Montana State University

Department of Mathematical Sciences

Writing Project

TITLE

Author:
Meaghan Winder

Supervisor:
Dr. Andrew Hoegh

Spring 2020

A writing project submitted in partial fulfillment
of the requirements for the degree

Master's of Science in Statistics

APPROVAL

of a writing project submitted by

Meaghan Winder

This writing project has been read by the writing project advisor and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Statistics Faculty.

Date

Andrew Hoegh
Writing Project Advisor

Date

Mark C. Greenwood
Writing Projects Coordinator

Contents

Abstract

abstract text here

1 Introduction

In early 2020, the City of Austin, Texas approved the spending of four million dollars over the next five years in an attempt to remove zebra mussels from the city's source of drinking water with a liquid copper sulfate pentahydrate released into the water intake pipes (Bontke, 2020). This is one of many pursuits to remove dreissenid mussels[1] from water bodies across the United States, and four million dollars is only a small fraction of what is spent annually on control and mitigation efforts.

Zebra mussels are native to the Caspian and Black Seas, but have become widespread in both Europe and the United States; they were discovered in the Great Lakes in the late 1980s and have since spread rapidly across the United States. The United States National Park Service stated that "[o]nce a population of zebra mussels has become established in a water body, there is very little to be done to remove them. Prevention, therefore, is the best way to keep a water body clean of zebra mussels" (U.S. National Park Service, 2017); hence, early detection of invasive species, such as dreissenid mussels, has become a priority, so that organizations can plan, budget, and install necessary technologies before colonization has occurred (Holser, 2017).

[1] Zebra mussels (*Dreissena polymorpha*) and quagga mussels (*Dreissena rostriformis bugensis*) collectively.

Zebra mussels live between two and five years; they start as microscopic veligers but mature to thumbnail sized adults; they begin reproduction at two years of age, after which, females can release up to one million eggs per year (U.S. National Park Service, 2017). Dreissenid mussel veligers free-swim in the water; often, they travel to uninfested waters on boats or through other aquatic recreational activities, however, sometimes they are moved by nature and travel downstream to uninfested waters. Adult dreissenids attach and colonize hard surfaces in the water, this process of accumulation of adult zebra mussels on rocks, native mussels, docks, boats, or other hard surfaces is referred to as "biofouling," and objects that are in the water for long periods of time become difficult and costly to clean. Once a water body is infested with dreissenid mussels, water supply and delivery facilities, water recreation sites, and other water dependent economies in that body of water become much more expensive to maintain and operate (Bureau of Reclamation, 2019). Dreissenid infestations result not only in economic impacts, but in environmental ones as well. Dreissenid mussels are filter feeders and siphon plankton from the water, which can lead to changes the water body ecosystem by increasing water clarity; a single adult dreissenid can filter about a liter of water per day, which reduces the availability of algae for native mussels and bottom feeding fish (Bureau of Reclamation, 2019). Additionally, "biofouling" can prevent native mussels from moving, feeding, reproducing, or regulating the water system. Several actions, such as the 2017 initiative, *Safeguarding the West from Invasive Species*, by the

Department of the Interior, have been taken to protect water bodies in the western United States from the economic and ecological threats posed by the invasive dreissenid mussels. Early detection of dreissenid mussel species can reduce the economic and ecological repercussions of dreissenid infestations, however there are issues with the available early detection methods.

The established standard for early detection of dreissenids in the western United States is plankton tow sampling for mussel veligers. Using a fine mesh net, water and debris are collected at multiple sampling sites within each water body; the debris from each net collected at the same sampling site on the same day is aggregated and examined, using cross-polarized light microscopy, for the free-swimming veligers. Following the microscopic examination, positive species identification is confirmed using polymerase chain reaction (PCR). This early detection method requires a breeding population, so is limited to the weeks immediately following a spawning event (Nichols, 1996); spawning begins at water temperatures above 10°C for quagga mussels and above 12°C for zebra mussels (McMahon, 1996; Mills et al., 1996). This suggests that veliger availability in northern latitude water bodies is typically limited to warmer months (Sepulveda et al., 2019).

An alternative method for detection of rare, endangered, or invasive species, one growing in popularity, is environmental DNA (eDNA) surveys (Schmelzle and Kinziger, 2016). Environmental DNA methods can detect DNA diffused from the target species from water sampled from a water body. Multiple water samples are collected from each sampling site within a lake, the samples are then analyzed using one of several types of PCR chemistry. Sepulveda, Amberg, and Hanson (2019) suggest the use of eDNA surveys may widen the seasonal sampling window over plankton tow methods, since eDNA does not rely on a breeding population. Environmental DNA surveys are more time and cost effective than traditional sampling methods for species of low abundance (Rees et al., 2014). However, a positive eDNA result does not necessarily mean the target species is present or alive at the site; positive eDNA results can be obtained from "a failed introduction, from external sources, or from field contamination, rather than fresh DNA from mussel colonization" (Sepulveda et al., 2019). One criticism of the detection of dreissenid mussels using eDNA is there is a possibility of obtaining false-positive results; since control efforts for invasive species are costly, there is some hesitation in using eDNA surveys as the sole decision-making tool for the management of invasive species. When these two methods result in conflicting answers, decision making can be even more complicated, since a positive eDNA result only suggests that the DNA of the target species is present, regardless of whether the species is alive or even present at all, but when veligers are detected, positive eDNA results indicate a potential colonization, which is useful to managers (Holser, 2017).

Occupancy models allow the occurrence of a species to be accurately estimated, even when the species is imperfectly detected. For both the plankton tow surveys and the eDNA surveys, there is a non-zero probability of a false negative result. Since plankton tow survey methods are restricted to capturing only mussel veligers, dreissenid mussels can be present in the lake and not be captured in the plankton tow nets at one or more of the sampling sites, either because the veligers are missed with the nets or because there are no free-swimming veligers available in the water; even if the veligers are captured in the net, there is a possibility that they are not detected using cross-polarized light microscopy. Similarly for eDNA methods, dreissenid mussels could be present in the lake, but their DNA could be missed in one or more of the samples from each of the sampling sites; even if dreissenid mussel DNA is present in the sample, it could be missed in the PCR replicate. Replication in the survey design can help researchers learn about the detection probabilities for each of the early detection methods.

The primary goal is to determine how the detection probabilities and false negative rates of dreissenid mussels compare for plankton tow and eDNA survey methods, and explore which environmental factors influence them. Furthermore, recommendations on the sampling effort needed to ensure high detection probabilities for each survey method are of subsequent interest. *A few sentences here about the structure of the rest of the paper.*

2 Data

A sentence or two here about the both data sets.

2.1 eDNA Survey Data

The eDNA data were collected during the spring, summer, and fall of 2019 as part of a study on the detection of dreissenid mussels and Asian clams in northeastern lakes using environmental DNA. There were six lakes surveyed across Maine, New Hampshire, New York, and Vermont; each lake was sampled for dreissenid mussels, Asian clams, or both. The data available are from three of the lakes sampled for dreissenid mussels. There are multiple sites within each lake, from each, five one-liter water samples were collected. Most sites were visited on more than one day throughout the course of the study. Each of the samples were analyzed using a DNA amplification technique, known as digital droplet PCR (ddPCR), on a Biorad Digital Droplet PCR system using dreissenid mussel assays. This PCR technique divides the sample aliquot into micro droplets to be analyzed separately. A field blank of clean lab water accompanied the set of five samples from each site on a given day to ensure no cross contamination in the lab. Droplet replications for dreissenid mussel DNA were negative for all field blank samples; these samples were removed from the data set.

The lakes are coded "BOM," "LG," and "MG," and these names will be used throughout. There are three sites in lake BOM (BOM1, BOM2, BOM3), each of which was sampled five times on three different days (May 28, July 8, and October 21), for a total of 15 samples per site and 45 total samples from lake BOM. There are five sites in lake LG; the first three sites (LG1, LG2, LG3) were sampled five times on both May 28 and July 8, and

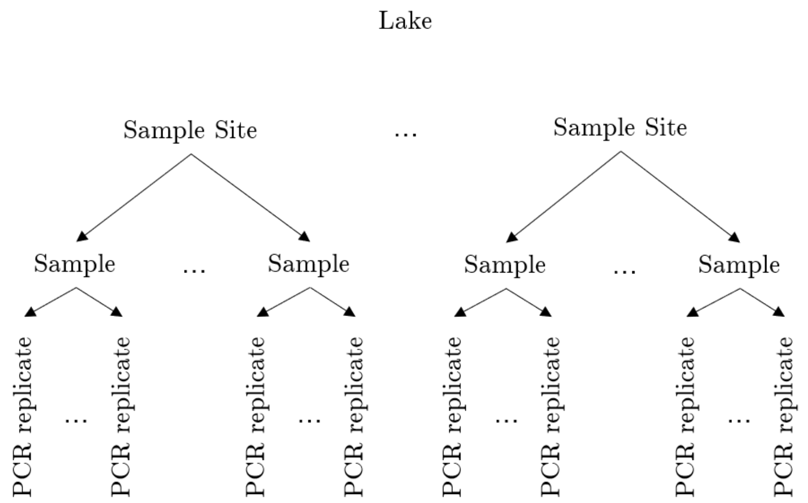

Figure 1: A diagram displaying the hierarchical structure of the eDNA data.

the final two sites (LG4, LG5) were each sampled five times on September 25, for a total of 40 samples from lake LG. There were originally five sites in lake MG, but one (MG2) was inaccessible on the study day, so there are data for four sites in lake MG (MG1, MG3, MG4, MG5), each of which was sampled five times on August 23, for a total of 20 samples from lake MG. In all, there are twelve sites and 105 samples. Each of the samples were divided into 20,000 droplets for ddPCR analysis. Figure 1 illustrates the hierarchical structure of the eDNA data within a lake. Ten sample rows of the data set are displayed in Table 1, where “Positive Droplets” is the number of droplets where dreissenid mussel DNA was amplified and “Concentration” is a proxy measurement of how many droplets were amplified in the sample. Water temperatures (°C) were also recorded at each site on each study day, with the exception of temperatures at sites in lake BOM on October 21.

Table 1: 10 sample rows of the eDNA data.

Lake	Site	Sample	Date	Water Temperature	Concentration	Positive Droplets
BOM	BOM1	BOM1w0528195	5/28/2019	11.0	8.51	60
BOM	BOM1	BOM1w0708194	7/8/2019	25.5	0.81	7
BOM	BOM2	BOM2w1021191	10/21/2019	NA	1.14	8
BOM	BOM3	BOM3w0528193	5/28/2019	6.0	178.72	1154
BOM	BOM3	BOM3w0708193	7/8/2019	24.5	175.99	1388
BOM	BOM3	BOM3w1021191	10/21/2019	NA	2.04	13
LG	LG1	LG1w0528194	5/28/2019	12.0	0.00	0
LG	LG2	LG2w0708194	7/8/2019	23.0	0.00	0
LG	LG2	LG2w0708195	7/8/2019	23.0	0.00	0
MG	MG5	MG5w082319D	8/23/2019	18.3	0.00	0

The proportion of positive droplets per sample range between 0.00000 (0 of 20,000 positive droplets) and 0.16655 (3331 of 20,000 positive droplets) for lake BOM, between 0.00000 and 0.00015 (3 of 20,000 positive droplets) for lake LG, and between 0.00000 and 0.00010 (2 of 20,000 positive droplets) for lake MG. It is noted that samples with three or fewer positive droplets should be treated with caution, as they have potential to be false positives. This is immediately noticeable in lakes LG and MG, as all the values of positive droplets per sample are low for these two lakes. Of the 40 samples from lake LG, 38 of them resulted in no positive droplets, and the remaining two samples resulted in two or three positive droplets per sample. Similarly for lake MG, 16 of the 20 samples resulted in no positive droplets, and the remaining samples resulted in one or two positive droplets per sample. While larger proportions of positive droplets were observed in samples from lake BOM, there are still four samples which resulted in some positive droplets, but three or fewer positive droplets. Based on Figure 2 it is clear that the samples taken on May 28 and July 8 from sites BOM2 and BOM3 tend

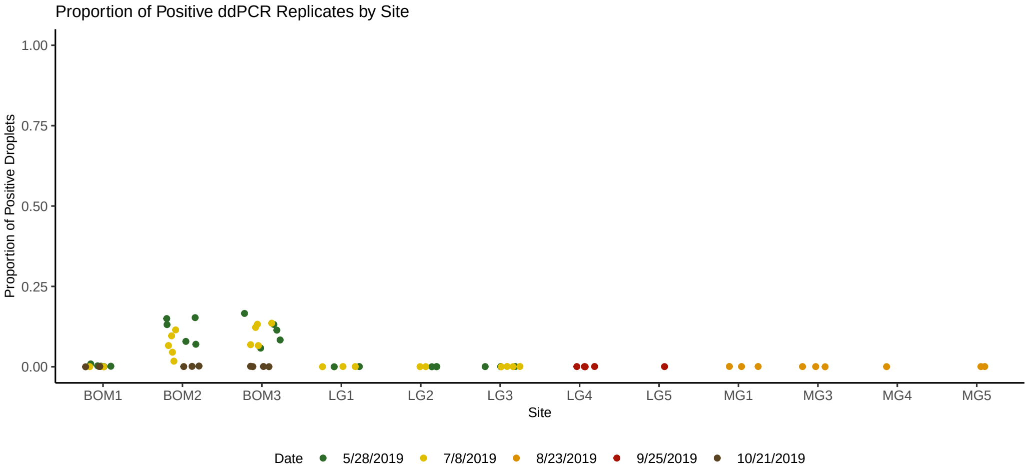

Figure 2: A plot of the proportion of positive droplets from each sample by site, colored by the date on which the sample was taken.

to have the highest proportion of positive droplets, while the samples at the remaining sites across their respective sampling dates, and the samples taken on October 21 at sites BOM2 and BOM3 all have proportions of positive droplets close to 0.

Water temperature could play a role in the availability of dreissenid mussel eDNA in the water, or the probability that the DNA is able to be detected using ddPCR. The water temperatures observed in lake BOM range from 6°C to 25.5°C, but temperatures were not recorded at any of the sites in lake BOM on October 21. The temperatures from a USGS gauge at a nearby lake, were positively adjusted by 2°C to obtain an estimated temperature of 14°C for all sites at lake BOM on October 21, 2019. The observed water temperatures in lake LG range from 12°C to 23°C, and from 17.2°C to 18.9°C in lake MG. Based on Figure 3, it seems that the water temperatures on the same day

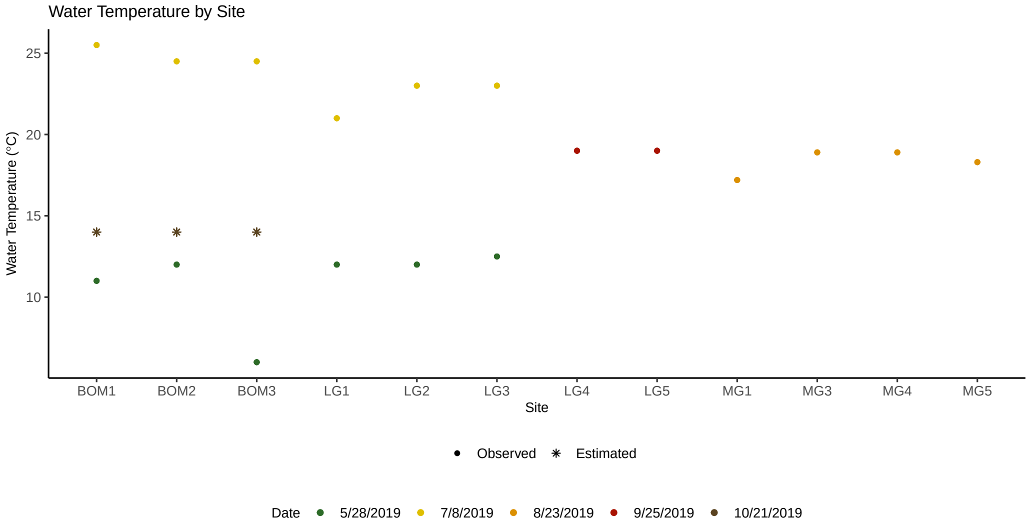

Figure 3: A plot of the water temperatures by site, colored by the date on which the sample was taken. The estimated water temperatures are denoted with stars.

are similar across sites within the same lake. For sites that were visited more than once, it is clear that water temperatures were cooler in the spring but warmed up in the summer.

A sentence here to transition to the plankton tow data

2.2 Plankton Tow Survey Data

Plankton tow data description

- data source: BOR
- region
- number of lakes (number of lakes where dreissenid mussel veligers were

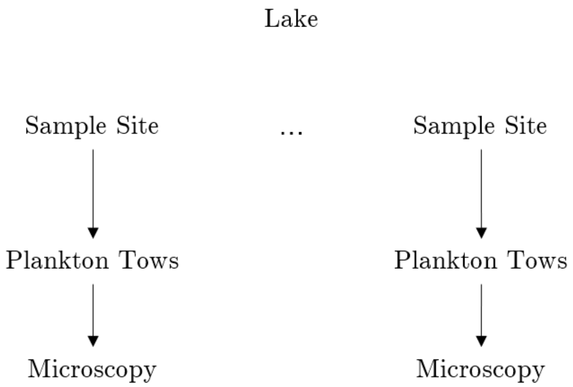

Figure 4: A diagram displaying the structure of the plankton tow data.

detected)

- number of sites
- dates
- potential covariates

Plankton tow data visualization

3 Methods

3.1 Bayesian Modeling Background

In Bayesian statistics, inferences are based on the posterior distributions of the unknown parameters. The posterior distribution is a function of the likelihood of the observed data given a sampling model and some prior belief about the unknown parameters. Let $p(\boldsymbol{y}|\boldsymbol{\theta})$ denote the likelihood for some sampling model, and $p(\boldsymbol{\theta})$ denote the prior distribution on the unknown parameters $\boldsymbol{\theta}$. Then, using Bayes' rule, the posterior distribution, $p(\boldsymbol{\theta}|\boldsymbol{y})$, is:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\Theta} p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}$$

If the resulting posterior distribution is of the same distributional family as the prior distribution for the unknown parameter, then the prior is said to be conjugate for the sampling model. For example, if the observed data are n independent Binomial trails, with some unknown probability of success, θ, and a Beta prior distribution is used for θ, it can be shown that the resulting posterior distribution is also a Beta distribution. The complete model definition is:

$$p(y_1, ..., y_n|\theta) \sim Binomial(n, \theta)$$
$$p(\theta) \sim Beta(\alpha, \beta)$$

Using this, it can be shown that,

$$p(\theta|y_1, ..., y_n) \sim Beta\left(\sum_{i=1}^{n} y_i + \alpha, n - \sum_{i=1}^{n} y_i + \beta\right)$$

Therefore, since the prior distribution and the posterior distribution are of the same distributional family, then the Beta prior is conjugate for the Binomial sampling model.

However, the resulting posterior distribution is not always a named statistical distribution. In these situations, it is impossible to sample directly from the posterior distribution, but iterative sampling mechanisms can be used to approximate the unknown posterior distribution in order to make inferences.

3.1.1 Gibbs Sampler

The most efficient iterative sampling technique available is the Gibbs Sampler. In order to draw samples from an approximation of the joint posterior distribution using a Gibbs Sampler, the full conditional posterior distribution of each of the unknown parameters must have a closed-form solution. The full conditional posterior distribution of a parameter is the distribution of that parameter, conditional on all other unknown parameters, the data, and the prior distributions. Once a full conditional posterior distribution is calculated for each unknown parameter, the iterative sampling can proceed in the following manner.

Gibbs Sampler

Let $\boldsymbol{\theta} = \{\theta_1, ..., \theta_p\}$ denote the vector of unknown parameters and $p(\theta_i|.)$ denote the full conditional distribution of θ_i. The Gibbs Sampler generates the s^{th} iteration as follows:

1. Sample $\theta_1^{(s)} \sim p(\theta_1|\theta_2^{(s-1)}, \theta_3^{(s-1)}, \ldots, \theta_p^{(s-1)}, y_1, \ldots, y_n)$.

2. Sample $\theta_2^{(s)} \sim p(\theta_2|\theta_1^{(s)}, \theta_3^{(s-1)}, \ldots, \theta_p^{(s-1)}, y_1, \ldots, y_n)$.

   $\vdots$

p. Sample $\theta_p^{(s)} \sim p(\theta_p|\theta_1^{(s)}, \theta_2^{(s)}, \ldots, \theta_{p-1}^{(s)}, y_1, \ldots, y_n)$.

This process generates a dependent sequence of $\boldsymbol{\theta}$ vectors for each iteration, which together, converge to the joint posterior distribution, $p(\boldsymbol{\theta}|\boldsymbol{y})$. The Gibbs Sampler is a basic Markov Chain Monte Carlo (MCMC) Algorithm, where the current state only depends on the previous state, and additionally, the results should not depend on the starting values of $\boldsymbol{\theta}^{(0)}$.

One common use of a Gibbs Sampler is for a Normal sampling with unknown mean and variance. With this sampling model, a Normal prior on the mean term and an Inverse-Gamma prior on the variance term enable the use of a Gibbs Sampler to approximate the joint posterior distribution. However, in some cases, such as generalized linear models, semi-conjugate priors or closed-form solutions for the full conditional distributions are not available, and in those cases a Gibbs Sampler cannot be used to sample from the joint posterior distribution of the unknown parameters.

3.1.2 Metropolis-Hastings Algorithm

In cases that do not permit the use of a Gibbs Sampler, a Metropolis-Hastings Algorithm is often used to sample from the target distribution. Unlike a Gibbs Sampler, a Metropolis-Hastings Algorithm requires tuning of the algorithm, but proceeds in the following manner.

Metropolis-Hastings Algorithm

Let $\boldsymbol{\theta}^{(s)}$ be the current set of parameter estimates, $\boldsymbol{\theta}^{*}$ denote a new proposed set of parameters, and $J(\boldsymbol{\theta}^{*}|\boldsymbol{\theta}^{(s)})$ denote the proposal distribution, which is usually a random walk distribution. For example, $J(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(s)}) \sim MVN(\boldsymbol{\theta}^{(s)}, \gamma^2 \boldsymbol{I}_p)$ where p is the number of unknown parameters, and γ is thought of as the step size, or the average distance the proposed set of parameters falls from the current set. To complete an iteration:

1. Sample $\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(s)} \sim J(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(s)})$.
2. Calculate the acceptance ratio $r = \frac{p(y|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)}{p(y|\boldsymbol{\theta}^{(s)})p(\boldsymbol{\theta}^{(s)})}$.
3. Set $\boldsymbol{\theta}^{(s+1)}$:
   - If $r \geq 1$ then the proposed set, $\boldsymbol{\theta}^*$, is more attractive than the current set, $\boldsymbol{\theta}^{(s)}$, so $\boldsymbol{\theta}^{(s+1)} = \boldsymbol{\theta}^*$.
   - If $r < 1$ then the proposed set, $\boldsymbol{\theta}^*$, is less attractive than the current set, $\boldsymbol{\theta}^{(s)}$, however the relative frequency of samples of $\boldsymbol{\theta}^*$ to $\boldsymbol{\theta}^{(s)}$ should be r, so with probability r, $\boldsymbol{\theta}^{(s+1)} = \boldsymbol{\theta}^*$.

The Metropolis Algorithm is a generalization of the Metropolis-Hastings Algorithm in which the proposal distribution is symmetric, meaning that $J(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(s)}) = J(\boldsymbol{\theta}^{(s)}|\boldsymbol{\theta}^*)$. The step size portion of the algorithm requires tuning and trace plots or other convergence tools should be used to ensure that the algorithm has efficiently explored the entire parameter space and converged to the true joint posterior distribution.

3.1.3 Bayesian Hierarchical Modeling

Oftentimes data is collected in a hierarchical structure, that is data with a multilevel structure, such as students within classes, or sampling sites within a larger area of interest. For example, with two levels, the responses are independent observations from group j and follow some distribution with unknown parameter θ_j, meaning that $y_{1,j}, \ldots, y_{n_j,j}|\theta_j \sim p(y|\theta_j)$; but the unknown parameters in each group are related to each other, such that they are independent samples from some distribution with parameter ϕ, such that $\theta_1, \ldots, \theta_m|\phi \sim p(\theta|\phi)$. With this representation, $p(y|\theta)$ is the variability among measurements within a group, and $p(\theta|\phi)$ is the sampling variability across groups. In order to completely specify the model, a prior distribution, $\phi \sim p(\psi)$, is needed for ϕ. These models can be adapted to include covariates at each level, or to describe a more complicated structure with more levels.

3.2 Occupancy Models

In ecological studies, there are several different state variables which may be of interest: abundance, vital rates, and occupancy are a few examples. Both mark-recapture and occupancy studies can be used to learn about the previously mentioned state variables, but there are advantages and limitations with each study type. In general, mark-recapture surveys are studies where individuals are capture or observed, given a unique mark, their identities recorded, and finally released; on subsequent occasions, both marked and unmarked individuals are captured, their identities recorded, unmarked individuals are marked, and they are all released. From this, capture histories, or encounter histories (typically a series of 0's and 1's) are recorded for each individual that is captured over the course of the study. There are several variations of mark-recapture studies, in which apparent survival, abundance, or both can be estimated depending on the study design. However, mark-recapture methods cannot be used when individuals of a species cannot be marked or uniquely identified. In this scenario, occupancy of a particular species can be recorded at each occasion. Occupancy methods are useful when studying a species over a large spatial scale for many years, particularly when the sampling sites are large relative to the species of interest. Though there is a loss in information about individuals of a species when occupancy studies are used, sometimes researchers are interested in the range of a species or the spread of invasion and therefore occupancy is the state variable of interest, and occupancy studies are most appropriate. Occupancy surveys are often used to track the mitigation efforts or spread of invasive species, but are plagued with imperfect detection probabilities.

Occupancy models are a way to move past logistic regression for occupancy data in order to account for imperfect detection of the target species. However, multiple samples per site must be taken in order to learn about detection probabilities. For example, suppose five samples are taken from a site and two return a positive result, indicating that the species is present at the site, but the species was missed in three of the five samples, which provides

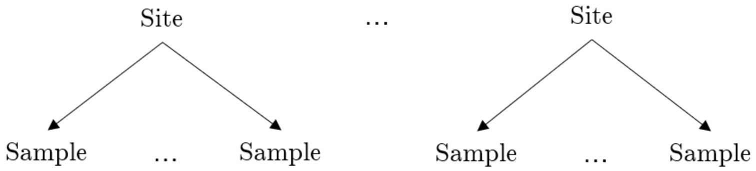

Figure 5: A diagram displaying the structure of a traditional occupancy study.

information about the detectability of the species. There are several different types of occupancy models, but in the traditional model the parameters are as follows: ψ_i denotes the probability that the target species is present at the i^{th} site, and p_{ij} denotes the probability that the target species is detected on the j^{th} survey, given the species occupies site i. Figure 5 illustrates the structure of a traditional occupancy study, where each sample results in a binary response.

For more illusive species, the samples from each site are split into subsamples, or replicates, and analyzed on a smaller scale. This introduces a hierarchical structure in the data creating dependencies which need to be accounted for (MacKenzie et al., 2018). One modeling approach to these types of data is the hierarchical latent state multi-scale occupancy model defined by Dorazio and Erickson which accounts for false negatives at the site, sample, and replicate levels (Dorazio and Erickson, 2018). At the site level, Z_i is defined as the presence ($Z_i = 1$) or absence ($Z_i = 0$) of the target species at the i^{th} site ($i = 1, \ldots, M$). Then, $Z_i \sim Bernoulli(\psi_i)$, where ψ_i denotes the probability that the target species occupies the i^{th} site. At the sample level, A_{ij} is defined as the presence or absence of the target species in the j^{th} sample from the i^{th} site ($j = 1, \ldots, J_i$). Then, $A_{ij}|z_i \sim Bernoulli(z_i\theta_{ij})$, where z_i is the latent occupancy state for the i^{th} site, and θ_{ij} is the conditional probability that the target species is present in the j^{th} sample from the i^{th} site, given the target species is present at the site. At the replicate level, Y_{ijk} is defined as whether the target species is detected or not in the k^{th} replicate of the j^{th} sample collected at the i^{th} site ($k = 1, \ldots, K_{ij}$). It follows that, $Y_{ijk}|a_{ij} \sim Bernoulli(a_{ij}p_{ijk})$, where a_{ij} is the latent occupancy state for sample j from site i, and p_{ijk} is the conditional the probability that the target species is detected in the k^{th} replicate of the j^{th} sample collected at the i^{th} site, given the target species is present in that sample. If p_{ijk} does not differ among the replicates and the replicates are statistically independent, the replicate level can be simplified in the following manner: $Y_{ij}|a_{ij} \sim Binomial(K_{ij}, a_{ij}p_{ij})$, where $Y_{ij} = \sum_{k=1}^{K_{ij}} Y_{ijk}$, and p_{ij} is the conditional probability of detection in each replicate of the j^{th} sample collected at the i^{th} location, given that the target species is present in that sample. One assumption of these models is that there is no unmodeled heterogeneity in the detection and occupancy probabilities, so covariates can be included in each level of the model to account for differences between individual sites, samples, or replicates. At the site level, $\boldsymbol{x}_i$ are the site specific covariates and $\boldsymbol{\beta}$ are corresponding regression parameters for ψ_i. At the sample level, $\boldsymbol{w}_{ij}$ are the sample specific covariates and $\boldsymbol{\alpha}$ are the corresponding regression parameters for θ_{ij}. At the replication level, where there are no replicate level covariates, $\boldsymbol{v}_{ij}$ are the sample specific covariates and $\boldsymbol{\delta}$ are the corresponding regression parameters for p_{ij}. Site specific covariates can also be included for θ_{ij} and p_{ij}. Another assumption of these models is that the species is not misidentified, however, if false positive results are a possibility, there are alternative methods for dealing with that. Additionally, each site is assumed to be closed to changes in occupancy status over the sampling period, meaning that if a site is occupied at the beginning of the study, it is assumed to remain occupied throughout. The final assumption of these models is that the detection of the species is independent for each replicate.

3.3 Implementation

There are several options for fitting occupancy models; the methods used to analyze the eDNA survey data and the plankton tow data, as well as some alternative methods are described below.

3.3.1 `msocc` package

The multi-scale occupancy model described above (Dorazio and Erickson, 2018) can be easily fit in a Bayesian framework with the `msocc` package (Stratton, 2020). This package implements the Pólya Gamma data augmentation strategy described by Polson et. al (2013) to enable the use of a Gibbs Sampler in order to speed up the computation and avoid the tuning necessary for a Metropolis Algorithm. The model is specified in a familiar way to users of `R` at each level of the hierarchy in the model fitting function `msocc_mod()`.

To make the package more computationally efficient, there is an option to use the Beta-Binomial sampler when possible. Built into the package are posterior summary tools (`posterior_summary()`) and a function for generating credible interval plots (`cred_plot()`) at all levels of the hierarchy.

3.3.2 **implementation method for plankton tow**

3.3.3 Alternative Options

There are several alternative options for fitting occupancy models. For example, the `eNDAoccupancy` (Dorazio and Erickson, 2018) package fits the Bayesian multi-scale occupancy model defined above using a Metropolis-Hastings algorithm, which is less efficient and requires tuning. These models can also be written in `STAN` or `JAGS` and run in `R` using the packages `rstan` (Stan Development Team, 2019) and `rjags` (Plummer, 2019) respectively, but again these methods are less computationally efficient than the methods described above. There are various frequentist methods for fitting occupancy models as well. A commonly used `R` package for fitting frequentist occupancy models is `unmarked` (Fiske and Chandler, 2011). Ecologists often utilize Program MARK, or access its functionalities through the R package `RMark` (Laake, 2013) to fit occupancy models, and other various models popular in ecology.

4 Analysis

4.1 Analysis of eDNA Data

analysis and results

```
eDNA_m1 <- msocc_mod(eDNA_wide,
                     site = list(model = ~ 1, cov_tbl = site.cov),
                     sample = list(model = ~ 1, cov_tbl = sample.cov),
                     rep = list(model = ~ 1, cov_tbl = rep.cov),
                     seed = 03202020)

burnin <- 100

# site summary
## posterior_summary function does not work b/c matrix too large
psi.post_m1 <- eDNA_m1$psi[-(1:burnin)]
summary(psi.post_m1)
quantile(psi.post_m1, c(0.025, 0.975))
plot(psi.post_m1, type = "l") # REDO THIS IN GGPLOT IF KEEPING IT

# sample summary
## posterior_summary function does not work b/c matrix too large
theta.post_m1 <- eDNA_m1$theta[-(1:burnin)]
```

```
summary(theta.post_m1)
quantile(theta.post_m1, c(0.025, 0.975))
plot(theta.post_m1, type = "l") # REDO THIS IN GGPLOT IF KEEPING IT

# rep summary
## posterior_summary function does not work b/c matrix too large
p.post_m1 <- eDNA_m1$p[-(1:burnin)]
summary(p.post_m1)
quantile(p.post_m1, c(0.025, 0.975))
plot(p.post_m1, type = "l") # REDO THIS IN GGPLOT IF KEEPING IT
```

```
eDNA_m2 <- msocc_mod(eDNA_wide,
                     site = list(model = ~ lake, cov_tbl = site.cov),
                     sample = list(model = ~ date, cov_tbl = sample.cov),
                     rep = list(model = ~ 1, cov_tbl = rep.cov),
                     seed = 03202020)

burnin <- 100

# site summary
## posterior_summary function does not work b/c matrix too large
X <- eDNA_m2$model.info$X
beta <- eDNA_m2$beta
```

```
eta <- beta %*% t(X)
psi <- exp(eta)/(1 + exp(eta))

psi.mcmc <- psi[-c(1:burnin), ]

mean <- apply(psi.mcmc, 2, mean)
median <- apply(psi.mcmc, 2, median)
quantiles <- apply(psi.mcmc, 2, quantile, probs = c(0.025, 0.975))
sum_tbl <- eDNA_m2$model.info$df %>%
  dplyr::select(-sample, -rep) %>%
  dplyr::distinct() %>%
  dplyr::mutate(median = median,
                mean = mean,
                lwr = quantiles[1, ],
                upr = quantiles[2, ])
sum_tbl

# sample summary
## posterior_summary function does not work b/c matrix too large
### theta.post_m2
W <- eDNA_m2$model.info$W
alpha <- eDNA_m2$alpha
```

```r
nu <- alpha %*% t(W)
theta <- exp(nu)/(1 + exp(nu))

theta.mcmc <- theta[-c(1:burnin), ]

mean <- apply(theta.mcmc, 2, mean)
median <- apply(theta.mcmc, 2, median)
quantiles <- apply(theta.mcmc, 2, quantile, probs = c(0.025, 0.975))
sum_tbl <- eDNA_m2$model.info$df %>% dplyr::select(-rep) %>%
  dplyr::distinct() %>%
  dplyr::mutate(median = median,
                mean = mean,
                lwr = quantiles[1, ],
                upr = quantiles[2, ])
sum_tbl

# rep summary
## posterior_summary function does not work b/c matrix too large
p.post_m2 <- eDNA_m2$p[-(1:burnin)]
summary(p.post_m2)
quantile(p.post_m2, c(0.025, 0.975))
plot(p.post_m2, type = "l") # REDO THIS IN GGPLOT IF KEEPING IT
```

```
eDNA_m3 <- msocc_mod(eDNA_wide,
                     site = list(model = ~ site, cov_tbl = site.cov),
                     sample = list(model = ~ 1, cov_tbl = sample.cov),
                     rep = list(model = ~ 1, cov_tbl = rep.cov),
                     seed = 03202020)

burnin <- 100



# site summary
## posterior_summary function does not work b/c matrix too large
X <- eDNA_m3$model.info$X
beta <- eDNA_m3$beta

eta <- beta %*% t(X)
psi <- exp(eta)/(1 + exp(eta))

psi.mcmc <- psi[-c(1:burnin), ]

mean <- apply(psi.mcmc, 2, mean)
median <- apply(psi.mcmc, 2, median)
quantiles <- apply(psi.mcmc, 2, quantile, probs = c(0.025, 0.975))
```

```
sum_tbl <- eDNA_m3$model.info$df %>%
  dplyr::select(-sample, -rep) %>%
  dplyr::distinct() %>%
  dplyr::mutate(median = median,
                mean = mean,
                lwr = quantiles[1, ],
                upr = quantiles[2, ])
sum_tbl
```

4.2 Analysis of Plankton Tow Data

analysis and results

5 Discussion

5.1 Further Investigations

6 References

Bontke, J. (2020). City spends $4 million on liquid compound to stop spread of zebra mussels. `https://cbsaustin.com/news/local/city-spends-4-million-on-liquid-compound-to-stop-spread-of-zebra-mussels`. Date accessed: January 29, 2020.

Bureau of Reclamation (2019). Invasive mussels. `https://www.usbr.gov/mussels/index.html`. Date accessed: Feburary 11, 2020.

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

Fiske, I. and Chandler, R. (2011). *unmarked: An R Package for Fitting Hierarchical Models of Wildlife Occurrence and Abundance.*

Holser, D. M. (2011). Early detection of dreissenid species: Zebra/Quagga mussels in water systems. *Aquatic Invasions*, 8(2):217 – 222.

Holser, D. M. (2017). Where is the body? Dreissenid mussels, raw water testing, and the real value of environmental DNA. *Management of Biological Invasions*, 8(3):335 – 341.

Laake, J. (2013). *RMark: An R Interface for Analysis of Capture-Recapture Data with MARK.*

MacKenzie, D. I., Nichols, J. D., Royle, J. A., and Pollock, K. H. (2018). *Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence*. Academic Press.

McMahon, R. F. (1996). The physiological ecology of the zebra mussel, *Dreissena polymorpha*, in North America and Europe. *American Zoologist*, 36(3):339 – 363.

Mills, E. L., Rosenberg, G., Spidle, A. P., Ludyanskiy, M., Pligin, Y., and May, B. (1996). A review of the biology and ecology of the quagga mussel (*Dreissena bugensis*), a second species of freshwater dreissenid introduced to North America. *American Zoologist*, 36(3):271 – 286.

Nichols, S. J. (1996). Variations in the reproductive cycle of *Dreissena polymorpha* in Europe, Russia, and North America. *American Zoologist*, 36(3):311 – 325.

Plummer, M. (2019). *rjags: Bayesian Graphical Models using MCMC*. R package version 4-10.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rees, H. C., Maddison, B. C., Middleditch, D. J., Patmore, J. R., and Gough, K. C. (2014). The detection of aquatic animal species using environmental dna – a review of edna as a survey tool in ecology. *Journal of Applied Ecology*, 51:1450 – 1459.

Schmelzle, M. C. and Kinziger, A. P. (2016). Using occupancy modelling to compare environmental DNA to traditional field methods for regional-scale monitoring of an endangered aquatic species. 16:1 – 14.

Sepulveda, A. J., Amberg, J. J., and Hanson, E. (2019). Using environmental DNA to extend the window of early detection for dreissenid mussels. *Managment of Biological Invasions*, 10(2):342 – 358.

Stan Development Team (2019). *RStan: the R interface to Stan.* R package version 2.19.2.

Stratton, C. (2020). *msocc: Package for Fitting and Analyzing Computationally Efficient Multi-scale Occupancy Models.* R package version 1.1.0.

U.S. National Park Service (2017). Invasive Zebra Mussels. `https://www.nps.gov/articles/zebra-mussels.htm`. Date Accessed: February 11, 2020.

7 Appendix - R Code

```
# packages used
library(car)
library(dplyr)
library(tidyr)
library(kableExtra)
library(ggplot2)
library(gridExtra)
library(tm)
library(readxl)
## library(devtools)
## devtools::install_github("StrattonCh/msocc")
library(msocc)


# load eDNA data
eDNA <- read.csv(
  "C:/Users/mwind/OneDrive/Writing Project_EXTRA/eDNA.csv")


# rename lakes
levels(eDNA$Lake) <- c("BOM", "LG", "MG")


# remove field blank samples
eDNA <- eDNA %>%
  filter(Site != "tb")
eDNA$Site <- droplevels(eDNA$Site)


# reorder dates in chronological order
eDNA$Date.Collected <- factor(eDNA$Date.Collected,
                          levels = c("5/28/2019", "7/8/2019",
                                     "8/23/2019", "9/25/2019",
                                     "10/21/2019"))
```

```
# generate table of 10 sample rows of eDNA data
set.seed(03142020)
knitr::kable(some(eDNA), 'latex', booktabs = T, linesep = "",
             caption = "\\label{tab:eDNA_data}
             10 sample rows of the eDNA data.",
             align = 'c', row.names = F,
             col.names = c("Lake", "Site", "Sample", "Date",
                           "Water Temperature", "Concentration",
                           "Positive Droplets")) %>%
  kable_styling(latex_options =
                  c("scale_down", "hold_position"))


# BOM
eDNA %>% filter(Lake == "BOM") %>% summary


# LG
eDNA %>% filter(Lake == "LG") %>% summary


# MG
eDNA %>% filter(Lake == "MG") %>% summary


# number of samples with 0 positive droplets, by lake
eDNA %>%
  group_by(Lake) %>%
  count(Positive.Droplets == 0)


# number of samples with (0, 3] positive droplets
eDNA %>%
  group_by(Lake) %>%
  count(Positive.Droplets <= 3  & Positive.Droplets > 0)
```

```
# number of samples with (3, 10] positive droplets
eDNA %>%
  group_by(Lake) %>%
  count(Positive.Droplets > 3 & Positive.Droplets <= 10)


# visualization of eDNA data
eDNA.df <- data.frame(eDNA)

date.cols <- c("5/28/2019" = "#2d6b28",
               "7/8/2019" = "#dfbf00",
               "8/23/2019" = "#db9002",
               "9/25/2019" = "#a91303",
               "10/21/2019" = "#5a4320")

# eDNA positive replicate plot
eDNA.df %>%
  ggplot(aes(x = Site,
             y = Positive.Droplets/20000,
             colour = Date.Collected)) +
  labs(title = 'Proportion of Positive ddPCR Replicates by Site',
       x = 'Site',
       y = 'Proportion of Positive Droplets',
       color = 'Date') +
  ylim(c(0, 1)) +
  geom_point(position = position_jitter(width = 0.25,
                                        height = 0.001,
                                        seed = 03172020)) +
  scale_colour_manual(values = date.cols) +
  theme_bw() +
  theme(title = element_text(size = 9),
        panel.border = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "black"),
        legend.position = 'bottom')
```

```
# replace missing water temps with 14C
eDNA$Water.Temp <- replace(eDNA$Water.Temp,
                           is.na(eDNA$Water.Temp), 14)

est.water.temp <- as.numeric(eDNA$Date.Collected == "10/21/2019"
                             & eDNA$Lake == "BOM")

eDNA.df2 <- data.frame(eDNA,
                       Est.Water.Temp = factor(est.water.temp))

levels(eDNA.df2$Est.Water.Temp) <- c("Observed", "Estimated")

# eDNA water temperature plot
eDNA.df2 %>%
  ggplot(aes(x = Site,
             y = Water.Temp,
             colour = Date.Collected)) +
  labs(title = 'Water Temperature by Site',
       x = 'Site',
       y = expression(paste('Water Temperature (', degree,
                            'C)')),
       color = 'Date') +
  geom_point(aes(shape = Est.Water.Temp)) +
  scale_colour_manual(values = date.cols) +
  scale_shape_manual(values = c(16, 8)) +
  theme_bw() +
  guides(shape = guide_legend(title = NULL)) +
  theme(title = element_text(size = 9),
        panel.border = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "black"),
        legend.position = 'bottom',
        legend.box = 'vertical')


# load tow data
```

```
tow <- read_excel(
  "C:/Users/mwind/OneDrive/Writing Project_EXTRA/tow.xlsx",
  sheet = 3)


# format eDNA data for use in msocc package
## USING ALL POSITIVE DROPLETS
eDNA_wide <- data.frame(site = eDNA$Site,
                        sample = eDNA$Sample.ID)
rep.names <- paste("PCR", 1:20000, sep = "")
eDNA_wide[, rep.names] <- NA

set.seed(03202020)
for(i in 1:nrow(eDNA_wide)){
  samp.id <- sample(1:20000, eDNA$Positive.Droplets[i])
  for(j in 1:20000){
  if(j %in% samp.id){
    eDNA_wide[i, j + 2] <- 1
    }
  }
}

eDNA_wide[is.na(eDNA_wide)] <- 0

# site level covariates
site.cov <- data.frame(site = levels(eDNA$Site))
site.cov[, 2] <- removeNumbers(as.character(site.cov$site)) %>%
  factor()
names(site.cov)[2] <- "lake"

# sample level covariates
sample.cov <- data.frame(site = eDNA$Site,
                         sample = eDNA$Sample.ID,
                         date = eDNA$Date.Collected)
## water temp too but there are missing values, so need to impute... later

# replicate level covariates
```

```
rep.cov <- data.frame(site = eDNA$Site,
                      sample = eDNA$Sample.ID,
                      date = eDNA$Date.Collected)
## water temp too but there are missing values, so need to impute... later
```