BAYESIAN HIERARCHICAL LATENT VARIABLE MODELS FOR ECOLOGICAL DATA TYPES

by

Christian Alexander Stratton

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

Doctor of Philosophy

in

Statistics

MONTANA STATE UNIVERSITY
Bozeman, Montana

May, 2022

ACKNOWLEDGEMENTS

I would like to sincerely thank my advisor Andrew Hoegh for his support, patience, and guidance over my time at Montana State University. In every way, Dr. Hoegh is an exemplary advisor and embodies the researcher I aspire to become. I am extremely grateful for my committee members, Jennifer Green, Kathryn Irvine, Katharine Banner, and Mark Greenwood, who were so generous with their time, support, and guidance. I would also like to thank my fellow graduate students for their help along the way.

I thank Eli Meyer for the support of a true friend, and for making me laugh often enough to stay sane. I will not forget lunches at TNC or long nights at the school, where perhaps the least amount of work was done. I am also grateful for Sarah Stratton for her humor and support, and am happy to have had her as a college roommate. Finally, I thank Meaghan Winder for her limitless support and help making it to the finish line.

I am also grateful for many other people at Montana State that impacted my time here. I thank Stacey Hancock and Jade Schmidt for my development as a teacher and Katie Sutich and Jane Crawford for their tireless support in the math office. Finally, I thank my parents for their heartfelt encouragement.

Without the support of these people, none of this would have been possible. I am truly grateful for my time at Montana State University, and appreciate all the relationships I have forged in that time.

TABLE OF CONTENTS

TABLE OF CONTENTS – CONTINUED

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES – CONTINUED

LIST OF FIGURES – CONTINUED

LIST OF FIGURES – CONTINUED

Figure Page

ABSTRACT

Ecologists and environmental scientists employ increasingly complicated sampling designs to address research questions that can help explain the impacts of climate change, disease, and other emerging threats. To understand these impacts, statistical methodology must be developed to address the nuance of the sampling design and provide inferences about the quantities of interest; this methodology must also be accessible and easily implemented by scientists. Recently, hierarchical latent variable modeling has emerged as a comprehensive framework for modeling a variety of ecological data types. In this dissertation, we discuss hierarchical modeling of multi-scale occupancy data and multi-species abundance data. Within the multi-scale occupancy framework, we propose new methodology to improve computational performance of existing modeling approaches, resulting in a 98% decrease in computation time. This methodology is implemented in an `R` package developed to encourage community uptake of our method. Additionally, we propose a new modeling framework capable of simultaneous clustering and ordination of ecological abundance data that allows for estimation of the number of clusters present in the latent ordination space. This modeling framework is also extended to accommodate hierarchical sampling designs. The proposed modeling framework is applied to two data sets and code to fit our model is provided. The software and statistical methodology proposed in this dissertation illustrate the flexibility of hierarchical latent variable modeling to accommodate a variety of data types.

CHAPTER ONE

INTRODUCTION

As ecological change continues to grow in complexity and pace, ecologists and environmental scientists must employ increasingly complicated sampling designs in order to address research questions that investigate the impacts of climate change, wildfires, and other emerging threats to ecosystems of interest. To understand these impacts, researchers often seek to make inferences about quantities that are not directly observed, including the occupancy status of multiple species at sample locations or the relationships between entire species assemblages at various sample locations. To estimate these latent quantities, ecologists and environmental scientists often employ hierarchical sampling designs that provide spatial replication by sampling secondary sampling units within primary sampling units (Eiler et al. 2018; Erickson, Merkes, and Mize 2019; Hunter et al. 2019; Yeo et al. 2009).

Statistical methodology must be developed that can appropriately address the hierarchical sampling designs that are commonly used in ecological applications, while simultaneously providing inferences about the quantities of interest. Additionally, this methodology must be accessible and easily implemented by scientists, if it is to be impactful. Recently, hierarchical latent variable models have emerged as a statistical technique capable of both accounting for the hierarchical sampling designs that are common in ecological applications and allowing inferences on important, yet often unobserved, quantities (Jain and Dubes 1988). This dissertation is the compilation of three manuscripts that use hierarchical latent variable models to make inferences about species distributions and draw comparisons between ecological communities. Additionally, the manuscripts herein focus on making models

accessible to practicing scientists. Below, additional background information is provided for each manuscript.

1.0.1 Primer for Chapter 2

Chapter 2, "`msocc`: Fit and analyze computationally efficient multi-scale occupancy models in `R`," describes the development of an `R` package that fits computationally expedient multi-scale occupancy models. Multi-scale occupancy models have recently gained traction as the primary analytical framework for environmental DNA (eDNA) surveys. These surveys rely on detection of DNA associated with target organisms to claim presence of those organisms at a survey location, but presence of an organism at a survey location does not imply that its DNA will be present in every sample taken from that survey location (Dorazio and Erickson 2018). To account for potential false-negative detections (when a species is truly present but not detected) at the sample level, multiple samples are taken from each survey location. Then, each sample is tested for presence of the target organism's DNA using polymerase chain-reaction (PCR) chemistry on multiple replicates from each sample. This second level of detection is also prone to false-negative detections, as presence of DNA in a sample does not imply presence of DNA in a replicate from that sample (Dorazio and Erickson 2018).

The sampling design used in eDNA surveys induces a hierarchical dependence structure in the occupancy state of the target organism at the survey location, sample, and replicate level that must be appropriately modeled. Dorazio and Erickson (2018) describe a hierarchical model in which the occupancy states at the survey location and sample levels are treated as latent Bernoulli random variables; they also provide an `R` package to fit their proposed model from a Bayesian perspective. Their package relies on a Markov chain Monte Carlo (MCMC) technique known as Metropolis-Hastings (Hastings 1970) to sample from the posterior distribution of their proposed model. This technique can be slow to converge

to the posterior distribution, requires time-consuming tuning, and is not well-suited to estimate parameters on vastly different scales (Robert 2015; Roberts and Rosenthal 2001). However Gibbs sampling, a more computationally efficient posterior sampling technique, is not directly available for their model as there is no known closed-form full conditional posterior distribution for Bernoulli sampling models parameterized by log-odds.

Our novel contribution is the development of an `R` package that leverages a data augmentation strategy to afford Gibbs sampling of all parameters in the multi-scale occupancy model proposed by Dorazio and Erickson (2018). This augmentation strategy, developed by Polson, Scott, and Windle (2013), relies on introducing Pólya-gamma distributed random variables within the MCMC routine, resulting in a conditionally Gaussian likelihood for which Gibbs sampling techniques are well defined. Below, we briefly outline the data augmentation strategy, but full details are provided in Polson, Scott, and Windle (2013).

Consider the standard logistic regression model, where y_i represents the number of successes, n_i represents the number of trials, and x_i represents the vector of regressors for observation i ($i = 1, 2, \ldots, N$), and

$$y_i \sim \text{Binomial}\left(n_i, \frac{\exp(x_i'\boldsymbol{\beta})}{1 + \exp(x_i'\boldsymbol{\beta})}\right). \tag{1.1}$$

The Pólya-gamma distribution is constructed such that the following is true:

$$\frac{(e^{\eta})^a}{(1 + e^{\eta})^b} = 2^{-b} e^{\kappa\eta} \int_0^{\infty} e^{-\omega\eta^2/2} p(\omega) d\omega \tag{1.2}$$

where $\kappa = a - b/2$ and $\omega \sim \text{Pólya-gamma}(b, 0)$. Using equation 1.2, we write the likelihood

contribution of observation i as

$$\begin{aligned} L_i(\boldsymbol{\beta}) &\propto \left(\frac{\exp(x_i'\boldsymbol{\beta})}{1+\exp(x_i'\boldsymbol{\beta})}\right)^{y_i}\left(1-\frac{\exp(x_i'\boldsymbol{\beta})}{1+\exp(x_i'\boldsymbol{\beta})}\right)^{n_i-y_i} \\ &= \frac{(\exp(x_i'\boldsymbol{\beta}))^{y_i}}{(1+\exp(x_i'\boldsymbol{\beta}))^{n_i}} \\ &\propto e^{\kappa_i x_i'\boldsymbol{\beta}}\int_0^\infty e^{-\omega_i(x_i'\boldsymbol{\beta})^2/2}p(\omega_i)d\omega_i, \end{aligned} \tag{1.3}$$

where $\kappa_i = y_i - n_i/2$. To derive the full conditional posterior distribution of $\boldsymbol{\beta}$, we consider all observations and condition on $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)$:

$$\begin{aligned} p(\boldsymbol{\beta}|\boldsymbol{\omega}, \boldsymbol{y}) &\propto p(\boldsymbol{\beta})\prod_{i=1}^N L_i(\boldsymbol{\beta}|\omega_i) \\ &= p(\boldsymbol{\beta})\prod_{i=1}^N \exp\left(\kappa_i x_i'\boldsymbol{\beta} - \omega_i(x_i'\boldsymbol{\beta})^2/2\right) \\ &\propto p(\boldsymbol{\beta})\prod_{i=1}^N \exp\left(\frac{\omega_i(x_i'\boldsymbol{\beta})^2}{2} - \kappa_i x_i'\boldsymbol{\beta} + \frac{\omega_i}{2}\left(\frac{\kappa_i}{\omega_i}\right)^2\right) \\ &\propto p(\boldsymbol{\beta})\prod_{i=1}^N \exp\left(\frac{\omega_i}{2}(x_i'\boldsymbol{\beta} - \kappa_i\omega_i)^2\right) \\ &\propto p(\boldsymbol{\beta})\exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{\Omega}(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{\beta})\right), \end{aligned} \tag{1.4}$$

where $\boldsymbol{z} = (\kappa_1/\omega_1, \ldots, \kappa_n/\omega_n)$ and $\boldsymbol{\Omega} = \text{diag}(\omega_1, \ldots, \omega_n)$. In equation 1.4, we recognize the kernel of a $\mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Omega}^{-1})$ distribution with working responses $\boldsymbol{z}$. If the prior on $\boldsymbol{\beta}$ is chosen to be $\mathcal{N}(\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0})$, then $\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{\omega} \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{V})$, where $\boldsymbol{V} = (\boldsymbol{X'\Omega X} + \boldsymbol{\Sigma_0^{-1}})^{-1}$ and $\boldsymbol{m} = \boldsymbol{V}(\boldsymbol{X'\Omega z} + \boldsymbol{\Sigma_0^{-1}\mu_0})$.

This data augmentation strategy is implemented within the `R` package described in Chapter 2 at multiple levels to allow for Gibbs sampling of all parameters in the multi-scale occupancy model. Additionally, Chapter 2 fully details the computational advantage afforded through data augmentation and showcases an example analysis using the package. Finally, Chapter 2 describes a supplemental web application that leverages the computational

expediency afforded by the data augmentation strategy to conduct simulation-based precision analyses for multi-scale occupancy models.

1.0.2 Primer for Chapter 3

Chapter 3, "Assessing spatial and temporal patterns in sagebrush steppe vegetation communities, 2012-2018," describes a collaborative effort to assess sample size requirements for the monitoring of sagebrush steppe vegetation communities in Grand Teton National Park. Between 2012 and 2018, cover class data were collected on over 80 species across 30 sampling frames in Grand Teton National Park. Within each sample frame, approximately 50 one-square-meter quadrats were randomly selected and visual cover class estimates, defined as the proportion of the quadrat obstructed by the canopy associated with each species, were provided for all plant species present within each quadrat. The analytical goal for this work was to use this existing data pipeline to assess the similarity in species composition between sample frames over both time and space. Then, based on those assessments, recommendations regarding whether a reduction in sampling effort was possible were provided.

At the direction of our collaborators, distance-based ordination techniques were used to assess similarity in species composition between sample frames over both time and space. These techniques rely on first constructing a dissimilarity matrix describing the dissimilarity in sample locations based on species composition. Then, that matrix is projected into a lower dimensional space that can be used to visualize the similarity in sample locations based on species composition by assessing each sample locations' proximity to one another in the lower dimensional space. Finally, sample locations are clustered based on species composition by applying algorithm-based clustering techniques, such as k-means clustering, in the lower dimensional space (Roberts 2020).

A large part of the work described in Chapter 3 focuses on developing simulation studies

to assess the impact of choices built into the distance-based ordination framework, including the choice of dissimilarity index. Using simulation, we investigated how the choice of dissimilarity index impacted the resulting ordination, and which dissimilarity index was best suited to correctly identify dissimilar communities. Based on these simulations, ordination results for the Grand Teton data were provided and summarized in a supplemental web application. The ordination results were then used to provide recommendations regarding a reduction in sampling effort for the monitoring program.

Through this work, we identified a number of potential shortcomings associated with distance-based ordination techniques. Among those shortcomings is the inability of distance-based techniques to appropriately account for the hierarchical sampling design that was used in Grand Teton National Park. Consequently, distance-based techniques result in ordination of species composition at the quadrat level, rather than the sample frame level, which was the primary inferential unit of interest to managers. Therefore, to make inferences at the sample frame level, ordination results must be aggregated *post-hoc*. Additionally, it can be difficult to cluster ordination results in the lower dimensional space, as the number of clusters present is seldom known *a priori*, yet this value is required for all algorithm-based clustering techniques. And finally, the distance-based framework does not provide a likelihood. Without a likelihood, it is not possible to formally assess the choice of distance measure, projection method, clustering technique, or number of clusters present in the latent space. The culmination of these observations motivated us to consider alternatives to distance-based ordination techniques, which is the subject of Chapter 4.

1.0.3 Primer for Chapter 4

Chapter 4, "Clustering and unconstrained ordination with Dirichlet process mixture models," describes the development of a hierarchical ordination model capable of simultaneous clustering and ordination that allows for estimation of the number of distinct ecological

communities present in a monitored region. This modeling framework builds on the model-based ordination framework developed by Hui (2016), which uses latent variables drawn from a finite mixture model to cluster sample locations along a latent ecological gradient based on species composition. The Hui (2016) model-based approach provides a number of advantages over the distance-based approach described in Section 1.0.2, including a likelihood with which to assess model fit. Additionally, by drawing the latent variables representing each sample locations' position along the underlying ecological gradient from a finite mixture distribution, Hui (2016) provide a model-based approach to assigning sample units to clusters in the latent space. However, their method still requires prior specification of the number of clusters present in the latent space, and does not explicitly discuss a way to account for the hierarchical sampling design that was present in the Grand Teton data described in Section 1.0.2.

Our novel contribution with this work is two-fold. First, we extend the model-based ordination framework developed by Hui (2016) to draw latent variables representing each sample locations' position in the underlying ecological gradient from an *infinite* mixture distribution, allowing for estimation of the number of clusters present in the latent space. As a result, researchers are able to make probabilistic statements about the number of clusters present in the latent space, rather than relying on expert knowledge or information criterion to make that determination. Second, we extend our model to accommodate hierarchical sampling designs by including random effects for each secondary sampling unit. As a result, ordination results are aligned with the primary sampling unit. In the context of the Grand Teton data described in Section 1.0.2, our proposed model results in ordination of the 30 sampling frames, rather than the approximately 1500 quadrat locations, allowing researchers to make inferences regarding the primary sampling unit without aggregation of ordination results *post-hoc*.

Chapter 4 fully details this modeling framework and showcases its implementation on

two example data sets. The first data set concerns presence-absence records of fish in the Doubs river in Eastern France, and the second data set describes presence-absence records of plant species in Craters of the Moon National Monument and Preserve (CRMO) in Idaho, USA. The former data set is used as a historical benchmark for ordination techniques, and the latter is used to showcase how the proposed model accomodates hierarchical sampling designs. Tools to summarize the posterior distribution and visualize model results are also provided for each data set. Code to fit our proposed model is provided in Appendix B.

CHAPTER TWO

`MSOCC`: FIT AND ANALYZE COMPUTATIONALLY EFFICIENT MULTI-SCALE OCCUPANCY MODELS IN R

2.1 Contribution of Authors and Co-Authors

Manuscript in Chapter 2

Author: Christian Stratton

Contributions: Responsible for writing of manuscript and submission, authored `msocc` package and supplemental web application, led data analysis.

Author: Adam Sepulveda

Contributions: Provided motivation for conceptualization of work and feedback on manuscript draft, assisted in response to reviewer comments.

Author: Andrew Hoegh

Contributions: Provided general guidance and feedback on manuscript draft, contributed to code used in `msocc` package, contributed to development of modeling framework, assisted in response to reviewer comments.

2.2 Manuscript Information

Christian Stratton, Adam Sepulveda, Andrew Hoegh

Methods in Ecology and Evolution

Status of Manuscript:
____ Prepared for submission to a peer-reviewed journal
____ Officially submitted to a peer-reviewed journal
____ Accepted by a peer-reviewed journal
_X_ Published in a peer-reviewed journal; this chapter is presented as the manuscript appears in the journal

British Ecological Society
Submitted 21 February 2020
Published 02 July 2020
Volume 11, Number 9
DOI: 10.1111/2041-210X.13442

Abstract

1. Environmental DNA (eDNA) sampling is a promising tool for the detection of rare and cryptic taxa, such as aquatic pathogens, parasites, and invasive species. Environmental DNA sampling workflows commonly rely on multi-stage hierarchical sampling designs that induce complicated dependencies within the data. This complex dependence structure can be intuitively modeled with Bayesian multi-scale occupancy models. However, current software for such models are computationally demanding, impeding their use.

2. We present an R package, `msocc`, that implements a data augmentation strategy to fit fully Bayesian, computationally efficient multi-scale occupancy models. The `msocc` package allows users to fit multi-scale occupancy models, to estimate and visualize posterior summaries of site, sample, and replicate level occupancy, and to compare different models using Bayesian information criterion. Additionally, we provide a supplemental web application that allows users to investigate study design for multi-scale occupancy models and acts as a graphical user interface to the `msocc` package.

3. The utility of the `msocc` package is illustrated on a published data set and the functions in `msocc` are compared to the primary Bayesian toolkit for multi-scale occupancy modeling, `eDNAoccupancy`, using various computational benchmarks. These benchmarks indicate that `msocc` is capable of fitting models 50 times faster than `eDNAoccupancy`.

4. We hope that access to software that efficiently fits, analyzes, and conducts study design investigations for multi-scale occupancy models facilitates their implementation by the research and wildlife management communities.

2.3 Introduction

Environmental DNA (eDNA) surveys continue to gain popularity for detecting invasive, cryptic, and rare species (Klymus et al. 2015; Lodge et al. 2012; Schmelzle and Kinziger 2016; Sepulveda et al. 2019), as these techniques are often easier, less expensive, and more sensitive than non-molecular detection tools (Eiler et al. 2018; Hunter et al. 2019; Sengupta et al. 2019; Sepulveda et al. 2019; Williams et al. 2018). These surveys rely on hierarchical sampling techniques to accommodate spatial heterogeneity in the occurrence of DNA within a study region. This hierarchical design is necessary, as presence of a species within a site does not imply that its DNA will be present in every sample taken from that site (Dorazio and Erickson 2018). Therefore, multiple samples are typically taken from each site. These samples are then assessed for the presence of DNA using polymerase chain reaction (PCR) chemistry on multiple replicates from each sample. However, even if DNA is present in the sample, it may not be present in all PCR replicates (Dorazio and Erickson 2018).

This sampling design induces a hierarchical dependence structure in the occupancy state of the target species at the site, sample, and replicate levels that must be appropriately modeled. Following Nichols et al. (2008) and Mordecai et al. (2011), Dorazio and Erickson (2018) proposed a hierarchical model of latent state variables to handle this dependence structure. The resulting multi-scale occupancy model provides a flexible platform for modeling multiple levels of uncertainty in the detection of the target species by accounting for false negatives at the site, sample, and replicate levels while simultaneously respecting the hierarchical dependence.

Dorazio and Erickson (2018) provide an R package, `eDNAoccupancy`, for fitting their model from a Bayesian perspective. Bayesian techniques provide a natural way to quantify the uncertainty in the estimated occupancy parameters, and are valuable tools for hierarchical modeling of ecological data (Dorazio 2015). While Bayesian hierarchical techniques

are valuable for understanding complex data structures, fitting hierarchical models can be computationally demanding. Fitting such models for binary data, with a logistic link, has historically required the use of a Markov chain Monte Carlo (MCMC) technique known as Metropolis-Hastings; this sampling technique is implemented in `eDNAoccupancy`. This technique can be slow to converge to the target posterior distribution and also requires time-consuming tuning (Robert 2015). Additionally, the algorithm is not well suited to estimate parameters on vastly different scales (Roberts and Rosenthal 2001), which often forces users to center and scale their continuous predictors, resulting in cumbersome interpretations. Due to these limitations, software packages that rely on this technique can be time-consuming to use, especially when fitting multiple models with the intent of model comparison. The `msocc` package provides an alternative to these packages, implementing a Gibbs sampler that quickly converges to the posterior distribution and features a web application capable of investigating study design, fitting models, and analyzing the results.

The `msocc` package implements the data augmentation strategy described by Polson, Scott, and Windle (2013) to fit the hierarchical model described by Dorazio and Erickson (2018) using a Gibbs sampler; this procedure requires no tuning, easily handles covariates on differing scales, and quickly converges to the posterior distribution. The remainder of this article is organized as follows: Section 2.4 contains a description of our package's features, including an example analysis on published data. Section 2.5 describes the supplemental web application developed to conduct precision analyses and act as a graphical user interface to the `msocc` package. Section 2.6 illustrates the computational advantage of the `msocc` package over `eDNAoccupancy` by comparing computational benchmarks.

2.4 Package Overview

The main function in the `msocc` package is `msocc_mod()`, which fits the multi-scale occupancy model described by Dorazio and Erickson (2018); for complete model notation,

readers are referred to Dorazio and Erickson (2018). The syntax of `msocc_mod()` is as follows:

```
> msocc_mod(wide_data, num.mcmc = 1000,
+  site = list(model = ~1, cov_tbl),
+  sample = list(model = ~1, cov_tbl),
+  rep = list(model = ~1, cov_tbl),
+  priors = list(
+    site = list(mu0 = 0, Sigma0 = 9),
+    sample = list(mu0 = 0, Sigma0 = 9),
+    rep = list(mu0 = 0, Sigma0 = 9),
+    a0 = 1, b0 = 1
+  ),
+  progress = T, print = NULL, seed = NULL, beta_bin = T
+ )
```

The `site`, `sample`, and `rep` arguments each take lists containing items named `model` and `cov_tbl`, respectively. The `model` item is used to specify the formula which determines the model frame for each level of the hierarchy; the `cov_tbl` item is used to specify the data frame containing the covariates used in `model` at each level. This design allows users to specify models using familiar `lm` and `glm` syntax, resulting in a function that is widely accessible.

The remaining arguments of `msocc_mod()` are the number of iterations in the Gibbs sampler (`num.mcmc`), the detection data (`wide_data`), whether progress should be printed (`progress`), how often progress should be printed (`print`), whether a faster beta-binomial sampler should be used when applicable (`beta_bin`), specification of priors (`priors`), and an optional seed to set for reproducibility (`seed`). The `msocc_mod()` function returns a list (of class `msocc`) with the following elements:

- `beta` = matrix of posterior samples of the regression coefficients at the site level
- `psi` = vector of posterior samples of the site level occupancy probability parameter; only returned if `beta_bin = TRUE` and `site$model =` ~ 1
- `alpha` = matrix of posterior samples of the regression coefficients at the sample level

- `theta` = matrix of posterior samples of the sample level occupancy probability parameter; only returned if `beta_bin = TRUE` and `sample$model =` ~ 1 or `sample$model =` $\sim$ `site`
- `delta` = matrix of posterior samples of the regression coefficients at the replicate level
- `p` = vector of posterior samples of the replicate level occupancy probability parameter; only returned if `beta_bin = TRUE` and `rep$model =` ~ 1
- `model.info` = list of model information, including the design matrices for each level of the hierarchy, the number of sites, the number of samples per site, and the number of replicates per sample.

2.4.1 Posterior summary tools

The `msocc` package is equipped with multiple functions to summarize the joint posterior distribution of the derived probability parameters. The `posterior_summary()` function provides a numerical summary table of the derived probability parameters at each level of the hierarchy. The syntax of `posterior_summary()` is as follows:

```
> posterior_summary(
+  msocc_mod, burnin = 0, level = "overall",
+  quantiles = c(0.025, 0.975), unique = T
+ )
```

The arguments of this function include a fitted model of class `msocc` (`msocc_mod`), the number of samples to discard as warm-up (`burnin`), the level of the model to summarize which may be one of "`overall`," "`site`," "`sample`," or "`rep`" (`level`), quantiles defining the credibility intervals to be provided (`quantiles`), and whether only unique rows of the summary table should be printed (`unique`).

The `cred_plot()` function provides a graphical summary of credibility intervals for each of the derived probability parameters at each level of the hierarchy. The syntax of `cred_plot` is as follows:

```
> cred_plot(
+  msocc_mod, level = "site", truth = NULL, n = "all",
+  quantiles = c(0.025, 0.975), burnin = 0
+ )
```

The arguments of this function include a fitted model of class `msocc` (`msocc_mod`), the level of the model to visualize which may be one of "`site`," "`sample`," or "`rep`" (`level`), the true values of the probability parameters which may be used during simulation (`truth`), the number of credibility intervals to plot at a time (`n`), quantiles defining the credibility intervals to be provided (`quantiles`), and the number of samples to discard as warm-up (`burnin`).

2.4.2 Example analysis

The functions in the `msocc` package are demonstrated on an eDNA survey of tidewater goby, found along the coast of California, USA (Schmelzle and Kinziger 2016). In this survey, water samples were collected from 39 sites along the California coast. The number of samples collected at each site ranged from 2 to 23, and six PCR replicates were tested for the presence of goby DNA from each sample. In addition to detection data, environmental covariates were collected at all 39 sites; the first five rows of the data are provided in Table 2.1.

In these data, `twg` represents an index of goby density (catch-per-effort), `sal` represents the salinity in parts per thousand, `turb` represents the turbidity in filtration time, `fish` represents an index of non-goby fish density (catch-per-effort), and `veg` is a logical indicator for the presence of vegetation. Schmelzle and Kinziger (2016) originally fit a suite of models to these data using the `WINBUGS` package, and Dorazio and Erickson (2018) recreated those

site	sample	pcr1	pcr2	pcr3	pcr4	pcr5	pcr6	twg	sal	turb	fish	veg
Big_Lagoon	1	1	1	1	1	1	1	26.63	1.75	132.00	80.00	1
Big_River	1	0	0	0	0	0	0	0.00	26.00	78.80	17.20	0
Caspar_Creek	1	0	0	0	0	0	0	0.00	20.70	413.00	1.20	0
Elk_Creek	1	0	0	0	0	0	0	0.00	30.00	144.67	19.30	0
Garcia_River	1	0	0	0	0	0	0	0.00	23.50	41.00	0.00	0

Table 2.1: First five rows of tidewater goby data.

results using their `eDNAoccupancy` package. Those results are again recreated here for comparative purposes.

```
> # prep data frames
> site.df <- goby %>%
+  distinct(site, .keep_all = TRUE)
> sample.df <- goby
> detect.df <- goby %>%
+  select(-c(twg:veg))
> # fit model
> goby_mod <- msocc_mod(
+  detect.df, num.mcmc = 11000,
+  site = list(model = ~ veg, cov_tbl = site.df),
+  sample = list(model = ~ sal + twg, cov_tbl = sample.df),
+  rep = list(model = ~ sal + fish + turb, cov_tbl = sample.df)
+ )

Iteration 11000 of 11000; 100% done. Current runtime of 0.66 minutes.
|============================================================| 100%
```

An overall summary of occupancy at all three levels is provided using `posterior_summary()`;

the first six sites are presented below. Note that by default, `posterior_summary()` returns only unique combinations of site, sample, and replicate probabilities.

```
> # numerical summary
> head(posterior_summary(goby_mod, level = "overall", burnin = 1000))
                         site sample rep       psi     theta         p
1                   Big_Lagoon      1   1 0.7706495 0.8845290 0.8669207
2                    Big_River      1   1 0.2231190 0.6650525 0.5446235
3                 Caspar_Creek      1   1 0.2231190 0.7222719 0.8575822
4                   Davis_Lake      1   1 0.7706495 0.8920972 0.8278296
5            Dead_Mouse_Marsh      1   1 0.2231190 0.7242514 0.8292023
6 Eel_River_Estuary_Preserve      1   1 0.7706495 0.6399665 0.6679964
```

These results match those provided by the `eDNAoccupancy` package. Credibility intervals are available for each estimate by specifying a particular level of the model using the `level` argument. These intervals can also be visualized using the `cred_plot()` function; credibility intervals for the sample level occupancy parameter for the first six sites are provided in Figure 2.1.

2.4.3 Additional tools

In addition to the functions described in Sections 2.4.1 and 2.4.2, the `msocc` package contains tools to calculate Bayesian information criterion and simulate data from multi-scale occupancy models. The `waic()` function can be used to calculate the Watanabe-Akaike information criterion (WAIC) on a suite of models (Gelman, Hwang, and Vehtari 2013), while the `msocc_sim()` function is used to simulate data from a multi-scale occupancy model. For details on each functions' use, please see the GitHub page for `msocc` or the R help page for either function.

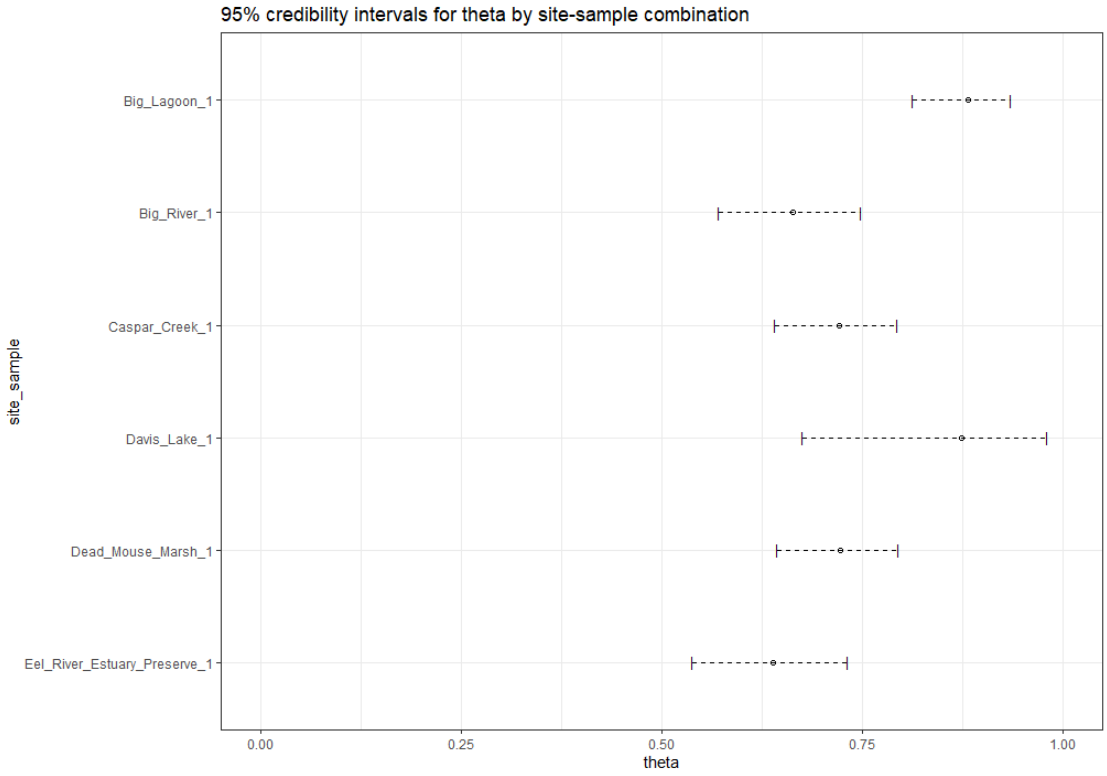

Figure 2.1: Plot of 95% credibility intervals for the sample level occupancy probability parameters of the first six sites of the tidewater goby data.

2.5 Supplemental Web Application

The `msocc` package is also equipped with an R Shiny web application capable of fitting models, visualizing the results from fitted models, and conducting precision analyses. The web application accommodates the first two tasks by providing a point-and-click interface into the `msocc_mod()`, `posterior_summary()`, and `cred_plot()` functions. To fit a multi-scale occupancy model with the `msocc` web application, users must upload the detection and covariate data frames in either .Rdata or .csv format. Once uploaded, the web application allows the user to visualize each data frame and specify the model to be fit (Figure 2.2). The application also allows the user to download the fitted model in .Rdata format.

To analyze a fitted model, the user must upload the model in .Rdata format. Once

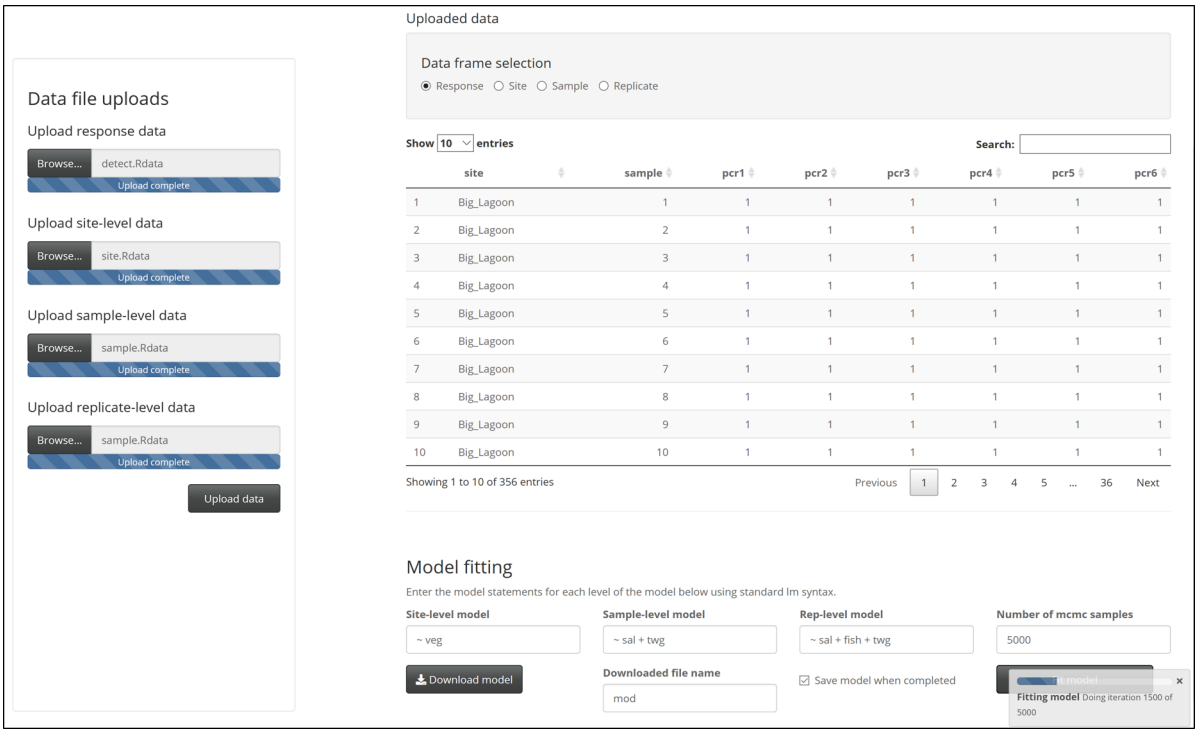

Figure 2.2: Model fitting tab of the `msocc` web application.

uploaded, the application provides a review of the fitted model and prompts the user to view table summaries of the occupancy parameters or visualize credibility intervals for the occupancy parameters (Figure 2.3). Users are also able to visualize traceplots for each parameter, though more formal assessments of convergence are available through the `coda` package (Plummer et al. 2006).

The web application conducts precision analyses by repeatedly calling `msocc_sim()` in the background for varying sample sizes and fitting a model to each simulated data set; the user determines at which level of the model to vary the sample size. Credibility intervals are then calculated for each of the parameters in these models and the widths are stored. These credibility interval widths are then plotted for each level of the model (Figure 2.4). This process allows users of the web application to assess the value of increasing the sample size at any level of the model in terms of precision, which is an important aspect of eDNA-based

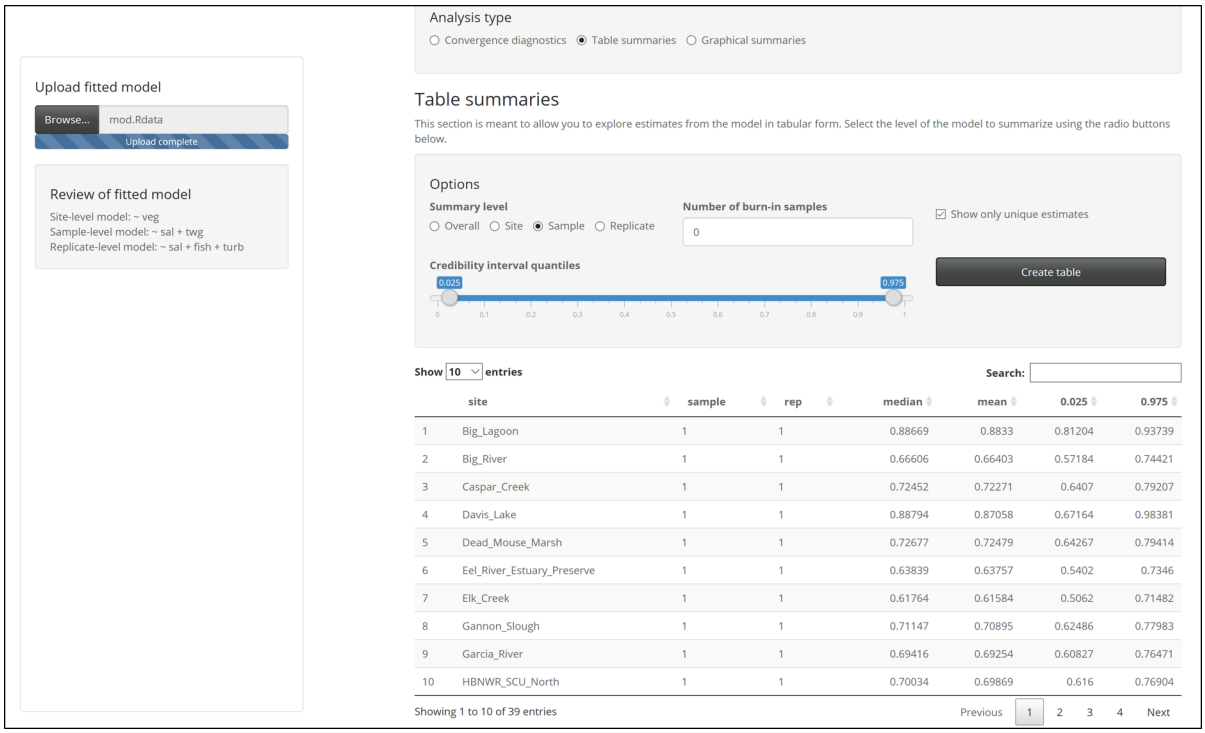

Figure 2.3: Table summaries for the fitted goby model from the `msocc` web application. Convergence diagnostics and graphical summaries of the posterior distribution are also available.

work (Erickson, Merkes, and Mize 2019). Figure 2.4 suggests little to no additional precision is gained by collecting more than five replicates from each sample for the specified design.

2.6 Comparison to `eDNAoccupancy`

The `msocc` package is designed to be more expedient than `eDNAoccupancy`. The `msocc_mod()` function relies on a Gibbs sampler, and therefore converges to the posterior distribution much quicker than `occModel()`, the model fitting function from `eDNAoccupancy`. Additionally, it is not necessary to center and scale covariates before fitting a model with `msocc_mod()`, thereby allowing users to interpret results on the original scale of the data. Gradient-based sampling frameworks, such as Stan or Greta, also generally provide quicker convergence to the posterior distribution (Carpenter et al. 2017). However, many of these

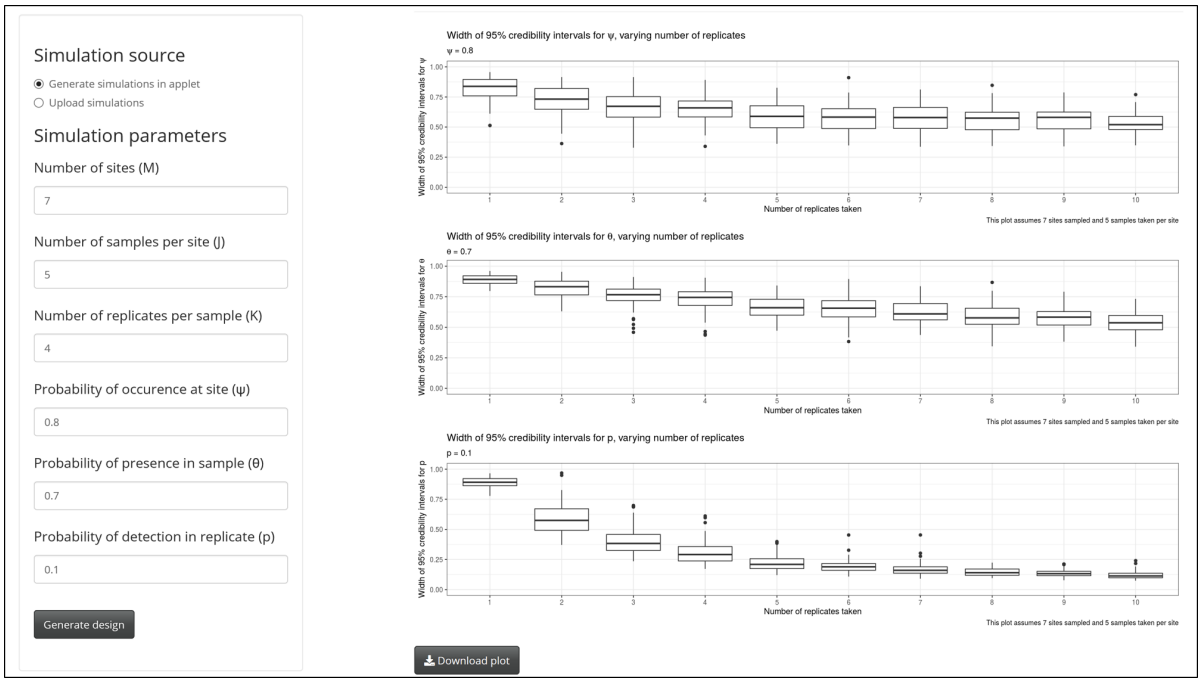

Figure 2.4: Credibility width analysis from the `msocc` web application. This analysis assumes seven sites are surveyed and five samples are taken from each site. The number of replicates taken varies from 1 to 10.

frameworks, including Stan and Greta, do not allow sampling of discrete parameters or latent variables (Golding 2019; Stan Development Team 2018). Consequently, to fit multi-scale occupancy models within these frameworks, practitioners must integrate out the latent variables prior to defining the likelihood (for example, see Mize et al. (2019)). In a hierarchical modeling framework with three levels, this task is non-trivial and not likely to be widely accessible to wildlife managers and other eDNA practitioners. Therefore, we have not included these techniques for comparison, as they lack the convenience and ease-of-implementation associated with canned `R` packages.

To compare these functions, we provide the amount of time taken to draw 11000 samples from the joint posterior of the goby model defined above, and the effective sample size for the `msocc_mod()` and `occModel()` functions per minute using both scaled and unscaled covariates. The effective sample size represents the number of uncorrelated posterior samples

to which an MCMC chain is equivalent, and therefore describes the degree of autocorrelation present in that chain; the `effectiveSize()` function in the `coda` package (Plummer et al. 2006) was used to calculate this value. This function provides effective sample sizes for each parameter in the model; the minimum effective sample size across all parameters for each model was chosen to summarize the fit. This process was repeated 10 times for each model and the results are summarized in Table 2.2. These models were fit on a Surface Book 2 laptop running Windows 10 with an i7-8650U CPU and 16GB of RAM.

Package	Data	Average time	Average ESS	Average ESS/min
msocc	Unscaled	36.41 seconds	4029.61	6640.39
msocc	Scaled	36.07 seconds	4161.01	6921.56
eDNAoccupancy	Unscaled	2198.72 seconds	2958.19	80.71
eDNAoccupancy	Scaled	1839.45 seconds	2908.97	94.87

Table 2.2: Summary of the effective sample size comparison. The values in this table describe the results from drawing 11000 samples from the the joint posterior distribution of the goby model 10 times.

Table 2.2 suggests that `msocc` is far more efficient than `eDNAoccupancy`, allowing models to be fit in seconds as opposed to half-hours. Additionally, models fit using `msocc` tend to have larger effective sample sizes than those fit by `eDNAoccupancy`. Consequently, users are required to take fewer samples from the posterior distribution when using `msocc`, and can do so much quicker than when using `eDNAoccupancy`.

2.7 Discussion

As eDNA surveys continue to gain popularity as a sensitive monitoring strategy for rare and cryptic species, the need for efficient modeling techniques of multi-scale occupancy data increases; the `msocc` package provides an efficient alternative to existing methods of

fitting Bayesian multi-scale occupancy models. This computational advantage allows the research and wildlife management communities the flexibility to fit multiple models when investigating scientific hypotheses, an otherwise time-consuming task. Additionally, `msocc` requires no tuning when fitting models, easily handles covariates on non-standardized scales, and is equipped with a web application capable of conducting precision analyses, fitting models, and exploring model results. The culmination of these factors eases the burden of ecologists working with eDNA data, thereby improving their ability to assess their research questions and disseminate that information. Moreover, *in situ* eDNA workflows continue to gain traction. Such analyses require computationally expedient techniques to understand these data structures in real-time; the ability to do so results in up-to-date information that can be used to minimize negative outcomes and improve management decision-making.

The `msocc` package continues to be developed. In the future, we hope to add a coherent dynamic modeling framework that accommodates longitudinal eDNA surveys. Continued research into dynamic multi-scale occupancy models is essential as longitudinal eDNA monitoring programs gain popularity (Bálint et al. 2018; Hutchins et al. 2019; Pilliod et al. 2019; Uchii et al. 2017). Once these methods have been developed, we intend to add them to `msocc`.

Author contributions

C.S. wrote the `msocc` package, wrote the supplemental web application, and led the writing of the manuscript. A.H. and A.S. provided intellectual guidance on the development of `msocc` and the web application. All co-authors assisted with edits and approved submission.

Acknowledgments

This work was supported by the USGS Ecosystems Mission Area. Any use of trade, firm or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. The time and effort of Christian Stratton and Dr. Andy Hoegh was made possible with a CESU agreement (G17AC00147) between USGS-NOROCK and MSU.

Data accessibility

Data

The tidewater goby data were originally archived by Schmelzle and Kinziger 2016 at Dryad Digital Repository: `http://dx.doi.org/10.5061/dryad.6rs23`. These data files are also accessible in the `msocc` package.

Package, web application, and simulation scripts

The `msocc` package and documentation are hosted at `https://github.com/StrattonCh/msocc`; the package is also archived at the Zenodo online repository (Stratton 2020a). The supplemental web application is hosted at `https://christianstratton.shinyapps.io/eDNAapp/`. The script file used to compare `msocc` and `eDNAoccupancy` is hosted at `https://github.com/StrattonCh/msocc_supplemental_code`; this file is also archived at the Zenodo online repository (Stratton 2020b).

CHAPTER THREE

ASSESSING SPATIAL AND TEMPORAL PATTERNS IN SAGEBRUSH STEPPE VEGETATION COMMUNITIES, 2012-2018

3.1 Contribution of Authors and Co-Authors

Manuscript in Chapter 3

Author: Christian Stratton

Contributions: Responsible for writing of manuscript and submission, authored supplemental simulations, led data analysis, authored supplemental web application for visualization of results.

Author: Andrew Hoegh

Contributions: Provided general guidance and feedback on manuscript draft, contributed to code used to create ordination and visualize results, contributed to development of simulation studies, assisted in response to reviewer comments.

Author: Kathryn M. Irvine

Contributions: Provided general guidance and feedback on manuscript draft, contributed to development of simulation studies, assisted in response to reviewer comments.

Author: Kristin Legg

Contributions: Provided general guidance and feedback on manuscript draft.

Author: Kelly McCloskey

Contributions: Provided general guidance and feedback on manuscript draft.

Author: Erin Shanahan

Contributions: Provided general guidance and feedback on manuscript draft.

Author: Mike Tercek

Contributions: Provided general guidance and feedback on manuscript draft.

Author: David Thoma

Contributions: Provided general guidance and feedback on manuscript draft.

3.2 Manuscript Information

Christian Stratton, Andrew Hoegh, Kathryn M. Irvine, Kristin Legg, Kelly McCloskey, Erin Shanahan, Mike Tercek, David Thoma

Natural Resource Reports

Status of Manuscript:
____ Prepared for submission to a peer-reviewed journal
____ Officially submitted to a peer-reviewed journal
____ Accepted by a peer-reviewed journal
_X_ Published in a peer-reviewed journal; this chapter is presented as the manuscript appears in the journal

National Park Service
Submitted 06 June 2019
Published 17 October 2019
NPS/GRYN/NRR-2019/2020

Abstract

Visual cover class data were collected on over 80 species across 30 permanent sampling frames in sagebrush steppe vegetation communities in Grand Teton National Park from 2012 to 2018. In this report, temporal and spatial patterns in species composition were assessed and used to inform potential sampling strategies for future monitoring. Specifically, the viability of a reduction in sampling effort was evaluated based on the similarity in species composition within each frame over time and among frames within each year. Using distance based ordination techniques, we found little to no evidence of differences in species composition within each frame over time. Furthermore, there was little evidence of heterogeneity in species composition among frames within each year, though there was some evidence of differences in composition between the two principle sagebrush community types (sagebrush dryland shrub and sagebrush-bitterbrush) aggregated across frames. Based on these results, we propose that a reduction in sampling effort is viable and suggest a new monitoring schedule.

3.3 Introduction

In 2010, the National Park Service (NPS) selected indicators in high elevation parks to monitor in the face of a changing global climate (Bingham et al. 2011). Sagebrush vegetation communities was one of the indicators selected as it is considered one of the most threatened ecosystems in the United States and home to sensitive wildlife species, such as sage grouse (*Centrocercus* spp.). In 2012, the NPS implemented a long-term, sagebrush steppe monitoring program in Grand Teton National Park. This program was adapted from the Upper Columbia Basin Network sagebrush steppe monitoring protocol (Yeo et al. 2009). Details of the monitoring and standard operating procedures specific to Grand Teton National Park are described by Yeo and Rodhouse (2013), with key features of the sampling

design summarized in Section 3.4 of this report. This program also aligns with the upland vegetation monitoring implemented by the Greater Yellowstone Network (Tercek et al. 2015). The overall intent of this monitoring program is to evaluate the composition and abundance of both native and invasive plant species over time in sagebrush communities as well as provide reference conditions to assess restoration efforts in former agricultural fields in Grand Teton National Park.

The goals of this report are to assess the temporal and spatial changes in composition and abundance of principal sagebrush steppe plant species in Grand Teton National Park based on data collected from 2012 to 2018 (National Park Service 2018) and determine if future sampling effort can be reduced while maintaining the ability to identify future changes. Specifically, we explored evidence of whether temporal and/or spatial heterogeneity existed in species composition. Answers to these questions informed sampling frequency moving forward. For example, a lack of spatial heterogeneity in species composition and abundance suggests that the number of sampling frames visited yearly can be reduced; where sampling frames exhibit similar community composition, there is not much information to be gained by sampling a greater spatial area. Our exploratory analysis focused on addressing these questions. Consequently, we present alternative sampling strategies to guide future data collection in an effort to ensure a sustainable monitoring program while still maintaining the ability to measure change in species composition and abundance over time.

The report is organized as follows: in Section 3.4, we include a description of the sampling design; in Section 3.5, we first summarize and then visualize the data; in Section 3.6, we present the statistical methods used in this analysis; in Section 3.7, we discuss our results with respect to temporal and spatial patterns; and in Section 3.8, we conclude with a discussion of these results and propose a reduction in sampling effort moving forward.

3.4 Sampling Design

Since 2012, 30 permanent sampling frames have been monitored in Grand Teton National Park. Of these 30 frames, 10 were identified as priority sampling frames (Figure 3.1). Five of these 10 frames were sampled annually (14, 22, 25, 29, and 30), and the remaining 25 frames were sampled at yearly intervals in order to support regular status and trend assessments to address management objectives (Table 3.1; Yeo and Rodhouse 2013). The frame locations were based on their spatial representation within two underlying principle sagebrush steppe community types: sagebrush dry shrubland and sagebrush-bitterbrush. Sagebrush dry shrubland accounts for approximately 75% of the sagebrush steppe in the park, with the remaining 25% comprised largely of sagebrush-bitterbrush. To account for the differences between these communities, 21 of the 30 frames (70%) were selected in the sagebrush dry shrubland and nine frames were selected in the sagebrush-bitterbrush habitat. All frames are located on the Snake River plains east of the Teton Range.

The number of frames sampled annually varied from 12 to 14, with the exception of 2013, when field crews surveyed 22 frames. Within each permanent frame, a spatially-balanced sample of non-permanent locations was selected every year using the generalized random tessellation stratified (GRTS) design described by Stevens and Olsen (2004). With this design, between fifty and fifty-five 1-m^2 quadrats were sampled from each of the targeted frames.

Field technicians visually estimated the canopy cover class of live or current year foliage for principal sagebrush steppe plant species in all quadrats. Canopy cover is defined as the percentage of the ground covered by a vertical projection of the outermost perimeter of the natural spread of foliage of plants (Society of Range Management 1999). The cover class scheme used is a modified Daubenmire scale (Daubenmire 1959), provided in Table 3.2. Appendix A contains a list of all species monitored, the year they were first monitored, and

the species group to which they belong.

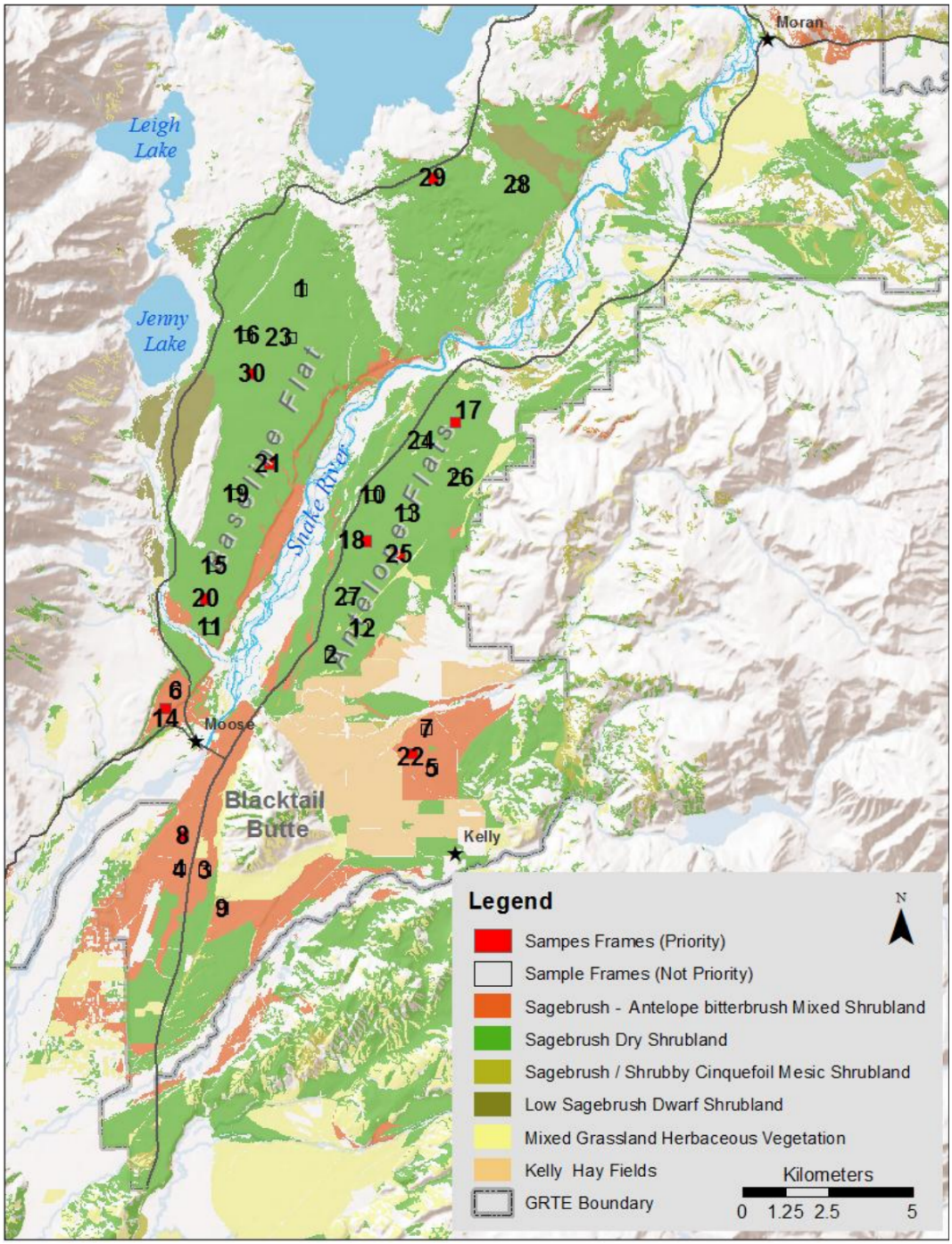

Figure 3.1: Sampling frame locations from Yeo and Rodhouse (2013); priority frames are denoted by red squares.

Plot	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
1	1	-	2	-	-	3	-	-	4	-	-
2	-	1	-	2	-	-	3	-	-	4	-
3	1	-	2	-	-	3	-	-	4	-	-
4	-	1	-	-	2	-	-	3	-	-	4
5	-	1	-	-	2	-	-	3	-	-	4
6	-	1	-	2	-	-	3	-	-	4	-
7	1	-	2	-	-	3	-	-	4	-	-
8	1	-	2	-	-	3	-	-	4	-	-
9	-	1	-	2	-	-	3	-	-	4	-
10	1	-	2	-	-	3	-	-	4	-	-
11	-	1	-	-	2	-	-	3	-	-	4
12	-	1	-	2	-	-	3	-	-	4	-
13	-	1	-	-	2	-	-	3	-	-	4
14	E	E	E	E	E	E	E	E	E	E	E
15	-	1	-	-	2	-	-	3	-	-	4
16	-	1	-	-	2	-	-	3	-	-	4
17	1	-	2	-	-	3	-	-	4	-	-
18	1	-	2	-	-	3	-	-	4	-	-
19	-	1	-	-	2	-	-	3	-	-	4
20	1	-	2	-	-	3	-	-	4	-	-
21	-	1	2	-	-	3	-	-	4	-	-
22	E	E	E	E	E	E	E	E	E	E	E
23	-	1	-	2	-	-	3	-	-	4	-
24	-	1	-	2	-	-	3	-	-	4	-
25	E	E	E	E	E	E	E	E	E	E	E
26	-	1	-	2	-	-	3	-	-	4	-
27	-	1	-	-	2	-	-	3	-	-	4
28	-	1	-	-	2	-	-	3	-	-	4
29	E	E	E	E	E	E	E	E	E	E	E
30	E	E	E	E	E	E	E	E	E	E	E

Table 3.1: Current monitoring schedule. The numbers denote how many times the frame has been sampled since the start of the study; frames labeled "E" are sampled every year.

To understand community composition, visual estimation of plant cover classes is widely recommended (Beck, Connelly, and Reese 2009; Mitchell, Bartling, and O'Brien 1988; Peet, Wentworth, and White 1998). Furthermore, visual coverage is easily interpreted by managers and assessed rapidly in the field. These factors motivated the use of visual estimations of

plant cover in this study, despite its sensitivity to seasonal variation in precipitation and plant phenology (Elzinga et al. 2001). To limit impacts in plant phenology, frames were sampled approximately the same time (late June, early July) every year.

This sampling design and response type are robust to many of the challenges posed by vegetation monitoring across vast, rugged landscapes. For example, it allows for rapid assessment of coverage and thus permits large sample sizes, which are necessary to describe the status of plant species across large areas. For a more detailed discussion of this design and its advantages, see Yeo and Rodhouse (2013).

Cover Class	Range
1	>0-5%
2	>5-25%
3	>25-50%
4	>50-75%
5	>75-95%
6	>95%

Table 3.2: Daubenmire coverage classes and their implied percent coverage.

3.5 Exploratory Data Summary and Analysis

Visualizing a multivariate response (80 species cover classes) over seven years and across 30 frames was challenging. Furthermore, this analysis required summarizing spatial and temporal patterns on each individual species, as well as aggregated at the species group level (Appendix A). Consequently, there were hundreds of graphics of interest. To aid in this endeavor, an internal collaborative web-tool was developed to compare the distribution of coverage classes within frames over time and within a year across frames, for both individual species and aggregated at the species group level. Due to the inherent complexity of this analysis, a subset of plots predominantly focused on sagebrush and invasive forbs are

presented in this report, though all of the graphical comparisons were provided to the agency partners via a desktop web tool. For access to the web app, the reader is asked to contact the Greater Yellowstone Monitoring Network.

To visually assess patterns in species composition in these data, each Daubenmire coverage class was aggregated at either the species or species group level to obtain the relative frequency of occurrence. To calculate these relative frequencies, the total count of each coverage class for the target species (or species group) within a frame and year was divided by the total number of measurement occasions for the target species (or species group). For example, in priority Frame 30 (monitored on an annual basis), 50 quadrats were sampled in 2012. Across those 50 quadrats, up to five species of sagebrush could be observed for a potential total of 250 observations of sagebrush species. Of these 250 observations, 180 had a coverage class of zero, yielding a relative frequency for the zero coverage class of 0.72 (see Figure 3.2 for an example).

In addition to assessing spatial and temporal patterns of native plant species, these relative frequencies are useful in measuring the abundance of invasive plant species. Figure 3.3 displays the distribution of coverage classes for the invasive forbs species group over time in Frame 30. This figure suggests that invasive forbs are seldom observed in this frame over the course of the study. This relationship holds for most other frames, as 96% of the coverage measurements on members of the invasive forbs species group were in the zero class.

3.5.1 Assessing temporal patterns

Species composition and abundance within each frame were examined to identify temporal heterogeneity. To visualize the temporal patterns in the sagebrush-steppe vegetation community, plots were created for each species group within a frame across time. For example, Figure 3.2 depicts the distribution of coverage class over time for the sagebrush species group in Frame 30. Based on this figure, there is little evidence of heterogeneity in

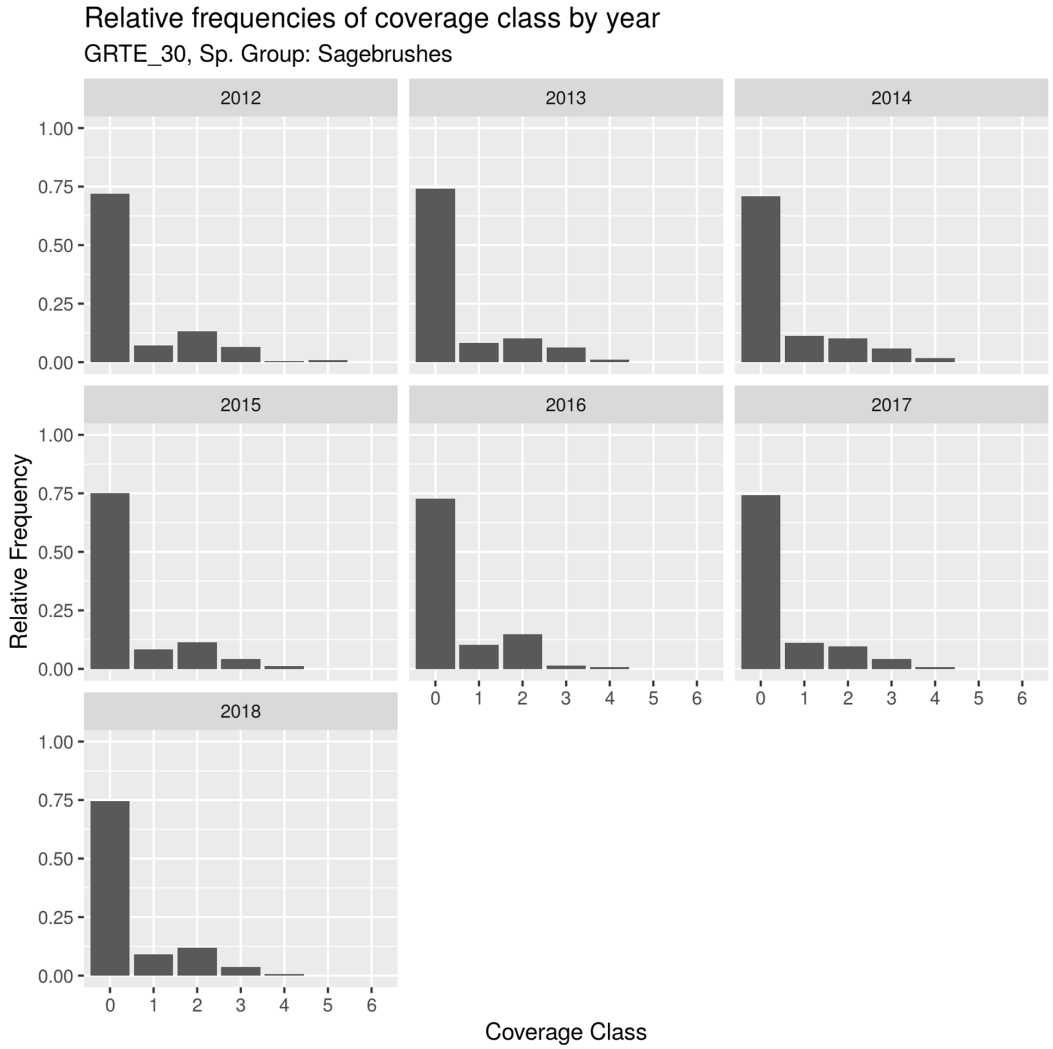

Figure 3.2: Relative frequencies of coverage classes by year for the sagebrush species group in Frame 30. This figure suggests little heterogeneity in coverage for this species group in Frame 30 over time.

coverage among the sagebrush species group over time in this frame; a similar relationship is present in the other 29 frames (plots available on the internal web-tool). Furthermore, this lack of temporal trend across frames held for six of the seven remaining species groups provided in Appendix A. The exception to this was the species group named "other" that contains bare ground, cryptobiotic soil crust, litter, moss, and rock.

The distribution of coverage classes for the "other" species group in Frame 30 over time

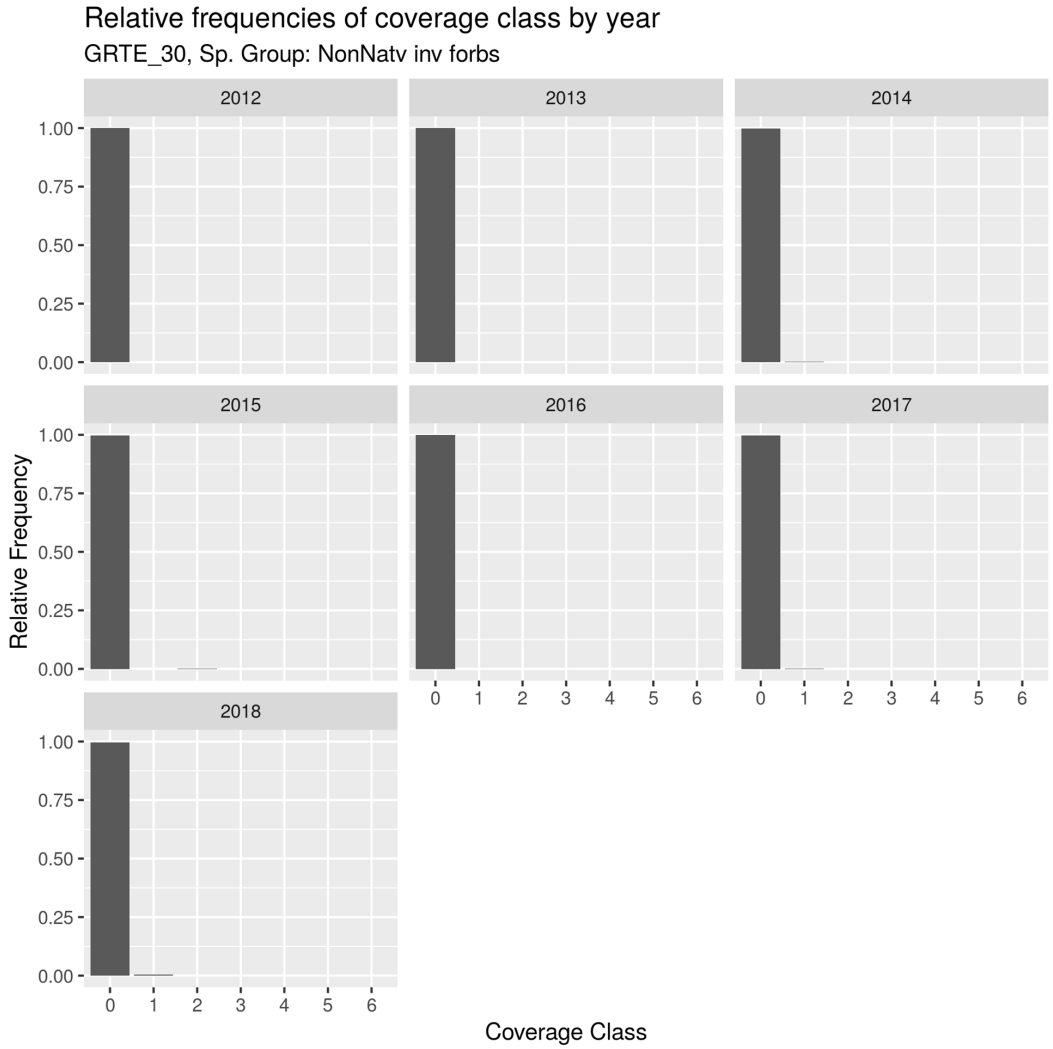

Figure 3.3: Relative frequencies of coverage classes by year for the invasive forbs group in Frame 30. Based on this figure, we see little heterogeneity in coverage over time for this species group in this frame. Furthermore, this figure highlights the low abundance of invasive forbs in this frame over time.

is illustrated in Figure 3.4. Here there is evidence of a significant shift in coverage beginning in 2016, which is ubiquitous across all frames during this time (not plotted). This large shift coincides with an adaptation in the monitoring protocol where coverage of rocks (in 2016) and moss and cryptobiotic crust (in 2017) were incorporated. Therefore, the 2016 shift in composition in the "other" species group (Figure 3.4) is likely due to the change in the

monitoring protocol, as opposed to a major shift in the biological community. As such, our analysis excluded the "other" species group. However, once an adequate data record exists, comparisons that include the "other" species group would be appropriate.

Patterns in individual species were also examined. Figure 3.5 shows the distribution of coverage class for *Artemisia tridentata* in Frame 30 over time. Again, there is little evidence of any meaningful change in coverage. This holds for most monitored species and reflects the evidence of little change present at the species group level. Additional plots to support these conclusions are available on the internal web-tool. Excluding the "other" species group, our exploratory data analysis suggested that there was little to no change in composition or abundance across the 30 frames between 2012 and 2018.

3.5.1.1 Assessing spatial patterns To visualize the spatial patterns in species composition, the relative frequency of each coverage class across the 30 frames within a year for each species and species group were plotted. Figure 3.6 provides little to no evidence of spatial heterogeneity among the sagebrush species group in 2013. A very similar relationship exists for the remaining years in the data set and this lack of trend holds for the remaining species groups.

Species level patterns were also considered and there was weak evidence of trends. As an example, the relative frequency of each coverage class across frames in 2013 for *Artemesia tridentata* are displayed in Figure 3.7. While there is more variability in this set of plots than in those depicted in Figure 3.6, there is no indication of large, systematic differences among the frames. In general, across all seven years and most species, there was little to no support for shifts in species composition and abundance among frames. Based on the exploratory analysis, there was little evidence of temporal or spatial heterogeneity in species composition. These patterns, or lack thereof, are evaluated using more formal tools in Section 3.6.

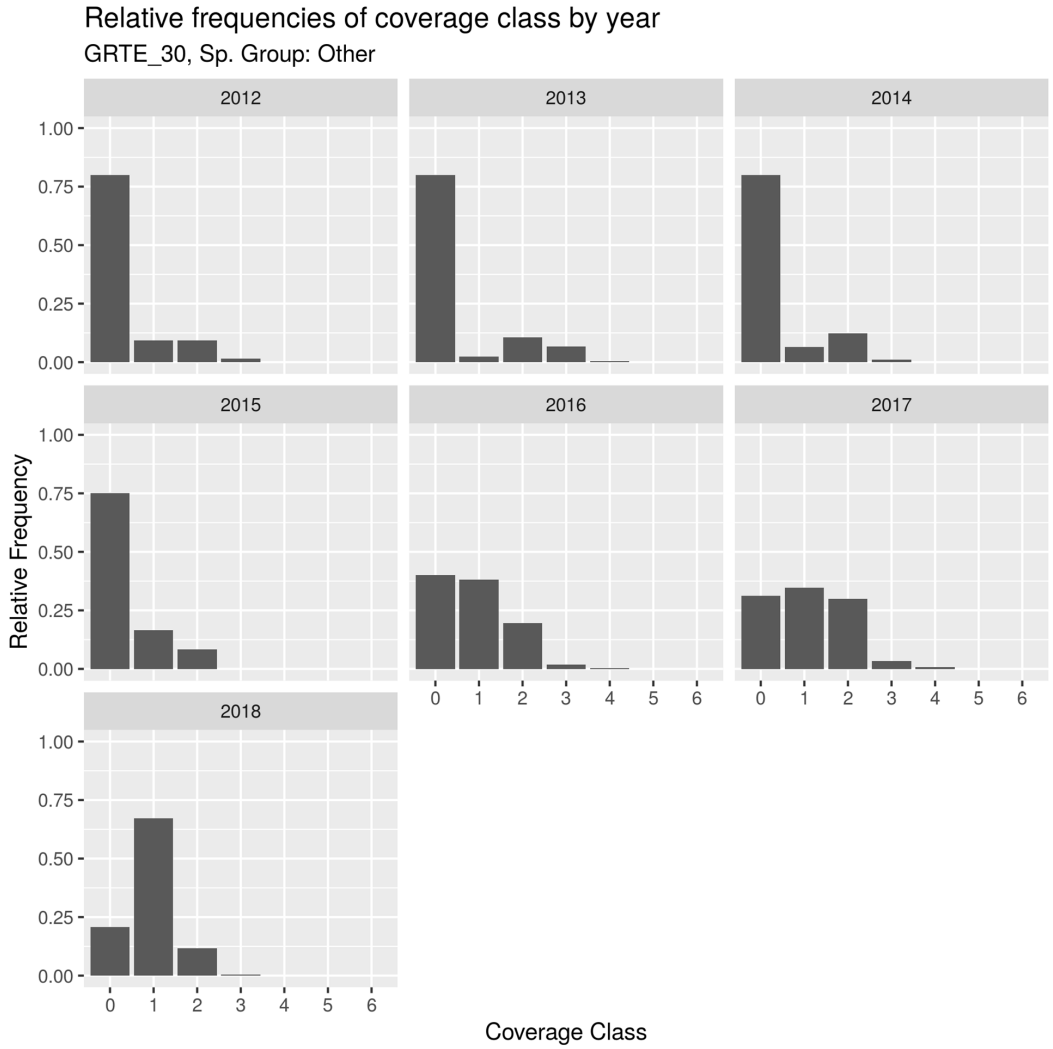

Figure 3.4: Relative frequencies of coverage classes by year for the other species group in Frame 30. This figure suggests a large shift in coverage in 2016 for this species group in Frame 30; this trend holds across all frames and coincides with the addition of rocks to the monitoring protocol, suggesting that the shift is due to a change in monitoring protocol, rather than a fundamental shift in biology.

3.5.1.2 Assessing climate and soils As part of the exploratory analysis, frames were studied for differences in climate conditions and soil water holding capacity. Soil maps for GRTE, obtained from gSSURGO, were used to determine the water holding capacity of the soil where the frames are located. Three types of soils are present (Soil Survey Staff 2019); 13

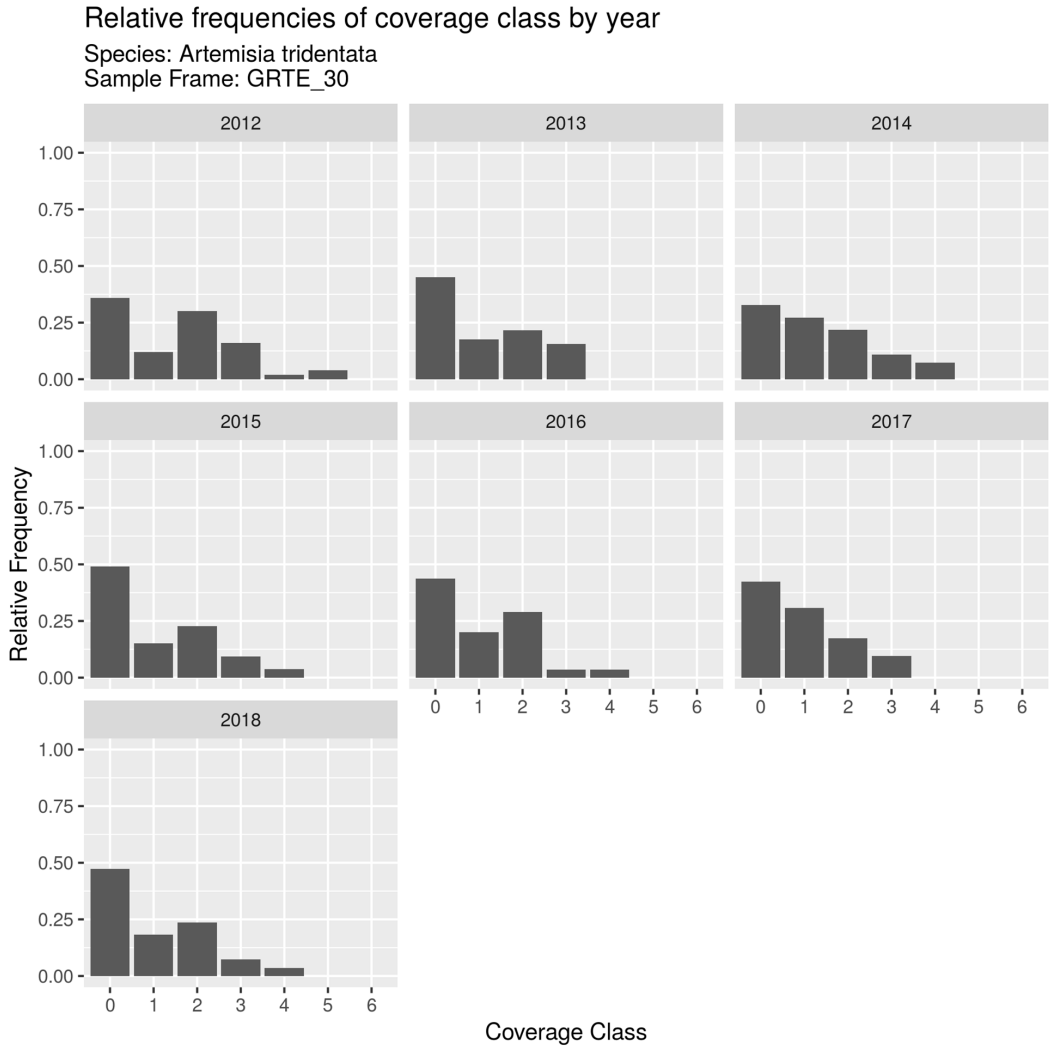

Figure 3.5: Relative frequencies of coverage classes by year for *Artemisia tridentata* in Frame 30. This figure suggests little to no change in species composition and abundance in this frame across time.

frames are located in Tineman-Bearmouth gravelly loams with 0 to 3 percent slopes, which allow for 62.2 mm water holding capacity in the top meter; 16 frames are located in Tineman gravelly loam, also with 0 to 3 percent slopes, allowing for 63.5 mm water holding capacity in the top meter; and one plot (Frame 9) is located in Tetonia-Lantonia silt loams with 0 to 3 percent slopes, which allows for 200 mm water holding capacity in the top meter.

Given that 29 of the 30 frames have gravelly soils with the same water holding

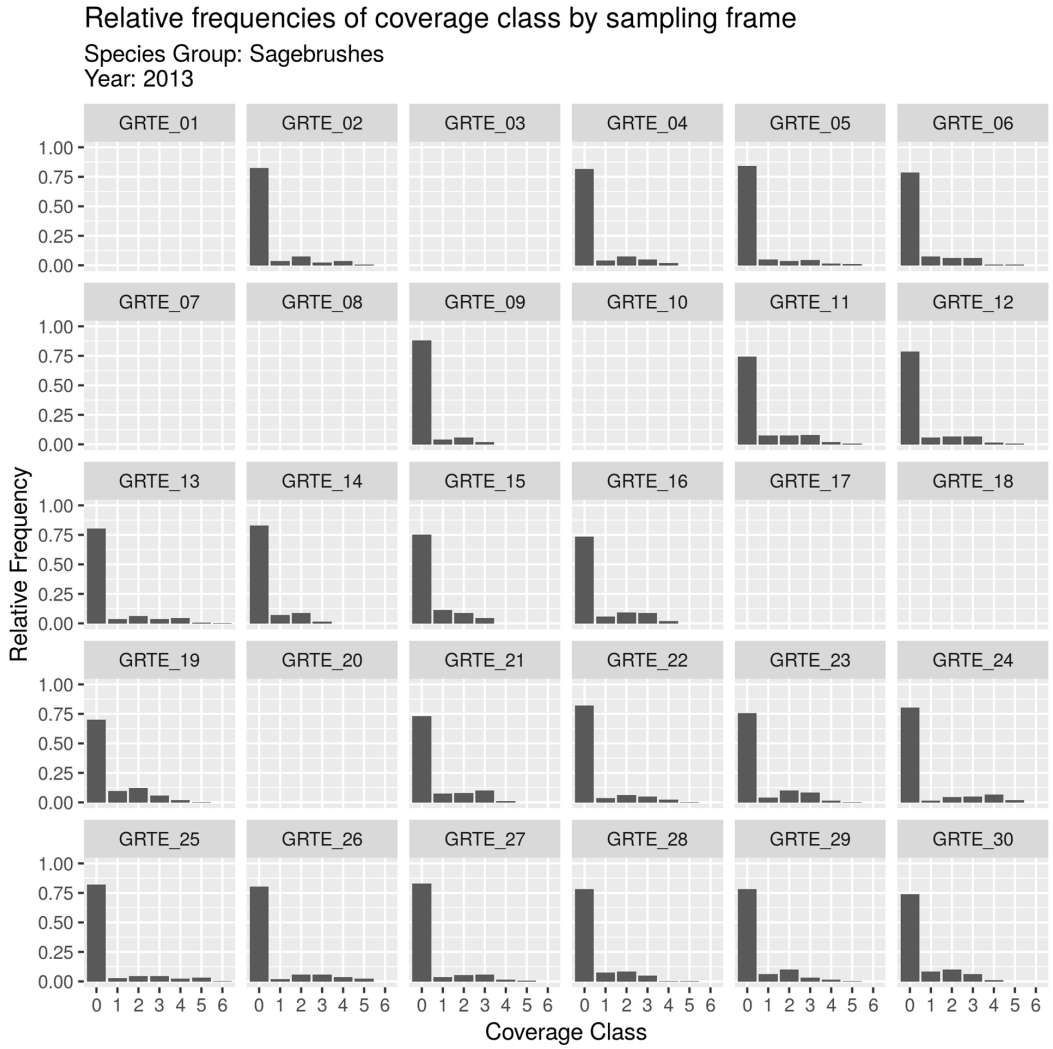

Figure 3.6: Relative frequencies of coverage classes by sampling frame for the sagebrush species group in 2013. Based on this figure, we have little to no evidence of heterogeneity in composition and abundance of the sagebrush species group across frames in 2013. Empty panels were not sampled in 2013.

capacity, the frames may only differ by climate drivers such as precipitation and temperature. However, the seasonal water balance for Frame 9 (located in Tetonia soil) is expected to be quite different due to greater water holding capacity. This soil type stores more water which is then available to plants longer into the growing season, which decreases late season drought stress. Therefore, some differences in the species composition of Frame 9 relative to the other

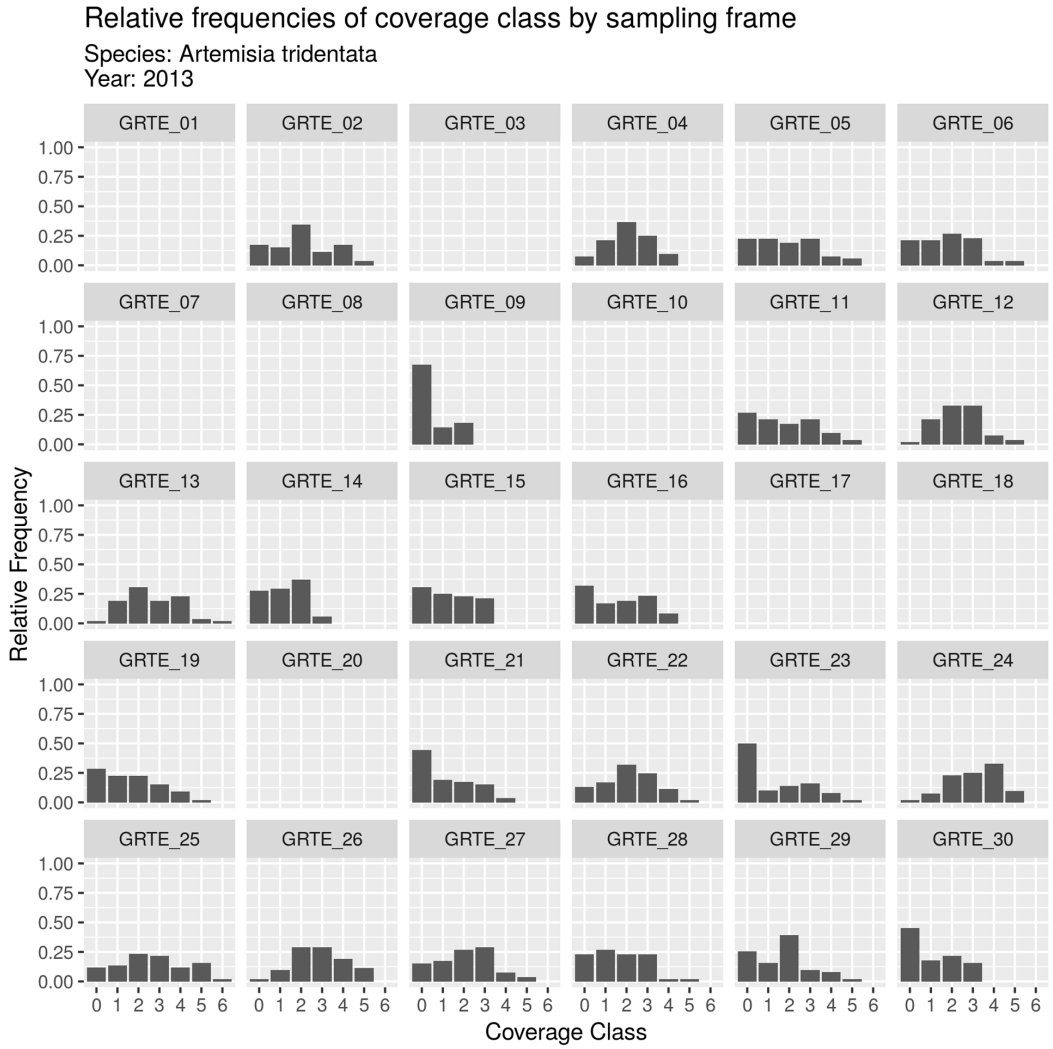

Figure 3.7: Relative frequencies of coverage classes by sampling frame for *Artemesia tridentata* in 2013. This figure suggests little difference in composition and abundance of this species across frames in 2013. There may be some evidence of a difference in Frame 9 (located in a fundamentally different soil type) relative to the other frames. Empty panels were not sampled in 2013.

frames are anticipated, particularly in dry years.

In regards to climatic conditions, park staff hypothesized that GRTE had an east-west gradient in soil moisture that may affect plant phenology. To examine this hypothesis, monthly moisture deficit and soil moisture values were extracted from a gridded water

balance model for the frame sampling locations during the summer months in years 2015 - 2018. No east-west or north-south patterns were apparent and only a very small elevation gradient was identified. Therefore, any apparent differences in species composition cannot be explained by climatic differences across frames.

3.6 Statistical methods

To formally assess the similarity of the quadrats across time and space, non-metric multi-dimensional scaling (NMDS) was implemented in a distance-based ordination framework. This ordination technique assesses both the temporal and spatial patterns in these data. To assess the temporal patterns, sampled quadrats from each frame across the years that a frame was monitored were ordinated and clusters related to year were identified. As an example, for Frame 30, which was measured all seven years, this process results in ordinating approximately 50 quadrats per year for a total of 350 quadrats. To assess spatial patterns, all quadrats measured each year were ordinated and clusters related to frame were identified.

Distance based ordination methods require specifying a distance measure to assess differences between sampling quadrats. These measures are discussed in Section 3.6.1. In Section 3.6.2, the ordination process and the advantages of NMDS in modeling these data are discussed.

3.6.1 Discussion of dissimilarity measures

A number of different distance measures including Bray-Curtis, Gower, Kendall's tie-adjusted tau, Goodman and Kruskal's gamma, and relative rank difference were originally considered for this analysis; the latter three of these distance measures being specifically tailored for ordinal data. To assess how well each measure differentiated between similar and dissimilar quadrats, a simulation study was conducted in which canopy coverages

for 10 species across three hypothetical quadrats were generated from two different beta distributions. Quadrats 1 and 2 were simulated from the same beta distribution (beta($\mu = 1/10$, $\delta = 1/10$)) and represent similar communities. Quadrat 3 was generated from a very different beta distribution (beta($\mu = 3/4$, $\delta = 3/4$)) and represents a different community from quadrats 1 and 2. These simulated coverages were then converted to Daubenmire coverage classes based on Table 2. Finally, to assess how well the measures differentiate similar and dissimilar quadrats, pairwise dissimilarity between quadrats 1 and 2 and 1 and 3 respectively were calculated.

The results of this simulation are provided in Figure 3.8. Based on this figure, the Bray-Curtis, Gower, and relative rank difference measures appear to best differentiate between similar and dissimilar quadrats. Furthermore, Bray-Curtis seems to outperform Gower and relative rank difference, which perform identically as they are equivalent with ordinal data. These results are consistent with Ricotta and Feoli (2013), who showed that there is no need to apply dissimilarity coefficients specifically designed for ordinal scales to coverage class data.

To choose between the Bray-Curtis and Gower distance measures, their performance was further investigated under varying degrees of zero-inflated responses in another simulation study. This simulation replicated the approach of the previous simulation, but included varying proportions of zeros in each quadrat and increased the number of species to 80. The results are provided in Figure 3.9, which suggest that the Bray-Curtis dissimilarity index better differentiates similar and dissimilar quadrats in the presence of many zeros, as is the case with the data used in this analysis. Therefore, the Bray-Curtis distance measure was chosen to assess the dissimilarity between quadrats in this analysis.

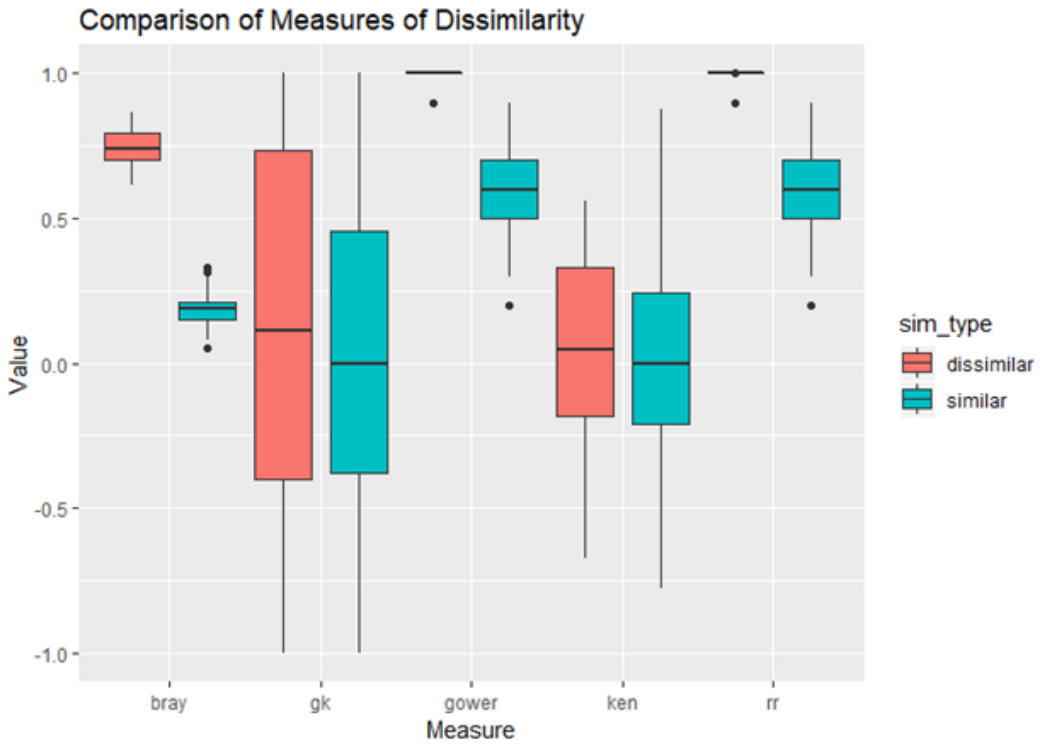

Figure 3.8: Results from dissimilarity index simulation study. Based on this figure, the Bray-Curtis, Gower, and relative rank difference measures tend to best preserve the original similarity or dissimilarity in species abundance and coverage. The Gower and relative rank indices are identical in this case, as the response is ordinal.

3.6.2 Ordination techniques

Distance measures like Bray-Curtis allow for the calculation of pairwise dissimilarity in species composition within a set of quadrats of interest. These pairwise dissimilarities are used to build a dissimilarity matrix, whose entries represent the dissimilarity between each distinct pair of quadrats. To visualize these dissimilarities, the distance matrix is often projected into two dimensions to create an ordination plot. This plot attempts to preserve the original dissimilarity contained in the matrix, though some is lost through the projection process.

To project the dissimilarity matrices in this analysis, non-metric multi-dimensional scaling (NMDS) was implemented through the `vegan` package in R. NMDS is a flexible tool

Figure 3.9: Results of simulation study comparing the Bray-Curtis index to the Gower index. Based on this figure, Bray-Curtis tends to better preserve the original dissimilarity in the presence of many zero coverages, as is the case with these data.

that can handle non-normal data types, as it is a rank based ordination. Furthermore, it does not make any assumptions of linearity. Consequently, it is a popular choice for handling ordinal community data (Oksanen et al. 2020), in which these features are common. To assess how well the lower dimensional projection preserves the high-dimensional dissimilarities, a statistic known as stress is calculated. This statistic is calculated by first determining the Euclidean distances between the points in the ordination plot, then these values are regressed against the original dissimilarities observed in the data. Stress is then calculated as $1 - R^2$, where R^2 is the coefficient of determination for this regression. In general, stress values less

than 0.05 are deemed very good fits, while values between .05 and .10 are considered good, values between .10 and .20 are considered fair, and values greater than .20 are considered poor (Kruskal 1964).

To address the questions of interest in this analysis, multiple dissimilarity matrices were required. To obtain these matrices, the data were filtered to contain only the quadrats of interest, which varied depending on whether the interest was in spatial or temporal patterns. For example, to examine the temporal patterns in Frame 30, the data were filtered to contain only the 350 quadrats from Frame 30 over the course of this study. Similarly, to examine spatial patterns in 2013, the data were restricted to the 1133 quadrats sampled across 22 frames in 2013. Using these quadrats, distance matrices were created based on the Bray-Curtis dissimilarity index using the `vegan` package in R (Oksanen et al. 2020).

These matrices were created for each frame over time and for each year across frames. To assess the temporal and spatial patterns in species composition over the course of this study, these matrices were projected into ordination plots.

3.7 Results

The ordination results reflected the exploratory data analysis in that there was little evidence of differences in species composition both across time within each frame and across frames within each year. In the subsections below, these results are discussed in greater detail.

3.7.1 Temporal ordination

To determine whether there was evidence of temporal shifts in species composition or abundance, dissimilarity matrices were created for each frame over time and projected into 30 different ordination plots. The stress statistics for these ordinations varied from 0.21 to 0.30, suggesting relatively poor preservation of the original dissimilarity. However, lower

stress values required higher dimensional projections which are not easily visualized, and so the 2D projections were used. As an example, the ordination plot for Frame 30 is provided in Figure 3.10. In this figure, there is no clear separation between any of the points, suggesting that there is little evidence of dissimilarity across any of the quadrats in Frame 30 over time. The remaining temporal ordination plots look similar, also suggesting a lack of dissimilarity over time. These plots are available on the internal web-tool.

These results suggest that there is negligible heterogeneity in species composition or abundance within the frames over time as expected. Given the short monitoring period (7 years), we would not expect large shifts in species composition or abundance without significant disturbance (such as fire), of which none occurred between 2012 and 2018.

3.7.2 Spatial ordination

To evaluate the evidence of spatial variation in species composition or abundance within each year across frames, dissimilarity matrices based on all quadrats measured each year were created and projected into seven different ordination plots (stress values varied from .22 to .25). The ordination plot for 2013 is provided in Figure 3.11 as an example. Again, there is no clear separation in the points based on frames in this plot. Frame 9 appears to lurk on the periphery, which was expected based on the soil hydrology analysis discussed in Section 3.5.1.2 Nonetheless, the separation between the quadrats from Frame 9 and the remaining frames is not particularly extreme. This relationship held true for the remaining six years.

Based on the seven comparisons by year discussed above, there did not appear to be convincing evidence of heterogeneity in species composition across the frames within each year. However, when quadrats are aggregated across frame and sorted by the underlying sagebrush community type (shrubland or bitterbrush), there is some evidence of differences in species composition based on the separation between the two community types. Figure 3.12 displays this relationship for the 1133 quadrats sampled in 2013, and is characteristic of

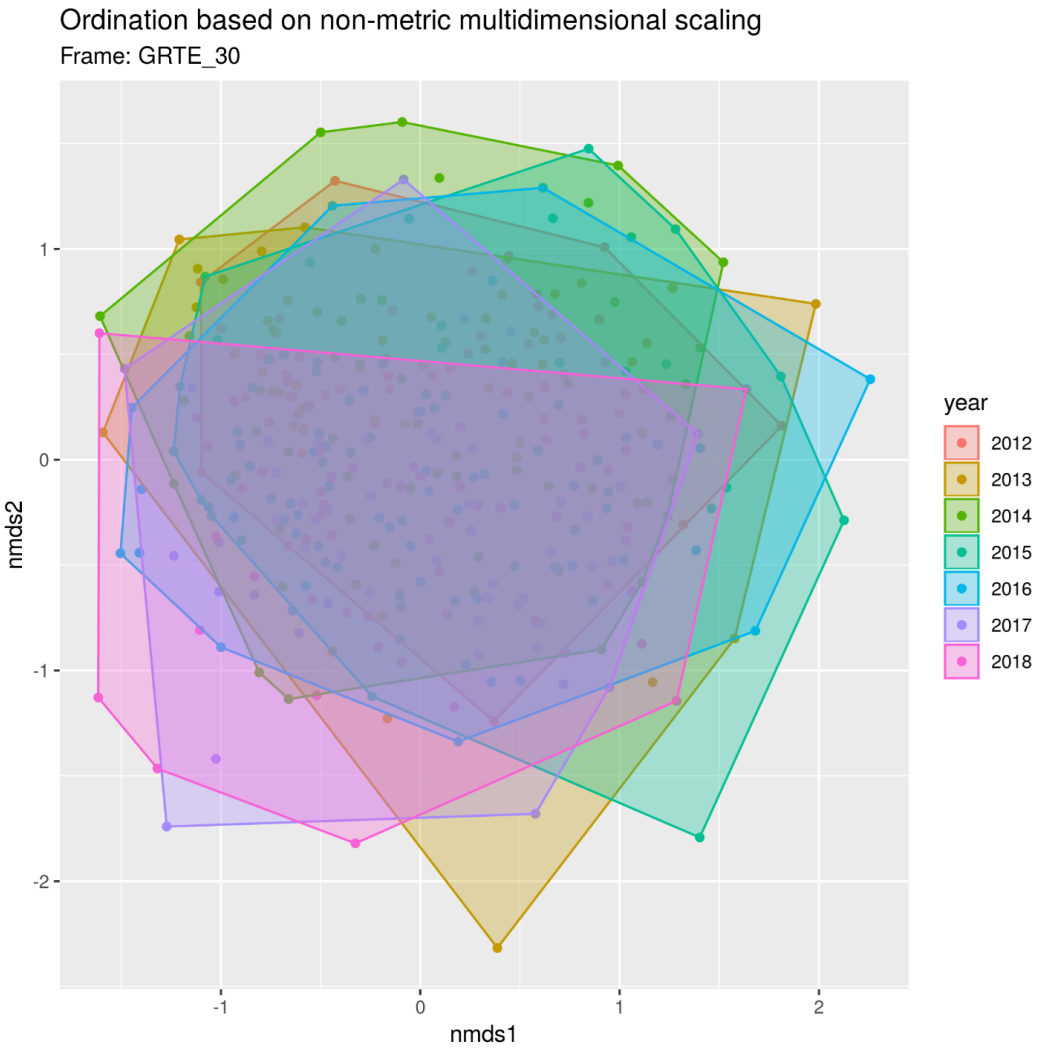

Figure 3.10: Ordination plot for Frame 30 over time. Based on the overlap between quadrats across years, there is little evidence of heterogeneity in coverage over time in Frame 30.

the other six years. These results suggest that there is not much heterogeneity in coverage class across the frames within each year. However, when quadrats are aggregated across frames and compared by the underlying sagebrush community type, there does appear to be some evidence of differences in composition and abundance.

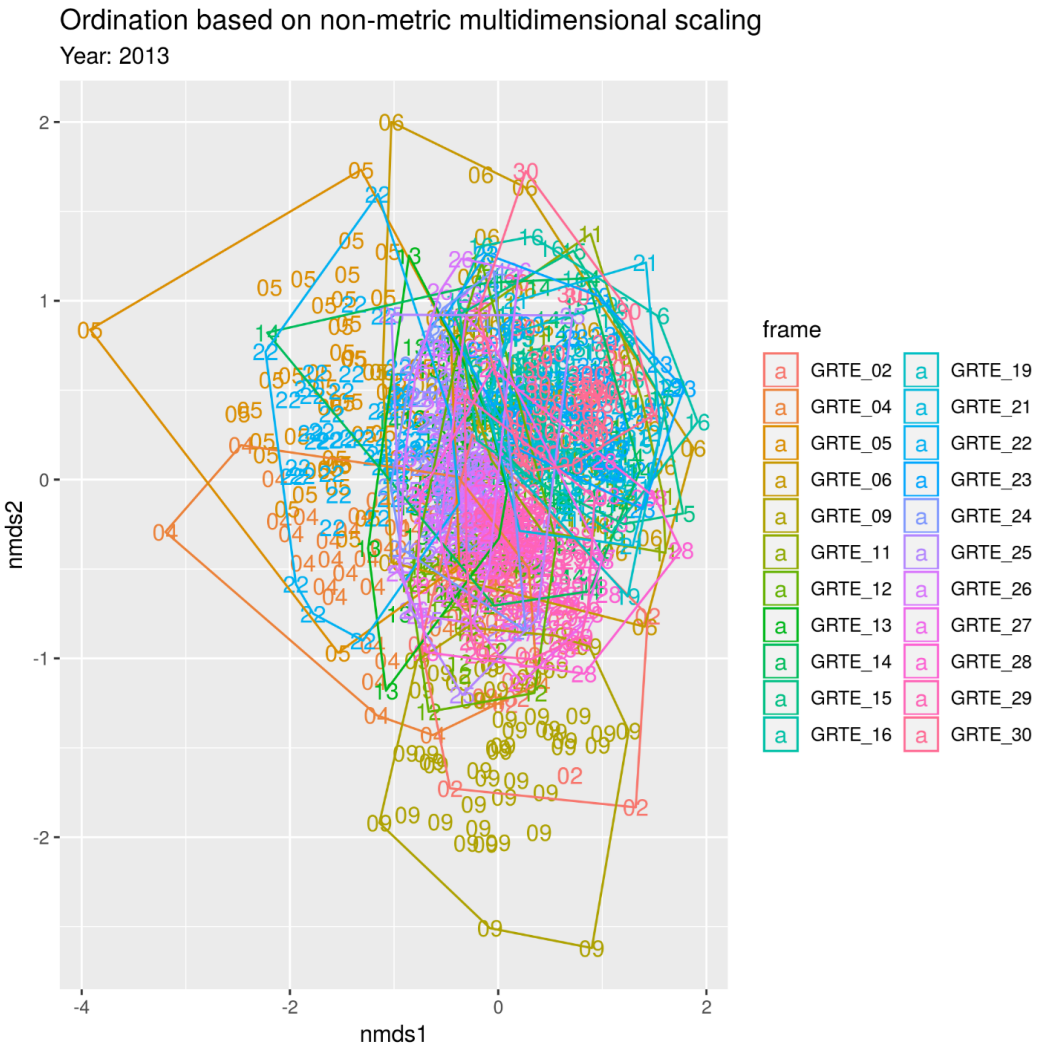

Figure 3.11: Spatial ordination plot for 2013 by frame. Based on the high degree of overlap across frames, there is little evidence of heterogeneity in coverage across frames in 2013. The one exception to this statement being Frame 9, whose quadrats lie on the periphery of the figure; this was expected due to the fundamentally different soil type on which Frame 9 is located.

3.8 Discussion

This analysis assessed the status of two sagebrush steppe types (sagebrush-bitterbrush and sagebrush dry shrubland) in GRTE with respect to temporal and spatial patterns in species composition and abundance. Based on the results of this analysis, there is little to no

Figure 3.12: Spatial ordination plot for 2013 by sagebrush community type. Based on the slight separation between the quadrats from each of these community types, there is some evidence to suggest that the coverage tends to differ between the two.

evidence to suggest meaningful shifts in the vegetation community in GRTE over time, which is consistent with the relatively narrow monitoring period (7 years) and lack of disturbances. Furthermore, there is little evidence to suggest substantial differences in species composition and abundance among the frames. However, there is some evidence to suggest differences in community composition between the underlying sagebrush community types.

Therefore, based on this assessment, it appears reasonable to reduce the sampling effort

in this monitoring protocol. With little evidence of differences in composition and abundance over both time and space, adequate monitoring of the park sagebrush communities can be accomplished across fewer frames. However, we do recommend sampling frames in both underlying community types each year, as there is some evidence of differences in community composition between the two sagebrush community types. Fortunately, the priority frames in the original sampling design were chosen to be representative of each of these community types (Yeo and Rodhouse 2013). Continued sampling of the five annual priority frames is suggested each year to ensure adequate coverage of both community types and to preserve annual data collection on those five frames.

To sample the remaining 25 frames, we propose cycling five frames each year in a panel design (Table 3.3) for a total of 10 frames sampled per year. Ten frames can reasonably be monitored by two crews of field technicians in five days. The remaining 25 frames were assigned to a panel based on when they were last sampled. Frames last sampled in 2017 were assigned to the first two panels, while frames last sampled in 2018 or 2019 were assigned to the last three panels. This design guarantees every frame is visited in a five-year window. Based on the lack of temporal patterns found in these data, this should be sufficient to detect any shifts that may occur. Note that the sampling schedule for 2019 reflects the original schedule provided in Table 3.1. The fieldwork for the 2019 field season has been devised to complete the third cycle of the original monitoring schedule, thereby maintaining a balanced sampling design. The proposed monitoring schedule that reflects this analysis is scheduled to begin in 2020. It is worth noting that these suggestions are contingent on the current status quo. In the event that any disturbances or changes in climatic conditions arise, this sampling frequency should be revaluated.

Plot	2017	2018	2019	2020	2021	2022	2023	2024
1	3	-	-	4	-	-	-	-
2	-	3	-	-	-	4	-	-
3	3	-	-	4	-	-	-	-
4	-	-	3	-	-	-	4	-
5	-	-	3	-	-	-	-	4
6	-	3	-	-	-	4	-	-
7	3	-	-	4	-	-	-	-
8	3	-	-	-	4	-	-	-
9	-	3	-	-	-	4	-	-
10	3	-	-	4	-	-	-	-
11	-	-	3	-	-	-	-	4
12	-	3	-	-	4	-	-	-
13	-	-	3	-	-	-	-	4
14	E	E	E	E	E	E	E	E
15	-	-	3	-	-	-	-	4
16	-	-	3	-	-	-	4	-
17	3	-	-	-	4	-	-	-
18	3	-	-	-	4	-	-	-
19	-	-	3	-	-	-	-	4
20	3	-	-	-	4	-	-	-
21	3	-	-	4	-	-	-	-
22	E	E	E	E	E	E	E	E
23	-	3	-	-	-	4	-	-
24	-	3	-	-	-	-	4	-
25	E	E	E	E	E	E	E	E
26	-	3	-	-	-	4	-	-
27	-	-	3	-	-	-	4	-
28	-	-	3	-	-	-	4	-
29	E	E	E	E	E	E	E	E
30	E	E	E	E	E	E	E	E

Table 3.3: Proposed monitoring schedule. The numbers denote how many times the frame has been sampled since the start of the study; frames labeled "E" are sampled every year. The schedule for 2019 reflects the original protocol to maintain balance.

Acknowledgements

We thank the numerous biological technicians for their field data collection efforts. The time and effort of Christian Stratton and Dr. Andrew Hoegh was made possible with a CESU

agreement (G17AC00147) between USGS Northern Rocky Mountain Science Center and Montana State University. Funding from agreement P18PG00390 with the National Park Service allowed for Dr. Kathi Irvine to participate in this collaborative project.

CHAPTER FOUR

CLUSTERING AND UNCONSTRAINED ORDINATION WITH DIRICHLET PROCESS MIXTURE MODELS

4.1 Contribution of Authors and Co-Authors

Manuscript in Chapter 4

Author: Christian Stratton

Contributions: Responsible for writing of manuscript and submission, authored code to fit and summarize models, led data analysis, contributed to method framework development.

Author: Andrew Hoegh

Contributions: Provided general guidance and feedback on manuscript draft, contributed to code to summarize and visualize fitted models, contributed to method development.

Author: Kathryn M. Irvine

Contributions: Provided general guidance and feedback on manuscript draft, method framework development.

Author: Thomas J. Rodhouse

Contributions: Provided interpretation of ecological results and general guidance and feedback on manuscript draft.

Author: Jennifer L. Green

Contributions: Provided general guidance and feedback on manuscript draft, contributed to method development.

Author: Katharine M. Banner

Contributions: Provided general guidance and feedback on manuscript draft, method development.

4.2 Manuscript Information

Christian Stratton, Andrew Hoegh, Kathryn M. Irvine, Thomas J. Rodhouse, Jennifer L. Green, Katharine M. Banner

Computational Statistics & Data Analysis

Status of Manuscript:
_X_ Prepared for submission to a peer-reviewed journal
____ Officially submitted to a peer-reviewed journal
____ Accepted by a peer-reviewed journal
____ Published in a peer-reviewed journal

Elsevier

Abstract

Assessment of similarity in species composition or abundance across sampled locations is a common goal in multi-species monitoring programs. Existing ordination techniques provide a framework for clustering sample locations based on species composition by projecting high-dimensional community data into a low-dimensional, latent ecological gradient representing species composition. However, these techniques require specification of the number of distinct ecological communities present in the latent space, which can be difficult to determine *a priori*. We develop a hierarchical ordination model capable of simultaneous clustering and ordination that allows for estimation of the number of clusters present in the latent ecological gradient. This model draws latent coordinates for each sample location from a Dirichlet process mixture model, affording researchers with probabilistic statements about the number of clusters present in the latent ecological gradient. Additionally, the model is extended to accommodate hierarchical sampling designs, providing ordination results that are aligned with primary sampling units. This model is applied to two empirical data sets; the first data set concerns presence-absence records of fish in the Doubs river in Eastern France, and the second data set describes presence-absence records of plant species in Craters of the Moon National Monument and Preserve (CRMO) in Idaho, USA. Code to fit the model using NIMBLE is provided in Appendix B. Application of the Dirichlet process ordination model to the Doubs river data provided evidence of two distinct ecological communities corresponding to upstream and downstream sample locations, consistent with previous analyses. Application of the model to the CRMO data provided evidence of four ecological regions in the latent space, corresponding to various features of the ecological gradient in CRMO, including elevation and proximity to volcanic features. Development of the Dirichlet process ordination model provides ecologists and wildlife managers with data-driven inferences about the number of distinct ecological communities present across monitored

locations. This information can be leveraged to develop more cost-effective monitoring strategies, supporting reliable decision-making for wildlife and conservation management.

4.3 Introduction

Long-term monitoring programs have recently turned their focus from historic single-species monitoring towards monitoring entire species assemblages (Choi et al. 2019; Damgaard, Hansen, and Hui 2020; Inoue, Stoeckl, and Geist 2017; Zurell et al. 2020). To better understand the impacts of climate change, disease, wildfires, and other emerging threats, many of these monitoring programs seek to assess similarity in species composition among sample locations across the monitored landscape or assess change in species composition over time (Jean, Stella, and Daley 2014; Nicolli 2019; Stratton et al. 2019). To make these assessments, multivariate abundance data are typically collected at multiple locations within the monitored region. These abundance data can then be used to identify trends in species composition or cluster sample locations based on species composition. There exist many different techniques for making inferences about high-dimensional community data, including generalized linear mixed models, species distribution models, and ordination techniques (Pollock et al. 2014; Veen et al. 2021). This research focuses on unconstrained ordination methods as a tool for assessing similarity in species composition among sample locations as it allows researchers to cluster sample locations based on species composition and visualize results in a low-dimsional space, does not require environmental covariates, and is widely implemented by ecologists and wildlife managers (Hui et al. 2015).

Unconstrained ordination may be used to assess similarity between sample locations based on species composition without environmental covariates (Jain and Dubes 1988, ch. 8). Traditional unconstrained ordination techniques rely on first constructing a pairwise dissimilarity matrix between sites based on some distance measure, e.g., Bray-Curtis dissimilarity (Bray and Curtis 1957), then projecting that matrix into a lower-dimensional space meant to represent the underlying dominant ecological gradient using some algorithm-based projection technique (Braak and Prentice 1988). Common choices of projection

techniques include classical multidimensional scaling (MDS, Kruskal and Wish 1978) or non-metric multidimensional scaling (NMDS, Kruskal 1964). In the lower-dimensional space, algorithmic clustering techniques, e.g., k-means clustering (Macqueen 1967), are applied to cluster sites based on proximity to one another in the latent ecological gradient. These algorithm-based ordination techniques can best be described as a two-step process. First, the community matrix is projected into a lower dimensional space. Then, sample locations are clustered in the lower dimensional space.

While algorithm-based ordination and clustering techniques are easily implemented and can provide ecologists with interpretable results, they present challenges when conducting inference. One such challenge is the inability to account for hierarchical sampling designs that are common when monitoring vegetation, where ordination is often the objective (Esposito and Rodhouse 2015; Jean, Stella, and Daley 2014; Nicolli 2019; Stratton et al. 2019). In many of these monitoring programs, vegetation assemblage data is collected by randomly selecting secondary sampling units (e.g., one square meter quadrats) from larger primary sampling units, which are selected based on known ecological gradients (e.g., elevation). Most algorithm-based ordination techniques, including PCA and NMDS, do not provide any way to account for the expected similarity in responses within a primary sampling unit, nor do they allow for ordination of the primary sampling units without first aggregating responses across secondary sampling units. Additional challenges arise when clustering ordination results, as it is usually impossible to determine the number of clusters present in the latent ecological gradient *a priori*, yet this value is required for most algorithm-based clustering techniques. And finally, the algorithm-based framework does not provide a likelihood. The lack of a likelihood precludes formal assessments of distance measures, projection methods, clustering techniques, or numbers of clusters present in the latent space. Consequently, these algorithm-based ordination and clustering techniques are often used as exploratory tools.

Recently, model-based approaches for unconstrained ordination have been explored that

allow for measures of uncertainty on parameter estimates (Hui et al. 2015; Veen et al. 2021). These techniques use latent variables to represent sample locations along an underlying ecological gradient. Hui et al. (2015) describe a number of advantages of model-based ordination, including: 1) the ability to account for spurious data properties; 2) access to model assessments via residuals analyses; and 3) the ability to compare models formally and quantify uncertainty in findings. The first of these three advantages allows for appropriate modeling of mean-variance relationships that are common in ecological data; failing to account for these relationships can result in misleading ordination results (Warton, Wright, and Wang 2012). The second and third advantages are a result of representing the data with a probabilistic model, affording likelihood-based comparisons and uncertainty on parameter estimates. While these model-based ordination techniques provide many advantages over the algorithmic ordination techniques, they still do not provide a model-based approach to cluster sample locations in the latent ecological gradient, or any way to determine the number of clusters of sites present in that space.

Following Hui et al. (2015), Hui (2016) developed a hierarchical latent variable model that allows for simultaneous Clustering and Ordination of ecological Abundance data (known as the CORAL model) by drawing the latent variables representing the positions of sample locations in the latent ecological gradient from a finite mixture distribution. Through this process, the authors were able to leverage the joint information available for dimension reduction and clustering, and in doing so, outperform competing algorithm-based clustering and ordination strategies (Hui 2016). Additionally, the model allows for probabilistic statements about cluster membership across sites, allowing ecologists to qualify impacts of management actions with appropriate uncertainty on cluster assignments. However, the CORAL model does not afford probabilistic statements about the number of clusters present in the latent ecological gradient, as the latent coordinates of each site along the underlying gradient are drawn from a finite mixture model, requiring prior specification of the number of

groups present. Instead, prior information and/or model selection tools (such as information criterion) are required to select the number of clusters present, prohibiting probabilistic statements about the number of clusters in the latent space.

Motivated by two empirical data sets, we propose a hierarchical ordination model, hereafter the "Dirichlet process ordination model" or DPORD, that simultaneously performs clustering and ordination of species assemblage data and is capable of estimating the number of groups that are present in the latent ecological gradient with appropriate measures of uncertainty. This model draws the coordinates of each sample location in the latent ecological gradient from a Dirichlet process mixture model, thereby affording probabilistic statements about the number of groups present in the latent space. Additionally, we extend the DPORD model to accommodate hierarchical sampling designs by incorporating random effects for each secondary sampling unit. The DPORD model is fit to two empirical data sets using Markov chain Monte Carlo (MCMC) methods and code is provided in the appendix. For comparative purposes, the first data set considered is the Doubs river data set analyzed by Hui (2016); this data set describes detection or non-detection of 27 species of fish along the Doubs river in eastern France. The second data set considered describes detection or non-detection of 25 plant species collected from sagebrush steppe ecosystems in Craters of the Moon National Monument and Preserve (CRMO) in Idaho, USA in 2019.

4.4 Methods

4.4.1 Dirichlet process ordination model

The DPORD model builds upon the CORAL model proposed by Hui (2016). Both models are designed to fit multivariate community data, which are typically represented by an $n \times s$ community matrix, where n represents the total number of sample locations (sites) and s represents the number of observed species across all sample locations. Entries in the community matrix represent measures of species abundance, typically as binary presence-

absence responses (truncated abundance), counts, or ordinal classes.

Similar to Hui (2016), the DPORD model represents abundance data with a two-stage hierarchy. The first stage of the hierarchy relates the mean response to d latent variables through a generalized linear model framework:

$$y_{ij}|\mu_{ij}, \phi_j \sim f(y_{ij}|\mu_{ij}, \phi_j), \quad g(\mu_{ij}) = \alpha_i + \beta_j + \boldsymbol{z}_i\boldsymbol{\theta}_j^T, \tag{4.1}$$

where y_{ij} is the observed response for species j from site i, $f()$ represents the assumed distribution for response y_{ij}, μ_{ij} represents the mean response for species j from site i, ϕ_j represents the dispersion parameter for species j, and $g()$ represents an appropriate link function. Within the link function, α_i represents a site effect that accounts for overall differences in abundance across sites, β_j represents a species effect that accounts for overall differences across species, $\boldsymbol{z}_i$ represents the $1 \times d$ row-vector of latent coordinates describing the location of site i in the latent ecological ecological gradient, and $\boldsymbol{\theta}_j$ represents the $1 \times d$ row-vector of species-specific coefficients.

We extend the Hui (2016) model to accommodate hierarchical sampling designs by including secondary sampling unit level random effects. Let i index the primary sampling unit, j index the species, and k index the secondary sampling unit within the primary sampling unit. The hierarchical formulation of the DPORD model is as follows:

$$y_{ijk}|\mu_{ijk}, \phi_j \sim f(y_{ijk}|\mu_{ijk}, \phi_j), \quad g(\mu_{ijk}) = \alpha_i + \beta_j + \boldsymbol{z}_i\boldsymbol{\theta}_j^T + \gamma_{ik}, \tag{4.2}$$

where $\gamma_{ik} \sim N(0, \sigma_i^2)$. By adding random effects for each secondary sampling unit to the linear predictor, the DPORD model is able to account for the hierarchical sampling design and align the ordination with the primary sampling unit. Consequently, the resulting ordination can be used to visualize the location of the primary sampling unit in the latent space and cluster primary sampling units based on their proximity to one another in the

latent space.

For both versions of the DPORD model, the second stage of the hierarchy draws the latent vectors from a Dirichlet process mixture of multivariate normal distributions:

$$\begin{aligned} \boldsymbol{z}_i|c_i, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_n &\sim \mathcal{N}(\boldsymbol{\mu}_{c_i}, \boldsymbol{\Sigma}_{c_i}), \\ \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_n &\sim G, \\ G|\delta, G_0 &\sim DP(\delta, G_0), \end{aligned} \tag{4.3}$$

where c_i denotes the latent cluster label associated with site i, δ represents the concentration parameter for the Dirichlet process, and G_0 represents the base distribution; in the context of ordination, δ affects the prior probability of forming new clusters in the latent ecological gradient and G_0 is typically chosen to be a multivariate normal distribution with mean $\boldsymbol{\mu}_0$ and covariance matrix $\boldsymbol{\Sigma}_0$. We note the definition of a "site" depends on the sampling model. In the case of equation 4.1, the latent vectors are indexed by total number of sampling locations. In the case of equation 4.2, the latent vectors are indexed by the total number of primary sampling units.

Similar to other latent variable techniques for ordination, the latent vectors $\boldsymbol{z}_i$ can be thought of as coordinates describing each sample location's position along the d-dimensional underlying ecological gradient. In the presence of a hierarchical sampling design, this vector instead represents the location of the primary sampling unit along the underlying ecological gradient. If the site-specific intercepts, α_i, are omitted, this space represents relative species abundance; if the site-specific intercepts are included, the model is adjusted for site-specific abundance and this space represents species composition (Hui 2016). Posterior mean estimates of the latent vectors associated with each sample location may be plotted to visualize patterns in species composition or relative abundance across sample locations. By drawing these latent vectors from an infinite mixture of normal distributions, sample

locations are clustered into groups of similar composition based on proximity to one another in the underlying ecological gradient. Additionally, the use of an infinite mixture allows the data to inform the number of clusters present in the ecological gradient, rather than specifying that quantity *a priori*.

As with all latent variable models, constraints must be placed on the elements of $\boldsymbol{\theta}$ ($s \times d$) and $\boldsymbol{\Sigma}$ ($d \times d$) to ensure parameter identifiability. For example, to avoid scale and rotation invariance, $\boldsymbol{\Sigma}$ is set to a d-dimensional identity matrix, the upper diagonal elements of $\boldsymbol{\theta}$ are set to zero, and the diagonal elements of $\boldsymbol{\theta}$ must be positive (Hui 2016; Skrondal and Rabe-Hesketh 2004). Notably, the constraints placed on the cluster-specific covariance matrices do not prevent the DPORD model from accounting for correlation between species. Hui (2016) shows that, conditional on the cluster, the distribution of the linear predictor at site i is multivariate normal with covariance matrix $\boldsymbol{\theta\Sigma\theta}'$. Imposing the constraint that $\boldsymbol{\Sigma}$ be an identity matrix results in a covariance matrix of $\boldsymbol{\theta\theta}'$, which can take any form and flexibly model correlation between species.

4.4.2 Data analyses

4.4.2.1 Daubs river data

The first data set considered describes detection or non-detection of 27 species of fish at 30 sample locations along the Doubs river in eastern France; these data are available in the ade4 R package (Dray and Dufour 2007). Previous analysis of the Doubs river data with the CORAL model found evidence of two distinct ecological communities, partly corresponding to spatial separation in upstream and downstream sample locations (Hui 2016). That analysis also suggested that there was comparable support for three distinct communities, with analyses resulting in Deviance Information Criterion (DIC_c) values (Spiegelhalter et al. 2002) of 221.51 and 223.02 for two or three clusters in the latent space, respectively. These data are again analyzed with the DPORD model to allow comparison and explore performance when the number of clusters in the latent space is

known *a priori*. Following Hui (2016), site eight was removed prior to analysis, as no species were recorded at that sample location; a map of sample locations is provided in Figure 4.1.

When fitting the DPORD model, prior distributions must be specified for the site-specific intercepts α_i, species-specific intercepts β_i, species-specific coefficients $\boldsymbol{\theta}_j$, concentration parameter of the Dirichlet process α_0, and the mean of the base distribution of the Dirichlet process $\boldsymbol{\mu}_0$. Independent standard normal prior distributions were placed on all site-specific intercepts, species-specific intercepts, and unconstrained elements of the species-specific coefficients matrix; standard half-normal distributions were placed on the diagonal elements of the species-specific coefficients matrix. A standard multivariate normal prior distribution was placed on the mean of the base distribution of the Dirichlet process. A gamma hyper-prior distribution with shape parameter of one and rate parameter of two was placed on the Dirichlet process concentration parameter; choosing the rate parameter to be large relative to the shape parameter encourages clustering in Dirichlet process mixture models (Frühwirth-Schnatter and Malsiner-Walli 2018). In this case, this hyper-prior distribution implies with high prior probability that there exists fewer than six groups in the latent ecological gradient (Antoniak 1974). This hyper-prior was chosen to reflect previous analyses of the Doubs River data, which found little evidence of greater than three groups exisiting in the latent space.

The DPORD model was fit to the Doubs river data using the probabilistic programming language NIMBLE (de Valpine et al. n.d., 2017); code is provided in Appendix B. To allow comparison to previous analyses of these data, the DPORD model was fit with $d = 2$ dimensions in the latent space. Three independent MCMC chains with random starting values were run for 100000 iterations and the first 50000 iterations of each chain were discarded as warm-up; the MCMC routine was assessed for convergence both visually and through the Gelman-Rubin statistic (Brooks and Gelman 1998). To account for the label-switching problem that is inherent to MCMC sampling of mixture distributions

(Frühwirth-Schnatter 2011), the label-switching algorithm developed by Malsiner-Walli, Frühwirth-Schnatter, and Grün (2017) was applied post-hoc. This algorithm is especially useful for infinite mixture models, as it can accommodate variable numbers of groups in MCMC procedures. Code to implement this label-switching algorithm is provided at `https://github.com/StrattonCh/dpord`.

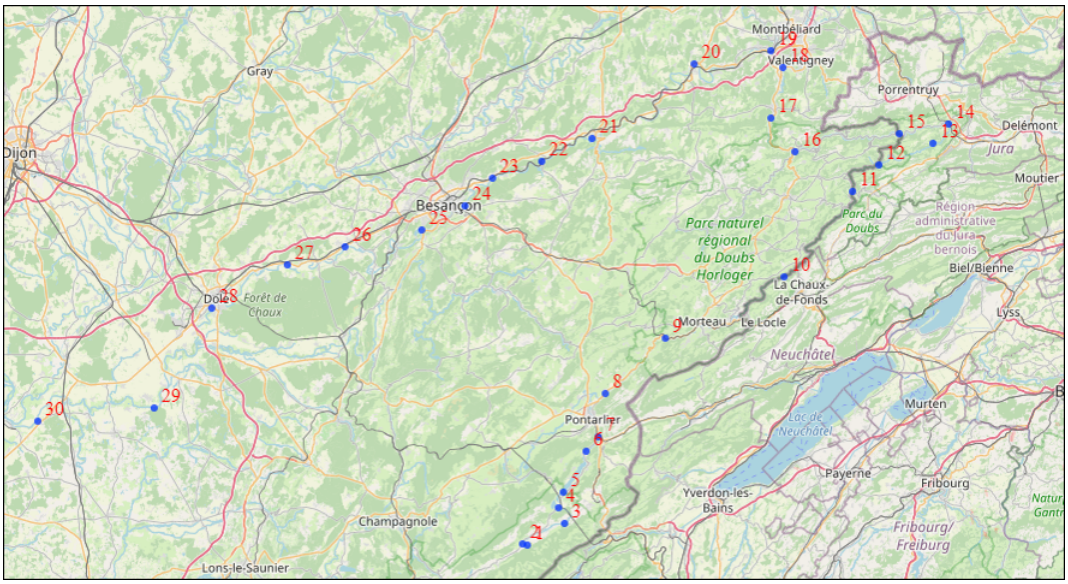

Figure 4.1: Map of 30 sample locations along the Doubs river in eastern France. Mapping layers provided by U.S. Geological Survey base mapping services.

4.4.2.2 Craters of the Moon data The second data set considered describes presence or absence of 78 plant species across 28 approximately 10 hectare sample frames, or primary sampling units, in the Craters of the Moon National Monument and Preserve (CRMO) in Idaho, USA. Craters of the Moon is located in a sagebrush steppe ecosystem, which covers much of the Great Basin and Snake River Plain that extends across the Interior West of the United States. Vegetation monitoring in CRMO presents a unique opportunity to explore ordination and clustering methods, as the park is comprised of a unique combination of volcanic features and sagebrush steppe along an elevational gradient, leading to anticipated

differences in plant composition based on proximity to volcanic features and elevational gradient position.

Sample frame locations were chosen to represent the range of community types within the CRMO sagebrush steppe landscape and were arranged along the proximity to volcanic features and elevation gradients (Yeo et al. 2009). Within each delineated sample frame, approximately 50 one square meter quadrat locations were selected according to the Generalized Random Tessellation Stratified (GRTS) spatially-balanced sampling design (Stevens and Olsen 2004). A total of 1518 vegetation quadrat locations were included in our analysis. For each quadrat location, ordinal cover classes for all plant species present in the one square meter quadrat were recorded. These ordinal classes were converted to presence-absence records prior to analysis. Sample frame locations are provided in Figure 4.2. The primary analytical goal for the CRMO data was to assess the similarity in species composition across the 28 sample frames in order to better understand community types within the monitoring program and to inform management strategies for preserving the sagebrush communities. Additionally, sample frames 19, 20, 25, 26, 27, 28, and 35 were of particular interest, as they are located in kipukas. Kipukas are areas of deeper soiled vegetation within the lava fields in CRMO that are completely surrounded by younger lava flows. Consequently, these sample frames have been physically isolated from livestock grazing and represent potentially quasi-pristine plant communities, making them of particular conservation interest (Nicolli 2019).

Prior to analysis, all species that did not appear in at least five percent of sample locations were excluded, resulting in 25 species considered for analysis. Removal of rare species prior to ordination is common in many practical settings, as common species are often most important for assessing relationships between species composition or abundance and underlying environmental gradients (Brasil et al. 2020). The same prior distributions implemented with the Daubs river data were used with the CRMO data, resulting in the prior

belief that fewer than six clusters existed in the latent space. While the number of groups present in the latent ecological gradient was not known *a priori*, prior research (Esposito et al. 2019) and expert opinion suggested that there existed a relatively small number of distinct ecological community types in the sagebrush ecosystems in CRMO. The gamma(1,2) prior placed on the Dirichlet process concentration parameter both reflected this prior belief and encouraged clustering in the latent space.

Four models were fit to the CRMO data: the hierarchical DPORD model (defined by equations 4.2, 4.3) and the standard DPORD model (defined by equations 4.1 and 4.3), each with two and three dimensions in the latent space. The Watanabe-Akaike information criterion (WAIC) was calculated for each model (Gelman, Hwang, and Vehtari 2013). Watanabe-Akaike information criterion values for each of the four models considered are provided in Table 4.1. Table 4.1 suggests that the CRMO data provide the most support for the hierarchical DPORD model with three dimensions in the latent space. As a result, inference concerning the CRMO data is hereafter based on the hierarchical DPORD model with three dimensions in the latent space. All models were fit using NIMBLE (de Valpine et al. n.d., 2017), and code is provided in Appendix B. Three independent MCMC chains with random starting values were run for 150,000 iterations and the first 75,000 iterations of each chain were discarded as warm-up; the MCMC routine was again assessed for convergence both visually and through the Gelman-Rubin statistic (Brooks and Gelman 1998). Posterior samples were post-processed for label-switching using the algorithm developed by Malsiner-Walli, Frühwirth-Schnatter, and Grün (2017).

4.5 Results

4.5.1 Daubs river data

The posterior modal number of clusters in the latent ecological gradient for the Doubs river data was two (posterior probability of 0.38), with posterior probabilities of 0, 0.31,

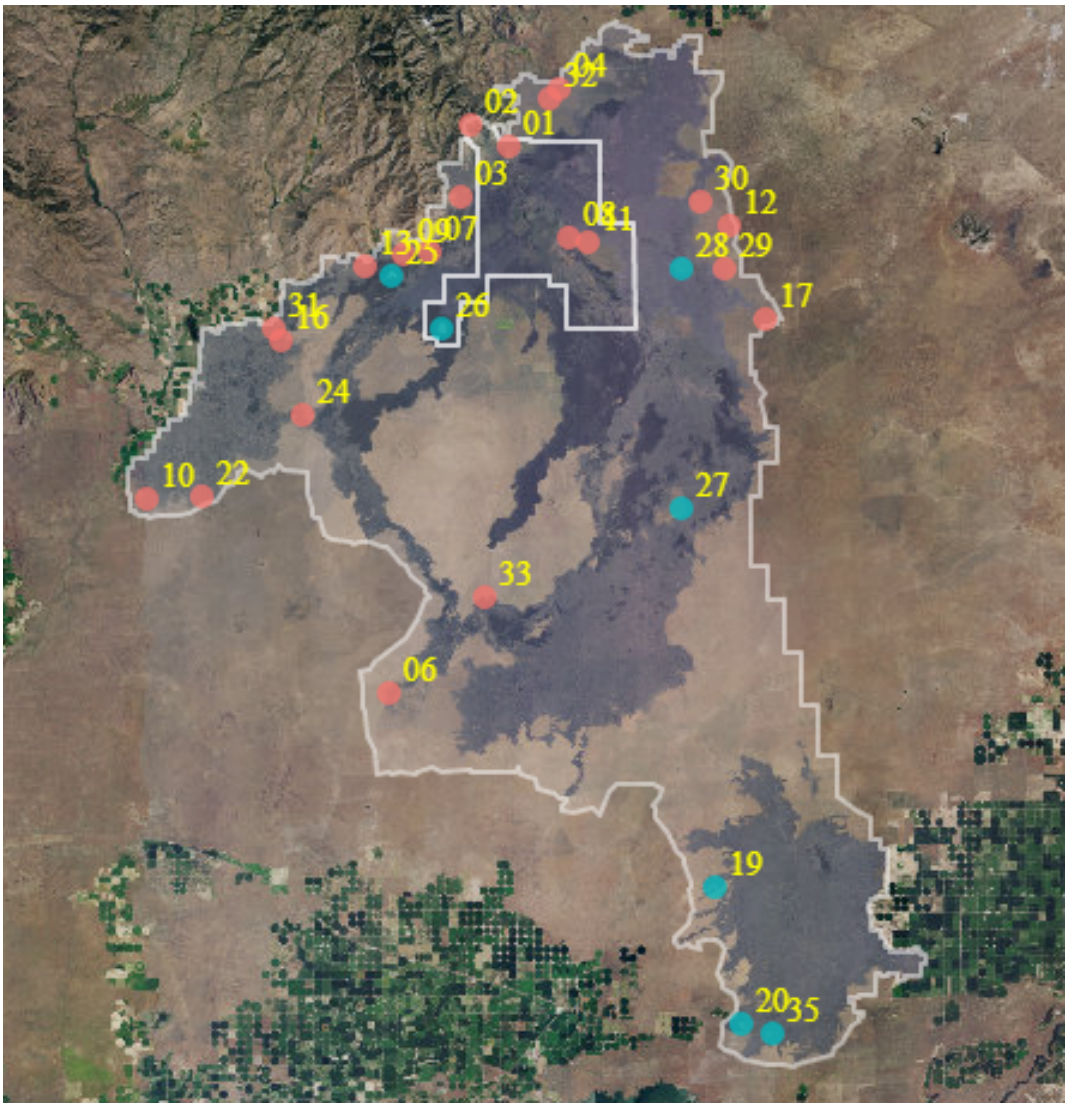

Figure 4.2: Map of 28 sample frame locations in Craters of the Moon National Monument and Preserve in Idaho, USA. Sample frames 19, 20, 25, 26, 27, 28, and 35 are located in unique vegetated islands within lava flows called kipukas (colored in blue), and are of particular conservation interest with potentially pristine communities not physically disturbed by grazing or other anthropogenic activities (Yeo et al. 2009).

0.17, and 0.14 for one, three, four, and five or more clusters respectively. These results are consistent with the previous analysis by Hui (2016), which found more support for two clusters in the latent space ($\text{DIC}_c = 221.51$) than three clusters in the latent space ($\text{DIC}_c = 223.02$). The DPORD modeling framework is advantageous here as it allows for

	$d = 2$	$d = 3$
standard DPORD	26971.99	25764.28
hierarchical DPORD	24585.86	**23353.19**

Table 4.1: WAIC values for each of the four models fit to the DPORD model. These results suggest the CRMO data provide the most support for the hierarchical DPORD model with three dimensions in the latent space.

probabilistic statements about the number of clusters in the latent ecological gradient, rather than relying on information criterion to select the appropriate number of clusters. Posterior mean estimates of the latent coordinates for each sample location colored by posterior modal cluster assignments are provided in Figure 4.3. The two distinct clusters in this figure roughly correspond to upstream and downstream sample locations, which reflects previous analyses of the Doubs river data (Hui 2016).

In addition to posterior modal cluster assignments, it can be informative to look at pairwise, posterior cluster membership between sample locations. Each cell in Figure 4.4 represents the posterior probability that the sample location on the horizontal and vertical axis share a cluster in the latent space. This figure further elucidates the separation between upstream and downstream sample locations, but also describes the degree of certainty in pairwise cluster membership. For example, sample locations 1 through 15 all share a high degree of certainty in their pairwise cluster membership. Conversely, sample location 16, which shares a posterior modal group assignment with locations 1 through 15, has a much lower pairwise cluster membership probability with locations 1 through 15. This result is consistent with the discovered separation between upstream and downstream sample locations in the latent space, as site 16 is located geographically in the transition between upstream and downstream sample locations.

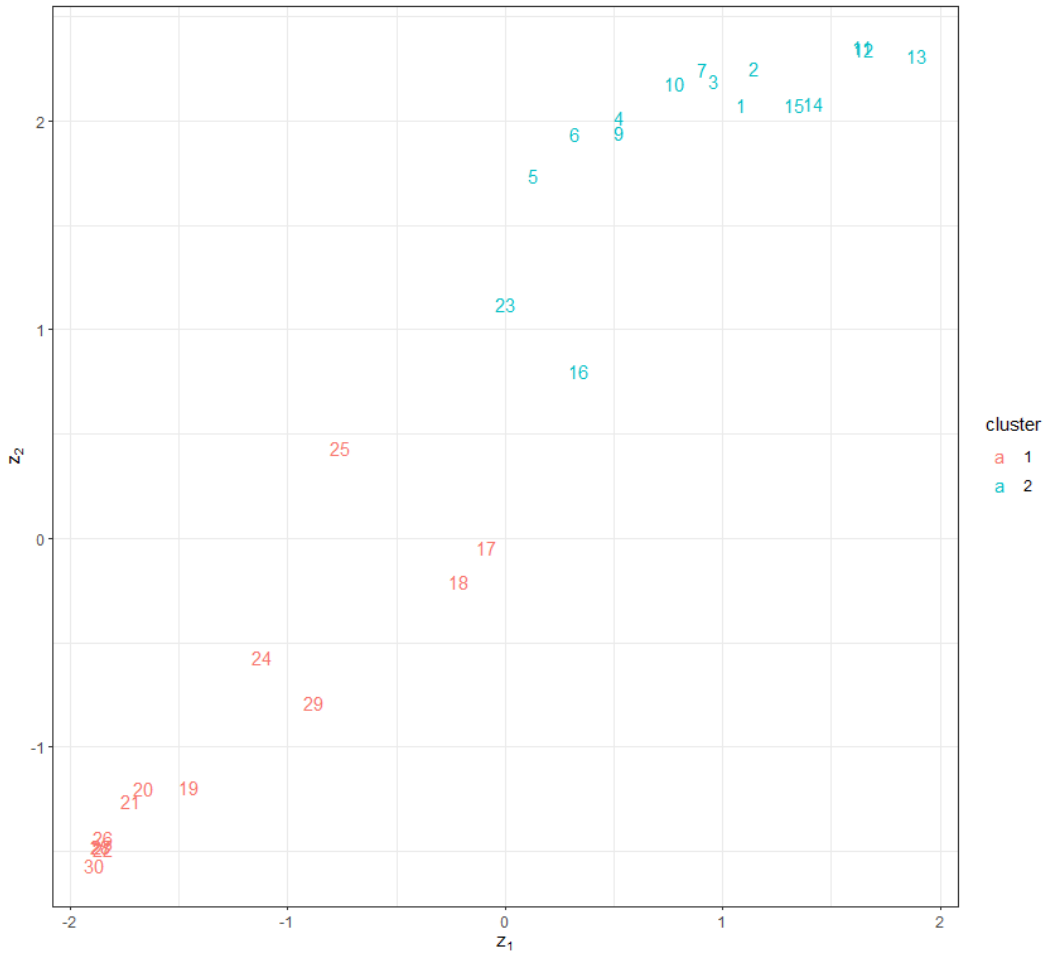

Figure 4.3: Posterior mean latent coordinates of each sample location in the Doubs river data; clustering in the latent space is denoted with color. The two clusters in the latent space roughly correspond to upstream and downstream sample locations, with the blue cluster representing upstream sites and the red cluster representing downstream sites.

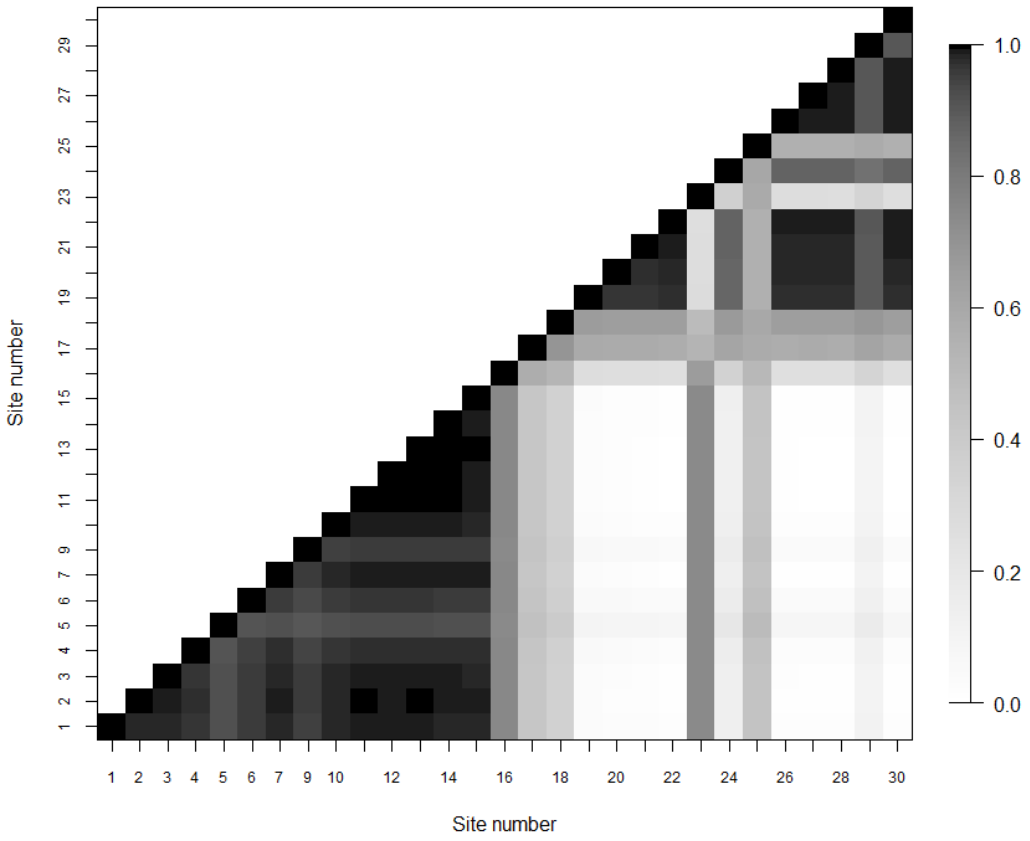

Figure 4.4: Pairwise, posterior cluster membership probabilities between sample locations in the Doubs river data. Each cell in this plot represents the posterior probability that the sample locations on the horizontal and vertical axis share a cluster in the latent space.

4.5.2 Craters of the Moon Monument and Preserve data

The posterior modal number of clusters in the latent space for the CRMO data was four (posterior probability of 0.31), with posterior probabilities of 0.22, 0.24, and 0.14, and 0.09 for less than four, five, six, and seven or more clusters respectively. Posterior mean estimates of the latent coordinates for each sample frame colored by posterior modal cluster assignments are provided in Figure 4.5. Pairwise, posterior cluster membership probabilities are summarized in Figure 4.6. Each cell in Figure 4.6 represents the posterior probability the sample frame on the horizontal and vertical axis share a cluster in the latent space. Figure 4.6 can be used to identify groups of sample frames that tend to be similar in species composition. For example, sample frames 12, 13, 16, and 17 tend to share latent cluster assignments with sample frames 3, 4, 6, and 7. A map of sample frames colored by cluster assignment is provided in Figure 4.7. Cluster labels for each sample frame were determined by their posterior modal cluster assignment.

Clustering results largely align with ecological gradients that exist within the park. For example, sample frames 1, 2, and 8, which define group four located within the interior monument portion of the park, are three of the four highest elevation sample frames in the park. These three sample frames are located along the highest ridge in CRMO, which spans from the Pioneer mountains north of the monument down to the visitor center. These sample frames were largely dominated by big sagebrush (*Artemisia tridentata*, present in 88% of quadrat locations) and wild buckwheat (*Eriogonum spp.*, present in 60% of quadrat locations), though bitterbrush (*Purshia tridentata*) and cheatgrass (*Bromus tectorum*) were common as well, present in 56% and 50% of quadrat locations, respectively.

Clustering results also corroborated anticipated differences in species composition between the sample frames located in kipukas and those elsewhere in the park. Group three (frames 19, 20, 26, 27, 32, and 35) accounts for five of the seven sample frames located in kipukas, and group two (frames 11, 25, and 28) accounts for the remaining two frames.

Quadrat locations from sample frames in group three were largely dominated by cheatgrass (*Bromus tectorum*, present in 99% of quadrat locations) and wild bluegrass (*Poa secunda*, present in 66% of quadrat locations), but tumbleweed mustard (*Sisymbrium altissimum*) and *Allium spp.* were also fairly common, present in 49% and 50% of quadrat locations, respectively. Our discovery of heavy weed infestation (both cheatgrass and tumbleweed mustard) in the cluster of kipuka frames within group three was surprising and suggests a process of degradation that has occurred in spite of physical barriers to grazing and other land use activities. The two kipuka frames (25 and 28) clustered in group two are quite distinctive with low amounts of big sagebrush but relatively high abundance of two other native sagebrush shrubs (*Artemisia arbuscula* and *Artemisia tripartita*), and they also remain relatively uninfested by cheatgrass and other weeds. These frames were dominated mostly by needlegrass (*Achnatherum spp.*, present in 69% of quadrat locations), and wild bluegrass (*Poa secunda*, present in 65% of quadrat locations), with Indian parsley (*Lomatium spp.*) also fairly prevalent, present in 53% of quadrat locations.

The driving relationships behind posterior clustering assignments may also be visualized with an alluvial diagram (Figure 4.8). In Figure 4.8, each alluvia describes the number of quadrats within a group for which the species described on the x-axis was present (1) or absent (0). This plot may be used to visualize the relationships previously described. For example, group four is unique in that it contains relatively low presence of *Poa secunda* and *Artemisia tridentata*, which were commonly observed in other groups. Additionally, group three contains a large proportion of quadrats occupied by *Poa secunda* and near ubiquitous presence of *Bromus tectorum*. Conversely, group two is defined by relatively common presence of *Poa secunda*, yet relatively low presence of *Bromus tectorum*.

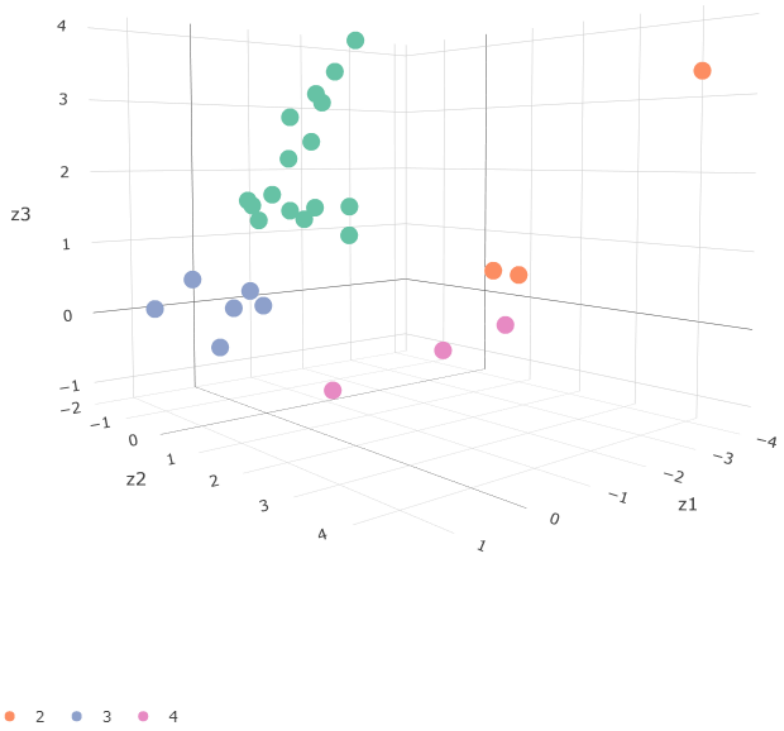

Figure 4.5: Posterior mean latent coordinates of each sample frame in the CRMO data. Coordinates are colored by posterior modal cluster assignment.

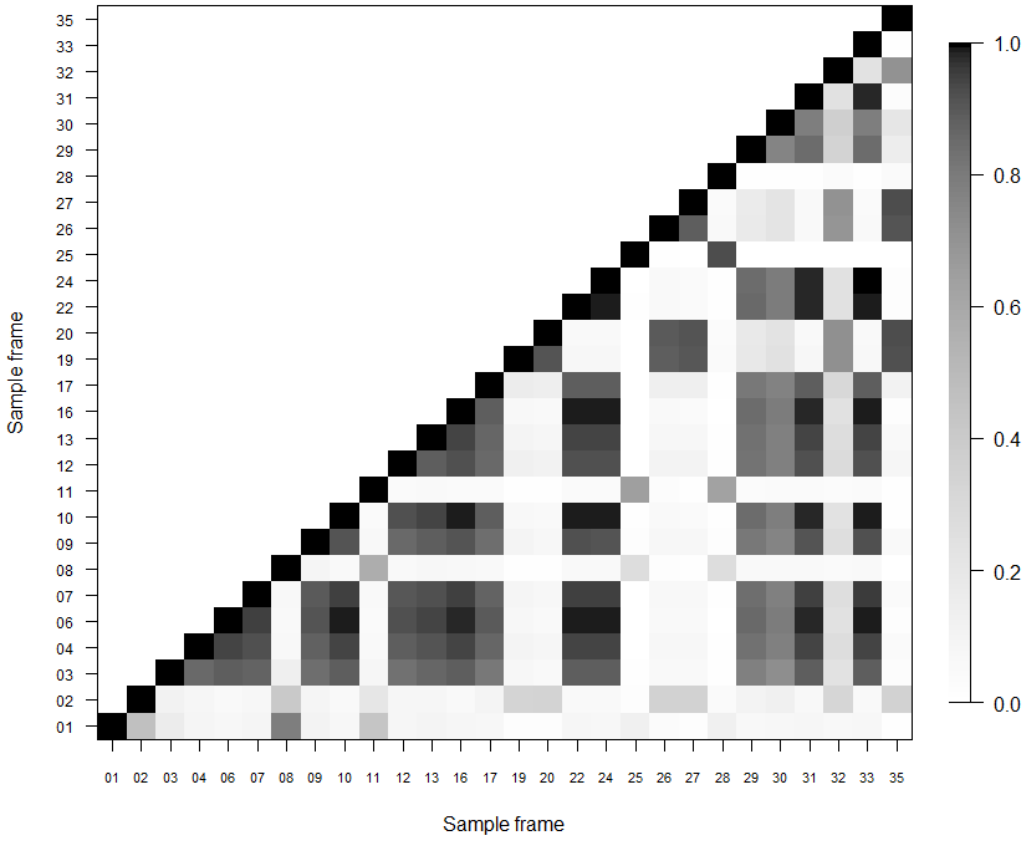

Figure 4.6: Pairwise, posterior cluster membership probabilities between sample locations for the CRMO data. Each cell in this plot represents the posterior probability that sample frames on the horizontal and vertical axes share a cluster assignment.

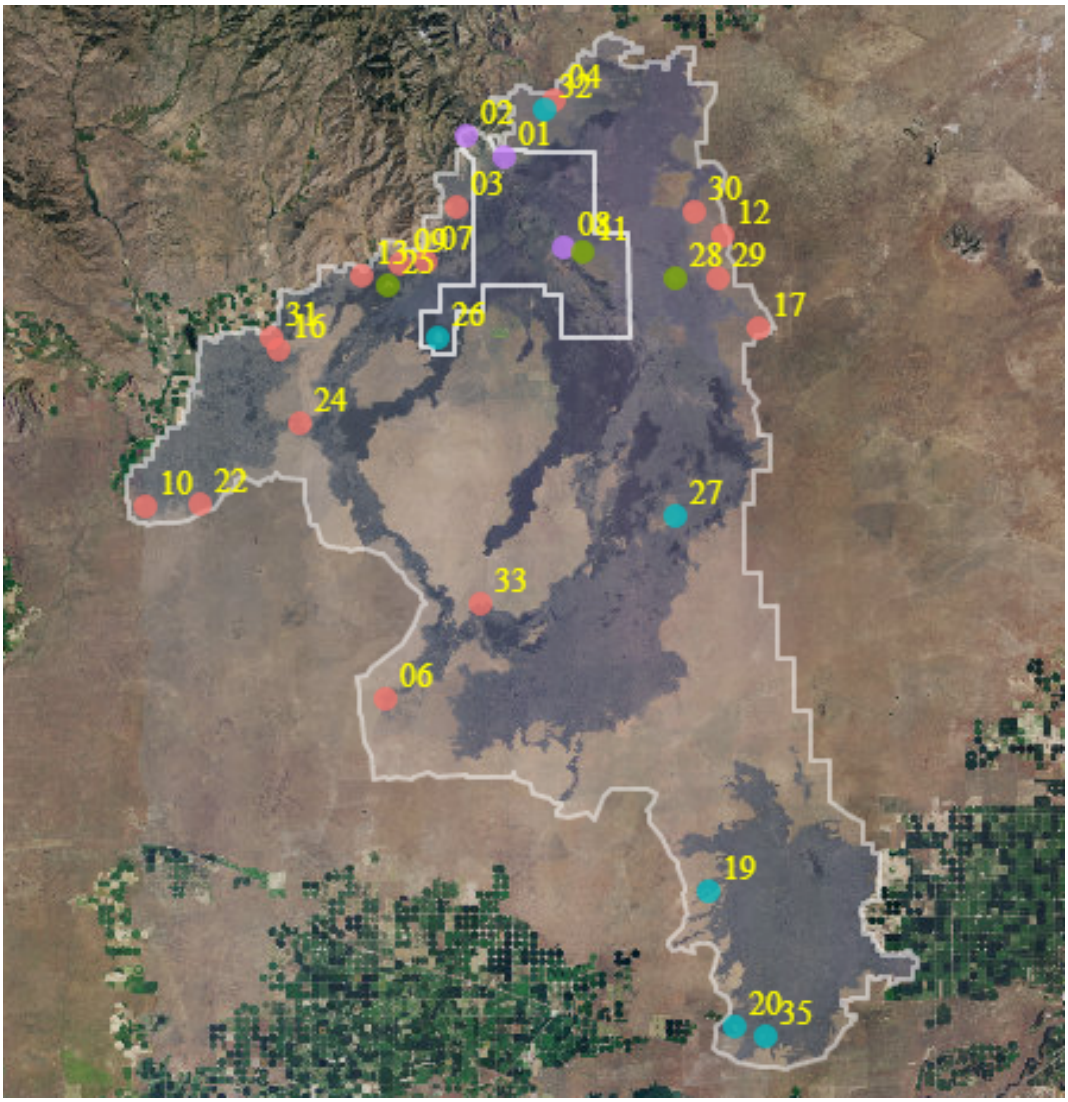

Figure 4.7: Sample locations and ordination results for the CRMO data. Cluster labels for each sample frame (denoted by color) were determined by posterior modal cluster labels for each sample frame. Group three is denoted in blue (frames 19, 20, 26, 27, 32, and 35), group two is denoted in green (frames 11, 25, 28), group four is denoted in purple (frames 1, 2, and 8), and group one is denoted in red (all remaining frames).

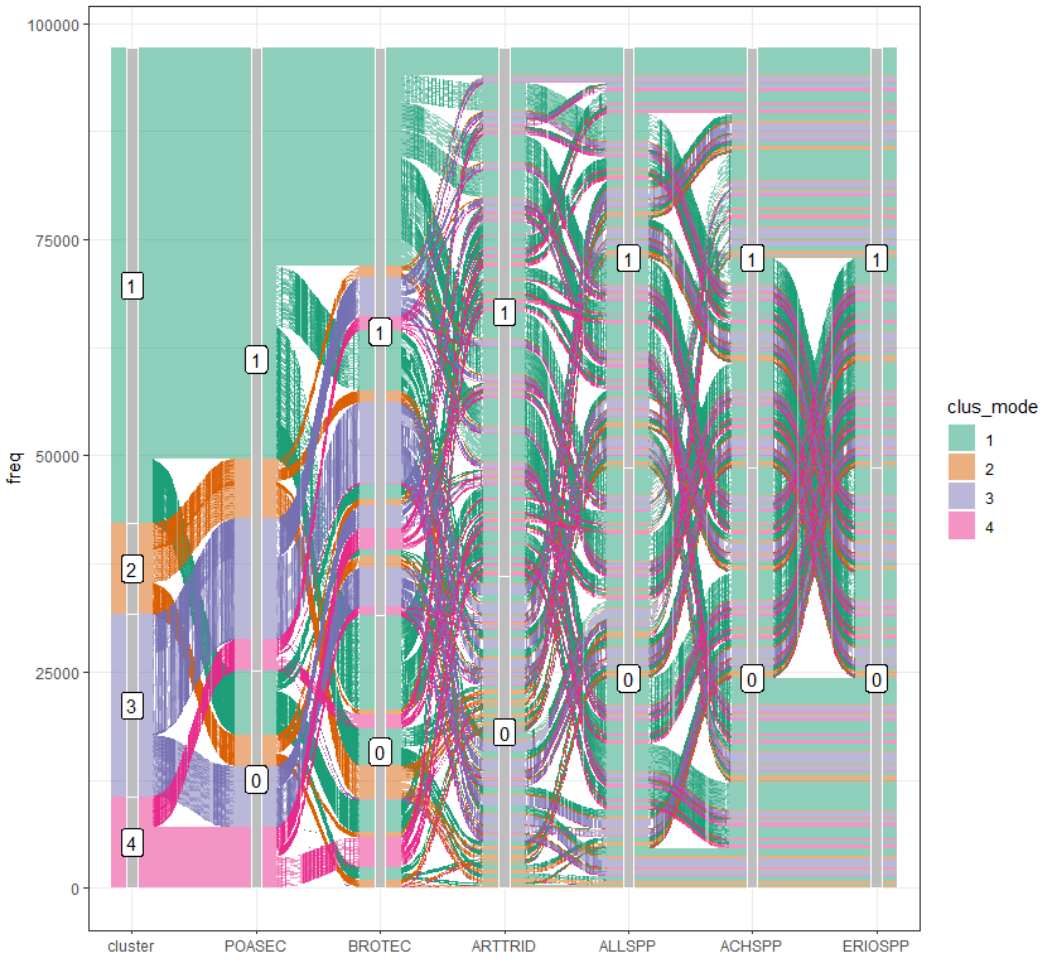

Figure 4.8: Alluvial diagrams explaining posterior cluster assignments for the most commonly observed species (*Poa secunda*, *Bromus tectorum*, *Artemisia tridentata*, *Allium spp.*, *Achnatherum spp.*, *Eriogonum spp.*). Each alluvia describes the number of quadrats within a group for which the species on the x-axis was present (1) or absent (0). This figure suggests that group four differs from the other groups in that it contains relatively low presence of *Poa secunda* and *Artemisia tridentata*, which were otherwise commonly observed. Additionally, groups two and three contain a large proportion of quadrats occupied by *Poa secunda* but differ substantially in prevalence of *Bromus tectorum*.

4.6 Discussion

We developed a statistical framework for simultaneous clustering and ordination of abundance data that allows for estimation of the number of clusters present in the latent ecological gradient. Additionally, that framework was extended to align ordination results with the primary sampling unit in hierarchical sampling designs. Code is provided for fitting and summarizing the DPORD model, facilitating easier uptake and application among ecologists and wildlife researchers. Existing methods for simultaneous clustering and ordination of abundance data rely on expert knowledge, information criterion, or other model-selection tools to choose the appropriate number of clusters present in the latent ecological gradient; in these cases, inference then proceeds conditional on the number of clusters specified in a two-step process. Conversely, by incorporating Dirichlet process priors on mixture-component-level distributions in a hierarchical ordination model, the DPORD model allows for data-driven estimation of the appropriate number of clusters in the latent space in a single step. As a result, ecologists and wildlife managers are afforded probabilistic statements about the number of clusters present in the latent ecological gradient. This information can be valuable when creating management strategies, as it informs the number of distinct communities present in the surveyed region.

Using the DPORD model, we provided ordination results for the historic Doubs river fish data set. Results suggested strong evidence of two distinct regions in the latent ecological gradient corresponding to upstream and downstream locations, consistent with previous analyses of these data. A hierarchical formulation of the DPORD model was also applied to plant cover data from Craters of the Moon National Monument and Preserve in Idaho, USA, demonstrating the flexibility of this modeling framework to handle nested sampling designs. Results suggested there exist four distinct ecological regions in the latent ecological gradient corresponding to various features of the ecological gradient in CRMO, including

elevation, proximity to volcanic features, and degree of ecological degradation from weed invasion. Furthermore, clustering results were consistent with anticipated differences in species composition between sample frames located in kipukas and elsewhere in the park. These clustering results provide insight into differences between various plant communities in CRMO, supporting more reliable and cost effective monitoring of sagebrush steppe throughout the park.

One criticism of using Dirichlet process mixture models for model-based clustering is their tendency to form few large clusters and many "singleton" clusters (Müller and Mitra 2013). This issue can be partially addressed with appropriate specification of hyper-priors on the concentration parameter of the Dirichlet process, or by clustering sites via sparse finite mixtures (Frühwirth-Schnatter and Malsiner-Walli 2018). In the future, we plan to investigate how inferences regarding the number of clusters in the latent space are influenced by the choice of hyper-parameters for the gamma distribution prior on the concentration parameter in the Dirichlet process or by the use of sparse finite mixtures as a clustering mechanism. Additionally, implementation of the DPORD model on longitudinal data sets remains an area of active research. In this context, the DPORD model could be used to estimate changes in latent clustering over time by allowing the number of clusters and cluster assignments to vary with time. Finally, we note that this model can be easily extended to accommodate non-binary data sets by changing the assumed sampling model.

Multi-species monitoring programs continue to gain traction as a cost-effective means for wildlife managers to assess the impact of climate change, disease, wildfires, and other emerging threats. Model-based clustering and ordination techniques provide researchers with a comprehensive modeling framework for understanding how species composition varies across sample locations, allowing for data-driven management decisions. The development of the DPORD model further supports data-driven management decisions by allowing for estimation of the number of clusters present in the latent ecological gradient, providing

further insight into differences in species composition throughout monitored regions.

Acknowledgments

We thank Dr. Mark Greenwood for his helpful comments and discussion during writing of this manuscript. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data accessibility

Data

The Doubs river data are available through the `ade4` R package (Dray and Dufour 2007). The CRMO data are publicly available and archived at `https://github.com/StrattonCh/dpord`.

Modeling scripts

All R code used in analyses discussed in this manuscript at provided at `https://github.com/StrattonCh/dpord`.

CHAPTER FIVE

CONCLUSION

As the global climate continues to change, there is renewed focus on the development of statistical methods capable of accommodating the increasing complexity in ecological sampling designs and analytical goals. Hierarchical latent variable models provide a flexible framework for modeling ecological data of varying degrees of complexity, providing a means to both account for hierarchical sampling designs that are common in ecological applications and provide inferences regarding latent, unobserved quantities. This dissertation focuses on the use of hierarchical latent variable modeling of both multi-scale occupancy data and multi-species assemblage data.

Chapter 2 describes the development of an `R` package that implements a data augmentation strategy to fit computationally efficient multi-scale occupancy models. This advancement allows multi-scale occupancy models to be fit in a matter of seconds, rather than hours, and scales well for larger data sets. As a result, Bayesian multi-scale occupancy modeling is more accessible for natural resource managers. Additionally, the development of computationally expedient methods paves the way for online estimation of multi-scale occupancy models, allowing for real-time monitoring of occupancy within this framework. This is especially valuable for monitoring water born pathogens, and is a direction of future work.

Chapter 3 introduced distance-based ordination techniques as a means of assessing similarity in species composition between sample locations. Using simulation, we investigated how various dissimilarity indices affected ordination results and explored how well each measure preserved dissimilarity in the presence of zero-inflated responses. However, through that process, we identified multiple deficiencies within the distance-based ordination

framework. Among those deficiencies was the inability of a distance-based framework to accommodate hierarchical sampling designs and the lack of likelihood with which to assess choices of dissimilarity indices and clustering mechanisms. These shortcomings motivated our work in Chapter 4.

Chapter 4 describes the development of a hierarchical model that uses latent variables to conduct simultaneous clustering and ordination of ecological abundance data. Additionally, the proposed model allows for estimation of the number of groups present in the latent ordination space, rather than relying on expert opinion or choosing that value prior to analysis. Motivated by vegetation monitoring data, the proposed model was also extended to accommodate hierarchical sampling designs in which secondary sampling units are selected within primary sampling units.

The Dirichlet process ordination model, as described in Chapter 4, requires careful consideration of the implied prior distribution on the number of clusters in the latent space. While the clustering results considered in Chapter 4 seemed robust to the choice of that prior distribution, further research is required to fully understand the implication of the Dirichlet process as a clustering mechanism within the ordination framework. Prior research has indicated that the Dirichlet process has a tendency to form many "singleton" clusters and a few large clusters (Müller and Mitra 2013), but that appropriate specification of hyper-prior distributions on Dirichlet process parameters can help alleviate this effect (Frühwirth-Schnatter and Malsiner-Walli 2018). In the future we plan to further explore this relationship and develop guidelines to encourage implementation of this modeling framework among natural resource managers.

REFERENCES

Antoniak, C. E. (1974). "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems". In: *The Annals of Statistics* 2 (6), pp. 1152 –1174.

Beck, J., J. Connelly, and K. Reese (2009). "Recovery of greater sage-grouse habitat features in Wyoming big sagebrush following prescribed fire". In: *Restoration Ecology* 17, pp. 393–403.

Bingham, B. et al. (2011). *Enhanced monitoring to better address rapid climate change in high-elevation parks: A multi-network strategy.* Natural Resource Report NPS/IMR/NRR-2011/285. National Park Service, Fort Collins, Colorado.

Braak, C. ter and I. Prentice (1988). "A theory of gradient analysis". In: *Advances in Ecological Research* 18, pp. 271–317.

Brasil, L. et al. (2020). "The importance of common and the irrelevance of rare species for partition the variation of community matrix: Implications for sampling and conservation". In: *Scientific Reports* 10 (19777).

Bray, J. R. and J. T. Curtis (1957). "An ordination of the upland forest communities of southern Wisconsin". In: *Ecological Monographs* 27 (4), pp. 325–349.

Brooks, S. and A. Gelman (1998). "General methods for monitoring convergence of iterative simulations". In: *Journal of Computational and Graphical Statistics* 7, pp. 434–455.

Bálint, M. et al. (2018). "Environmental DNA time series in ecology". In: *Trends in Ecology and Evolution* 33 (12), pp. 945 –957.

Carpenter, B. et al. (2017). "Stan: A probabilistic programming language". In: *Journal of Statistical Software* 76 (1), pp. 1 –32.

Choi, S.-W. et al. (2019). "Long-term (2005–2017) macromoth community monitoring at Mt. Jirisan National Park, South Korea". In: *Ecological Research* 34 (4), pp. 443–443.

Damgaard, C., R. R. Hansen, and F. Hui (2020). "Model-based ordination of pin-point cover data: Effect of management on dry heathland". In: *bioRxiv*.

Daubenmire, R. (1959). "A canopy-coverage method". In: *Northwest Science* 33, pp. 43–64.

de Valpine, P. et al. (n.d.). *NIMBLE: MCMC, particle filtering, and programmable hierarchical modeling*.

de Valpine, P. et al. (2017). "Programming with models: Writing statistical algorithms for general model structures with NIMBLE". In: *Journal of Computational and Graphical Statistics* 26 (2), pp. 403–413.

Dorazio, R. (2015). "Bayesian data analysis in population ecology: Motivations, methods, and benefits". In: *Population Ecology* 58, pp. 31 –44.

Dorazio, R. M. and R. A. Erickson (2018). "eDNAoccupancy: an R package for multiscale occupancy modelling of environmental DNA data". In: *Molecular Ecology Resources* 18 (2), pp. 368 –380.

Dray, S. and A.-B. Dufour (2007). "The ade4 Package: Implementing the duality diagram for ecologists". In: *Journal of Statistical Software* 22 (4), 1–20.

Eiler, A. et al. (2018). "Environmental DNA (eDNA) detects the pool frog (*Pelophylax lessonae*) at times when traditional monitoring methods are insensitive". In: *Scientific Reports* 8 (1), pp. 1 –9.

Elzinga, C. et al. (2001). *Monitoring plant and animal populations*. Blackwell Sciences, Malden, MA.

Erickson, R., C. Merkes, and E. Mize (2019). "Sampling designs for landscape-level eDNA monitoring programs". In: *Integrated Environmental Assessment and Management* 15.

Esposito, D. et al. (2019). "Differential species responses to aspects of resistance to invasion in two Columbia Plateau - Protected Areas". In: *Rangeland Ecology and Management*.

Esposito, D. and T. Rodhouse (2015). *Sagebrush steppe vegetation monitoring in John Day Fossil Beds National Monument: 2014 annual report*. Natural Resource Data Series NPS/UCBN/NRDS—2015/797. National Park Service, Fort Collins, Colorado.

Frühwirth-Schnatter, S. (2011). *Finite mixture and Markov switching models*. Springer.

Frühwirth-Schnatter, S. and G. Malsiner-Walli (2018). "From here to infinity: Sparse finite versus Dirichlet process mixtures in model-based clustering". In: *Advances in Data Analysis and Classification* 13, pp. 1–32.

Gelman, A., J. Hwang, and A. Vehtari (2013). "Understanding predictive information criteria for Bayesian models". In: *Statistics and Computing* 24, pp. 1 –30.

Golding, N. (2019). "greta: Simple and scalable statistical modelling in R". In: *Journal of Open Source Software* 4, p. 1601.

Hastings, W. K. (1970). "Monte Carlo sampling methods using Markov chains and their applications". In: *Biometrika* 57 (1), pp. 97–109.

Hui, F. (2016). "Model-based simultaneous clustering and ordination of multivariate abundance data in ecology". In: *Computational Statistics and Data Analysis* 105.

Hui, F. K. et al. (2015). "Model-based approaches to unconstrained ordination". In: *Methods in Ecology and Evolution* 6 (4), pp. 399–411.

Hunter, M. et al. (2019). "Efficacy of eDNA as an early detection indicator for Burmese pythons in the ARM Loxahatchee National Wildlife Refuge in the greater Everglades ecosystem". In: *Ecological Indicators* 102, pp. 617 –622.

Hutchins, P. et al. (2019). "The Yellowstone river fish-kill: Fish health informs and is informed by vital signs monitoring". In: *Yellowstone Science* 27 (1), pp. 55 –57.

Inoue, K., K. Stoeckl, and J. Geist (2017). "Joint species models reveal the effects of environment on community assemblage of freshwater mussels and fishes in European rivers". In: *Diversity and Distributions* 23 (3), pp. 284–296.

Jain, A. K. and R. C. Dubes (1988). *Algorithms for clustering data*. Vol. 1. Prentice Hall.

Jean, C., K. Stella, and R. Daley (2014). *Sagebrush steppe vegetation monitoring in Grand Teton National Park: 2013 data summary*. Natural Resource Data Series NPS/GRYN/NRDS-2014/680. National Park Service, Fort Collins, Colorado.

Klymus, K. E. et al. (2015). "Quantification of eDNA shedding rates from invasive bighead carp *Hypophthalmichthys nobilis* and silver carp *Hypophthalmichthys molitrix*". In: *Biological Conservation* 183, pp. 77 –84.

Kruskal, J. (1964). "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis". In: *Psychometrika* 29 (1), pp. 1–27.

Kruskal, J. and M. Wish (1978). *Multidimensional scaling*. Vol. 1. Sage Publications.

Lodge, D. et al. (2012). "Conservation in a cup of water: Estimating biodiversity and population abundance from environmental DNA". In: *Molecular Ecology* 21, pp. 2555 –2558.

Macqueen, J. (1967). "Some methods for classification and analysis of multivariate observations". In: *5^{th} Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297.

Malsiner-Walli, G., S. Frühwirth-Schnatter, and B. Grün (2017). "Identifying mixtures of mixtures using Bayesian estimation". In: *Journal of Computational and Graphical Statistics* 26 (2), pp. 285–295.

Mitchell, J., P. Bartling, and R. O'Brien (1988). "Comparing cover class macroplot data with direct estimates from small plots". In: *American Midland Naturalist* 120, pp. 70–78.

Mize, E. L. et al. (2019). "Refinement of eDNA as an early monitoring tool at the landscape-level: Study design considerations". In: *Ecological Applications* 29 (6).

Mordecai, R. et al. (2011). "Addressing challenges when studying mobile or episodic species: Hierarchical Bayes estimation of occupancy and use". In: *Journal of Applied Ecology* 48, pp. 56 –66.

Müller, P. and R. Mitra (2013). "Bayesian nonparametric inference – why and how". In: *Bayesian Analysis* 8 (2), pp. 269 –302.

National Park Service (2018). *Sagebrush steppe vegetation and ground cover class values collected for long term monitoring at Grand Teton National Park starting in 2012 through 2018.* Data provided to Christian Stratton, Montana State University November 19, 2018. National Park Service Inventory and Monitoring Division - Greater Yellowstone Network. Bozeman, MT.

Nichols, J. et al. (2008). "Multi-scale occupancy estimation and modelling using multiple detection methods". In: *Journal of Applied Ecology* 45, pp. 1321 –1329.

Nicolli, M. M. (2019). *Sagebrush steppe vegetation monitoring in Craters of the Moon National Monument and Preserve*. Natural Resource Report NPS/UCBN/NRR-2019/2020. National Park Service, Bend, Oregon, USA.

Oksanen, J. et al. (2020). *vegan: Community ecology package*. R package version 2.5-7.

Peet, R., T. Wentworth, and P. White (1998). "A flexible multipurpose method for recording vegetation composition and structure". In: *Castanea* 63, pp. 262–274.

Pilliod, D. et al. (2019). "Integration of eDNA-based biological monitoring within the U.S. Geological Survey's national streamgage network". In: *JAWRA Journal of the American Water Resources Association*, pp. 1505 –1518.

Plummer, M. et al. (2006). "CODA: Convergence diagnosis and output analysis for MCMC". In: *R News* 6 (1), pp. 7 –11.

Pollock, L. et al. (2014). "Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM)". In: *Methods in Ecology and Evolution* 5.

Polson, N., J. Scott, and J. Windle (2013). "Bayesian inference for logistic models using Polya-Gamma latent variables". In: *Journal of the American Statistical Association* 108, pp. 1339 –1349.

Ricotta, C. and E. Feoli (2013). "Does ordinal cover estimation offer reliable quality data structures in vegetation ecological studies?" In: *Folia Geobotanica* 48 (4), pp. 437–447.

Robert, C. P. (2015). "The Metropolis-Hastings algorithm". In: *arXiv*.

Roberts, D. W. (2020). "Comparison of distance-based and model-based ordinations". In: *Ecology* 101 (1).

Roberts, G. and J. Rosenthal (2001). "Optimal scaling for various Metropolis-Hastings algorithms". In: *Statistical Science* 16, pp. 351 –367.

Schmelzle, M. and A. Kinziger (2016). "Using occupancy modeling to compare environmental DNA to traditional field methods for regional-scale monitoring of an endangered aquatic species". In: *Molecular Ecology Resources* 16, pp. 1 –14.

Sengupta, M. et al. (2019). "Environmental DNA for improved detection and environmental surveillance of schistosomiasis". In: *Proceedings of the National Academy of Sciences* 116 (18), pp. 8931 –8940.

Sepulveda, A. J. et al. (2019). "Adding invasive species biosurveillance to the U.S. Geological Survey streamgage network". In: *Ecosphere* 10 (8), pp. 1 –17.

Skrondal, A. and R. Rabe-Hesketh (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. Boca Raton: Chapman and Hall.

Society of Range Management (1999). *A glossary of terms used in range management.* Denver, CO.

Soil Survey Staff (2019). *Gridded Soil Survey Geographic (gSSURGO)*. Database for Wyoming. United States Department of Agriculture, Natural Resources Conservation Service. Available online at https://gdg.sc.egov.usda.gov/.

Spiegelhalter, D. J. et al. (2002). "Bayesian measures of model complexity and fit". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64 (4), pp. 583–639.

Stan Development Team (2018). *Stan modeling language users guide and reference manual, version 2.18.0*.

Stevens, D. and A. Olsen (2004). "Spatially balanced sampling of natural resources". In: *Journal of the American Statistical Association* 99, pp. 262–278.

Stratton, C. (2020a). *StrattonCh/msocc: msocc release (version 1.0.0)*. `http://doi.org/10.5281/zenodo.3911239`. Zenodo.

— (2020b). *StrattonCh/msocc_supplemental_code: msocc supplemental code release*. `https://doi.org/10.5281/zenodo.3908568`. Zenodo.

Stratton, C. et al. (2019). *Assessing spatial and temporal patterns in sagebrush steppe vegetation communities, 2012-2018: Grand Teton National Park*. Natural Resource Report NPS/GRYN/NRR-2019/2020. National Park Service, Fort Collins, Colorado.

Tercek, M. et al. (2015). *Upper Columbia Basin Network sagebrush steppe vegetation monitoring protocol: Narrative and standard operating procedures (bound separately) version 1.0*. Natural Resource Report NPS/UCBN/NRR—2009/142. National Park Service, Fort Collins, Colorado.

Uchii, K. et al. (2017). "Distinct seasonal migration patterns of Japanese native and non-native genotypes of common carp estimated by environmental DNA". In: *Ecology and Evolution* 7, pp. 8515 –8522.

Veen, B. van der et al. (2021). "Model-based ordination for species with unequal niche widths". In: *Methods in Ecology and Evolution* 12 (7), pp. 1288–1300.

Warton, D. I., S. T. Wright, and Y. Wang (2012). "Distance-based multivariate analyses confound location and dispersion effects". In: *Methods in Ecology and Evolution* 3 (1), pp. 89–101.

Williams, K. et al. (2018). "Detection and persistence of environmental DNA from an invasive, terrestrial mammal". In: *Ecology and Evolution* 8, pp. 688 –695.

Yeo, J. and T. Rodhouse (2013). *Sagebrush steppe vegetation monitoring protocol narrative: Grand Teton National Park*. Unpublished Natural Resource Report, Upper Columbia Basin Network, Moscow, ID.

Yeo, J. et al. (2009). *Sagebrush steppe vegetation monitoring protocol narrative: Grand Teton National Park*. Unpublished Natural Resource Report, Upper Columbia Basin Network, Moscow, ID.

Zurell, D. et al. (2020). "Testing species assemblage predictions from stacked and joint species distribution models". In: *Journal of Biogeography* 47 (1), pp. 101–113.

APPENDICES

APPENDIX A

SUPPLEMENTAL MATERIALS FOR CHAPTER 3

Table of species observed in GRTE

Table A.1 provides a list of the species observed in GRTE, their species group, and the year first observed.

Species Name	Species group	Year First Targeted
Achnatherum lettermanii	Natv perenn grasses	2015
Achillea millefolium	Natv persist forbs	2012
Achnatherum nelsonii	Natv perenn grasses	2012
Acroptilon repens	NonNatv inv forbs	2012
Agoseris spp.	Natv other forbs	2012
Agropyron cristatum	NonNatv inv grasses	2012
Alyssum spp.	NonNatv inv forbs	2016
Amelanchier alnifolia	Shrubs	2012
Antennaria spp.	Natv persist forbs	2012
Arabis spp.	Natv other forbs	2012
Arenaria spp.	Natv persist forbs	2012
Artemisia arbuscula	Sagebrushes	2012
Artemisia cana	Sagebrushes	2012
Artemisia frigida	Sagebrushes	2012
Artemisia tridentata	Sagebrushes	2012
Artemisia tripartita	Sagebrushes	2012
Aster spp.	Natv persist forbs	2012
Astragalus spp.	Natv persist forbs	2012
Balsamorhiza sagittata	Natv persist forbs	2012
Bare ground	Other	2012
Bromus inermis	NonNatv inv grasses	2012
Bromus japonicus	NonNatv inv grasses	2012
Bromus marginatus	Natv perenn grasses	2012
Bromus tectorum	NonNatv inv grasses	2012
Cardaria draba	NonNatv inv forbs	2012
Carduus nutans	NonNatv inv forbs	2012
Carex spp.	Natv perenn grasses	2012
Castilleja spp.	Natv persist forbs	2012
Centaurea diffusa	NonNatv inv forbs	2012
Centaurea maculosa	NonNatv inv forbs	2012
Chrysothamnus viscidiflorus	Shrubs	2012
Cirsium vulgare	NonNatv inv forbs	2012
Clematis hirsutissima	Natv persist forbs	2012
Comandra umbellata	Natv persist forbs	2012

Table A.1 Continued

Crepis spp.	Natv persist forbs	2012
Cryptobiotic soil crust	Other	2017
Cynoglossum officinale	NonNatv inv forbs	2016
Dactylis glomerata	NonNatv inv grasses	2012
Danthonia unispicata	Natv perenn grasses	2012
Dasiphora floribunda	Shrubs	2012
Elymus elymoides	Natv perenn grasses	2012
Elymus repens	NonNatv inv grasses	2012
Elymus trachycaulus	Natv perenn grasses	2012
Erigeron spp.	Natv persist forbs	2012
Ericameria nauseosa	Shrubs	2012
Eriogonum spp.	Natv persist forbs	2012
Festuca idahoensis	Natv perenn grasses	2012
Frasera speciosa	Natv persist forbs	2012
Geranium viscosissimum	Natv persist forbs	2012
Geum triflorum	Natv other forbs	2012
Helianthella spp.	Natv persist forbs	2012
Hesperostipa comata	Natv perenn grasses	2012
Isatis tinctoria	NonNatv inv forbs	2012
Kochia scoparia	NonNatv inv forbs	2012
Koeleria macrantha	Natv perenn grasses	2012
Leymus cinereus	Natv perenn grasses	2016
Linaria dalmatica	NonNatv inv forbs	2012
Linaria vulgaris	NonNatv inv forbs	2012
Lithospermum ruderale	Natv persist forbs	2012
Litter	Other	2015
Lomatium spp.	Natv persist forbs	2012
Lupinus spp.	Natv persist forbs	2012
Mahonia repens	Shrubs	2012
Medicago sativa	NonNatv inv forbs	2012
Melilotus spp.	NonNatv inv forbs	2012
Melica spp.	Natv perenn grasses	2012
Microseris spp.	Natv other forbs	2017
Moss	Other	2017
Oryzopsis hymenoides	Natv perenn grasses	2012
Penstemon spp.	Natv persist forbs	2012
Perideridia montana	Natv persist forbs	2012
Phacelia hastata	Natv persist forbs	2012
Phlox hoodii	Natv persist forbs	2015
Phlox longifolia	Natv persist forbs	2015
Phleum pratense	NonNatv inv grasses	2012

Table A.1 Continued

Phlox spp.	Natv persist forbs	2012
Poa bulbosa	NonNatv inv grasses	2012
Poa secunda	Natv perenn grasses	2012
Poa spp.	NonNatv inv grasses	2012
Potentilla recta	NonNatv inv forbs	2012
Potentilla spp.	Natv persist forbs	2012
Prunus virginiana	Shrubs	2012
Pseudoroegneria spicata	Natv perenn grasses	2012
Purshia tridentata	Shrubs	2012
Ribes spp.	Shrubs	2012
Rock	Other	2016
Sedum spp.	Natv other forbs	2012
Symphoricarpos spp.	Shrubs	2012
Taraxacum officinale	NonNatv inv forbs	2012
Tragopogon dubius	NonNatv inv forbs	2012
Viola spp.	Natv other forbs	2012

Table A.1: Table of species observed in GRTE.

APPENDIX B

SUPPLEMENTAL MATERIALS FOR CHAPTER 4

Code for fitting DPORD model

Below, NIMBLE code for fitting the standard DPORD model is provided. Full implementation of this model to each data set is provided electronically at `https://github.com/StrattonCh/dpord`.

```
# priors
## site effects
for(site in 1:nsites){
    alpha[site] ~ dnorm(0, 1)
}

## species effects
for(species in 1:nspecies){
    beta[species] ~ dnorm(0, 1)
}

## z priors
### Dirichlet process mixture parameters
clus_id[1:nsites] ~ dCRP(dp_con, size = nsites)
dp_con ~ dgamma(1, rate = 2)

### table parameters - fix covariance as identity
for(i in 1:max_clus){
    mu[i, 1:d] ~ dmnorm(mu0[1:d], Lambda0[1:d, 1:d])
}

for(site in 1:nsites){
# identity matrix for constraint
    z[site, 1:d] ~ dmnorm(mu[clus_id[site], 1:d], cov = S[1:d, 1:d])
}

# theta prior
## upper triangle = 0
for(row in 1:(d-1)){
    for(col in (row+1):d){
        theta[row, col] <- 0
    }
}
```

```
## diag > 0
for(diag_element in 1:d){
    theta[diag_element, diag_element] ~ T(dnorm(0, sd = 1), 0, Inf)
}

## lower diag of first d rows
for(row in 2:d){
    for(col in 1:(row-1)){
      theta[row, col] ~ dnorm(0, sd = 1)
    }
}

## all other elements
for(row in (d+1):nspecies){
    for(col in 1:d){
        theta[row, col] ~ dnorm(0, sd = 1)
    }
}

# likelihood
for(site in 1:nsites){
    for(species in 1:nspecies){
        logit(pi[site, species]) <- alpha[site] + beta[species]
        + inprod(z[site,1:d], theta[species, 1:d])
        Y[site, species] ~ dbern(pi[site, species])
    }
}
```