Missing data summary

Matt T

2021-04-06

Missing data in the context of Appling et al. dataset

I filled in the dates with missing data in the Appling et al. paper by site · year and made a historgram of missing data frequencies. I am making the assumption that if data were measured at the beginning and end of the year then they were likley collecting data for the whole year. There are obviously flaws to that assumption but it was the easiest way to address whether data was missing because it was actively being collected or not. The red line is the percent of missing data for entire 10 year Shatto Ditch dataset. The blue line is the percent of missing data from 2 years of Kalamath River dataset (from Laurel).

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   20.00   33.00   40.67   58.00  100.00
```

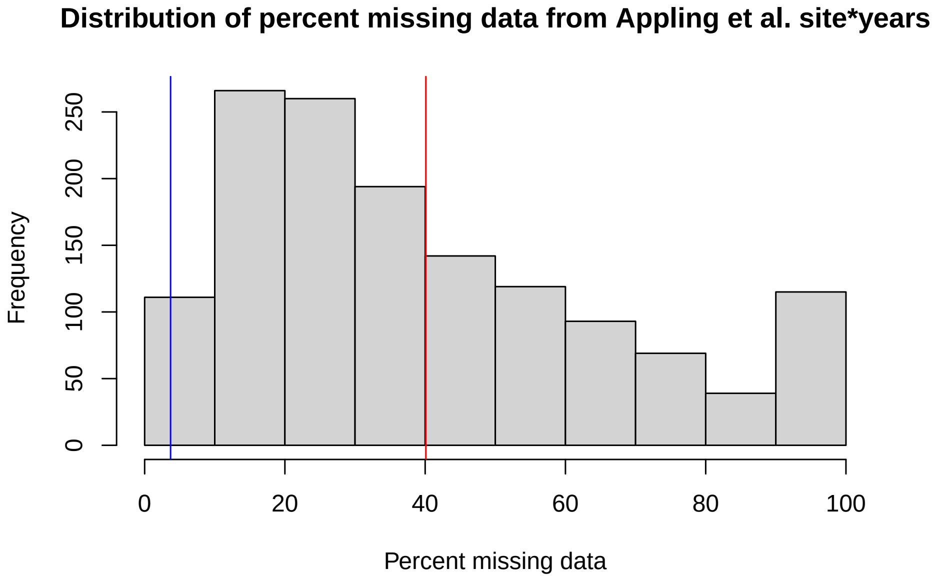
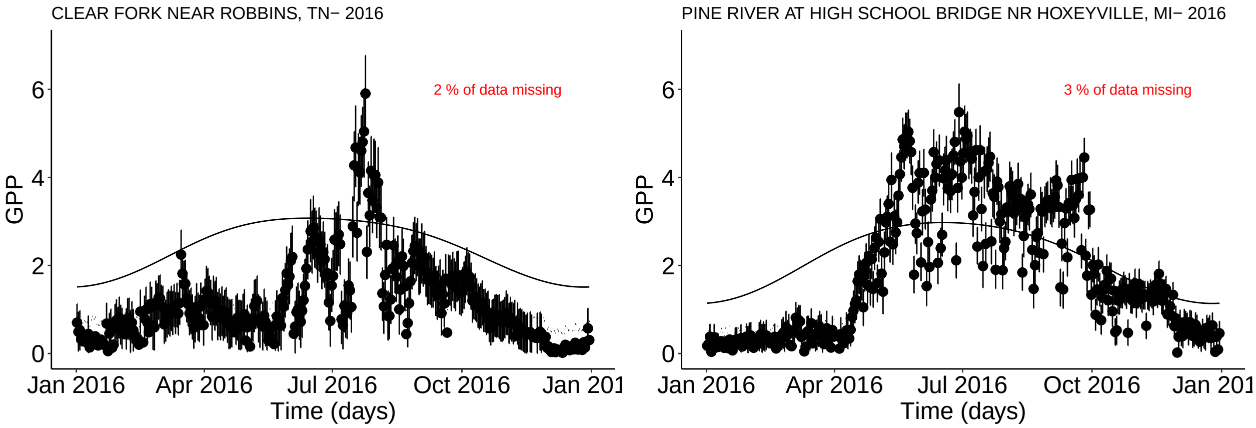


Figure 1: Figure caption: The histogram shows the percent of missing data for each site x year with data measured on the first and last day of the year.

Next, I searched the Appling et al dataset for the site · year(s) with least amount of missing data. There was not a site · year combination with a full year of data, but there were two with less than 3% of data missing.

Site	Year	Missing data (count)	Prop. of year missing
nwis_03409500	2016	7	2
nwis_04125460	2016	12	3

Fortunately, they each have somewhat unique annual GPP trends.



\begin{figure}
\caption{Figure caption: Two sites with low missing data from Appling et al. Data points represent the estimate and the error bars are the 95% credible interval. The line is modeled light using Yard et al. 1995 (NOTE: Light is scaled to fit on the graph) } \end{figure} ### Estimating missing data on real datasets–Bayesian parameter estimation I randomly removed data in 7-day blocks ranging from 2 to 60% of the original data.

Model

Then used an AR(1) missing data process-model with an intercept and light covariate to model the parameters and missing data.

$$GPP = \beta_0 + \phi \times GPP_{t-1} + \beta_1 \times X_{light} + \epsilon_{sdp}$$

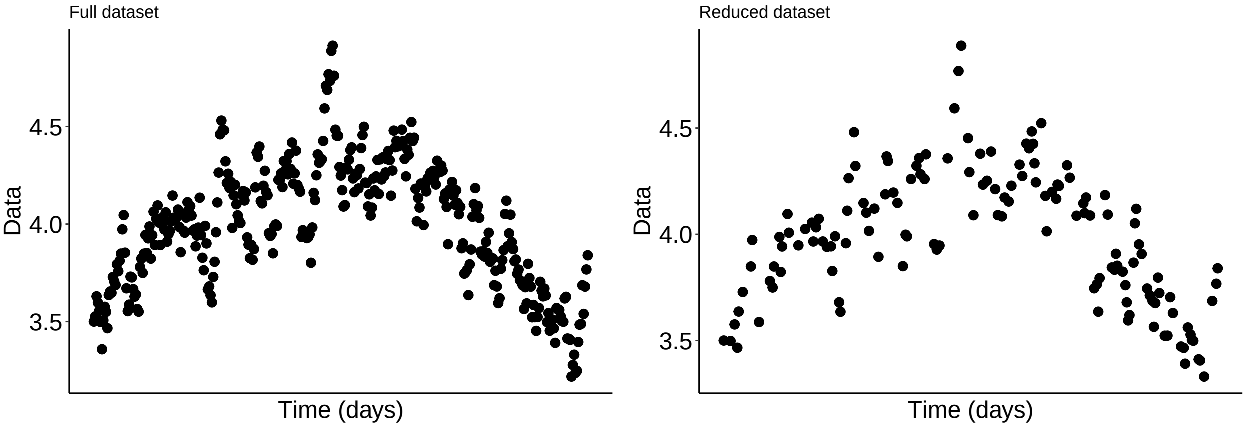
Stan Code

```
"

/*----------------------- Data -------------------------*/
  /* Data block: defines the objects that will be inputted as data */
  data {
    int N; // Length of state and observation time series
    vector[N] y_miss; // Observations
    real z0; // Initial state value
    vector[N] light; //log of light observations
    int y_nMiss; // number of missing values
    int y_index_mis[y_nMiss]; // index or location of missing values within the dataset
  }
/*----------------------- Parameters -------------------------*/
  /* Parameter block: defines the variables that will be sampled */
  parameters {
    vector<lower = 0>[y_nMiss] y_imp;// Missing data
    real<lower=0> sdp; // Standard deviation of the process equation
    //real<lower=0> sdo; // Standard deviation of the observation equation
    real b0;
    real b1;
    real<lower = 0, upper=1 > phi; // Auto-regressive parameter
    //vector[N] z; // State time series
  }
  transformed parameters {
    vector[N] y;
    y=y_miss; // makes the data a transformed variable
    y[y_index_mis] =y_imp; // replaces missing data in y with estimated data
    }
  /*----------------------- Model --------------------------*/
  /* Model block: defines the model */
  model {
    // Prior distributions
    sdp ~ normal(0, 1);
    phi ~ beta(1,1);
    b0 ~ normal(0,5);
    b1 ~ normal(0,5);

    // Distribution for the first value
    y[1] ~ normal(z0, sdp);

    // Distributions for all other states
    for(t in 2:N){
      y[t] ~ normal(b0+y[t-1]*phi+light[t]*b1, sdp);
    }

   generated quantities {
    vector[N] y_rep; // replications from posterior predictive dist
    y_rep[1]=normal_rng(y[1], 0.1);

    for (t in 2:N) {
    y_rep[t]=normal_rng(b0+y[t-1]*phi+light[t]*b1, sdp);
 }
  }
"
```

Simulated-randomly missing days

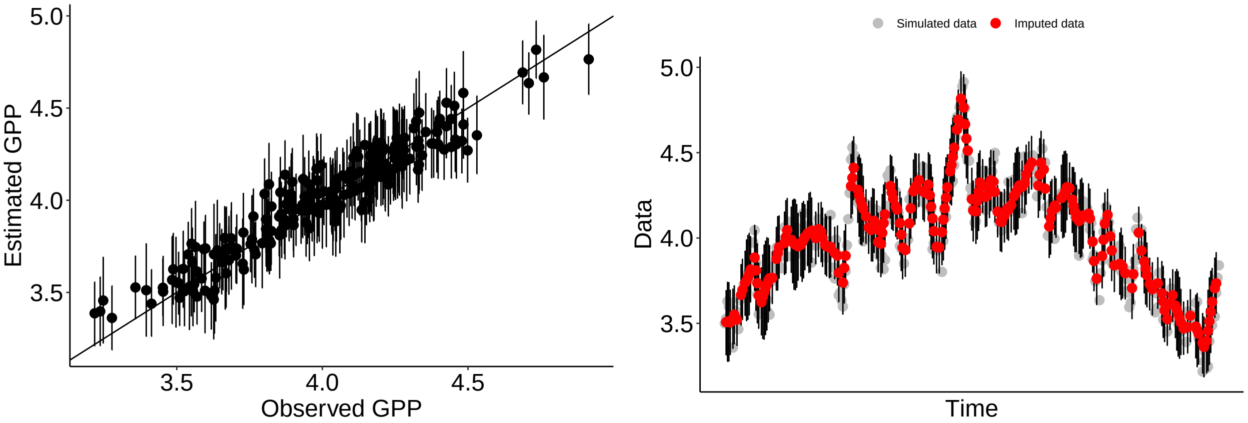
$$GPP = \beta_0 + \phi \times GPP_{t-1} + \beta_1 \times X_{light} + \epsilon_{sdp}$$

β_0=0.1
ϵ_{sdp}= 0.1
ϕ=0.8
β_1=0.1
X_{light}=modeled from Yard et al. (1995) Ecological Modeling

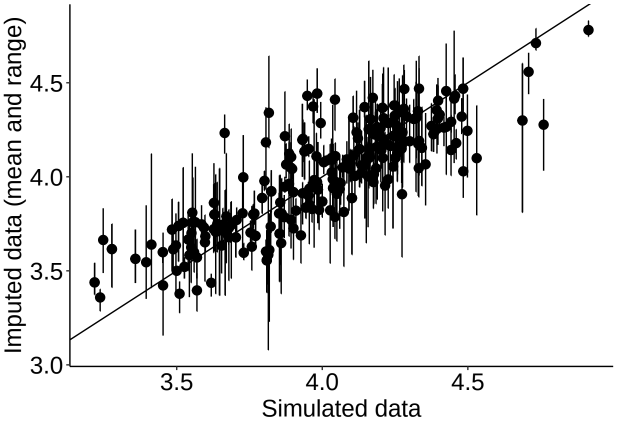


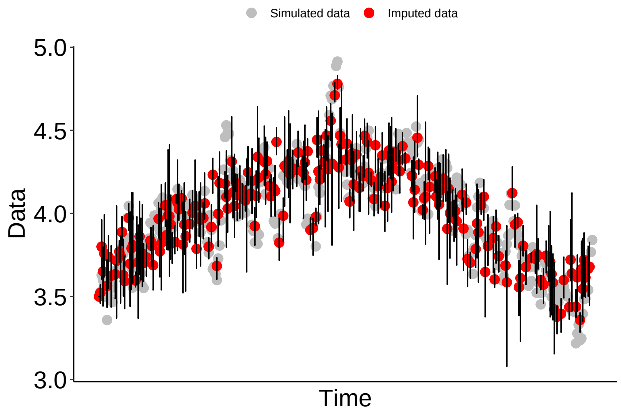
\begin{figure}
\caption{Figure caption: Two sites with low missing data from Appling et al. Data points represent the estimate and the error bars are the 95% credible interval. The line is modeled light using Yard et al. 1995 (NOTE: Light is scaled to fit on the graph) } \end{figure}

Bayes estimated missing data-simulated missing days



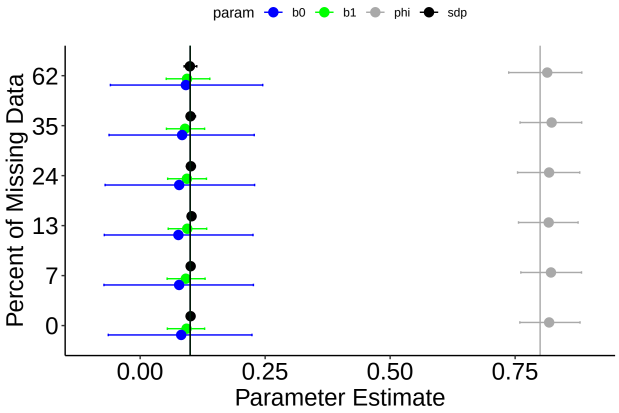
\begin{figure}
\caption{Figure caption: Two sites with low missing data from Appling et al. Data points represent the estimate and the error bars are the 95% credible interval. The line is modeled light using Yard et al. 1995 (NOTE: Light is scaled to fit on the graph) } \end{figure} ### Amelia estimated missing data-simulated missing days \begin{figure}





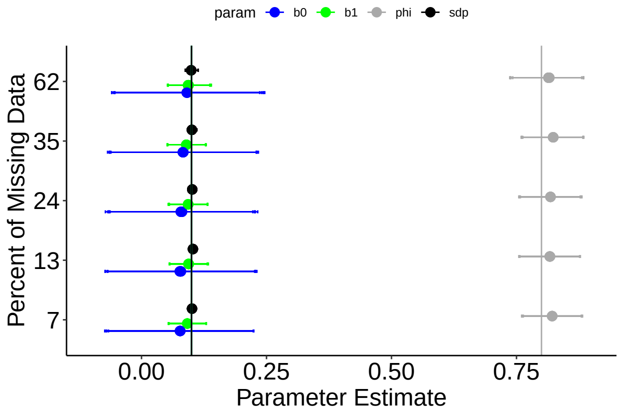
\caption{Figure caption: Two sites with low missing data from Appling et al. Data points represent the estimate and the error bars are the 95% credible interval. The line is modeled light using Yard et al. 1995 (NOTE: Light is scaled to fit on the graph) } \end{figure}

Bayes parameter-simulated missing days

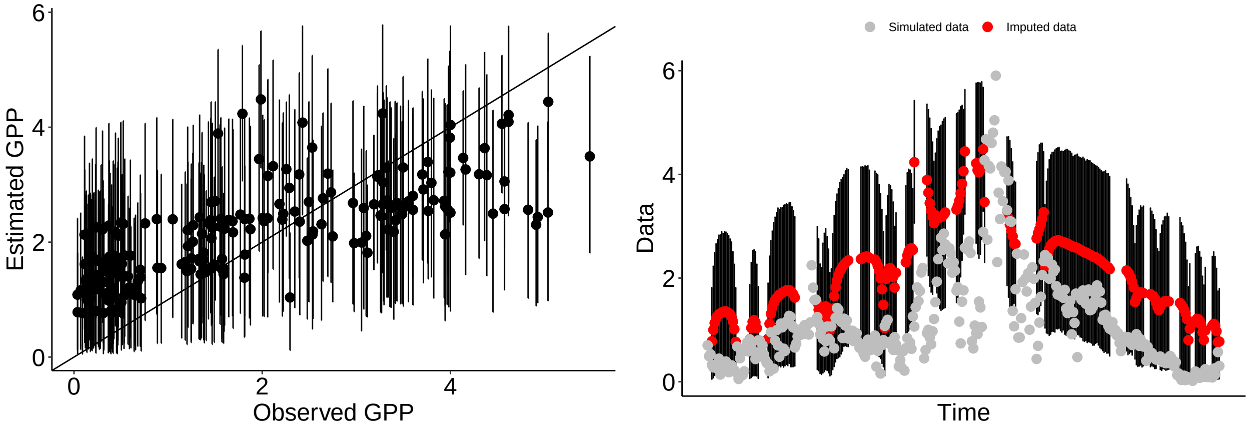


\begin{figure} \caption{Figure caption: Two sites with low missing data from Appling et al. Data points represent the estimate and the error bars are the 95% credible interval. The line is modeled light using Yard et al. 1995 (NOTE: Light is scaled to fit on the graph) } \end{figure}

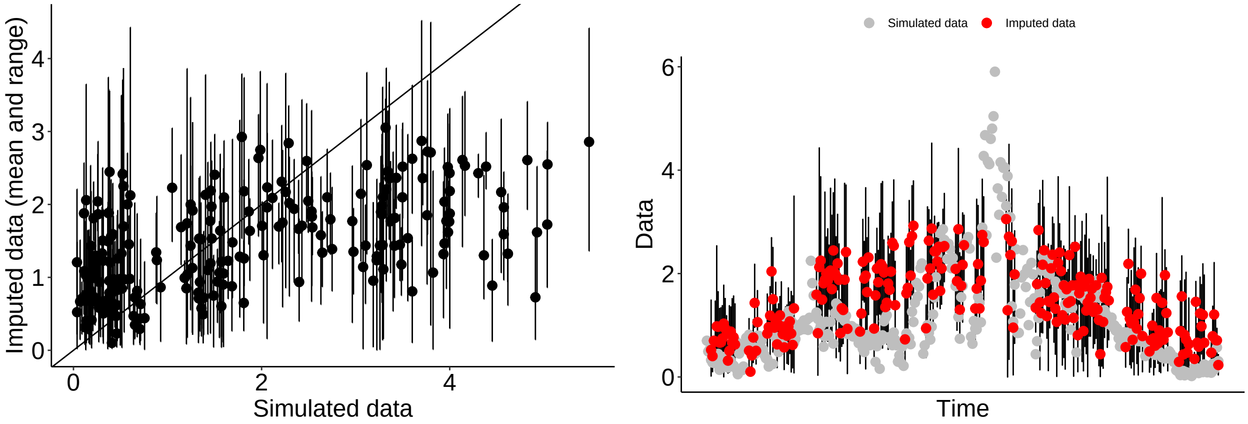
Amelia parameter-simulated missing days



\begin{figure} \caption{Figure caption: Two sites with low missing data from Appling et al. Data points represent the estimate and the error bars are the 95% credible interval. The line is modeled light using Yard et al. 1995 (NOTE: Light is scaled to fit on the graph) } \end{figure} ### Bayes estimated data-low miss 1 \begin{figure}



\caption{Figure X: Estimated missing data with 95% CI. (A)=scatter plot of observed vs. expected. (B)= time-series of estimated data overlapping known data.} \end{figure} ### Amelia estimated data-low miss 1



\begin{figure}

\caption{Figure X: Estimated missing data with 95% CI. (A)=scatter plot of observed vs. expected. (B)= time-series of estimated data overlapping known data.} \end{figure}

Bayes parameter-low miss 1

Amelia parameter-low miss 1

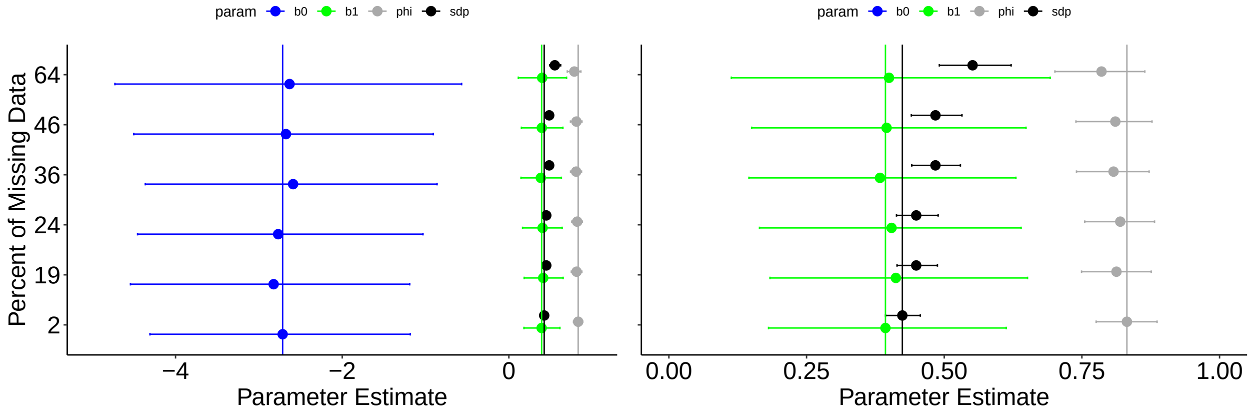


Figure 2: Figure X: Parameter estimates from Bayesian missing-data model. (A)= all data, (B)= positive data

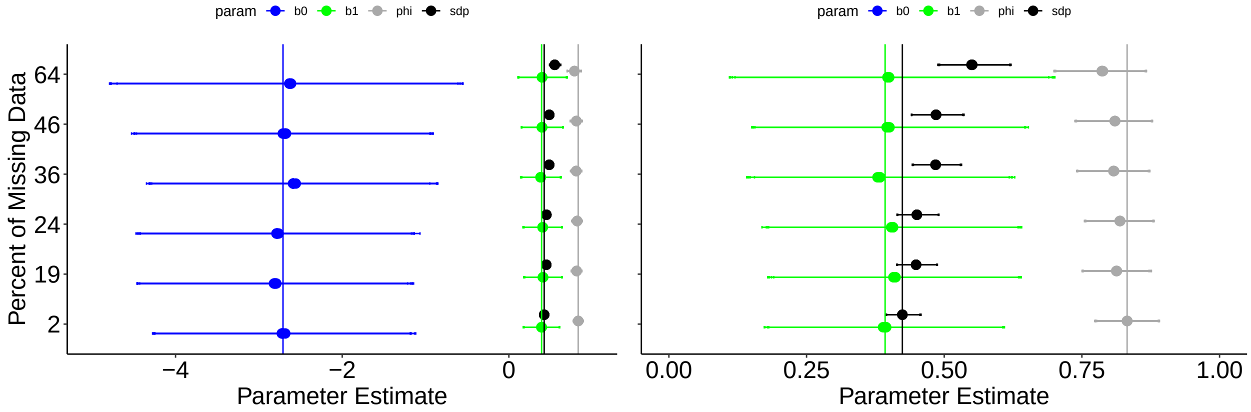
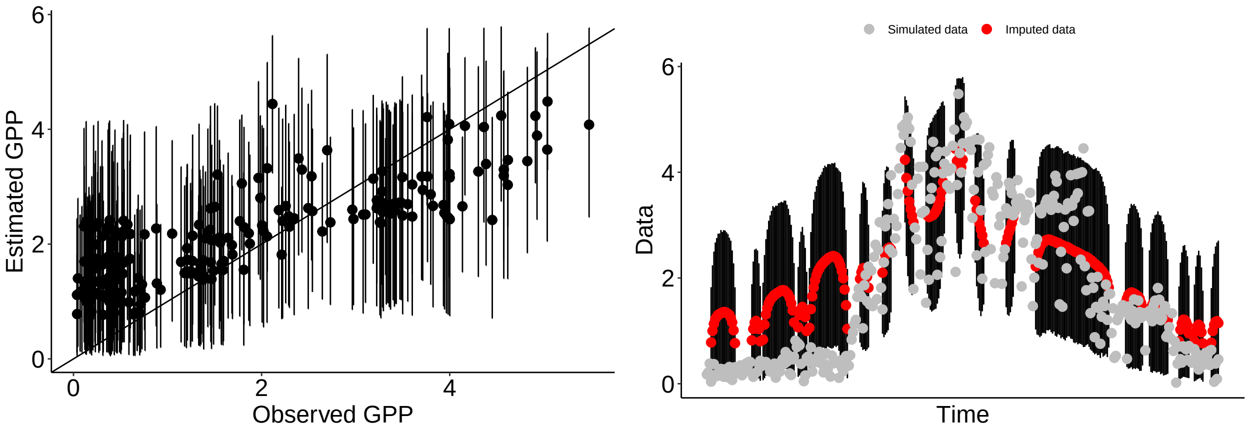


Figure 3: Figure X: Parameter estimates from Bayesian missing-data model. (A)= all data, (B)= positive data

Bayes estimated data-low miss 2



\begin{figure}
\caption{Figure X: Estimated missing data with 95% CI. (A)=scatter plot of observed vs. expected. (B)= time-series of estimated data overlapping known data.} \end{figure}

Amelia estimated data-low miss 2

Bayes parameter-low miss 2

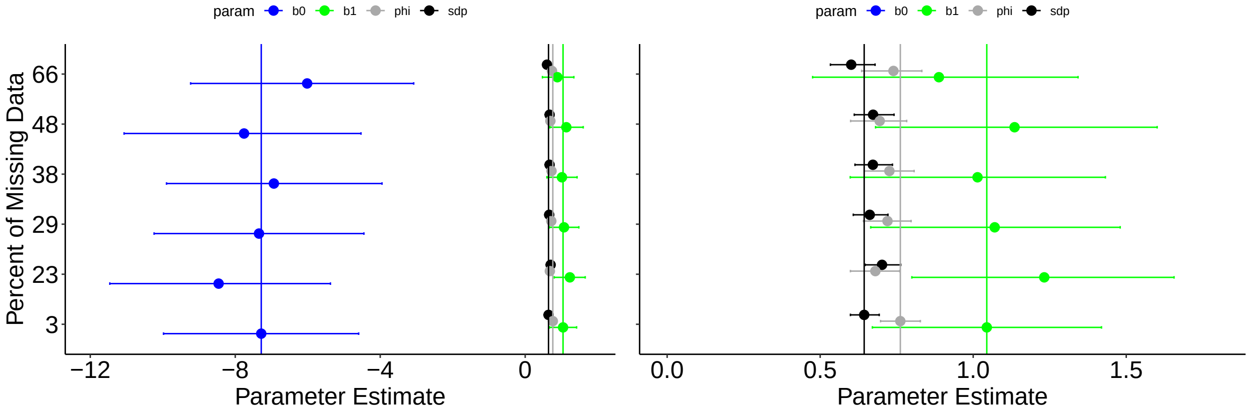


Figure 4: Figure X: Parameter estimates from Bayesian missing-data model. (A)= all data, (B)= positive data

Amelia parameter-low miss 2

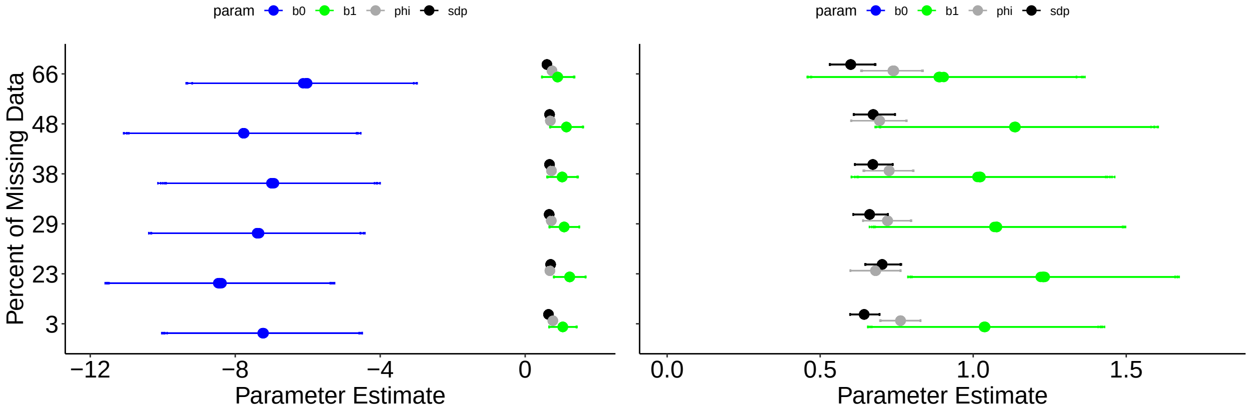


Figure 5: Figure X: Parameter estimates from Bayesian missing-data model. (A)= all data, (B)= positive data