Vignette: How to simulate and fit data with stratified-by-species sampling for validation

Jacob Oram

2024-01-18

Introduction

This document accompanies the manuscript "Investigation into stratified-by-species validation of species labels for acoustic surveys", so that the reader may conduct similar simulations on their own. The manuscript provides a complete example analysis; here, we focus on the mechanics of setting up and running the necessary code to recreate the simulation study and explore alternative validation designs for different species assemblages.

[**A brief paragraph describing the necessary terms and key workflow attributes for this vignette to make sense.**]

Conducting your own simulation study

Step 1: Installing and loading required packages

After cloning this repo, the next step is loading the necessary packages in R. To simulate, fit and visualize models used with a stratified-by-species validation design using our code, the following packages are necessary:

- `tidyverse`
- `nimble`
- `coda`
- `rstan`
- `parallel`
- `here`

If you do not have one or more of these packages installed, run the following, with the name of the missing package in place of `your_package_name_here`:

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

After installing the necessary packages, load these libraries by calling

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

Step 2: Set up your working directory

In the course of the simulation study, several objects are saved:

- simulated datasets (optional)
- simulated datasets after validation (optional)
- model fits (optional)
- individual summaries for one dataset/validation scenario combination (optional)
- overall summaries for each validation scenario (always saved)

Functions that save any part of the simulation require a `directory` argument be specified. If you choose to save model fits or individual summaries, our simulation functions expect your working directory to contain the folders `your/working/directory/ThetaID/fits` and `your/working/directory/ThetaID/individual_summaries`, where `ID` is replaced with the classifier scenario ID. Visually, this file structure should appear as follows:

- Your working directory
  - ThetaID
    - fits
    - individual_summaries

The Testing folder in this repo provides an example. Here, we use numbers for the classifier scenarios. The first scenario has `ID=1`, giving the paths to the necessary directories `your/working/directory/Testing/Theta1/fits` and `your/working/directory/Testing/Theta1/individual_summaries`.

Step 3: Simulate data

Step 4: Fit the data

Step 5: Visualize simulations

Using saved datasets and model fits