# Report on Fossil/Not Fossil

January 22, 2018

## 1 Set-Up

We want to predict the expected diversity of a Macrostrat geological unit. We are focused on the diversity of individual biological classes. Geological units are described in terms of their lithological descriptions, areal extent, location, "connectedness" (units above/below), and other values. Lithological description is a compositional variable and thus has annoying properties which can prove difficult to interpret.

The dataset is split into two sets: the late Ordovician (460.4–445.6) and the Hirnantian (445.6–443.8); the former is the training dataset while the latter is the testing dataset.

## 2 Model

The basics of the models used in this analysis is a hurdle model which is a mixture of a Bernoulli distribution and either a Poisson or Negative-Binomial distribution. The Bernoulli aspect describes the probability of observing 0 species in a geological unit, while the Poisson or Negative-Binomial describes the expected number of species present in that geological unit if there are more than 0.

Both parts of the mixture are modeled as regressions with all of the geological unit covariates as predictors.

The models are fit to the training dataset and the results of which are used to predict the diversity of the Hirnantian geological units. The approximate expected out-of-sample predictions were evaluated through 5 rounds of 5-fold cross-validation.

The models are fit in a fully Bayesian context using the Stan probabilistic programming language. Model fit for the complete training dataset is evaluated through a series of posterior predictive checks. Additionally, models are compared using WAIC and LOOIC which are estimates of comparative out-of-sample predictive performance.

# 3 Results

## 3.1 Full data posterior predictive checks

figures 1, 2, 3, 4, 5. waic 1 and looic 2.

## 3.2 K-fold cross-validation results

figures 6. CV RMSE estimates 3.

## 3.3 Regression coefficients

figures 7, 8, 9.

## 3.4 Predictive results

figures 10, 11, 12, 13, 14. rmse from prediction table 4.

Table 1

| Taxonomic group | Poisson model WAIC | Poisson model SE WAIC | NegBin model WAIC | NegBin model SE WAIC |
|---|---|---|---|---|
| Bivalvia | 1485.42 | 124.66 | 1317.38 | 102.66 |
| Brachiopoda | 2602.90 | 179.58 | 2134.50 | 123.81 |
| Gastropoda | 2089.17 | 176.01 | 1636.84 | 114.12 |
| Trilobita | 1828.36 | 139.25 | 1585.80 | 105.04 |

Table 2

| Taxonomic group | Poisson model LOOIC | Poisson model SE LOOIC | NegBin model LOOIC | NegBin model SE LOOIC |
|---|---|---|---|---|
| Bivalvia | 1485.45 | 124.66 | 1319.02 | 102.88 |
| Brachiopoda | 2602.95 | 179.58 | 2135.65 | 123.91 |
| Gastropoda | 2089.20 | 176.01 | 1639.43 | 114.45 |
| Trilobita | 1828.38 | 139.25 | 1586.98 | 105.18 |

Table 3

| Taxonomic group | Poisson Model Mean CV RMSE | Poisson Model SD CV RMSE | NegBin Model Mean CV RMSE | NegBin M |
|---|---|---|---|---|
| Bivalvia | 3.51 | 0.13 | 4.18 | 0.08 |
| Brachiopoda | 7.07 | 0.13 | 7.47 | 0.07 |
| Gastropoda | 5.23 | 0.15 | 6.37 | 0.17 |
| Trilobita | 3.52 | 0.07 | 4.36 | 0.05 |

Table 4

| Taxonomic group | Poisson model Mean CV RMSE | Poisson model SD CV RMSE | NegBin model Mean CV RMSE | NegBin m |
|---|---|---|---|---|
| Bivalvia | 2.21 | 1.30 | 5.11 | 2.89 |
| Brachiopoda | 5.05 | 3.14 | 7.81 | 6.18 |
| Gastropoda | 3.24 | 0.89 | 6.37 | 2.20 |
| Trilobita | 1.81 | 0.99 | 4.18 | 1.07 |

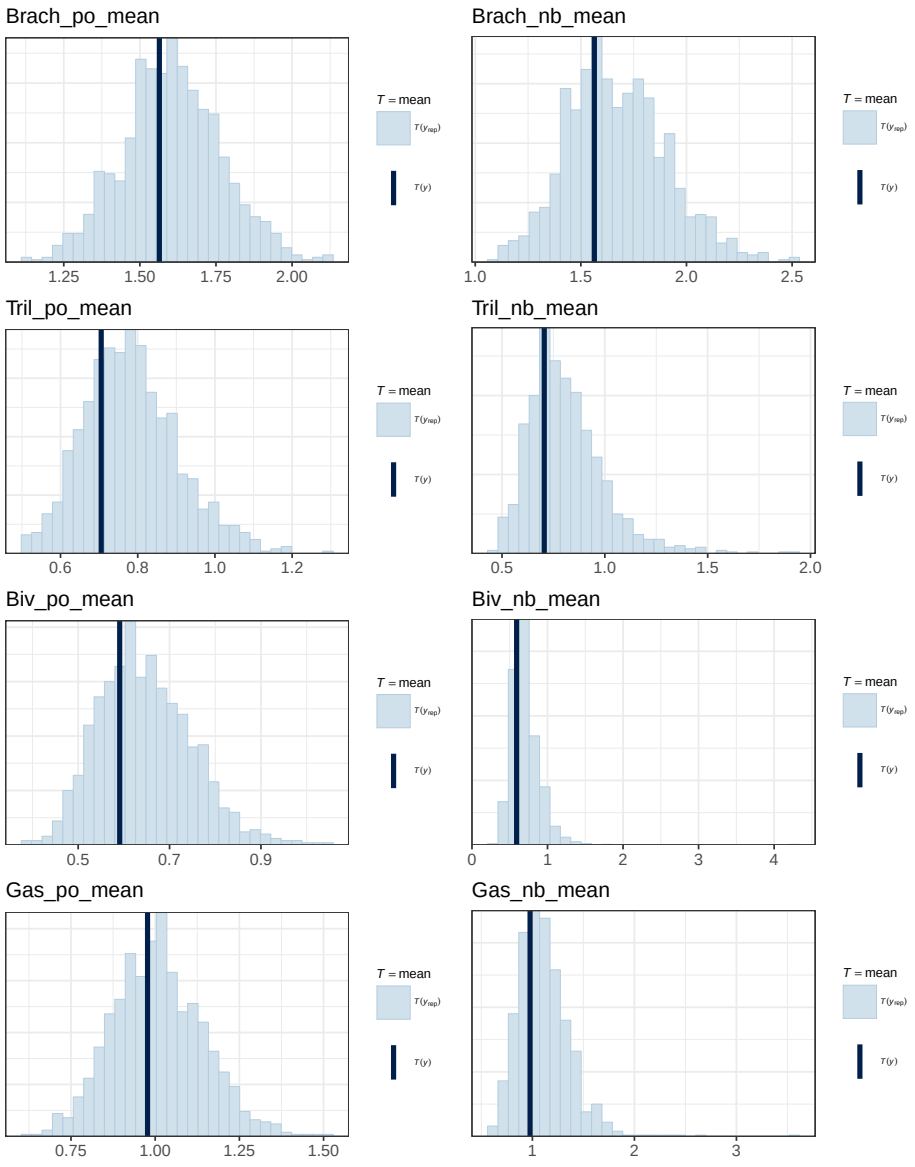


Figure 1: Mean of unit diversity. Distribution is estimated from posterior, vertical line is observed. Columns are by model type, rows are by taxonomic group.

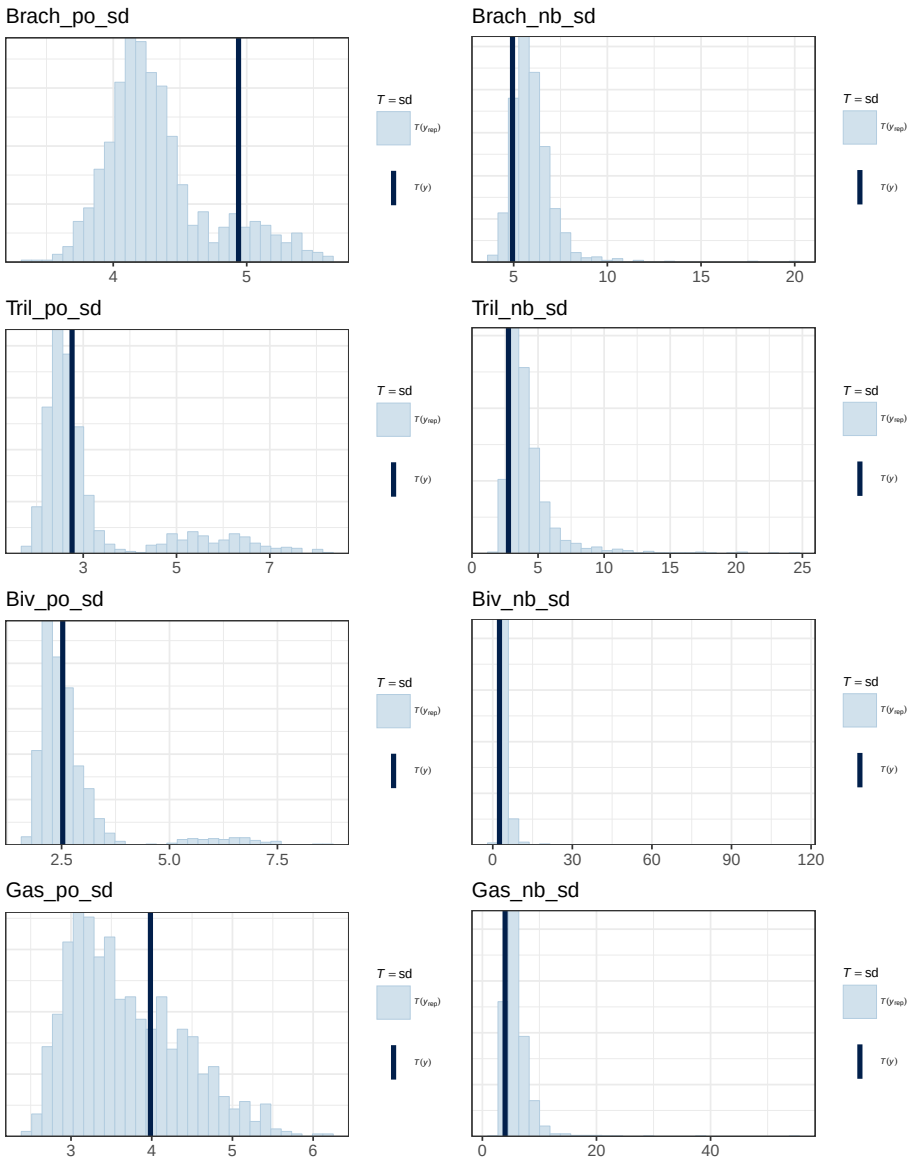


Figure 2: Standard deviation of unit diversity. Distribution is estimated from posterior, vertical line is observed. Columns are by model type, rows are by taxonomic group.

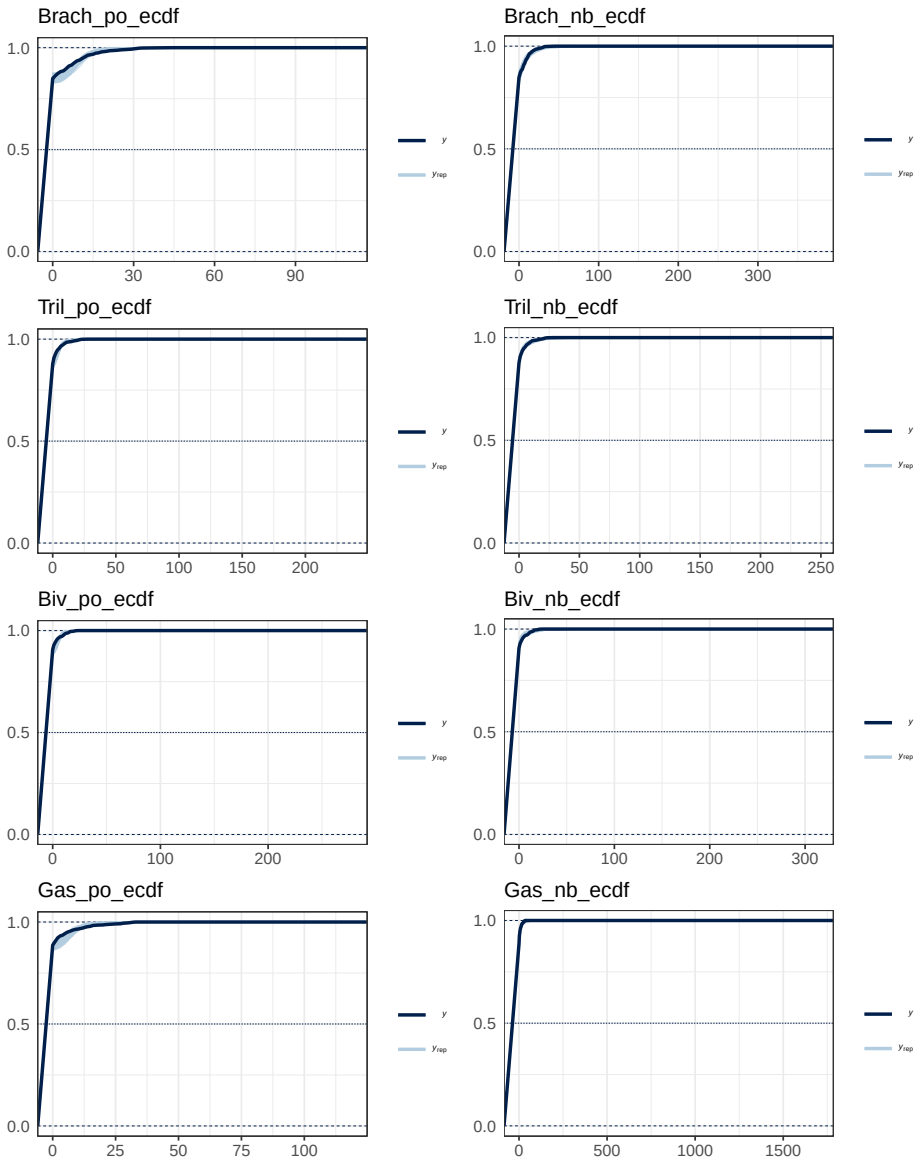


Figure 3: Empirical cumulative distribution function of unit diversity; that is, rank order proportional accumulation. Distribution is estimated from posterior, dark line is observed. Columns are by model type, rows are by taxonomic group.

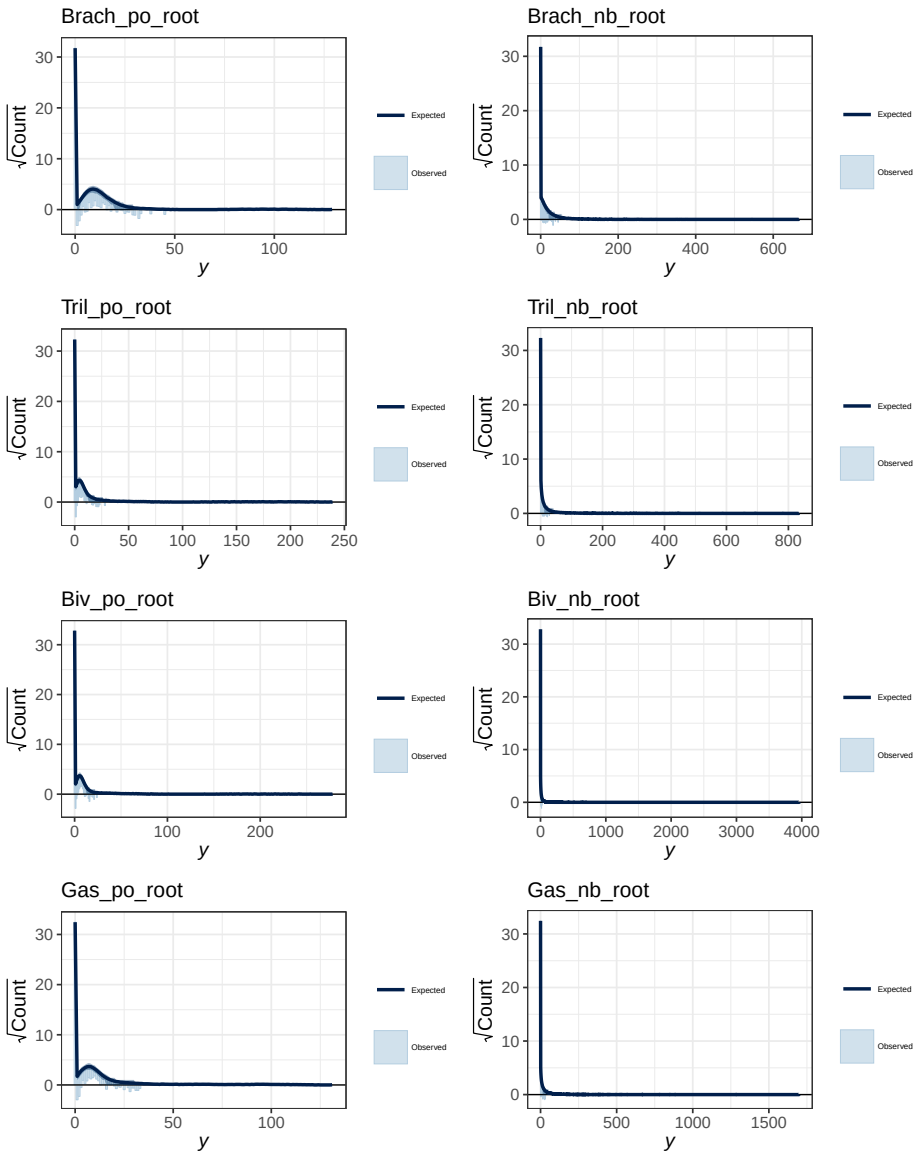


Figure 4: Rootogram of unit diversity. Distribution is estimated from posterior, hanging histogram is is observed. If histogram is above x-axis, overestimate; if histogram is below x-axis, underestimate. Columns are by model type, rows are by taxonomic group.

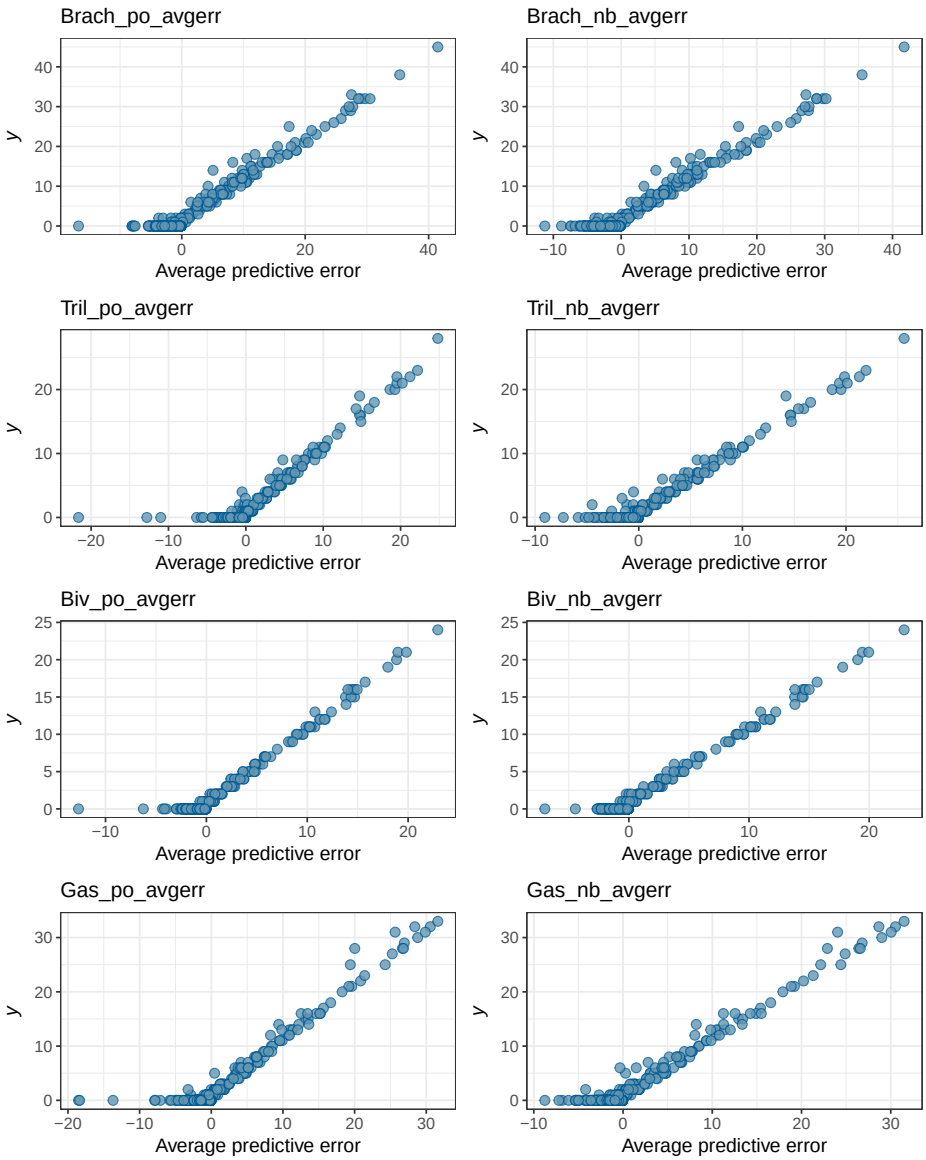


Figure 5: Average error of estimated unit diversity. Comparison is between observed diversity and estimated diversity. Columns are by model type, rows are by taxonomic group.

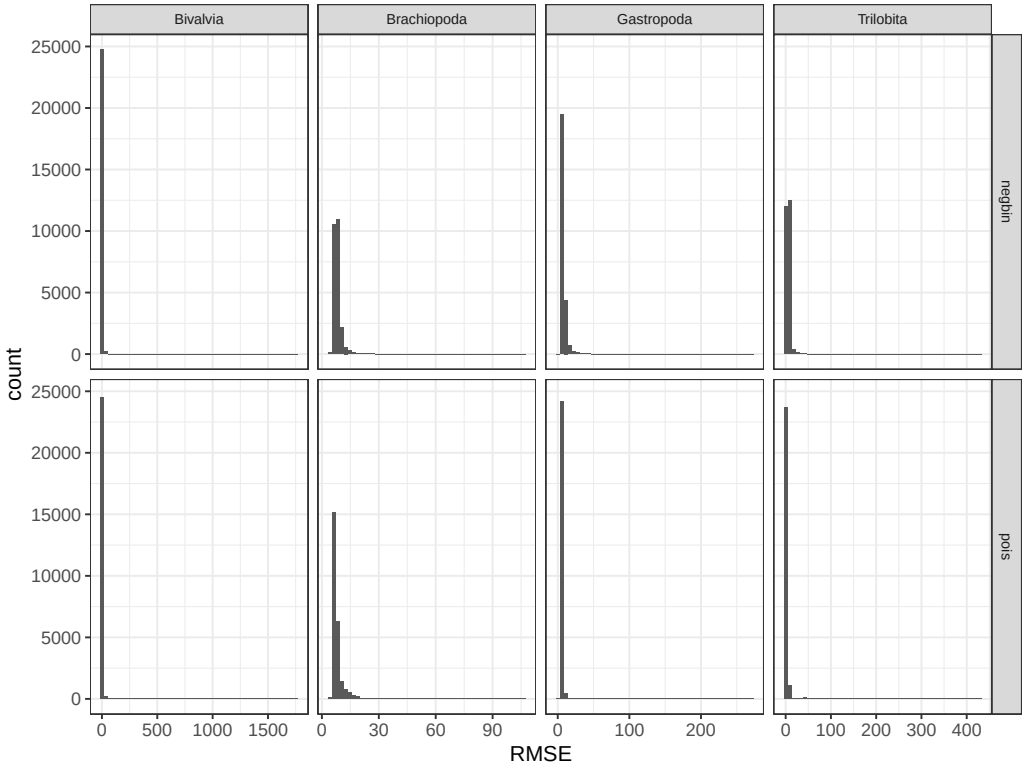


Figure 6: Posterior estimates of the expected out-of-sample error based on 5 rounds of 5-fold cross-validation. Columns are by model type, rows are by taxonomic group.

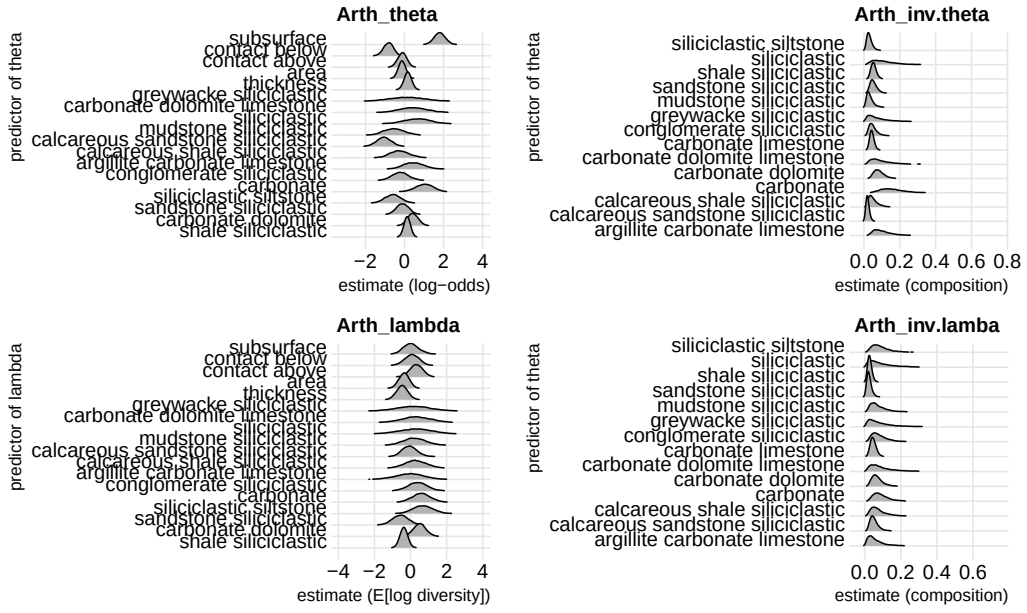


Figure 7: left: coefficients from the arthropod analysis. right: regression coefs for compositional variables back-transformed into composition units. theta is for binomial part (i.e. zeroes), lambda is for poisson or negative binomial part (i.e. non-zeroes)

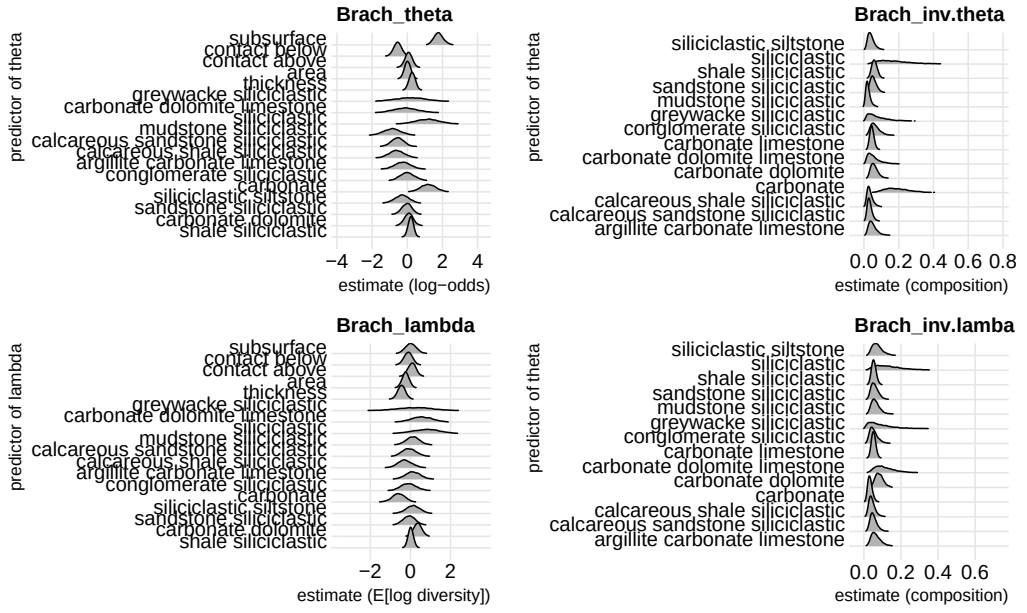


Figure 8: left: coefficients from the brachiopod analysis. right: regression coefs for compositional variables back-transformed into composition units. theta is for binomial part (i.e. zeroes), lambda is for poisson or negative binomial part (i.e. non-zeroes)

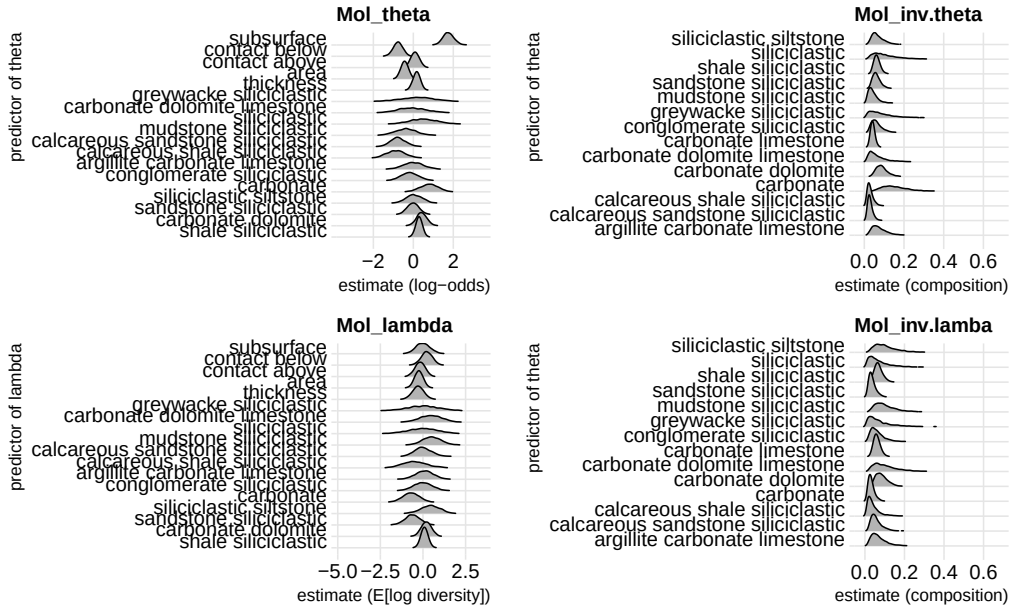


Figure 9: left: coefficients from the mollusc analysis. right: regression coefs for compositional variables back-transformed into composition units. theta is for binomial part (i.e. zeroes), lambda is for poisson or negative binomial part (i.e. non-zeroes)

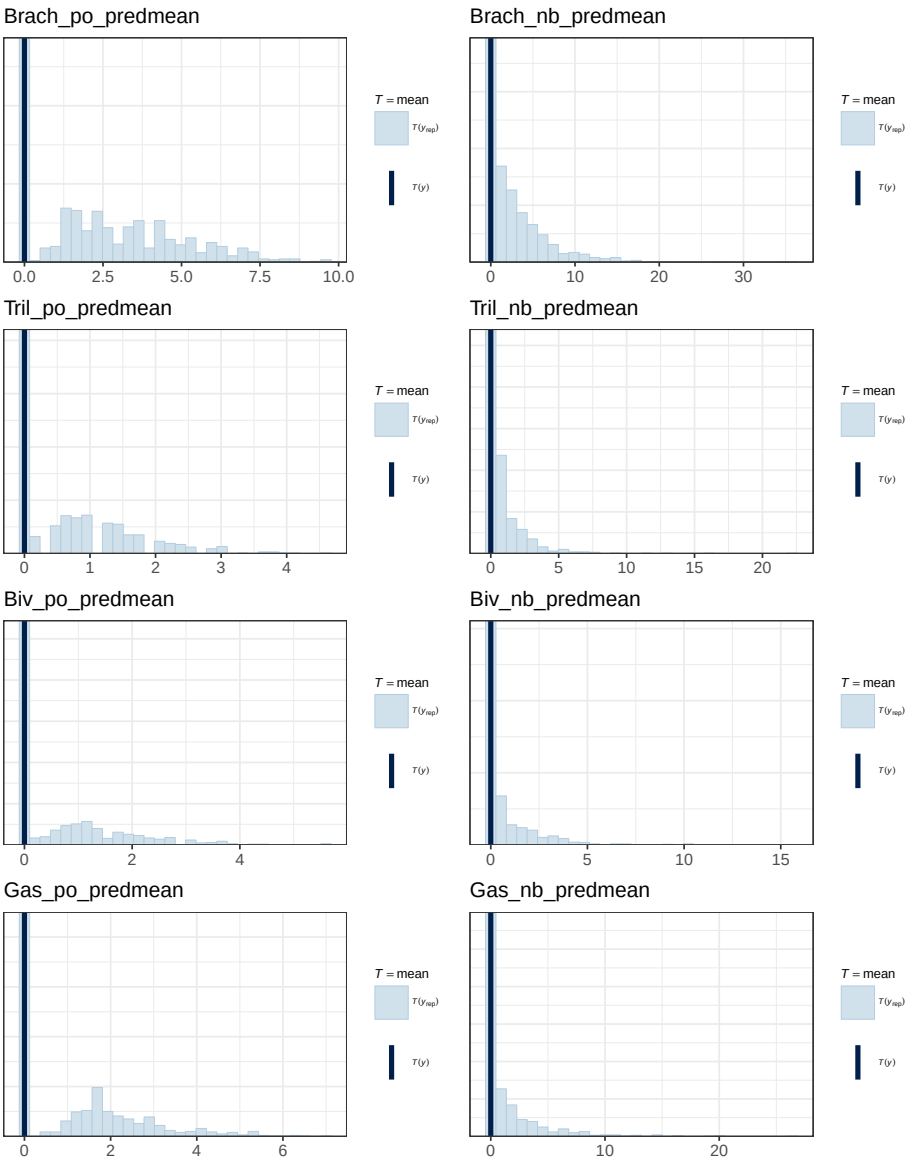


Figure 10: Mean of unit diversity for testing dataset. Distribution is estimated from posterior, vertical line is observed. Columns are by model type, rows are by taxonomic group.

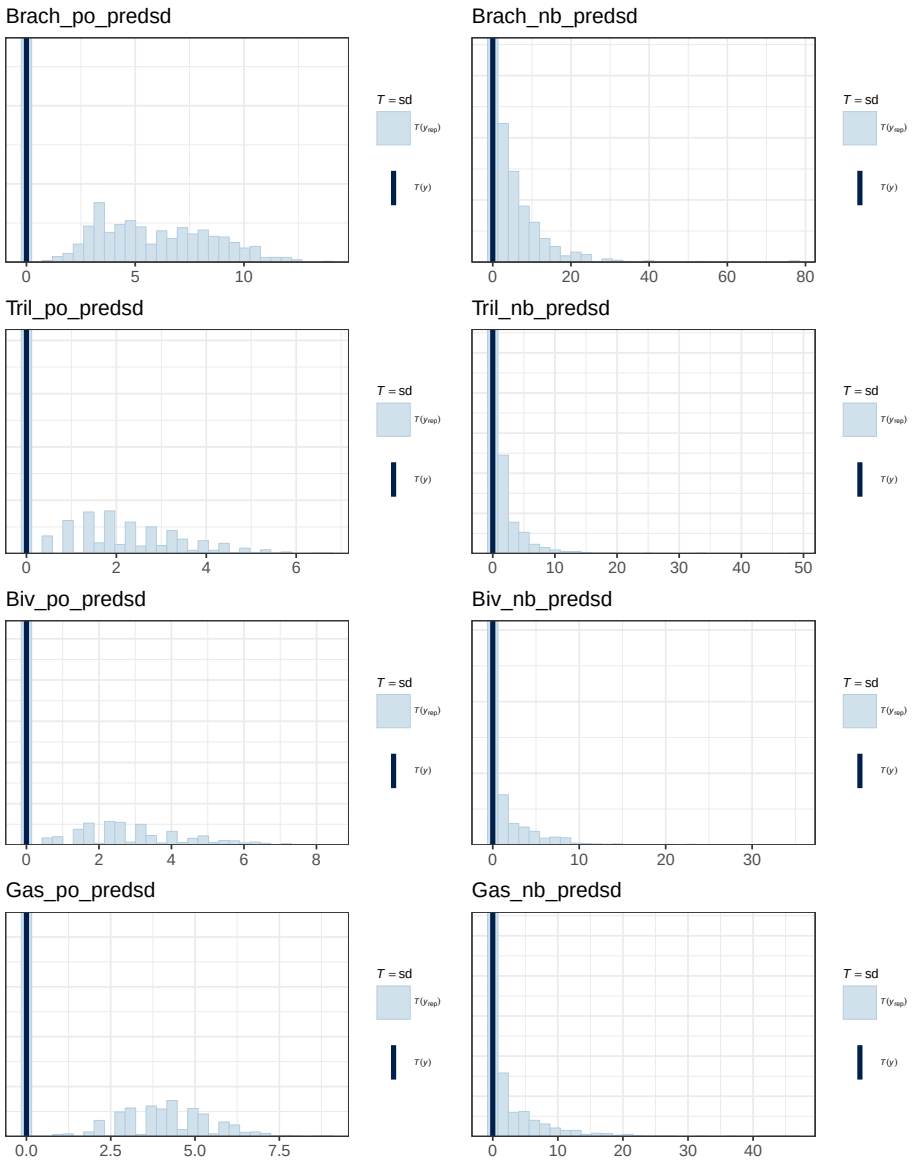


Figure 11: Standard deviation of unit diversity for testing dataset. Distribution is estimated from posterior, vertical line is observed. Columns are by model type, rows are by taxonomic group.

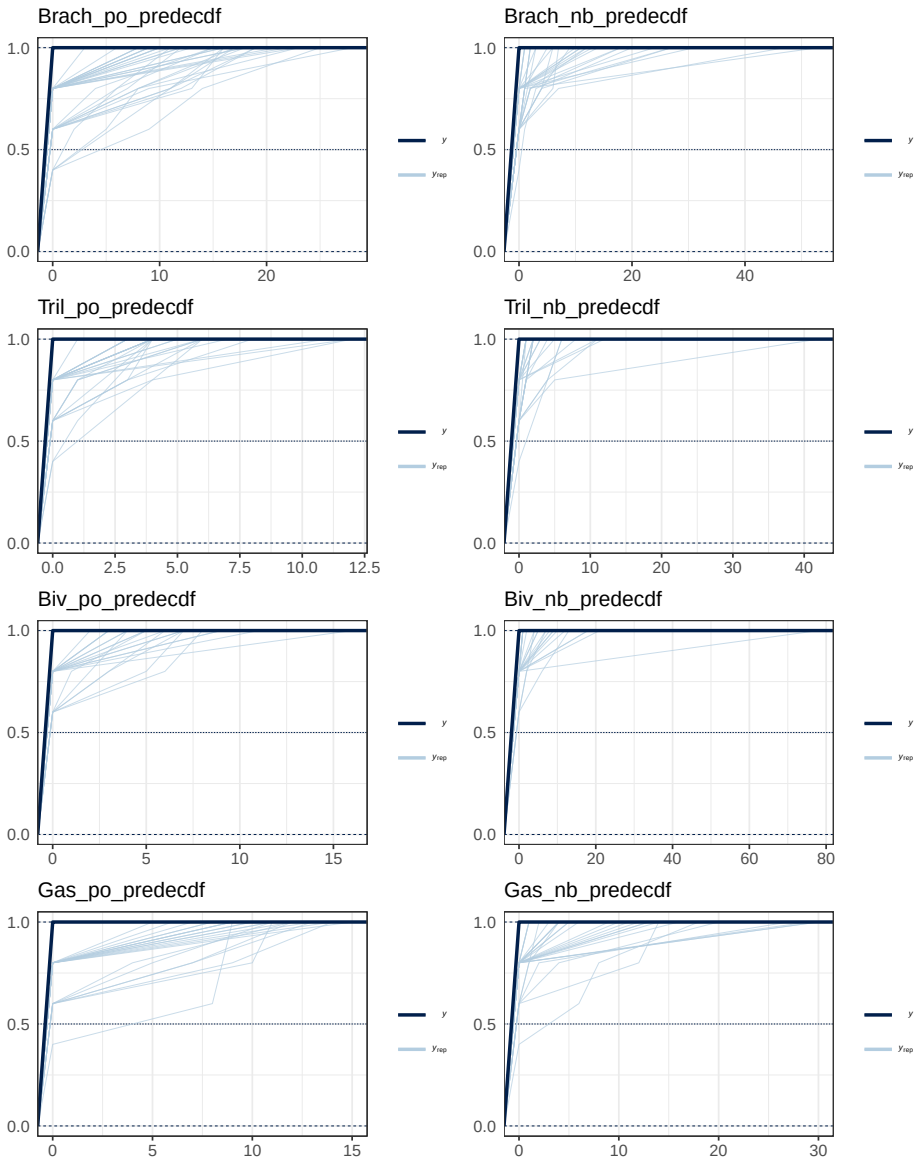


Figure 12: Empirical cumulative distribution function of unit diversity for testing dataset; that is, rank order proportional accumulation. Distribution is estimated from posterior, dark line is observed. Columns are by model type, rows are by taxonomic group.

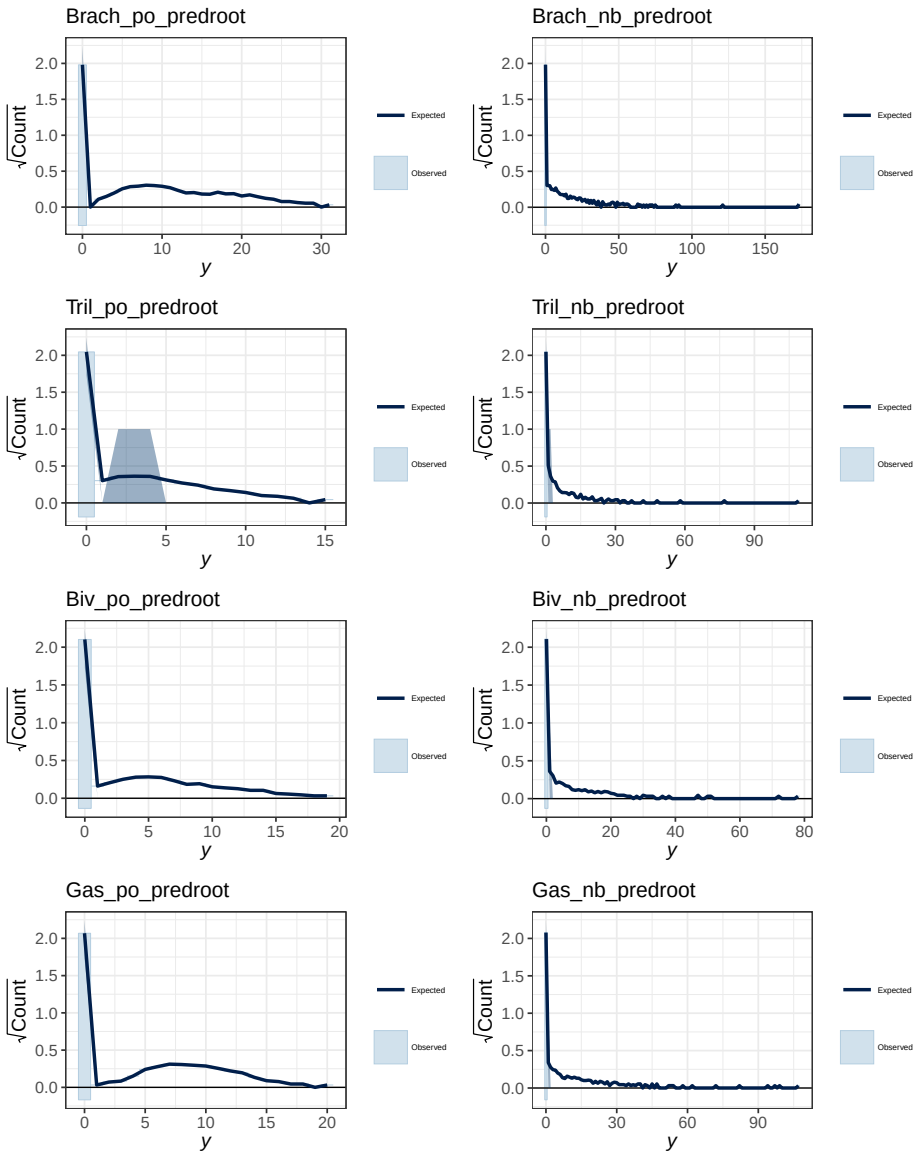


Figure 13: Rootogram of unit diversity for testing dataset. Distribution is estimated from posterior, hanging histogram is is observed. If histogram is above x-axis, overestimate; if histogram is below x-axis, underestimate. Columns are by model type, rows are by taxonomic group.

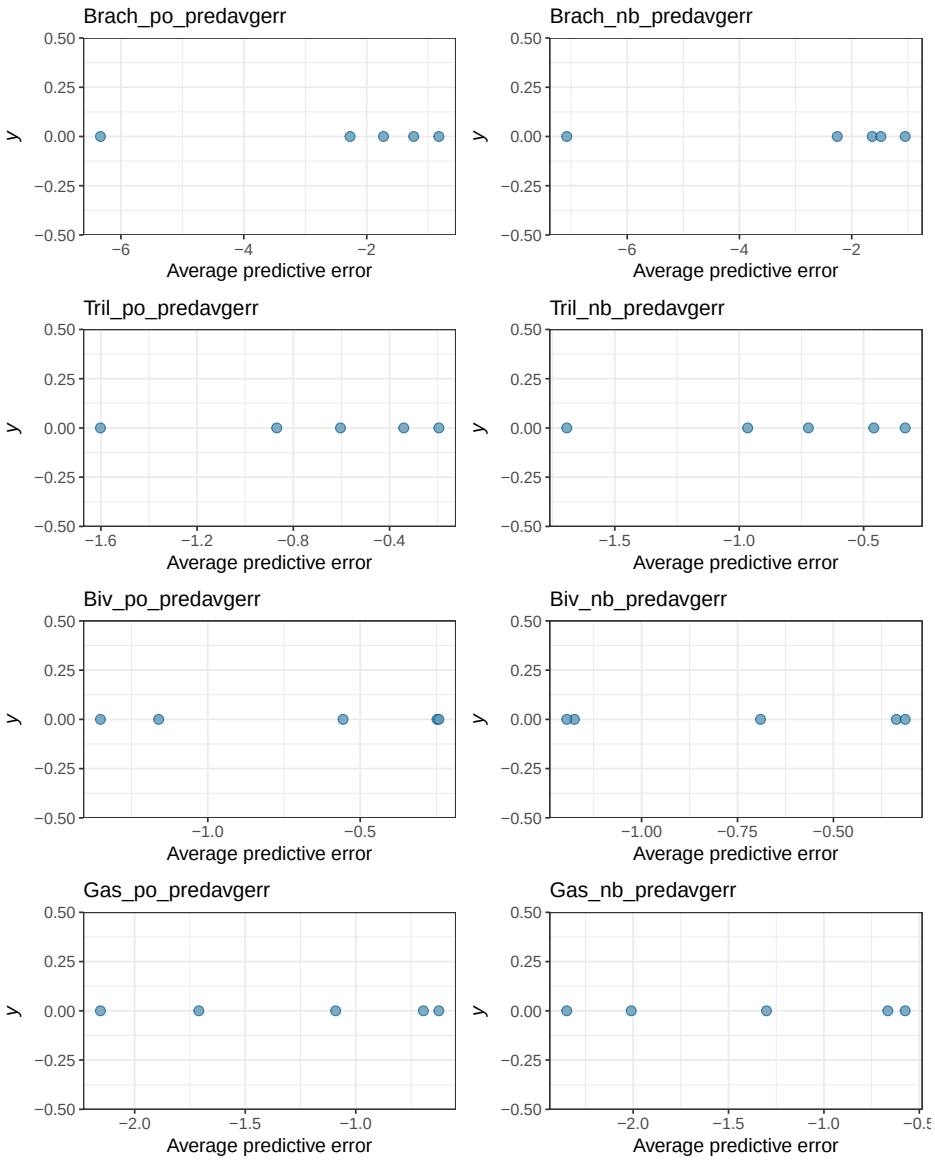


Figure 14: Average error of estimated unit diversity for testing dataset. Comparison is between observed diversity and estimated diversity. Columns are by model type, rows are by taxonomic group.