Hazelnut Data Analysis

Sitong Liu
Supervisor: Prof. Geoff Nicholls and Dr. Chieh-Hsi Wu

September 2, 2024

1 Introduction

In this project, we will analyze the hazelnut data using Bayesian inference and investigate how canopy density changes over time. In Section 2, we summarise the modern and archaeological data. The LAI values in the archaeological data are missing and must be imputed. In Section 3, we present a single imputation analysis, which is straightforward but is a very crude method. In Section 4, we use Bayesian inference to analyze the whole data and show that deforestation did occur.

2 Data Summary

2.1 The Modern data

The modern dataset lists 192 observations on six variables of interest. See Figure 13 in Appendix B for scatter plots of these data. The variables are as follows.

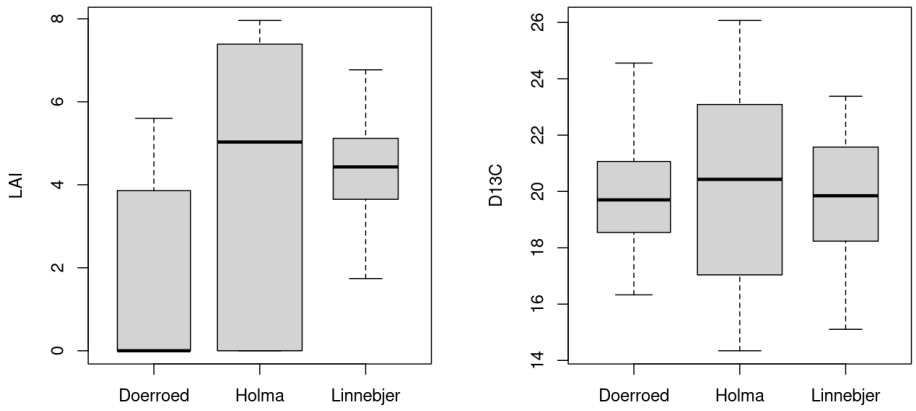
- SampleID: Unique sample identifier.
- Site: The hazelnut sample site name. A 3-level categorical variable.
- Intra-tree/nut: This indicates whether the hazelnut samples come from the same tree. If the value is "Y", this indicates that hazelnut D13C values from the same location came from the same tree rather than potentially coming from multiple trees in that location. The values '1', '2' and '3' indicate samples that come from different positions within the same nut. We define our group index variable based on this information in data analysis.
- LAI (l_M): Leaf Area Index, which quantifies the amount of leaf material in a canopy. By definition, it is the ratio of one-sided leaf area per unit ground area. Values are continuous and vary from 0 (no leaf cover, so open) to infinity. This is our independent variable of interest.
- Date: The year in which the hazelnut sample was gathered.
- D13C (y_M): This is our dependent variable, measuring the carbon isotope value of hazelnut shells.

See Figure 1 for the distributions of LAI and D13C values for each site in the modern data.

2.2 The Archaeological data

The archaeological data list 72 observations on four variables of interest. The variables are as follows.

- SampleID: Unique sample identifier
- Site: The hazelnut sample site name. An 18-level categorical variable.
- Date: The time period in which samples came from. We know the lower and upper bound of each time interval. $\tau_i \in [\tau_i^-, \tau_i^+]$, where τ_i is the exact date of sample i, and τ_i^-, τ_i^+ are the associated lower and upper bound, respectively.



(a) (b)

Figure 1: The distribution of (a) LAI (b) D13C by Site in the Modern data. Box width is proportional to the number of observations.

- D13C (y_A): This is our dependent variable, measuring the carbon isotope value of hazelnut shells.

See Figure 2 for the time of archaeological data. See Figure 3 for the distribution of D13C in each time period recorded in the archaeological data.

2.3 Notation

In this section, we will introduce the mathematical notation we use for each physical quantity in this project.

- $y_M = \{y_{M,i} : i \in \mathcal{M}\}$ is the D13C value for modern data samples, where $\mathcal{M}$ is the set of modern data index.
- $y_A = \{y_{A,i} : i \in \mathcal{A}\}$ is the D13C value for archaeological data samples, where $\mathcal{A}$ is the set of archaeological data index.
- $l_M = \{l_{M,i} : i \in \mathcal{M}\}$ is the LAI value for modern data samples, where $\mathcal{M}$ is the set of modern data index.
- $l_A = \{l_{A,i} : i \in \mathcal{A}\}$ is the LAI value for archaeological data samples, where $\mathcal{A}$ is the set of archaeological data index. This is what we want to predict.
- $\tau = \{\tau_i : i \in \mathcal{A}\}$ is the time for archaeological data samples, where $\mathcal{A}$ is the set of archaeological data index.
- L is the lower bound of time in the archaeological data.
- U: is the upper bound of time in the archaeological data.

3 Single Imputation

The LAI values in the archaeological data are missing. To figure it out, we use the modern data to model the relation between *LAI* and *D13C* and then invert that relation to impute the missing *LAI* values in the archaeological data from their observed *D13C* values.

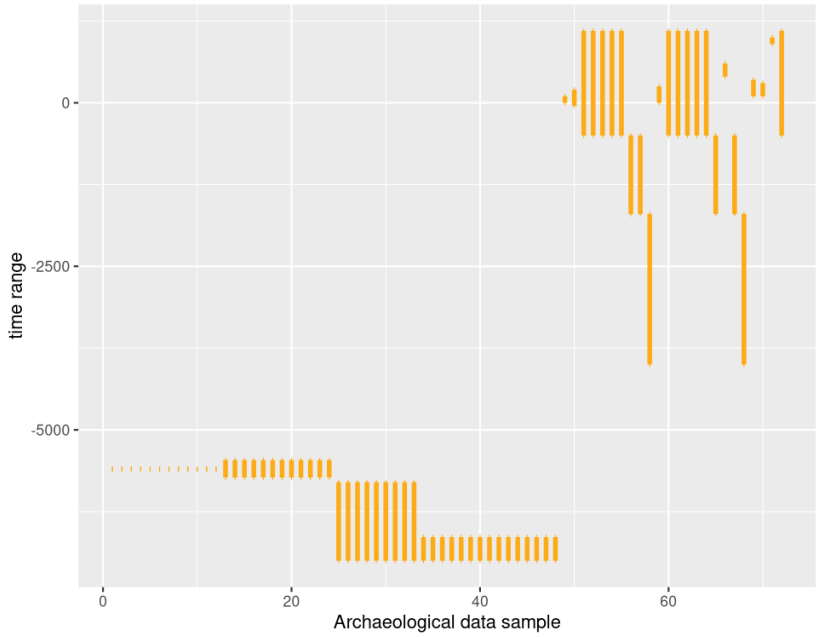


Figure 2: Time of archaeological data, characterized by upper and lower bound.

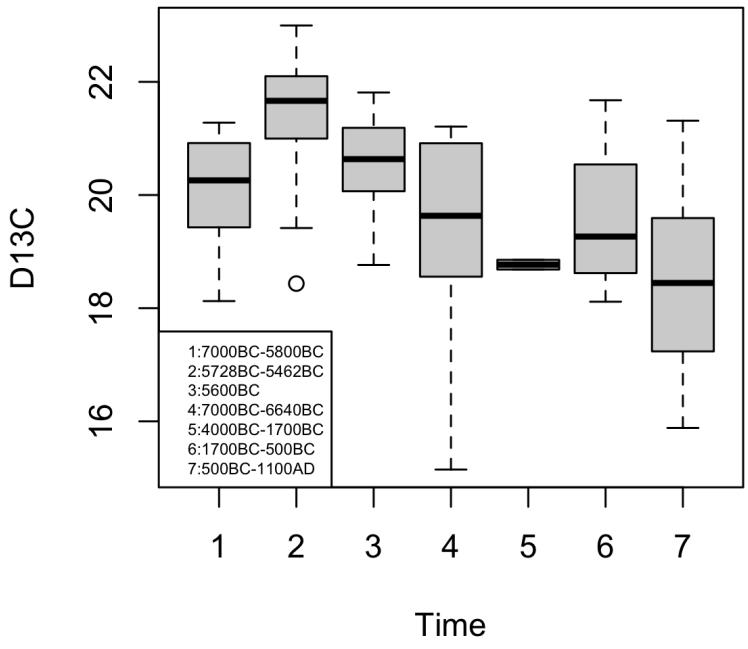


Figure 3: Distribution of D13C for each time period recorded in the archaeological data.

3.1 Methods and Results for Single Imputation

We first fit a normal linear model regressing y_M on l_M to the modern training data

$$y_{M,i} = \alpha + \beta \cdot l_{M,i} + \epsilon_i, \;\; i \in \mathcal{M}.$$

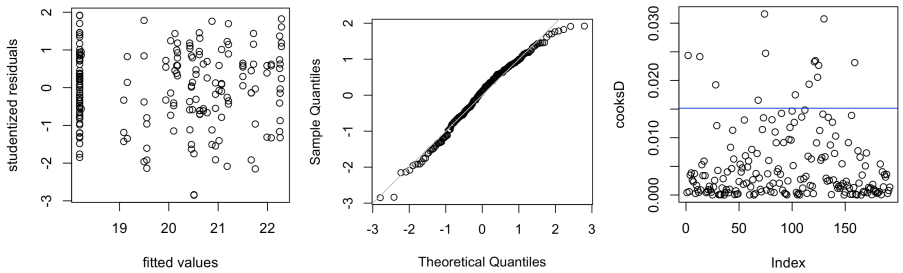
Random effects due to site are omitted as they do not generalise to new sites. Variation due to site is absorbed into the variance of ϵ.

The estimated result is $\hat{\alpha} = 18.19$ with standard deviation 0.24, and $\hat{\beta} = 0.51$ with standard deviation 0.05. Then, we impute l_A on the archaeological data using the predicted mean D13C value and inverting,

$$\hat{l}_{A,i} = \frac{y_{A,i} - \hat{\alpha}}{\hat{\beta}}, \;\; i \in \mathcal{A}.$$

See Figure 4 for diagnostic plots of the model.

Figure 5 is the boxplot of $\hat{l}_A$ in each time period.



(a) (b) (c)

Figure 4: Diagnostic plots for the single linear regression model for single imputation (a) fitted values versus studentized residuals (b) Q-Q plot of studentized residuals (c) Cook's distance, where the blue horizontal line is three times the mean distance.

4 Bayesian Inference

Our goal is to figure out how canopy density has changed over time. We write down the posterior of interest and explain each part in detail separately.

$$\begin{aligned}\pi(\alpha, \beta, \sigma_0^2, \sigma_1^2, l_A, \tau, \gamma, A, B | y_A, y_M) \propto\; & p(y_M | \alpha, \beta, \sigma_0^2, \sigma_1^2) p(y_A | \alpha, \beta, \sigma_0^2, \sigma_1^2, l_A) \\ & \times \pi(\alpha, \beta, \sigma_0^2, \sigma_1^2, \gamma, A, B) \pi(l_A | \tau, A, B, \gamma) \pi(\tau)\end{aligned}$$

4.1 Likelihood model

4.1.1 $p(y_M | \alpha, \beta, \sigma_0^2, \sigma_1^2)$

The observation model for y_M (D13C value in the modern data) is

$$y_{M,i} = \alpha + \beta \cdot l_{M,i} + \epsilon_i, \;\; \text{for} \;\; i \in \mathcal{M}, \tag{1}$$

$$\epsilon_i \sim N(0, \sigma_{y,i}^2), \tag{2}$$

$$\sigma_{y,i}^2 = \sigma_0^2 + \sigma_1^2 \mathbb{1}\{l_{M,i} > 0\}, \tag{3}$$

where $\mathcal{M}$ is the set of modern data index. ϵ captures randomness in the data, which includes both measurement error of the linear regression model and randomness that comes from LAI values. We notice zero-inflation of l_M. If $l_{M,i} = 0$, the randomness only comes from measurement error, which we use σ_0^2 to express. If $l_{M,i} > 0$, variance will increase, which is measured by σ_1^2.

We perform a Box-Cox transformation on y_M, and the optimal λ is approximately 1 (see Figure 6), which indicates that there is no need to do extra transformation for y_M.

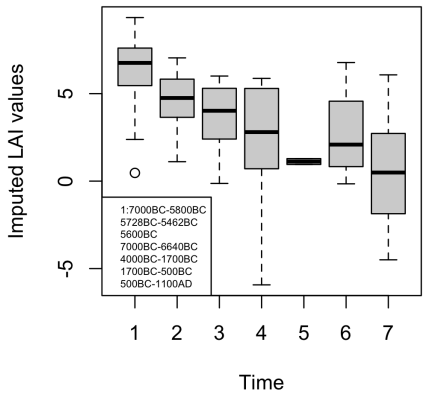


Figure 5: $\hat{l}_A$ (imputed LAI) grouped by time in the archaeological data.

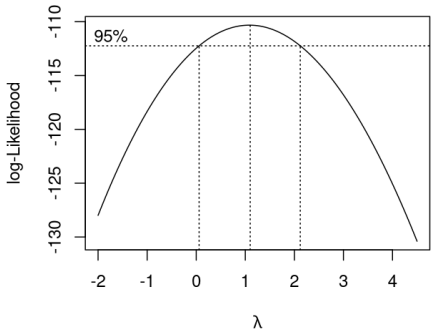


Figure 6: Box-Cox plot of D13C in the modern data. The optimal λ is approximately 1.

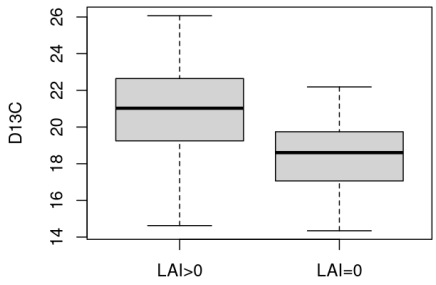


Figure 7: Distribution of D13C given the LAI values in the Modern data. D13C with positive LAI has a larger variance than D13C with LAI equal to zero.

In this model, we assume different variances for $\{y_{M,i} : l_{M,i} = 0\}$ and $\{y_{M,i} : l_{M,i} > 0\}$. We can use F-test to test the equality of two variances. The p-value is 0.01, so we reject the null hypothesis at a significance level of 0.95. We also notice that the data is not normally distributed, so F-test is not a proper choice. Alternatively, we can use Levene's test, and it again shows that the two groups have different variances. We give a boxplot of D13C grouped by LAI value in Figure 7.

4.1.2 $p(y_A|\alpha, \beta, \sigma_0^2, \sigma_1^2, l_A)$

We use the same model structure for both modern and archaeological data sets. For the archaeological data, we have

$$\begin{aligned} y_{A,i} &= \alpha + \beta \cdot l_{A,i} + \epsilon_i, \text{ for } i \in \mathcal{A}, \\ \epsilon_i &\sim N(0, \sigma_{y,i}^2), \\ \sigma_{y,i}^2 &= \sigma_0^2 + \sigma_1^2 \mathbb{1}\{l_{A,i} > 0\}. \end{aligned}$$

l_A is LAI value for the archaeological data, and $\mathcal{A}$ is the set of archaeological data index.

4.2 Prior elicitation

We observe that there is a non-negligible probability that LAI takes value zero (zero inflation), which happens when the hazelnut shell samples are collected in an open area (outside the forest). We assume

$$l_{A,i} \sim p_{\tau,i} \cdot \delta_0(\cdot) + (1 - p_{\tau,i})\delta_\gamma(\cdot).$$

At time t, with probability p_t, the hazelnut shell i is collected in an open area, so $l_{A,i} = 0$. Otherwise, the hazelnut shell i is collected in the forest, and $l_{A,i}$ takes value γ. Of course, the canopy cover in the forest will vary, so we could imagine that γ should be a separate random value for each sample. However, this variance can be absorbed into the observation model for y and is the reason for the two-level variance in Equation 3.

p is parameterized by t, a, b as

$$p_t := \frac{1}{1 + \exp(a + bt)}$$

To have a better physical interpretation, we reparameterize p_t. Let $A := -\frac{a}{b}$, which is the t value when $p = \frac{1}{2}$. We expect p_t to change continuously and smoothly over time, so we impose a prior $\pi(A) = N(-2000, 2000^2)$, which centers at -2000, and covers approximately the whole time range in the data, varying from -6000 to 2000. Let $B := -\frac{b}{4}$, which is the derivative of p_t when $t = -\frac{a}{b}$. We suppose p_t change by 10% in a century, so we let $\pi(B \cdot S) = t(5)$, where $S = 1000$. Intuitively, S is the number of years that p increases from 0 to 1 (informally, it is the timescale over which deforestation occurred). In this case, $a = 4AB, b = -4B$ and $p_t = \frac{1}{1+\exp(4AB-4Bt)}$.

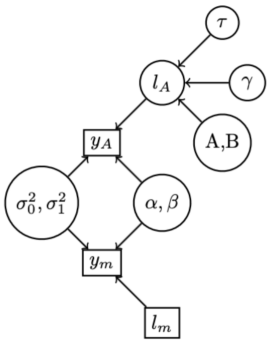


Figure 8: Graphical representation of the model

As mentioned before, we only know which time interval archaeological samples belong to, but not the exact date. For sample i in the archaeological data, we are given that it belongs to the time interval $[\tau_i^-, \tau_i^+]$. Let τ_i be the exact date of sample i, then we assume τ_i is uniformly distributed over $[\tau_i^-, \tau_i^+]$. Therefore, $\pi_\tau(\tau) = \prod_{i \in \mathcal{A}} \frac{\mathbb{1}\{\tau_i \in [\tau_i^-, \tau_i^+]\}}{\tau_i^+ - \tau_i^-}$, where $\mathcal{A}$ is the set of sample index for the archaeological data and $\tau = \{\tau_i : i \in \mathcal{A}\}$.

We also specify priors for other parameters. $\pi(\alpha) = N(20, 1^2), \pi(\beta) = N(0.5, 0.1^2), \pi(\gamma) = \Gamma(4, 1)$, and Jeffreys prior for σ_0^2, σ_1^2.

Figure 8 is the graphical representation of the problem we analyse. Let $\phi = (\sigma_0^2, \sigma_1^2, \alpha, \beta), \theta = (l_A, \gamma, \tau, A, B)$, then this model is equivalent to a two-module system as shown in Figure 9.

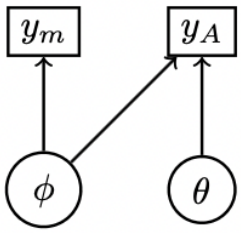


Figure 9: Two-module simplified representation.

4.3 MCMC sampling

We use Metropolis-within-Gibbs to sample from the posterior. We have 151 variables to sample $\alpha, \beta, \sigma_0^2, \sigma_1^2, \gamma, l_A, \tau, A, B$ (l_A and τ contain 72 variables each). For each iteration, we randomly update one variable among them. For variables other than l_A, τ, we use random walk proposals. We update l_A by sampling either 0 or γ with probability proportional to p_τ to choose 0. We update τ by uniformly sampling between the lower and upper bound. After checking our MCMC chains have converged, we plot p with time (see Figure 10).

5 Conclusion

We allowed for uncertainty in the dating of sample specimens. The data adds very little information to the original bounds provided by dating, so the posterior distribution for these dates remains near uniform in the intervals in which they are constrained to lie.

From Figure 10, we see that p_t is around 0 before 4000 BC, increases rapidly and stabilizes at 1 after around 2000 BC. Before 4000 BC, there was no open area, and after 2000 BC, all areas were open. We can infer that deforestation occurs between 4000 BC and 2000 BC. The pattern of p_t depends heavily on the data. As we only have data from either very early or very recent times and few in between, it

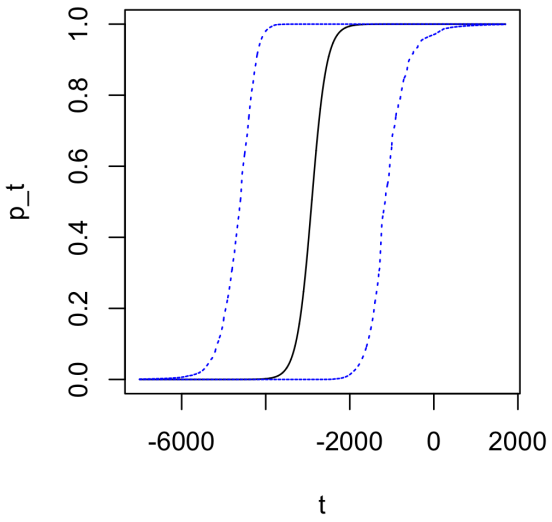


Figure 10: Probability $p_t,\ t \in [L, U]$ of getting samples in open areas as time went by. We run MCMC for 5000000 iterations and use the last 2000000 iteration samples with a thinning interval of 5000. The blue dotted lines are 95 % confidence intervals.

is unclear to figure out what happens between 4000 BC and 2000 BC. In future, we can analyze the data using semi-modular inference and see whether our model is a good fit to the data.

6 Appendix

6.1 Appendix A: MCMC Convergence Check

In this section, we show that our MCMC algorithm has reached convergence. We run MCMC for 5000000 iterations and take samples with a thinning interval of 5000. We show trace plots in Figure 11 and the corresponding histogram in Figure **??**.

6.2 Appendix B: Scatter plot of the Modern data

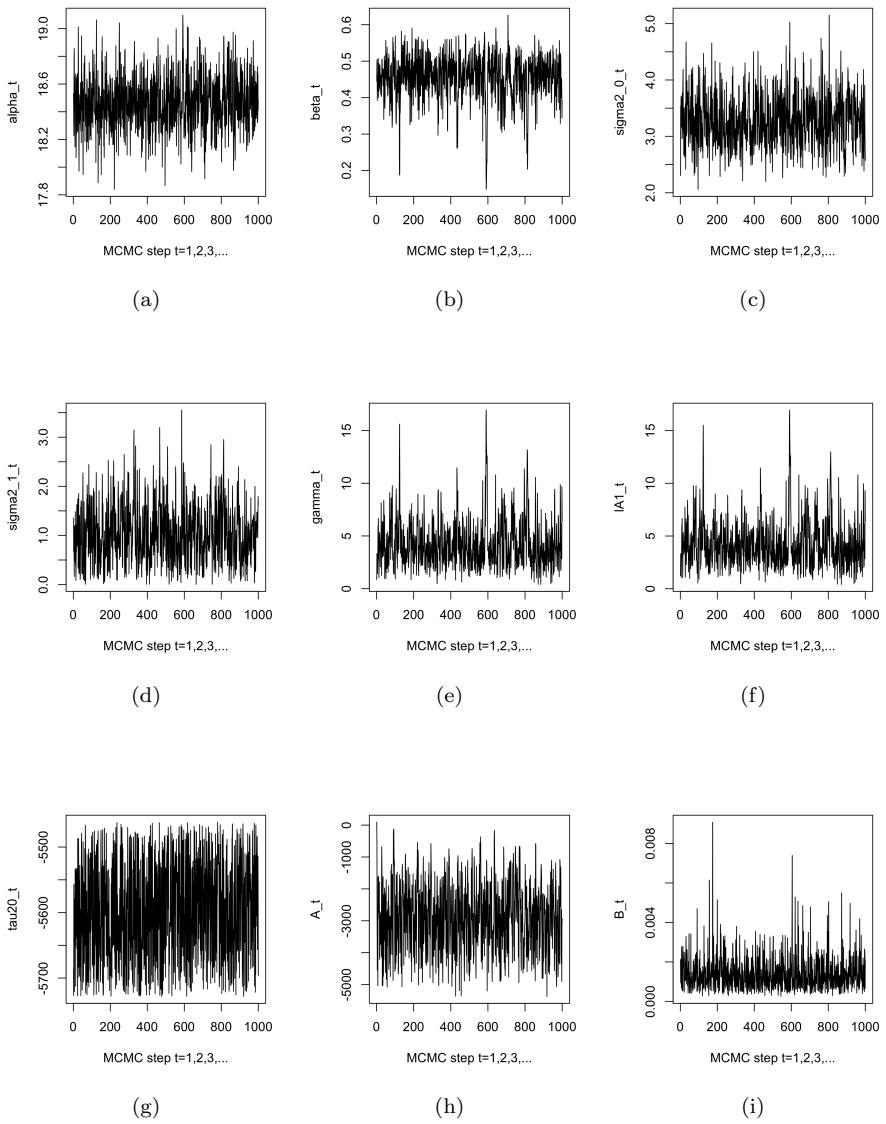
In Figure 13, we present scatter plots of a selection of key variables from the modern data.

7 Semi-Modular Inference

As mentioned in Section 5, the model is equivalent to a two-module system, and Figure 8(b) is the graphical representation. In this section, we will use semi-modular inference (SMI) to analyze the data. We write the SMI posterior.

$$
\begin{aligned}
p_{smi,\eta}(\phi, \theta, \tilde{\theta} | y_M, y_A) &= p_{pow,\eta}(\phi, \tilde{\theta} | y_M, y_A) p(\theta | y_A, \phi), \\
p_{pow,\eta}(\phi, \tilde{\theta} | y_M, y_A) &\propto p(y_M | \phi) p(y_A | \phi, \tilde{\theta})^{\eta} p(\phi, \tilde{\theta}), \\
p_{smi,\eta}(\phi, \theta | y_M, y_A) &= \int p_{smi,\eta}(\phi, \theta, \tilde{\theta} | y_M, y_A) d\tilde{\theta},
\end{aligned}
$$

where $\phi = (\sigma_0^2, \sigma_1^2, \alpha, \beta)$, $\theta = (l_A, \gamma, \tau, A, B)$ in our model.



(a) (b) (c)

(d) (e) (f)

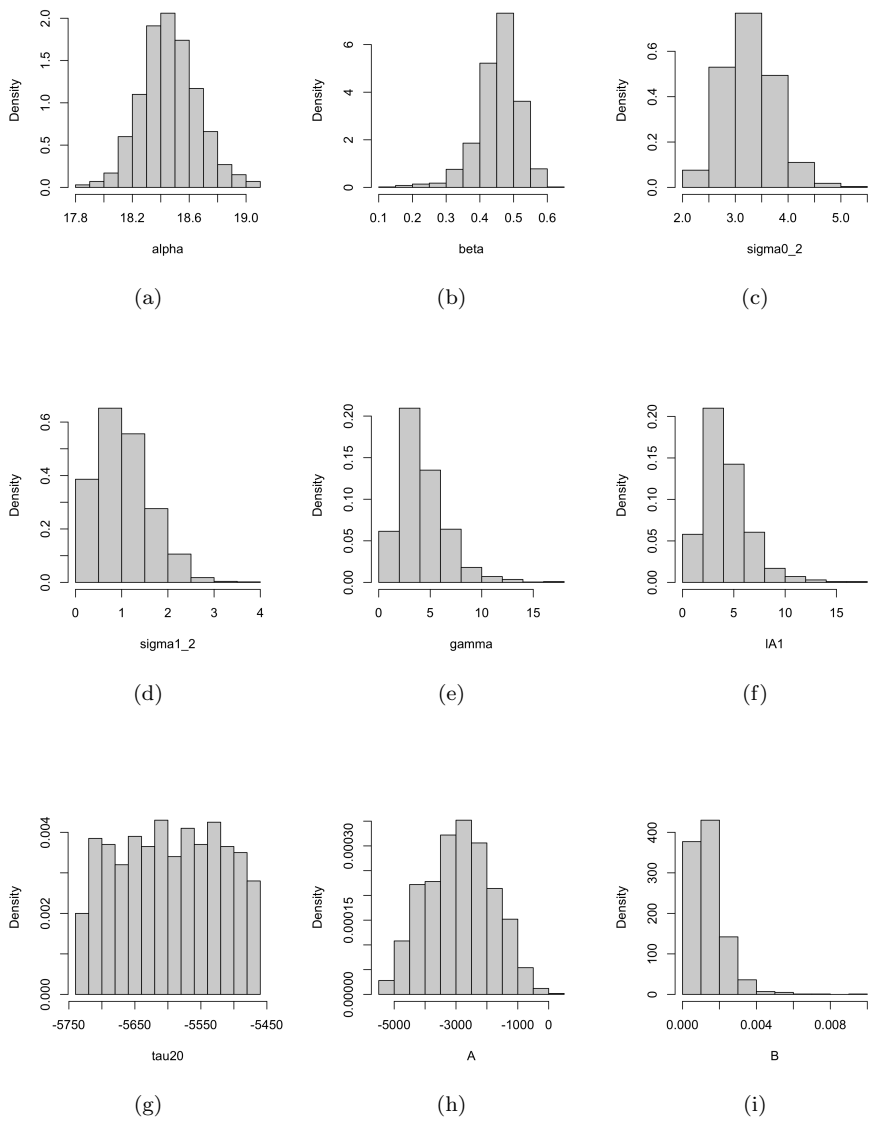
(g) (h) (i)

Figure 11: MCMC with 5000000 iterations. The trace plots show MCMC samples after thinning (thinning interval 5000). LA1 means the LAI value for the first sample, and tau20 means the time for the twentieth sample.

We use nested MCMC to sample from the η-SMI posterior. First, we sample N_1 draws from $p_{pow,\eta}(\phi, \tilde{\theta}|y_M, y_A)$; for each sampled value of ϕ, run a sub-chain targeting $p(\theta|\phi, y_A)$ for N_2 steps; keep only the last sampled value in this sub-chain. The resulting joint samples $(\phi, \tilde{\theta}, \theta)$ are approximately distributed according to the SMI posterior.

We choose η, the influence parameter, according to elpd (expected log pointwise predictive density). As we do not know the true data-generating process, we use WAIC estimators to approximate the elpd.

In practice, we set $N_1 = 200000$ and thin the MCMC samples with thinning interval 4000. We then run the sub-chain under these 50 samples and let $N_2 = 100000$. After obtaining SMI posterior samples $\{\phi^{(s)}, \theta^{(s)}\}_{s=1}^{5}0$, we use WAIC to estimate elpd, as shown in Figure 13. We only use 50 samples and the variance might affect the result. From the figure, we see that there is no apparent trending of elpd. This might be due to the reason that the prior dominates more than the likelihood in our model, so changing η does not change much in the predictive performance.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 12: MCMC with 5000000 iterations. The histograms show MCMC samples after thinning (thinning interval 5000). LA1 means LAI value for the first sample, and tau20 means the time for the twentieth sample.

8 Power Posterior (Update in August, 2024)

We only perform the first stage and put η-power on the second module.

$$p_{pow,\eta}(\phi, \theta | y_M, y_A) \propto p(y_M|\phi)p(y_A|\phi, \tilde{\theta})^{\eta} p(\phi, \tilde{\theta})$$

We use LOOCV to estimate ELPD on Y_A. For MCMC, I use 5,000,000 iterations and thinning interval is 5000. Let the posterior samples be $\{(\phi^i_{-j}, \theta^i_{-j})\}_{i=1}^n \sim \pi(\phi, \theta | y_M, y_{A,-j}),\ \ j = 1, \ldots, |\mathcal{A}|$, and n is the number of posterior samples. Then

$$ELPD_{loocv} = \sum_{j=1}^{|\mathcal{A}|} log(\frac{1}{n} \sum_{i=1}^{n} p(y_{A,j}|\phi^i, \theta^i))$$

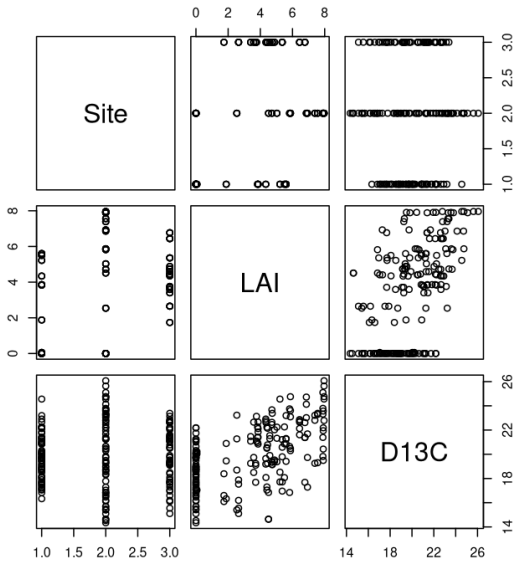


Figure 13: Scatter plot of some of the key variables in the modern data.

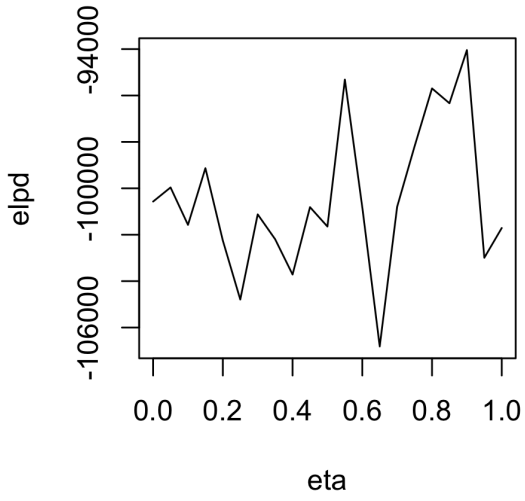


Figure 14: Elpd estimated by WAIC across values of $\eta = \{0, 0.05, 0.10, \ldots, 1.0\}$.

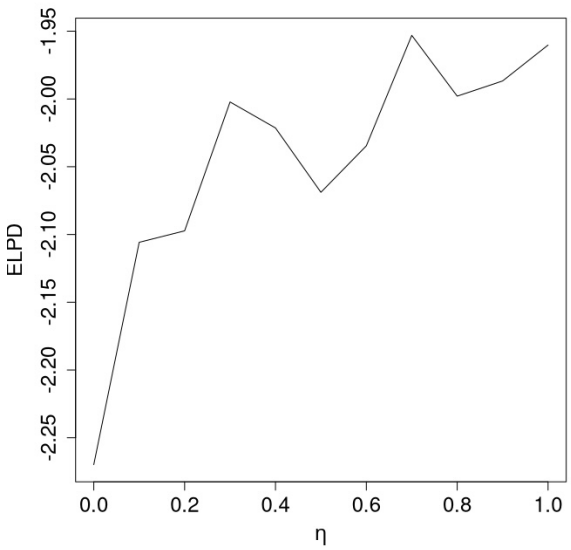


Figure 15: ELPD of Y_A, estimated using LOOCV.