6-2006

Data augmentation for a Bayesian spatial model involving censored observations

Brooke L. Fridley
Mayo Clinic

Philip M. Dixon
Iowa State University, pdixon@iastate.edu

Data augmentation for a Bayesian spatial model involving censored observations

Abstract
Spatial environmental data sometimes include below detection limit observations (i.e. censored values reported as less than a level of detection). Historically, the most common practice for analysis of such data has been to replace the censored observations with some function of the level of detection (LOD), like LOD/2. We show that estimates and standard errors found using this single substitution method are biased. In particular, the spatial variance and variability in estimation is underestimated. We develop a measurement error Bayesian spatial model for the analysis of spatial data with censored values. Parameter estimation and predictions at observed and unobserved locations are computed using a data augmentation method using a Markov chain Monte Carlo algorithm. The data augmentation method is illustrated using data from a dioxin contaminated site in Missouri. We also use simulation to investigate the small sample properties of predictions and parameter estimates and the robustness of the data augmentation method.

Keywords
below detection limit observations, censored data, spatial correlation, varioram, kriging

Disciplines
Statistics and Probability

Comments
This preprint was published as Brooke L. Fridley and Philip Dixon, "Data augmentation for a Bayesian spatial model involving censored observations", *Environmetrics* (2007): 107-123, doi: 10.1002/env.806

Data augmentation for a Bayesian spatial model involving censored observations

Brooke L Fridley[1], Mayo Clinic

Philip Dixon, Iowa State University

Short title: Prediction with below-detection limit observations

[1]200 First Street SW, Rochester, MN 55905, fridley.brooke@mayo.edu, 507-538-3646, 507-284-9542 (fax)

SUMMARY

Spatial environmental data sometimes include below detection limit observations (i.e. censored values reported as less than a level of detection). Historically, the most common practice for analysis of such data has been to replace the censored observations with some function of the level of detection (LOD), like $LOD/2$. We show that estimates and standard errors found using this single substitution method are biased. In particular, the spatial variance and variability in estimation is underestimated. We develop a measurement error Bayesian spatial model for the analysis of spatial data with censored values. Parameter estimation and predictions at observed and unobserved locations are computed using a data augmentation method using a Markov chain Monte Carlo algorithm. The data augmentation method is illustrated using data from a dioxin contaminated site in Missouri. We also use simulation to investigate the small sample properties of predictions and parameter estimates and the robustness of the data augmentation method.

Keywords: below detection limit observations, censored data, spatial correlation, variogram, kriging

1 Introduction

Environmental studies often include some observations falling below a level of detection (LOD). These values are reported as $< LOD$, where the LOD is a specified value for each observation. The values reported as $< LOD$ are either left censored or interval censored $0 < x < LOD$. Censored spatial data are often analyzed by ignoring spatial correlation and using one of many methods available for independent observations (Helsel, 2005; Gibbons, 1995; Porter, Ward and Bell, 1988). Or, the censoring is ignored by substituting some function of the level of detection (e.g. $LOD/2$, LOD) for the censored values and then using a commonly available spatial method, e.g. variogram estimation and kriging. This substitution simplifies the spatial analysis but results in biased estimates of the mean and variance (Helsel 2005), and, as we show later, a biased estimate of the overall spatial variability.

Alternatives to substitution, such as maximum likelihood estimation, are difficult because direct evaluation of the likelihood for correlated data with censored values involves computationally intractable integrals (Abrahamsen and Benth 2001). Although Militino and Ugarte (1999) develop an EM algorithm for kriging censored spatial data, most approaches have used Monte-Carlo approximation of the integral, e.g. Stein 1992. Data augmentation provides a mechanism that eliminates the need to evaluate the high dimensional integral. Abrahamsen and Benth (2001) combined data augmentation, inequality constraints and universal kriging to map a spatial process. Lockwood et al. (2004) construct a Bayesian model for the joint distribution of seven

groundwater contaminants on a spatial lattice. And, Hopke, Liu and Rubin (2001) analyzed pollutant data with spatio-temporal correlation using multiple imputation, i.e. using data augmentation to construct a few complete data sets. However, most implementations of Monte-Carlo approximation for spatial censored data have been based on Bayesian kriging or prediction (Kitanidis 1986).

In Bayesian prediction, the posterior predictive distribution of values at unobserved locations is estimated by a Markov chain Monte Carlo (MCMC) algorithm given the model for the observations, the data and the specified prior distributions for the parameters. Bayesian prediction has been used to map contaminant concentrations and define hot spots, areas of extreme contamination (de Oliveira and Ecker 2001). Recently, de Oliveira (2005) developed a Markov chain Monte Carlo algorithm to fit a Bayesian spatial model to data with censored values.

In this paper, we develop a measurement error Bayesian spatial model for which data augmentation is an especially convenient way to analyze spatially correlation data in which some observations are censored. After describing the model, we derive the conditional distributions and describe a Markov chain Monte Carlo algorithm to estimate the posterior distribution of parameters and posterior predictive distribution. Results from a simulation study are used to evaluate the small sample performance and the robustness to misspecification of the spatial covariance function. Analysis of data from a dioxin contaminated site in Missouri are used to illustrate the method and compare Bayesian data augmentation to substituting half the level of detection ($LOD/2$) for the censored observations.

2 Bayesian spatial measurement-error model and prediction

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

$$Y(s_i) = \mu + W(s_i) + \varepsilon(s_i),$$

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 $\varepsilon(s_i) \sim$ independent $N(0, \tau^2)$, and $W(s_i)$ represents the random spatial effect at location s_i with $\boldsymbol{W} \sim MVN(\mathbf{0}, \boldsymbol{V}(\boldsymbol{\Theta}))$ (Cressie, 1993; Carlin and Louis, 1996; Ecker and Gelfand, 1997). Hence, we have $\boldsymbol{Y} \sim MVN(\boldsymbol{\mu}, \boldsymbol{V}(\boldsymbol{\Theta}) + \tau^2 \boldsymbol{I})$ and $\boldsymbol{Y}|\boldsymbol{W} \sim MVN(\boldsymbol{\mu} + \boldsymbol{W}, \tau^2 \boldsymbol{I})$. There are various ways to parameterize $\boldsymbol{V}(\boldsymbol{\Theta})$. One isotropic parameterization for the spatial covariance matrix is the exponential, in which $V(\sigma^2, \phi)_{ij} = \sigma^2 \exp\{-d_{ij}/\phi\}$ and $d_{ij} = \|s_i - s_j\|$, where σ^2 represents the variability in the spatial process and ϕ represents the spatial range parameter. Other common isotropic parameterizations for the spatial covariance matrix are presented in Table 1 This model differs from others proposed for spatial censored data by the introduction of an unobserved latent spatial process, $\boldsymbol{W}$ (de Oliveira V, 2005).

Insert Table 1: Isotropic parameterizations for the spatial covariance matrix here

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 conjugate proper priors to non-informative or improper priors (Ecker and Gelfand, 1987; Berger, de Oliveira, and Sanso, 2001). Whatever the choice, sensitivity analysis for the final inferences with respect to the prior distributions is recommended. A possible proper prior specification for a spatial Bayesian model with isotropic exponential spatial covariance structure would be $\sigma^2 \sim IG(\alpha, \beta)$, $\tau^2 \sim IG(\gamma, \delta)$, $\mu \sim N(\lambda, \psi^2)$, $\phi \sim G(\eta, \theta)$.

Producing a map of contaminated areas requires predicting values at unobserved or ungauged locations. Let $\boldsymbol{Y}_u$ represent the vector of values at ungauged (unobserved) locations and $\boldsymbol{Y}_g$ represent the vector of values at gauged (observed) locations. Assume temporarily that there are no censored observations. The joint distribution of $\boldsymbol{Y}_u$ and $\boldsymbol{Y}_g$ can be written as

$$\begin{pmatrix} \boldsymbol{Y}_u \\ \boldsymbol{Y}_g \end{pmatrix} \sim MVN\left(\begin{pmatrix} \boldsymbol{\mu}_u \\ \boldsymbol{\mu}_g \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{uu} & \boldsymbol{\Sigma}_{ug} \\ \boldsymbol{\Sigma}_{gu} & \boldsymbol{\Sigma}_{gg} \end{pmatrix}\right),$$

with $\boldsymbol{\mu}_g$ and $\boldsymbol{\mu}_u$ representing the mean response of the ungauged and gauged locations, $\boldsymbol{\Sigma}_{uu}$, $\boldsymbol{\Sigma}_{gg}$, $\boldsymbol{\Sigma}_{ug}$ and $\boldsymbol{\Sigma}_{gu}$ represent the partitioning of the covariance matrix for the ungauged and gauged locations. Bayesian prediction then uses the posterior predictive distribution, $p(\boldsymbol{Y}_u|\boldsymbol{Y}_g)$, as the method for prediction at ungauged locations (Carlin and Louis, 1995; Gelman et al, 1995) . The resulting conditional distribution of the ungauged locations given the gauged locations, $p(\boldsymbol{Y}_u|\boldsymbol{Y}_g)$, is a multivariate normal distribution (Johnson and Wichern, 1982).

Data augmentation, as proposed by Tanner and Wong (1987), can be used to incorporate information from censored observations. Within a Markov chain Monte Carlo, the idea is as follows. Given the current value of the parameters $\boldsymbol{\Theta}^{(m)}$, draw a vector $\boldsymbol{Y}_c^{(m+1)}$ for the censored data from $p(\boldsymbol{Y}_c|\boldsymbol{Y}_o, \boldsymbol{\Theta}^{(m)})$, where $\boldsymbol{Y}_c$ represents the censored values and $\boldsymbol{Y}_o$ represent the observed or uncensored values. Then based on $\boldsymbol{Y}_c^{(m+1)}$, draw $\boldsymbol{\Theta}^{(m+1)}$ from $p(\boldsymbol{\Theta}|\boldsymbol{Y}_o, \boldsymbol{Y}_c^{(m+1)})$, the complete data posterior for $\boldsymbol{\Theta}$.

At every iteration of the chain 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 $\{\boldsymbol{\Theta}^{(m)}, \boldsymbol{Y}_c^{(m)} : m = 1, 2, ...\}$ whose stationary distribution is $p(\boldsymbol{\Theta}, \boldsymbol{Y}_c|\boldsymbol{Y}_o)$ (Shafer, 1997; Gilks, Richardson and Spiegelhalter, 1996).

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

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

with $\boldsymbol{\mu}_{u.g}^{(m)} = \boldsymbol{\mu}_u^{(m)} + \boldsymbol{\Sigma}_{ug}^{(m)}\boldsymbol{\Sigma}_{gg}^{-1(m)}(\boldsymbol{Y}^{*(m)} - \boldsymbol{\mu}_g^{(m)})$, $\boldsymbol{\Sigma}_{u.g}^{(m)} = \boldsymbol{\Sigma}_{uu}^{(m)} - \boldsymbol{\Sigma}_{ug}^{(m)}\boldsymbol{\Sigma}_{gg}^{-1(m)}\boldsymbol{\Sigma}_{gu}^{(m)}$, and $\boldsymbol{Y}^{*(m)} = (\boldsymbol{Y}_{go}^T, \boldsymbol{Y}_{gc}^{(m)T})^T$, for various MCMC iterations m, where $\boldsymbol{Y}_{gc}^{(m)}$ represents the augmented data for the censored observations at iteration m of the MCMC (Fridley,

2003; de Oliveira, 2005; de Oliveira and Ecker, 2002; Gelman, Carlin, Stern and Rubin, 1995). When censored spatial data are modeled by a measurement error Bayesian spatial model with proper priors, data augmentation can be completed within a Gibbs sampler (Geman and Geman, 1984; Fridley, 2003).

3 Markov chain Monte Carlo algorithm to approximate the posterior distributions

The MCMC algorithm for an isotropic exponential spatial covariance matrix is a combination of Gibbs sampling and Metropolis-Hastings steps. Recall, the ij entry of the isotropic exponential spatial covariance matrix is defined to be $V(\sigma^2, \phi)_{ij} = \sigma^2 \exp\{-d_{ij}/\phi\}$ with $V^*_{ij}(\phi) = \exp\{-d_{ij}/\phi\}$ where d_{ij} is the Euclidean distance between location s_i and s_j. Details of the MCMC algorithm for the isotropic spherical and isotropic Gaussian spatial covariance models are presented in the Appendix.

Set $m = 0$. Begin by setting starting values for the mean $\mu^{(0)}$, random error variance component $\tau^{2(0)}$,spatial process variance component $\sigma^{2(0)}$, vector of random spatial effects $\boldsymbol{W}^{(0)}$, and spatial range parameter $\phi^{(0)}$ and set the censored values equal to their level of detection or half their level of detection. Let $\boldsymbol{Y}^{(m)} = (\boldsymbol{Y}_c^{(m)T}, \boldsymbol{Y}_o^T)^T$ represent the augmented-complete data at iteration m, where $\boldsymbol{Y_c}$ and $\boldsymbol{Y_o}$ represent the censored data and observed data, respectively.

First, generate $\mu^{(m+1)}$ from $N(\mu_1^{(m+1)}, \sigma_1^{2(m+1)})$, where $\mu_1^{(m+1)} = (\frac{\psi^2\tau^{2(m)}}{\tau^{2(m)}+\psi^2})[\frac{1}{\psi^2}\lambda +$

$\frac{1}{\tau^{2(m)}}(\bar{Y}^{(m)} - \bar{W}^{(m)})]$ and $\sigma_1^{2(m+1)} = (\frac{1}{n})(\frac{\psi^2\tau^{2(m)}}{\tau^{2(m)}+\psi^2})$. Next, draw $\tau^{2(m+1)}$ from $IG(n/2 + \gamma, (1/2)(\boldsymbol{Y}^{(m)} - (\boldsymbol{\mu}^{(m+1)} + \boldsymbol{W}^{(m)}))^T(\boldsymbol{Y}^{(m)} - (\boldsymbol{\mu}^{(m+1)} + \boldsymbol{W}^{(m)})) + \delta)$ and $\sigma^{2(m+1)}$ from $IG(n/2 + \alpha, (1/2)\boldsymbol{W}^{T(m)}V^*(\phi^{(m)})^{-1}\boldsymbol{W}^{(m)} + \beta)$, where $\boldsymbol{\mu}^{(m+1)} = \mu^{(m+1)}\mathbf{1}$.

The random spatial effects, $\boldsymbol{W}^{(m+1)}$, are then generated from a $MVN(\boldsymbol{\mu}_w^{(m+1)}, \boldsymbol{\Sigma}_w^{(m+1)})$, where $\boldsymbol{\mu}_w^{(m+1)} = [\boldsymbol{V}^{-1}(\sigma^{2(m+1)}, \phi^{(m)}) + \frac{1}{\tau^{2(m+1)}}\boldsymbol{I}]^{-1}[\frac{1}{\tau^{2(m+1)}}(\boldsymbol{Y}^{(m)} - \boldsymbol{\mu}^{(m+1)})]$ and $\boldsymbol{\Sigma}_w^{(m+1)} = [\boldsymbol{V}^{-1}(\sigma^{2(m+1)}, \phi^{(m)}) + \frac{1}{\tau^{2(m+1)}}\boldsymbol{I}]^{-1}$, where $\boldsymbol{I}$ is a $n \times n$ identity matrix. Lastly, using a Metropolis-Hastings step(s), $\phi^{(m+1)}$ is simulated from its full conditional distribution which is proportional to $\frac{\phi^{\eta-1}}{|\boldsymbol{V}^*(\phi)|^{1/2}} \exp\{\frac{-1}{2\sigma^{2(m+1)}}\boldsymbol{W}^{T(m+1)}\boldsymbol{V}^*(\phi)^{-1}\boldsymbol{W}^{(m+1)} - \theta\phi\}$ (Hastings, 1970; Metropolis & Ulam, 1949; Metropolis et al, 1953). This concludes the posterior step at iteration m.

The imputation step at iteration m is then completed as follows. Let $\boldsymbol{Y}_c^T = (Y_{1c}, Y_{2c}, .., Y_{kc})$ represent the k censored values and $LOD_1, LOD_2, ..., LOD_k$ represent the level of detections for the k censored values. For each censored value, generate $Y_{ic}^{(m+1)}$ from $N(\mu^{(m+1)} + W_i^{(m+1)}, \tau^{2(m+1)})$, truncated at the level of detection LOD_i, for $i = 1, ..., k$. Prediction can then be completed for a set of locations based on the augmented-complete data and parameter values at iteration m as outlined in section 2. Set $m = m + 1$ and repeat a large number of times. Note that the imputation of censored values is especially easy for the measurement error model proposed here. Given the current values of $\boldsymbol{W}^{(m+1)}$, $\mu^{(m+1)}$, and $\tau^{2(m+1)}$, the censored values are imputed by generating independent realizations from univariate truncated normal distributions.

4 Simulation Studies

Three simulation studies were conducted to investigate properties of the estimates, properties of predictions, and robustness to misspecification of the spatial covariance function. In addition to assessing the validity of the data augmentation procedure, the simulation studies investigating estimation and prediction also compare the data augmentation method to the method of replacing the censored observations with half their level of detection ($LOD/2$) (Fridley, 2003).

4.1 Estimation

The first simulation study assessed properties of the parameter estimates produced by the data augmentation (DA) and LOD/2 methods. One thousand generated datasets were constructed containing 100 observations on a 10x10 regular grid or lattice. The data were simulated using the isotropic exponential parameterization of the spatial covariance matrix as outlined in section 2 with parameter values of μ = 0, τ^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.

Estimates for the parameters μ, τ^2, σ^2 and ϕ were taken to be the median of the simulated posterior distributions. Summary of the estimates for μ, τ^2, σ^2 and ϕ across the 1000 simulated datasets are displayed in Table 2. From the table of results, 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. 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, data augmentation method producing more variability in the estimates for the parameters τ^2 and σ^2 in relation to the LOD/2 method. Hence, the method of subsituting LOD/2 for the censored observations is underestimating the error in estimation (i.e. the confidence intervals are too small).

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.

Insert Table 2: Summary of estimates for the 1000 simulated datasets

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

4.2 Prediction

The second simulation study compared 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. To investigate, 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 using the spatial exponential model described in section 3 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 were used in parameter estimation along with prediction. To illustrate further, Figure 1 displays the locations used in estimation and the locations set aside for future comparisons of predictions. Note, that the locations for prediction represent the best possible scenario for prediction, since most locations are surrounded by four observed locations.

Insert Figure 1: Locations for simulation study investigating prediction error and robustness

The prediction stage of the analysis was completed using the Bayesian prediction method outlined in section 2. 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 estimated 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 $\sum_{i=1}^{n}(\hat{y}_i - y_i)^2/n$). 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.

Insert Table 4: Summary of mean prediction error and mean squared prediction error

Table 4 illustrates 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. 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.

4.3 Robustness to model misspecification

The last simulation study investigated robustness to misspecification of the spatial covariance model. Data sets were generated with 225 observations on a regular 15 x 15 lattice, with 5 units between nearest neighbors, using either an exponential, a Gaussian or a spherical covariance model. Parameters were set to $\mu = 0$, $\tau^2 = 2$, $\sigma^2 = 5$, and $\phi = 20$. 20% of the observations were censored. Three hundred data sets were simulated for each covariance structure, wherein a third of the data sets were

analyzed according to either a Gaussian, exponential or spherical model (analysis model). Thus, one hundred data sets where analyzed in each combination of data model and analysis model.

Data from half of the locations (113 locations) were used to estimate parameters and approximate the posterior predictive distributions and the remaining half of the locations (112 locations) were used for future comparison of predictions (Figure 1). The median of each posterior distribution was used as the prediction at that location. Overall prediction accuracy was summarized using the mean square prediction error (MSPE), computed for the 112 locations not used in parameter estimation. Results are reported as the median MSPE across the 100 data sets (Table 5).

The median MSPE is smallest when data are analyzed using the correct model, i.e. the model used to generate the data. However, the increase in median MSPE is small (8% or less) when the wrong model is used. Predictions are reasonably robust to misspecification of the spatial covariance function, at least among the three isotropic models considered here.

Insert Table 5: Median of the estimated mean squared prediction errors

5 Missouri dioxin contamination site

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. They sampled various areas, including the shoulder of the road, to determine their contamination levels. The data reported in Zirschky and Harris (1986), from an 3600 m x 65 m along the two shoulders 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.

Insert Figure 2: Missouri study locations here

Forty-three percent 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. All samples were analyzed according to USEPA approved procedures. The clean-up criteria for dioxin is 1 μg/kg (Zirschky & Harris, 1986).

5.2 Model specification

The Bayesian spatial model with the exponential correlation structure described in section 2 assumes normality and spatial isotropy. The distribution of dioxin concentrations was skewed and thus a log transformation was applied to the original observations. Exploratory analysis of the spatial correlation suggested geometric anisotropy. The dependency at lag distances of 10 m in the Y direction (perpendicular to the road) was similar to that at lag distances of 1000 m in the X direction (parallel to the road). After dividing the X coordinate by 100 the isotropy assumption seemed reasonable.

The analysis was completed using an exponential covariance structure with prior distributions of $\mu \sim N(0, 50)$, $\sigma^2 \sim IG(2.1, 6.6)$, $\phi \sim G(2, 0.1)$, and $\tau^2 \sim IG(2.1, 0.55)$. These priors have large, but finite, variance with the distributions centered roughly around the means estimated by replacing the censored values with their levels of detection in a non-Bayesian geostatistical analysis. Alternatively, a fully Bayesian analysis could be applied involving the specification of hyper-priors. The use of improper or 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), a gamma candidate generating distribution of $G(2X, 2)$ was used, where X represents the current value of ϕ. By choosing $G(2X, 2)$, the mean of the candidate generating distribution for the current iteration of the chain is the current value for ϕ. 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.

For comparison to the DA estimates that account for spatial correlation, the mean and variance of the log transformed observations were estimated using two non-spatial estimators. Maximum likelihood estimates assuming a normal distribution were computed by numerically maximizing the log-likelihood function (Helsel, 2005). A 95% confidence interval for the mean was computed using a Z quantile and the asymptotic standard error from the numerical Hessian matrix. Nonparametric estimates of the mean, the variance, and the standard error of the mean were computed from the Kaplan-Meier estimate of the cumulative frequency distribution (Helsel, 2005). Computations were done in R, using the optim function and the NADA package (Lee, 2005).

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 6. Table 6 displays medians and 95% credible intervals for the parameters μ, τ^2, σ^2, and ϕ. 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 the estimated posterior distribution of the spatial variability parameter σ^2. Data augmentation suggests a much larger amount of spatially associated variation. 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.

Insert Table 6: Dioxin: Median and 95% credible intervals based on the simulated marginal posterior distributions here

The nonparametric Kaplan-Meier (KM) estimate of the mean is close to the spatial DA estimate, but the KM estimate of the variance is less than half of the comparable quantity, $\tau^2 + \sigma^2$ from the spatial model (Table 7). The underestimation of the variance is likely a consequence of the large number of values censored at the smallest observed dioxin concentration, 0.1. In contrast, the parametric log-normal MLE of the mean is smaller than the spatial estimate, but the MLE of the variance is about

the same. As expected, 95% ML and KM confidence intervals for the mean dioxin concentration (Table 7) are much narrower than the DA intervals because the ML and KM estimates ignore the positive spatial dependence.

Insert Table 7: Dioxin: Estimated mean, 95% confidence interval for the mean, and variance of log transformed concentrations using maximum likelihood (ML) and Kaplan-Meier (KM) here

Since the goal of this study is the identification of areas requiring clean-up based on a criteria of 0 ln(μ/kg), since 1 μ/kg is the clean-up criteria on the original scale, Bayesian prediction results are presented in Figure 3. Figure 3 displays the median of the approximated data augmentation Bayesian posterior predictive distribution. Based on this plot 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.

Insert Figure 3: Dioxin: Posterior median of the Bayesian predictive distribution using data augmentation for censored values here

Figure 4 provides comparison of predictions produced by the DA and LOD/2 methods. The figure portrays the difference in medians of posterior predictive distribution produced by using the DA and LOD/2 methods. The figure shows 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).

Insert Figure 4: Dioxin: Difference in posterior medians for DA and

LOD/2 methods for handling censored values (LOD/2 - DA) here

To illustrate the difference in the clean-up regions determined by the DA and LOD/2 methods, Figure 5 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 between the DA and LOD/2 methods.

Insert Figure 5: Dioxin: Difference in clean-up regions between the DA method (solid line) and the LOD/2 method (dashed line) based on different probabilities of being above clean-up cut-off values

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 8. As Table 8 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.

Insert Table 8: Dioxin: Median and 95% credible intervals based on the simulated marginal posterior distributions for three different prior specifications here

6 Discussion and conclusions

We have illustrated the use of data augmentation for the analysis of spatially correlated data in which some of the observation are censored within a measurement error Bayesian spatial model. 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 an environmental contamination site in Missouri. In addition, three simulation study were conducted to investigate properties of estimation, prediction and robustness of data augmentation to misspecification of the spatial dependency.

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 geometric anisotropy present in the Missouri dataset lead to simplification of the analysis. The data augmentation method does not require isotropy or geometric anisotropy. The procedure can be extended to cases involving directional dependence where simple techniques/solutions to handle directional dependence are not applicable, such as modeling the directional dependency (Cressie, 1993; Ecker and Gelfand, 1999; Ecker and Gelfand, 2003). Also, trend could be accommodated by using a more complicated model for the mean. The procedure could also be extended to other Gaussian Bayesian spatial models and other forms of censoring (e.g. right censoring, interval censoring). In addition to extension to various forms of censoring and Bayesian spatial models, the data augmentation method can be extended to non-Gaussian models and conditionally specified models (Diggle, Tawn, Moyeed, 1998; Fridley, 2003).

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, especially when augmenting missing or censored values (Schafer, 1997).

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 the 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, 2003). Data augmentation can be easily applied to analyze censored spatial data, producing more accurate marginal posterior distributions and predictions. The difference in predicted contamination levels between the ad hoc methods and the data augmentation method can vary (as seen in the simulation studies and the dioxin study), with moderate differences in the size of the regions requiring clean-up of the contamination between the two methods, as seen in the dioxin study, to more extreme differences in the size of the regions requiring clean-up, depending on the amount of censored data, level of detection(s) and required clean-up level for the contaminate.

Appendix

Below is the Markov chain Monte Carlo algorithm with a data augmentation step for the analysis of censored spatial data within a measurement error Bayesian spatial model with spherical spatial covariance matrix and prior distributions set to be $\tau^2 \sim IG(\gamma, \delta)$, $\sigma^2 \sim IG(\alpha, \beta)$, $\phi \sim G(\eta, \theta)$ and $\mu \sim N(\lambda, \psi^2)$. Recall, that $V(\Theta)_{ij} = V(\sigma^2, \phi)_{ij} = \sigma^2 \frac{1}{2}(\frac{d_{ij}^3}{\phi^3} - \frac{3d_{ij}}{\phi} + 2)$ and let $V^*(\phi)_{ij} = \frac{1}{2}(\frac{d_{ij}^3}{\phi^3} - \frac{3d_{ij}}{\phi} + 2)$, where d_{ij} is the Euclidean distance between location s_i and s_j.

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

   $\boldsymbol{Y}^{T(m)} = (\boldsymbol{Y}_c^{T(m)}, \boldsymbol{Y}_o^T)^T$, where $\boldsymbol{Y_c}$ and $\boldsymbol{Y_o}$ represent the censored data and observed data, respectively.

3. Generate $\mu^{(m+1)}$ from $N(\mu_1^{(m+1)}, \sigma_1^{2(m+1)})$, 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

   $IG(n/2 + \gamma, (1/2)(\boldsymbol{Y}^{(m)} - (\boldsymbol{\mu}^{(m+1)} + \boldsymbol{W}^{(m)}))^T(\boldsymbol{Y}^{(m)} - (\boldsymbol{\mu}^{(m+1)} + \boldsymbol{W}^{(m)})) + \delta)$.

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

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

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

$\boldsymbol{\Sigma}_w^{(m+1)} = [\boldsymbol{V}^{-1}(\sigma^{2(m+1)}, \phi^{(m)}) + \frac{1}{\tau^{2(m+1)}}\boldsymbol{I}]^{-1}$, where $\boldsymbol{I}$ represents a identity matrix of appropriate dimensions.

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

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

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

8. Using $\mu^{(m+1)}, \tau^{2(m+1)}, \sigma^{2(m+1)}, \phi^{(m+1)}, \boldsymbol{W}^{(m+1)}$ and $\boldsymbol{Y}^{(m)}$, impute values for $\boldsymbol{Y}_c$ to produce $\boldsymbol{Y}_c^{(m+1)}$ as follows. Let k represent the number of censored observations and let $\boldsymbol{Y}_c = (Y_{1c}, Y_{2c}, .., Y_{kc})$ and let LOD_i represent the level of detection for the i^{th} censored value, $i = 1, ..., k$.

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

   ....

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

9. Complete prediction for a set of locations based on current values of the augmented-complete data and the parameters.

10. Repeat the algorithm a large number of times.

The Markov chain Monte Carlo algorithm with data augmentation step for the analysis of censored spatial data with a Gaussian spatial covariance matrix is similar to the above outlined MCMC, with $V(\Theta)_{ij} = V(\sigma^2, \phi)_{ij} = \sigma^2 \exp\{-(d_{ij}/\phi)^2\}$ and $V^*(\phi)_{ij} = \exp\{-(d_{ij}/\phi)^2\}$.

References

Abrahamsen P, Benth FE. 2001. Kriging with Inequality Constraints. *Mathematical Geology* **33**: 719-744.

Berger JO, de Oliveira V, Sanso B. 2001. Objective Bayesian Analysis of Spatially Correlated Data. *Journal of the American Statistical Association* **96**: 1361-1374.

Carlin BP, Louis TA. 1996. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall: London.

Cressie NAC. 1993. *Statistics for Spatial Data, Revised Edition*. John Wiley & Sons, Inc.: New York.

de Oliveira V. 2005. Bayesian inference and prediction of Gaussian random fields based on censored data. *Journal of Computational and Graphical Statistics* **14**: 95-115.

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

Diggle PJ, Tawn JA, Moyeed RA. 1998. Model-based geostatistics. *Applied Statistics* **47**: 299-350.

Ecker MD, Gelfand AE. 1997. Bayesian Variogram Modeling for an Isotropic Spatial Process. *Journal of Agricultural, Biological, and Environmental Statistics* **2**: 347-369.

Ecker MD, Gelfand AE. 1999. Bayesian Modeling and Inference for Geometrically Anisotropic Spatial Data. *Mathematical Geology* **31**(1): 67-83.

Ecker MD, Gelfand AE. 2003. Spatial modeling and prediction under stationary non-geometric range anisotropy. *Environmental and Ecological Statistics* **10**: 165-178.

Fridley BL. 2003. Data augmentation for the handling of censored spatial data. Ph.D. dissertation, Dept. Statistics, Iowa State Univ.

Gelman A, Carlin JB, Stern HS, Rubin DB. 1995. *Bayesian Data Analysis*. Chapman & Hall: London.

Geman S, 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 WR, Richardson S, Spiegelhalter DJ. 1996. *Markov Chain Monte Carlo in Practice*. Chapman & Hall: London.

Hastings WK. 1970. Monte Carlo Sampling Methods Using Markov

Helsel DR. 2005. *Nondetects and Data Analysis*. Wiley & Sons: New York.

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

Johnson RA, Wichern DW. 1982. *Applied Multivariate Statistical Analysis,* 2^{nd} *Ed.* Prentice Hall: New Jersey.

Kitanidis PK. 1986. Parameter uncertainty in estimation of spatial functions: Bayesian analysis. *Water Resources Reseach* **22**: 499-507.

Lee L. 2005. NADA: Nondetects And Data Analysis for environmental data. R package version 1.2-2.

Lockwood JR, Schervish MJ, Gurian PL, Small MJ. 2004. Analysis of Contaminant Co-Occurrence in Community Water Systems. *Journal of the American Statistical Association* **99**: 45- 56.

Metropolis N, Ulam S. 1949. The Monte Carlo Method. *Journal of the American Statistical Association* **44**:335-341.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, and Teller E. 1953. Equation of state calcuations by fast computing machines. *Journal of Chemical Physics* **21**: 1087-1092.

Militino AF, Ugarte, MD. 1999. Analyzing Censored Spatial Data. *Mathematical Geology* **31**: 551-562.

Porter PS, Ward RC, Bell HF. 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 JL. 1997. *Analysis of Incomplete Multivariate Data.* Chapman & Hall: London.

Stein ML. 1992. Prediction and Inference for Truncated Spatial Data. *Journal of*

Computational and Graphical Statistics **1**: 91-110.

Tanner MA, Wong WH. 1987. The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association* **82**: 528-540.

United States Unvironmental Protection Agency (USEPA). *Determination of 2,3,7,8-TCDD in Soil and Sediment.* USEPA Region VII Laboratory, Kansas City, Kansas.

Zirschky JH, Harris DJ. 1986. Geostatistical Analysis of Hazardous Waste Site Data. *Journal of Environmental Engineering* **112**: 770-784.

Table 1: Common istropic parameterizations of the spatial covariance matrix with $\sigma^2 > 0$, $\phi > 0$ and d_{ij} representing the Euclidean distance between location s_i and s_j.

Name	Parameterization of $V(\cdot)$
Spherical	$\sigma^2 \frac{1}{2}(d^3/\phi^3 - 3d/\phi + 2)$ if $d \leq \phi$
	0 if $d > \phi$
Exponential	$\sigma^2 \exp\{-d/\phi\}$
Gaussian	$\sigma^2 \exp\{(-d/\phi)^2\}$
Power Exponential	$\sigma^2 \exp\{(-d/\phi)^\alpha\}$ with $0 < \alpha \leq 2$

Table 2: Summary of estimates using data augmentation (DA) or substitution of LOD/2 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 3: Summary of lengths of 95% credible intervals using data augmentation (DA) or substitution of LOD/2 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

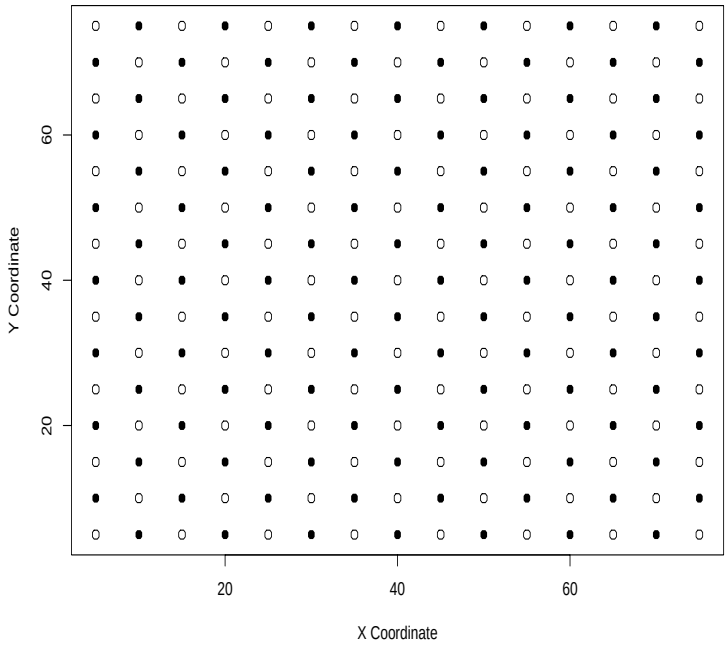

Figure 1: Locations for simulation study investigating prediction error and robustness, ∘ represent locations used in parameter estimation and • represent locations used for comparision of predictions

Table 4: Summary of mean prediction error (MPE) and mean squared prediction error (MSPE) for the 50 simulated datasets using data augmentation (DA) or substitution of LOD/2. We also report the mean squared difference (MSD) in the two predictions (LOD/2 - DA)

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
	MSD	-0.138	0.342	0.567	0.575	0.728	1.543

Table 5: Median of the estimated mean squared prediction error

	Analysis Model		
Data Model	Exponential	Gaussian	Spherical
Exponential	383.7	397.3	393.9
Gaussian	279.1	267.6	289.2
Spherical	504.8	478.7	474.1

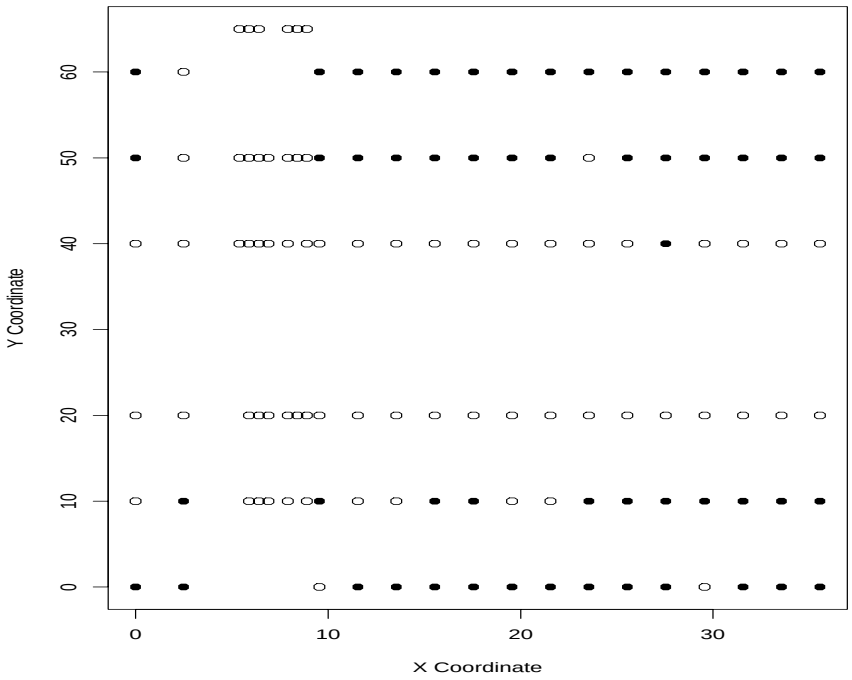

Figure 2: Missouri study locations, ∘ represents an observed value and • represents a censored value

Table 6: Dioxin: Median and 95% credible intervals based on the simulated marginal posterior distributions using data augmentation (**DA**), replacement of censored values by **LOD/2** and replacement of censored values by **LOD**.

	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)

Table 7: Dioxin: Estimated mean, 95% confidence interval for the mean, and variance of log transformed concentrations using maximum likelihood (**ML**) and Kaplan-Meier (**KM**), both assuming independent observations.

	ML		**KM**	
	Estimate	Interval	Estimate	Interval
μ	-1.42	(-1.97, -0.88)	-0.68	(-0.99, -0.37)
σ^2	7.29		3.16	

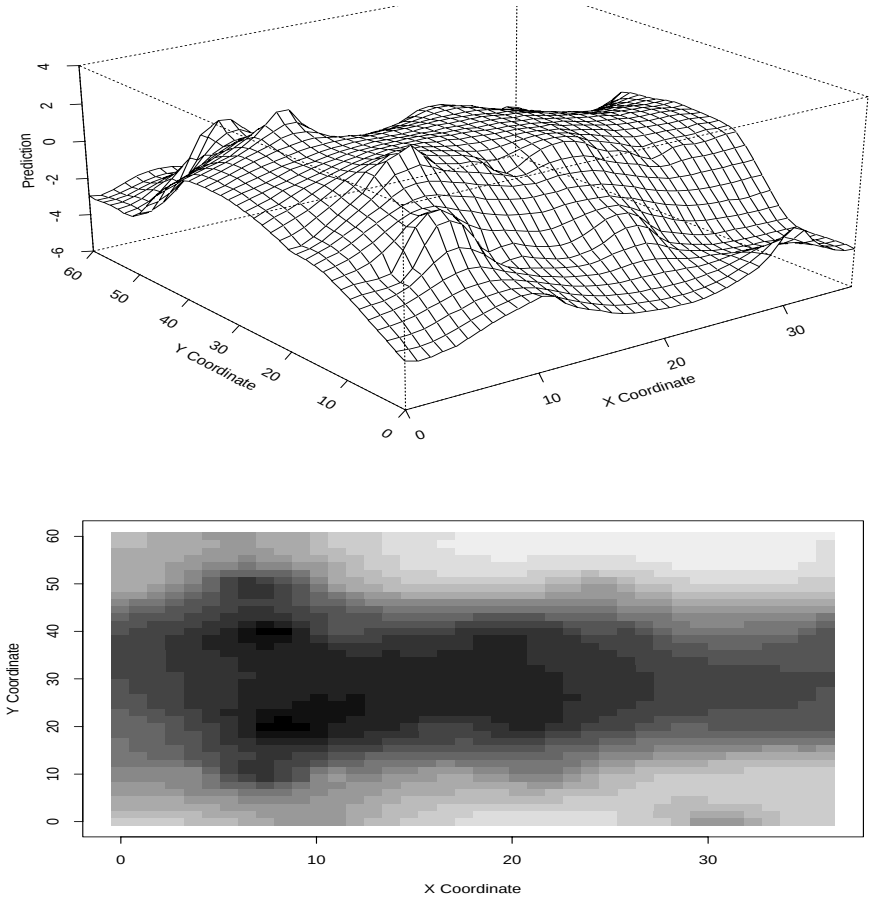

Figure 3: Dioxin: Posterior median of the Bayesian predictive distribution using data augmentation for censored values

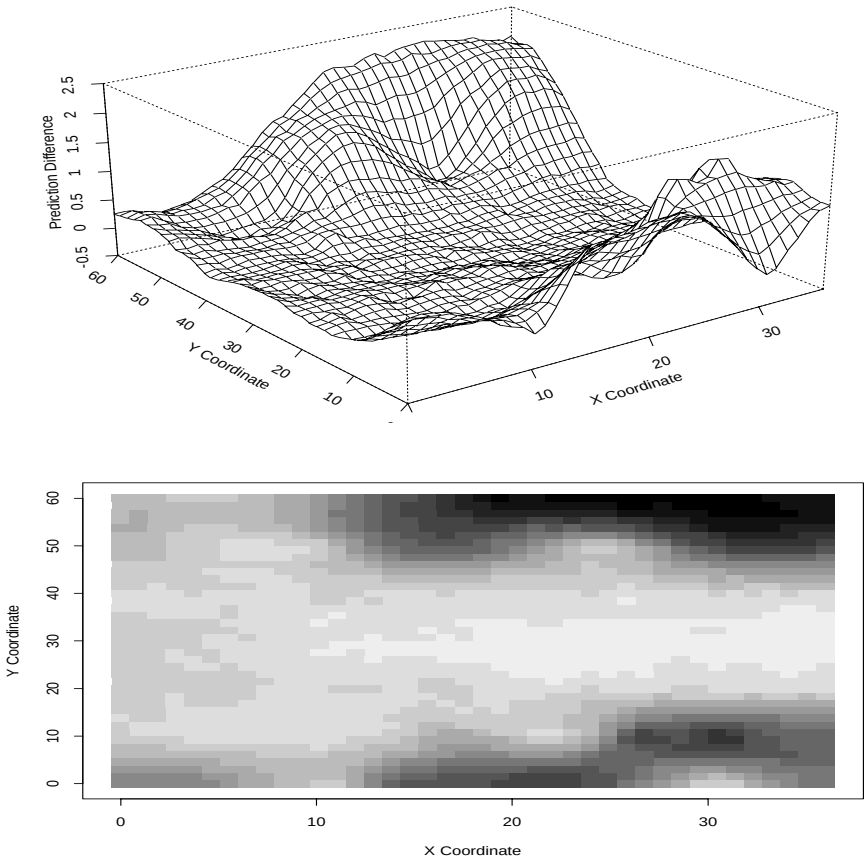

Figure 4: Dioxin: Difference in posterior medians for DA and LOD/2 methods for handling censored values (LOD/2 - DA)

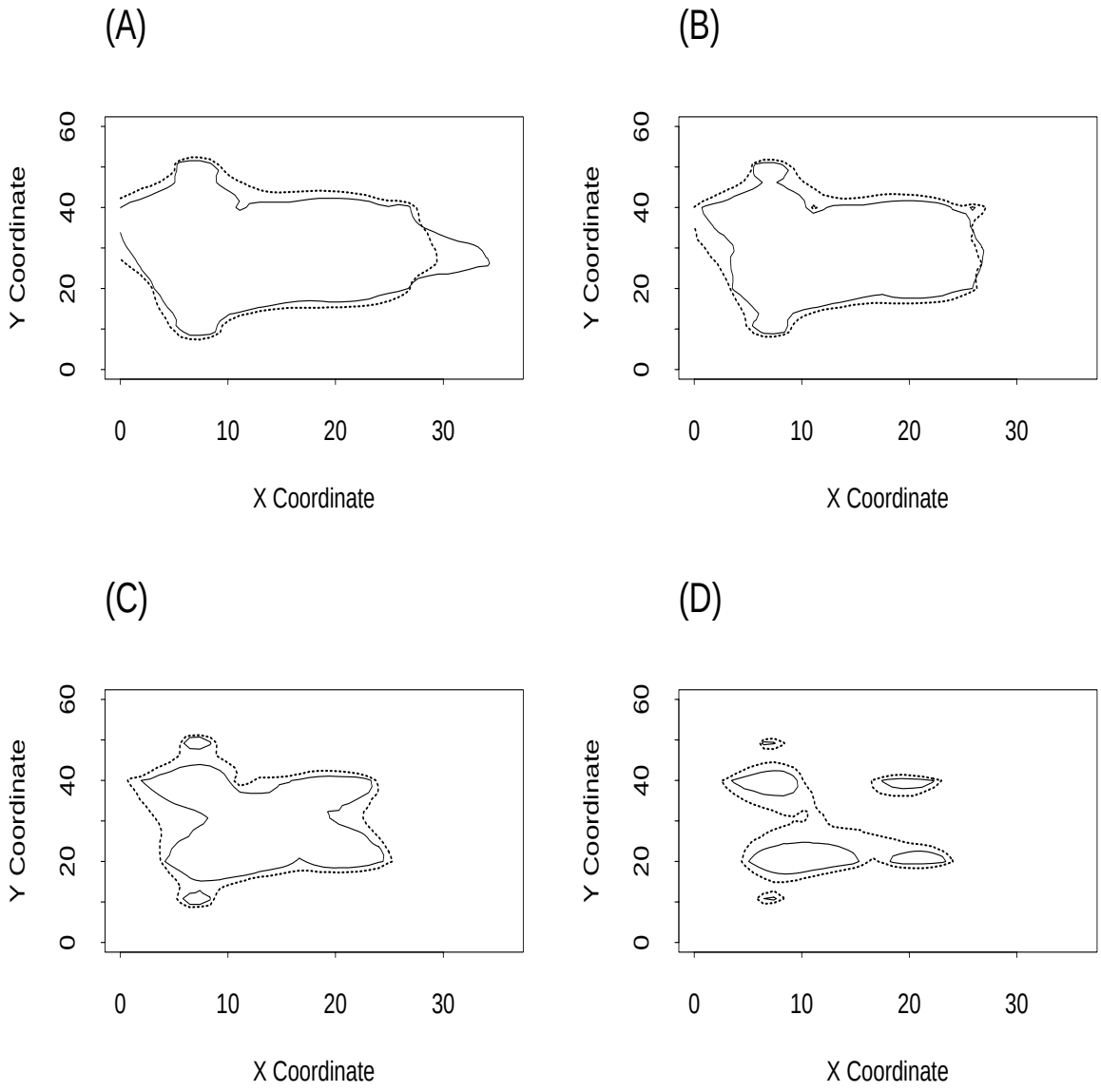

Figure 5: Dioxin: Difference in clean-up regions between the DA method (solid line) and the LOD/2 method (dashed line) based on different probabilities of being above clean-up cut-off values, (A) probability > 0.60, (B) probability > 0.70, (C) probability > 0.80, (D) probability > 0.90

Table 8: 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)