LOD/2$. The goal of Simulation Study III was to investigate the estimates and predictions produced by imputing values for the censored observations at each iteration of the Markov chain Monte Carlo, as outlined in Section 4.1. In addition to investigating the data augmentation procedure in general, analyses replacing the censored observations with half their level of detection ($LOD/2$) were also completed. Hence, the data augmentation procedure and the common method of replacing the censored values with $LOD/2$ are compared.

4.2.1 Estimation

To assess the parameter estimates produced by using data augmentation, 1000 simulated datasets of size 100 (10x10 regular grid) were simulated. For the conditionally specified Gaussian model to be valid, η is restricted to be between -6.258 and 0.376. The 1000 datasets were simulated using $\alpha = 0$, $\tau^2 = 2$, and η=0.25 with 20% of the observations censored. To finish the specification of the Bayesian model, the proper diffuse priors used were

$$\alpha \sim \text{NOR}(0, 50),$$
$$\tau^2 \sim \text{INGAM}(2.1, 2.2),$$
$$\eta \sim \text{Transformed BETA}(1.0, 1.0),$$

with -6.258 $< \eta <$ 0.376. The Gibbs sampler was run for 3,000 iterations with four Metropolis-Hastings steps for the simulation of η. The candidate generating distribution used in the Metropolis-Hastings steps was the transformed BETA$(2X, 2(1 - X))$,

Table 8 Summary of estimates for the 1000 simulated datasets for the data augmentation and LOD/2 methods.

DA	Parameter	Min	Q1	Median	Mean	Q3	Max
	α	-0.854	-0.174	-0.012	-0.009	0.140	0.783
	τ^2	1.095	1.671	1.901	1.898	2.103	3.023
	η	-3.858	-0.818	-0.426	-0.605	-0.201	0.102
LOD/2	Parameter	Min	Q1	Median	Mean	Q3	Max
	α	-0.564	0.099	0.245	0.250	0.386	0.984
	τ^2	0.678	0.982	1.106	1.117	1.242	1.799
	η	-2.864	-0.765	-0.396	-0.558	-0.195	0.055

where X represents the current value of η. Estimation was based on the last 2,000 iterations. For comparison purposes, censored observations were handled using both the data augmentation method (DA) and the method which replaces the censored observations with half their level of detection (LOD/2).

Comparison of the 1000 estimates produced via DA and the 1000 estimates produced using the LOD/2 method are illustrated in Figure 13. The scatterplots show the DA method consistently producing smaller estimates of α and larger estimates of τ^2 as compared to the LOD/2 method. Hence, for this particular scenario, replacing the censored values with $LOD/2$ resulted in the variability being under-estimated and the mean being over-estimated. To summarize the scatterplots, estimates of $P(\hat{\alpha}_{DA} < \hat{\alpha}_{LOD/2})$, $P(\hat{\tau}^2_{DA} < \hat{\tau}^2_{LOD/2})$ and $P(\hat{\eta}_{DA} < \hat{\eta}_{LOD/2})$ were found to be 1, 0, and 0.543, respectively.

Table 8 and Figure 14 show the data augmentation procedure tended to produce estimates closer to the truth as compared to the LOD/2 method. Using the LOD/2 method, the average estimates for α and τ^2 were 0.250 and 1.117, respectively. Conversely, data augmentation produced estimates of α and τ^2 closer to the true values of 0 and 2. In addition to the difference in estimation accuracy, the LOD/2 method also under-estimated the variability in estimating τ^2. This can also be seen in Table 9, which displays summary information of the 1000 95% equal-tailed credible intervals. The aver-

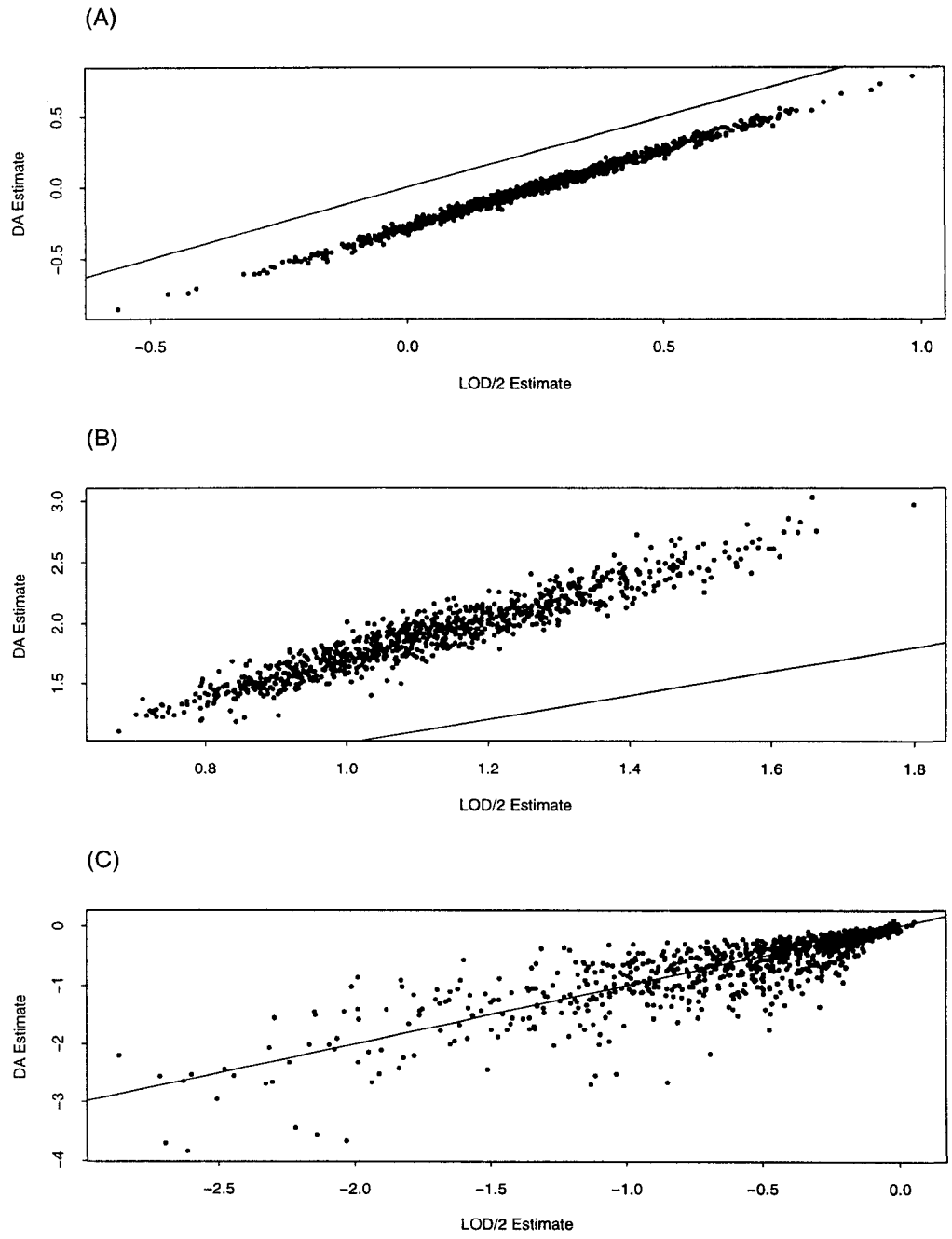

Figure 13 Scatterplot of estimates found via DA and LOD/2; (A) α (B) τ^2 (C) η

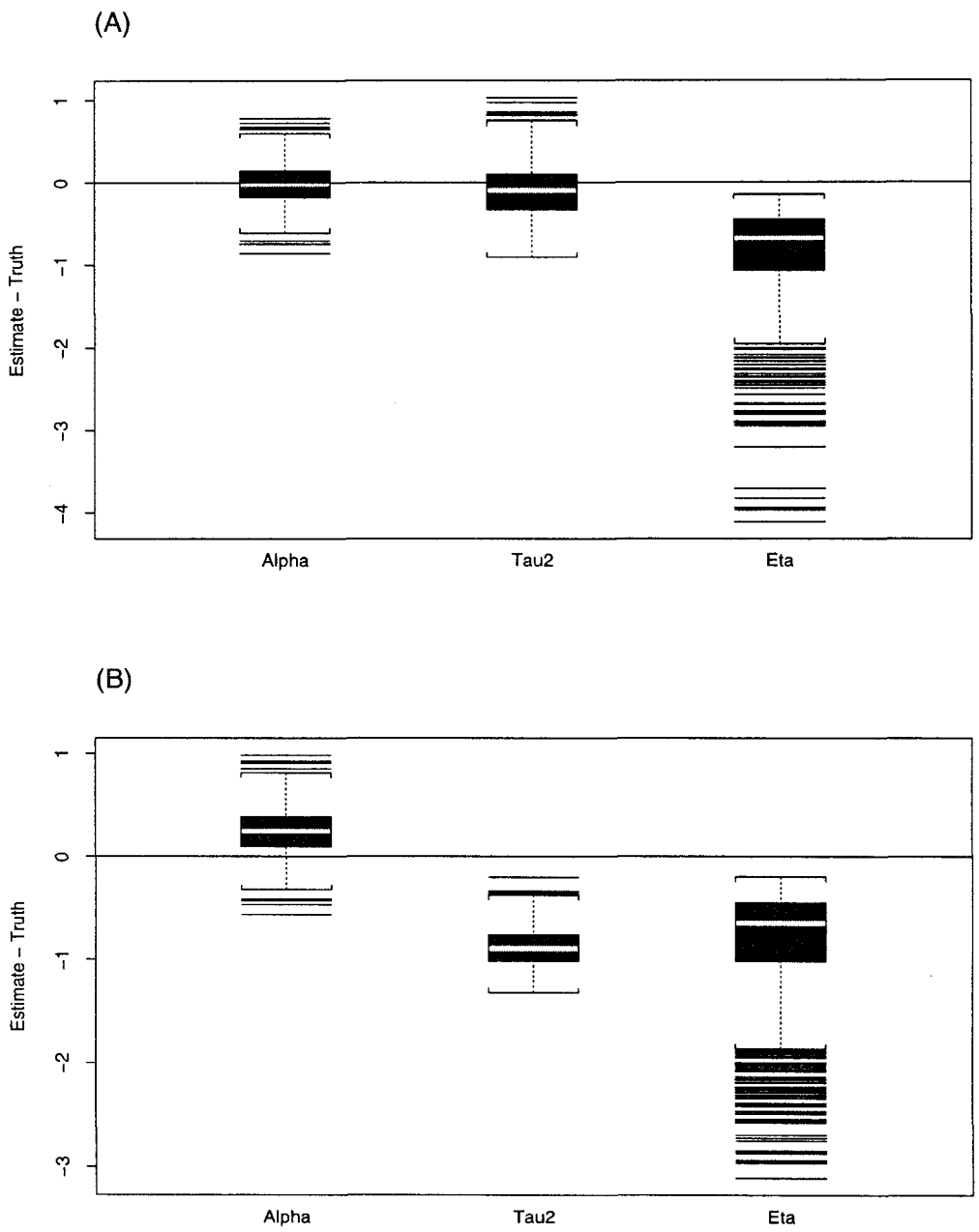

Figure 14 Boxplots of $\hat{\alpha}$ - α, $\hat{\tau}^2$ - τ^2, and $\hat{\eta}$ - η using (A) data augmentation and (B) LOD/2 method

Table 9 Summary of lengths for 95% credible intervals for the 1000 simulated datasets using data augmentation and LOD/2 methods.

DA	Parameter	Min	Q1	Median	Mean	Q3	Max
	α	0.192	0.383	0.445	0.448	0.513	0.808
	τ^2	0.724	1.114	1.257	1.263	1.402	2.037
	η	0.563	1.565	2.125	2.199	2.758	4.257
LOD/2	Parameter	Min	Q1	Median	Mean	Q3	Max
	α	0.141	0.284	0.336	0.335	0.386	0.597
	τ^2	0.378	0.563	0.631	0.638	0.708	1.021
	η	0.608	1.495	2.011	2.073	2.577	3.906

age interval length for τ^2 using the LOD/2 method was 0.638, while data augmentation on average produced intervals of length 1.263. While there seems to be differences with regards to the estimation of α and τ^2, there does not seem to be much difference in the estimation of η between the two methods. This may be due to the fact that a sample size of 100 with 20% of the observations censored is too small of a sample to estimate η with any accuracy or precision. The effect of sample size when estimating η was examined in section 4.3, were N is shown to have an impact on the estimation of η.

Quantification of the estimation procedures in terms of the mean square error (MSE) was also completed. Estimates of bias and variance were computed by finding the sample mean and variance of the 1000 estimates. Estimates of MSE, bias and variance for both the DA method and the LOD/2 method are displayed in Table 10. Once again, we see that DA produced smaller estimates of MSE for α and τ^2, with MSEs of 0.0587 and 0.1117 as compared to the estimates 0.1111 and 0.8147 produced using the LOD/2 method.

4.2.2 Prediction

Along with parameter estimation, predictions between the DA and the LOD/2 methods were compared. To compare the two methods, 50 simulated datasets were con-

Table 10 Estimates of bias, variance and mean square error for estimation of α, τ^2 and η using data augmentation and LOD/2 methods.

DA	Parameter	Bias	Variance	MSE
	α	-0.0092	0.0586	0.0587
	τ^2	-0.1019	0.1013	0.1117
	η	-0.8552	0.3288	1.0602
LOD/2	Parameter	Bias	Variance	MSE
	α	0.2498	0.0487	0.1111
	τ^2	-0.8828	0.0354	0.8147
	η	-0.8080	0.2597	0.91245

structed on a 15x15 regular lattice with 5 units distance between adjacent locations. Each dataset of sample size 225 was generated with 20% of the observations being coded as censored using parameter values of α=0, $\tau^2 = 2$ and η =0.05. The choice of η was due to the fact that η was restricted to the range -3.11 to 0.12. The simulated data were then split-up into two parts; a dataset for analysis and estimation of parameters (observed locations) and a dataset for prediction (predicted locations). The prediction dataset will be used as the "truth" to be compared to predictions produced by the DA and the LOD/2 methods. Figure 3 displays the locations used in the estimation of parameters and the locations used for prediction. The prediction stage of the analysis was completed by approximating the Bayesian posterior predictive distribution as described in Section 3.2.2 (Haining and Griffith, 1989; Carlin and Louis, 1996; de Oliveira and Ecker, 2002; Fridley and Dixon, 2003).

The parameter estimation and prediction for the 50 simulated datasets were conducted using the same proper diffuse priors outlined in Section 4.2.1. The Gibbs sampler was ran for 3000 iteration with 4 Metropolis-Hastings steps at each iteration for the estimation of η. The posterior predictive distribution was then approximated using the last 2000 iterations. To eliminate any possible correlation between iterations, every 5th iteration was used for prediction. So, for each prediction location, the posterior predictive

Table 11 Summary of mean prediction error (MPE) and mean squared prediction error (MSPE) for the 50 simulated datasets using data augmentation and LOD/2 method for the handling of censored data

DA	Measure	Min	Q1	Median	Mean	Q3	Max
	MPE	-0.409	-0.109	0.003	0.036	0.210	0.551
	MSPE	1.661	1.859	2.122	2.140	2.318	2.888
LOD/2	Measure	Min	Q1	Median	Mean	Q3	Max
	MPE	-0.082	0.179	0.328	0.347	0.503	0.894
	MSPE	1.680	2.025	2.195	2.235	2.390	3.146
LOD/2 - DA	Measure	Min	Q1	Median	Mean	Q3	Max
	MSPE	-0.124	-0.017	0.104	0.095	0.177	0.690

distribution was approximated via 400 simulated predictions. From these approximated posterior predictive distributions, point estimates ($\hat{y}_i$) were computed as the median of the posterior predictive distribution for each of the i locations. Following the simulation of the posterior predictive distributions, two measures were computed for each of the 50 simulated datasets; the mean prediction error (MPE), $\sum_{i=1}^{n}(\hat{y}_i - y_i)/n$, and the mean squared prediction error (MSPE), $\sum_{i=1}^{n}(\hat{y}_i - y_i)^2/n$. This process was complete for the two different methods of handling censored spatial data, that of the DA and the LOD/2 methods. Results are displayed in Table 11 and Figures 15 through 17.

Out of the 50 simulated datasets, the data augmentation method produced lower MSPE 70% of the time. Table 11 shows a similar finding as does Table 4 in Section 3.2.2; the LOD/2 method over-estimates when it comes to prediction. Lastly, Figure 17 displays the case when data augmentation was vastly superior to LOD/2 and likewise, the case in which LOD/2 outperformed data augmentation. Figure 17 (A) is the case when MSPE was 2.456 for data augmentation while the LOD/2 method produced a value of 3.146. Figure 17 (B) shows the converse case with the LOD/2 and data augmentation methods producing MSPEs of 1.810 and 1.935, respectively.

As with Simulation Study I, Simulation Study III looked at general properties of the

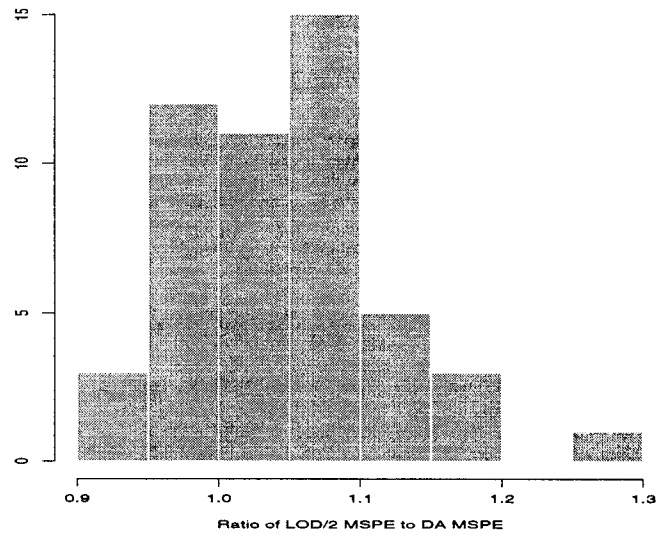

Figure 15 Histogram of the ratio of LOD/2 MSPE to DA MSPE

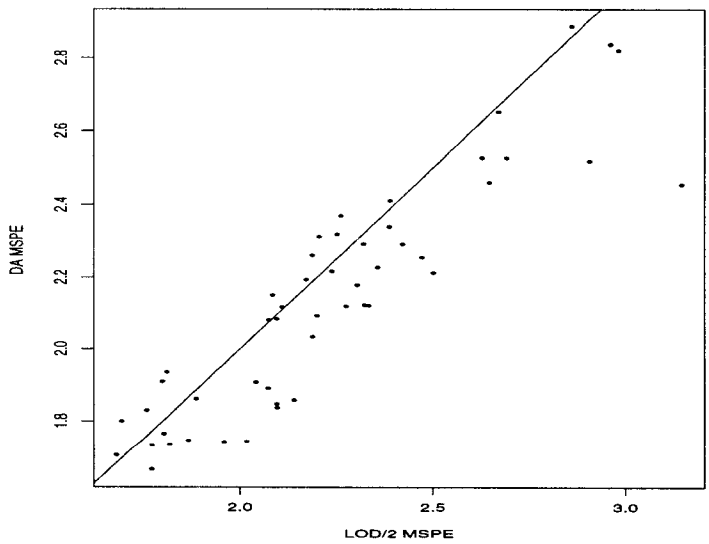

Figure 16 Scatterplot of DA MSPE and LOD/2 MSPE

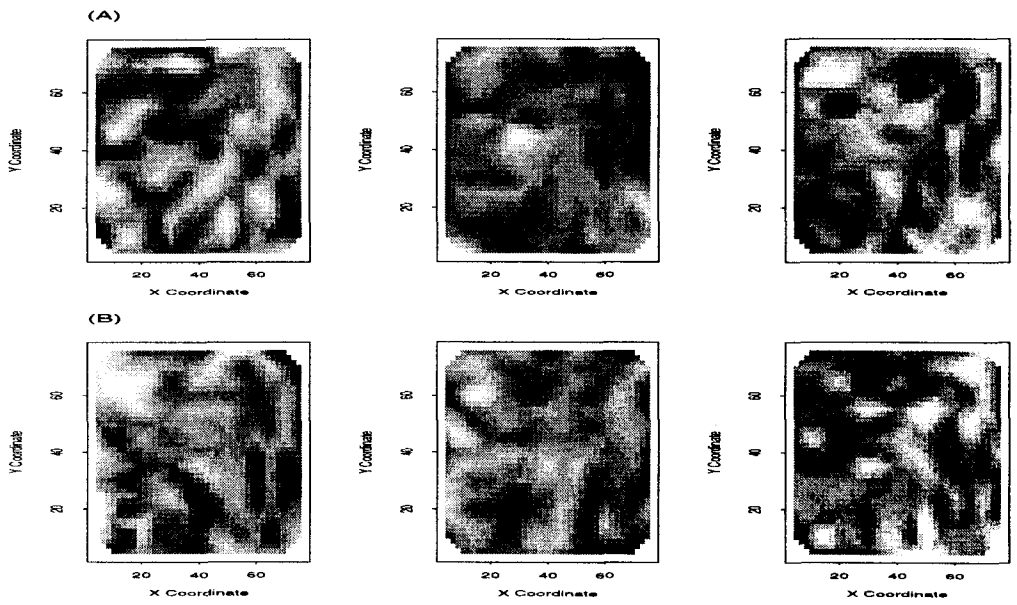

Figure 17 Map of the truth, predicted surface using DA and predicted surface using LOD/2 method; (A) Simulated dataset that resulted in largest superior performance with DA (B) Simulated dataset that resulted in superior performance via the LOD/2 method

data augmentation procedure in the context of a conditionally specified Gaussian model for the analysis of one particular situation. The study showed the data augmentation procedure to be superior to the LOD/2 method in terms of estimation and prediction. Data augmentation produced smaller estimate of MSE for the overall mean (α) and the level of variability (τ^2) as compared to the LOD/2 method. In terms of prediction, data augmentation also produced smaller MSPE 70% of the time. The question becomes, "Do these results hold in general?". This question is addressed in Simulation Study IV, where investigation into which factor(s) may impact the data augmentation procedure was conducted.

4.3 Simulation Study IV

The fourth simulation study was focused on answering the question "What factor(s), if any, impact the data augmentation procedure for the analysis of censored spatial data in the context of a Bayesian conditionally specified Gaussian model?". To answer this question, Simulation Study IV was designed to look at four possible factors: sample size, percent censored, variability and spatial dependence. The factor levels used in the study are displayed in Table 12. The 13 different scenarios were produced by changing the necessary simulation parameters. For each scenarios 50 simulated datasets were generated using the conditionally specified Gaussian model outlined in Section 4.1. The data were simulated with the default parameter values of $\alpha = 0$, $\tau^2 = 2$, η=0.25 with 20% censored data on a 10x10 regular lattice (10 units between nearest neighbors), yielding a sample size of 100. For example, to investigate the factor of variability, simulation parameters would be α=0, η=0.25, N=100 and % censored = 20% with values of τ^2 fixed to be 0.5, 1.5 or 5.0.

There was an added complication for the factor sample size. The factor sample size took levels of 7x7 (N=49), 10x10 (N=100) and 15x15 (N=225). Due to model restrictions, η had an upper bound of 0.570 for N=49, 0.376 for N=100 and 0.239 for

Table 12 Factor levels used for Simulation Study IV

Factor	Level 1	Level 2	Level 3	Level 4
Sample Size (N)	7x7	10x10	15x15	—
% Censored	0	20	40	60
Variability (τ^2)	0.5	1.5	5.0	—
Spatial Dependence (η)	0.00	0.15	0.30	—

N=225. Thus, for investigation of the factor sample size, η was fixed to be 0.15 instead of 0.25. For instance, to simulate data to look at the factor sample size, data were generated using $\alpha = 0$, $\tau^2 = 2$, η=0.15 and % censored = 20% with sample sizes of 49, 100 and 225.

To apply the data augmentation procedure, portions of the data were coded as falling below the level of detection (LOD). The level of detection was determined by the level of censoring. If 40% of the observations were to be censored, the LOD value was determined to be the value in which 40% of the observations fell below. This value would then be the LOD and any observation falling below the LOD would be recorded as "$< LOD$".

To complete the model, proper diffuse priors were placed on all parameters. Let τ^{2*} represent the true value of τ^2 used to simulate the data. The priors used for Simulation Study IV, based on the current level of the factor being investigated, were

$$\alpha \sim \text{NOR}(0, 50),$$
$$\tau^2 \sim \text{INGAM}(2.1, 1.1(\tau^{2*})),$$
$$\eta \sim \text{Transformed BETA}(1.0, 1.0),$$

for $\eta \in (1/h_1, 1/h_n)$, where $1/h_1$ and $1/h_n$ are the smallest and largest eigenvalues of H (C=ηH), respectively. The prior $\eta \sim$ transformed BETA(1.0, 1.0) results in an uniform prior over the range $(\frac{1}{h_1}, \frac{1}{h_n})$. For instance, the priors used to analyze data created to address the lowest level of the factor variability (τ^2) would be $\alpha \sim \text{NOR}(0, 50)$, $\tau^2 \sim$ INGAM(2.1, 1.1(0.5)) and $\eta \sim$ Transformed BETA(1.0, 1.0).

Table 13 Average parameter estimates for the various factor levels

Factor	Level 1	Level 2	Level 3	Level 4
N	49	100	225	—
α	-0.06	0.00	-0.05	—
τ^2	1.84	1.93	2.01	—
η	-0.86	-0.69	-0.16	—
% Censored	0	20	40	60
α	0.02	-0.06	0.01	0.00
τ^2	1.97	1.98	1.71	1.84
η	-0.49	-0.53	-0.71	-0.74
Variability, τ^2	0.5	1.5	5.0	—
α	0.02	0.00	-0.09	—
τ^2	0.48	1.46	4.96	—
η	-0.38	-0.45	-0.45	—
Spatial Dependence, η	0	0.15	0.30	—
α	0.00	-0.03	0.00	—
τ^2	1.92	1.87	2.00	—
η	-0.62	-0.60	-0.41	—

The Gibbs sampler was run for 3,000 iterations. For estimation purposes, 2500 iterations were used, ignoring the first 500 iterations for burn-in. For the simulation of η, 4 Metropolis-Hastings steps were implemented at each iteration of the Gibbs sampler, using a transformed BETA$(2X, 2(1-X))$ as the candidate generating distribution. Results of the simulation study are displayed in Tables 13 and 14 and Figures 18 to 21.

Estimates of bias were computed for each scenario as a measure of accuracy. Table 13 and Figures 18 and 19 show the average estimates and confidence intervals for the estimated bias for the 13 scenarios. The factors of sample size and percent censored impacted the estimation of ϕ and τ^2, but not α. The amount of variability and spatial dependence were estimated more accurately at the highest level of sample size while the higher levels of censoring resulted in larger bias in the estimation of τ^2 and η. As for the factor of variability, as the level of variability increased so did the confidence interval for the bias in estimating the parameters α and τ^2. The level of spatial dependence (η)

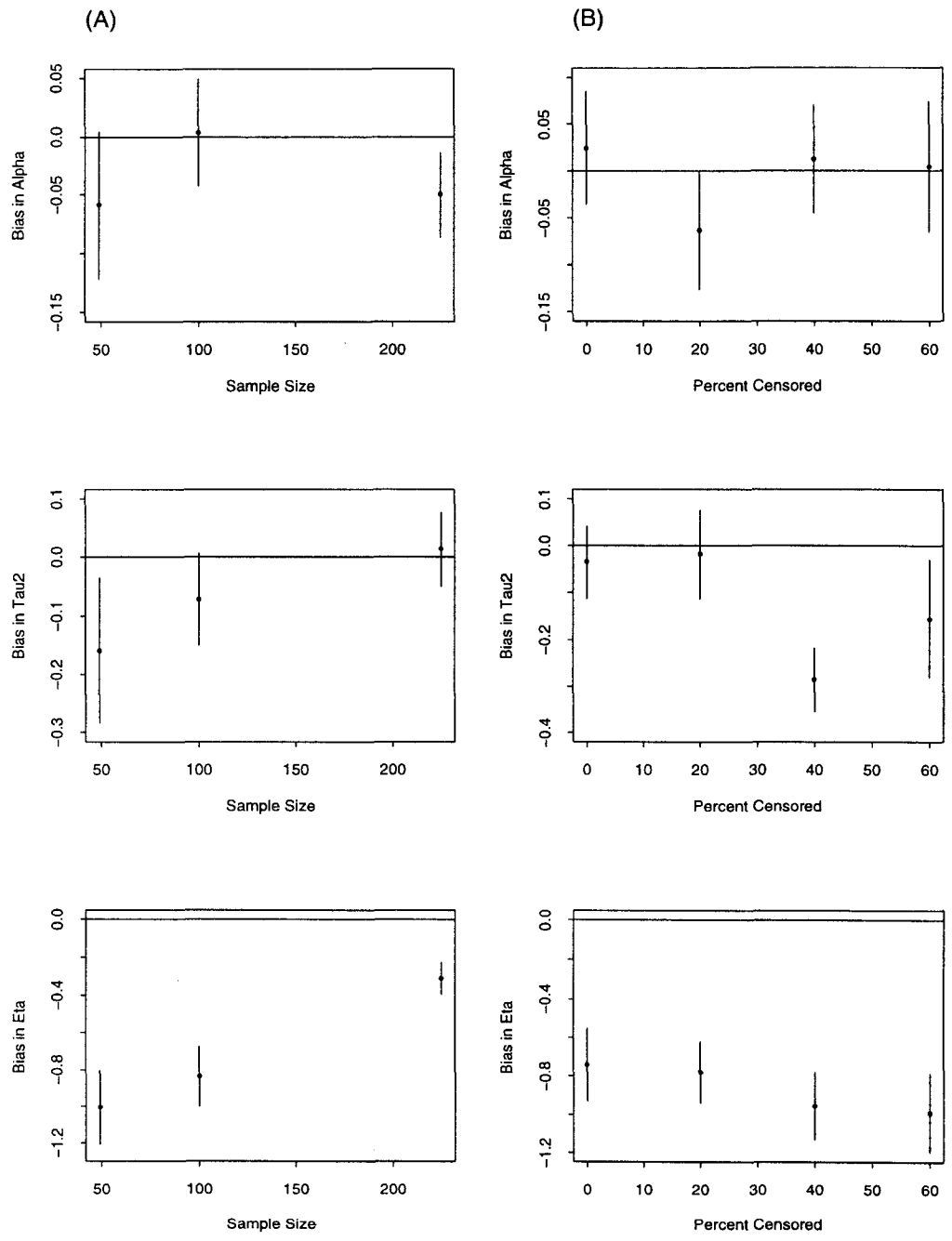

Figure 18 Plot of 95% confidence intervals for the average estimate for the various factor levels; (A) Sample Size (B) Percent Censored

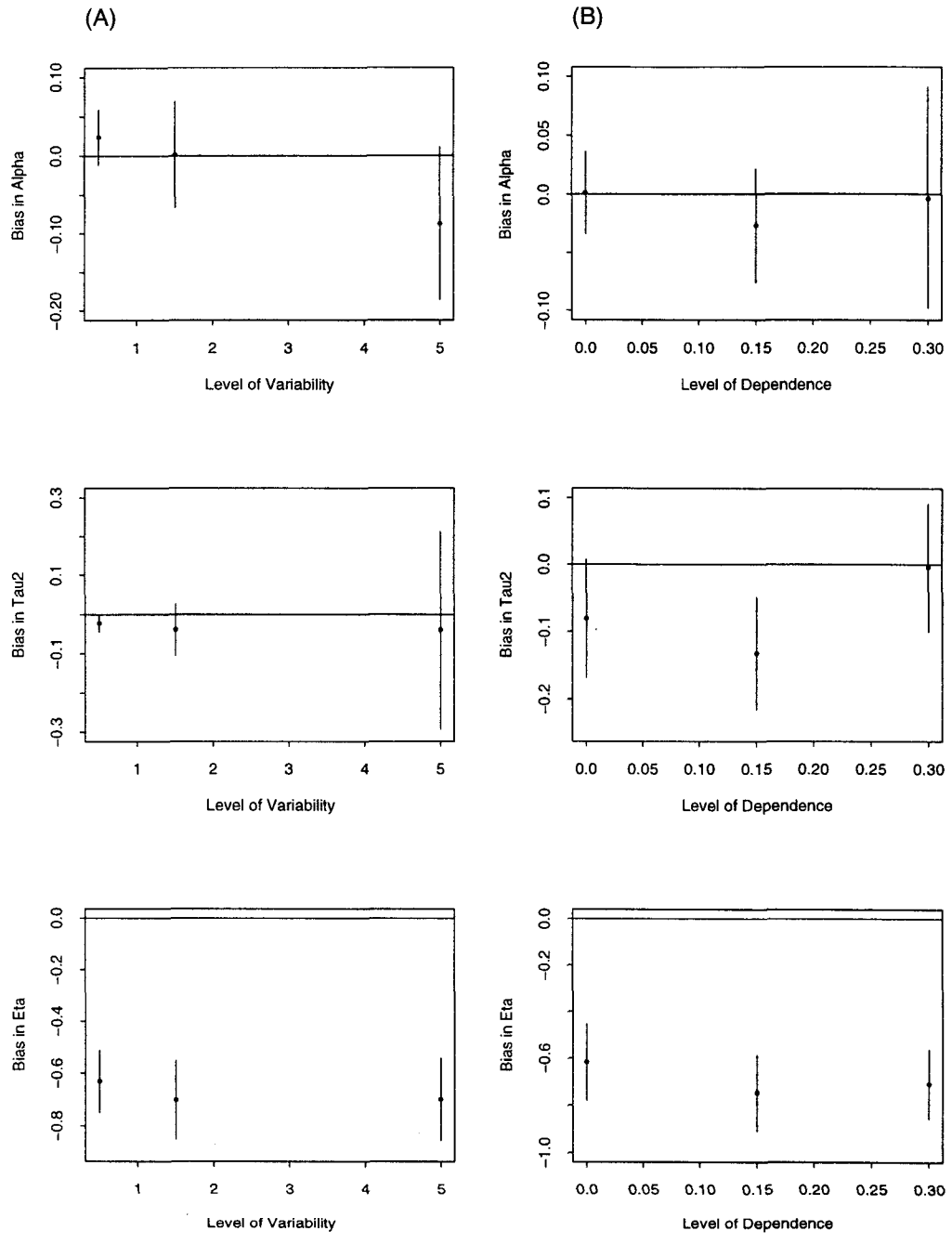

Figure 19 Plot of 95% confidence intervals for the average estimate for the various factor levels; (A) Variability (B) Spatial Dependence

Table 14 Average length of credible intervals for the various factor levels

Factor	Level 1	Level 2	Level 3	Level 4
N	49	100	225	—
α	0.83	0.56	0.48	—
τ^2	1.77	1.31	0.89	—
η	3.68	2.59	1.17	—
% Censored	0	20	40	60
α	0.66	0.60	0.60	0.80
τ^2	1.13	1.33	1.40	1.92
η	2.14	2.41	2.86	3.14
Variability, τ^2	0.5	1.5	5.0	—
α	0.40	0.60	0.98	—
τ^2	0.32	0.98	3.36	—
η	2.21	2.25	2.26	—
Spatial Dependence, η	0	0.15	0.30	—
α	0.62	0.58	0.74	—
τ^2	1.30	1.26	1.36	—
η	2.56	2.48	2.19	—

seemed to have little or no impact on estimation accuracy.

Along with interest in estimation bias or accuracy, investigation into which factors may possibly effect estimation precision was completed. The measure of precision used was length of equal-tailed 95% credible intervals. Results for the various factor levels are displayed in Table 14 and Figures 20 and 21. The length of intervals for all parameters are greatly impacted by the sample size and the amount of censoring. As sample size increased, the average width of intervals decreased, while interval widths increased as the amount of censoring increased. In addition to the factors of sample size and percent censoring, the level of variability also impacted precision in estimating α and τ^2. As the amount of variability increased, the width of 95% credible intervals for α and τ^2 increased. Lastly, the level of spatial dependence seems to have little impact on the precision in estimating α, τ^2 and η.

The practical implications of these results for investigators are the following: (1)

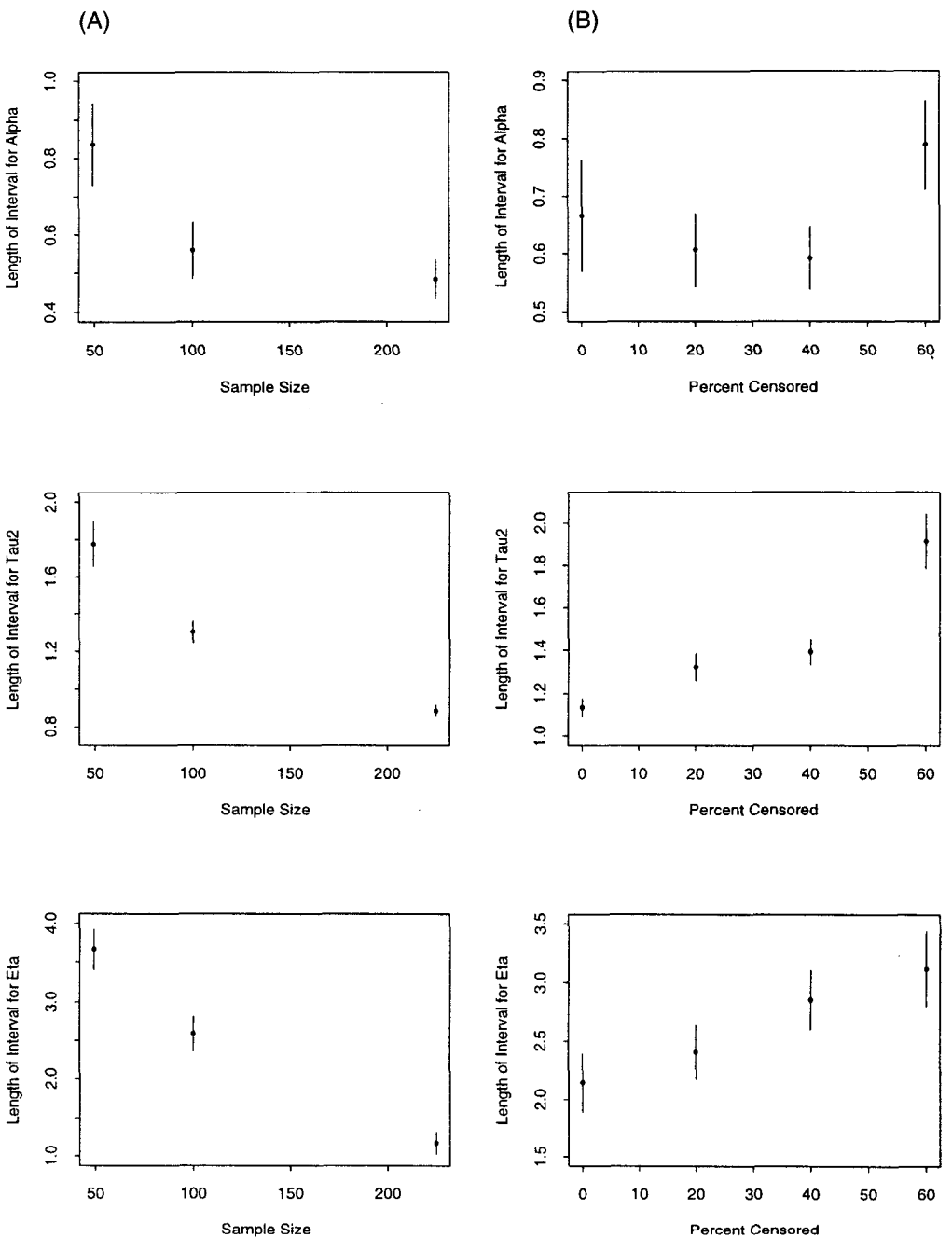

Figure 20 Plot of 95% confidence intervals for the average length of credible intervals for the various factor levels; (A) Sample Size (B) Percent Censored

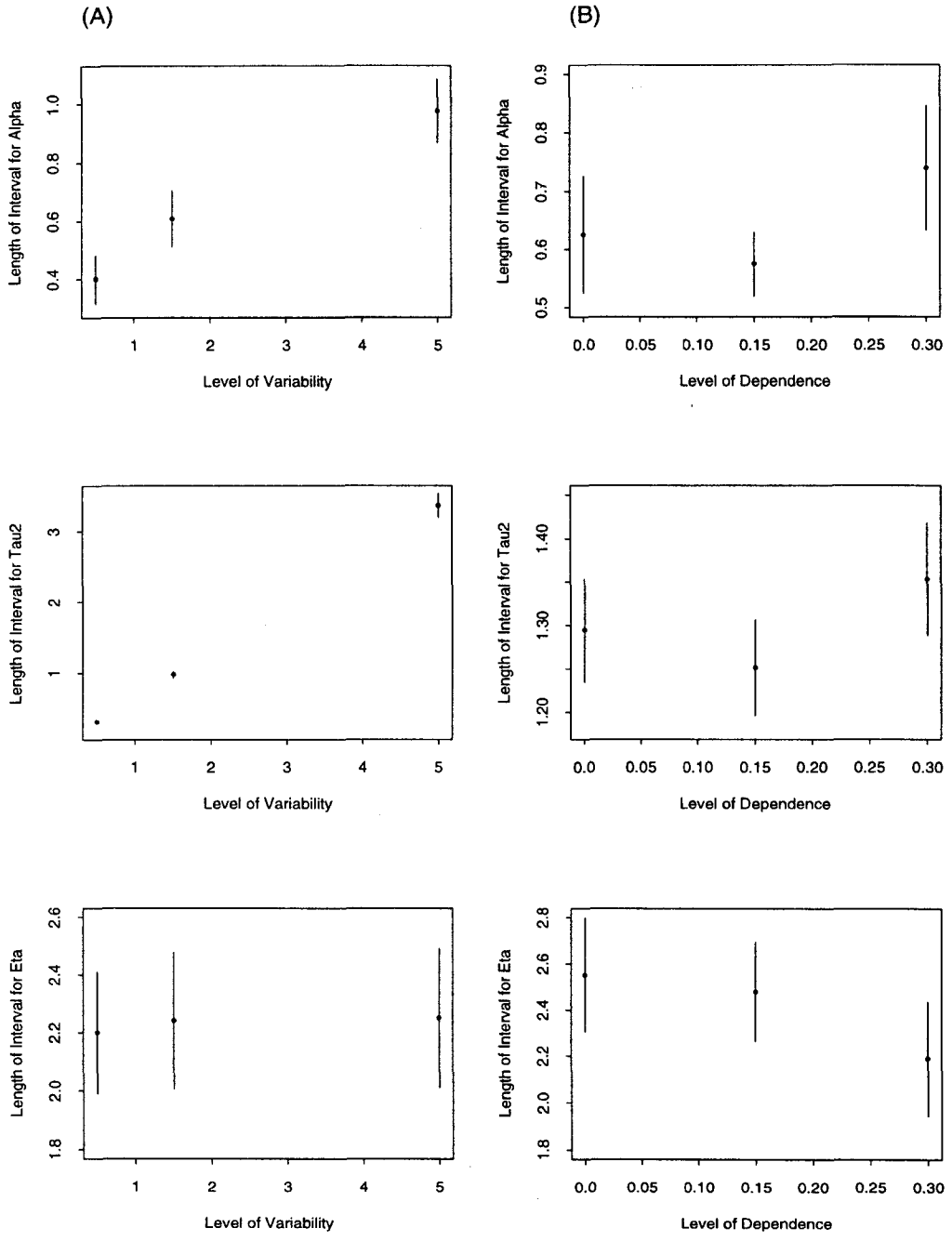

Figure 21 Plot of 95% confidence intervals for the average length of credible intervals for the various factor levels; (A) Variability (B) Spatial Dependence

avoid using small sample sizes ($N < 100$), since sample size impacts both accuracy and precision in estimation, (2) the amount of censored data impacts the precision in estimation; increase the sample size if the amount of censoring is expected to be large, (3) the amount of variability present in the data impacts the precision in estimation; increase the sample size to increase the precision in estimation, if cost permits. Simulation studies can be used to aid in the choice of sample size based on different study scenarios.

5 Conclusions

This paper presented simulation results assessing the accuracy and precision when using data augmentation for the analysis of censored spatial data. The simulation studies were conducted using both a Bayesian spatial or geostatistical model and a Bayesian conditionally specified Gaussian or CAR model. These simulation studies or others can be used to aid experimenters in study design. For example, simulation studies can be used to help determine the sampling design or layout (e.g. grid verses non-grid).

The first part of the paper (Simulation Study I and II) addressed the application of data augmentation procedure within a Bayesian spatial model. Simulation Study I focused properties of parameter estimates and predictions produced when using the data augmentation procedure for the handling of censored spatial data. In addition, comparison of the data augmentation method to the LOD/2 method was also completed. The second simulation study focused on answering the question "What factor(s), if any, impact the performance of data augmentation for the analysis of censored spatial data?". In doing so, the factors of sample size, percent censored, level of variability, level of spatial variability and spatial dependence were investigated at various levels.

The results showed the data augmentation procedure to out-perform the LOD/2 method in terms of both parameter estimation and prediction. When compared to the LOD/2 method, the data augmentation method produced smaller mean square predic-

tion errors 98% of the time. In addition to demonstrating the accuracy of parameter estimation and prediction when using the data augmentation procedure, Simulation Study II showed the factor of percent censored had little effect on the accuracy of estimation. The other factors (sample size, variability, spatial variability and spatial dependence) all impacted the amount of bias in estimation for one or more parameters. In terms of precision in estimation, all factors seem to impact the level of precision in one way or another.

The second half of the paper focused on the conditionally specified Gaussian or CAR model and the use of data augmentation. Simulation Study III once again focused on estimation and prediction with comparison of the data augmentation method to the LOD/2 method. Likewise, Simulation Study IV focused on factors that may impact the performance of the data augmentation procedure. For the conditionally specified Gaussian model, the factors used in the simulation study were sample size, percent censored, variability and spatial dependence.

For the conditionally specified Gaussian model, Simulation Study III showed data augmentation to be superior to the common method of replacing the censored values with $LOD/2$. The largest difference between the two methods was in regards to the estimation of α and τ^2. The LOD/2 method consistently under-estimated τ^2 and over-estimated α. Both methods performed about the same in terms of estimating η. In terms of prediction, data augmentation produced smaller mean squared prediction errors 70% of the time. Simulation Study IV found the amount of bias in estimation to be influenced by the sample size, percent censored, and to a small extent the factors of variability and spatial dependence. The factors sample size, percent censored and level of variability all seemed to influence the precision in estimation (i.e. length of credible intervals).

Overall, the simulation studies demonstrated the data augmentation method to be superior to the common method of replacing the censored observations with half the level of detection, particular in the estimation of parameters representing variability.

Since the parameter estimates go directly into prediction, more accurate estimation lead to better predictions and smaller mean square prediction errors. Not accounting for the censored observations in a adequate manner leads to inaccurate predictions that may have severe health, political and cost ramifications. A good study design can also help produce data of higher quality and consequently better estimation and predictions. Simulation Studies II and IV presented factors that seem to impact estimation accuracy and precision. Investigators can use these results or results from other simulation studies to aid them in the construction of study designs. Practical implications of these results for study design are:

1. If possible under time and cost constraints, avoid using small sample sizes, since sample size impacts both accuracy and precision in estimation.

2. If a high amount of variability or spatial dependence is present, a larger sample size will be needed to produced precise results.

3. Depending on the amount of censoring, one may wish to increase the sample size to produce more reliable and precise estimates.

4. Make sure the sample design is able to estimate the spatial dependence parameter accurately. As seen in Simulation Study I, by using a grid sampling design with 10 units between adjacent locations, accuracy in estimating the spatial dependence parameter was poor when the level of spatial dependence was low.

Lastly, with data augmentation completed conditional on the model specified, studies looking at parameter estimation and prediction using an incorrect model for the augmentation and subsequent analysis are needed. Further work is needed to investigate the robustness to model misspecification and diagnostics in the context of data augmentation for censored spatial data.

Appendix I

This appendix presents the derivation of the full conditional distributions required for the Gibbs sampler involving a data augmentation step for the Bayesian spatial model.

Full conditional distribution for σ^2:

The full conditional distribution for σ^2 is

$$p(\sigma^2|\tau^2,\phi,\mu,\boldsymbol{W},\boldsymbol{X}) \propto p(\boldsymbol{W}|\sigma^2,\phi)p(\sigma^2)$$
$$\propto |\sigma^2 V^*(\phi)|^{-1/2}\exp\{\tfrac{-1}{2}\boldsymbol{W}^T(\sigma^2 V^*(\phi))^{-1}\boldsymbol{W}\}(\sigma^2)^{-1(\alpha+1)}\exp\{-\beta/\sigma^2\},$$

where $V^*(\phi) = \exp\{-d/\phi\}$. Therefore, the full conditional distribution for σ^2 is

$$\sigma^2|\tau^2,\mu,\phi,\boldsymbol{W},\boldsymbol{X} \sim \text{INGAM}(n/2+\alpha,(1/2)\boldsymbol{W}^T V^*(\phi)^{-1}\boldsymbol{W}+\beta).$$

Full conditional distribution for τ^2:

The full conditional distribution for τ^2 is

$$p(\tau^2|\sigma^2,\mu,\phi,\boldsymbol{W},\boldsymbol{X}) \propto p(\boldsymbol{X}|\boldsymbol{W},\mu,\tau^2)p(\tau^2)$$
$$\propto \tfrac{1}{(\tau^2)^{n/2}(\tau^2)^{\gamma+1}}\exp\{\tfrac{-1}{2\tau^2}(\boldsymbol{X}-(\boldsymbol{\mu}+\boldsymbol{W}))^T(\boldsymbol{X}-(\boldsymbol{\mu}+\boldsymbol{W}))-\tfrac{\delta}{\tau^2}\}.$$

Therefore, the full conditional distribution for τ^2 is

$$\tau^2|\sigma^2,\mu,\phi,\boldsymbol{W},\boldsymbol{X} \sim \text{INGAM}(n/2+\gamma,(1/2)(\boldsymbol{X}-(\boldsymbol{\mu}+\boldsymbol{W}))^T(\boldsymbol{X}-(\boldsymbol{\mu}+\boldsymbol{W}))+\delta).$$

Full conditional distribution for μ:

The full conditional distribution for $\boldsymbol{\mu}$ is

$$p(\boldsymbol{\mu}|\tau^2,\sigma^2,\phi,\boldsymbol{W},\boldsymbol{X}) \propto p(\boldsymbol{X}|\boldsymbol{W},\boldsymbol{\mu},\tau^2)p(\boldsymbol{\mu}).$$

We will first find the full conditional distribution for $\boldsymbol{\mu}$ and then the full conditional distribution for μ. Thus,

$$p(\boldsymbol{\mu}|\tau^2,\sigma^2,\phi,\boldsymbol{W},\boldsymbol{X})$$
$$\propto \exp\{\tfrac{-1}{2}((\boldsymbol{X}-\boldsymbol{W})-\boldsymbol{\mu}))^T(\tau^2 I)^{-1}((\boldsymbol{X}-\boldsymbol{W})-\boldsymbol{\mu}))+\tfrac{-1}{2}(\boldsymbol{\mu}-\boldsymbol{\lambda})^T(\psi^2 I)^{-1}(\boldsymbol{\mu}-\boldsymbol{\lambda})\}.$$

By completing the square, we have

$$\boldsymbol{\mu}|\tau^2,\sigma^2,\phi,\boldsymbol{W},\boldsymbol{X}\sim \mathrm{MVN}(\boldsymbol{\mu_o},\Sigma_o),$$

where $\boldsymbol{\mu_o}=(\frac{\psi^2\tau^2}{\tau^2+\psi^2})(\frac{1}{\psi^2}\boldsymbol{\lambda}+\frac{1}{\tau^2}(\boldsymbol{X}-\boldsymbol{W}))$ and $\Sigma_o=\frac{\psi^2\tau^2}{\tau^2+\psi^2}I$. Since $(\mathbf{1}^T/n)\boldsymbol{\mu}=\mu$, the full conditional distribution for μ is

$$\mu|\sigma^2,\tau^2,\phi,\boldsymbol{W},\boldsymbol{X}\sim \mathrm{NOR}(\mu_1,\sigma_1^2),$$

where $\mu_1=(\frac{\psi^2\tau^2}{\tau^2+\psi^2})[\frac{1}{\psi^2}\lambda+\frac{1}{\tau^2}(\bar{X}-\bar{W})]$ and $\sigma_1^2=(\frac{1}{n})(\frac{\psi^2\tau^2}{\tau^2+\psi^2})$.

Full conditional distribution of $\boldsymbol{W}$:

The full conditional distribution for the spatial random effects, $\boldsymbol{W}$, is

$$p(\boldsymbol{W}|\boldsymbol{X},\mu,\tau^2,\sigma^2,\phi)\propto p(\boldsymbol{X}|\boldsymbol{W},\mu,\tau^2)p(\boldsymbol{W}|\sigma^2,\phi)$$
$$\propto \exp\{\tfrac{-1}{2}(\boldsymbol{X}-(\boldsymbol{\mu}+\boldsymbol{W}))^T(\tau^2 I)^{-1}(\boldsymbol{X}-(\boldsymbol{\mu}+\boldsymbol{W}))\}\times\exp\{\tfrac{-1}{2}\boldsymbol{W}^T V(\sigma^2,\phi)^{-1}\boldsymbol{W}\}$$
$$=\exp\{\tfrac{-1}{2}((\boldsymbol{X}-\boldsymbol{\mu})-\boldsymbol{W})^T(\tau^2 I)^{-1}((\boldsymbol{X}-\boldsymbol{\mu})-\boldsymbol{W})+\tfrac{-1}{2}\boldsymbol{W}^T V(\sigma^2,\phi)^{-1}\boldsymbol{W}\}.$$

By completing the square, we have the full conditional distribution for $\boldsymbol{W}$ to be

$$\boldsymbol{W}|\boldsymbol{X},\mu,\tau^2,\sigma^2,\phi\sim \mathrm{MVN}(\boldsymbol{\mu}_w,\Sigma_w),$$

where $\boldsymbol{\mu}_w=[V^{-1}(\sigma^2,\phi)+\frac{1}{\tau^2}I]^{-1}[\frac{1}{\tau^2}(\boldsymbol{X}-\boldsymbol{\mu})]$ and $\Sigma_w=[V^{-1}(\sigma^2,\phi)+\frac{1}{\tau^2}I]^{-1}$.

Full conditional distribution of ϕ:

The full conditional distribution for ϕ is

$$p(\phi|\mu,\tau^2,\sigma^2,\phi,\boldsymbol{W},\boldsymbol{X})\propto p(\boldsymbol{W}|\sigma^2,\phi)p(\phi)$$
$$\propto \frac{\phi^{\eta-1}}{|V^*(\phi)|^{1/2}}\exp\{\tfrac{-1}{2\sigma^2}\boldsymbol{W}^T V^*(\phi)^{-1}\boldsymbol{W}-\theta\phi\}.$$

Hence, there is no closed form (i.e. known distribution) for the full conditional for ϕ. The full conditional distribution for ϕ is only known up to a proportional constant.

Appendix II

This appendix presents the derivation of the full conditional distributions required for the Gibbs sampler involving a data augmentation step within the Bayesian conditionally specified Gaussian model. The derivation of the transformed beta distribution is also presented in this appendix.

Full conditional distribution for τ^2:

The full conditional distribution for τ^2 is

$$p(\tau^2|\boldsymbol{y},\alpha,\eta) \propto p(\boldsymbol{y}|\tau^2,\alpha,\eta)p(\tau^2)$$
$$\propto (\tau^2)^{-(n/2+\gamma_o+1)} \exp\{\tfrac{-1}{\tau^2}(\tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{\alpha})^T(I-C)(\boldsymbol{y}-\boldsymbol{\alpha})+\beta_o)\}.$$

Hence, the full conditional distribution for τ^2 is

$$\tau^2|\boldsymbol{y},\alpha,\eta \sim INGAM(\tfrac{n}{2}+\gamma_o, \tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{\alpha})^T(I-C)(\boldsymbol{y}-\boldsymbol{\alpha})+\beta_o).$$

Full conditional distribution for α:

The full conditional distribution for α is

$$p(\alpha|\boldsymbol{y},\tau^2,\eta) \propto p(\boldsymbol{y}|\alpha,\tau^2,\eta)p(\alpha)$$
$$\propto \exp\{\tfrac{-1}{2}(\boldsymbol{y}-\boldsymbol{\alpha})^T M^{-1}(I-C)(\boldsymbol{y}-\boldsymbol{\alpha}) + \tfrac{-1}{2}(\boldsymbol{\alpha}-\mu_o\mathbf{1})^T(\sigma_o^2 I)^{-1}(\boldsymbol{\alpha}-\mu_o\mathbf{1})\}.$$

We will first find the full conditional distribution for $\boldsymbol{\alpha}$ and then the full conditional distribution for α, where $\boldsymbol{\alpha} = \mathbf{1}\alpha$. Completing the square, we have the full conditional distribution for $\boldsymbol{\alpha}$ to be

$$\boldsymbol{\alpha}|\boldsymbol{y},\tau^2,\eta \sim MVN(\boldsymbol{\mu}_\alpha, \Sigma_\alpha),$$

where $\boldsymbol{\mu}_\alpha = (\frac{1}{\sigma_o^2}I + \frac{1}{\tau^2}(I-C))^{-1}(\frac{\mu_o}{\sigma_o^2}\mathbf{1} + \frac{1}{\tau^2}(I-C)\boldsymbol{y})$ and $\Sigma_\alpha = (\frac{1}{\sigma_o^2}I + \frac{1}{\tau^2}(I-C))^{-1}$. Therefore, the full conditional distribution for α is

$$\alpha|\boldsymbol{y}, \tau^2, \eta \sim N(\mu_\alpha, \sigma_\alpha^2),$$

where $\mu_\alpha = \frac{1}{n}\mathbf{1}^T\boldsymbol{\mu}_\alpha$ and $\sigma_\alpha^2 = \frac{1}{n^2}\mathbf{1}^T\Sigma_\alpha^2\mathbf{1}$.

Full conditional distribution for η:

The full conditional distribution for η is

$$p(\eta|\boldsymbol{y}, \alpha, \tau^2) \propto p(\boldsymbol{y}|\alpha, \tau^2, \eta)p(\eta)$$
$$\propto |(I-C)^{-1}M|^{-1/2}\exp\{\tfrac{-1}{2}(\boldsymbol{y}-\boldsymbol{\alpha})^T M^{-1}(I-C)(\boldsymbol{y}-\boldsymbol{\alpha})\}$$
$$\times(\tfrac{h_n h_1}{h_1-h_n})^{\psi_o}\tfrac{1}{B(\psi_o,\phi_o)}(\eta-\tfrac{1}{h_1})^{\psi_o-1}[1-(\eta-\tfrac{1}{h_1})(\tfrac{h_n h_1}{h_1-h_n})]^{\phi_o-1}$$
$$\propto [\textstyle\prod_{i=1}^n(1-\eta h_i)]^{1/2}\exp\{\tfrac{\eta}{2\tau^2}(\boldsymbol{y}-\boldsymbol{\alpha})^T H(\boldsymbol{y}-\boldsymbol{\alpha})\}\times(\eta-\tfrac{1}{h_1})^{\psi_o-1}[1-(\eta-\tfrac{1}{h_1})(\tfrac{h_n h_1}{h_1-h_n})]^{\phi_o-1}.$$

There is no closed form for η's full conditional distribution (i.e. no known distribution). The full conditional distribution is only known up to a proportional constant. That is,

$$p(\eta|\boldsymbol{y}, \alpha, \tau^2) \propto$$
$$[\textstyle\prod_{i=1}^n(1-\eta h_i)]^{1/2}\exp\{\tfrac{\eta}{2\tau^2}(\boldsymbol{y}-\boldsymbol{\alpha})^T H(\boldsymbol{y}-\boldsymbol{\alpha})\}(\eta-\tfrac{1}{h_1})^{\psi_o-1}[1-(\eta-\tfrac{1}{h_1})(\tfrac{h_n h_1}{h_1-h_n})]^{\phi_o-1}.$$

Derivation of transformed Beta distribution:

If the support of x is $1/h_1 \le x \le 1/h_n$ and $y = (x-\frac{1}{h_1})(\frac{h_n h_1}{h_1-h_n})$, we have $0 \le y \le 1$. Likewise, if the support of y is $0 \le y \le 1$ and $x = y(\frac{h_1-h_n}{h_n h_1}) + \frac{1}{h_1}$, we have $\frac{1}{h_1} \le x \le \frac{1}{h_n}$. Let $y \sim Beta(\alpha, \beta)$ and $x = g(y) = y(\frac{h_1-h_n}{h_n h_1}) + \frac{1}{h_1}$. Hence, we have $y = g^{-1}(x) = (x-\frac{1}{h_1})(\frac{h_n h_1}{h_1-h_n})$. By transformation, we have

$$f_x(x) = f_y((x-\tfrac{1}{h_1})(\tfrac{h_n h_1}{h_1-h_n})) \times |\tfrac{d}{dx}(x-\tfrac{1}{h_1})(\tfrac{h_n h_1}{h_1-h_n})|.$$

Now, $f_y((x-\frac{1}{h_1})(\frac{h_n h_1}{h_1-h_n})) = \frac{1}{B(\alpha,\beta)}[(x-\frac{1}{h_1})(\frac{h_n h_1}{h_1-h_n})]^{\alpha-1}[1-(x-\frac{1}{h_1})(\frac{h_n h_1}{h_1-h_n})]^{\beta-1}$ and $|\frac{d}{dx}(x-\frac{1}{h_1})(\frac{h_n h_1}{h_1-h_n})| = \frac{h_n h_1}{h_1-h_n}$. Thus, the distribution for the transformed beta random variable x is

$$f_x(x) = \frac{1}{B(\alpha,\beta)}(\frac{h_n h_1}{h_1 - h_n})^{\alpha}[x - \frac{1}{h_1}]^{\alpha-1}[1 - (x - \frac{1}{h_1})(\frac{h_n h_1}{h_1 - h_n})]^{\beta-1},$$

with $1/h_1 \leq x \leq 1/h_n$.

References

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, B*, **36**, 192-236.

Carlin, B.P. and Louis, T.A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman & Hall, London.

Cressie, N.A.C. (1993). *Statistics for Spatial Data, Revised Edition.* John Wiley & Sons, Inc., New York.

Daniels, M.J., Lee, Y.D., and Kaiser, M.S. (2001). Assessing sources of variability in measurement of ambient particular matter. *Environmetrics*, **12**, 547-558.

de Oliveira, V., and Ecker, M.D. (2002). Bayesian hot spot detection in the presence of spatial trend: application to total nitrogen concentration in Chesapeake Bay. *Environmetrics*, **13**, 85-101.

Ecker, M.D., and Gelfand, A.E. (1997). Bayesian Variogram Modeling for an Isotropic Spatial Process. *Journal of Agricultural, Biological, and Environmental Statistics*, **2**, 347-369.

Fridley, B.L., and Dixon, P. (2003). Data Augmentation for a Bayesian Spatial Model involving Censored Observations. In preparation.

Fridley, B.L. and Dixon, P. (2003). Data Augmentation for a Conditionally Specified Gaussian Spatial Model involving Censored Observations. In preparation.

Gelfand, A.E., and Smith, A.F.M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, **85**, 398-409.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). *Bayesian Data Analysis.* Chapman & Hall, London.

Geman, S., and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Transactions on Patter Analysis and Machine Intelligence*, **6**, 721-741.

Gilks, W.R., Richardson, S., and Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice.* Chapman & Hall, London.

Haining, R., and Griffith, D. (1989). Maximum Likelihood Estimation with Missing Spatial Data and with an Application to Remotely Sensed Data. *Communications in Statistics: Theory and Methods*, **18(5)**, 1875-1894.

Kaiser, M.S., and Cressie, N. (2000). The Construction of Multivariate Distributions form Markov Random Fields. *Journal of Multivariate Analysis*, **73**, 199-220.

Li, K.H. (1988). Imputations Using Markov Chains. *Journal of Statistical Computation and Simulation*, **30**, 57-79.

Robert, C.P. (1995). Simulation of truncated normal variables. *Statistics and Computing*, **5**, 121-125.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data.* Chapman & Hall, London.

Tanner, M.A., and Wong, W.H. (1987). The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, **82**, 528-540.

GENERAL CONCLUSIONS

Censored observations occur often in environmental studies where one is measuring trace amounts of contaminants at different spatial locations. This research presented a method for the analysis of censored spatial data through the use of data augmentation. This procedure allows for more accurate modeling of the spatial processes involving censored observations. The data augmentation method for handling spatial censored data is implemented through a Markov chain Monte Carlo, with an additional step for the imputation or augmentation of the censored observations.

This dissertation presented data augmentation for the analysis of censored spatial data in terms of a traditional geostatistical model and a conditionally specified Gaussian or auto-regressive (CAR) model. In addition, a Bayesian framework was used in which proper priors were specified for all parameters. By using the Bayesian approach in estimation and prediction, we were able to incorporate the uncertainty of parameter estimation into the the posterior predictive distribution. In addition, the Bayesian approach produces an entire distribution for the prediction at each location, as opposed to a prediction point estimate and prediction error.

Overall, the data augmentation method for the analysis of censored spatial data was found to be superior to the common method of replacing the censored observations with a function of the level of detection. Along with producing biased parameter estimates, the common practice of replacing censored observations with a function of the level of detection under-estimates the variability in the approximated marginal densities. The data augmentation procedure produced more accurate marginal posterior distributions

and predictions as compared to the method of replacing the censored observations with half their level of detection. The largest difference in parameter estimation was in terms of the variability parameters. By replacing all censored observations with half their level of detection, the variability is vastly under-estimated. The data augmentation approach resulted in much larger variability parameter estimates, along with more variability in their marginal posterior densities.

The last section of this dissertation presented results from four simulation studies looking at the general properties of the data augmentation method, along with comparison to the method of replacing the censored observations with half their level of detection ($LOD/2$). Two of the four simulation studies were conducted using a Bayesian spatial or geostatistical model and the remaining two studies used a Bayesian conditionally specified Gaussian model. Data augmentation was shown to out-perform the LOD/2 method, in terms of both parameter estimation and prediction. In addition to investigation of the general accuracy of the data augmentation method, the simulation studies also investigated which factors may impact the procedure. Investigators can use the results from these simulation studies or other simulation studies to aid them in the construction of studies designed to investigate censored responses taken at various spatial locations.

ACKNOWLEDGMENTS

I would like to thank Dr. Dixon for all his kindness, wisdom and advice. The completion of this dissertation has been an invaluable learning experience for the both of us. I would also like to thank Dr. Bob Stephenson and Dr. Isaacson for always taking the time out of their busy days for a needed word of advice.

I would like to thank the New Jersey Department of Environmental Protection for their kind permission to use the site 15 data in the dissertation.

Many thanks to the members on my committee; Dr. Vardeman, Dr. Kaiser, Dr. Koehler and Dr. Glanville. In addition to these committee member, I would like to thank my classmates, the faculty members and the support staff at Iowa State University Department of Statistics - you have made my time at Iowa State University more enjoyable and rewarding. In addition to these individuals at Iowa State University, I would also like to thank Dr. Mariza de Andrade at the Mayo Clinic for the opportunities and support she provided me over the past few years.

Lastly, I would like to thank my family for their unconditional love and support over the years. I would like to thank my mother Carol for her encouragement and support, my sisters (and best friends) Becky and Erin for all the laughs and my brother Patrick. Finally, I would like to thank Troy for reminding me on many occasions to "breathe".