**Journal**: Ecology; **Article Type**: Statistical Innovations

# Accounting for missing data in autoregressive models of ecological time series

Alice E. Stears[*, 1, 2], Melissa DeSiervo[*, 3, 2], Dustin Gannon[4, 2], Amy Patterson[5, 2], Alice M. Carter[6,7], Joanna R. Blaszczak[8], Matt Trentman[9], Eliza Grames[10], Robert O. Hall, Jr[7], Joshua P. Jahner[11, 2], Saheed O. Jimoh[2], Courtenay A. Ray[2], Christa L. Torrens[7], Lauren Shoemaker[°2], Christopher Weiss-Lehman[°2]

**Author affiliations**

[1] Center for Adaptable Western Landscapes, Northern Arizona University, Flagstaff, AZ

[2] Botany Department, University of Wyoming, Laramie, WY

[3] Biology Department, Union College, Schenectady, NY

[4] Department of Forest Ecosystems and Society, Oregon State University, Corvallis, OR

[5] Department of Biology, University of Maryland, College Park, MD

[6] Department of Mathematics and Statistics, Utah State University, Logan, UT

[7] Flathead Lake Biological Station, University of Montana, Polson, MT

[8] Department of Natural Resources and Environmental Sciences, University of Nevada, Reno, Reno, NV

[9] O'Connor Center for the Rocky Mountain West, University of Montana, Missoula, MT

[10] Biological Sciences, Binghamton University, State University of New York, Binghamton, NY

[11] Department of Biology, New Mexico Institute of Mining and Technology, Socorro, NM

[*] Denotes equal contribution as lead author

[°] Denotes equal contribution as primary investigator

**Corresponding author**: Alice Stears, alice.e.stears@gmail.com

# Appendix S1

## Introducing Missingness

We created MCAR datasets with varying proportions of missing data and degrees of autocorrelation in missingness (Fig. S1 B–E) by viewing a time series as a Markov-modulated Bernoulli process where the variable could have two states: missing or not missing (Edwards, 1960; Gharib et al., 2014). The probability that an observation in a time series at time $t+1$ was missing depended on both the specified proportion of non-missing values in the entire time series ($p$) and the specified degree of autocorrelation in missingness ($\omega$). In a time series $X_1, X_2, ..., X_n$, the transition matrix that describes the probability of an observation at $X_{t+1}$ being missing, based on whether the observation at $X_t$ was missing is defined as:

$$\begin{array}{ccc} & X_{t+1} & \\ X_t & \text{Present} & \text{Missing} \\ \text{Present} & 1-(1-\omega)p & (1-\omega)p \\ \text{Missing} & (1-\omega)(1-p) & \omega+(1-\omega)p \end{array} \tag{1}$$

We created MNAR datasets with various proportions of msising data by first calculating the mean and standard deviation of the time series with no missing data, then using these point estimates as the mean and standard deviation of a normal distribution. We then identified the quantiles of that normal distribution above and below which the density of the normal distribution corresponded to the desired proportion of missingness. We replaced any values above and below those quantiles with an NA.

## Missing Data Approaches

**Simple and Complete Data Deletion**: The "simple data deletion" approach involves removing missing values from a time series, compressing the dataset, and running the model as if the time

intervals between observations were all equal (Fig. 1 A). This method violates the assumption of equal temporal spacing between observations, an assumption implicit in most time series models. We include it here as a reference because it is simple and commonly used in published studies. We also include "complete case data deletion," which maintains equal spacing between observations by removing a missing value itself as well as the subsequent observation(s) that is predicted by the missing value (Fig. 1 A). However, those observations after a missing value are retained as predictors of the subsequent observation(s).

**Multiple Imputation**: Multiple imputation (MI) is an approach that systematically fills in missing observations with imputed values, and creates several versions of complete data sets that can be used to estimate uncertainty around each imputed value (Fig. 1 B). MI is commonly used in ecology, with multiple studies evaluating methods and approaches to conduct MI for functional traits (Johnson et al., 2021; Penone et al., 2014; Taugourdeau et al., 2014), population biology (Onkelinx et al., 2017), time series (Hui et al., 2004), and meta-analyses (Ellington et al., 2015). Multiple Imputation's (MI) effectiveness can depend on the number of imputed datasets (*m*). It is often assumed that *m*=5 is a minimum value (Honaker & King, 2010); however, researchers have used *m*=200 when comparing methods in the ecological sciences (Onkelinx et al., 2017). In general, larger values of *m* result in more accurate estimates of both parameter values and uncertainty. However, increasing *m* results in a trade-off between accuracy and computation time; this can be particularly problematic for data-rich (e.g., long time series) or complex (e.g., hierarchical) models. After imputing the *m* data sets, the analyses of interest are confronted with each data set, and the estimated parameters from the *m* analyses are averaged using Rubin rules of averaging to get the parameter(s), and associated uncertainty, from which inference can be made. We implemented multiple imputation with the Amelia II package in R (Honaker et al., 2011), which uses an expectation maximization algorithm (see below) in combination with a bootstrapping technique for deciding what values to impute. We used $m = 5$ in order to provide decent estimates without excessive run times.

For both the simulated and empirical population count time series, since we did not have any

covariates, the only variables used for imputation were the population size at time *t* and population size at time *t-1*. For time series with chunks of missing data, the Amelia multiple imputation function had to be run iteratively, with missing values filled in from the edges of the missing chunks. In addition, while the recommended settings for dealing with time series data using the Amelia package include incorporating preceding and proceeding time points by specifying the "lags" and "leads" options (Honaker et al., 2011), it was not possible to use the lags option, since the current population was already using the population at the previous point as its only predictor for imputation. Instead, we included only the leads option, which still resulted in occasional failure of the method at extremely high levels ($> 70\%$) of missing data due to excessive collinearity between the preceding and proceeding time points. The lack of error handling for extremely collinear variables is an unfortunate issue for this method when using data sets without covariates, or data sets with highly collinear covariates.

For both the simulated and empirical Gaussian series, implementing multiple imputation was more straightforward since an observation at time *t* was informed by two covariates in addition to the observation at the time *t-1*. In this case, we were able to use both the "lags" and "leads" options in `amelia`.

Following execution of MI using Amelia II, we fit statistical models to time series following the methods described in **Main Text: Comparing missing data approaches**.

**Kalman Filter**: The Kalman Filter (KF) was developed to estimate the state of a dynamic system that is observed with error but can be used to derive the likelihood function of a time series with missing observations (Fig. 1 C). To illustrate the approach, assume a state-space model

$$\begin{aligned} X_t &= \phi X_{t-1} + \epsilon_t \\ Y_t &= X_t + e_t \end{aligned} \tag{2}$$

where $X_t$ is the true "state" of the system at time $t$, $Y_t$ is the observed value at time $t$, and $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ and $e_t \sim \mathcal{N}(0, \tau^2)$ are IID white noise error terms for the process and observation error, respectively. The Kalman Filter is primarily focused on estimating the unobserved state of the

system, $X_t$, and can be conceptualized as a two-step procedure in which, given an initial state $X_0$, we can forecast the next state $X_1$. Then, following data collection at the next time point, $y_1$, we update the forecast using Bayes' theorem. Specifically, the forecast distribution for $X_1$ is

$$p(x_1) = \int p(x_1|x_0)p(x_0)dx_0 \tag{3}$$

where $p(\cdot)$ denotes the probability density function. Assuming IID Gaussian errors, $p(x_1)$ is normal with mean $\tilde{x}_1 = \phi x_0$ and variance $v_1 = \phi^2 \frac{\sigma^2}{1-\phi^2} + \sigma^2$. Given the observed value $y_1$, we update the estimate of $X_1$ using Bayes theorem

$$p(x_1|y_1) \propto p(y_1|x_1)p(x_1) \quad = \mathcal{N}\Big(\tilde{x}_1 + K_1(y_1 - \tilde{x}_1),\ (1 - K_1)v_1\Big) \tag{4}$$

where $K_1 = v_1/(v_1 + \tau^2)$ is the ***Kalman gain*** and creates a weighted average of the forecast and observation. For our focus on missing data, we assume the process is observed without error such that $Y_t = X_t$ and $\tau^2 = 0$. Without observation error, the Kalman gain $K_t = 1$ for all $t$ since $\tau^2 = 0$, and $p(x_1|y_1) = \mathcal{N}(y_t, 0)$. Thus, the update step gives complete information about $X_t$, and the likelihood function can be defined based on the data $y_1, ..., y_n$. However, if data are missing, the update step cannot occur. So, in the case of missing data without observation error, the Kalman Filter alternates between pure forecast steps when data are missing and pure "update" steps when data and the state of the system are completely observed, but the forecast steps yield a method for computing the likelihood function recursively without needing to know the states of $X_t$ in which we were unable to observe the process and therefore have no associated $y_t$.

The Kalman filter assumes a Gaussian error distribution, so we only used this method with the simulated and empirical real-valued time series. We implemented KF missing data approach at the same time as the model fitting process, where we fit an AR(1) model with two covariates using the `arima` function from the `stats` package in R (R Core Team, 2021) (KF is the default algorithm used to handle missing values in this R function).

**Expectation Maximization**: The expectation maximization (EM) algorithm is an iterative

algorithm that is conceptually similar to KF, and recursively computes the likelihood of a time series with missing data (Fig. 1 D). Given an initial guess for the parameter vector we wish to estimate, $\boldsymbol{\theta}_0$, the first step (Expectation step) proceeds to "fill in" the missing observations with their expectation given the observed data and the initial parameter vector $\boldsymbol{\theta}_0$ , which is equivalent to the forecast step of the Kalman filter conditioned on $\boldsymbol{\theta}_0$. In the second step (maximization step), we compute the maximum likelihood estimate of $\boldsymbol{\theta}$ using the filled-in time series as data to give an updated estimate $\hat{\boldsymbol{\theta}}_1$. We then iterate this process, updating the forecasts of the missing data using their expectations conditional on $\hat{\boldsymbol{\theta}}_1$, then maximizing the likelihood with respect to $\boldsymbol{\theta}$ using the time series filled-in with the updated forecasts. This process is iterated until the difference between successive estimates is acceptably small, indicating convergence (that is, $||\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_{i-1}||_1 < \delta$ for some small $\delta > 0$).

Given its similarity to KF, we only used this missing data approach for the simulated and empirical times series of counts. We constructed an approximate EM algorithm to estimate the parameters of the Ricker model in which missing data were rounded to the nearest integer value during the expectation step such that the likelihood was well-defined for the filled-in series. As such, the missing data were dealt with at the same time as the model fitting process. We used the `optim` function from the `stats` package in `R` for the maximization step (R Core Team, 2021). Note that the algorithms required to estimate standard error of parameter estimates generated from EM are notoriously unstable and difficult to implement in R, so the results of models we fit using this missing data approach do not include standard error estimates.

**Data Augmentation**: Data augmentation (DA) provides a model-based framework for estimating missing observations as well as the parameters of interest, but comes with the added benefit of standard errors for the estimates of all the unknown quantities by treating the missing observations as additional parameters to be estimated (Fig. 1 D). We fit the Gaussian AR(1) models with DA and Stan (Carpenter et al., 2017) by using the rstan (Stan Development Team, 2024) and brms (Bürkner, 2017) packages in R (R Core Team, 2021). Data augmentation for the population model is not possible with Stan, however, due to the requirement of continuous parameter space for the

Hamiltonian Monte Carlo (HMC) methods Stan uses to sample the posterior distribution (at least not without marginalizing out the discrete parameters, which proved intractable). Treating missing integer data as parameters was therefore not possible with Stan, and partially-known parameter vectors are not supported in JAGS. We therefore designed a Gibbs sampler with Metropolis updates for the log growth factor ($r$) and intra-specific competition coefficient ($\alpha$), and Gibbs sampling of any missing observations, $N_t^{(0)}$, conditional on $(r, \alpha, \mathbf{N}_{t-})$, where $\mathbf{N}_{t-}$ is the vector of abundances (both observed and unobserved) up to but not including time $t$. We used weakly informative Gaussian priors for $r$ and $\ln(\alpha)$ and fit the models using custom functions written in R.

# References


Bürkner, P.-C. (2017). Brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, *80*, 1–28.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Edwards, A. W. F. (1960). The meaning of binomial distribution. *Nature*, *186*. https://doi.org/10.1038/1861074a0

Ellington, E. H., Bastille-Rousseau, G., Austin, C., Landolt, K. N., Pond, B. A., Rees, E. E., Robar, N., & Murray, D. L. (2015). Using multiple imputation to estimate missing data in meta-regression. *Methods in Ecology and Evolution*, *6*(2), 153–163. https://doi.org/10.1111/2041-210X.12322

Gharib, M., Ramadan, M. M., & Al-Ajmi, K. A. (2014). Characterization of markov-bernoulli geometric distribution related to random sums. *Journal of Mathematics and Statistics*, *10*, 186–191. https://doi.org/10.3844/jmssp.2014.186.191

Honaker, J., & King, G. (2010). What to do about missing values in time-series cross-section data. *American Journal of Political Science*, *54*(2), 561–581. https://doi.org/10.1111/j.1540-5907.2010.00447.x

Honaker, J., King, G., & Blackwell, M. (2011). Amelia ii: A program for missing data. *Journal of Statistical Software*, *45*, 1–47.

Hui, D., Wan, S., Su, B., Katul, G., Monson, R., & Luo, Y. (2004). Gap-filling missing data in eddy covariance measurements using multiple imputation (MI) for annual estimations. *Agricultural and Forest Meteorology*, *121*(1), 93–111. https://doi.org/10.1016/S0168-1923(03)00158-8

Johnson, T. F., Isaac, N. J. B., Paviolo, A., & González-Suárez, M. (2021). Handling missing values in trait data. *Global Ecology and Biogeography*, *30*(1), 51–62. https://doi.org/10.1111/geb.13185

Onkelinx, T., Devos, K., & Quataert, P. (2017). Working with population totals in the presence of missing data comparing imputation methods in terms of bias and precision. *Journal of Ornithology*, *158*(2), 603–615. https://doi.org/10.1007/s10336-016-1404-9

Penone, C., Davidson, A. D., Shoemaker, K. T., Marco, M. D., Rondinini, C., Brooks, T. M., Young, B. E., Graham, C. H., & Costa, G. C. (2014). Imputation of missing data in life-history trait datasets: Which approach performs the best? *Methods in Ecology and Evolution*, *5*(9), 961–970. https://doi.org/10.1111/2041-210X.12232

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Stan Development Team. (2024). RStan: The R interface to Stan [R package version 2.32.6]. https://mc-stan.org/

Taugourdeau, S., Villerd, J., Plantureux, S., Huguenin-Elie, O., & Amiaud, B. (2014). Filling the gap in functional trait databases: Use of ecological hypotheses to replace missing data. *Ecology and Evolution*, *4*(7), 944–958. https://doi.org/10.1002/ece3.989

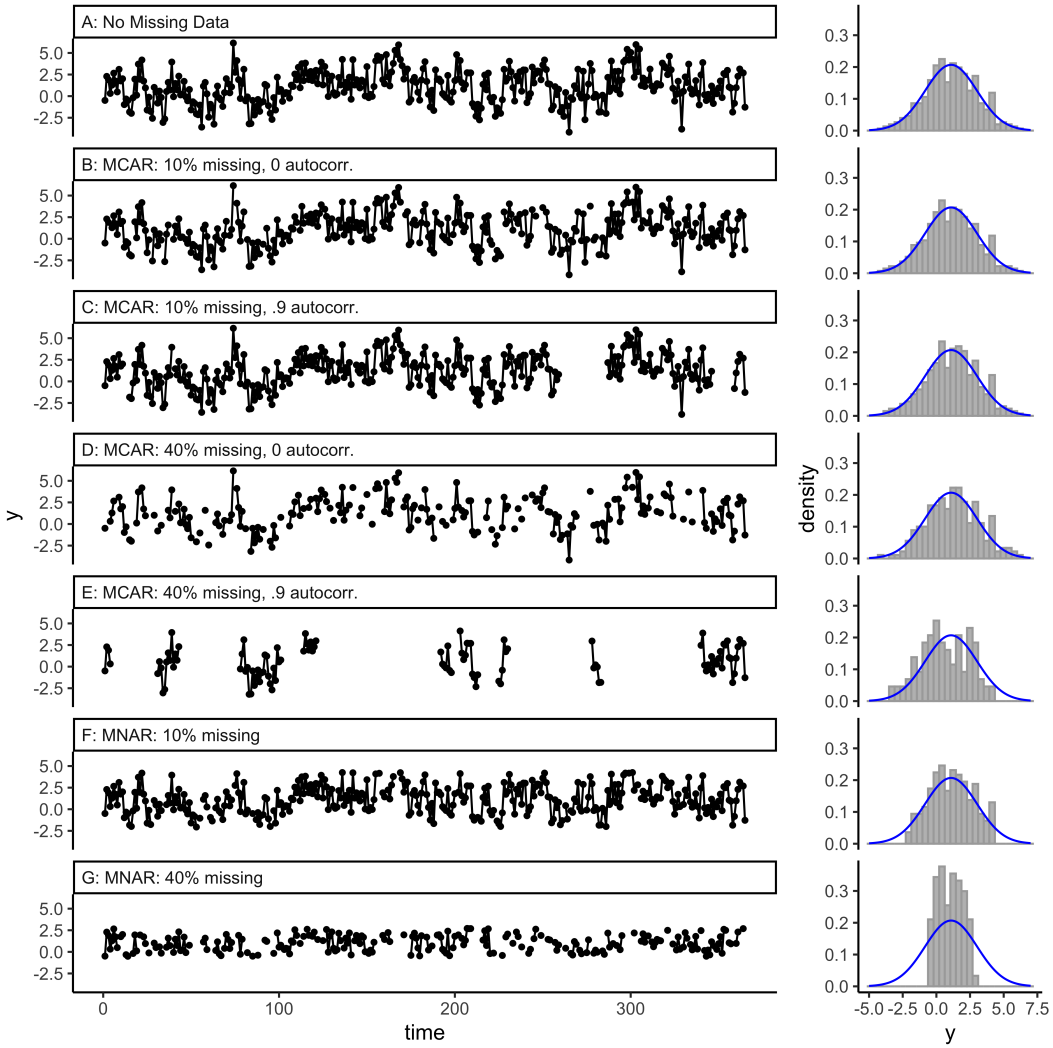


Figure S1: An example time series demonstrating different types and amounts of missing data. The left column shows the same time series with different amounts and types of missingness and right column shows the distribution of data points in each resultant time series. A. Complete time series with no missing data. Rows B through E show the time series with 10% (B and C) or 40% (D and E) of data missing completely at random (MCAR), with either low autocorrelation in missing data (B and D) or high autocorrelation (C and E). Rows F and G show the time series with data missing not at random (MNAR) for 10% missing data (F) and 40% missing data (G).

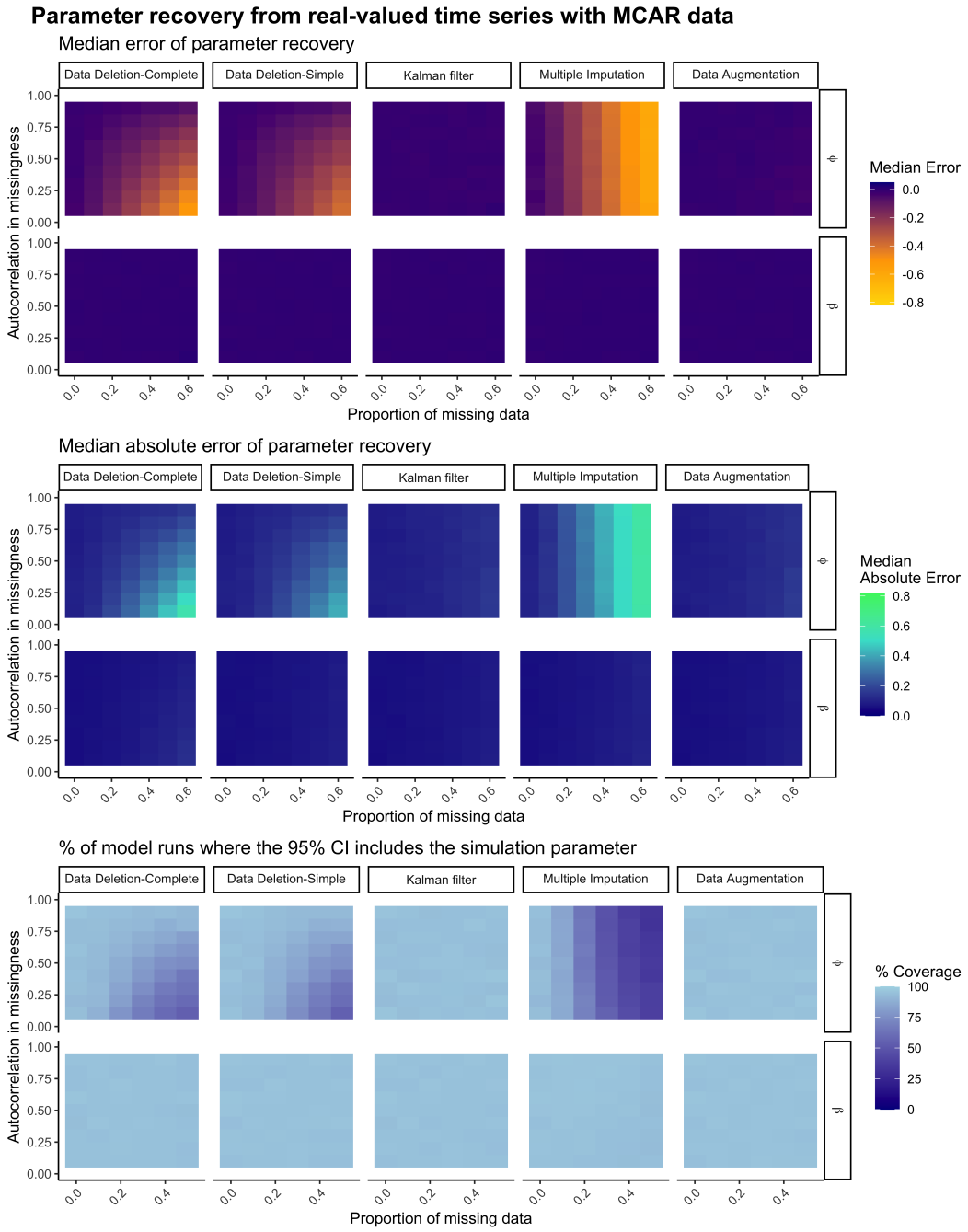


Figure S2: Median error of parameter recovery, median absolute error, and model coverage of $\phi$ and $\beta$, depending on the proportion of missing data and autocorrelation in missingness for each of five missing data approaches, using simulated, real-valued datasets with data missing completely at random (MCAR).

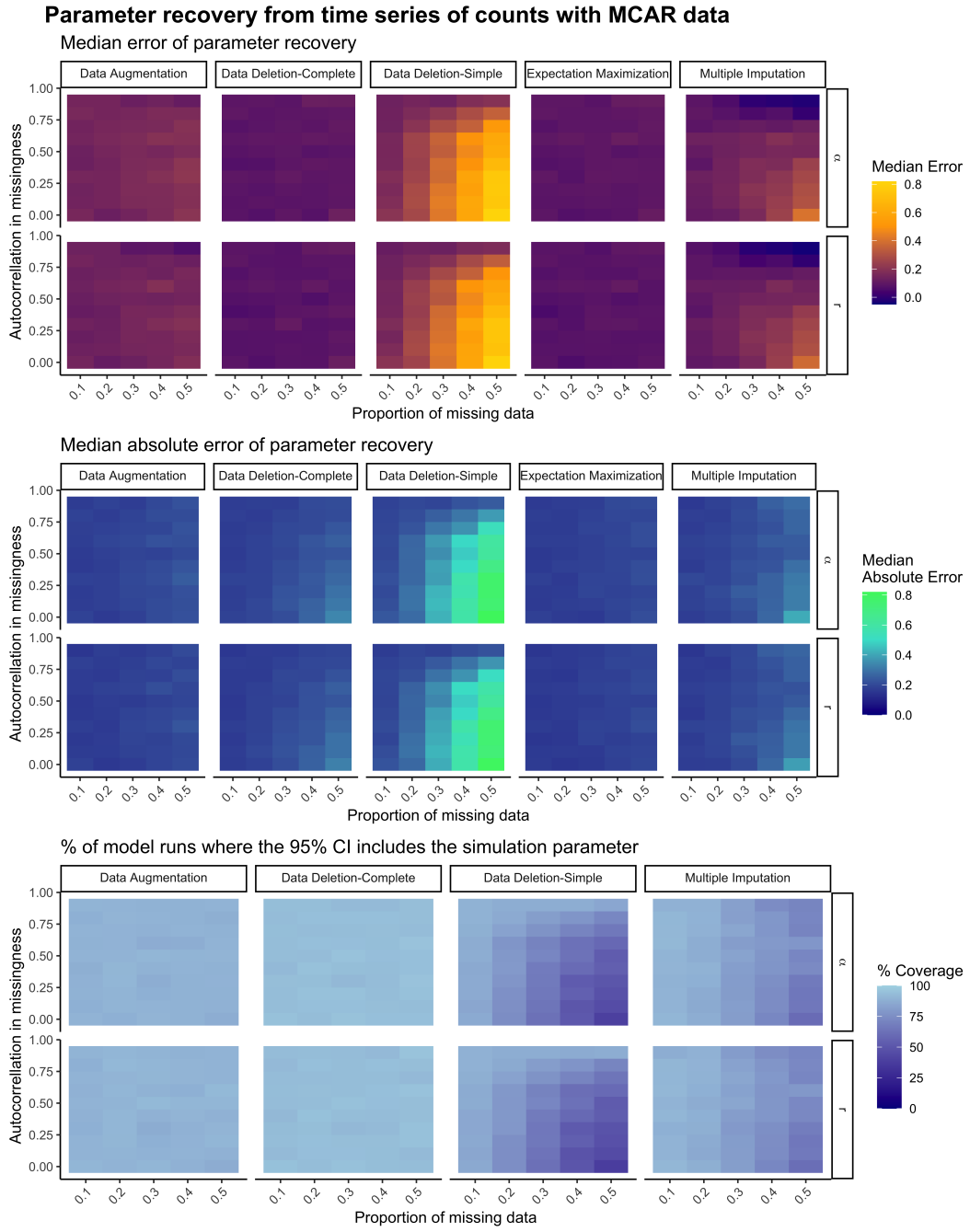


Figure S3: Median error of parameter recovery, median absolute error, and model coverage of $\alpha$ and $r$, depending on the proportion of missing data and autocorrelation in missingness for each of five missing data approaches, using simulated time series of counts with data missing completely at random (MCAR). Note that coverage is not shown for the Expectation Maximization approach, since most implementations of this method do not include estimates of standard error.