DEPARTMENT OF ENVIRONMENTAL SCIENCE & TECHNOLOGY

Brian W. Davis[1] brianwdavis@gmail.com
Steven B. Mirsky[2], *W. Dean Hively*[2,3], *Greg McCarty*[2], *Brian A. Needelman*[1]
[1]University of Maryland, [2]USDA-ARS, [3]USGS

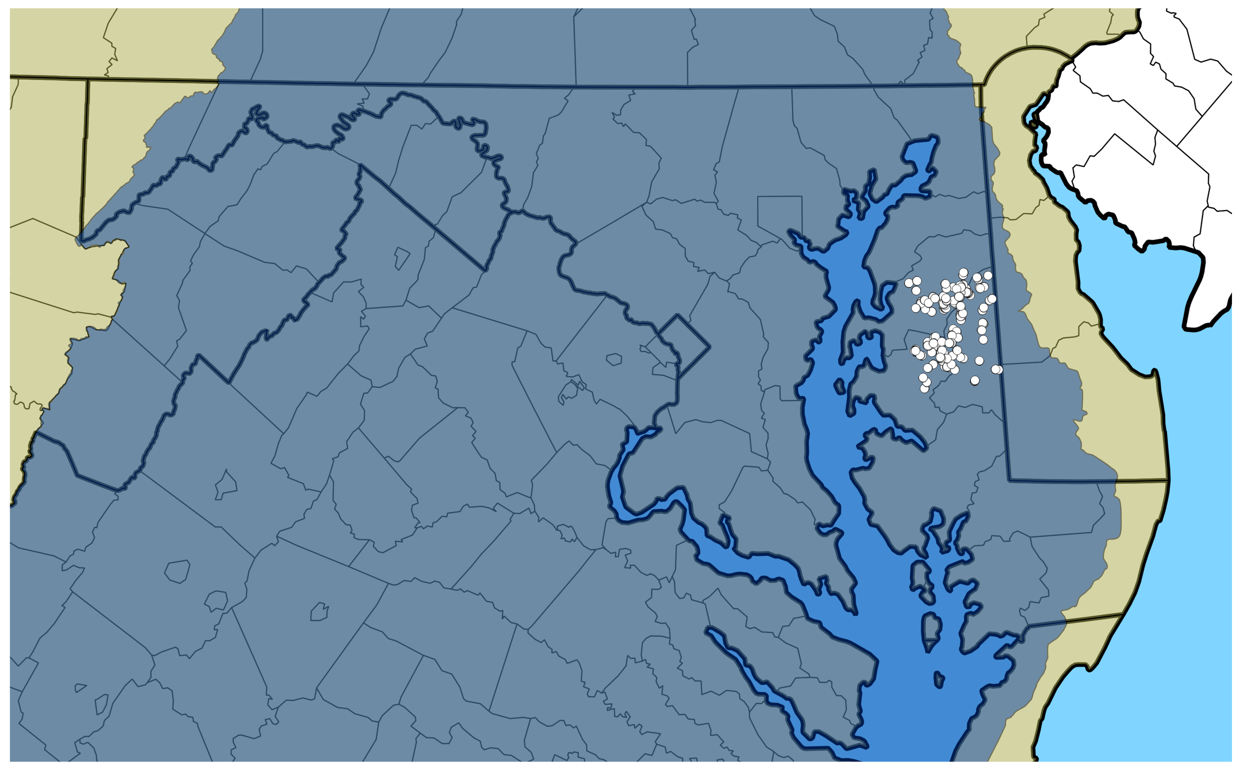
Bayesian Estimation with Unbalanced Data in the Maryland Cover Crop Program

Introduction

Scientists are often trained in frequentist analysis, which can present conceptual challenges to adopting Bayesian techniques. In particular, they often express unease with assigning priors for models, preferring to "let the data speak for itself".

In this use case however, we are modeling biomass performance in the Maryland Cover Crop Program based on remote sensing. We also have the original physical calibration data that were collected to develop the remote sensing models. Thus we have a perfect opportunity to apply Bayesian regression using paired priors.

This study is observational, so causal relationships cannot be inferred, but this technique could be used in many similar datasets.



Calibration data:

2005-2011; n = 224 fields

Paired collection of biomass and remote imagery in winter, around cover crop dormancy, and in spring, near termination.

5 replicates of 0.5 m² quadrats were weighed and analyzed for shoot N content.

Predictive models were then developed to estimate biomass from NDVI.

Performance data:

2005-2011; n = 9,384 fields

Remote imagery was collected at the same timepoints as the calibration data.

Collection area represented 85,273 hectare·years.

Management variables: sampling season (2), species (8), planting date (3), establishment method (4), previous cash crop (4), commodity program (2).

Model

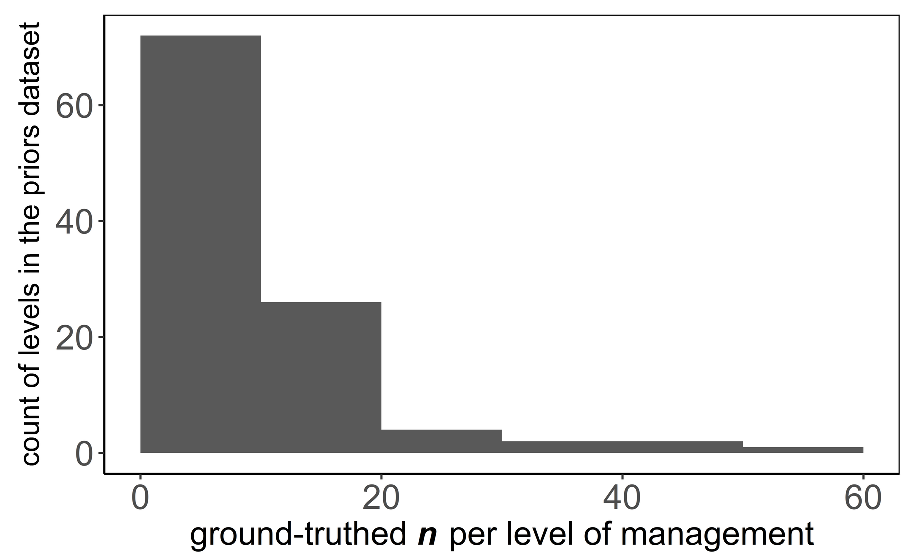
We fit a relatively simple mixed model that considers each combination of management variables as a level (i) of a single factor and each year (j) as a random effect:

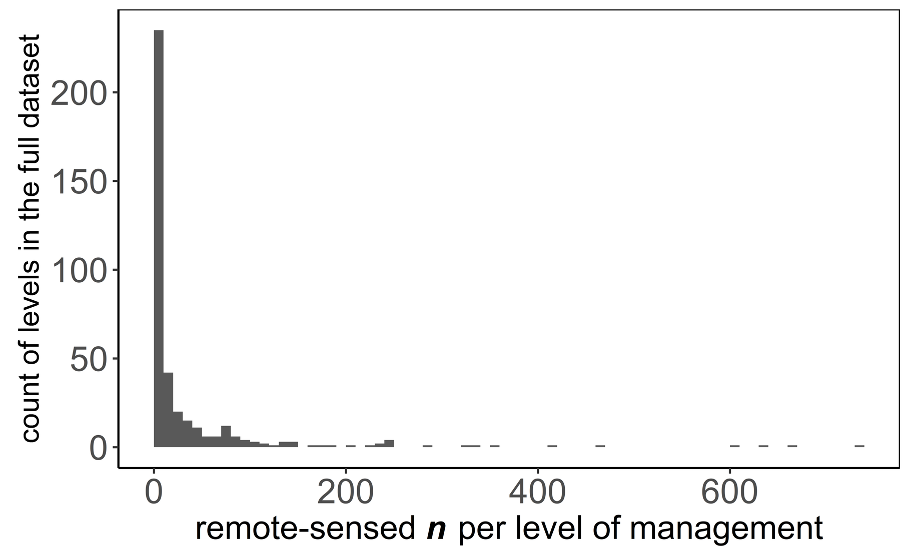
$$\begin{aligned} \log(\text{Biomass}) &\sim \beta_i + b_j + \varepsilon \\ \beta_i &\sim N({}^{\text{prior}}\beta_i,\ 2.5 \cdot {}^{\text{prior}}\sigma_i) \\ b_j &\sim N(0,\ {}^{\text{prior}}b_j) \\ \varepsilon &\sim N(0,\ {}^{\text{prior}}\varepsilon) \end{aligned}$$

The standard deviation of the prior estimates are scaled 2.5× to make them more weakly informative. This scaling factor was chosen arbitrarily based on defaults in `rstanarm`, but results in plausible shrinkage.

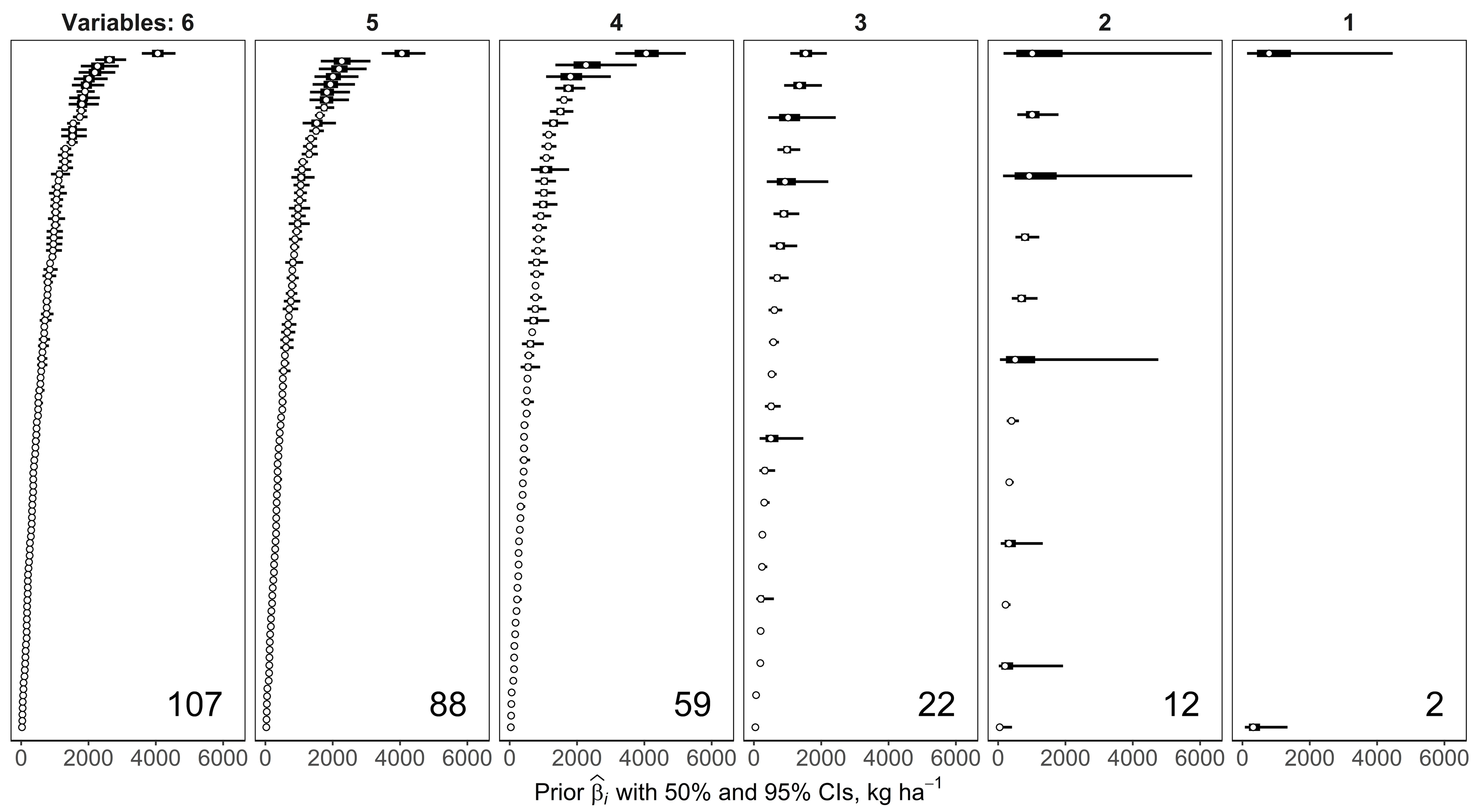
However, not every level of the remote-sensed data is present in the ground-truthed data. Therefore we used a hierarchical algorithm to construct successively simpler, weaker priors by removing management variables from the ground-truthed model.

Highly unbalanced observations





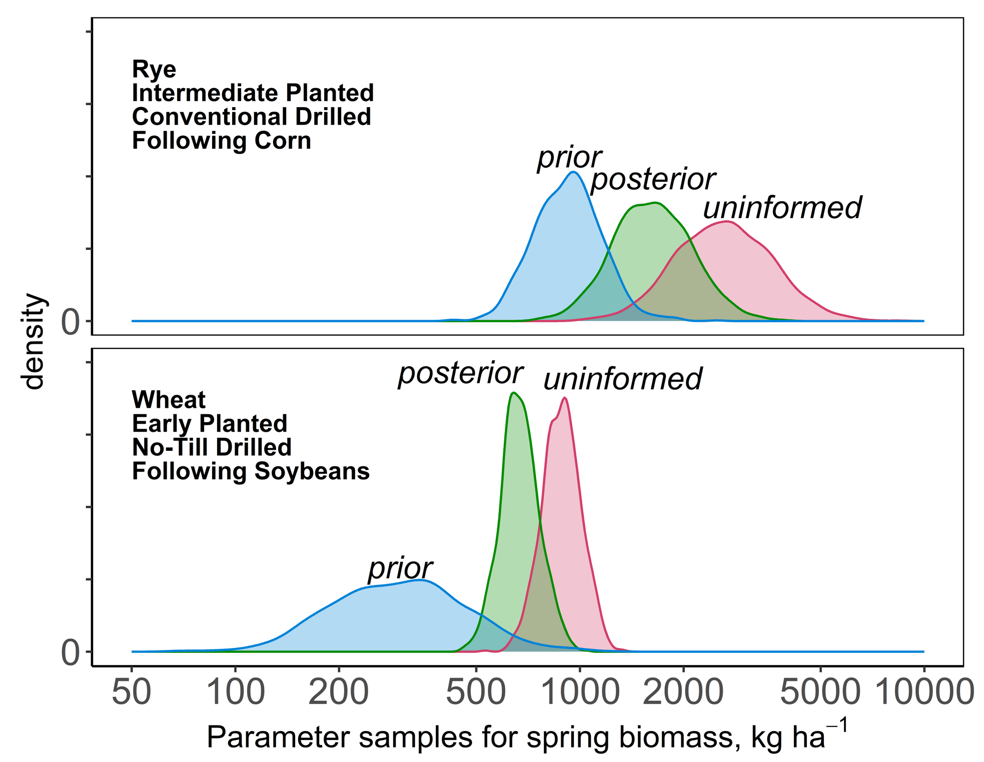
Hierarchically-constructed prior distributions

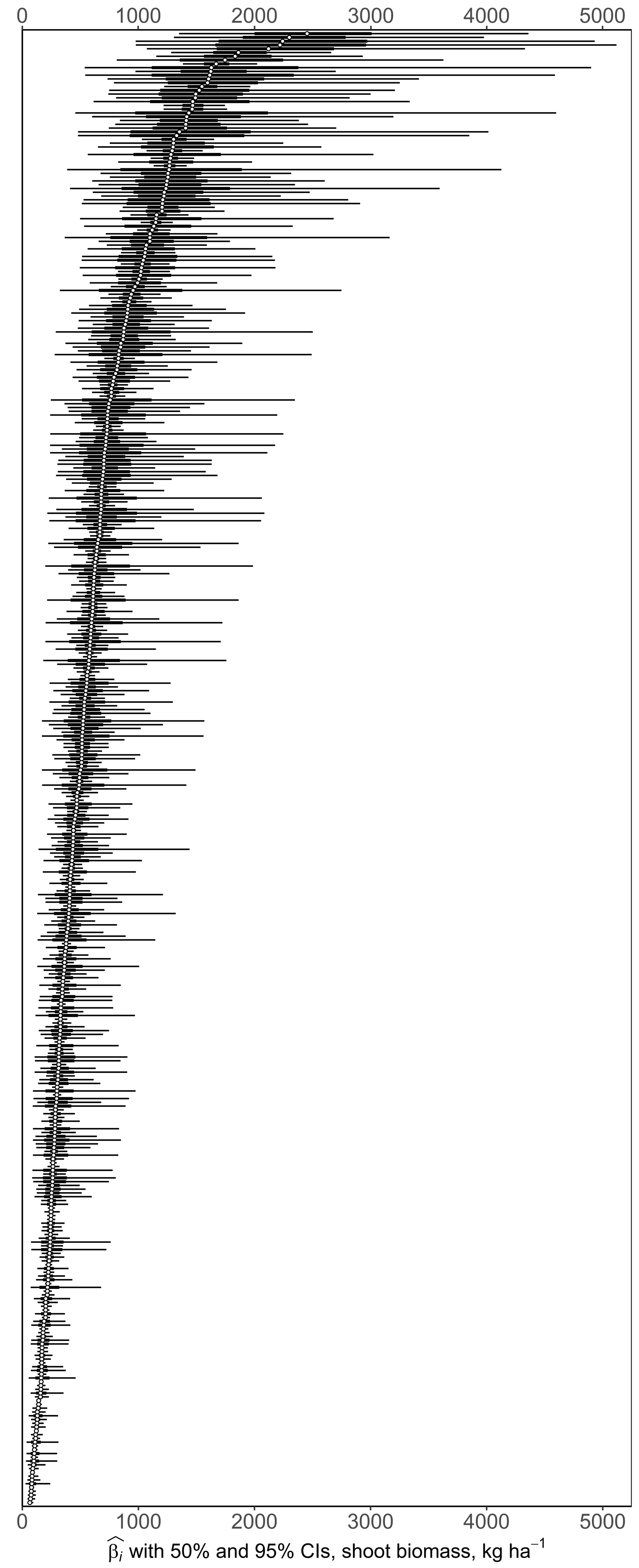


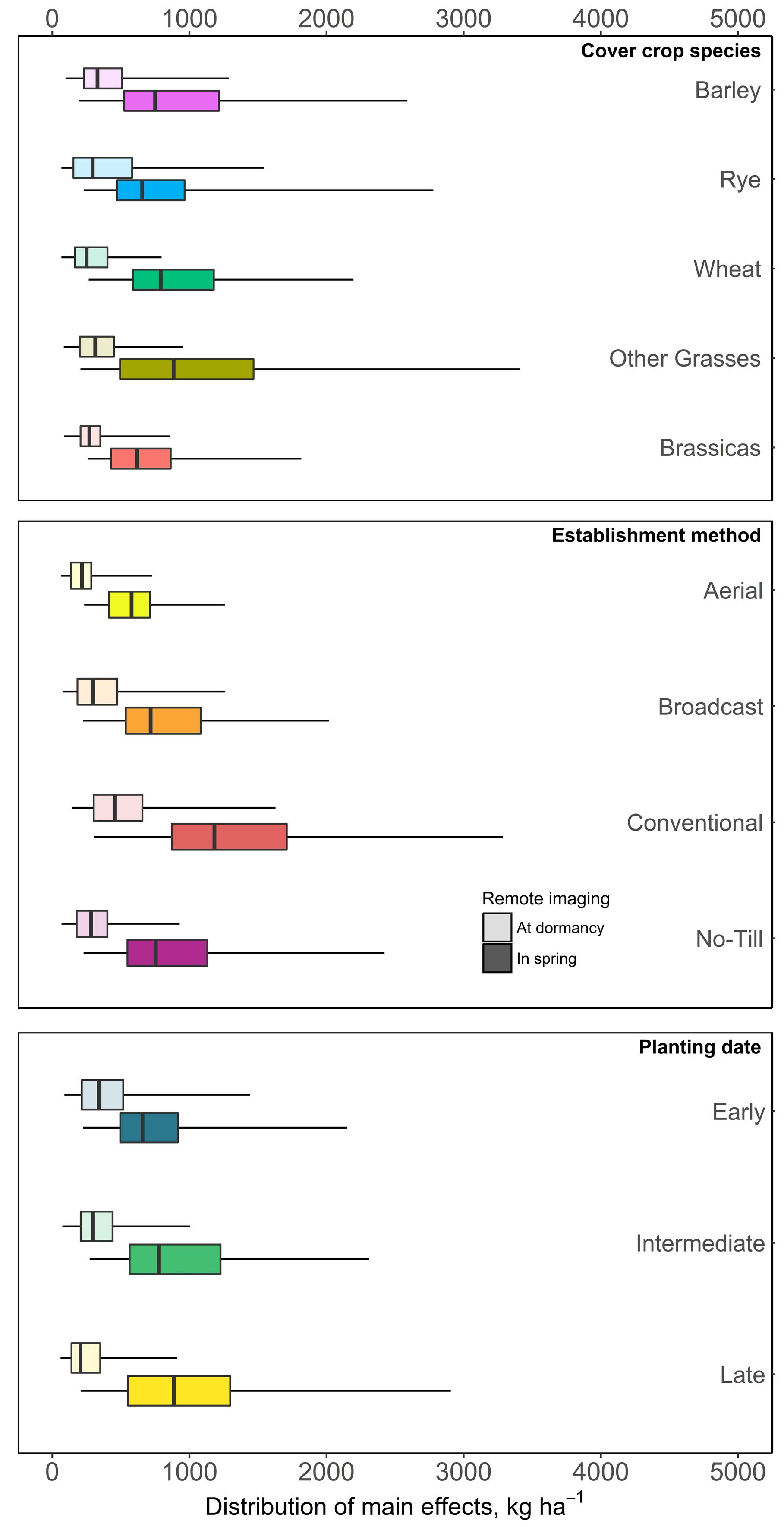
Example parameters

At management combination levels where the calibration data has many observations, or where observations have a small standard deviation, the prior distribution will be strongly informative, and the posterior samples will shrink toward it.

For management combination levels where calibration data was not present, a larger sample of observations was taken from a simpler model. This population has a larger S.D., which generates a weakly informative, nearly flat prior with little shrinkage.







Goodness of fit

In total, 390 coefficients were estimated, ranging from 64.1-2453 kg ha^{-1}.

These coefficients explained over half of the variance in the raw data: 0.538 (0.531, 0.546).

Median RMSE (95%CI) for this model was 0.367 (0.0128, 1.450), while for the equivalent frequentist mixed and OLS models, RMSE was 0.658 and 0.723.

The range of standard errors around the estimates for each coefficient were smaller:

`brms`:	0.253 (0.062, 0.575)
`lmer`:	0.300 (0.168, 0.687)
`lm`:	0.277 (0.046, 0.733)

Conclusions

Calibration data is one of many opportunities in agricultural research to apply Bayesian principles. Nearby long-term trials, regional NASS values, previous runs of an assay, and estimates from literature all make excellent choices for informative priors.

Unbalanced data is especially well suited to this type of analysis, since strong priors can provide context to treatments with small sample numbers.

Multilevel priors can be used to estimate new treatment levels in ongoing experiments, even when that treatment has not been observed previously.

Software packages for Bayesian analysis (such as `brms` with `Stan`) now feature familiar syntax to traditional R analyses.

It's always useful to refit your models using both frequentist and Bayesian approaches to do a plausibility check.