Vignette: Simulation studies with `ValidationExplorer`

Jacob Oram[1] Katharine Banner[1] Christian Stratton[2] Kathryn M. Irvine[3,*]

2025-01-10

Abstract

Our vignette demonstrates the use of the `ValidationExplorer` package to conduct statistical simulation studies that explore the costs and inferential properties (e.g., near nominal coverage and/or minimal estimation error) of alternative validation designs in the context of the count detection model framework. Our functions allow the user to specify a suite of candidate validation designs using either a stratified sampling procedure or a fixed-effort design type. An example of the former is provided in the manuscript entitled '`ValidationExplorer`: Streamlined simulations to aid informed management decisions using bioacoustic data in the presence of misclassification', which was submitted to the Applications series of *Methods in Ecology and Evolution*. In this vignette, we provide further details not covered in the manuscript and an additional example of data simulation, model fitting, and visualization of simulation results when using a fixed-effort design type. Our demonstration here is intended to aid researchers and others to tailor a validation design that provides useful inference while also ensuring that the level of effort meets cost constraints.

[1] Department of Mathematical Sciences, Montana State University, Bozeman, MT, USA
[2] Department of Mathematics and Statistics, Middlebury College, Middlebury, VT, USA
[3] U.S. Geological Survey, Northern Rocky Mountain Science Center, Bozeman, MT, USA

[*] Correspondence: Kathryn M. Irvine <kirvine@usgs.gov>

Contents

1 Introduction

Automated recording units (ARUs) provide one of the main data sources for many contemporary monitoring programs that aim to provide inference about status and trends for assemblages of species (Loeb et al. 2015). As described in the main text, entitled, "`ValidationExplorer`: Streamlined simulations to inform management decisions using bioacoustic data in the presence of misclassification," substantial practical interest lies in identifying cost-effective validation designs that will allow a monitoring program to obtain its measurable objectives. We believe that statistical simulation studies are a valuable tool for evaluating the relative merits of candidate validation designs prior to gathering and validating ARU data, and it is our goal for `ValidationExplorer` to provide those tools in an approachable package.

This vignette aids exploration of possible validation designs within a count-detection framework using the `ValidationExplorer` package. As described in the main text, a validation design is composed of two parts: a random mechanism for selecting observations to be manually reviewed by experts ("validated"), and a percentage or proportion that controls the sample size. We refer to the random mechanism as the *design type* and the proportion as the level of validation effort (LOVE). In the following section we consider five possible LOVEs and show an example of a fixed effort design.

We emphasize that results from any simulation study – including those produced using the `ValidationExplorer` package – are conditional on the settings and assumptions of the study. In the case of the count-detection model framework that is implemented in `ValidationExplorer`, the assumptions are

1. The occurrence of species within a site are independent; the presence of one species carries no information about the presence or absence of another.
2. For any one species, its occurrence at one location is independent of its occurrence at any other location (independence across sites).
3. Visits to a site (i.e., detector nights) are independent.
4. Recordings within the same site-visit are independent.
5. Validation of one recording does not influence the probability that another will be validated
6. At sites where at least one species is present, the count of recordings generated by each species per night is a Poisson random variable.
7. All species in the assemblage can co-occur and are capable of being confused (no structural zeros in the classification matrix).

8. The configuration of visits to sites is balanced. Note that this is not required to use our method but it is assumed in the simulation studies. For example, the data used by Oram et al. (in submission) had a varying number of detector nights at each site. To model these data, Oram et al. (in submission) made the additional assumption that the number of visits to a site is independent of the presence/absence and expected relative activity of species there. That is, locations where detectors were deployed for more nights were not more or less likely than other sites to be occupied by species of interest and were not more or less likely than other sites to have high levels of activity.

In addition, the simulations produced by `ValidationExplorer` are unbiased, meaning that the model that is fit is the same as the data-generating model.

2 Conducting a simulation study with a fixed effort design type

2.1 What is a fixed effort design?

In this section, we assume a fixed-effort (the design type) validation design, under which $x\%$ of recordings obtained from each visit to a site are validated by experts. The level of validation effort is controlled by the value of x. We begin the process of simulating under this design by defining the real-world objectives and constraints we anticipate.

2.2 Step 0: Define measurable objectives and constraints

Recall from the main text that the first step – before opening R and loading `ValidationExplorer` – is to identify and write down the set of measurable objectives that the data will be used for. Suppose that, for this example, the measurable objectives and cost constraints are the same as those in Section 3 of the main text:

- The measurable objective is to estimate relative activity parameters (denoted as λ_k) for each species with estimation error less than 1 call per night and with the expected width of 95% posterior intervals less than 3 calls per night.
- The monitoring program can pay their expert bat biologists to validate at most 4000 recordings. WE make the assumption that all recordings are approximately the same cost to validate (i.e., rare species autoIDs are not necessarily more expensive or time consuming than extremely common ones).

Table 1: Prior knowledge of relative activity rates and occurrence probabilities for the six bat species of interest. These values will be used to simulate data.

Species	ψ	λ
Eptesicus fuscus (EPFU)	0.63	5.9
Lasiurus cinereus (LACI)	0.61	4.2
Lasionycteris noctivagans (LANO)	0.85	14.3
Myotis californicus (MYCA)	0.70	6.2
Myotis ciliolabrum (MYCI)	0.24	11.9
Myotis evotis (MYEV)	0.70	2.4

Suppose further that the species assemblage is the same as in the main text. That is, we have six species of interest that co-occur. The existing prior knowledge (perhaps from another study) about the relative activity rates and occurrence probabilities for each species are summarized in Table 1.

2.3 Step 1: Installing and loading required packages

Once measurable objectives and constraints are clearly defined, the next step is to load the required packages. For first time users, it may be necessary to install a number of dependencies, as shown by Table 1 in the main text. If you need to install a dependency or update the version, run the following, with `your_package_name_here` replaced by the name of the package:

```
install.packages("your_package_name_here")
```

After installing the necessary packages, load these libraries by calling

```
library(tidyverse)
library(nimble)
library(coda)
library(rstan)
library(parallel)
library(here)
```

Finally, install and load `ValidationExplorer` by running

```
devtools::install_github(repo = "j-oram/ValidationExplorer")
library(ValidationExplorer)
```

With `ValidationExplorer` installed, users have access to all the functions outlined in the table of functions found in Section 5.

2.4 Step 2: Simulate data

The first step in a simulation study is to simulate data under each of the candidate validation designs, which is accomplished with the `simulate_validatedData` function in `ValidationExplorer`. We begin by assigning values for the number of sites, visits, and species, as well as the parameter values in Table 1, which are existing estimates obtained by Stratton et al. (2022):

```
# Set the number of sites, species and visits
nsites <- 30
nspecies <- 6
nvisits <- 4

psi <- c(0.6331, 0.6122, 0.8490, 0.6972, 0.2365, 0.7036)
lambda <- c(5.9347, 4.1603, 14.2532, 6.1985, 11.8649, 2.4050)
```

Additionally, `simulate_validatedData` requires that the user supply misclassification probabilities in the form of a matrix, subject to the constraint that rows in the matrix sum to one. An easy way to simulate a matrix of probabilities that meet these criteria is to leverage the `rdirch` function from the `nimble` package:

```
# Simulate a hypothetical confusion matrix
set.seed(10092024)
Theta <- t(apply(diag(29, nspecies) + 1, 1, function(x) {nimble::rdirch(alpha = x)}))
```

Note that the above definition of `Theta` places high values on the diagonal of the matrix, corresponding to a high probability of correct classification. To lower the diagonal values, change the specification of `diag(29, nspecies)` to a smaller value. For example:

```
another_Theta <- t(apply(diag(5, nspecies) + 1, 1, function(x) {
    nimble::rdirch(alpha = x)
  }))
```

`another_Theta` has lower values on the diagonal, and greater off-diagonal values (i.e., higher probability of misclassification). If you have specific values you would like to use for the assumed classification probabilities (e.g., from an experiment), these can be supplied manually:

```
manual_Theta <- matrix(c(0.9, 0.05, 0.01, 0.01, 0.02, 0.01,
                         0.01, 0.7, 0.21, 0.05, 0.02, 0.01,
                         0.01, 0.01, 0.95, 0.01, 0.01, 0.01,
                         0.05, 0.05, 0.03, 0.82, 0.04, 0.01,
                         0.01, 0.015,  0.005,  0.005, 0.95, 0.015,
                         0.003, 0.007, 0.1, 0.04, 0.06, 0.79),
```

```
                        byrow = TRUE, nrow = 6)

print(manual_Theta)
```

```
##       [,1]  [,2]  [,3]  [,4] [,5]  [,6]
## [1,] 0.900 0.050 0.010 0.010 0.02 0.010
## [2,] 0.010 0.700 0.210 0.050 0.02 0.010
## [3,] 0.010 0.010 0.950 0.010 0.01 0.010
## [4,] 0.050 0.050 0.030 0.820 0.04 0.010
## [5,] 0.010 0.015 0.005 0.005 0.95 0.015
## [6,] 0.003 0.007 0.100 0.040 0.06 0.790
```

If you define the classifier manually, make sure the rows sum to 1 by running

```
all(rowSums(manual_Theta) == 1) # want this to return TRUE
```

```
## [1] TRUE
```

```
# If the above returns FALSE, see which one is not 1:
rowSums(manual_Theta)
```

```
## [1] 1 1 1 1 1 1
```

With the required inputs defined, we can simulate data:

```
sim_data <- simulate_validatedData(
    n_datasets = 10, # For demonstration -- use 50+ for real simulation studies
    nsites = nsites,
    nvisits = nvisits,
    nspecies = nspecies,
    design_type = "FixedPercent",
    scenarios = c(0.05, 0.1, 0.15, 0.3),
    psi = psi,
    lambda = lambda,
    theta = Theta,
    save_datasets = FALSE, # default value is FALSE
    save_masked_datasets = FALSE, # default value is FALSE
    directory = here::here("Vignette", "Fixed_Effort")
)
```

Note that we specified the design type through the argument `design_type = "FixedPercent"`, with the possible scenarios defined as the vector `scenarios = c(0.05, 0.1, 0.15, 0.3)`. These two arguments specify the set of alternative validation designs we will compare in our simulation study. Under the first validation design, 5% of recordings from each visit to a site are validated, while in the second validation design 10% of recordings from each visit to a site are validated, and so on.

To understand the output from `simulate_validatedData`, we can investigate `sim_data`. The output is a list, containing three objects:

```
length(sim_data)
```

```
## [1] 3
```

```
names(sim_data)
```

```
## [1] "full_datasets" "zeros"         "masked_dfs"
```

We examine each of these objects below:

- `full_datasets`: A list of length `n_datasets` with unmasked datasets. These are the datasets if all recordings were validated so that every recording has an autoID and a true species label. We opted to not save these datasets to the working directory by setting `save_datasets = FALSE`. If we had specified `save_datasets = TRUE`, then these will be saved individually in `directory` as `dataset_n.rds`, where `n` is the dataset number. As an example of one element in `sim_data$full_datasets`, we examine the third simulated full dataset:

```
full_dfs <- sim_data$full_datasets
head(full_dfs[[3]]) # Dataset number 3 if all recordings were validated
```

```
## # A tibble: 6 x 10
## # Groups:   site, visit [1]
##    site visit true_spp id_spp lambda   psi  theta     z count     Y.
##   <int> <int>    <int>  <int>  <dbl> <dbl>  <dbl> <int> <int> <int>
## 1     1     1        1      1   5.93 0.633 0.954      1     4    24
## 2     1     1        1      1   5.93 0.633 0.954      1     4    24
## 3     1     1        1      1   5.93 0.633 0.954      1     4    24
## 4     1     1        1      1   5.93 0.633 0.954      1     4    24
## 5     1     1        4      2   6.20 0.697 0.0289     1     1    24
## 6     1     1        3      3  14.3  0.849 0.826      1    10    24
```

Notice that in addition to the site, visit, true species and autoID (`id_spp`) columns, the parameter values (`lambda`, `psi`, and `theta`) are given for each true species-autoID combination. In addition, the occupancy state `z` for the true species is given and the `count` of calls at that site visit with a specific true species-autoID label combination. For example, at site 1, visit 1, there are four calls from species 1 that are assigned autoID 1, yielding 4 rows with `count = 4`. There is also one call from species 4 that was assigned a species 2 label with probability 0.02893559. Because this happened once, it is documented with `count = 1` and only

occupies a single row. Next, we can see that there were 10 calls that were correctly identified as species 3; ten rows will have `true_spp = 3` and `autoID = 3`. Finally, the `Y.` column tells us how many observations were made from all species at that site visit; for visit 1 to site 1, the unique value is 24. That is, the 25th row of this dataset will contain the first observation from visit 2 to site 1.

- `zeros`: A list of length `n_datasets` containing the true species-autoID combinations that were never observed at each site visit. The `count` column, which, again, contains the count of each site-visit-true_spp-id_spp combination, is zero for all entries. For example, in dataset 3 (below), species 2 was not present at site 1, so it could not be classified as species 1 (first row). Additionally, species 3 was present at site 1, but it was never classified as species 1 on visit 1 (second row). These zero counts are necessary for the model to identify occurrence probabilities and relative activity rates. If `save_datasets = TRUE`, the zeros for each dataset will also be saved in `directory` individually as `zeros_in_dataset_n.rds`, where `n` is the dataset number.

```
zeros <- sim_data$zeros

# The site-visit-true_spp-autoID combinations that were never observed in
# dataset 3. Notice that count = 0 for all rows!
head(zeros[[3]])
```

```
## # A tibble: 6 x 10
## # Groups:   site, visit [1]
##    site visit true_spp id_spp lambda   psi   theta     z count     Y.
##   <int> <int>    <int>  <int>  <dbl> <dbl>   <dbl> <int> <int> <int>
## 1     1     1        2      1   4.16 0.612 0.0286      0     0    24
## 2     1     1        3      1  14.3  0.849 0.0137      1     0    24
## 3     1     1        4      1   6.20 0.697 0.0426      1     0    24
## 4     1     1        5      1  11.9  0.236 0.00559     0     0    24
## 5     1     1        6      1   2.40 0.704 0.00312     1     0    24
## 6     1     1        1      2   5.93 0.633 0.0159      1     0    24
```

- `masked_dfs`: A nested list containing each dataset masked under each scenario. For example, `masked_dfs[[4]][[3]]` contains dataset 3, assuming that it was validated according to scenario 4 (30% of recordings randomly sampled from each site-visit for validation). If `save_masked_datasets = TRUE`, then each dataset/scenario combination is saved individually in `directory` as `dataset_n_masked_under_scenario_s.rds`, where `n` is the dataset number and `s` is the scenario number.

```
masked_dfs <- sim_data$masked_dfs

# View dataset 3 subjected to the validation design in scenario 4:
# randomly select and validate 30% of recordings from the first visit
```

```
# to each site
head(masked_dfs[[4]][[3]], 10)
```

```
## # A tibble: 10 x 12
##     site visit true_spp id_spp lambda   psi  theta     z count    Y.
##    <int> <int>    <int>  <int>  <dbl> <dbl>  <dbl> <int> <int> <int>
##  1     1     1       NA      1   5.93 0.633 0.954      1     4    24
##  2     1     1       NA      1   5.93 0.633 0.954      1     4    24
##  3     1     1       NA      1   5.93 0.633 0.954      1     4    24
##  4     1     1       NA      1   5.93 0.633 0.954      1     4    24
##  5     1     1       NA      2   6.20 0.697 0.0289     1     1    24
##  6     1     1        3      3  14.3  0.849 0.826      1    10    24
##  7     1     1        3      3  14.3  0.849 0.826      1    10    24
##  8     1     1        3      3  14.3  0.849 0.826      1    10    24
##  9     1     1       NA      3  14.3  0.849 0.826      1    10    24
## 10     1     1        3      3  14.3  0.849 0.826      1    10    24
## # i 2 more variables: unique_call_id <chr>, scenario <int>
```

The name of this nested list comes from the way that validation effort is simulated: recordings that are not selected for validation have their `true_spp` label masked with an `NA`. Notice that in the example output, all entries in the dataset are identical to the unmasked version output in `full_dfs` above, with the exception of the `true_spp` column. From this column we can see that calls 1-5 and 9 were not selected for validation (because `true_spp = NA`), while recordings 6-8 and 10 were (because the true species label is not marked as `NA`).

2.4.1 Summarize validation effort

For most simulations, it will be useful to summarize the number of recordings that are validated under a given validation design and scenario. This can be accomplished using the `summarize_n_validated` function:

```
summarize_n_validated(
  data_list = sim_data$masked_dfs,
  theta_scenario = "1",
  scenario_numbers = 1:4
)
```

```
## # A tibble: 4 x 3
##   theta_scenario scenario n_validated
##   <chr>          <chr>          <dbl>
## 1 1              1               218.
## 2 1              2               375.
## 3 1              3               540.
## 4 1              4              1021.
```

We can see here that any of the validation designs considered in our simulations will remain well within budget.

2.5 Step 3: MCMC tuning

Running a complete simulation study can be time consuming. In an effort to help users improve the efficiency of their simulations, we provide the `tune_mcmc` function, which provides the user with information about the warmup and number of iterations required for the MCMC to reach approximate convergence. This function takes in a masked dataset and the corresponding zeros, fits a model to these data, and outputs an estimated run time for 10,000 iterations, as well as the estimated number of required warmup and total iterations. These are intended to assist tuning of the MCMC algorithm, which is done by the user in the following steps, which we walk through in greater detail below:

1. Use `tune_mcmc` to fit a model with multiple long chains.
2. Create trace plots for all model parameters.
3. Examine effective sample sizes n_{eff} and Gelman-Rubin statistics $\hat{R}$ for all parameters.
4. Choose values for the number of iterations and warmup that are slightly larger than what is needed based on steps 1-3. This may help ensure that a greater number of model fits are available to inform simulation study results.

2.5.1 Fit a model

As in the main text, we use a dataset from the scenario with the lowest number of validated recordings, as we expect the greatest number of iterations for this scenario. In our example, this is scenario 1 , in which an average of ≈ 218 recordings are validated per dataset (Section 2.4.1).

```
scenario_number <- 1
dataset_number <- sample(1:length(masked_dfs[[scenario_number]]), 1)

tune_list <- tune_mcmc(
  dataset = sim_data$masked_dfs[[scenario_number]][[dataset_number]],
  zeros = sim_data$zeros[[dataset_number]],
)
```

```
## [1] "Fitting MCMC in parallel ... this may take a few minutes"
```

The output from `tune_mcmc` is a list containing draws from the fitted model, the time required to fit the model with 10,000 draws, MCMC diagnostics and guesses for the number of iterations and warmup required to reliably fit a model. If the guessed values for total iterations and/or warmup are greater than 10,000 draws, an error is issued. We can see the names for each object by running the following block:

```r
names(tune_list)
```

```
## [1] "max_iter_time"    "min_warmup"       "min_iter"         "fit"
## [5] "MCMC_diagnostics"
```

The first element is the time required to fit a model with three chains of 10,000 iterations each:

```r
tune_list$max_iter_time
```

```
## Time difference of 2.11137 mins
```

This may seem insignificant, but over the course of an entire simulations study with 5 scenarios $\times$ 50 datasets, that corresponds to around 8 hours of run time. Using fewer than 10,000 iterations will substantially reduce the time to run a simulation study.

2.5.2 Create trace plots

To decide the number of iterations and warmup for use in a simulation study, we recommend beginning by creating trace plots, which show how the sampled values of a parameter evolve over the course of each Markov chain. To ensure the MCMC algorithm will characterize the posterior distribution well, we need to check that chains are stationary and mixing well, and that effective sample sizes are sufficiently large to characterize the posterior. Trace plots are especially useful for assessing the first of these. To create a trace plot using the bayesplot package (Gabry et al. 2019) for a single parameter, run the following:

```r
# Load bayesplot package specifically designed for visualizing
library(bayesplot)
```

```
## This is bayesplot version 1.11.1
```

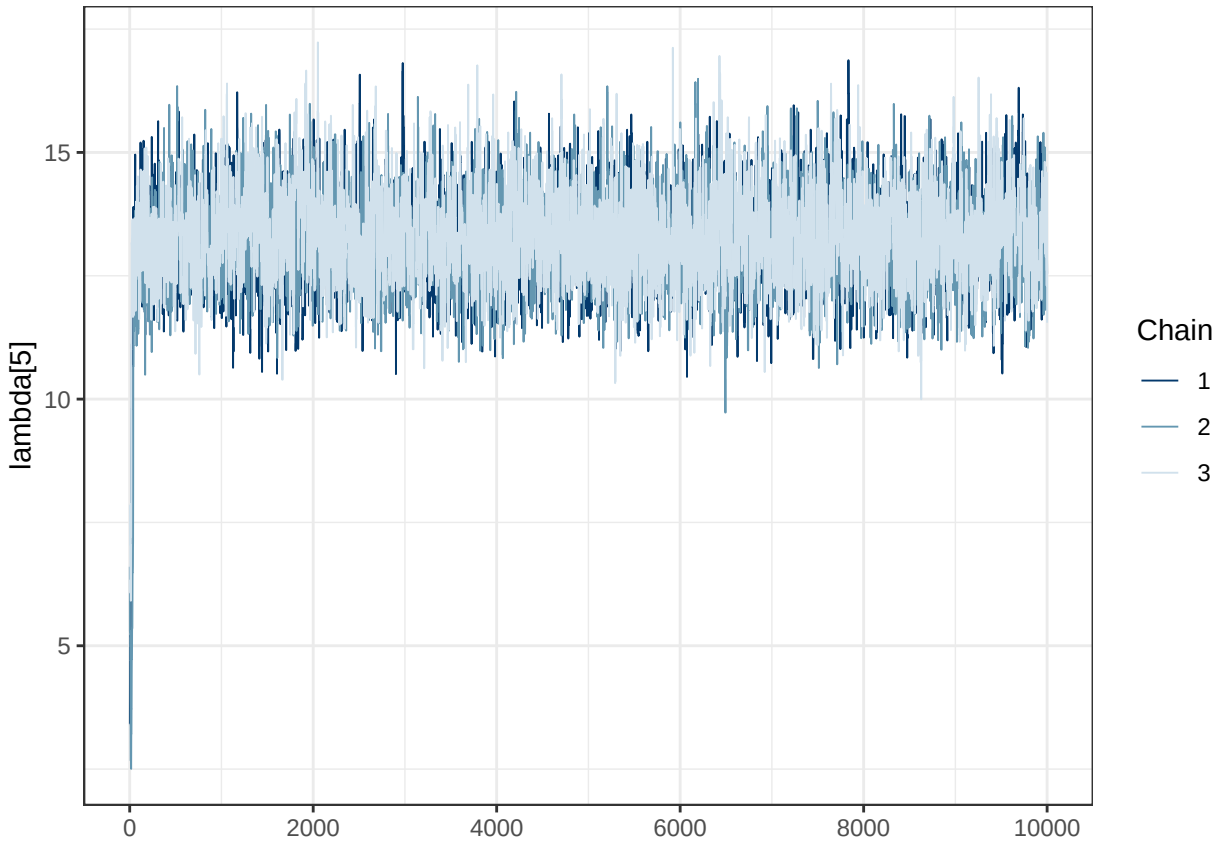
```
## - Online documentation and vignettes at mc-stan.org/bayesplot
```

```
## - bayesplot theme set to bayesplot::theme_default()
```

```
##    * Does _not_ affect other ggplot2 plots
```

```
##    * See ?bayesplot_theme_set for details on theme setting
```

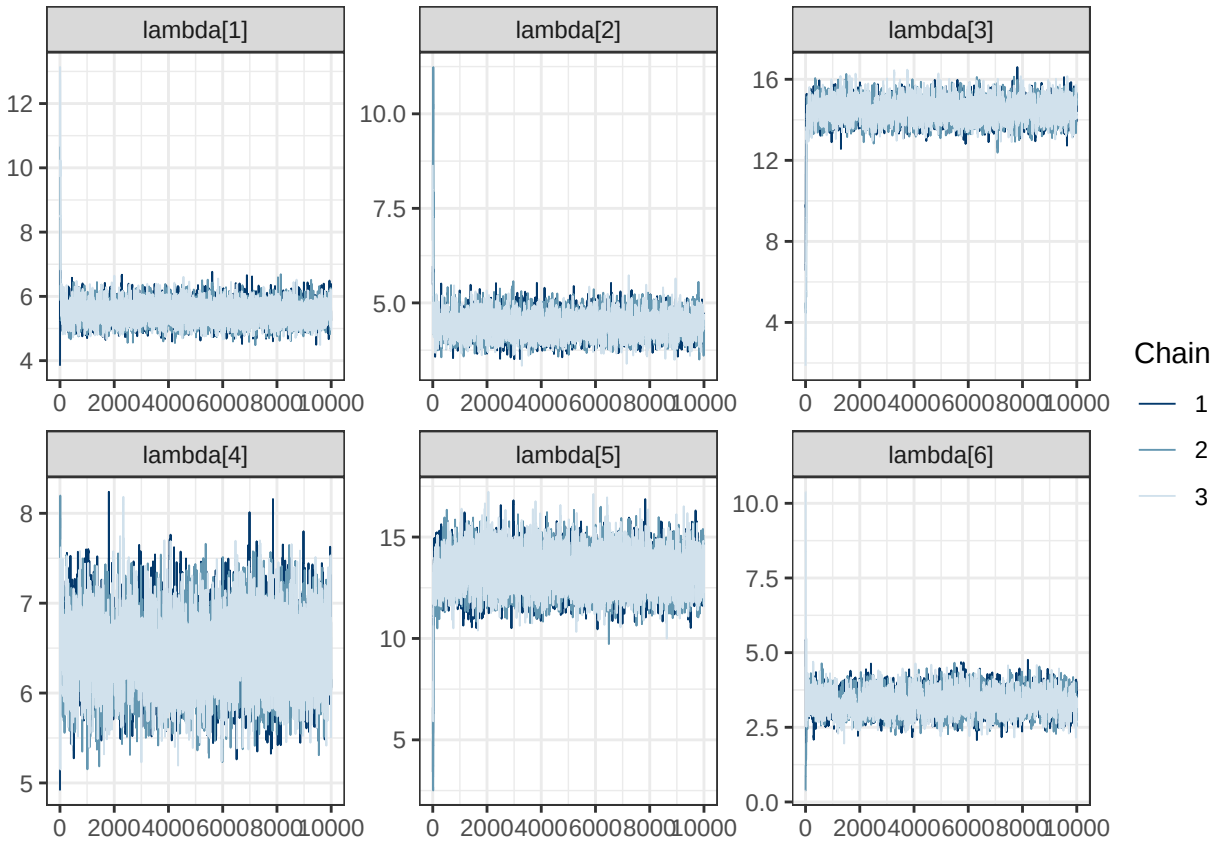
```r
# extract the fitted model from tune_mcmc output
fit <- tune_list$fit

# create traceplot
mcmc_trace(fit, pars = "lambda[5]")
```



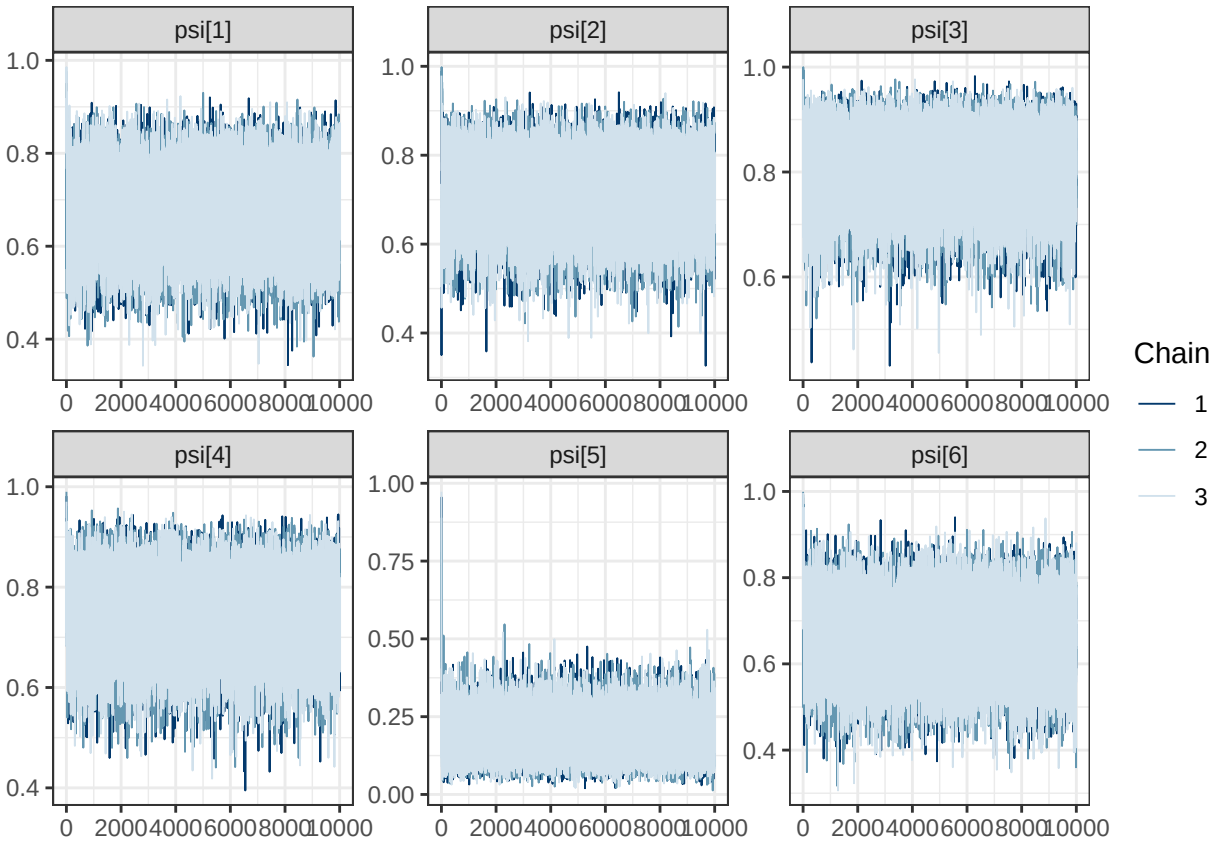
We can see that far fewer than 10,000 iterations are likely required because the chains are mixing well after a few hundred iterations. This is shown by strongly overlapping chains, with no one chain appearing by itself in a region of the parameter space. Chains also appear to be stationary because they do not wander vertically substantially after the first few hundred iterations.

We need to check that chains are stationary and mixing for all parameters. One way to accelerate visual inspection this is through the `regex_pars` argument, which shows relative activity parameters for all species as in the following three code blocks.

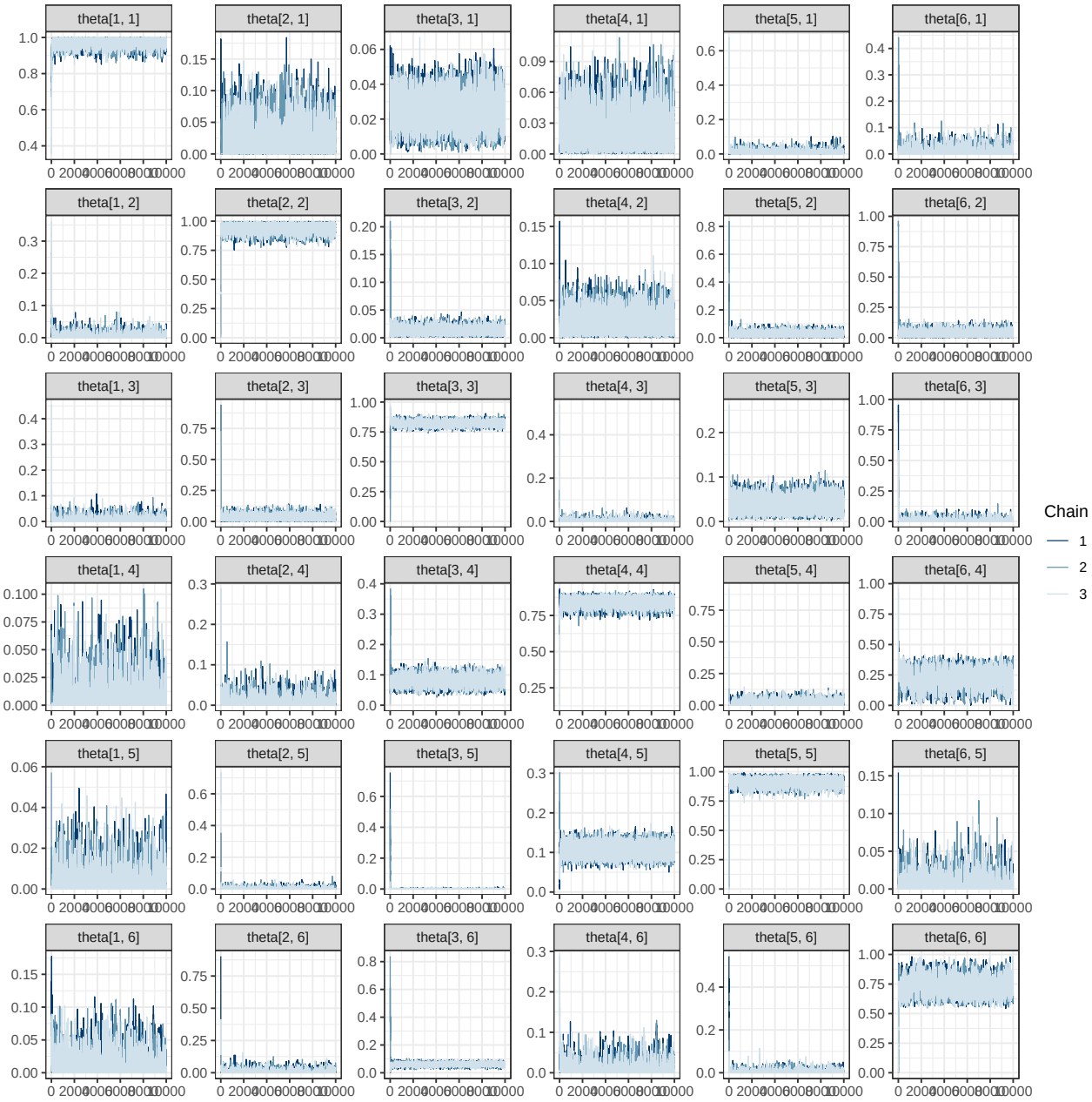
```r
mcmc_trace(fit, regex_pars = "lambda")
```




```
mcmc_trace(fit, regex_pars = "psi")
```




```
# create a traceplot for all elements of the confusion matrix
mcmc_trace(fit, regex_pars = "theta")
```



In all of the trace plots for all of the parameters, chains appear to mix quickly, meaning that it is possible a far smaller number of iterations may be suitable for a simulation study. A good place to start for reducing the number of iterations is from the output given by `tune_mcmc`. In our case, we can see that a guess for the minimum number of iterations is 1500 with a warmup of 500:

```
tune_list$min_iter
```

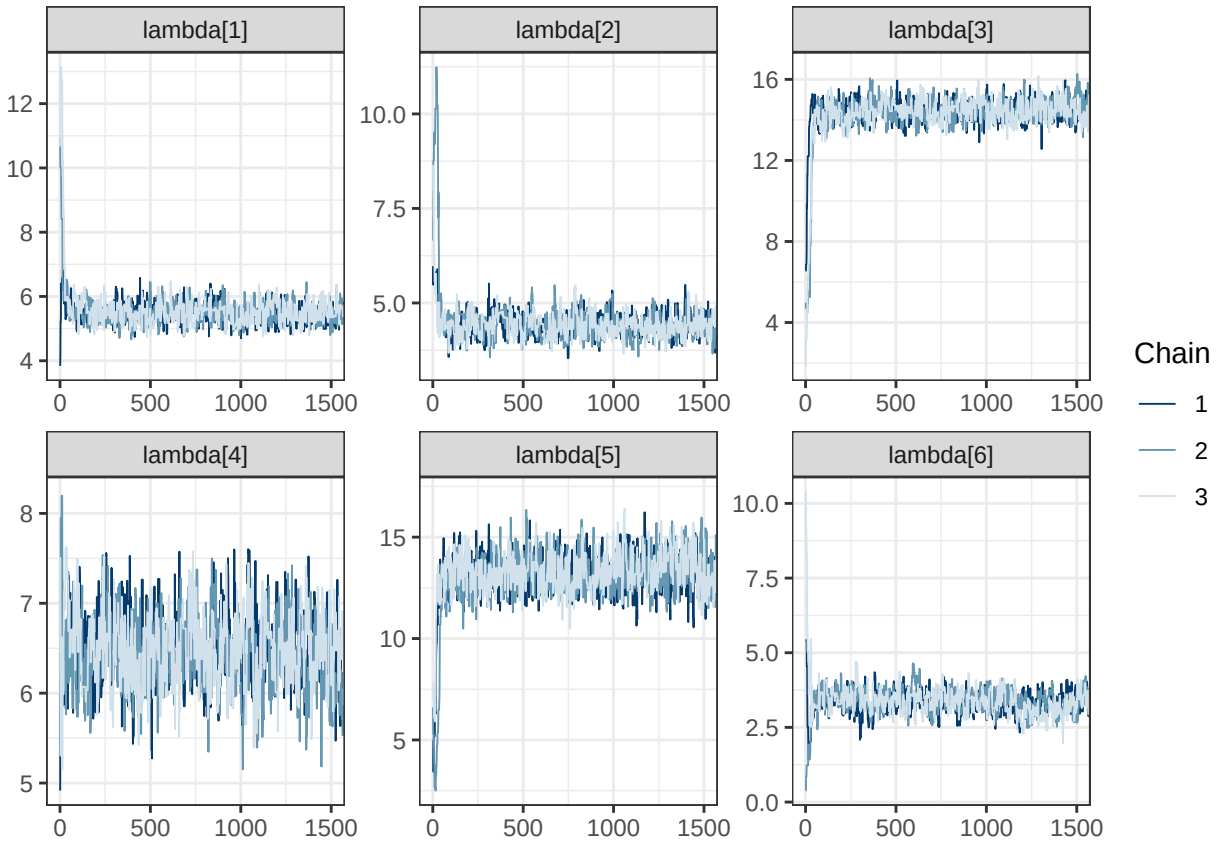
```
## [1] 1500
```

```
tune_list$min_warmup
```

```
## [1] 500
```

It is possible to use these values to zoom in on the trace plots by using the `window` argument:

```
mcmc_trace(fit, regex_pars = "lambda", window = c(0, tune_list$min_iter))
```



Once again, we see the three chains overlap strongly after around 500 iterations and sample around a horizontal line.

After making traceplots and zooming in for all parameters (not all plots are shown here), it appears that the guessed warmup value of 500output from `tune_mcmc` is a reasonable choice. Even so, we increase our iterations and warmup beyond these minimum values in Section 2.6 to avoid simulations failing due to convergence.

2.5.3 Examine effective sample size and Gelman-Rubin statistics

As a final step, we examine the effective sample sizes (out of 10,000 draws) and $\hat{R}$ values. The effective sample size statistics `ess_bulk` and `ess_tail` are MCMC diagnostic statistics that summarizes the number

of effectively independent draws from a parameter's posterior distribution the Markov chain contains. If a parameter has a large value in the `ess_bulk` column, then it is likely that inference based on sampled draws will characterize the center of the posterior distribution well. The `ess_tail` column, on the other hand, describes how much information is available about posterior tail probabilities. Once again, we want a large value for `ess_tail`, preferably `ess_tail`≥ 250.

```
tune_list$MCMC_diagnostics
```

```
##      parameter   ess_bulk   ess_tail     Rhat
## 1    lambda[1]  3863.6858  4230.4082 1.001030
## 2    lambda[2]  2849.7799  3298.1928 1.000767
## 3    lambda[3]  2211.1843  3372.2443 1.001422
## 4    lambda[4]  2892.4484  5067.8405 1.001677
## 5    lambda[5]  3049.8698  4787.3791 1.001435
## 6    lambda[6]  1567.0835  2513.8131 1.002113
## 7       psi[1] 29450.3326 24993.3025 1.000116
## 8       psi[2] 24340.6732 23041.6325 1.000023
## 9       psi[3] 29439.4885 27421.0153 1.000016
## 10      psi[4] 29244.8506 29870.8124 1.000172
## 11      psi[5] 25110.5107 23615.6458 1.000050
## 12      psi[6] 21438.4017 22227.4922 1.000006
## 13 theta[1, 1]  3604.0215  3442.8212 1.000615
## 14 theta[2, 1]  1161.3831   822.5634 1.002923
## 15 theta[3, 1]  2374.2719  3317.5425 1.002389
## 16 theta[4, 1]  3715.6326  4859.6665 1.000049
## 17 theta[5, 1]  1677.1046  2250.1213 1.001303
## 18 theta[6, 1]  2510.8993  3251.2743 1.002067
## 19 theta[1, 2]  3461.8377  3838.1656 1.000710
## 20 theta[2, 2]  2285.0520  4714.1941 1.001054
## 21 theta[3, 2]  4214.2886  5481.2434 1.000651
## 22 theta[4, 2]  3449.9943  3544.1028 1.000611
## 23 theta[5, 2]  4543.7921  3390.8561 1.000442
## 24 theta[6, 2]  3885.0650  3785.0452 1.000822
## 25 theta[1, 3]  2970.3840  3568.8895 1.000845
## 26 theta[2, 3]  4125.6538  5217.9836 1.000153
## 27 theta[3, 3]  2090.8262  5068.3749 1.001960
## 28 theta[4, 3]  2868.9266  2917.4013 1.001901
## 29 theta[5, 3]  6512.9124  6673.9483 1.000218
## 30 theta[6, 3]  2830.6770  3688.3438 1.000131
## 31 theta[1, 4]  3451.3564  3390.2174 1.000317
## 32 theta[2, 4]  2659.1885  2975.5334 1.001160
## 33 theta[3, 4]  2128.6855  4197.4873 1.001855
## 34 theta[4, 4]  3148.2456  4375.7213 1.000737
## 35 theta[5, 4]   936.8766  1656.1938 1.002355
## 36 theta[6, 4]  1264.6785  1776.8593 1.003249
## 37 theta[1, 5]  2022.6652   849.3750 1.002004
## 38 theta[2, 5]  3114.9491  3823.1458 1.000670
## 39 theta[3, 5]  2838.3373  3140.6124 1.001459
## 40 theta[4, 5]  6351.9712  9064.4706 1.000725
## 41 theta[5, 5]  2364.1657  4709.0542 1.001285
```

```
## 42 theta[6, 5]  3353.8191  3514.8509 1.000334
## 43 theta[1, 6]  2269.8186  2864.9995 1.000523
## 44 theta[2, 6]  2433.9124  3054.2535 1.001501
## 45 theta[3, 6]  2547.7273  2836.8061 1.002608
## 46 theta[4, 6]  1508.2564  1509.2496 1.001721
## 47 theta[5, 6]  2843.7472  3117.7029 1.000958
## 48 theta[6, 6]  1325.1647  2190.3602 1.001907
```

For all parameters, the bulk and tail effective sample sizes are fairly large, meaning that even with fewer than 10,000 draws, we could expect $n_{\text{eff}} \geq 250$, allowing us to characterize both the center and tails of the posteriors for all parameters with these draws. Furthermore, the $\hat{R}$ values for all parameters is near 1. In general, we want values of $\hat{R} \leq 1.1$ for chains to be considered converged. We can double check by recomputing these statistics on shortened chains:

```
# for each chain, extract iterations 1001:2500 for all parameters
shortened <- lapply(fit, function(x) x[1001:2500,])

# summarize the shortened chains and select the effective sample
# size columns. mcmc_sum is an internal function used inside of `run_sims`, but
# we use it here to quickly obtain MCMC diagnostics for each parameter
mcmc_sum(shortened, truth = rep(0, ncol(shortened[[1]]))) %>%
  select(parameter, ess_bulk, ess_tail, Rhat)
```

```
##      parameter  ess_bulk  ess_tail      Rhat
## 1    lambda[1]  607.8294  681.9628 1.0108505
## 2    lambda[2]  486.0025  601.2513 1.0075169
## 3    lambda[3]  375.4320  644.6767 1.0014910
## 4    lambda[4]  401.6998  663.0600 1.0048934
## 5    lambda[5]  291.0804  768.6158 1.0137366
## 6    lambda[6]  229.5343  515.1851 1.0196415
## 7       psi[1] 4532.4567 4351.4287 1.0002644
## 8       psi[2] 4480.9409 4475.5581 0.9997728
## 9       psi[3] 4273.0627 4276.6589 0.9999403
## 10      psi[4] 4805.4383 4427.7582 0.9999147
## 11      psi[5] 4319.7173 3995.7945 1.0005130
## 12      psi[6] 3902.3247 4319.3104 0.9998825
## 13 theta[1, 1]  518.2264  451.9967 1.0033104
## 14 theta[2, 1]  196.5426  202.3911 1.0148090
## 15 theta[3, 1]  352.8095  560.1102 1.0118391
## 16 theta[4, 1]  593.1919  741.8146 1.0054704
## 17 theta[5, 1]  314.5272  405.9030 1.0037749
## 18 theta[6, 1]  499.9994  655.7612 1.0077598
## 19 theta[1, 2]  598.1756  511.1293 1.0023073
## 20 theta[2, 2]  358.4724  863.2333 1.0098629
## 21 theta[3, 2]  732.3278  922.6721 1.0014873
## 22 theta[4, 2]  585.7739  830.0864 1.0078040
## 23 theta[5, 2]  849.5523  841.0106 1.0022997
## 24 theta[6, 2]  571.4574  565.7589 1.0026868
## 25 theta[1, 3]  549.0994  640.8539 1.0054837
```

```
## 26 theta[2, 3]  633.5005  854.4388 1.0109568
## 27 theta[3, 3]  296.5956 1010.3848 1.0032231
## 28 theta[4, 3]  575.3593  762.0131 1.0017522
## 29 theta[5, 3] 1059.6573 1347.5769 1.0016651
## 30 theta[6, 3]  520.7469  732.1743 1.0135429
## 31 theta[1, 4]  508.0144  577.4731 1.0074157
## 32 theta[2, 4]  535.9739  715.8251 1.0025120
## 33 theta[3, 4]  301.4109  933.5054 1.0062035
## 34 theta[4, 4]  601.4933  901.3706 1.0021416
## 35 theta[5, 4]  112.0073  283.0519 1.0368826
## 36 theta[6, 4]  125.6020  325.2078 1.0262094
## 37 theta[1, 5]  664.1465  588.5673 1.0030516
## 38 theta[2, 5]  482.8635  440.3829 1.0064307
## 39 theta[3, 5]  447.3579  408.2391 1.0058526
## 40 theta[4, 5] 1048.0148 1551.5007 1.0047974
## 41 theta[5, 5]  196.7384  996.7905 1.0171951
## 42 theta[6, 5]  511.7644  573.4703 1.0056833
## 43 theta[1, 6]  231.0497  343.0134 1.0087047
## 44 theta[2, 6]  354.1645  504.5090 1.0032238
## 45 theta[3, 6]  359.1055  606.1931 1.0084861
## 46 theta[4, 6]  260.5061  288.1928 1.0047416
## 47 theta[5, 6]  297.9489  232.4969 1.0159171
## 48 theta[6, 6]  159.6269  404.2219 1.0188543
```

These results appear satisfactory, with effective sample sizes in both the tail and bulk of the posterior distributions of more than 250 and $\hat{R}$ near 1. Based on the results of MCMC tuning, it appears that using an MCMC with at least 1500 iterations with at least 500 discarded as warmup should to produce good results for our simulation study.

2.5.4 Set iterations for simulation

Based on our findings in the MCMC tuning step, we set the number of iterations for simulation to be slightly higher to guard against convergence issues that preclude using a fitted model for inference:

```
# to be used in the following section
iters_for_sims <- tune_list$min_iter + 1000
warmup_for_sims <- tune_list$min_warmup + 500
```

2.5.5 A note about non-convergence while tuning

In some instances, we have run `tune_mcmc` with a dataset and received a series of error messages that convergence was not reached in under 10,000 iterations. If this persists after trying to fit several other datasets, we have several options:

1. We could increase the number of iterations in the simulation study to be above 10,000 – perhaps to 20,000 and settle for a longer run time of the simulation study.

2. We could take this as a sign that the level of effort is insufficient to identify model parameters. In this case, this scenario should not be considered.

In our experience fitting these models, the second option seems to often be the case, and we encourage users to remove scenarios from consideration if models fit during the tuning step do not reach convergence within 10,000 iterations.

We emphasize that the results from `tune_mcmc` are from a single model fit; they are supplied only as guidelines, and we encourage users to increase the number of iterations above the minimum values output from `tune_mcmc`. While each model fit will take slightly longer with an increased number of total iterations, this approach may save time in the long run by avoiding the need to re-run simulations.

2.6 Step 4: Fit models to simulated data

With the simulated dataset and some informed choices about tuning of the MCMC, we use `run_sims` to run the simulations:

```
sims_output <- run_sims(
         data_list = sim_data$masked_dfs,
         zeros_list = sim_data$zeros,
         DGVs = list(lambda = lambda, psi = psi, theta = Theta),
         theta_scenario_id = "FE", # for "fixed effort"
         parallel = TRUE,
         niter = iters_for_sims,
         nburn = warmup_for_sims,
         thin = 1,
         save_fits = TRUE,
         save_individual_summaries_list = FALSE,
         directory = here::here("Vignette", "Fixed_Effort")
)
```

```
## Beginning scenario 1.
```

```
## 2024-12-28 13:34:33.143357
```

```
##   |                                                                      |
```

```
## Beginning scenario 2.
```

```
## 2024-12-28 13:49:48.485313
```

```
##   |                                                                      |
## Beginning scenario 3.
## 2024-12-28 13:58:11.456341
##   |                                                                      |
## Beginning scenario 4.
## 2024-12-28 14:07:01.192863
##   |                                                                      |
```

The output object, `sims_output` is a dataframe with summaries of the MCMC draws for each parameter estimated from each dataset under each validation scenario. Summaries include the posterior mean, standard deviation, Naive standard error, Time-series standard error, quantiles for 50% and 95% posterior intervals, median, and MCMC diagnostics (Gelman-Rubin statistic, and effective samples sizes in the tails and bulk of the distribution).

```
str(sims_output)
```

```
## 'data.frame':    1920 obs. of  19 variables:
##  $ parameter     : chr  "lambda[1]" "lambda[2]" "lambda[3]" "lambda[4]" ...
##  $ Mean          : num  6 4.8 14.05 5.99 11.79 ...
##  $ SD            : num  0.348 0.343 0.449 0.492 0.826 ...
##  $ Naive SE      : num  0.00518 0.00512 0.00669 0.00733 0.01232 ...
##  $ Time-series SE: num  0.0129 0.015 0.019 0.0295 0.032 ...
##  $ 2.5%          : num  5.32 4.19 13.18 5.02 10.27 ...
##  $ 25%           : num  5.77 4.56 13.76 5.66 11.25 ...
##  $ 50%           : num  6 4.79 14.04 5.99 11.7 ...
##  $ 75%           : num  6.23 5.02 14.36 6.3 12.29 ...
##  $ 97.5%         : num  6.71 5.52 14.92 7.01 13.57 ...
##  $ Rhat          : num  1.01 1.01 1 1.02 1.01 ...
##  $ ess_bulk      : num  758 379 561 264 589 ...
##  $ ess_tail      : num  1022 642 1042 562 637 ...
##  $ truth         : num  5.93 4.16 14.25 6.2 11.86 ...
##  $ capture       : num  1 0 1 1 1 1 1 1 1 1 ...
##  $ converge      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ theta_scenario: chr  "FE" "FE" "FE" "FE" ...
##  $ scenario      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ dataset       : int  1 1 1 1 1 1 1 1 1 1 ...
```

Note that we fit all models in parallel to reduce simulation time; we encourage users to do the same. However, note that if you fit models with `parallel = FALSE` in `run_sims`, the console will display the messages NIMBLE prints as it compiles code for each model. Internally, we specify a custom distribution to compute the marginal probabilities for ambiguous autoID labels, and you will see a warning about overwriting a custom user-specified distribution if `parallel = FALSE`. These warnings can be safely ignored.

2.7 Step 5: Visualize simulations

Once models have been fit to all simulated datasets, you can visualize results using several functions. We recommend beginning with the most detailed functions, which are `visualize_parameter_group` and `visualize_single_parameter`. The plots output from these functions show many following features of the simulation study. These are

- Facet grids: parameters (only for `visualize_parameter_group`)
- X-axis: Manual verification scenario
- y-axis: parameter values
- Small grey error bars: 95% posterior interval for an individual model fit where all parameters were below `convergence_threshold`.
- Colored error bars: average 95% posterior interval across all converged models under that scenario.
- Color: Coverage, or the rate at which 95% posterior intervals contain the true data-generating parameter value.
- Black points: the true value of the parameter
- Red points: average posterior mean

The `visualize_parameter_group` function is useful for examining an entire set of parameters, such as all relative activity parameters. For example, we can visualize the inference for the relative activity parameters in the first three scenarios in our simulation study above by running the code below.

```
visualize_parameter_group(
  sim_summary = sims_output,
  pars = "lambda",
  theta_scenario = 1,
  scenarios = 1:4,
  convergence_threshold = 1.05
)
```

The output shown in Figure 1 indicates that under all validation scenarios we considered, the expected inference for relative activity rates of species 1-4 and 6 meets our measurable objectives: the posterior interval width is less than three for these species and there is minimal estimation error. However, note that under validation scenario 1 and 2, fewer of the models converged within 2500 iterations of the MCMC, which is visible from smaller number of grey intervals. This suggests that a higher level of validation effort could be beneficial. Additionally, average interval width is never below 3 for species 5; none of the designs

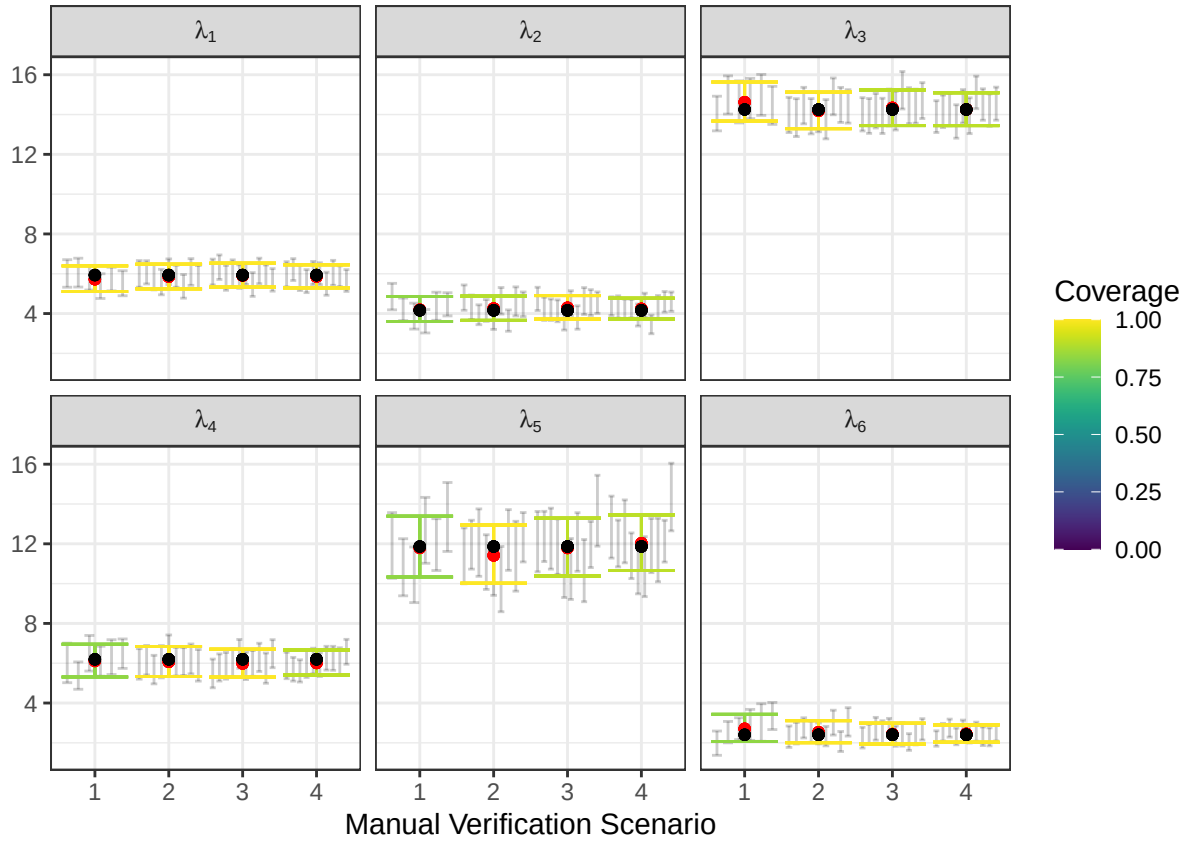


Figure 1: Example output from the `visualize_parameter_group` function. Results are faceted by parameter, with the x-axis indicating the validation scenario number, and the y-axis showing the data generating value (black point). Red points show average posterior means, small grey intervals show 95% posterior intervals from converged model fits and colored intervals are average 95% posterior intervals. Coverage is shown by color, with blue corresponding to low coverage.

considered here meet the measurable objective for this species. We can see this more clearly by examining output through alternative visualization functions.

A first step would be to use `visualize_single_parameter`, which takes the same arguments as the previous visualization function:

```
visualize_single_parameter(
  sims_output, par = "lambda[5]",
  theta_scenario = 1,
  scenarios = 1:4,
  convergence_threshold = 1.05
)
```

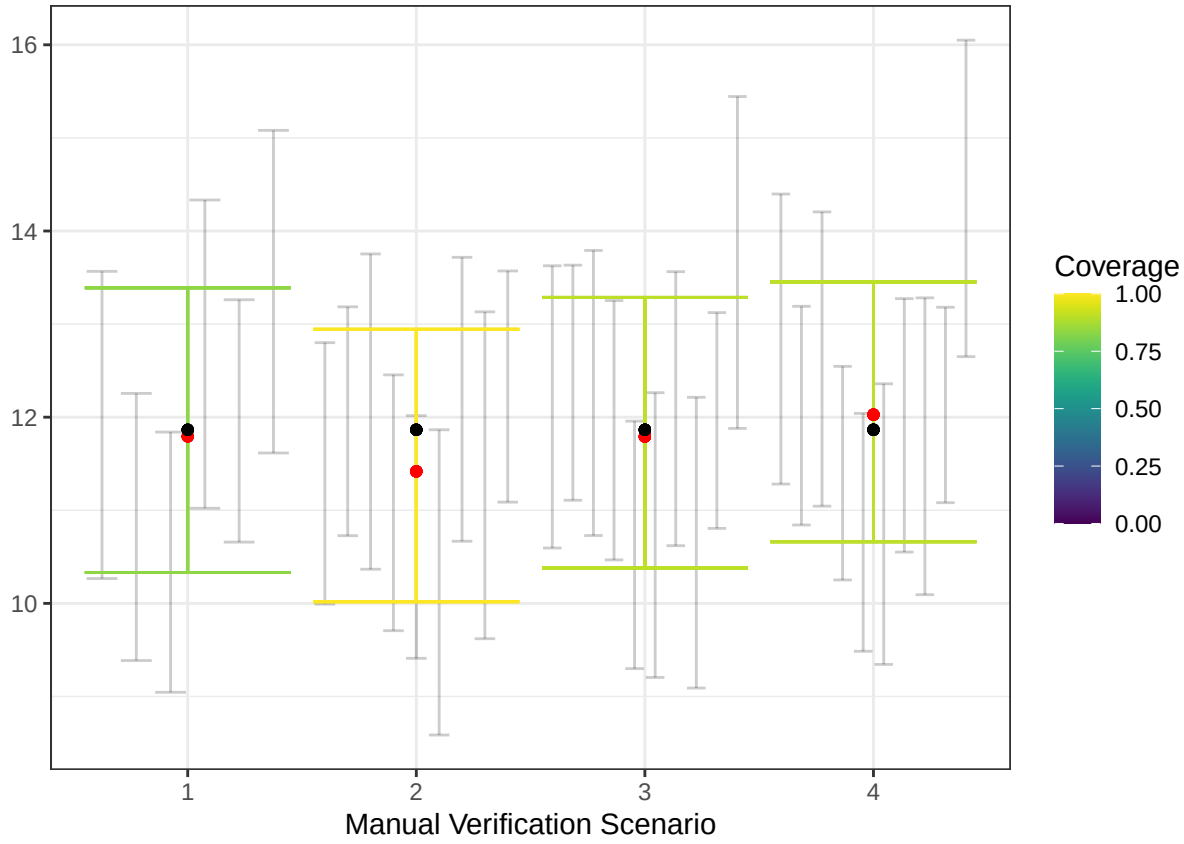


Figure 2: Example output from `visualize_single_parameter` is similar to that of `visualize_parameter_group`: the x-axis shows the validation scenario, and the y-axis shows the true value of the parameter (black point). Average posterior means are shown by red points, individual 95% posterior intervals from converged model fits are shown as small grey error bars and average 95% posterior intervals are shown by the larger colored error bars. The coverage is indicated by color.

The greater detail of `visualize_single_parameter` in Figure 2 underscores the difference in the number of converged models, shown by the number grey 95% posterior interval for each converged model fit. The visualization also helps clarify that while average interval width is near 3 in scenario 2, it is still acceptable

for our measurable objectives. Coverage for most relative activity parameters is also near-nominal under this validation scenario, which requires that around 375 recordings be validated in an average dataset. This can also be seen using the `plot_width_vs_calls` function (Figure 3), which compares average interval widths (and the IQR for interval widths) across scenarios based on the number of recordings validated. An example is provided below:

```
# obtain summary of validation effort as an object that can be used with
# the summary plot functions plot_x_vs_calls
s <-  summarize_n_validated(
  data_list = sim_data$masked_dfs,
  theta_scenario = "FE", # must not be NULL for the plot_X_vs_calls functions
  scenario_numbers = 1:4
)
```

```
plot_width_vs_calls(
  sim_summary = sims_output,
  calls_summary = s,
  regex_pars = "lambda",
  theta_scenario = "FE",
  scenarios = 2:5
) + # add horizontal line at target 95% CI width
  geom_hline(yintercept = 3, linetype = "dotted")
```

We provide analogous functions `plot_coverage_vs_calls` and `plot_bias_vs_calls` taking identical arguments that show how coverage and estimation error change with the number of calls validated. Note that in all of our visualization functions, if no fitted models have $\hat{R} \leq c$ for all parameters, where c is the specified convergence threshold under a given scenario, then the scenario will not appear on the x-axis. In our example, we used $c = 1.1$.

To ensure that there aren't substantial problems with estimation of other model parameters, we can create plots for ψ and Θ:

```
visualize_parameter_group(
  sim_summary = sims_output,
  pars = "psi",
  theta_scenario = 1,
  scenarios = 1:4,
  convergence_threshold = 1.05
)
```

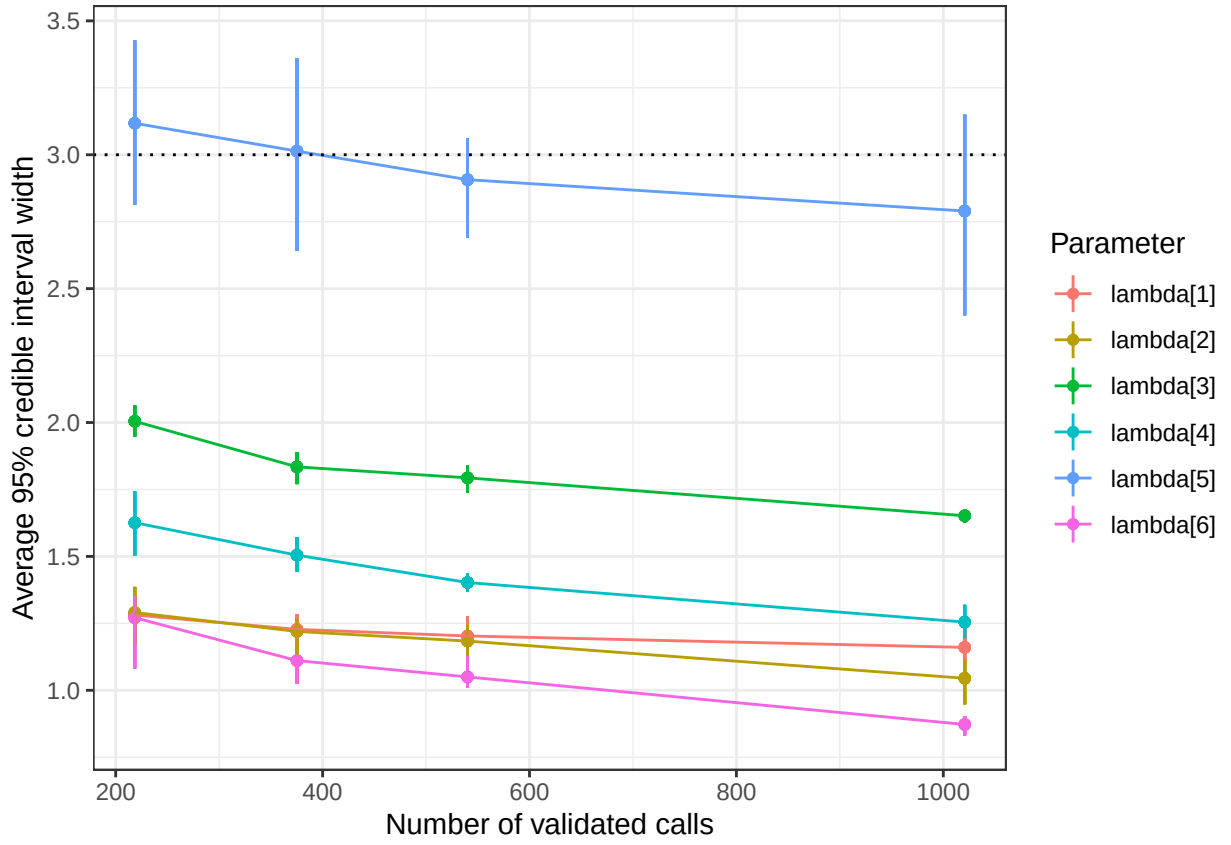


Figure 3: Example output from `plot_width_vs_calls`. The x-axis shows the number of validated recordings. Note that the level of effort increases with scenario number for the fixed-effort designs; this may not always be the case, for example, with a stratified-by-species design as in the main text. The y-axis shows the value of the average 95% posterior interval width (point). The middle 50% of interval widths for each parameter are shown by the error bars. Color indicates the parameter.

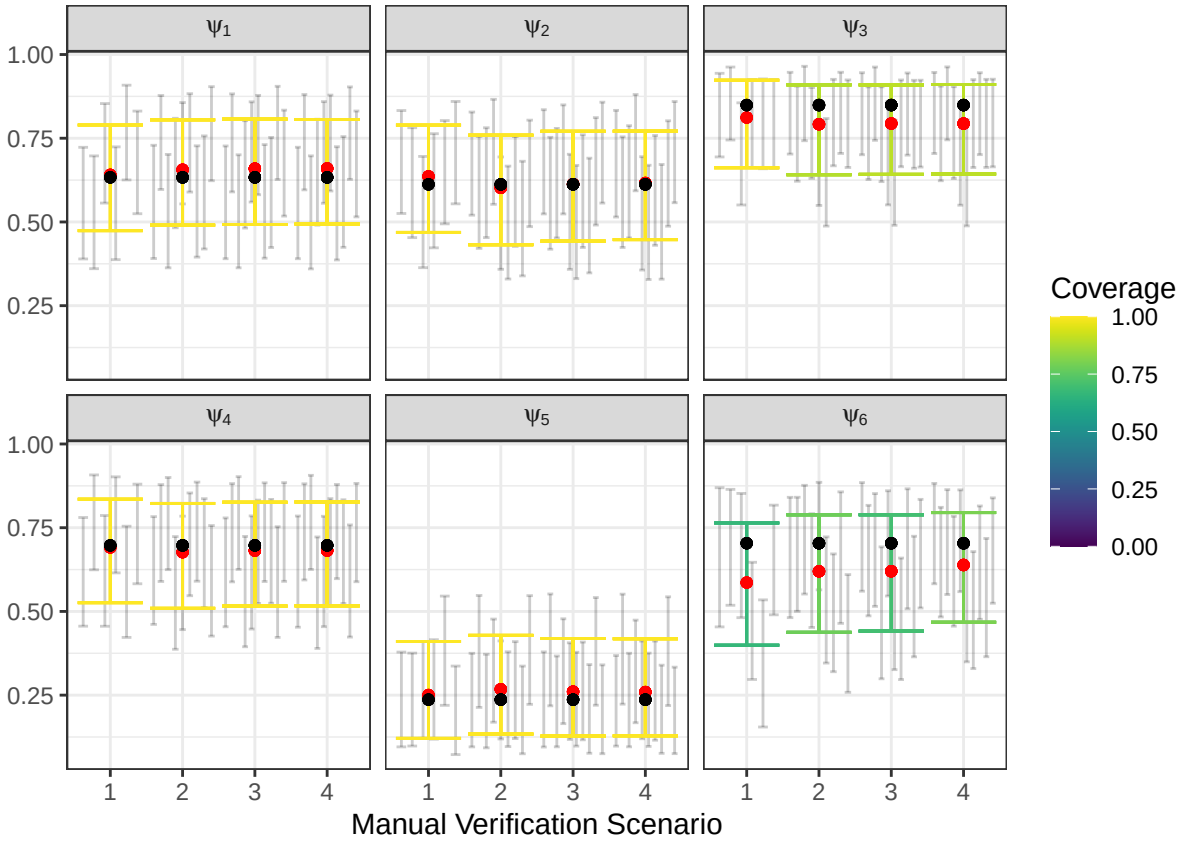


Figure 4: We encourage users to check inference for other model parameters, even if they are not explicitly of interest based on measurable objectives. Here, we examine $\boldsymbol{\psi}$. Note that coverage for several parameters is slightly low. For the purpose of exposition, we only fit 10 datasets and we expect that coverage would be near nominal levels if 50 or more datasets were used in simulation.

```r
visualize_parameter_group(
  sim_summary = sims_output,
  pars = "theta",
  theta_scenario = 1,
  scenarios = 1:4,
  convergence_threshold = 1.05
)
```

The simulation results shown in Figure 4 and 5 don't immediately cause concern that inference for relative activity rates will be severely compromised for any of these validation designs. Some parameters have slightly low coverage, but we believe this is mostly due to the small number of datasets used in this example; users should increase the number of datasets to at least 50, which we expect would lead to coverage near nominal levels.

2.8 Take-aways from the Fixed-Effort Simulations

Based on Figures 1 - 3, we can rule out validation scenario 1, because the expected width of 95% posterior intervals is greater than the threshold value of 3. Because of this, we might choose validation scenario 2 as one possible fixed effort validation design. This fixed effort design assumes that 10% of recordings from each site-night are randomly selected for validation, leading to around 375 recordings being validated per dataset. In our simulations, posterior means for each λ_k appear to be nearly unbiased, with near-nominal coverage and average posterior interval widths less than or equal to 3 for this scenario. However, it is possible that for a given dataset, the interval width for λ_5 will be larger than 3, regardless of the number of validated calls; error bars in Figure 3 overlap or exceed the dotted line in all scenarios.

In situations such as this one, a decision is required of the user. If a slightly larger interval width is acceptable, validating 10% percent of recordings from all site-visits is reasonable. If it is not, then the LOVE will need to be increased beyond that in scenario 4 (1021 recordings). The question is whether potentially tripling the number of validated recordings will be worth the benefit of ensuring the relative activity level for species 5 is less than 3. This is a question about measurable objectives that we leave to the practitioner and their monitoring priorities.

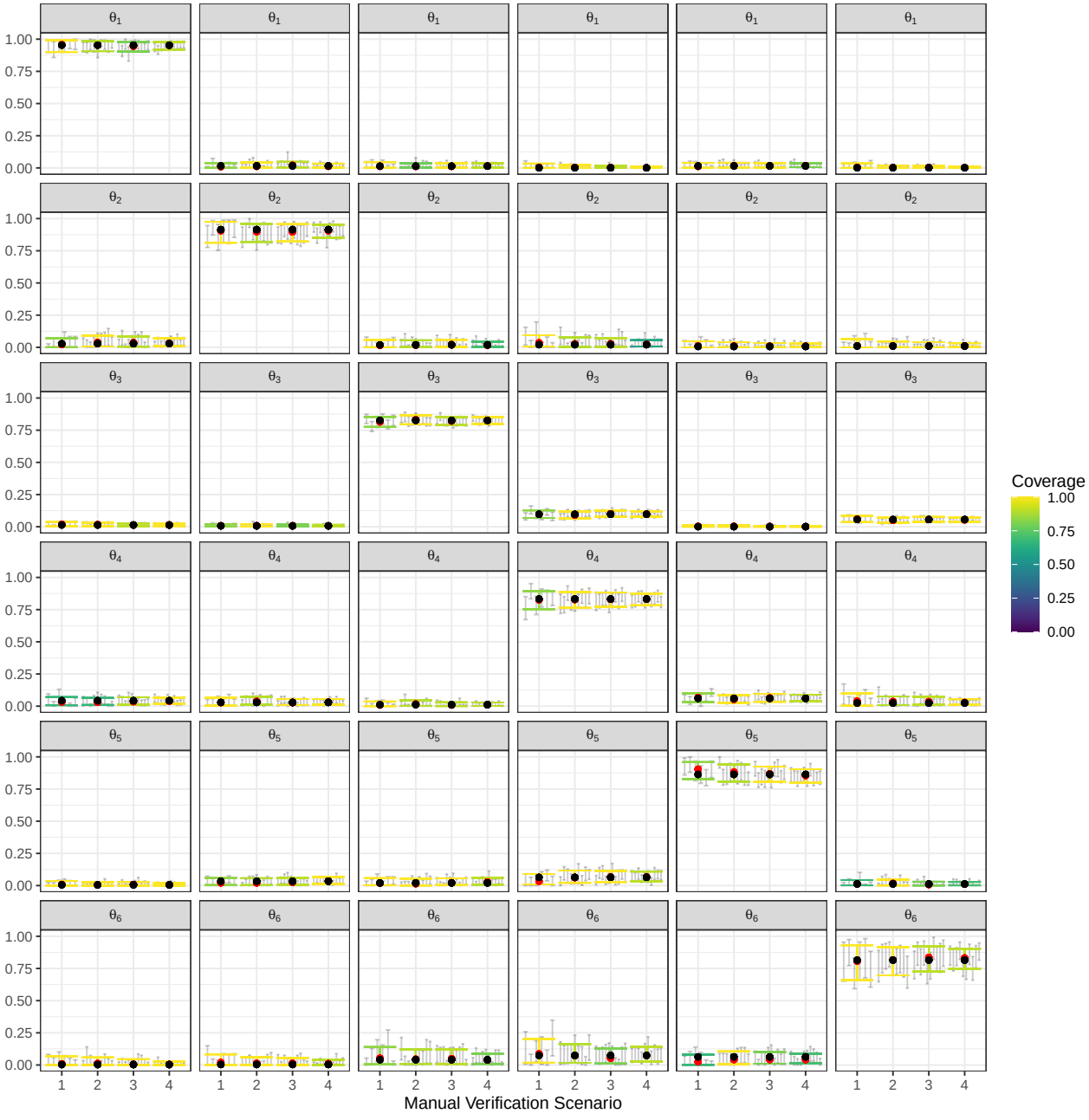


Figure 5: Checking inference for the elements of the classification matrix Θ. For most parameters (facets), validation scenario 2 appears to lead to minimal estimation error and near-nominal coverage.

3 Stratified-by-species example

[Katie, should I go as in depth as in the previous example? It feels a bit strange to show all of the output twice, since they are essentially the same, but maybe we want the sections to stand alone..? Sometimes that can be nice, but I could also imagine it being overly cumbersome.]

In this section, we provide further details of the example in the main manuscript, which assumed a stratified-by-species design. In a stratified-by-species design, the random mechanism is a stratified random sample, with strata formed by the species labels assigned by automated classification software (i.e., by autoID). In this kind of design, the LOVE is the proportion of recordings selected to be validated for each species label.

We repeat the simulation study conducted in the main text in this section, providing greater detail for some of the steps. We assume the same measurable objectives and cost constraints as in Section 2.

3.1 Simulate data

Using the same values for `psi`, `lambda` and `Theta` as in Section 2.4, we simulate data using `simulate_validatedData` with `design_type = "BySpecies"`.

```
sim_data <- simulate_validatedData(
  n_datasets = 10,
  design_type = "BySpecies",
  scenarios = list(
    spp1 = 0.15,
    spp2 = 0.15,
    spp3 = c(0.15, .5, 1),
    spp4 = 0.1,
    spp5 = c(0.5, 1),
    spp6 = 0.25
  ),
  nsites = nsites,
  nspecies = nspecies,
  nvisits = nvisits,
  psi = psi,
  lambda = lambda,
  theta = Theta,
  directory = here::here("Vignette", "BySpecies")
)
```

Note that in contrast with Section 2.4, `simulate_validatedData` expects the `scenarios` argument to be a list of proportions corresponding to the possible levels of effort for each species when specifying a stratified-by-species design. Internally, `simulate_validatedData` calls `base::expand.grid`, and considers all possible combinations of the various levels of effort for each species, meaning that the number of simulated scenarios

grows extremely quickly. The biggest implication of having a larger number of simulation scenarios to consider is increased computation time. For example, in the previous code block we fixed the level of effort for species 1, 2, 4 and 6 at .15, .15, .1, and .25, respectively. There are three possible proportions to validate for species 3 and two possible levels of effort for species 5, yielding six possible scenarios, which are summarized in the additional output `sim_data$scenarios_df` that is available when `design_type = "BySpecies"` is specified:

```
names(sim_data)
```

```
## [1] "full_datasets" "zeros"         "masked_dfs"    "scenarios_df"
```

```
sim_data$scenarios_df
```

```
##   scenario spp1 spp2 spp3 spp4 spp5 spp6
## 1        1 0.15 0.15 0.15  0.1  0.5 0.25
## 2        2 0.15 0.15 0.50  0.1  0.5 0.25
## 3        3 0.15 0.15 1.00  0.1  0.5 0.25
## 4        4 0.15 0.15 0.15  0.1  1.0 0.25
## 5        5 0.15 0.15 0.50  0.1  1.0 0.25
## 6        6 0.15 0.15 1.00  0.1  1.0 0.25
```

We can combine the `scenarios_df` output with the output of `summarize_n_validated` to understand what the scenarios are and how many recordings are validated under each.

```
summary1 <- sim_data$scenarios_df
summary2 <- summarize_n_validated(
  sim_data$masked_dfs,
  scenario_numbers = 1:6,
  theta_scenario = "BySpecies"
)

call_sum <- left_join(summary1, summary2, by = "scenario")

call_sum
```

```
##   scenario spp1 spp2 spp3 spp4 spp5 spp6 theta_scenario n_validated
## 1        1 0.15 0.15 0.15  0.1  0.5 0.25      BySpecies       603.8
## 2        2 0.15 0.15 0.50  0.1  0.5 0.25      BySpecies      1018.3
## 3        3 0.15 0.15 1.00  0.1  0.5 0.25      BySpecies      1610.4
## 4        4 0.15 0.15 0.15  0.1  1.0 0.25      BySpecies       778.3
## 5        5 0.15 0.15 0.50  0.1  1.0 0.25      BySpecies      1192.8
## 6        6 0.15 0.15 1.00  0.1  1.0 0.25      BySpecies      1784.9
```

In the resulting dataframe, we see that scenario 1 has the lowest overall level of validation effort: species 1-3 have 15% of their recordings validated, species 4 has 10% validated, species 5 has 50%, and species 6 has

25%, yielding around 604 recordings validated in an average dataset. We use one dataset simulated under scenario 1 to tune the MCMC.

3.2 Tune the MCMC

The MCMC tuning in this step is very similar to the procedures used in Section 2.5. We begin by fitting a model to one dataset using the `tune_mcmc` function.

```
i <- sample(1:length(sim_data$full_datasets), 1)

tune_list <- tune_mcmc(
  dataset = sim_data$masked_dfs[[1]][[i]],
  zeros = sim_data$zeros[[i]]
)
```

```
## [1] "Fitting MCMC in parallel ... this may take a few minutes"
```

As in Section 2.5, we visually inspect the chains for convergence using trace plots. We set trace plot windows to be wider than the suggested values from `tune_mcmc`, meaning a lower starting value and a higher ending value. Based on these plots, chains appear to be mixing well as no one chain stands out visually.

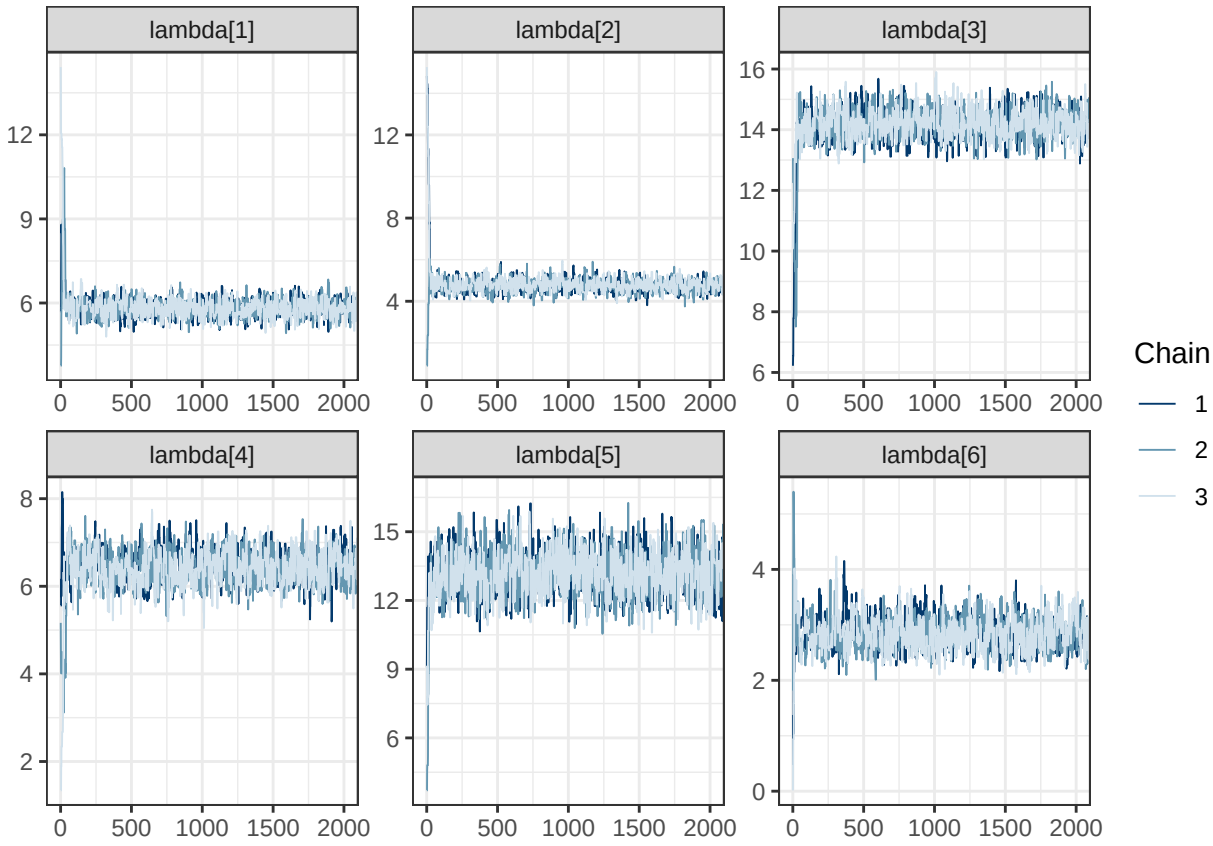
```
tune_list$min_warmup
```

```
## [1] 500
```

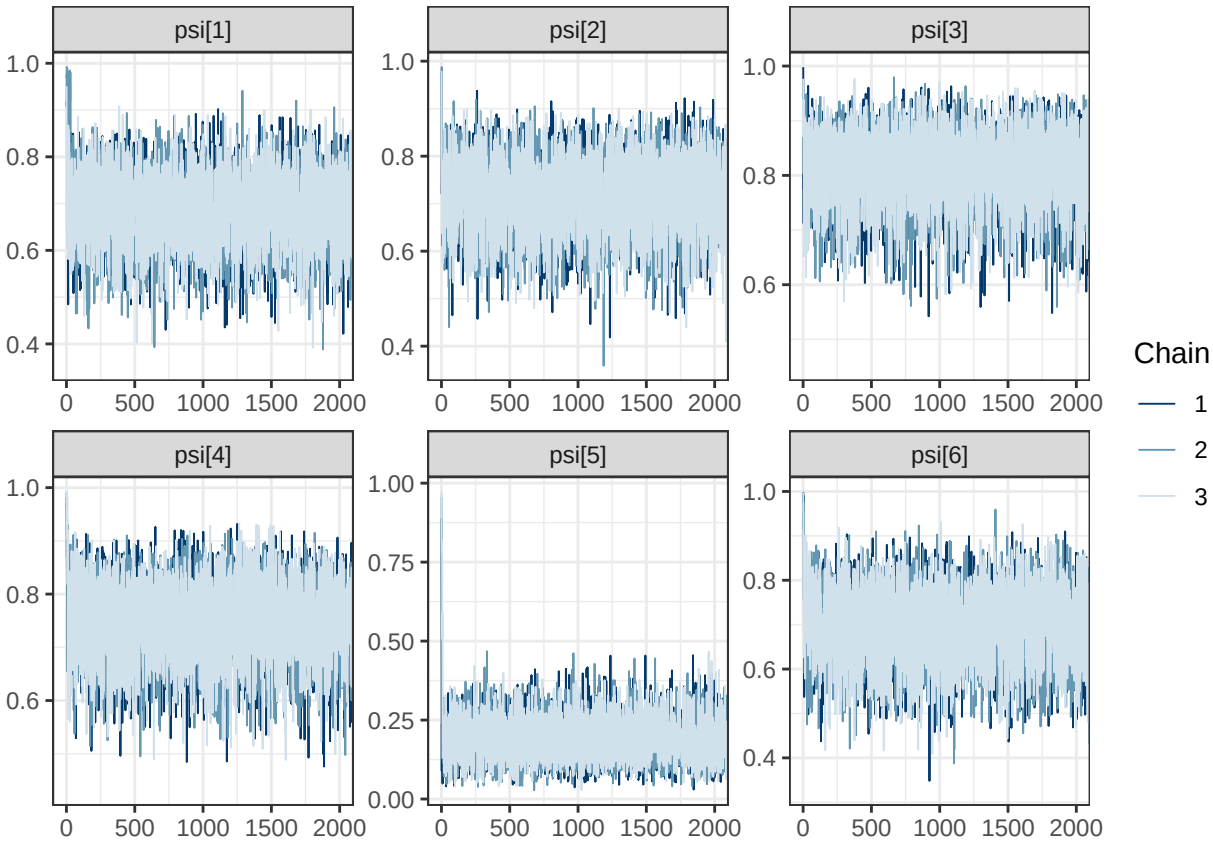
```
tune_list$min_iter
```

```
## [1] 1500
```

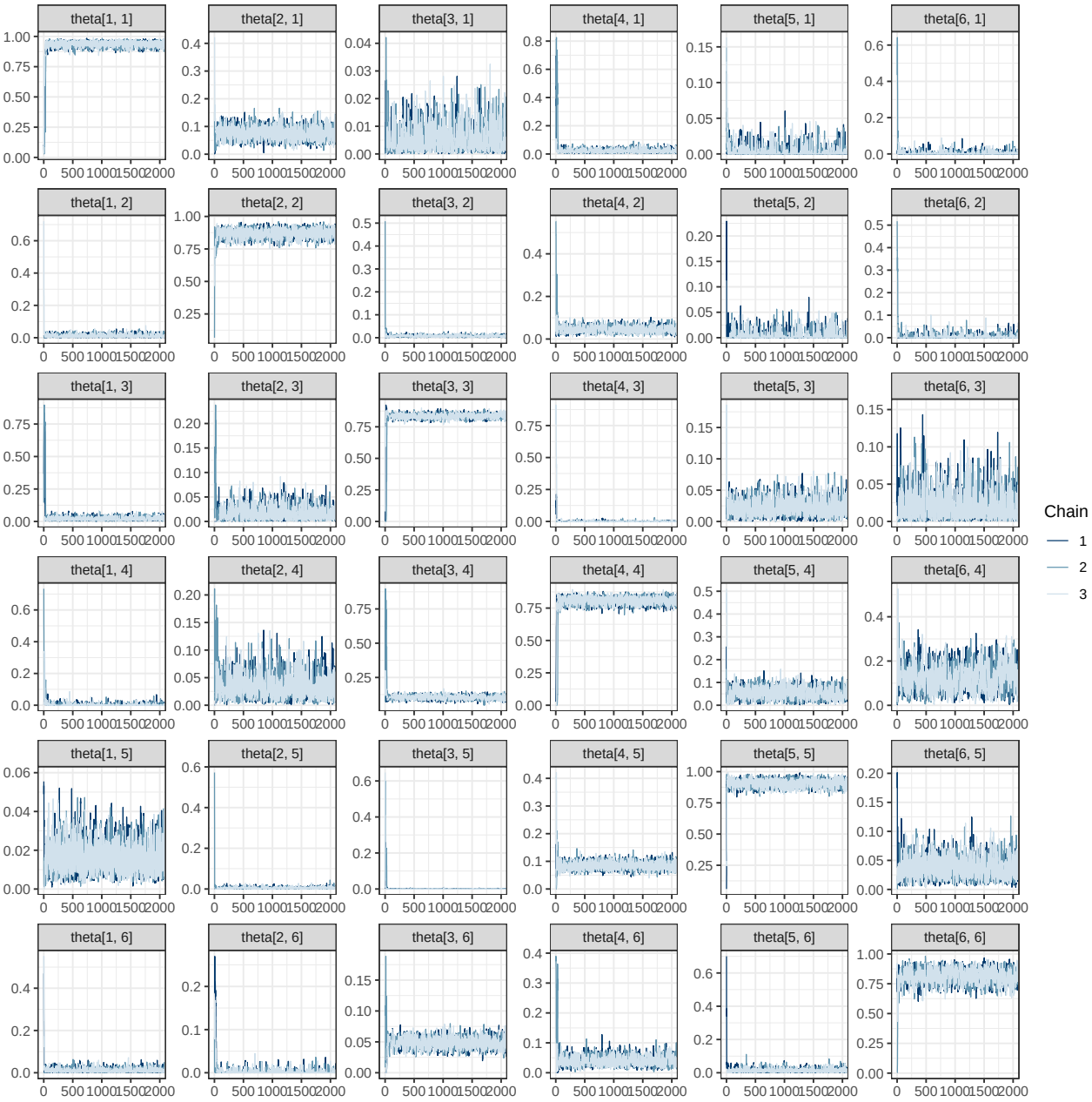
```
# increase iters to visualize beyond the warmup and
# iterations output from tune_mcmc
fit <- tune_list$fit

mcmc_trace(fit, regex_pars = "lambda", window = c(0, tune_list$min_iter + 500))
```




```
mcmc_trace(fit, regex_pars = "psi", window = c(0, tune_list$min_iter + 500))
```




```
mcmc_trace(fit, regex_pars = "theta", window = c(0, tune_list$min_iter + 500))
```



In all trace plots, it appears that MCMC chains have reached approximate convergence well before the warmup value of 500 draws output from `tune_mcmc`.

Next, we examine the effective sample sizes in the tails and bulk of the posteriors and the $\hat{R}$ values:

```
tune_list$MCMC_diagnostics
```

```
##      parameter  ess_bulk  ess_tail      Rhat
## 1    lambda[1]  4657.167  5457.071 1.0008453
## 2    lambda[2]  4258.668  5136.436 1.0018807
## 3    lambda[3]  3759.045  5231.923 1.0003233
```

```
## 4    lambda[4]  3616.319  4707.875 1.0011418
## 5    lambda[5]  4697.559  6401.291 1.0003968
## 6    lambda[6]  2770.656  4541.737 1.0002258
## 7       psi[1] 23316.530 20672.725 1.0000697
## 8       psi[2] 29744.963 29837.224 1.0000093
## 9       psi[3] 29383.485 29300.417 1.0000054
## 10      psi[4] 29042.613 25965.502 1.0001393
## 11      psi[5] 30262.992 29435.301 0.9999815
## 12      psi[6] 24076.063 25595.472 1.0000579
## 13 theta[1, 1]  5770.486  7297.346 1.0004596
## 14 theta[2, 1]  5756.939  7143.769 1.0005660
## 15 theta[3, 1]  4214.412  4282.661 1.0011440
## 16 theta[4, 1]  4382.339  5840.220 1.0013630
## 17 theta[5, 1]  3840.160  3881.033 1.0004450
## 18 theta[6, 1]  3524.445  4417.813 1.0010412
## 19 theta[1, 2]  5104.201  4492.158 1.0002552
## 20 theta[2, 2]  6198.730  8537.925 1.0006967
## 21 theta[3, 2]  5031.551  5226.067 1.0005384
## 22 theta[4, 2]  5168.674  7261.228 1.0002213
## 23 theta[5, 2]  2832.185  3226.998 1.0007657
## 24 theta[6, 2]  3321.046  4018.054 1.0003502
## 25 theta[1, 3]  4781.282  6609.743 1.0006503
## 26 theta[2, 3]  4095.401  3291.010 1.0004389
## 27 theta[3, 3]  4416.084  7939.014 1.0003155
## 28 theta[4, 3]  2837.780  2043.733 1.0006095
## 29 theta[5, 3]  4293.883  4779.353 1.0003034
## 30 theta[6, 3]  5227.736  5444.933 1.0004644
## 31 theta[1, 4]  2572.031  3239.633 1.0010872
## 32 theta[2, 4]  4940.939  5669.964 1.0013073
## 33 theta[3, 4]  3679.146  6486.391 1.0009714
## 34 theta[4, 4]  6121.785  7937.438 1.0012085
## 35 theta[5, 4]  2988.075  3599.707 1.0003774
## 36 theta[6, 4]  2437.743  5166.489 1.0008458
## 37 theta[1, 5]  6288.486  8394.523 1.0001072
## 38 theta[2, 5]  5558.291  5320.515 1.0007597
## 39 theta[3, 5]  3442.863  4193.477 1.0008817
## 40 theta[4, 5]  7013.708  9457.490 1.0000771
## 41 theta[5, 5]  3722.483  6767.135 1.0002251
## 42 theta[6, 5]  5999.325  7994.946 1.0008838
## 43 theta[1, 6]  5110.607  5364.993 1.0004374
## 44 theta[2, 6]  3359.262  4701.304 1.0007382
## 45 theta[3, 6]  4671.334  5827.765 1.0003612
## 46 theta[4, 6]  4077.511  5443.258 1.0016645
## 47 theta[5, 6]  5012.718  5773.624 1.0004617
## 48 theta[6, 6]  2645.472  5898.481 1.0007112
```

For all parameters, the bulk and tail effective sample sizes are fairly large: if we slightly decrease the number of post-warmup draws, we could expect to characterize both the center and tails of the posteriors well. Additionally, $\hat{R} \approx 1.00$ for all parameters, indicating good mixing for chains. As before, we check this suspicion by computing MCMC diagnostic statistics for truncated chains:

```r
# for each chain, extract iterations 501:2000 for all parameters
shortened <- lapply(fit, function(x) x[(tune_list$min_warmup):tune_list$min_iter + 1000,])

# summarize the shortened chains and select the effective sample
# size columns
mcmc_sum(shortened, truth = rep(0, ncol(shortened[[1]]))) %>%
  select(parameter, ess_bulk, ess_tail)
```

```
##      parameter  ess_bulk  ess_tail
## 1    lambda[1]  492.9109  631.3996
## 2    lambda[2]  401.5597  329.0051
## 3    lambda[3]  478.1999  629.0566
## 4    lambda[4]  403.3821  522.2753
## 5    lambda[5]  627.5047  831.3168
## 6    lambda[6]  364.8041  561.3582
## 7       psi[1] 2804.1022 2870.7717
## 8       psi[2] 3068.7076 3092.6423
## 9       psi[3] 2994.9267 2994.9622
## 10      psi[4] 3034.1430 3015.7601
## 11      psi[5] 3048.8035 2947.0121
## 12      psi[6] 2274.6952 2424.4337
## 13 theta[1, 1] 1010.2582 1380.1778
## 14 theta[2, 1]  682.6149  945.6879
## 15 theta[3, 1]  468.7398  411.4688
## 16 theta[4, 1]  646.2412  814.6341
## 17 theta[5, 1]  454.3240  581.9752
## 18 theta[6, 1]  465.7229  485.5060
## 19 theta[1, 2]  495.0805  699.6539
## 20 theta[2, 2]  655.2808  967.7691
## 21 theta[3, 2]  461.2923  432.1370
## 22 theta[4, 2]  482.1296  720.2454
## 23 theta[5, 2]  390.9667  372.6068
## 24 theta[6, 2]  368.2706  362.5135
## 25 theta[1, 3]  589.2350  731.4446
## 26 theta[2, 3]  456.1352  541.5844
## 27 theta[3, 3]  532.3858  824.0556
## 28 theta[4, 3]  389.1457  476.4701
## 29 theta[5, 3]  607.3436  764.1072
## 30 theta[6, 3]  498.6467  523.5634
## 31 theta[1, 4]  306.8974  343.9046
## 32 theta[2, 4]  496.2511  386.8291
## 33 theta[3, 4]  421.2301  598.3627
## 34 theta[4, 4]  652.4830 1096.0114
## 35 theta[5, 4]  438.9726  570.2416
## 36 theta[6, 4]  223.9105  646.8290
## 37 theta[1, 5]  561.3096  809.9009
## 38 theta[2, 5]  691.5540  695.1876
## 39 theta[3, 5]  338.7503  513.0897
## 40 theta[4, 5]  607.6179 1076.1083
## 41 theta[5, 5]  610.0461 1116.5192
## 42 theta[6, 5]  499.4247  653.4270
## 43 theta[1, 6]  633.7006  600.6828
## 44 theta[2, 6]  541.6620  437.8138
```

```
## 45 theta[3, 6]  545.3496  809.3776
## 46 theta[4, 6]  389.8838  553.9753
## 47 theta[5, 6]  514.6094  595.0496
## 48 theta[6, 6]  299.9923  929.8208
```

The results appear satisfactory, with effective sample sizes of more than 250 in both the tail and bulk of the posterior distributions for each parameter. Based on the results of MCMC exploration, it appears that using an MCMC with 2500 iterations with 500 discarded as warmup is likely to produce good results for our simulation study.

3.3 Fit models

```
sims_out <- run_sims(
  data_list = sim_data$masked_dfs,
  zeros_list = sim_data$zeros,
  DGVs = list(lambda = lambda, psi = psi, theta = Theta),
  theta_scenario_id = "BySpecies",
  parallel = TRUE,
  niter = tune_list$min_iter + 1000,
  nburn = tune_list$min_warmup + 500,
  thin = 1,
  save_fits = FALSE,
  save_individual_summaries_list = FALSE,
  directory = here::here("Vignette", "BySpecies")
)
```

```
## Beginning scenario 1.

## 2024-12-28 14:42:49.677064

##   |                                                                      |

## Beginning scenario 2.

## 2024-12-28 14:51:50.698516

##   |                                                                      |

## Beginning scenario 3.

## 2024-12-28 17:00:05.325756

##   |                                                                      |

## Beginning scenario 4.
```

```
## 2024-12-28 22:48:33.074712

##   |                                                                      |

## Beginning scenario 5.

## 2024-12-28 22:56:41.004401

##   |                                                                      |

## Beginning scenario 6.

## 2024-12-28 23:05:17.349226

##   |                                                                      |
```

3.4 Visualize results

Recall that the measurable objectives of our study are to estimate the relative activity rates with estimation error less than 1 call per night and a 95% posterior interval width of less than 3 calls per night. Furthermore, we assume that the monitoring program can afford to validate 4000 calls in total. All of the possible validation designs shown in `sim_data$scenarios_df` are feasible given this budget.

We begin with detailed plots of relative activity rates, occurrence probabilities, and classification probabilities.

```
visualize_parameter_group(
  sim_summary = sims_out,
  pars = "lambda",
  theta_scenario = "BySpecies",
  scenarios = 1:6,
)
```

Based on the simulation results shown in Figure 6, any of validation scenarios 1-6 is expected to produce a posterior mean estimate for each relative activity parameter that is near the true value. All models converged for all scenarios. Coverage varies by scenario for each parameter, but recall that we only fit models to 10 datasets.

```
visualize_parameter_group(
  sim_summary = sims_out,
  pars = "psi",
  theta_scenario = "BySpecies",
  scenarios = 1:6,
)
```

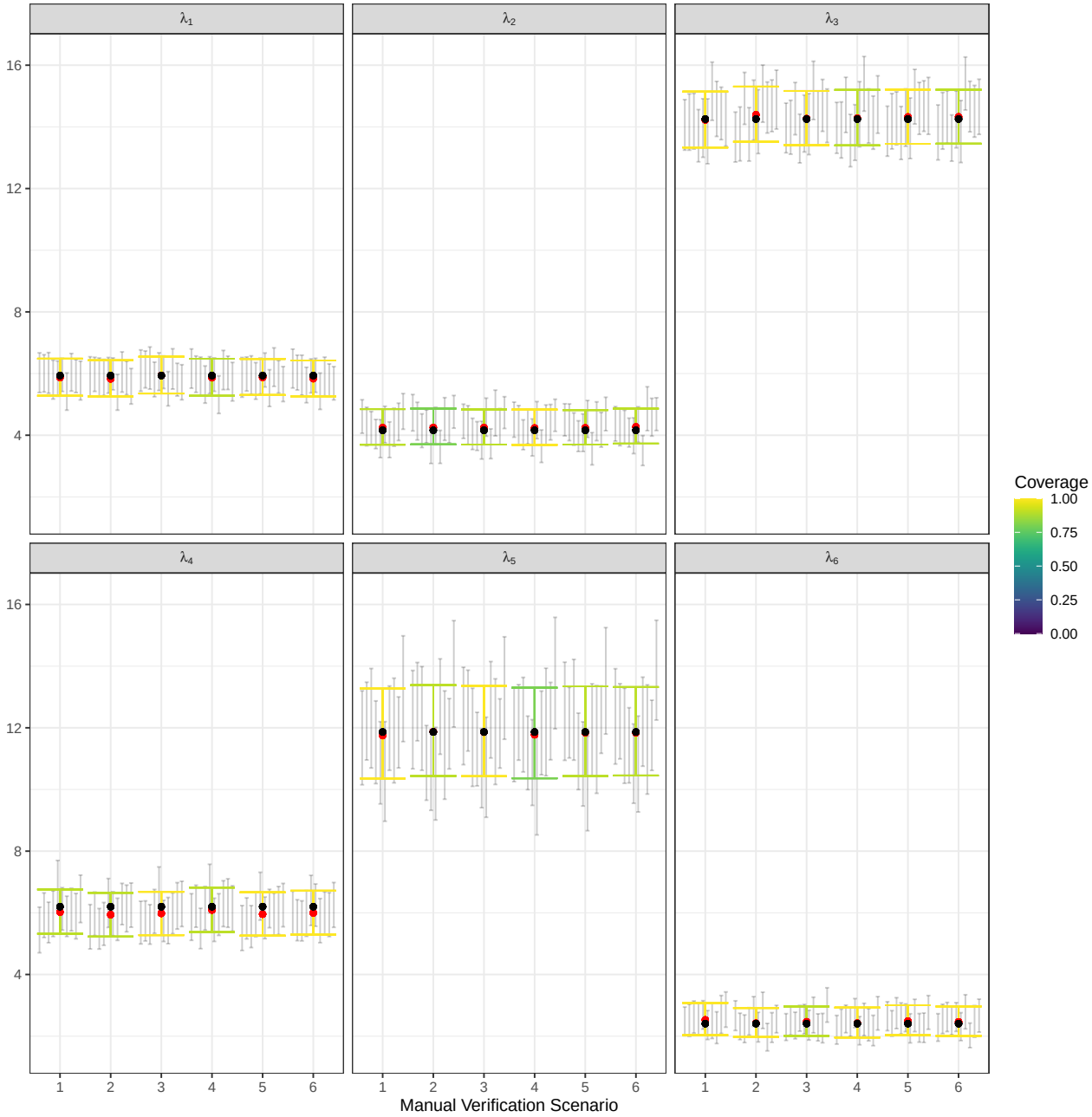


Figure 6: Output from `visualize_parameter_group` under six possible stratified-by-species scenarios for relative activity rates. Parameters are shown in each facet and validation scenario number is on the x-axis. Small grey intervals are 95% posterior intervals for each parameter from fitted models that converged. Larger colored error bars are average 95% posterior intervals with the color indicating the coverage. Black dots are the true parameter value and red dots are average posterior means.

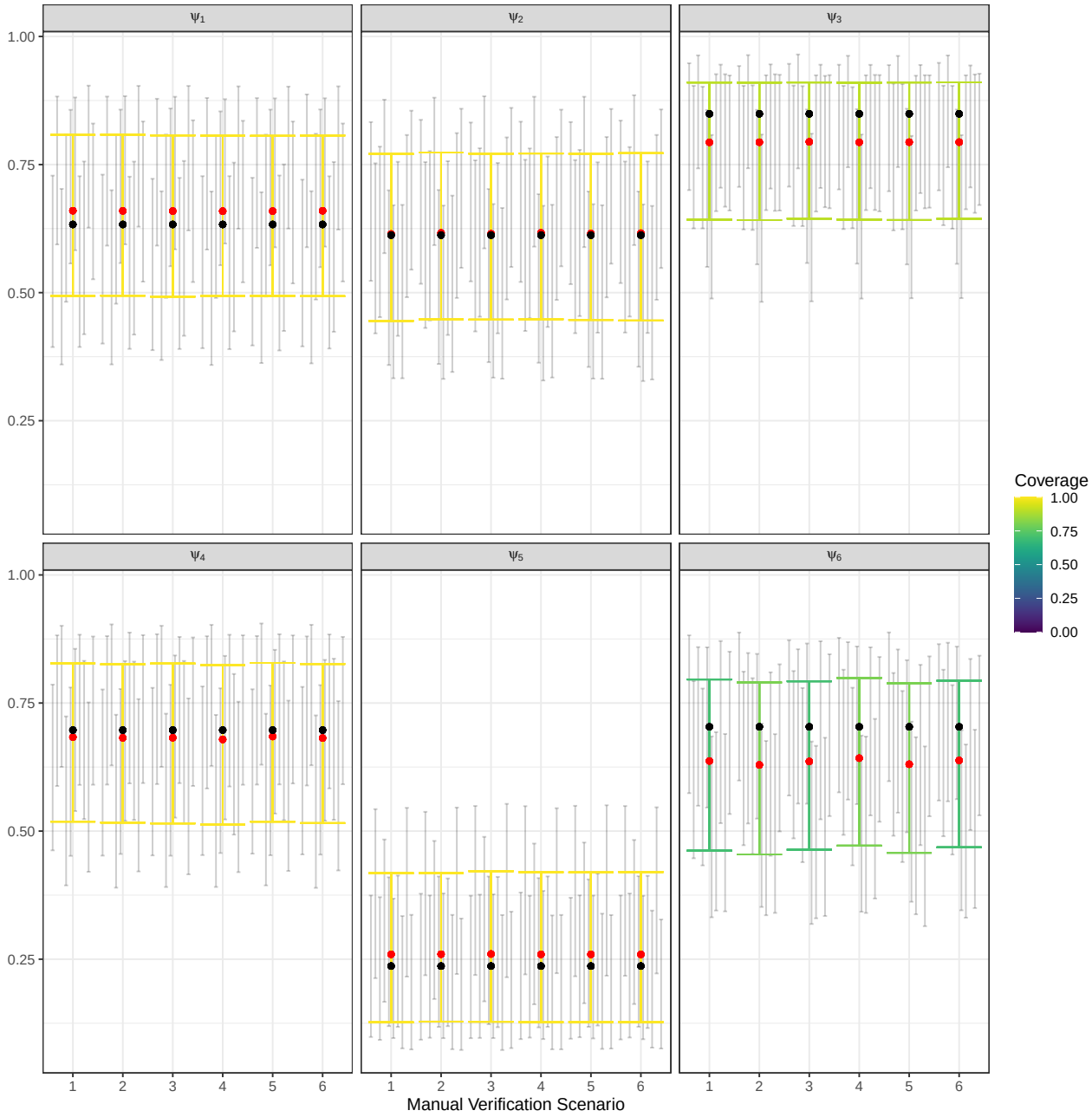


Figure 7: Output from `visualize_parameter_group` under six possible stratified-by-species scenarios for occurrence probabilities. Parameters are shown in each facet and validation scenario number is on the x-axis. Small grey intervals are 95% posterior intervals for each parameter from fitted models that converged. Larger colored error bars are average 95% posterior intervals with the color indicating the coverage. Black dots are the true parameter value and red dots are average posterior means.

The simulation results for occurrence probabilities show a small amount estimation error for all occurrence probabilities, with the size and direction varying depending on the species. For species with larger estimation error, coverage is also slightly low. However, since we are most concerned with relative activity, the goal is to check results for each ψ_k for severe estimation error and/or lack of coverage, which are not shown in Figure 7. Similarly, we do not observe alarming results in Figure 8.

```
visualize_parameter_group(
  sim_summary = sims_out,
  pars = "theta",
  theta_scenario = "BySpecies",
  scenarios = 1:6,
)
```

One measurable objective is for 95% posterior interval widths to be less than 3 calls per night for each species' relative activity rate. As in Section 2.7, the `plot_width_vs_calls` function provides Figure 9, which addresses this measurable objective directly.

```
plot_width_vs_calls(
  sim_summary = sims_out,
  calls_summary = call_sum,
  regex_pars = "lambda",
  theta_scenario = unique(sims_out$theta_scenario),
  scenarios = 1:6
)
```

Once again as in Section 2.7, the widest posterior interval width is for species 5. All validation scenarios have average 95% posterior interval widths near or slightly below 3 calls per night, but the error bars indicate an interval width greater than 3 is possible for any validation scenario. Scenario 6, in which 1785 recordings are validated, has the narrowest expected interval width.

3.5 Take-aways

The results in Section 3.4, imply that, of the stratified-by-species validation designs considered, scenario 6 offered the best results with respect to the measurable objectives we outlined in Section 2.2. In this scenario, the six species in our assemblage receive 15%, 15%, 100%, 10%, 100%, and 25% of their recordings validated, respectively.

To keep the number of scenarios in this example small, we fixed the LOVE for several species. However, if the measurable objectives specified the desired accuracy and precision for estimates of occurrence probability, it may be desirable to consider alternative validation scenarios. For example, we consistently underestimated

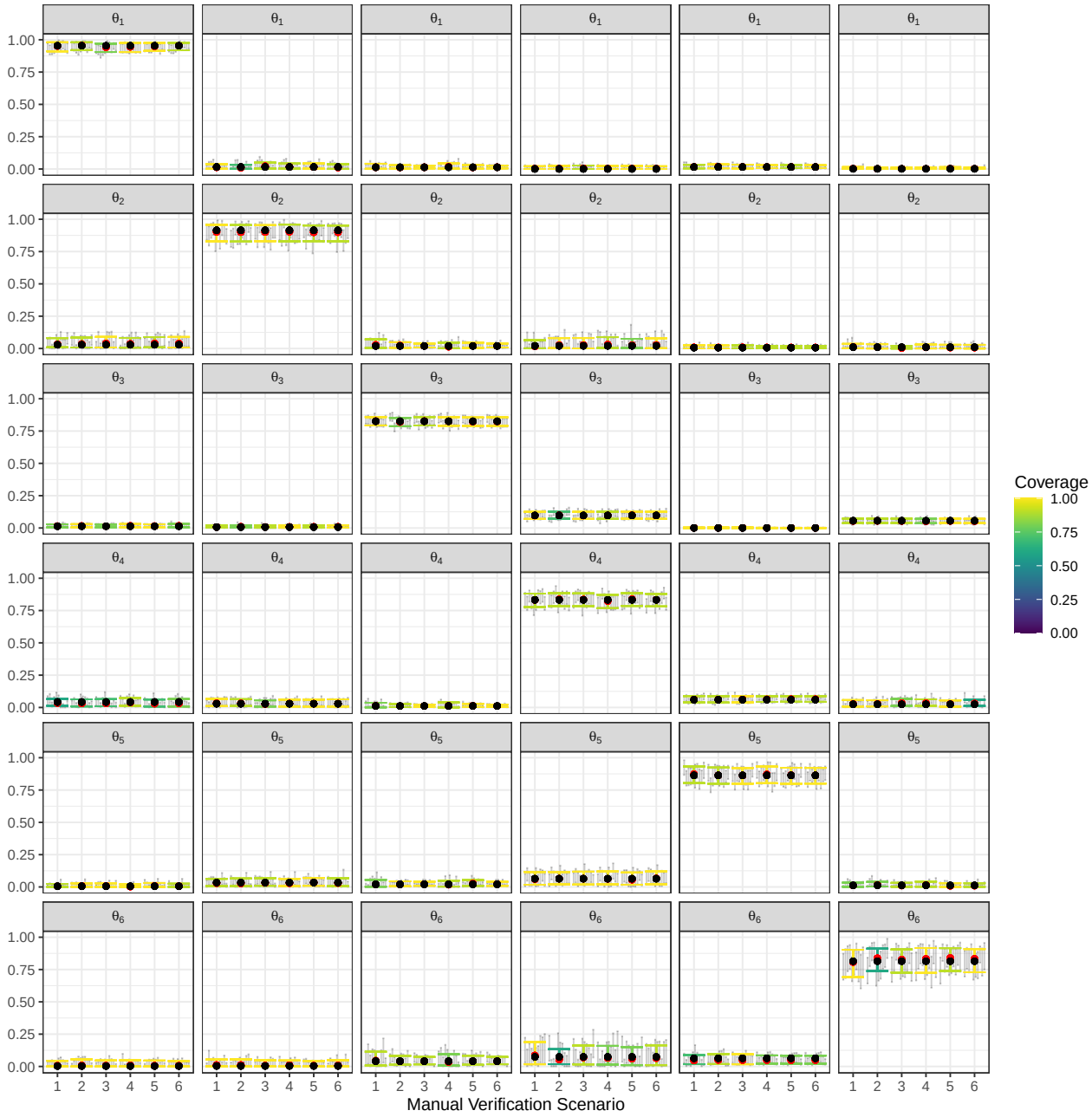


Figure 8: Output from `visualize_parameter_group` under six possible stratified-by-species scenarios for classification probabilities. Parameters are shown in each facet and validation scenario number is on the x-axis. Small grey intervals are 95% posterior intervals for each parameter from fitted models that converged. Larger colored error bars are average 95% posterior intervals with the color indicating the coverage. Black dots are the true parameter value and red dots are average posterior means.

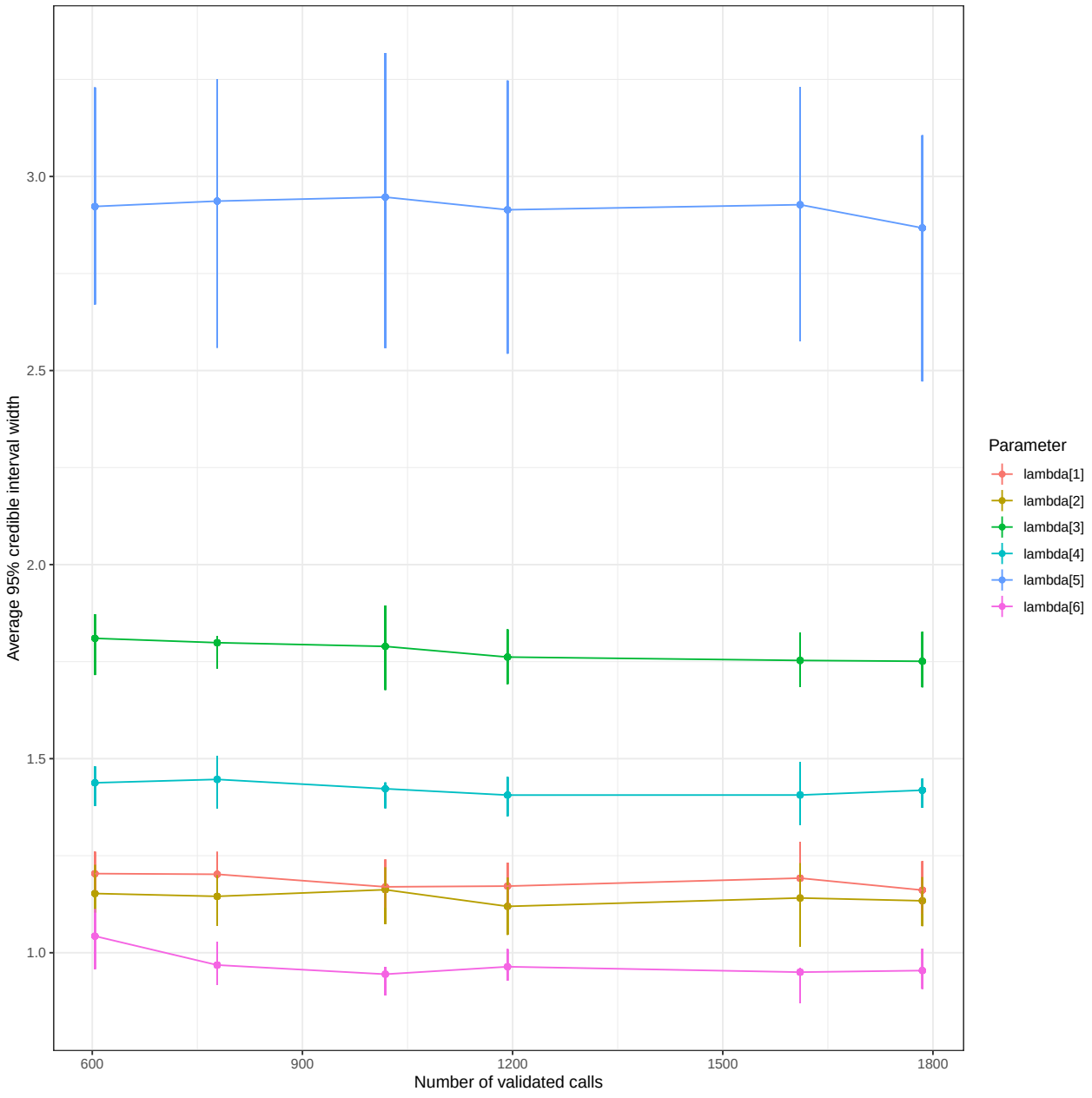


Figure 9: Plots of 95% posterior interval width vs the number of calls validated.

the occurrence probability for species 6 (Figure 7), and additional simulations under validation designs that varied the level of effort for species 6 would be warranted if measurable objectives related to occurrence probability for this species.

4 Conclusion

We have demonstrated the use of the `ValidationExplorer` package when the objective is to compare the merits of four competing LOVEs for fixed-effort and stratified-by-species designs. Specifically, in the fixed-effort design we considered validating a random sample of 5%, 10%, 15% or 30% of the recordings obtained during each visit to each site. In the stratified-by-species designs, we considered a suite of validation designs, fixing effort for species 1,2, 4, and 6 at less than or equal to 25%, and varied the level of effort for species 3 and 6. The exact simulations considered are summarized in the `call_sum` object in Section 3.1. The measurable objectives outlined in Section 2.2 were to estimate the relative activity parameters for each species, with estimation error less than 1 call per night and 95% posterior interval width less than 3 calls per night. Results, and their implications for these measurable objectives were summarized in Sections 2.8 and 3.5.

5 Table of Functions

Function	Argument	Description
simulate_validatedData	n_datasets	The number of datasets to be simulated
	design_type	The type of validation design. Must be one of "BySpecies" or "FixedPercent"
	scenarios	The possible levels of effort. If design_type = "BySpecies", this is provided as a list with vector-valued entries containing the possible percentages to validate for each species. If design_type = "FixedPercent", this argument is a vector of possible percentages.
	n_sites, n_visits, n_species	The number of sites, visits and species in the assemblage
	psi, lambda, theta,	Vectors of length nspecies containing the the parameter values for each species
	directory	The working directory where datasets are to be saved if the following arguments are set to TRUE
	save_datasets, save_masked_datasets	Logicals indicating whether to save each type of dataset
summarize_n_validated	data_list	A list of simulated (masked) datasets output from simulate_validatedData
	zeros_list	A list of true species/autoID combinations that were never observed at each site-visit.
	theta_scenario	An optional character string identifying the classifier scenario. If output from summarize_n_validated is to be used with any of the plot_X_vs_calls functions described below, this string must match the theta_scenario_id argument supplied to run_sims.
run_sims	data_list	A nested list of masked datasets in the format output from simualate_validatedData. The first layer of the list corresponds to scenarios, with each entry containing a list of n_datasets validated according to the scenario.

	`zeros_list`	A list of length `n_datasets` containing the site-visit-true-autoID combinations that were never observed.
	`DGVs`	A named list of the true data-generating values with entries `"psi", "lambda"` and `"theta"`.
	`theta_scenario_id`	An ID to show which classifier the scenario is under. This is provided as a convenience for the user if multiple simulation studies are to be conducted.
	`parallel`	A logical indicating whether MCMC sampling should be fit in parallel (default setting is TRUE). If you have many datasets and many scenarios, we recommend this setting.
	`n_iter, nburn, thin`	The number of iterations, warmup and thinning interval for each chain in the MCMC. Default values are 2000, 1000, and 1, respectively.
	`save_fits`	A logical denoting whether or not to save the draws from individual fitted models. If TRUE, you must have the file structure described in Step 2. Fits will be saved as RDS objects that can be read in later. Default value is FALSE.
	`save_individual_summaries_list`	A logical indicating whether to save individual summary lists that are output after each validation scenario. Default value is FALSE.
	`directory`	Where to save fits and summaries. Required if `save_fits = TRUE` or `save_individual_summaries_list = TRUE`. Default value is the current working directory given by `here::here()`
`visualize_parameter_group`	`sim_summary`	A dataframe in the format of the summaries output by `run_sims`. Column names must match those of the `run_sims` output.
	`pars`	The name of the parameter "group" to be visualized (e.g, `"psi"`, `"lambda"` or `"theta"`).
	`theta_scenario`	The Θ classifier ID.
	`scenarios`	Which scenarios to visualize?

	`convergence_threshold`	What value should $\hat{R}$ be below to be considered "converged"? Default value is 1.1. This value matters because only model fits where all parameter values are below the `convergence_threshold` are used for visualization.
`visualize_single_parameter`		Arguments are identical to `visualize_parameter_group`
`plot_bias_vs_calls`, `plot_coverage_vs_calls`, `plot_width_vs_calls`	`sims_summary`	The summary output in the format from `run_sims`
	`calls_summary`	A summary of the number of calls validated per scenario. Expected format is that of output from `summarize_n_validated`.
	`pars`	The parameters to visualize.
	`regex_pars`	A group of parameters to visualize. One of `"psi"`, `"lambda"` or `"theta"`.
	`theta_scenario`	The classifier scenario ID.
	`scenarios`	The scenarios to be compared.
	`convergence_threshold`	At what value is an MCMC algorithm considered "converged"?

Table 2: Argument descriptions for each function in the `ValidationExplorer` software. My hope is that in the eventual final paper, this will be a scrollable html table

References

Gabry, Jonah, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. 2019. "Visualization in Bayesian Workflow." *J. R. Stat. Soc. A* 182: 389–402. https://doi.org/10.1111/rssa.12378.

Loeb, Susan C., Thomas J. Rodhouse, Laura E. Ellison, Cori L. Lausen, Jonathan D. Reichard, Kathryn M. Irvine, Thomas E. Ingersoll, et al. 2015. "A Plan for the North American Bat Monitoring Program (NABat)." SRS-208. USDA Forest Service.

Stratton, Christian, Kathryn M. Irvine, Katharine M. Banner, Wilson J. Wright, Cori Lausen, and Jason Rae. 2022. "Coupling Validation Effort with in Situ Bioacoustic Data Improves Estimating Relative Activity and Occupancy for Multiple Species with Cross-Species Misclassifications." *Methods in Ecology and Evolution* 13 (6): 1288–1303. https://doi.org/https://doi.org/10.1111/2041-210X.13831.