A comprehensive approach to analyzing environmental data with non-detects

Benjamin F. Trueman,* Madison Gouthro, Amina K. Stoddart, and Graham A. Gagnon

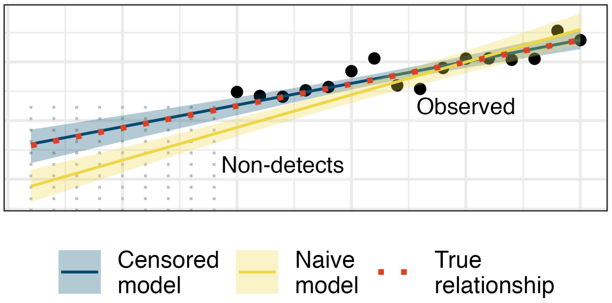
Centre for Water Resources Studies, Department of Civil & Resource Engineering,
Dalhousie University, 1360 Barrington St., Halifax, Nova Scotia, Canada B3H 4R2

*Corresponding author
E-mail: benjamin.trueman@dal.ca
Tel: 902.494.6070
Fax: 902.494.3105

Abstract

Non-detects—measurements reported as “below the detection limit”—are ubiquitous in environmental science and engineering. They are frequently replaced with a constant, but this biases estimates of means, regression slopes, and correlation coefficients. Omitting non-detects is worse, and has led to serious errors. Simple alternatives are available: rank-based statistics, maximum likelihood estimation, and re-purposed survival analysis routines. But many environmental datasets do not align well with the assumptions these methods make—it is often necessary to account for hierarchy (e.g., measurements nested within lakes), sampling strategy (e.g., measurements collected as time series), heterogeneity (e.g., site-dependent variance), and measurement error. Bayesian methods offer the flexibility to do this; incorporating non-detects is also easy and does not bias model parameter estimates as substitution does. Here we discuss Bayesian implementations of common bivariate and multivariate statistical methods relevant to environmental science. We use a dataset comprising time series of Ag, As, Cd, Ce, Co, Sb, Ti, U, and V concentrations in municipal biosolids that includes many non-detects. The models can be reproduced and extended to new problems using the data and code accompanying this paper.

Graphical abstract



Introduction

Non-detects—measurements recorded as less than a detection or reporting limit—are ubiquitous in environmental science and engineering. In the statistical literature, they are known as left-censored observations. A popular method of representing them in statistical routines is to replace them with one-half, or some other fraction, of the detection limit. But while common, this strategy can severely bias estimates based on the data when the censoring rate is high. Worse still is omission—leaving out non-detects has led to serious and well-documented errors.[1]

For basic tasks like comparing two groups or estimating a mean, linear regression slope, or correlation coefficient, there are simple alternatives to substitution and omission. These include rank-based methods, maximum likelihood estimation, and re-purposed survival analysis routines.[1] But many environmental datasets require more complex models that account for hierarchy, sampling strategy, heterogeneity, and measurement error. For instance, measurements may be collected across multiple lakes with different characteristics or they may be recorded as time series (i.e., serially dependent data). The standard toolbox for censored data analysis does not always accommodate these features.

Bayesian methods excel in this context—since the sampling techniques they rely on provide a near-universal approach to parameter estimation, they can be very flexible. In particular, it is straightforward to account for non-detects in almost any model. Here, we provide examples of common statistical models in environmental science and engineering whose Bayesian versions can easily accommodate non-detects. They are reproducible via the code and data that accompany this paper.

Materials and methods

Data collection

We fit models to a dataset comprising concentrations of Ag, As, Cd, Ce, Co, Sb, Ti, U, and V in municipal biosolids. Biosolids samples were collected from the clarifiers of three wastewater treatment facilities in 125 mL polypropylene bottles. Samples were autoclaved, desiccated by baking at 105°C for approximately 60 hours, digested according to EPA Method 3050B,[2] diluted serially, and quantified by inductively coupled plasma mass spectrometry according to Standard Method 3125.[3] A summary of the data is available in Table 1.

Table 1. *A summary of element concentrations in biosolids samples collected at three wastewater treatment facilities.*

Element	Median (μg g^{-1})	Lower quartile (μg g^{-1})	Upper quartile (μg g^{-1})	% censored
Ag	0.8	0.6	1.1	6.4
As	4.9	3.7	6.8	1.7
Cd	0.7	0.5	1.0	6.4
Ce	12.2	8.4	22.2	1.2
Co	2.3	1.6	3.1	1.7
Sb	0.2	0.1	0.2	77.5
Ti	30.1	17.3	47.1	1.2
U	0.7	0.5	1.0	9.2
V	4.4	3.2	5.9	1.7

Data analysis

The data and code necessary to reproduce the main results from this paper are available at https://github.com/bentrueman/censored-env-data-analysis; several functions used to fit the models in Stan are available in a separate R package.[4] We used R version 4.3.3 throughout,[5] along with a collection of contributed packages.[6–12]

Results and discussion

Bayesian modeling

Bayesian inference entails fitting a probability model to data, then summarizing it as the joint distribution of the model parameters, θ.[13] The model starts as a prior, $P(\theta)$, quantifying the plausibility of all possible parameter values. The prior reflects background knowledge and practical considerations.[14]

The data, x, are used to update the prior via Bayes' theorem. It relates the posterior or updated joint parameter distribution, $P(\theta|x)$, with the prior and the likelihood. The likelihood, $P(x|\theta)$, quantifies the compatibility of the data with the proposed model.

In practice, model fitting follows these basic steps:

1. Assign a probability distribution to the data.
2. Choose a model for each of the distributional parameters.
3. Choose a distribution of prior probability for each parameter.
4. Iterate the following steps to obtain a sample from the posterior:
   (a) Propose values for all parameters.
   (b) Quantify their plausibility without reference to the data (via the prior distributions).
   (c) Quantify the plausibility of each data point given the assumed data distribution and the proposed parameter values (i.e., the likelihood).
   (d) Obtain the relative posterior probability as the product of the likelihood and the prior (i.e., Bayes' theorem).

Iterating over steps 4 (a–d) may require searching a high-dimensional parameter space, which is often accomplished via the Hamiltonian Monte Carlo algorithm.[14] Fortunately, software packages make this straightforward: the R package *brms*[7], for instance, fits a huge variety of Bayesian regression models—including censored data models—with a standard and familiar syntax. Further customization is possible using Stan,[15] a programming platform for Bayesian statistics written in C++.

Substitution biases parameter estimates

Replacing non-detects with a constant can bias parameter estimates substantially, especially when the censoring rate is high. We show this using a small simulation study that compares substitution at one-half the detection limit with a parameter estimation strategy that relies on the cumulative distribution function.

When the dependent variable in a linear regression model includes left-censored observations, one method of accounting for them is to construct the likelihood for every censored observation using the appropriate cumulative distribution function in place of the probability density function. Here, the cumulative distribution function quantifies the probability that a data point is less than the detection limit—that is, the compatibility of a non-detect with the proposed model. The likelihood, then, becomes $P(x|\theta) = F(x_{observed}|\theta)G(x_{censored}|\theta)$, where F and G are the probability density and cumulative distribution functions, respectively.

We simulated from a simple linear regression model, $y_i \sim N(\mu_i = 3 + 0.15x_i, \sigma = 0.5)$, where the dependent variable was partially censored—here, N represents the normal distribution with mean μ_i and standard deviation σ. We fit a censored regression using *brms* where the simulated non-detects were modeled using the normal cumulative distribution function. We compared it to a naive model where the non-detects were replaced with one-half the detection limit.

Specifically, the i simulated observations, y_i, were modeled as follows. Except for the special handling of left-censored observations $(y_i|censored_i = 1)$, the naive model was identical.

$$
\begin{aligned}
&\text{likelihood:}\\
y_i|censored_i = 0 &\sim N(\mu_i, \sigma) \;\; [observed]\\
y_i|censored_i = 1 &\sim N\text{-}CDF(\mu_i, \sigma) \;\; [left\text{-}censored]\\[1em]
&\text{model for } \mu\text{:}\\
\mu_i &= \beta_0 + \beta_1 x_i\\[1em]
&\text{priors:}\\
\beta_j &\sim T\big(\mu_\beta = 0, \sigma_\beta = 2.5, \nu_\beta = 3\big), \text{for j} = 0{,}1\\
\sigma &\sim T(\mu_\sigma = 0, \sigma_\sigma = 2.5, \nu_\sigma = 3)
\end{aligned}
\tag{1}
$$

In equation (1), $censored_i$ is a binary variable $(0 = observed,\ 1 = left\text{-}censored)$, and $N\text{-}CDF$ is the normal cumulative distribution function (i.e., $P(X \leq x)$, the probability that a random variable X is less than or equal to some value x).[14,15] The parameters β_0 and

β_1 define the linear model for μ_i, and T denotes the t distribution with degrees of freedom—which controls probability density in the tails—parameterized by ν.

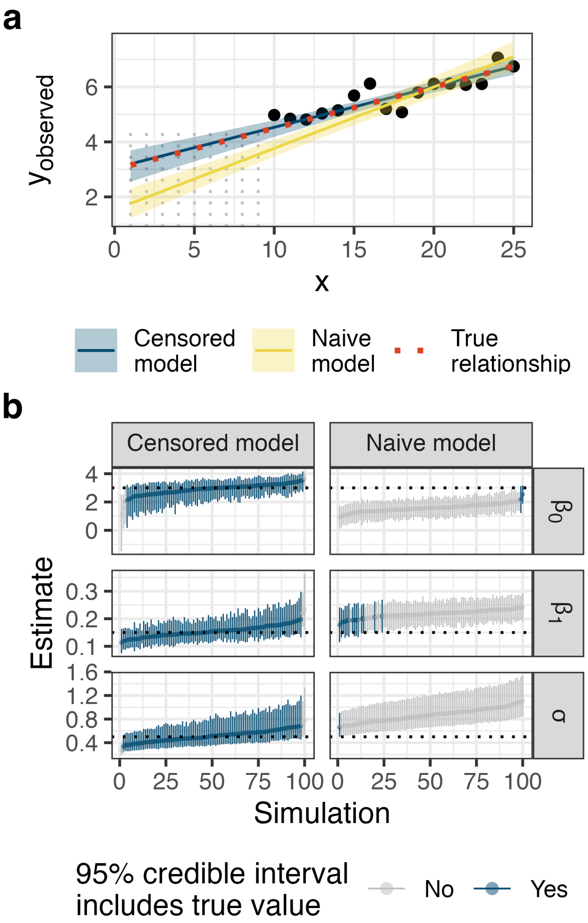


Figure 1. (a) *One iteration of the linear regression simulation. The model that accounts for left-censoring via the cumulative distribution function recovered the true model parameters well, whereas the naive model that used substitution at one-half the detection limit was biased. Points represent observations, and vertical dashed grey lines represent left-censored values.* ***(b)*** *The same pattern was evident across the entire simulation: the censored regression model recovered the true parameters well and the naive model was biased.*

The censored regression model recovered the true parameter values much more accurately than the naive model (Figure 1). That is, the censored model yielded 95% credible intervals on the intercept, slope, and residual standard deviation that included the true parameter values 96, 98, and 97% of the time, respectively. The naive model yielded intervals that included the true values just 2, 14, and 1% of the time.

Accounting for non-detects in a more complex model

The same strategy can be incorporated into more complex models: here, we use a dataset of metals concentrations in municipal biosolids to demonstrate fitting a smoothing spline, a popular method for characterizing environmental time series and other problems.[16–19] It was fitted using *bgamcar1*,[4] an extension of *brms* that accommodates continuous-time autoregression—accounting for the dependence of one observation on the previous one in unequally spaced time series.[16,20] Titanium concentrations at three wastewater treatment facilities (Sites 1–3) were modeled as follows:

$$\begin{aligned}
&\text{likelihood:}\\
&log([Ti]_t)|censored_t = 0 \sim T(\mu_t, \sigma, \nu)\ \ [observed]\\
&log([Ti]_t)|censored_t = 1 \sim T\text{-}CDF(\mu_t, \sigma, \nu)\ \ [left\text{-}censored]\\
\\
&\text{model for } \mu_t\text{:}\\
&\mu_t = \alpha + \beta_{site}X_{site} + f(t) + \phi^s r_{t-s}\\
&r_{t-s} = log([Ti]_{t-s}) - \alpha - f(t-s)\\
\\
&\text{priors:}\\
&\sigma \sim Half\text{-}T(\mu_\sigma = 0, \sigma_\sigma = 2.5, \nu_\sigma = 3)\\
&\beta_{site} \sim T\left(\mu_\beta = 0, \sigma_\beta = 2.5, \nu_\beta = 3\right)\\
&\nu \sim Gamma(\mu_\nu = 2, \alpha_\nu = 0.1)\\
&\phi \sim N\left(\mu_\phi = 0.5, \sigma_\phi = 0.5\right)\\
&\alpha \sim T(\mu_\alpha = 0, \sigma_\alpha = 2.5, \nu_\alpha = 3)
\end{aligned} \tag{2}$$

where, in addition to the symbols defined above, $Half\text{-}T$ represents the positive-valued t distribution and $Gamma$ the gamma distribution, parameterized by mean μ and shape parameter α. The linear model $\beta_{site}X_{site}$ estimates a separate intercept for each site, where X_{site} is the design matrix and β_{site} the coefficients. The autocorrelation coefficient, ϕ^s, and the residual at time $t - s$, r_{t-s}, define the dependence of each observation on the previous one, where s is the spacing between adjacent observations.[16,20]

The term $f(t)$ is a smooth spline function that captures nonlinear variation in the mean with time. It takes the following form:

$$f(t) = X_{spline}\beta_{spline} + Zb$$

priors on spline parameters:

$$(3)\quad \beta_{spline} \sim T\left(\mu_\beta = 0, \sigma_\beta = 2.5, \nu_\beta = 3\right)\ [unpenalized]$$
$$b \sim N(0, \sigma_b)\ [penalized]$$
$$\sigma_b \sim Half\text{-}T\left(\mu_{\sigma_b} = 0, \sigma_{\sigma_b} = 2.5, \nu_{\sigma_b} = 3\right)$$

where Z and X_{spline} are matrices representing the penalized and unpenalized basis functions, while β_{spline} and b represent the corresponding spline coefficient vectors.[21]

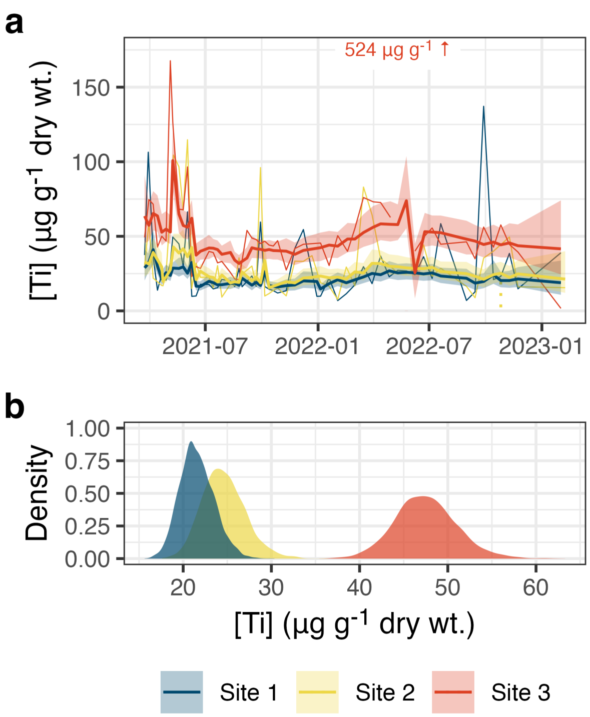


Figure 2. *Titanium concentration time series representing biosolids collected at three locations (light lines). Model predictions are superimposed in bold, the shaded regions represent a 95% credible interval on the posterior mean, and non-detects are shown as vertical dashed lines extending to the detection limit. A single value beyond the plot limits is annotated.*

Geometric mean titanium concentrations varied in a quasi-seasonal pattern (Figure 2), and samples collected at one facility, Site 3, had mean concentrations approximately 20–30 μg g^{-1} higher than those representing the other two facilities. Observations exhibited mild serial correlation, which quantifies the dependence of each observation on the previous one, after accounting for trends and site-specific variation. The serial correlation parameter in the model, ϕ, had a posterior median of 0.11 with a 95% credible interval spanning 0.03–0.26. In general, accounting for serial correlation improves the accuracy of predictions and helps avoid overfitting.[16,22]

A censored predictor

Left-censoring may also occur in the predictor variable. One potential application is building a linear regression model to predict missing values in one variable using another, partially censored, variable. Left-censored predictors, though, are not amenable to substituting a cumulative distribution function in the likelihood—another strategy is required.

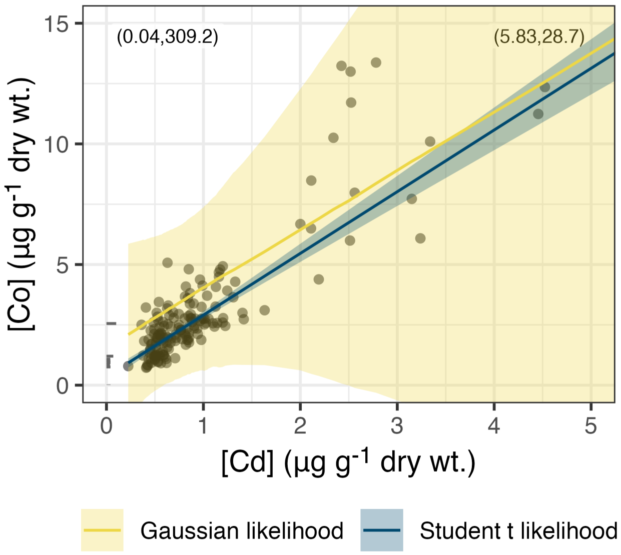


Figure 3. *Cobalt concentrations as a function of cadmium concentrations in biosolids from three treatment facilities. Nondetect cadmium concentrations are represented as horizontal lines extending to the detection limit. A robust (Student t) linear model is superimposed in blue and the shaded region represents a 95% credible interval on the posterior mean. The equivalent non-robust model yields an extremely wide credible interval due to the unusually high cobalt concentration of 309 µg g^{-1}. Coordinates outside the plotting limits are annotated in parentheses.*

Fortunately, there is a straightforward alternative: all the non-detects can be treated as missing values with an upper bound and represented by parameters in the model. Since Bayesian modeling results in a set of posterior draws—vectors of plausible parameter values—this is similar to multiple imputation of missing values with an upper—and optionally a lower—bound. But since it is done in one step, we get a joint distribution that quantifies the uncertainty and interrelationships among all the parameters, including the censored values.[23] This might be important, for instance, when there is serial dependence in the data.

Another advantage is the size of the imputed dataset: since the model is fitted to the data just once, it is straightforward to generate several thousand imputed values, even for complex models. This may not be practical if values are imputed before model fitting: the conventional imputation strategy entails fitting one model for each set of imputed values. Furthermore, it may be difficult to find a multiple imputation routine that meets all of the needs of a particular data analysis.

Using the Bayesian approach, we can specify any predictive model we would like to impute the censored values. Since there were only a few censored observations in this pair of variables (Figure 3), we chose a simple intercept-only imputation model, fitted using *bgamcar1*[4]. It was defined as follows:

$$
\begin{aligned}
&\text{likelihood:} \\
[Co]_i &\sim T\left(\mu_{[Co]_i}, \sigma_{[Co]}, \nu_{[Co]}\right) \\
[Cd]_i &\sim T\left(\mu_{[Cd]}, \sigma_{[Cd]}, \nu_{[Cd]}\right) \\
\\
&\text{model for } \mu: \\
\mu_{[Co]_i} &= \beta_{0_{[Co]}} + \beta_1 [Cd]_i \\
\mu_{[Cd]_i} &= \beta_{0_{[Cd]}} \\
\\
&\text{priors:} \\
\beta_j &\sim T\left(\mu_\beta = 0, \sigma_\beta = 2.5, \nu_\beta = 3\right), \text{ for j} = 0{,}1 \\
\sigma &\sim T(\mu_\sigma = 0, \sigma_\sigma = 2.5, \nu_\sigma = 3) \\
\nu &\sim Gamma(\mu_\nu = 2, \alpha_\nu = 0.1)
\end{aligned}
\tag{4}
$$

When the model was fitted with a Gaussian likelihood, the extreme cobalt concentration of 309 μg g^{-1} yielded a posterior mean with an extremely wide credible interval (Figure 3). The robust model, fitted with a Student t likelihood, yielded a much narrower credible interval and a posterior mean that was much less heavily influenced by the extreme value. A disadvantage of both models is that simulating from them may generate negative concentrations, even though the posterior mean remains positive over its range. This could be solved by modeling log-transformed Co concentrations instead, resulting in a slightly different interpretation: the model would then predict geometric mean concentrations on the scale of measurement.[24]

Multivariate models

In a multivariate context, the one-step multiple imputation strategy is often simpler to apply, since multivariate cumulative distribution functions can be difficult to implement. Two multivariate models with applications in environmental science are the intercept-only model, used to estimate a correlation matrix, and principal component analysis.

A Bayesian correlation matrix

In addition to handling non-detects, Bayesian correlation has the advantage that it can be readily applied to variables that are best described using distributions other than the Gaussian. A relevant example for environmental sciences is robust correlation, where a Student t distribution is assigned to each variable and its parameters estimated. This tends to limit the influence of extreme values, which might otherwise exert undue influence on the estimated correlation coefficients.

Given y, an $N \times D$ matrix containing N concentrations of D elements, we can estimate the pairwise correlations as follows:

$$\begin{gathered}\text{likelihood:}\\ y \sim MultivariateNormal\left(\begin{bmatrix}\mu_1\\ \dots\\ \mu_D\end{bmatrix}, \Sigma\right)\\ \Sigma = \begin{bmatrix}\sigma_1 & \dots & 0\\ \vdots & \ddots & \vdots\\ 0 & \dots & \sigma_D\end{bmatrix} R \begin{bmatrix}\sigma_1 & \dots & 0\\ \vdots & \ddots & \vdots\\ 0 & \dots & \sigma_D\end{bmatrix}\\ \\ \text{priors:}\\ \mu_j \sim T(\mu_\mu = 0, \sigma_\mu = 10, \nu_\mu = 3), \text{ for j} = 1, .., D\\ \sigma_j \sim T(\mu_\sigma = 0, \sigma_\sigma = 10, \nu_\sigma = 3), \text{ for j} = 1, .., D\\ R \sim LKJcorr(2)\end{gathered} \tag{5}$$

where the $\mu_{1 \dots D}$ are the column means of y, Σ is the covariance matrix, the $\sigma_{1 \dots D}$ are the column standard deviations of y, and R is the correlation matrix. $LKJcorr(2)$ is a regularizing prior that encodes mild skepticism of extreme correlation coefficients near -1 or 1.[14]

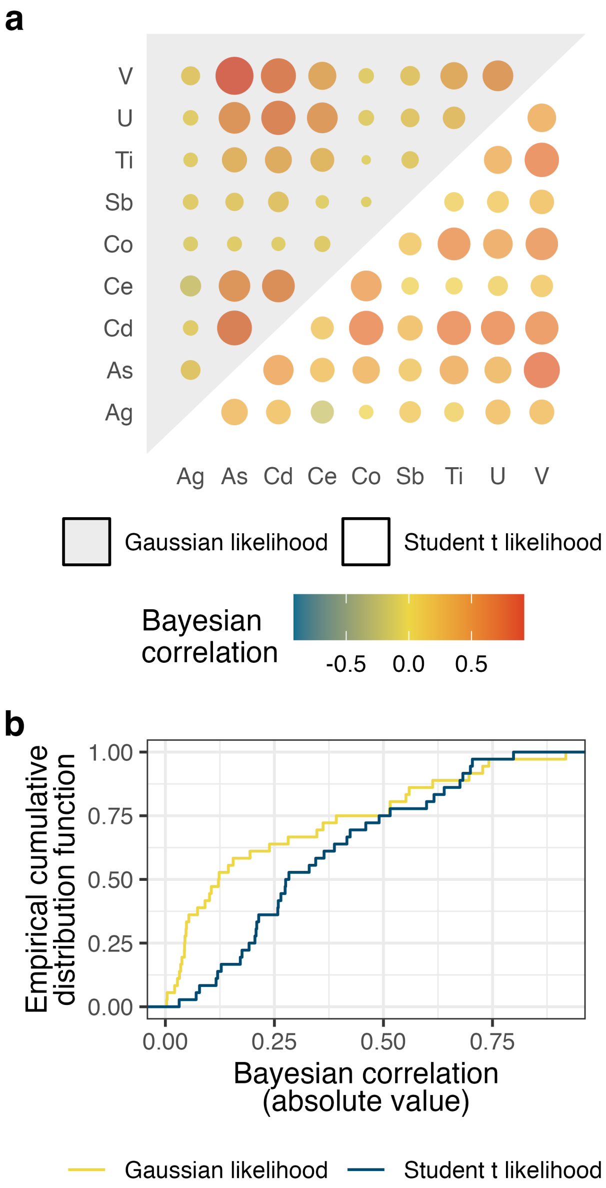


Figure 4. (a) *Pairwise Bayesian correlations among the elemental concentrations in the dataset, estimated using Student t and Gaussian likelihoods.* ***(b)*** *The robust model—fitted with Student t likelihoods—identifies more correlation than the non-robust model fitted with Gaussian likelihoods.*

Arsenic, vanadium, and cadmium concentrations were most strongly correlated in biosolids (Figure 4). And overall, the robust model—incorporating Student t likelihoods—identified more correlation among the variables than the conventional, non-robust model fitted with Gaussian likelihoods. This is due to the much smaller influence that extreme values have on the robust model.

Probabilistic principal components analysis

Principal component analysis is a method for summarizing a multivariate dataset using a subset of derived variables that capture the majority of the data's variance.[25] Here we implement probabilistic principal component analysis, a Bayesian generalization of the classical approach.[26] Our model is modified from the approach described in a recent paper[27] to accommodate left-censoring of the data and is written in Stan[8,15]. Given y, a $D \times N$ matrix containing N concentrations of D elements,

$$\begin{aligned} &likelihood: \\ &Y \sim MultivariateNormal(Wz + \mu, \sigma I) \\ \\ &priors: \\ &z \sim N(0, I) \\ &W \sim N(0, \sigma\alpha) \\ &\mu \sim Lognormal\left(\mu_{\mu} = 2.5, \sigma_{\mu} = 1\right) \\ &\sigma \sim Lognormal(\mu_{\sigma} = 0, \sigma_{\sigma} = 1) \\ &\alpha \sim Invgamma(\alpha_{\alpha} = 1, \beta_{\alpha} = 1) \end{aligned} \tag{6}$$

where z is a $k \times N$ matrix of latent (i.e., unobserved) variables with $k \leq D$, W is a $D \times k$ transformation matrix mapping from the latent space to the data space, σ is the standard deviation of the error (also a latent parameter), I is the identity matrix, and $InvGamma$ is the inverse gamma distribution.[13]

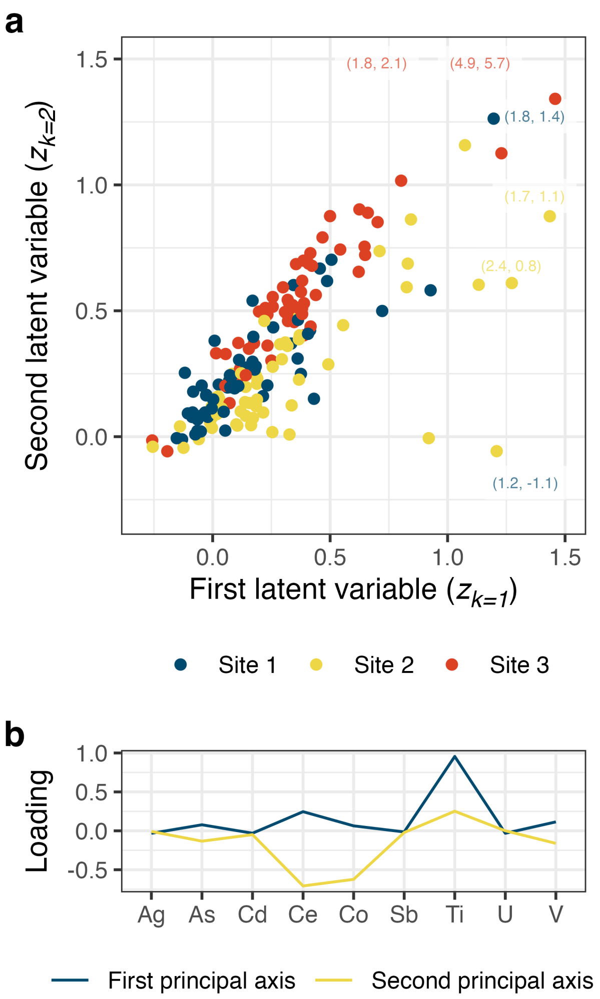


Figure 5. (a) *The dataset projected onto the first two probabilistic principal components (z in equation (6)). Values appearing outside the extents of the plot are annotated in parentheses at the margins.* ***(b)*** *The first two principal axes; that is, the orthonormalized columns of the transformation matrix W.*

Differences in metals concentrations among the three sites are apparent in Figure 5a. In particular, Site 3 scored differently on the two principal components, resulting in substantial separation in the two-component space from the data representing the other two sites. Differences in titanium concentrations among treatment facilities play a strong role here: the principal axes—that is, the orthonormalized columns of the transformation matrix W (equation 6)—load titanium concentrations heavily. And titanium concentrations were high in biosolids from Site 3 relative to Sites 1 and 2 (Figure 2).

Conclusion

Replacing non-detects with a constant—often one-half the detection limit—biases estimates of means, regression slopes, and correlation coefficients. Simple alternatives exist, but they are limited and not always applicable to complex environmental datasets that exhibit hierarchy, complex dependence structures, and heterogeneity. Bayesian methods have the flexibility to model all of these features, and they can easily accommodate left-censoring by either modifying the likelihood or one-step multiple imputation as a part of model fitting.

Acknowledgements

This work was supported by NSERC through an Alliance Grant (ALLRP 568507–21). We acknowledge utility staff for facilitating sample collection at the treatment facilities and Heather Daurie, Alyssa Chiasson, Evelyne Doré, Jorginea Bonang, Genevieve Erjavec, Sebastian Munoz, and Bofu Li for laboratory assistance.

References

(1) Helsel, D. R. *Statistics for Censored Environmental Data Using Minitab and R*, 2nd ed.; Wiley series in statistics in practice; Wiley: Hoboken, N.J, 2012.

(2) US Environmental Protection Agency. Method 3050B: Acid Digestion of Sediments, Sludges, and Soils. *Test methods for evaluating solid waste, physical/chemical methods* **1996**.

(3) American Public Health Association. 3125 Metals by Inductively Coupled Plasma—Mass Spectrometry. In *Standard methods for the examination of water and wastewater*; 2018. https://doi.org/10.2105/SMWW.2882.048.

(4) Trueman, B. *bgamcar1: Fit bayesian GAMs with CAR(1) errors to censored data*. https://github.com/bentrueman/bgamcar1.

(5) R Core Team. *R: A language and environment for statistical computing*. https://www.R-project.org/.

(6) Wickham, H.; Averick, M.; Bryan, J.; Chang, W.; McGowan, L. D.; François, R.; Grolemund, G.; Hayes, A.; Henry, L.; Hester, J.; Kuhn, M.; Pedersen, T. L.; Miller, E.; Bache, S. M.; Müller, K.; Ooms, J.; Robinson, D.; Seidel, D. P.; Spinu, V.; Takahashi,

K.; Vaughan, D.; Wilke, C.; Woo, K.; Yutani, H. Welcome to the tidyverse. *Journal of Open Source Software* **2019**, *4* (43), 1686. https://doi.org/10.21105/joss.01686.

(7) Bürkner, P.-C. *brms: Bayesian regression models using stan*. https://github.com/paul-buerkner/brms.

(8) Gabry, J.; Češnovar, R.; Johnson, A. *cmdstanr: R interface to CmdStan*. https://mc-stan.org/cmdstanr/.

(9) Fischetti, T. *assertr: Assertive programming for R analysis pipelines*. https://docs.ropensci.org/assertr/.

(10) Müller, K. *here: A simpler way to find your files*. https://here.r-lib.org/.

(11) Pedersen, T. L. *patchwork: The composer of plots*. https://patchwork.data-imaginist.com.

(12) Lawlor, J. *PNWColors: Color palettes inspired by nature in the US pacific northwest*. https://github.com/jakelawlor/PNWColors.

(13) Gelman, A.; Carlin, J. B.; Stern, H. S.; Dunson, B., David; Vehtari, A.; Rubin, D. B. *Bayesian Data Analysis*, Third edition.; Chapman & Hall/CRC texts in statistical science; CRC Press: Boca Raton, 2014.

(14) McElreath, R. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*; Chapman & Hall/CRC texts in statistical science series; CRC Press/Taylor & Francis Group: Boca Raton, 2016.

(15) Stan Development Team. *Stan modeling language users guide and reference manual, version 2.34*. http://mc-stan.org/.

(16) Simpson, G. L. Modelling Palaeoecological Time Series Using Generalised Additive Models. *Frontiers in Ecology and Evolution* **2018**, *6*, 149. https://doi.org/10.3389/fevo.2018.00149.

(17) Murphy, R. R.; Perry, E.; Harcum, J.; Keisman, J. A Generalized Additive Model Approach to Evaluating Water Quality: Chesapeake Bay Case Study. *Environmental Modelling & Software* **2019**, *118*, 1–13. https://doi.org/10.1016/j.envsoft.2019.03.027.

(18) Beck, M. W.; De Valpine, P.; Murphy, R.; Wren, I.; Chelsky, A.; Foley, M.; Senn, D. B. Multi-Scale Trend Analysis of Water Quality Using Error Propagation of Generalized Additive Models. *Science of The Total Environment* **2022**, *802*, 149927. https://doi.org/10.1016/j.scitotenv.2021.149927.

(19) Chen, T. Y.-J.; Guikema, S. D. Prediction of Water Main Failures with the Spatial Clustering of Breaks. *Reliability Engineering & System Safety* **2020**, *203*, 107108. https://doi.org/10.1016/j.ress.2020.107108.

(20) Trueman, B. F.; James, W.; Shu, T.; Doré, E.; Gagnon, G. A. Comparing Corrosion Control Treatments for Drinking Water Using a Robust Bayesian Generalized Additive Model. *ACS ES&T Engineering* **2022**, acsestengg.2c00194. https://doi.org/10.1021/acsestengg.2c00194.

(21) Wood, S. N. *Generalized Additive Models: An Introduction with R*, 2nd ed.; Chapman; Hall/CRC, 2017.

(22) Abokifa, A. A.; Sela, L. Integrating Spatial Clustering with Predictive Modeling of Pipe Failures in Water Distribution Systems. *Urban Water Journal* **2023**, *20* (4), 465–476. https://doi.org/10.1080/1573062X.2023.2180393.

(23) Hopke, P. K.; Liu, C.; Rubin, D. B. Multiple Imputation for Multivariate Data with Missing and Below-Threshold Measurements: Time-Series Concentrations of Pollutants in the Arctic. *Biometrics* **2001**, *57* (1), 22–33. https://doi.org/10.1111/j.0006-341X.2001.00022.x.

(24) Helsel, D. R.; Hirsch, R. M.; Ryberg, K. R.; Archfield, S. A.; Gilroy, E. J. *Statistical Methods in Water Resources*; U.S. Department of the Interior; U.S. Geological Survey, 2020; Vol. Techniques and Methods 4–A3.

(25) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer Series in Statistics; Springer New York: New York, NY, 2009. https://doi.org/10.1007/978-0-387-84858-7.

(26) Bishop, C. M. *Pattern Recognition and Machine Learning*; Information science and statistics; Springer: New York, 2006.

(27) Kucukelbir, A.; Tran, D.; Ranganath, R.; Gelman, A.; Blei, D. M. Automatic Differentiation Variational Inference. *Journal of Machine Learning Research* **2017**, *18* (14), 1–45.