Montana State University

Department of Mathematical Sciences

Writing Project

Occupancy Modeling of ddPCR Data for the Early Detection of Dreissenid Mussels using Environmental DNA

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

Zebra and quagga mussels are highly invasive species that have several negative impacts, both economically and environmentally. Once a water body is invaded, the dreissenid mussels attach to substrate or native mussels; as a result, water dependent economies become much more expensive to maintain and operate and native mussels are not able to properly regulate the water system. Currently, there are no eradication methods for established dreissenid mussel populations in open water systems, so prevention is important. Early detection has become a priority for land managers and researchers for several reasons. There are two survey methods used for early detection of dreissenid mussels: plankton tow surveys and environmental DNA surveys. In 2019, a study was conducted across several lakes in the northeastern United States using environmental DNA surveys of the water; the samples were analyzed with a DNA amplification technique known as digital droplet PCR (ddPCR). For this exploration, the data are modeled in a traditional occupancy model framework; the analysis is followed by a discussion of the results. The exploration concludes with a discussion about future work and improvements to the modeling framework.

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 merely 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 first discovered in the Great Lakes in the late 1980s and have since spread rapidly across the continental U.S. 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 in order for organizations to plan, budget, and install necessary technologies before colonization has occurred (Holser, 2017). If detected before they are able to establish, there is potential for eradication of the dreissenid mussels with methods such as Zequanox® or EarthTec®. Additionally, early detection can aid in containment efforts,

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

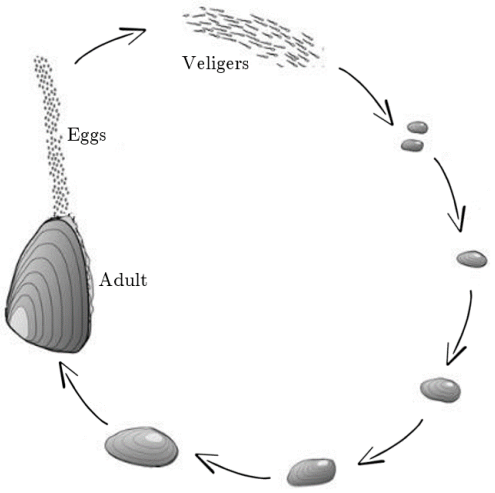

Figure 1: A diagram depicting the life cycle of dreissenid mussels.

limiting the potential for invasive mussel spread to other waters if a water body is known to be infested.

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). A diagram depicting the life cycle of dreissenid mussels is shown in Figure 1. 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 colonized by adult 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). For example, after dreissenid mussels were detected in Montana's Tiber Reservoir in 2016, with suspected detections in Canyon Ferry Reservoir as well, the state of Montana highlighted the economic impacts that a full dreissenid mussel colonization could have on the state; in January 2019, the Montana Department of Natural Resources and Conservation published a report titled *Enumeration of Potential Economic Costs of Dreissenid Mussels Infestation in Montana*. The report states that if dreissenid mussels were to infest every lake and river in Montana, it could cost the state more than 230 million dollars annually in mitigation costs, as well as lost revenue (Nelson, 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 in 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 (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 mussels 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) chemistry. 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, growing in popularity, for detection of rare, endangered, or invasive species, is environmental DNA (eDNA) surveys (Schmelzle and Kinziger, 2016). Environmental DNA methods can detect DNA diffused by the target species in 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 different types of PCR chemistry. Sepulveda, Amberg, and Hanson (2019) suggest the use of eDNA surveys may widen the seasonal sampling window of 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 indicate that 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 dreissenid mussels.

Occupancy models allow the uncertainties that arise in the detection process to be incorporated in the estimation of the occurrence probabilities. For both the plankton tow and 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; further, 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. For both of the survey methods, there is also potential for false positive results, but the initial focus of this exploration will be on the imperfect detection that arises from false negatives. Multiple visits to the same site (either spatially or temporally) can help researchers learn about the detection probabilities for each of the early detection methods.

Land managers are often interested in the probability of occupancy in a given lake, whereas researchers are typically interested in detection probabilities. Furthermore, the researchers are also interested in recommendations on the number of samples necessary to ensure a high probability of at least one detection using digital droplet PCR (ddPCR) when dreissenid mussel eDNA is present at the site. The remainder of this exploration proceeds as follows: Section 2 briefly describes the data collection process and provides summary and visualization of the raw data; Section 3 contains a brief description of Bayesian statistical methods, a discussion of occupancy models and multi-scale occupancy models, and a few implementation methods for fitting the previously described models in `R`; Section 4 describes the analysis and the various models fit to the eDNA data; Section 5 covers the results from the models and their ecological implications; and finally, Section 6 briefly

describes future work to be done on the project.

2 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 (*Corbicula fluminea*), 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 separately analyzed. 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; 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 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 sample was filtered, and a portion of each of the filtered samples was divided into thousands of droplets for ddPCR analysis. Figure 2 illustrates the hierarchical structure 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 the copies of target DNA per microliter of extraction (copies/μL). Water temperatures (°C) were also recorded at each sampling 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

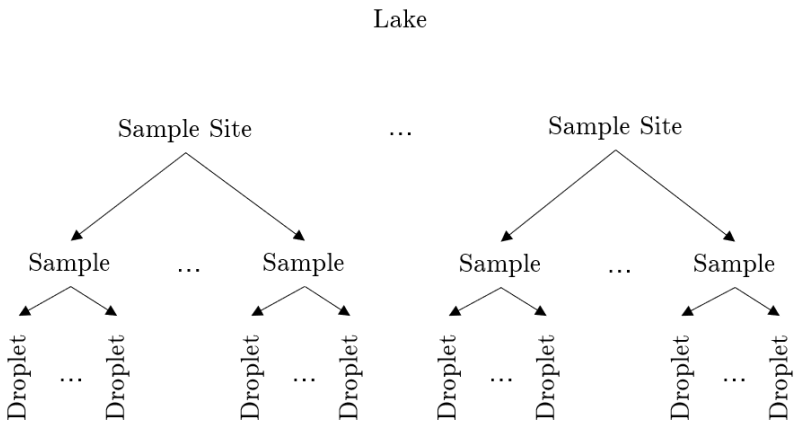

Figure 2: A diagram displaying the hierarchical structure of the eDNA data.

The number of positive droplets per sample range between 0 and 3331 for lake BOM, between 0 and 3 for lake LG, and between 0 and 2 for lake MG. Based on Figure 3 it is clear that the samples taken on May 28 and July 8 from sites BOM2 and BOM3 tend to have the highest numbers 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 numbers positive droplets near 0. It is noted that samples with fewer than three 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

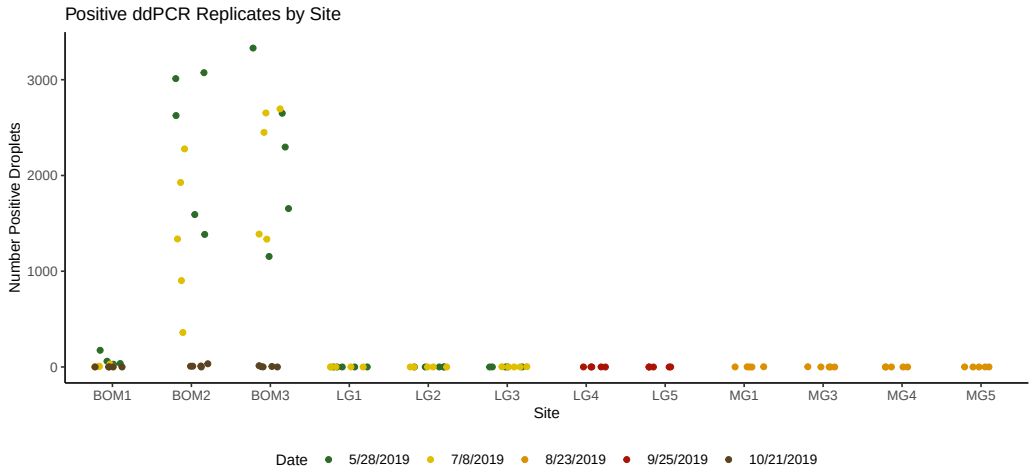

Figure 3: A plot of the positive droplets from each sample by site, colored by the date on which the sample was taken.

two positive droplets per sample. While larger numbers of positive droplets were observed in samples from lake BOM, there are still three samples which resulted in some, but fewer than three, positive droplets.

Somewhat related to the number of positive droplets in a sample, the concentrations in the study range between 0.00 and 733.80 copies/μL. Based on the top plot in Figure 4, it is clear that the sites with the highest concentrations occur in samples from sites BOM2 and BOM3, with average concentrations of 206 and 235 copies/μL in each of the respective sites. Sites LG1, LG2, LG4, LG5, and MG4 all have average concentrations of 0 copies /μL. The samples from the remaining sites tend to have low, but non-zero concentrations; the bottom plot in Figure 4 displays the samples by site for non-zero concentrations below 10 copies/μL.

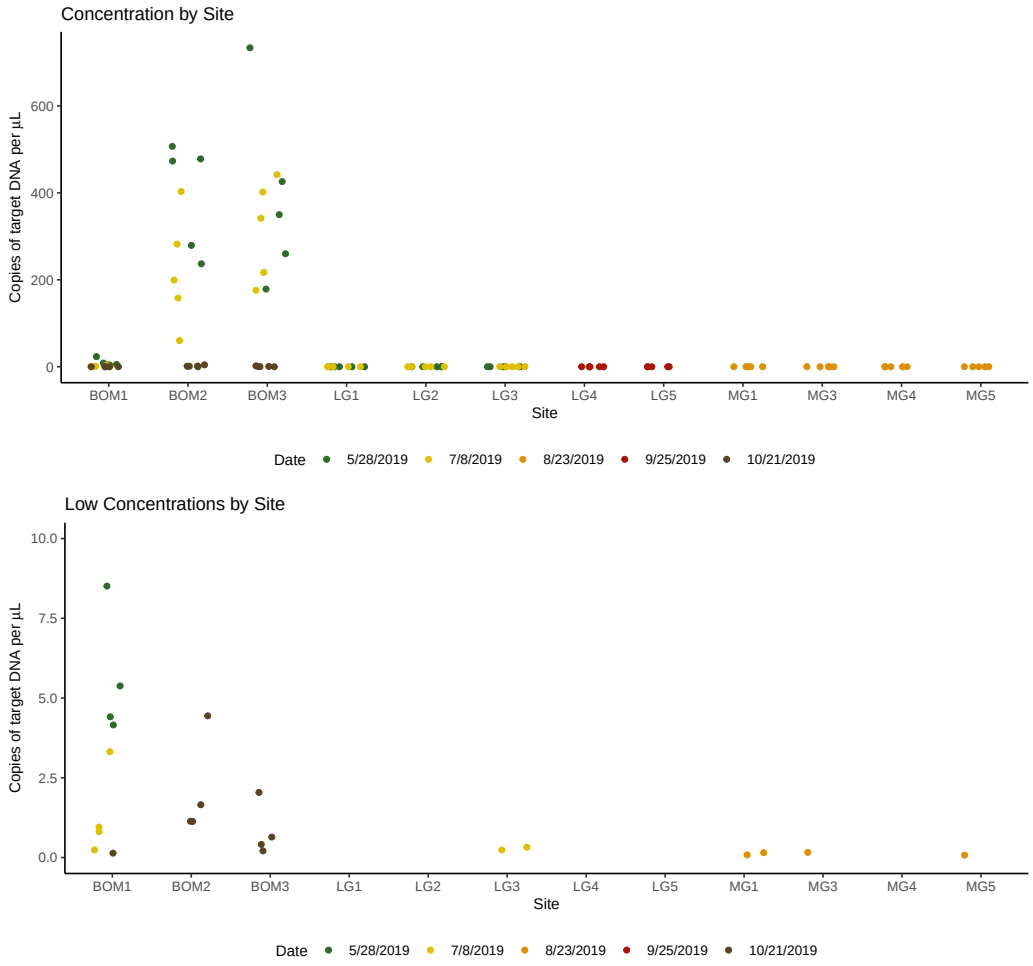

Figure 4: A plot displaying the number of copies of target DNA per μL per sample by site (top), and a plot displaying the number of copies of target DNA per μL for samples with low, but non-zero, concentrations (bottom).

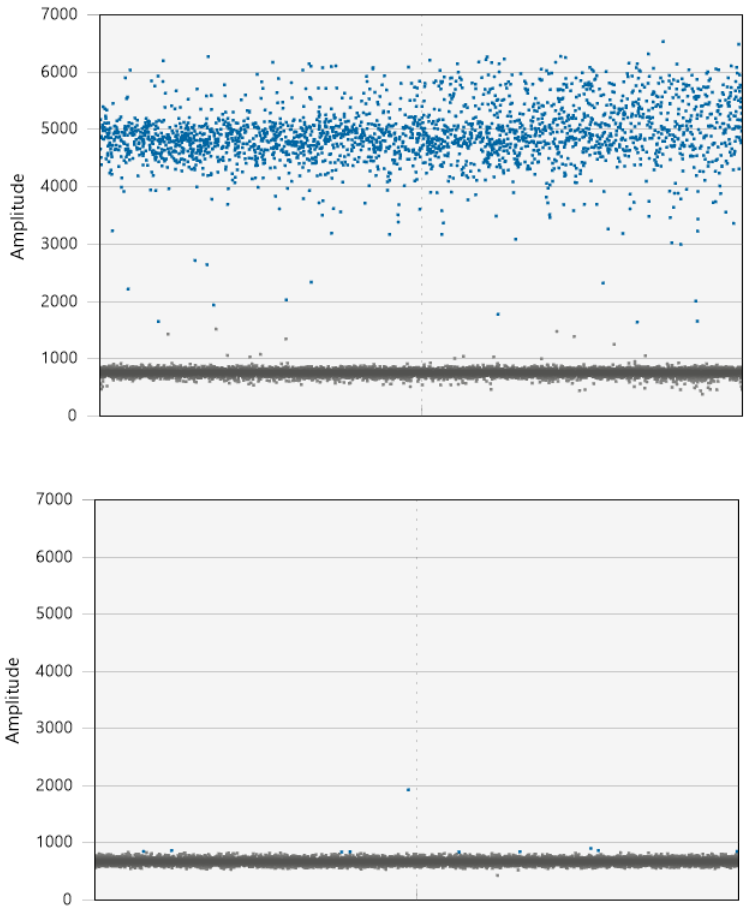

Figure 5: The droplet amplitudes for the second sample from site BOM3 on 7/8/2019 (top) and for the third sample from site LG5 on 9/25/19 (bottom).

Since the number of positive droplets per sample is based on a subjective choice of amplitude cutoff to distinguish between positive and negative droplets, the number of positive droplets in a sample is arbitrary. The top plot in Figure 5 display the amplitudes of each droplet from the second sample from site BOM3 on 7/28/2019 and the bottom plot displays the amplitudes of each droplet from the third sample from site LG5 on 9/25/2019; the droplets with high amplitudes tend to suggest positive results. The band of blue points with amplitudes around 5000 in the top plot in Figure 5 are likely positive droplets, whereas the band of grey points with amplitudes less than 1000 in the same plot are likely negative droplets, however where to draw the line to distinguish between positive and negative droplets will change the number of positive droplets, perhaps dramatically in some samples. In the bottom plot in Figure 5, the question becomes whether the droplet with an amplitude around 2000 should be considered positive or negative. To avoid basing the results on an arbitrary number of positive droplets, these data will be analyzed at the sample level. To limit the possibility of false positive detections for these analyses, to be considered a positive sample, the recommended threshold of three or more positive droplets will be used, therefore samples with three or more positive droplets will be considered positive samples. The top plot in Figure 6 displays the proportion of positive samples by site if all samples with positive droplets are treated as positive samples, and the bottom plot displays the proportion of positive samples by site using the above threshold to limit the positive sites to those with three or more positive

droplets; the plots aid in understanding the changes in the proportion of positive samples from each site when the threshold is used. The application of the threshold changes the proportion of positive samples in six of the twelve sites. When the threshold is used, there are no positive samples observed in eight of the twelve sites.

Water temperature could impact the probability that dreissenid mussel eDNA is able to be detected using ddPCR; spawning events are likely to increase eDNA in the water (10°C for quagga mussels and above 12°C for zebra mussels), and the persistence of eDNA can increase in lower water temperatures which would lead to an increase in detection probability (Gingera et al., 2017). 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 7, it seems that the water temperatures on the same day are similar across sites within the same lake. For sites that were visited more than once, water temperatures were cooler in the spring but warmed up in the summer, as expected.

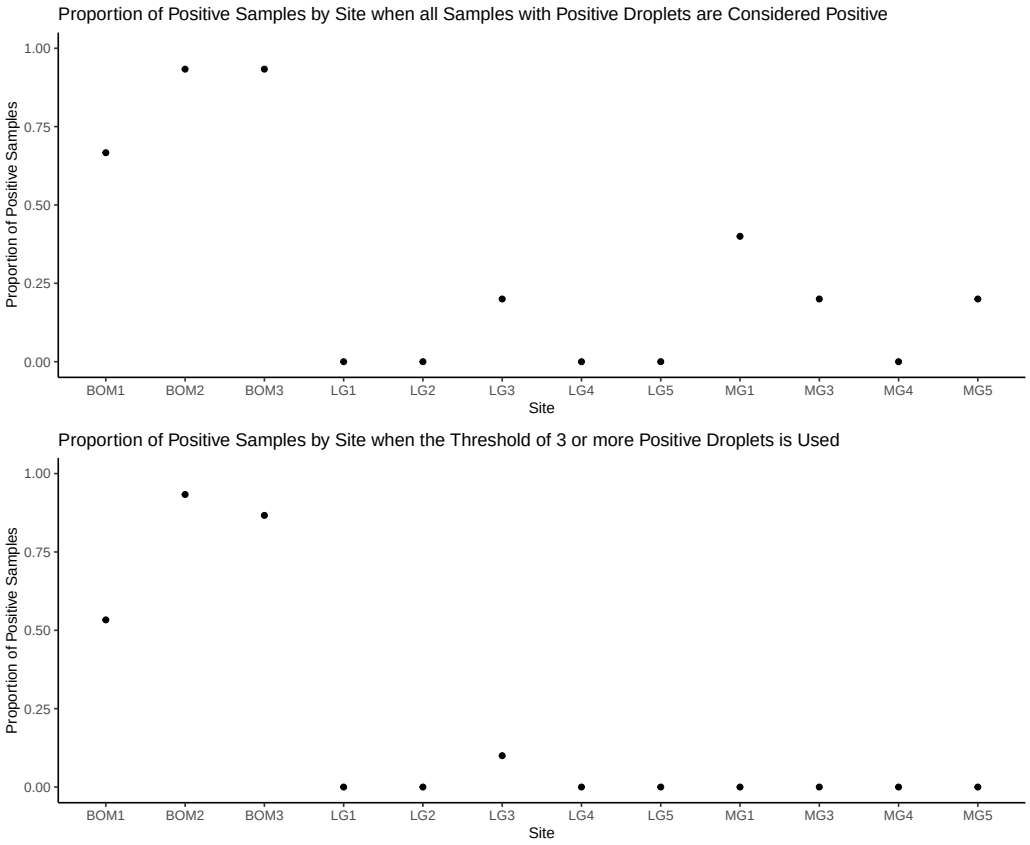

Figure 6: A plot of the proportion of positive droplets by site if all samples with any positive droplets are treated as positive detections (top), and a plot of the proportion by site if samples with fewer than three positive droplets are treated as negative (bottom).

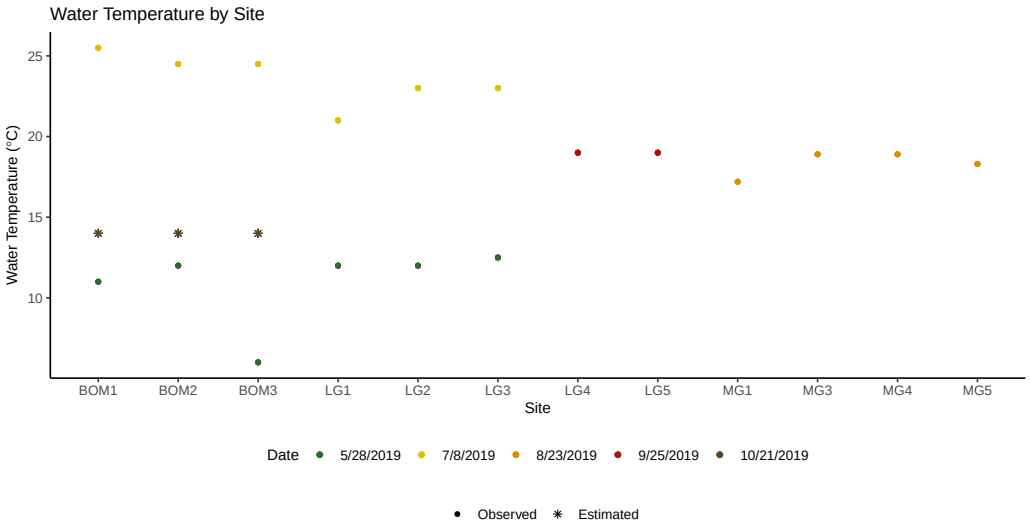

Figure 7: A plot of the water temperatures by site, colored by the date on which the sample was taken. The estimated water temperatures are denoted with stars.

3 Methods

To make inference about the occupancy and detection probabilities, an occupancy model will be used. Due to their hierarchical structure, occupancy models naturally lend themselves to a Bayesian framework.

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. The key distinction between Bayesian models and frequentist models is that is that, in Bayesian models the unknown

parameters are treated as random variables. 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:

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

The choice of prior distribution, $p(\boldsymbol{\theta})$, is up to the user. 'Subjective' or 'informative' priors the user wants to include known information about the unknown parameters, $\boldsymbol{\theta}$. Most often, 'uninformative', 'flat', or 'diffuse' prior distributions are used. The choice of the prior distribution is arbitrary and should be justified. 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 trials, with some unknown, and constant, 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 the model definition and Bayes' rule, 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.

In general, the resulting posterior distribution is not always a named statistical distribution. In situations where the posterior distribution is not a known distribution, 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 about the parameters of interest.

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 initial values of $\boldsymbol{\theta}^{(1)}$ can be chosen or randomly generated to be consistent with the parameter space for each θ_p. 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 sample from the approximate joint posterior distribution. However, in many 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, 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 special case 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 (i.e. the γ parameter) 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 are collected in a hierarchical, or multi-level, structure. Classic examples of a two-level hierarchy include: students within classes, or sampling sites within a larger area of interest. In these two level hierarchical data, 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. 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 studies are often used to track the mitigation efforts or spread of invasive species, but are plagued with imperfect detection probabilities. Occupancy surveys are an attempt

to learn about true presence/absence status based on the detection/non-detection data obtained.

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 information about the detectability of the species. There are several different types of occupancy models, but the traditional model is one that can be used when it is reasonable to assume that sites are closed to changes in occupancy status within a sampling season. The traditional occupancy model is defined in the following manner. 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, the model assumes, $Z_i \sim Bernoulli(\psi_i)$, where ψ_i denotes the probability that the target species occupies the i^{th} site. At the sample level, Y_{ij} is defined as whether the target species is detected or not in the j^{th} visit to the i^{th} site ($j = 1, \ldots, J_i$). It follows that, $Y_{ij}|z_{ij} \sim Bernoulli(z_i p_{ij})$, where z_i is the latent occupancy state for 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. The probability of detecting the target species at least one time given the species occupies the i^{th} site can be calculated in the following manner: $p_i^* = 1 - \prod_{j=1}^{J_i}(1 - p_{ij})$. Figure 8

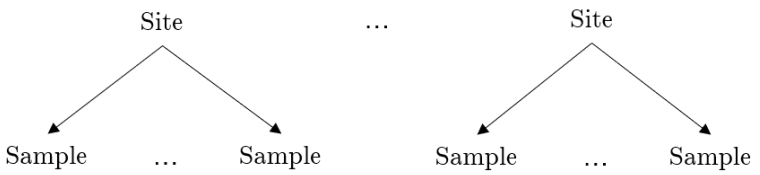

Figure 8: A diagram displaying the structure of a traditional occupancy study.

illustrates the structure of a traditional occupancy study, where each sample results in a binary response.

One assumption of occupancy models is that there is no unmodeled heterogeneity in the occupancy probabilities, or the detection probabilities, therefore covariates can be included at the site and sample levels. At the site level, $\boldsymbol{x}_i$ are the site-specific covariates for the i^{th} site and $\boldsymbol{\beta}$ is the corresponding vector of regression parameters. The $\boldsymbol{\beta}$ vector represents the logit (or probit) -linear relationship between $\boldsymbol{x}_i$ and the logit (or probit) transformed probability of presence at the i^{th} site, ψ_i, conditional on the other covariates in the model. At the sample level, $\boldsymbol{w}_{ij}$ are the sample-specific covariates and $\boldsymbol{\alpha}$ is the corresponding vector of regression parameters. Similarly, the $\boldsymbol{\alpha}$ vector represents the logit (or probit) -linear relationship between $\boldsymbol{w}_{ij}$ and the logit (or probit) transformed probability of detection on the j^{th} visit to the i^{th} site, conditional on presence at the i^{th} site and the other covariates in the model. It is also assumed that the sites are closed to changes in occupancy within a season, so a site that is occupied at the beginning of the study should remain occupied throughout; if an occupancy study takes

place over many years, there are models to account for the potential changes in occupancy status at the sites (i.e. dynamic or multi-season occupancy models). Another assumption of these models is that the species is not misidentified, however, there are models that can account for false negative and false positive detections. The final assumption of these models is that the detection of the species is independent across sites and within visits to a site. Violations of these assumptions can result in biased estimates of the occupancy and detection probabilities.

3.2.1 Multi-scale Occupancy Models

For more illusive species, the samples from each site are often split into sub-samples, 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). The model is defined similarly to the traditional occupancy model described above, with an additional replicate level, the notation is as follows. 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} sam-

ple 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} (sometimes deemed the availability parameter) 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 availability state for sample j from site i, and p_{ijk} is the conditional 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. Similarly to above, an 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; the notation for the covariates and corresponding regression parameters for the multi-scale occupancy model are as follows. At the site level, $\boldsymbol{x}_i$ are the site specific covariates and $\boldsymbol{\beta}$ is the corresponding vector of regression parameters for ψ_i. At the sample level, $\boldsymbol{w}_{ij}$ are the sample specific covariates and $\boldsymbol{\alpha}$ is the corresponding vector

of 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 $\boldsymbol{\delta}$ is the corresponding vector of regression parameters for p_{ij}. Site specific covariates can be included for θ_{ij} and p_{ij}. The remaining assumptions from traditional occupancy models, described above, hold for multi-scale occupancy models as well.

When summarized at the sample level, the data could be modeled in a traditional occupancy framework or a multi-scale occupancy framework. With the occupancy probability modeled at the sampling site level and detection probability modeled at the sample level, the data fit into a traditional occupancy model; however, with the occupancy probability modeled at the lake level, the availability modeled at the sampling site level, and the detection probability modeled at the sample level, the data fit into a multi-scale occupancy model. Finally, if analyzed at the droplet level the data could be analyzed with a multi-scale occupancy model where the occupancy probability is modeled at the sampling site level, availability modeled at the sample level, and the detection probability modeled at the droplet level. As previously mentioned, the data will be summarized at the sample level and analyzed in a traditional occupancy model framework.

3.3 Implementation

There are several options for fitting occupancy models; the `R` package used for this analysis as well as some alternative methods for fitting occupancy

models are described below.

3.3.1 `Rcppocc` Package for Fitting Occupancy Models

The traditional occupancy model described above can be fit in a Bayesian framework with the `PGocc4()` function from the `R` package Rcppocc (Clark, 2018); this package will be used for this analysis. The package relies on `Rcpp` and `RcppArmadillo`, so the models are computationally efficient. Additionally, the `PGocc4()` function introduces Pólya Gamma latent variables into the model to enable the use of a Gibbs sampler with the logit link function, while alternative options often use a Gibbs sampler with a probit link function, or a Metropolis-Hastings algorithm in conjunction with the logit link function. The use of a Gibbs sampler over a Metropolis-Hastings algorithm is generally preferred since the Gibbs sampler does not require tuning and is more computationally efficient. The interpretations of the estimated coefficients are not as straight-forward for the probit link function as they are when the logit link function is used, therefore this package provides the best of both worlds in terms of sampling algorithms and link functions for the regression parameters.

3.3.2 `msocc` Package for Fitting Multi-scale Occupancy Models

The multi-scale occupancy model described above (Dorazio and Erickson, 2018) can be easily fit in a Bayesian framework with the `msocc` package in `R` (Stratton, 2020). This package implements the Pólya Gamma data aug-

mentation strategy described by Polson et. al (2013) to enable the use of a Gibbs sampler in conjunction with a logit link function. 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. Other functionalities of the package are the posterior summary tools (`posterior_summary()`) and the function for generating credible interval plots (`cred_plot()`) at all levels of the hierarchy.

3.3.3 Alternative Options

There are several alternative options for fitting occupancy models. For example, the `eDNAoccupancy` (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. Occupancy 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

For these data, the 'sites' will be the twelve sampling sites across the three lakes, and the 'samples' will be the individual water samples from each of the sites. In this case, the occupancy probabilities can be estimated as constant across the sites, estimated by lake, or estimated individually for each site. Since the twelve sites are across three different lakes, a constant occupancy probability across sites does not make sense, because different bodies of water have different water chemistry that could impact the establishment of a dreissenid mussel population, additionally, lakes that are more frequented by water recreationists tend to be more vulnerable to introduction of dreissenid mussel veligers. However, estimating an individual occupancy probability for each site is likely more than necessary, because once a site is occupied with dreissenid mussels, it makes sense that the entire lake should be considered occupied, therefore occupancy probabilities will be estimated as constant for sites within the same lake. There are more options for covariates to impact the detection probabilities in each sample, conditional on site occupancy. It seems reasonable to assume that, conditional on site occupancy, the probability of detection depends on water temperature to allow for heterogeneity in detection probabilities across sites. The date on which the samples were taken could also impact the detection probabilities, but date and water temperature are related variables, so water temperature will be used, since it is a continuous measurement over the course of the study dates. The as-

sumption of no unmodeled heterogeneity in the occupancy probabilities or the detection probabilities seems reasonable here, though it would be nice to include water chemistry variables at the site level that are thought to impact whether a dreissenid mussel population can flourish. The sites in the study are assumed to be independent; in this case, there are multiple sites within each lake, which would violate the independence assumption if not accounted for in the model; the sites are not spatially, located within a lake, such that the independence assumption seems unreasonable. The likelihood of false positive detections were limited by using the recommended threshold of three or more positive droplets to be considered a positive sample, however the one site in lake LG with one positive sample only contained three positive droplets. The sites are also assumed to be closed to changes in occupancy throughout the course of the study, there is limited knowledge about how long eDNA remains in the water or how it moves throughout the water, so there is potentially some violation of the closure assumption, however, if the eDNA moves in and out of the sites randomly throughout the course of the study, the ψ parameter estimates will be unbiased. The model statements for the probability parameters are as follows :

$$logit(\psi_i) = \beta_0 + \beta_1 I_{lake=LG,i} + \beta_2 I_{lake=MG,i}$$

$$logit(p_{ij}) = \alpha_0 + \alpha_1 water.temp_{ij}$$

where the $I_{lake=LG,i}$ is an indicator variable that is 1 when the i^{th} site is in lake LG and 0 otherwise, $I_{lake=MG,i}$ is a similarly defined indicator variable,

which is 1 when the i^{th} site is in lake MG and 0 otherwise, and $water.temp_{ij}$ is the water temperature for the j^{th} sample from the i^{th} site.

Prior distributions are needed for each vector of regression coefficients, $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. The following prior distributions were used:

$$\boldsymbol{\beta} \sim MVN(\tilde{0}, 3^2 I)$$
$$\boldsymbol{\alpha} \sim MVN(\tilde{0}, 3^2 I).$$

As pointed out by Seaman et al. (2012), "if parameters with diffuse proper priors are subsequently transformed, the resulting induced priors can, of course, be far from diffuse, possibly resulting in unintended influence on the posterior of the transformed parameters," oftentimes this is a problem when non-linear transformations, such as the logit-link function, are used. For occupancy models, and other logit-link models, the problem is exacerbated when the probabilities are close to 0 or 1. Consequently, a multivariate normal distribution with a mean vector of $\tilde{0}$ and covariance matrix $3^2 I$ will be used as the prior distributions for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. The multivariate normal distribution centered at $\tilde{0}$ places equal prior probability on each of the β and α coefficients being positive or negative. A covariance matrix of $3^2 I$ is a diffuse prior variance on the probability scale.

5 Results

Based on the β estimates in Table 2, the estimated odds of dreissenid mussel eDNA being present in lake LG is between 0.004 times and 1.31 times the estimated odds of dreissenid mussel eDNA being present in lake BOM, with 95% probability. The estimated odds of dreissenid mussel eDNA being present in lake MG is between 0.0004 and 0.48 times the estimated odds of dreissenid mussel eDNA being present in lake BOM, with 95% probability.

Table 2: Summary table of the $\hat{\beta}$ coefficients for $logit(\hat{\psi}_i) = \hat{\beta}_0 + \hat{\beta}_1 I_{lake=LG,i} + \hat{\beta}_2 I_{lake=MG,i}$.

	Mean	2.5% Quantile	97.5% Quantile
$\hat{\beta}_0$	1.16224	-0.95616	3.54199
$\hat{\beta}_1$	-2.48269	-5.50485	0.26869
$\hat{\beta}_2$	-4.02276	-7.94814	-0.72410

Furthermore, based on the α estimates in Table 3, for a 1°C increase in water temperature, the estimated odds of detecting dreissenid mussel eDNA with ddPCR is between 0.95 and 1.25 times what it would have been previously, with 95% probability. This suggests that water temperature might not impact sample level detection probabilities, but it will be retained in the model to control for differences that could arise in detection probabilities.

Table 3: Summary table of the $\hat{\alpha}$ coefficients for $logit(\hat{p}_{ij}) = \hat{\alpha}_0 + \hat{\alpha}_1 water.temp_{ij}$.

	Mean	2.5% Quantile	97.5% Quantile
$\hat{\alpha}_0$	0.09334	-0.92009	1.12171
$\hat{\alpha}_1$	0.08638	-0.04641	0.22375

5.1 Ecological Implications

Perhaps more interesting than the α and β estimates themselves, especially to land managers and researchers are the probability estimates derived from those estimated α's and β's. Based on the model, there is a 95% chance that the probability dreissenid mussel eDNA is present at each of the sites in lake BOM is between 27.76% and 97.19% (posterior mean occupancy probability for lake BOM: 71.58%). Additionally, there is a 95% chance that the probability dreissenid mussel eDNA is present at each of the sites in lake MG is between 0.11% and 48.38% (posterior mean occupancy probability for lake MG: 11.42%). Finally, there is a 95% chance that the probability dreissenid mussel eDNA is present at the sites in lake LG is between 0.02% and 64.97% (posterior mean occupancy probability for lake LG: 25.60%). Figure 9 displays visually the 95% credible intervals and posterior means for each of the $\hat{\psi}_i$. These results suggest that, on average, the sites in lake BOM are most likely to be infested with dreissenid mussel eDNA, and the sites in lake MG are the least likely to be infested with dreissenid mussel eDNA.

Of further interest to researchers are the detection probabilities for each

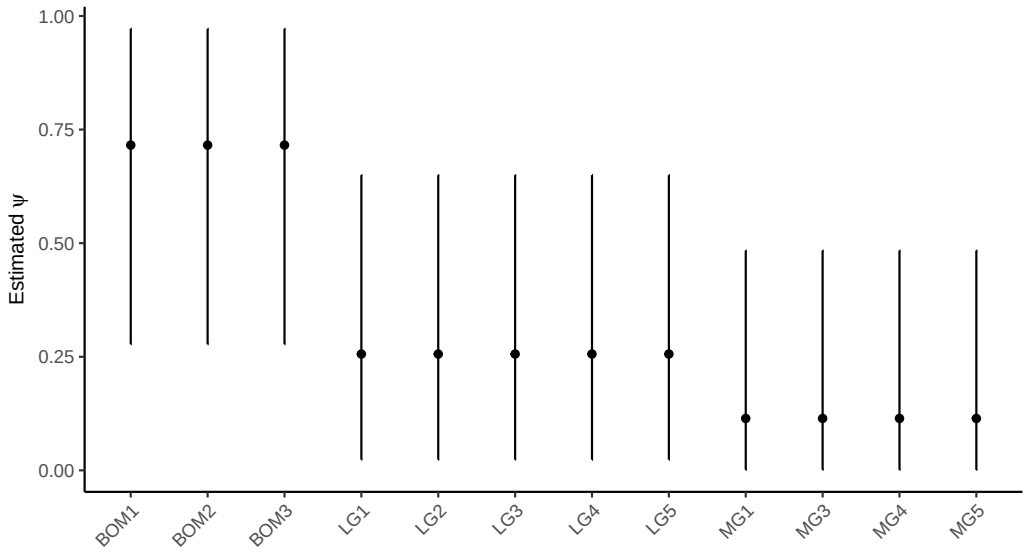

Figure 9: Plot of the 95% credible intervals for each of the $\hat{\psi}_i$, the posterior mean is plotted with a point.

sample. In this model, detection probabilities were allowed to vary by water temperature, so each of the five samples on the same day from each site should have the same detection probability estimates because multiple water temperatures were not recorded within the same site on the same day. A plot of the 95% credible intervals and posterior means for each uniquely estimated detection probability is plotted by site and date in Figure 10; based on the figure, it appears that the detection probabilities, conditional on site occupancy, vary slightly. Table 4 displays the numerical summaries of the estimated detection probabilities, conditional on site occupancy.

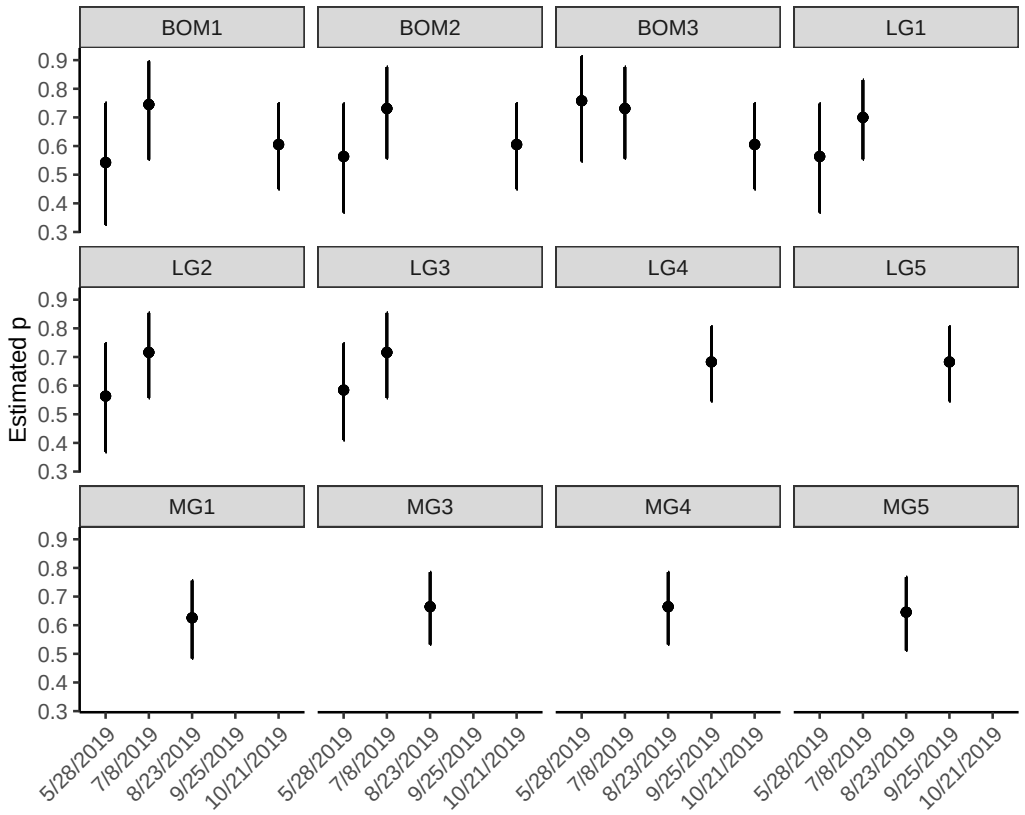

Figure 10: Plot of the 95% credible intervals for each of the uniquely estimated p_{ij}, the posterior mean is plotted with a point.

Table 4: Summary table of the estimated detection probabilities for each site and date combination; the water temperatures are provided and the temperatures that were estimated in lake BOM on 10/20/2019 are denoted with *.

Site	Date	Water Temp. (°C)	Mean	2.5% Quantile	97.5% Quantile
BOM1	5/28/2019	11	0.54257	0.32546	0.75032
BOM1	7/8/2019	22.5	0.74504	0.55279	0.89574
BOM1	10/21/2019	14*	0.60530	0.44990	0.74902
BOM2	5/28/2019	12	0.56353	0.36855	0.74766
BOM2	7/8/2019	24.5	0.73106	0.55636	0.87587
BOM2	10/21/2019	14*	0.60530	0.44990	0.74902
BOM3	5/28/2019	6	0.75797	0.54705	0.91316
BOM3	7/8/2019	24.5	0.73106	0.55636	0.87587
BOM3	10/21/2019	14*	0.60530	0.44990	0.74902
LG1	5/28/2019	12	0.56353	0.36855	0.74766
LG1	7/8/2019	21	0.69989	0.55403	0.83043
LG2	5/28/2019	12	0.56353	0.36855	0.74766
LG2	7/8/2019	23	0.71601	0.55707	0.85376
LG3	5/28/2019	12.5	0.58451	0.41016	0.74681
LG3	7/8/2019	23	0.71601	0.55707	0.85376
LG4	9/25/2019	19	0.68272	0.54582	0.80627
LG5	9/25/2019	19	0.68272	0.54582	0.80627
MG1	8/23/2019	17.2	0.62570	0.48420	0.75529
MG3	8/23/2019	18.9	0.66456	0.53314	0.78501
MG4	8/23/2019	18.9	0.66456	0.53314	0.78501
MG5	8/23/2019	18.3	0.64551	0.51195	0.76774

5.2 Sampling Recommendations

Of further interest to the researchers is sampling recommendations, such as the recommended number of samples to take to ensure a high probability of at least one detection at a site, if the site is occupied. A table of probabilities of detecting dreissenid mussel eDNA at least one time in the J_i samples from

each of the twelve sites, given that there is dreissenid mussel eDNA present at the i^{th} site is provided in Table 5. Based on this model, the estimated probabilities of detecting dreissenid mussel eDNA in at least one sample from the i^{th} site, given dreissenid mussel eDNA is present at the site, is high across all the sites, with a lower bound of 97%. This suggests that the sampling scheme used for this study is sufficient. However, these detection probabilities are driven by the detections at lake BOM. Since the occupancy probabilities and detection probabilities are interwoven in their estimation, more samples at lake LG and MG could help researchers get a better idea about the detection probabilities at sites that are potentially occupied but have low abundance of dreissenid mussel eDNA. Additionally, more samples across differing water temperatures could help better inform the detection probabilities, conditional on site occupancy status.

As previously mentioned, it would also be of interest to collect other water chemistry variables thought to impact whether a dreissenid mussel population is able to establish, such as dissolved calcium and pH (Benson et al., 2020); these covariates could be included in the modeling for the occupancy parameters, which would better inform why the sites differ in terms of occupancy.

After the samples are collected and filtered, it would be beneficial to take multiple sub-samples from the filtered water. This would hopefully lead to an understanding of where some of the missed detections are coming from. There are multiple ways a negative sample can arise: there is no dreissenid

Table 5: Summary table of the estimated probabilities of detecting dreissenid mussel eDNA at least one time in the J_i samples, given there is dreissenid mussel eDNA present at the i^{th} site.

Site	Number of Samples	2.5% Quantile	97.5% Quantile
BOM1	15	1.000	1.000
BOM2	15	1.000	1.000
BOM3	15	1.000	1.000
LG1	10	0.998	1.000
LG2	10	0.998	1.000
LG3	10	0.999	1.000
LG4	5	0.981	1.000
LG5	5	0.981	1.000
MG1	5	0.963	0.999
MG3	5	0.978	1.000
MG4	5	0.978	1.000
MG5	5	0.972	0.999

mussel eDNA at the site; there is dreissenid mussel eDNA at the site but it is not captured in the sample; the dreissenid mussel eDNA is present at the site, is captured in the sample, but not be captured in the sub-sample that was analyzed with ddPCR; or there the dreissenid mussel eDNA is present at the sites, is captured in the sample, is captured in the sub-sample, but is not detected with ddPCR. If multiple sub-samples, or replicates, were taken from the same sample, more could be learned about where the missed detections are occurring in the process.

6 Future Work

Oftentimes in occupancy surveys, there is potential for another species to be misidentified as the species of interest in the study, this would lead to a false positive results. In eDNA surveys, false positives likely occur as a result of the PCR technique. With digital droplet PCR, there is a chance that optical interference could lead to false positive results. False positives have the potential to positively bias the detection probabilities, which in turn, can negatively bias the occupancy probabilities. In the near future, the plan is to model these data with an occupancy model that accounts for false positive detections. If a false positive occupancy model is used, the threshold of the number of droplets to be considered a positive sample can be excluded, and all samples with any number of positive droplets can be considered positive samples, since the model itself would allow for false positive detections. There are various different options for false positive occupancy models. One example is the model defined by Royle and Link (2006), in which there is symmetry in the likelihood which leads to solutions that are not unique. To account for this, the parameter space is restricted in the model such that the true positive detection probability is greater than the false positive detection probability. However, this restriction was criticized by McClintock et al. (2010) because if the the target species is not present and is falsely detected, then the conclusion would be that the species is present due to the restriction imposed on the detection probabilities in the model

(McClintock et al., 2010). The Bayesian model developed by Ferguson et al. (2015) requires more information in the form of a confirmations; two models are discussed in the paper, one with confirmed and unconfirmed absences and confirmed and unconfirmed presences (CACP model), and one with only unconfirmed absences but confirmed and unconfirmed presences (CP model) (Ferguson et al., 2015; and many others).

After fitting the false positive occupancy model and getting back to the researchers, the issues with the amplitude cutoff will be addressed. Currently, there is a subjective choice of an amplitude cutoff by looking at plots such as those in Figure 5 and choosing a cutoff for which the droplets with amplitudes above that point should be considered positive. This is an arbitrary choice and can lead to potentially vastly different numbers of positive droplets per sample if a different cutoff is chosen. The choices in the amplitude cutoff are loosely based off a visual exploration of plots of the amplitudes for positive and negative controls; examples of amplitude plots for a positive and negative control from lake LG can be found in Figure 11. Ideally, we would like to avoid the specification of an arbitrary amplitude cutoff by instead using the distributions of amplitudes for the positive and negative controls to quantify the behavior of false positives and false negatives.

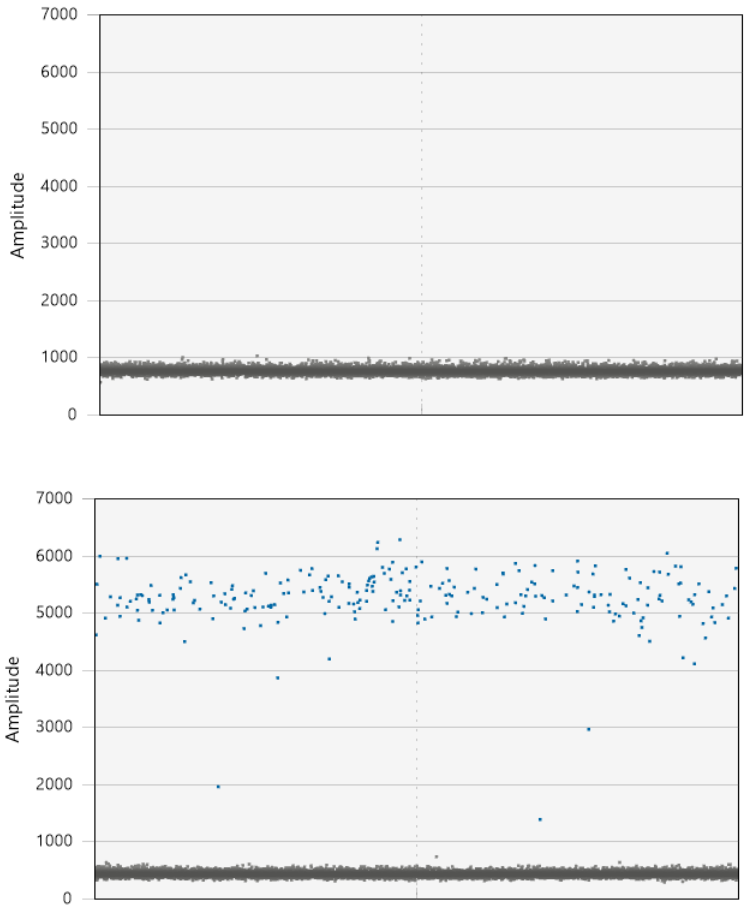

Figure 11: An example of amplitude plots for a positive control (bottom) and a negative control (top) from lake LG.

7 Acknowledgements

I would like to thank Dr. Andrew Hoegh for the time spent helping me with this project and furthering my understanding of statistics in general. I would also like to thank Dr. Adam Sepulveda from the Northern Rocky Mountain Science Center for his collaboration and willingness to share his resources. Additionally, I would like to extend a thank you to Alison Watts from the University of New Hampshire for sharing her data. Finally, I would like to thank the entire Department of Mathematical Sciences Statistics faculty at Montana State University for helping me grow both as a statistician and as a student.

8 References

Benson, A. J., Raikow, D., Larson, J., Fusaro, A., Bogdanoff, A. K., and Elgin, A. (2020). *Dreissena polymorpha* fact sheet. `https://nas.er.usgs.gov/queries/FactSheet.aspx?speciesID=5`. Date Accessed: April 28, 2020.

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.

Clark, A. (2018). *Rcppocc: Fits various Bayesian occupancy models assuming all regression effects uses a logit link function*. R package version 1.0.

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.

Ferguson, P. F., Conroy, M. J., and Hepinstall-Cymerman, J. (2015). Occupancy models for data with false positive and false negative errors

and heterogeneity across sites and surveys. *Methods in Ecology and Evolution*, 6(12):1395–1406.

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

Gingera, T., Bajno, R., Docker, M., and Reist, J. (2017). Environmental DNA as a detection tool for zebra mussels Dreissena polymorpha (pallas, 1771) at the forefront of an invasion event in Lake Winnipeg, Manitoba, Canada. *Management of Biological Invasions*, 8(3):287–300.

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.

McClintock, B. T., Bailey, L. L., Pollock, K. H., and Simons, T. R. (2010). Unmodeled observation error induces bias when inferring patterns and

dynamics of species occurrence via aural detections. *Ecology*, 91(8):2446–2454.

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.

Nelson, N. M. (2019). Conservation and resource development. urlhttp://dnrc.mt.gov/divisions/cardd. Date Accessed: February 15, 2020.

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.

Royle, J. A. and Link, W. A. (2006). Generalized site occupancy models allowing for false positive and false negative errors. *Ecology*, 87(4):835–841.

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.

Seaman, J. W. I., Seaman, J. W. J., and Stamey, J. D. (2012). Hidden Dangers of Specifying Noninformative Priors. *The American Statistician*, 66(2):77–84.

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.

9 Appendix - R Code

```
# packages used
library(car)
library(dplyr)
library(tidyr)
library(dataRetrieval)
library(kableExtra)
library(ggplot2)
library(grid)
library(gridExtra)
library(tm)
library(readxl)
library(Rcppocc)


# 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"))
```

```
# vector of site names
Site.names <- unique(eDNA$Site)

# number of sites
M <- length(Site.names)

# vector of lake names for each sample
lake <- eDNA$Site %>%
  as.character() %>%
  removeNumbers() %>%
  factor()


# 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)


# overall summary of concentrations
summary(eDNA$Conc)

# mean concentration by site
eDNA %>%
  group_by(Site) %>%
  summarise(mean(Conc))


# summary at sample level
## 3 or more positive droplets
eDNA$Detect3 <- rep(NA, nrow(eDNA))
for(i in 1:nrow(eDNA)){
  if(eDNA$Positive.Droplets[i] >= 3){
    eDNA$Detect3[i] <- 1
  } else {
      eDNA$Detect3[i] <- 0
      }
}


## number of postive samples in each site w/ threshold
```

```
site.detect3 <- eDNA %>%
  group_by(Site) %>%
  summarise(sum(Detect3))


## total number of samples in each site
count <- eDNA %>%
  group_by(Site) %>%
  count()


## proportions of positive samples in each site w/ threshold
cbind(Site = Site.names, site.detect3[, 2]/count[, 2])


## all positive droplets
eDNA$Detect <- rep(NA, nrow(eDNA))
for(i in 1:nrow(eDNA)){
  if(eDNA$Positive.Droplets[i] > 0){
    eDNA$Detect[i] <- 1
  } else {
      eDNA$Detect[i] <- 0
      }
}


site.detect <- eDNA %>%
  group_by(Site) %>%
  summarise(sum(Detect))


eDNA.df2 <- data.frame(Site.names,
                       site.detect[2]/count[2],
                       site.detect3[2]/count[2])
names(eDNA.df2) <- c("Site", "Prop", "Prop3")


# eDNA positive sample plot
prop.plot <- eDNA.df2 %>%
```

```
  ggplot(aes(x = Site,
             y = Prop)) +
  geom_point() +
  labs(title = paste0('Proportion of Positive Samples by Site when
      all Samples with Positive Droplets are Considered Positive'),
       x = 'Site',
       y = 'Proportion of Positive Samples') +
  ylim(c(0, 1)) +
  theme_bw() +
  theme(title = element_text(size = 10),
        panel.border = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "black"))

# eDNA positive sample plot w/ threshold
prop3.plot <- eDNA.df2 %>%
  ggplot(aes(x = Site,
             y = Prop3)) +
  geom_point() +
  labs(title = paste0('Proportion of Positive Samples by Site when
  the Threshold of 3 or more Positive Droplets is Used'),
       x = 'Site',
       y = 'Proportion of Positive Samples') +
  ylim(c(0, 1)) +
  theme_bw() +
  theme(title = element_text(size = 10),
        panel.border = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "black"))

# plot positive sample plots w/ and w/o threshold
# vertically stacked
grid.arrange(prop.plot, prop3.plot, nrow = 2)


# replace missing water temps from lake BOM with 14C
eDNA$Water.Temp <- replace(eDNA$Water.Temp,
```

```
                          is.na(eDNA$Water.Temp), 14)


# denote which water temperatures were estimated
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),
                       lake = lake)

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 = 10),
        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')


# format data for use in PGocc4 function
```

```
## number of samples per sites
J <- rep(0, M)

for(j in 1:length(J)){
  for(i in 1:nrow(eDNA)){
    if(eDNA$Site[i] == Site.names[j]){
      J[j] <- J[j] + 1
    }
  }
}


## vector of the sample number in each site
samp <- matrix(NA, nrow = M, ncol = max(J))

for(i in 1:M){
  if(J[i] == ncol(samp)){
    samp[i, ] <- seq(1:J[i])
  } else {
    samp[i, 1:J[i]] <- seq(1:J[i])
  }
}

samp <- na.omit(as.vector(t(samp)))


## detection/non-detection matrix
y <- matrix(NA, nrow = M, ncol = max(J))

for(i in 1:M){
  for(k in 1:nrow(eDNA)){
    if(eDNA$Site[k] == Site.names[i] & eDNA$Detect3[k] == 1){
      y[i, samp[k]] <- 1
    } else if(eDNA$Site[k] == Site.names[i] & eDNA$Detect3[k] == 0){
      y[i, samp[k]] <- 0
      }
  }
}
```

```
y <- matrix(as.integer(y), nrow = M, ncol = max(J))


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

X <- as.matrix(X)

colnames(X)[2] <- "lake"


## sample level covariates
### water temperature
W1 <- matrix(NA, nrow = M, ncol = max(J))

for(i in 1:M){
  if(J[i] == ncol(W1)){
    W1[i, ] <- eDNA$Water.Temp[eDNA$Site == Site.names[i]]
  } else {
    W1[i, 1:J[i]] <- eDNA$Water.Temp[eDNA$Site == Site.names[i]]
    }
}

### date
W2 <- data.frame(matrix(NA, nrow = M, ncol = max(J)))

for(i in 1:M){
  if(J[i] == ncol(W2)){
    W2[i, ] <- eDNA$Date.Collected[eDNA$Site == Site.names[i]]
  } else {
    W2[i, 1:J[i]] <- eDNA$Date.Collected[eDNA$Site == Site.names[i]]
    }
}
```

```
W2 <- as.matrix(W2)

### site
W3 <- data.frame(matrix(NA, nrow = M, ncol = max(J)))

for(i in 1:M){
  if(J[i] == ncol(W3)){
    W3[i, ] <- eDNA$Site[eDNA$Site == Site.names[i]]
  } else {
    W3[i, 1:J[i]] <- eDNA$Site[eDNA$Site == Site.names[i]]
    }
}

W3 <- as.matrix(W3)

### lake
W4 <- data.frame(matrix(NA, nrow = M, ncol = max(J)))

for(i in 1:M){
  if(J[i] == ncol(W4)){
    W4[i, ] <- rep(X[i, 2], ncol(W4))
  } else {
    W4[i, 1:J[i]] <- rep(X[i, 2], J[i])
  }
}

W4 <- as.matrix(W4)

## list of sample covariates
W <- list(W1 = W1, W2 = W2, W3 = W3, W4 = W4)


# merge for use in PGocc4
data <- vb_Designs(W = W, X = X, y = y)


# function to summarize the alpha, beta, p, and psi estimates
post_summary <- function(parameter.matrix, param, plot = T){
  mean <- apply(parameter.matrix, 1, mean)
```

```
median <- apply(parameter.matrix, 1, median)
min <- apply(parameter.matrix, 1, min)
lwr <- apply(parameter.matrix, 1, quantile, probs = 0.025)
upr <- apply(parameter.matrix, 1, quantile, probs = 0.975)
max <- apply(parameter.matrix, 1, max)
out <- cbind("mean" = mean,
             "median" = median,
             "min" = min,
             "2.5%" = lwr ,
             "97.5%" = upr,
             "max" = max)
if(missing(param)){
  out <- out
} else if(param == "psi"){
  out <- cbind.data.frame(out,
                          rownames = Site.names)
} else if(param == "p"){
  out <- cbind.data.frame(Site = eDNA$Site,
                          Sample = samp,
                          out)
} else if(param == "alpha"){
  rownames(out) <- paste("alpha", 0:(nrow(out) - 1))
} else if(param == "beta"){
  rownames(out) <- paste("beta", 0:(nrow(out) - 1))
}
if(plot == T){
  tr.plot <- list()
  for(i in 1:nrow(parameter.matrix)){
    df <- data.frame(param = parameter.matrix[i, ],
                     iter = 1:ncol(parameter.matrix))
    tr.plot[[i]] <- ggplot(df, aes(x = iter, y = param)) +
      geom_path() +
      xlab("Post Burn-in Iterations") +
      ylab(rownames(out)[i]) +
      theme_bw() +
        theme(title = element_text(size = 10),
              panel.border = element_blank(),
              panel.grid.major = element_blank(),
              panel.grid.minor = element_blank(),
```

```
                axis.line = element_line(colour = "black"))
    }
    if (missing(param)){
      print(grid.arrange(grobs = tr.plot, ncol = 2))
    } else if (param == 'beta'){
      print(grid.arrange(grobs = tr.plot, ncol = 2,
            top = textGrob(expression(paste("Trace Plots of the ",
                                            beta, "'s")))))
    } else if (param == 'alpha'){
      print(grid.arrange(grobs = tr.plot, ncol = 2,
            top = textGrob(expression(paste("Trace Plots of the ",
                                            alpha, "'s")))))
    } else if (param == 'p') {
      print(grid.arrange(grobs = tr.plot, ncol = 2,
            top = textGrob("Trace Plots of the p's")))
    } else if (param == 'psi'){
      print(grid.arrange(grobs = tr.plot, ncol = 2,
            top = textGrob(expression(paste("Trace Plots of the ",
                                            psi, "'s")))))
    }
  }
  return(out)
}


# function to generate nice looking plots
# of credible intervals for p and psi
cred_plot <- function(post_summary_out, parameter){
  df <- data.frame(post_summary_out)
  if(nrow(df) == 12){
    rownames(df) <- Site.names
  }
  if(parameter == "psi"){
    cred.plot <- ggplot(df, aes(x = rownames(df), y = mean)) +
      geom_point() +
      xlab("") +
      geom_errorbar(aes(ymax = X97.5., ymin = X2.5.), width = 0) +
      theme_bw() +
      theme(title = element_text(size = 10),
```

```
            panel.border = element_blank(),
            panel.grid.major = element_blank(),
            panel.grid.minor = element_blank(),
            axis.line = element_line(colour = "black"),
            axis.text.x = element_text(angle = 45, hjust = 1)) +
      ylab(expression(paste("Estimated ", psi)))
  } else if(parameter == "p"){
    cred.plot <- ggplot(df, aes(x = eDNA$Date.Collected, y = mean)) +
      geom_point() +
      xlab("") +
      geom_errorbar(aes(ymax = X97.5., ymin = X2.5.), width = 0) +
      theme_bw() +
      theme(title = element_text(size = 10),
            panel.border = element_blank(),
            panel.grid.major = element_blank(),
            panel.grid.minor = element_blank(),
            axis.line = element_line(colour = "black"),
            axis.text.x = element_text(angle = 45, hjust = 1)) +
      ylab("Estimated p") +
      facet_wrap(~Site)
  }
  print(cred.plot)
}


# model psi ~ lake, p ~ water.temp

# Priors

## beta ~ MVN(0, 3^2I)
### prior mean for beta
beta_m <- matrix(0, nrow = 3, ncol = 1)
#### prior precision for beta
sigma_inv_beta <- diag(nrow(beta_m))/9

## alpha ~ MVN(0, 3^2I)
### prior mean for alpha
alpha_m <- matrix(0, nrow = 2, ncol = 1)
### prior precision for alpha
```

```
sigma_inv_alpha <- diag(nrow(alpha_m))/9


# fit the model
set.seed(04172020)
m5 <- PGocc4(formula = V1 ~ lake ~ as.numeric(W1),
             design_mats = data,
             ndraws = 50000,
             alpha_m = alpha_m,
             beta_m = beta_m,
             sigma_inv_alpha_p = sigma_inv_alpha,
             sigma_inv_beta_p = sigma_inv_beta,
             percent_burn_in = 1/10)


# extract beta estimates from model output
m5.beta <- m5$beta

# posterior summary and trace plots of beta estimates
m5.beta.sum <- post_summary(m5.beta, 'beta')


# extract alpha estimates from model output
m5.alpha <- m5$alpha

# posterior summary and trace plots of alpha estimates
m5.alpha.sum <- post_summary(m5.alpha, 'alpha')


# table of beta estimates
m5.beta.sum
# table of alpha estimates
m5.alpha.sum
# summary of psi
## calculate estimates of psi from estimates of beta
m5.psi.mcmc <- exp(model.matrix(~data$X$lake) %*% m5.beta)/
  (1 + exp(model.matrix(~data$X$lake) %*% m5.beta))

## summary of psi
```

```
m5.psi.sum <- post_summary(m5.psi.mcmc, 'psi', plot = F)
m5.psi.sum

# credible interval plot for psi
cred_plot(m5.psi.sum, 'psi')


# summary of p
## calculate estimates of p from estimates of alpha
m5.p.mcmc <- exp(model.matrix(~as.numeric(data$W$W1)) %*% m5.alpha)/
  (1 + exp(model.matrix(~as.numeric(data$W$W1)) %*% m5.alpha))

## summary of p
m5.p.sum <- post_summary(m5.p.mcmc, 'p', plot = F)
m5.p.sum

## credible interval plot for p
cred_plot(m5.p.sum, 'p')


# table of p* for each site
names(m5.p.sum)[6:7] <- c("lwr", "upr")

m5.p.sum

tmp <- rep(NA, M)
p.star <- rep(NA, M)

m5.p.sum %>%
  group_by(Site) %>%
  summarise(p.star.lwr = 1 - prod(1 - lwr),
            p.star.upr = 1 - prod(1 - upr))
```