Fitting an AR(1) Model to Environmental Measurements with Non-Detects

John Rogers
Westat, 1600 Research Blvd, Rockville, MD 20850

Abstract

For concentration measurements, the detection limit (DL) is a small concentration below which, due to measurement error, a measurement is not statistically different from zero. Non-detects are measurements reported as below the detection limit (e.g., < 5), resulting in censored data. In many situations, the measurement distribution can be described as the sum of a lognormal distributed concentration and a normally distributed measurement error. The data may be modeled by replacing the non-detects by substitute values, such as DL/2, or using survival analysis assuming a lognormal distribution. Environmental time-series can often be approximated by an AR(1) time-series model for the log-transformed concentrations. However, time series models do not account for non-detects. Bayesian procedures provide adequate flexibility to fit time-series models to data with non-detects, whether assuming the concentration distribution is lognormal or the sum of a normal and lognormal distribution. This paper illustrates a Bayesian approach to fit water pollution data with non-detects and provides simulation results comparing the Bayesian fit to some simpler approximations.

Key Words: Time series, Bayesian, survival analysis, measurement error

1. Introduction

Environmental time-series, such as pollutant concentrations in water treatment effluent, may be serially correlated and can often be approximated by an AR(1) time-series model for the log-transformed concentrations. However, some concentrations may be reported as less than the detection (or reporting) limit for the measurement process (i.e., non-detects, e.g., < 5), resulting in censored data. For establishing water treatment effluent limitations, EPA needs to model the distribution of the pollutant concentrations including estimating the serial correlation. This paper illustrates a Bayesian approach to estimate the serial correlation in water pollution data with non-detects and provides simulation results comparing the Bayesian fit to some simpler approximations.

Standard time-series and survival models do not estimate serial correlation adjusted for censoring. Naïve approaches such as replacing the non-detect by half the detection limit tend to give biased estimates of parameters. Several approaches to this problem have been suggested including full likelihood estimation (Zeger and Brookmeyer, 1986), imputation based on maximum likelihood (Park, Genton, and Ghosh, 2007), and multiple imputation based on a Bayesian approach (Hopke, Liu & Rubin, 2001). This paper investigates Bayesian options for fitting an AR(1) model to serially correlated concentration data with non-detects. All computation was done using SAS/STAT® software.

1.1 Serial Correlation

Assume N normally distributed observations (Y_i, such as log-transformed effluent concentrations) collected over time: (Y_i, T_i, i = 1 to N). The time difference between observations is: $\Delta_i = T_i - T_{i-1}$. Let $\emptyset$ (Phi) be the correlation between observations separated by one time unit. Then the AR(1) model, can be represented as:

$$\delta_0 \sim N(\mu, \sigma_Y^2);\ \Delta_1 = 1$$
$$\mu_i = \mu + \emptyset^{\Delta_i}\delta_{i-1}$$
$$Y_i \sim N\left(\mu_i, \sigma_Y^2\left(1 - \emptyset^{2\Delta_i}\right)\right)$$
$$\delta_i = Y_i - \mu.$$

Although more complicated correlation patterns can be modeled, the AR(1) model provides a good first order approximation for many concentration time series for which measurements close together in time are more similar than observations farther apart in time.

1.2 Measurement Error

Environmental concentration measurements (X_i) might be approximated by a log-normal distribution for the true concentrations with two additional measurement error components: 1) an error with constant CV; and 2) an error with constant variance. These errors can be modeled by m_i and e_i (both normally distributed) where:

$$X_i = exp(Y_i + m_i) + e_i.$$

Without duplicate or QC measurements, the variances of m_i and Y_i cannot be estimated independently. This paper assumes that either 1) the variance of m_i is ignorable, or 2) the serial correlation of $Y_i + m_i$ is to be estimated by an AR(1) model. Thus, assume:

$$X_i = exp(Y_i) + e_i,\ \ e_i \sim N(0, \sigma_e^2).$$

The distribution of X_i will be referred to as the Normal+Lognormal (NLN) distribution. The additive measurement error spreads out the lognormal distribution ($exp(Y_i)$), resulting in possibly negative concentration estimates for X_i, as illustrated in Figure 1.

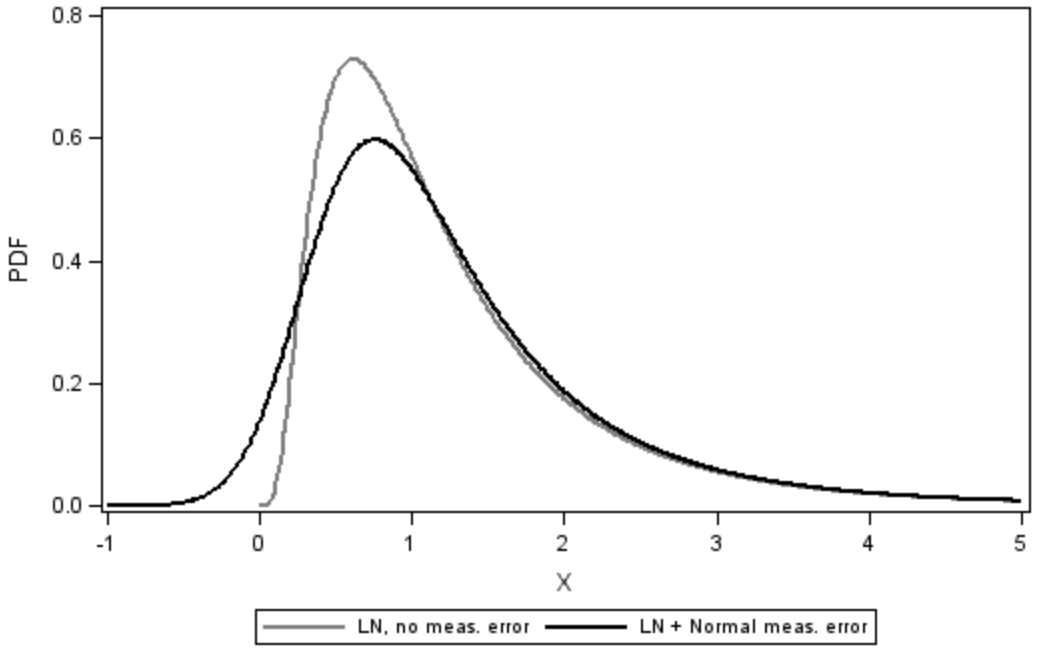

Figure 1: Example lognormal and NLN distribution

1.3 Detection Limits and Censoring by the Analytical Laboratory

The detection limit (DL) is a small concentration below which a measurement is not reliably different from zero. Detection limits may differ by sample. For some measurements, the detection limit is defined as M (perhaps 3 or 10) times the standard deviation of replicate QC measurements with zero or very low concentrations, i.e., $DL \cong M\sigma_e$. For each measurement, the lab reports if the concentration is censored (C_i) and reports a value, either the measured concentration or the sample specific detection limit (DL_i). The data available for analysis are: $C_i, D_i,$ and $T_i, i = 1\ to\ N$.

$$C_i = \begin{cases} 1 & if\ X_i < DL_i \\ 0 & Otherwise \end{cases}$$
$$D_i = \begin{cases} DL_i & if\ C_i = 1 \\ X_i & Otherwise. \end{cases}$$

1.4 Handling Non-Detects

When not using survival analysis, non-detects may be replaced by a substitute value (S_i) between zero and the detection limit, such as $DL_i/2$, $DL_i/\sqrt{2}$, or DL_i. The adjusted data are analyzed using standard statistical procedures. However, the analysis results may be sensitive to the choice of substitute value, particularly if there are many non-detects.

When using substitute values for non-detects, the data for analysis are: Z_i and $T_i, i = 1\ to\ N$ where:

$$Z_i = \begin{cases} S_i & if\ C_i = 1 \\ X_i & otherwise. \end{cases}$$

The choice of substitute value may depend on the distribution of the data. The substitute value can be described by a percentile (P) within the distribution of the non-detects. The

example in Figure 2 shows the distribution of the detected values in blue, the distribution of the non-detects in red, and P corresponding to various possible substitute values.

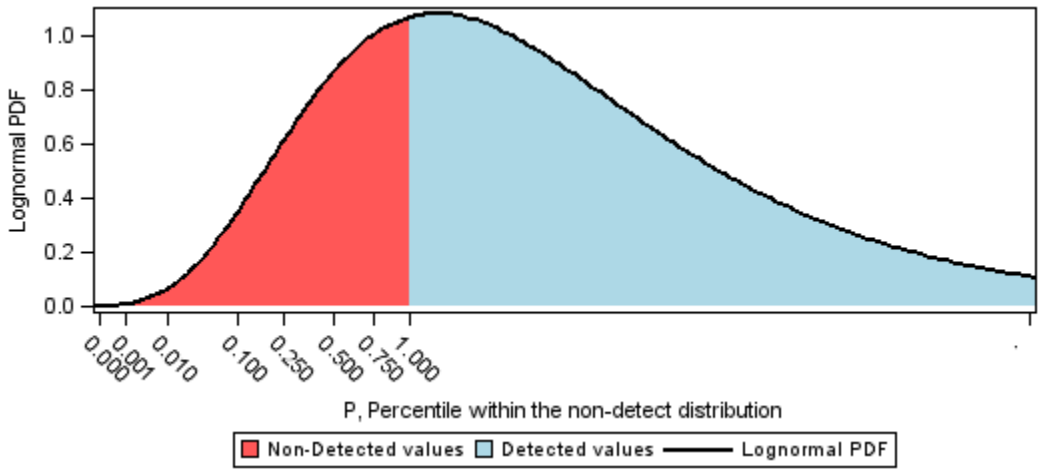

Figure 2: Example: Substitute value defined by percentiles within the non-detect distribution.

1.5 Serial Correlation Estimates Using Substitute Values

If a substitute value (defined by P) is to be used for non-detects, the best choice for P depends on the parameter being estimated. A roughly unbiased estimate of the mean log-transformed concentration can be achieved with P approximately equal to 0.45. An unbiased estimate of the standard deviation requires a lower P. When estimating the serial correlation, the estimation bias is relatively insensitive to the choice of P, as illustrated in Figure 3. The serial correlation estimates are generally lower than the true values, with increasing bias associated with high correlations and lower sample sizes. Figure 3 shows the mean estimate of serial correlation for N = 60 using 800 simulated data sets with ∅ equal to 0, .4, or .8 and 0%, 30%, or 60% non-detects. Note that, for N = 60, the maximum likelihood estimate of the serial correlation is biased low even when the true correlation is zero. This bias goes to zero as the sample size increases.

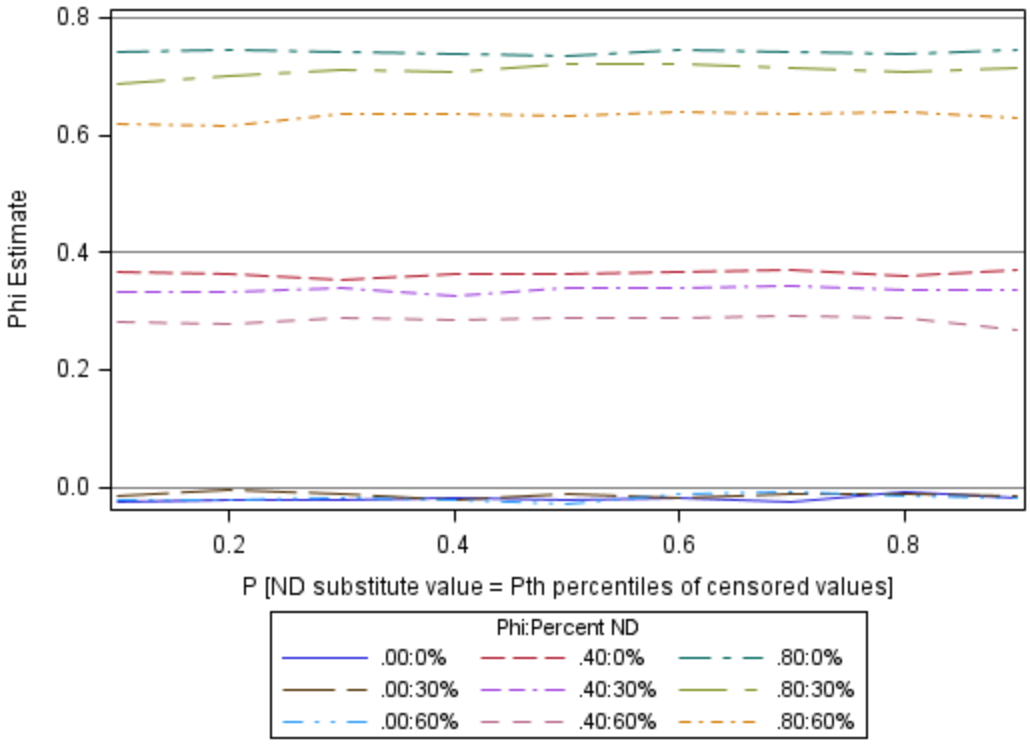

Figure 3: Mean phi estimate versus P, by true phi and percentage of non-detects (800 simulated data sets with N = 60 for each combination of parameters).

For the simulations above, substitute values were based on P and the known distribution of the data. Given field data, the parameters of the distribution must be estimated from the data.

2. Simulations

Simulations were used to evaluate five models for estimating serial correlation using seven different data distributions, three serial correlations, and three sample sizes. This section describes the models and parameters.

2.1 Models

Table 1 provides details of the five models, designated: ARIMA AR(1), BayesianNLN, BayesianNLN|DL, BayesianLN, and BayesianLN|P. The tilde notation defines the likelihood for the data.

Table 1: Models for fitting an AR(1) model to environmental measurements with non-detects

ARIMA AR (1)	
Use survival analysis to estimate the mean and standard deviation of Y_i assuming $\emptyset = 0$ (using the SAS LIFEREG procedure).	$\begin{cases} Y_i \sim N(\mu_N, \sigma_N^2) & C_i = 0 \\ C_i \sim Binary\left(\emptyset(Ln(DL), \mu_N, \sigma_N^2)\right) & C_i = 1 \end{cases}$
Replace non-detects by the estimated median (P=0.5) of the non-detects, creating Z_i.	$\begin{cases} Z_i = Y_i & C_i = 0 \\ Z_i = \Phi^{-1}(P * \Phi(Ln(DL_i), \mu_N, \sigma_N^2), \mu_N, \sigma_N^2) & C_i = 1 \end{cases}$
Using Z_i, estimate $\emptyset$ using time series models (using the SAS AUTOREG procedure).	$\delta_0 \sim N(\mu, \sigma_Z^2);\ \Delta_1 = 1$ $\mu_i = \mu + \emptyset^{\Delta_i}\delta_{i-1}$ $Z_i \sim N\left(\mu_i, \sigma_Y^2(1 - \emptyset^{2\Delta_i})\right)$ $\delta_i = Z_i - \mu$
BayesianNLN	
Fit the NLN distribution, estimating the parameters $(\mu, \sigma_Y^2, \emptyset, \sigma_e^2)$ and imputing δ_0, Y_i, and the non-detects. The model has an equation for Y_i, $X_i \vert Y_i$, and $C_i \vert X_i, Y_i$.	$\delta_0 \sim N(\mu, \sigma_Y^2);\ \Delta_1 = 1$ $\mu_i = \mu + \emptyset^{\Delta_i}\delta_{i-1}$ $Y_i \sim N\left(\mu_i, \sigma_Y^2(1 - \emptyset^{2\Delta_i})\right)$ $\delta_i = Y_i - \mu$ $X_i \vert Y_i \sim N(exp(Y_i), \sigma_e^2)$ $C_i \vert X_i, Y_i \sim Binary(X_i < DL_i)$
BayesianNLN|DL	
Assume the detection limit was defined as $DL = M\sigma_e$, removing one parameter from the BayesianNLN model.	Above with: $X_i \vert Y_i \sim N\left(exp(Y_i), \frac{DL_i}{M}\right)$
BayesianLN	
Assume a lognormal distribution, $\sigma_e = 0$, removing one parameter from the BayesianNLN model and removing the requirement to impute Y_i.	Above with: $X_i \vert Y_i = exp(Y_i)$
BayesianLN|P	
Assume a lognormal distribution, use the likelihood function from survival analysis, which does not require imputation of the non-detects. However, if the previous observation is a non-detect, use a substitute value (using P = .5) calculated using all of the parameters estimates.	$\delta_0 \sim N(\mu, \sigma_Y^2);\ \Delta_1 = 1$ $\mu_i = \mu + \emptyset^{\Delta_i}\delta_{i-1}$ $\sigma_{a_i}^2 = \sigma_Y^2(1 - \emptyset^{2\Delta_i})$ $P_{DL} = \Phi\left(Ln(DL_1), \mu_i, \sigma_{a_i}^2\right)$ $\begin{cases} Y_i \sim N\left(\mu_t, \sigma_{a_i}^2\right) & C_i = 0 \\ C_i \sim Binary(P_{DL}) & C_i = 1 \end{cases}$ $\begin{cases} \delta_i = Y_i - \mu & C_i = 0 \\ \delta_i = \Phi^{-1}\left(P * P_{DL}, \mu_i, \sigma_{a_i}^2\right) - \mu & C_i = 1 \end{cases}$

Note: the Bayesian models impute the missing values by modeling both the data and the mechanism creating the missing values. Bayesian models were fit using the SAS MCMC procedure.

2.1.1 Data Distributions

Data was simulated from seven distributions with either 0%, 30%, or 60% of the observations below the detection limit. Figures 4 and 5 show the distributions. All plots use the same scales.

Figure 4 shows the three lognormal distributions. The shaded portion of the distribution indicates the detected values. The models that assume the data has a lognormal distribution without additional measurement error (BayesianLN, BayesianLN|P, and ARIMA AR (1)) were fit using simulated data from these distributions.

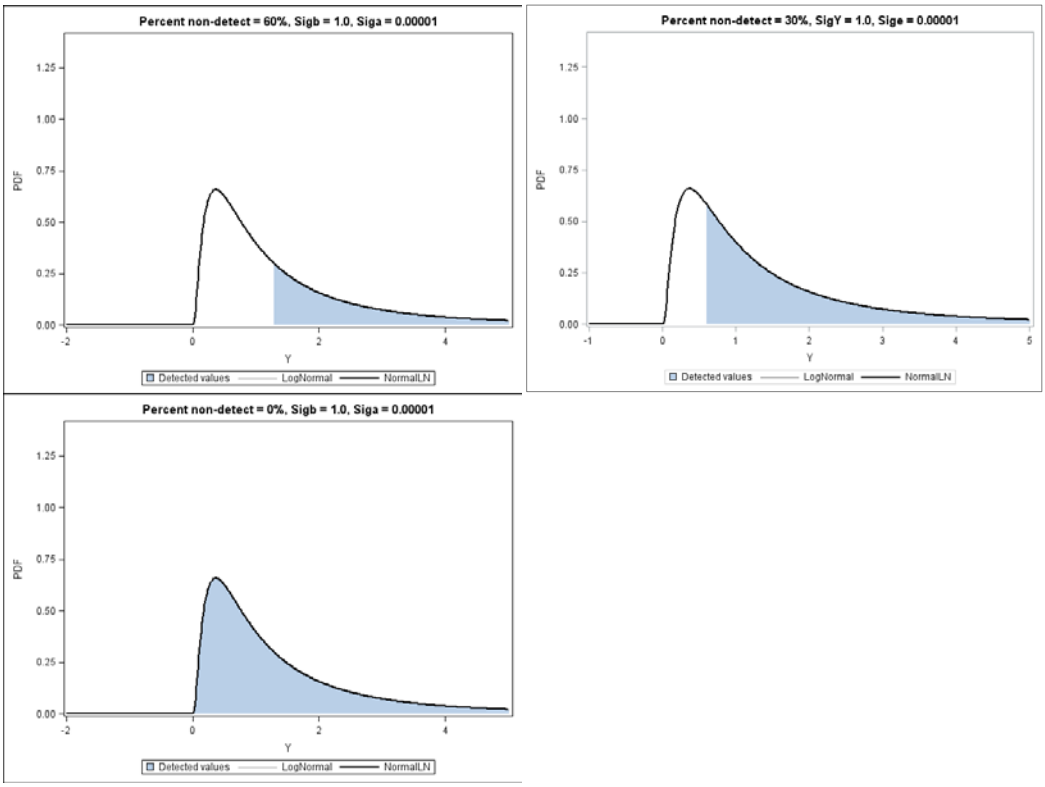

Figure 4: Lognormal distributions used for the simulations**.**

Figure 5 shows the four NLN distributions with either 30% or 60% non-detects. The shaded portion of the distributions indicates the detected values. The gray line shows the distribution of the lognormal data before adding normally distributed error. The NLN distributions were constructed so that the detection limit is three times the standard deviation of the normal additive error (M=3). All models were fit to data with these NLN distributions.

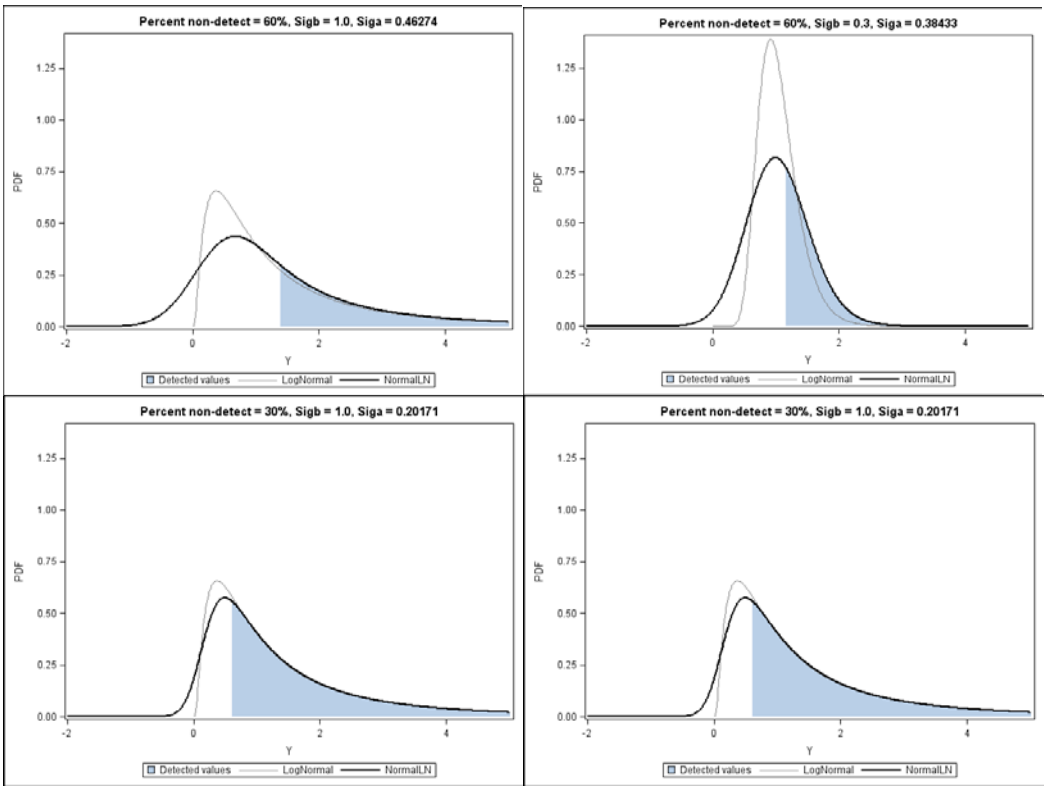

Figure 5: NLN distributions used for the simulations.

2.1.2 Simulation Parameters:

The simulations used three levels of serial correlation, Phi = 0, .4, and .8, and three different sample sizes (30 sequential observations, 100 sequential observations, and 50 observations spread across 100 time points, i.e., with half of the values missing at random).

2.2 Simulation Details:

The following provides some additional details regarding the simulations and Bayesian models:

- SAS macros were prepared to fit each model;
- MCMC parameters for variance are $Ln(\sigma)$;
- The BayesianNLN|DL model assumed $M = 3$, corresponding to the simulated data;
- Uninformative normal priors used for all parameters except for Phi which had a uniform [-1,1] prior;
- For each combination of simulation parameters, the number of simulations was set so that the mean effective sample size for posterior estimates was greater than 300;
- In rare cases the ARIMA AR (1) procedure did not converge or had numerical problems;
- In rare cases the BayesianNLN|DL procedure had numerical problems; and
- Probabilities of 0 and 1 were changed to 0.00001 and 0.99999 to avoid numerical problems.
- The number of simulations for each combination of simulation parameters was 160 for lognormal distributions and 56 for NLN distributions.

2.3 Outcome Measures

For analysis of bias, the estimates of Phi were transformed using Fisher's Z. Estimates in the Z scale were roughly normally distributed.

For each combination of the simulation parameters (model, distribution, percentage of non-detects, serial correlation, and sample size), the analysis of precision used the log transformed ratio of the pooled standard error of Phi to the standard deviation of Phi across simulations.

3. Simulation Results

On the Fisher's Z scale, the distribution of the serial correlation estimates is relatively normally distributed. For the results from the BayesianLN model, boxplots in Figure 6 illustrate the distribution of the estimated correlation for different values for N, the true serial correlation, and percentage of non-detects. The vertical axis scale uses Fisher's Z transformation with the labels showing the corresponding correlation. For N = 100 (the red boxplots) the serial correlation estimates are relatively unbiased. However, for N = 30, the serial correlation estimates are biased low when the correlation is high and the percentage of non-detects is high.

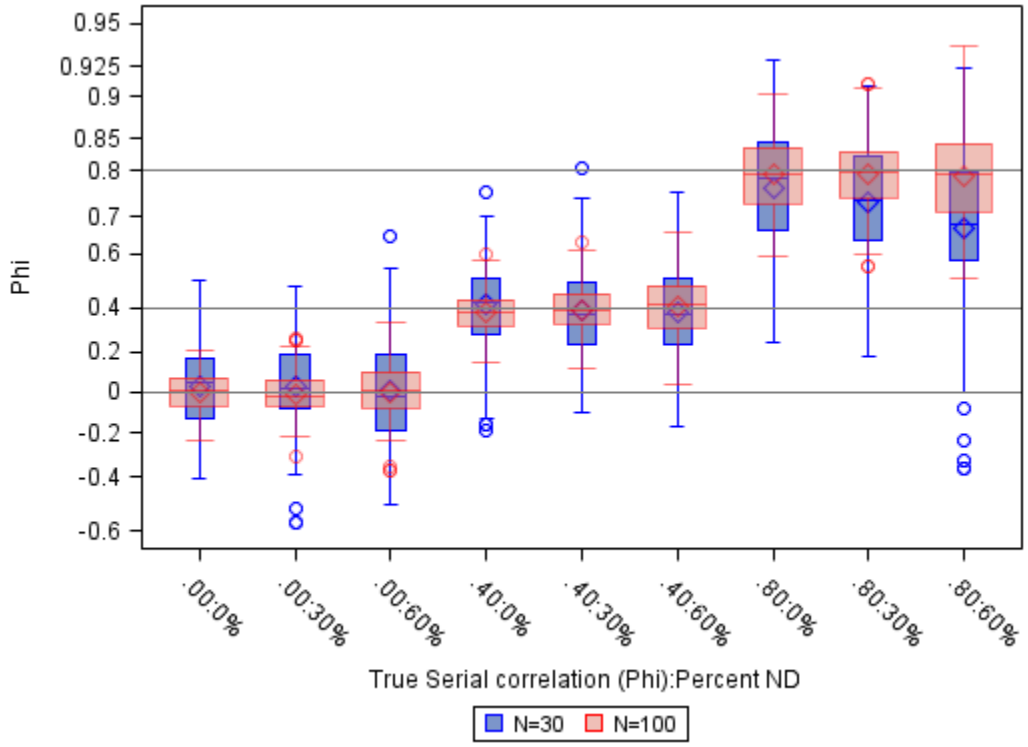

Figure 6: Boxplots illustrating the distribution of the estimated serial correlation when using the BayesianLN model.

3.1 Log-Normally Distributed Data

For each combination of simulation parameters (model, distribution, percentage of non-detects, serial correlation, and sample size), the following plot shows the geometric mean precision ratio on the vertical axis and the mean bias for Phi (calculated on the Fisher's Z

scale) on the horizontal axis. Separate plots are used for the lognormal and NLN distributions. Both plots use the same axis scales. Separate colors are used for each model. The ellipses help to show the distribution of the points for each model. For the preferred model, the ellipses would be centered close to zero bias on the horizontal axis and a ratio of 1.0 on the vertical axis and cover a small area.

Figure 7 shows the results for the models applies to lognormally distributed data. The BayesianLN and BayesianLN|P models performed better than the ARIMA AR (1) model, based on bias and precision.

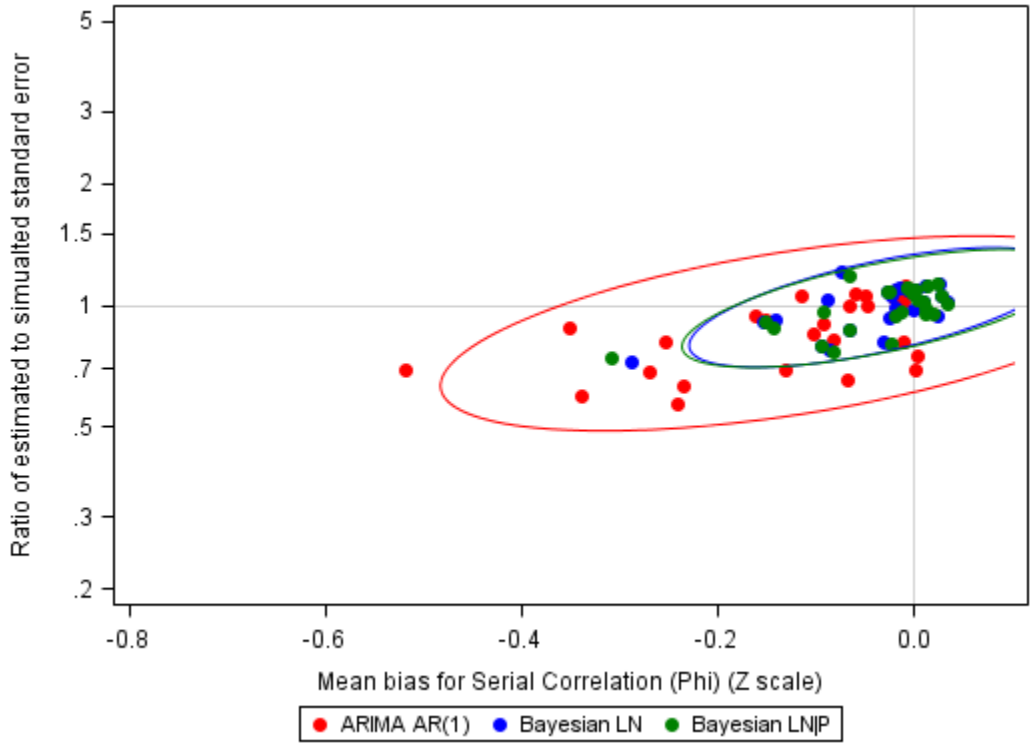

Figure 7: Bias and precision for serial correlation estimates for models applied to log-normally distributed data.

3.2 Normal + Log-Normal Data

Figure 8 show the results for the models applied to data with a NLN distribution. The BayesianLN and BayesianLN|P models performed better than the ARIMA AR (1) and BayesianNLN|DL models, based on bias and precision.

The BayesianNLN model, which estimates all parameters from the data, had some convergence problems when modeling data with no censoring and some negative concentrations and had serious convergence problems estimating the variance components for the censored data. As a result, results for the BayesianNLN model are not shown.

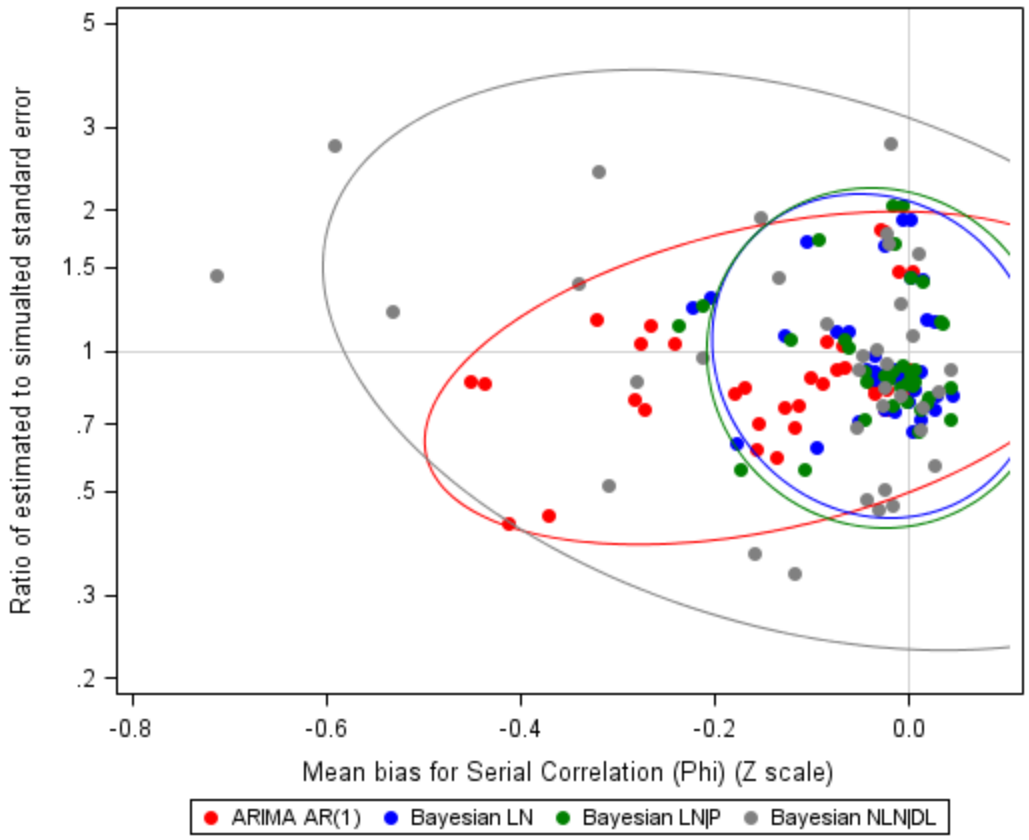

Figure 8: Bias and precision for serial correlation estimates for models applied to data with a NLN distribution.

4. Discussion and Conclusions

Environmental time-series, such as pollutant concentrations in water treatment effluent, may be serially correlated and can often be approximated by an AR(1) time-series model for the log-transformed concentrations. However, some concentrations may be reported as less than the detection (or reporting) limit for the measurement process (Non-detects, e.g., < 5), resulting in censored data. When estimating serial correlation in data series with non-detects, there are several possible analysis approaches. This paper looked at replacing the non-detects by a substitute value and using standard time series models versus using four different Bayesian models. Simulations were used to evaluate the performance of the different analysis approaches using data with either a lognormal distribution or a lognormal distribution with additional normally distributed measurement error (Normal+Lognormal (NLN) distribution). The simulations used different sample sizes, serial correlations, and percentages of non-detects. The Bayesian models assumed the data had either a lognormal distribution or an NLN distribution.

Across all analysis models and data distributions, serial correlation estimates tend to be biased low, with larger bias associated with high serial correlations, higher proportions of non-detects, and smaller sample sizes. Serial correlation estimates are reasonably normally distributed on the Fisher's Z scale. The presence of missing values appears to make the estimates more variable.

If a substitute value is used in conjunction with standard ARIMA models, the estimated serial correlation is relatively insensitive to how the substitute value is calculated. For the ARIMA AR(1) model, a survival model was run to find a substitute value that was consistent with the distribution of the data. The ARIMA procedure could be applied to data using a substitute value selected using another procedure. To the extent that the correlation estimates are insensitive to the choice of the substitute values, I would expect the results to be similar.

The BayesianNLN model assumed an NLN distribution for the concentration measurements, imputing both the true unknown concentration and, for the non-detects, the measured concentration. The BayesianNLN model had some convergence problems when modeling data with no censoring and had serious convergence problems estimating the variance components for censored data. However, for some measurement processes, the detection limit is defined as a multiple (M) of the standard deviation of the measurement error. The BayesianNLN|DL model used the known M and detection limits to fit the NLN distribution.

Although the assumptions behind the BayesianNLN|DL model correspond to the simulated NLN data, the BayesianNLN|DL model performed poorly compared to the simpler Bayesian models for some combinations of the simulation parameters. The poor performance of the BayesianNLN|DL model may be due to slower convergence and fixed limits on the number of Bayesian simulations.

The BayesianLN model assumed a lognormal distribution for the measured concentrations. The BayesianLN|P model also assumed a lognormal distribution for the measured concentrations; however, it used the likelihood from survival analysis and assumed a substitute value equal to the median of the distribution of the non-detects for estimating the serial correlation. The primary difference between the ARIMA AR (1) model and the BayesianLN|P model is that the substitute value in the BayesianLN|P model depends on the correlation in addition to the other parameters.

The BayesianLN|P and BayesianLN models perform similarly and have less bias than the BayesianNLN|DL and ARIMA AR(1) models even when the distribution of the data has additional measurement error not explicitly included in the model. The BayesianLN|P model converges substantially faster, making it preferable to the BayesianLN model.

Overall, for the simulated distributions, the BayesianLN|P procedure is preferred because it provides less biased or similar estimates as other approaches and requires less computing time. For all models, the estimates are biased low for higher serial correlations, higher proportions of non-detects, and smaller sample sizes.

References

G. E. P. Box and G. M. Jenkins. 1970. *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

P. Congdon. 2001. *Bayesian statistical modelling*. Chichester, England: John Wiley & Sons.

Gelman, et. al. 2004. *Bayesian data analysis*. Boca Raton, FL: Chapman & Hall/CRC. 2nd ed.

D. R. Helsel. 2005. *Nondetects and data analysis: statistics for censored environmental data*.. Hoboken, NJ: Wiley-Interscience.

P. K. Hopke, C. Liu, and D. B. Rubin. 2001. Multiple imputation for multivariate data with missing and below-threshold measurements: time-series concentrations of pollutants in the Arctic. In *Biometrics*. 57, 22–33.

J. W. Park, M. G. Genton, and S. K. Ghosh. 2007. Censored time series analysis with autoregressive moving average models. In *The Canadian Journal of Statistics*, Vol. 35, No. 1. 151–168.

S. L. Zeger and R. Brookmeyer. 1986. Regression analysis with censored autocorrelated data. In *Journal of the American Statistical Association*. 81, 722–729.