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. This method is 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). *new paragraph here?* 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 *(written such that lake is the top level for eDNA too, if it isn't analyzed like this then change this sentence a little bit)*. Replication in the survey design can help researchers learn about the detection probabilities for each of the early detection methods. One question of interest is how the detection probabilities and false negative rates of dreissenid mussels compare for plankton tow methods and eDNA survey methods. Given that information, the follow up question is how many samples *(I don't think I want to use the word samples since it is defined so clearly below)* should be taken with each method *(get a bit more specific here)*.

2 Data

A sentence or two here about the both data sets.

2.1 eDNA Survey Data

data source: part of larger study involving Asian clams region: north-eastern United States, VT and NY number of lakes: 3 with data, 4 total... the last one did not have any positive replicates data collected summer/fall 2019 water temperature information available

5 samples (at all sites) and a field blank (at most sites, all had 0 positive droplets)

20000 ddPCR replicates from each sample?? Figure 1 is a diagram of data structure within a lake.

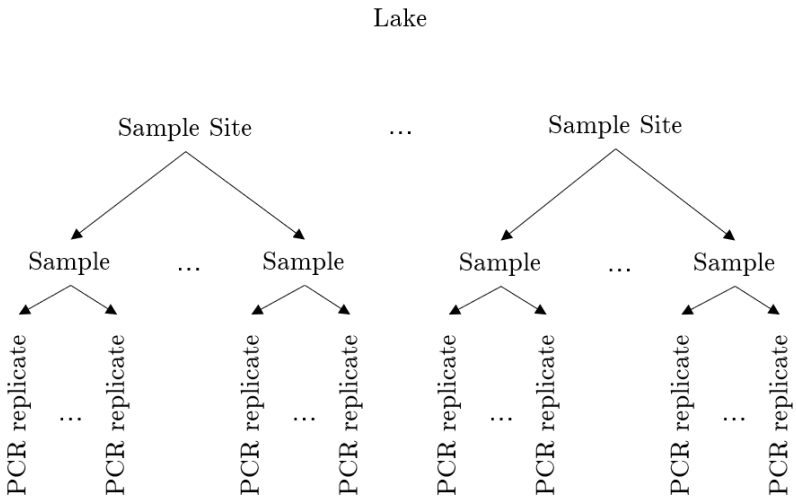

Figure 1: A diagram displaying the hierarchical structure of the eDNA data.

2.2 Plankton Tow Survey Data

Plankton tow data description

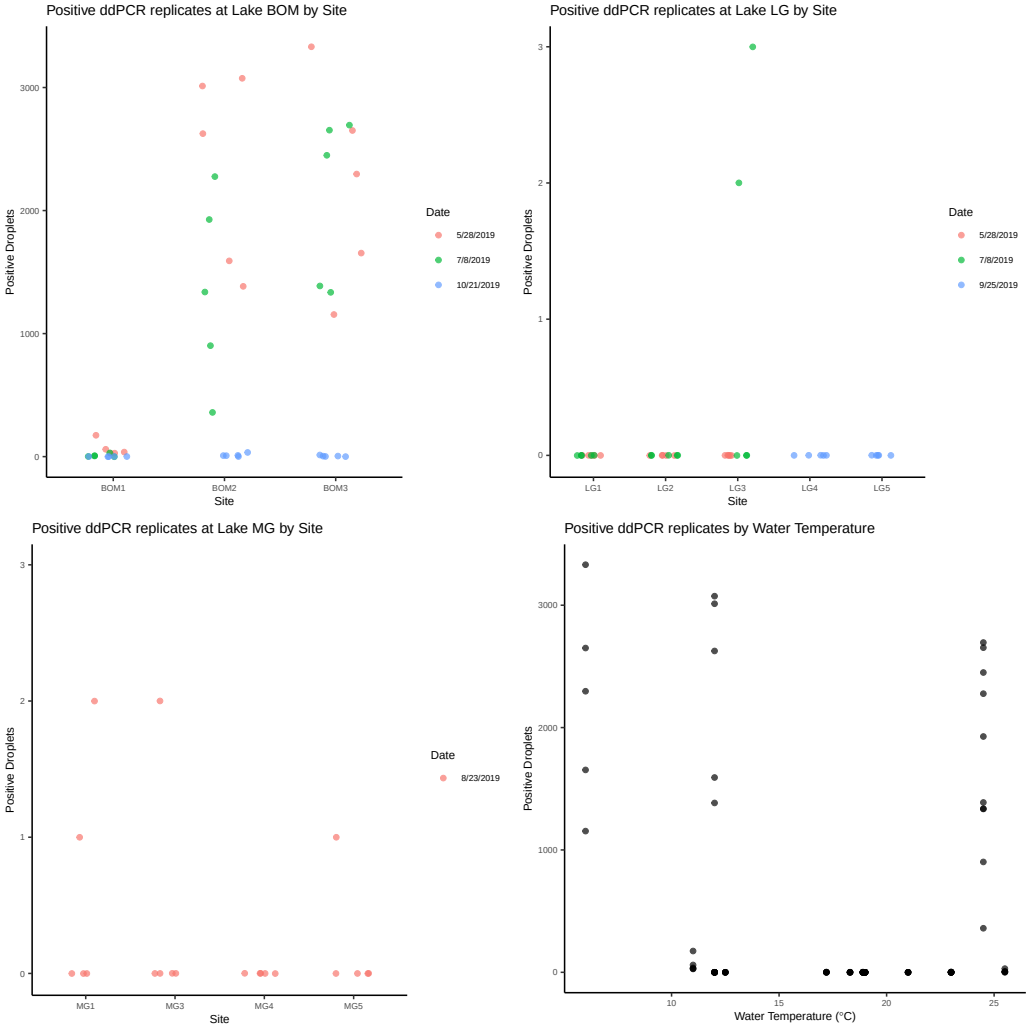
Positive ddPCR replicates at Lake BOM by Site
Positive Droplets
3000
2000
1000
0
BOM1
BOM2
BOM3
Site
Date
5/28/2019
7/8/2019
10/21/2019
Positive ddPCR replicates at Lake LG by Site
Positive Droplets
3
2
1
0
LG1
LG2
LG3
LG4
LG5
Site
Date
5/28/2019
7/8/2019
9/25/2019
Positive ddPCR replicates at Lake MG by Site
Positive Droplets
3
2
1
0
MG1
MG3
MG4
MG5
Site
Date
8/23/2019
Positive ddPCR replicates by Water Temperature
Positive Droplets
3000
2000
1000
0
10
15
20
25
Water Temperature (°C)

Table 1: 10 sample rows of the eDNA data.

Lake	Site	Sample.ID	Date.Collected	Water.Temp	Conc	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

- data source: BOR
- region
- number of lakes (number of lakes where dreissenid mussel veligers were detected)
- number of sites
- dates
- potential covariates

Plankton tow data visualization

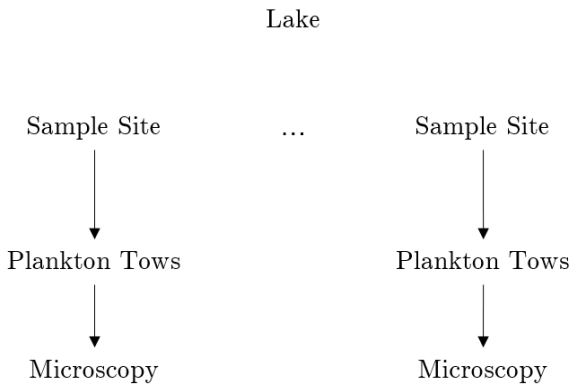

Figure 2: A diagram displaying the structure of the plankton tow data.

3 Methods

3.1 Bayesian Modeling Background

3.2 Occupancy Models

this is still in the works

Occupancy is the presence of a particular species on a given site, this may not be the first choice of state variables to ecologists but occupancy studies are useful when there is a large spatial scale or the study is conducted over many years, when abundance or vital rates are hard to measure. Occupancy studies are also useful over capture-recapture methods when individuals cannot be marked or uniquely identified. However, sometimes patterns of species occurrence are of interest, this happens when researchers are interested in the range of a species or the spread of invasion. The sampling units for occupancy studies are called 'sites'. We can learn about detection probabilities when

multiple site visits are used. Also, when using occupancy models we need to account for imperfect detection because it is possible that the researchers could miss the species even if it is present at the site.

Hierarchical structure of eDNA data creates dependencies that need to be accounted for (**?**)

in the hierarchical multi-scale occupancy model (with three nested levels) defined by (Dorazio and Erickson, 2017):

- learn about the probability of species occurrence at a site
- Z_i denotes the presence or absence of the target species at the i^{th} site $(i = 1, \ldots, M)$
- $Z_i \sim Bernoulli(\psi_i)$
- ψ_i is the probability that the target species is present at the i^{th} site
- $\boldsymbol{\beta}$ are the site level regression parameters for ψ_i
- $\boldsymbol{x}_i$ are the site level covariates for ψ_i
- learn about the conditional probability of species occurrence in a sample of a site given the species is present at the site
- A_{ij} denotes the presence or absence of the target species in the j^{th} sample from the i^{th} site $(j = 1, \ldots, J_i)$
- $A_{ij}|z_i \sim Bernoulli(z_i\theta_{ij})$

- z_i is a realized value of Z_i
- 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 location
- $\boldsymbol{\alpha}$ are the sample level regression parameters for θ_{ij}
- $\boldsymbol{w}_{ij}$ are the sample level covariates for θ_{ij}
- learn about the conditional probability of detection of the species in a sub-sample of a sample given that the species is present in the sample
- Y_{ijk} denotes the detected or not detected in the k^{th} replicate of the j^{th} sample collected at the i^{th} site ($k = 1, \ldots, K_{ij}$)
- $Y_{ijk}|a_{ij} \sim Bernoulli(a_{ij}p_{ijk})$
- a_{ij} is a realized value of A_{ij}
- p_{ijk} denotes 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 independent, then $Y_{ij} = \sum_{k=1}^{K_{ij}} Y_{ijk}$
- $Y_{ij}|a_{ij} \sim Binomial(K_{ij}, a_{ij}p_{ij})$

- p_{ij} is the conditional probability of detection of the target species in each replicate of the j^{th} sample collected at the i^th location, given that the target species is present in that sample
- $\boldsymbol{\delta}$ are the sample level regression parameters for p_{ij}
- $\boldsymbol{v}_{ij}$ are the sample level covariates for p_{ij}

The assumptions are:

- The species is not misidentified, no false positives
- no un-modeled heterogeneity in the probabilities of detection and occupancy
- each survey is closed to changes in occupancy over the sampling period
- the detection of the species is independent for each survey

In WILD 502 when talking about multi-season occupancy models, we talked about extirpation and colonization rates, but I think that these could be modeled with a latent variable(s)? I don't think they are of particular interest here.

3.3 Implementation

Package options for Multi-season (should we even be using these models or should we be modeling the time component in another way) single-species occupancy models:

- nimble.dynamic.occ
- STAN
- JAGS
- wiquid package??
- Frequentist Options:
  - unmarked
  - Program MARK
- write my own package?

Package for eDNA data:

`msocc` package

4 Analysis

4.1 Analysis of eDNA Data

analysis and results

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. (2017). `ednaoccupancy`: An `R` package for multiscale occupancy modelling of environmental DNA data. *Molecular Ecology Resources*, 18(2):368 – 380.

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.

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.

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.

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

Things to think about for the simulation or questions I have:

- think about how changing the detection probabilities and occupancy probabilities impact the results:
  - High occupancy, high detection
  - High occupancy, low detection
  - Low occupancy, high detection
  - Low occupancy, low detection
- How do we/can we include sample level covariates to account for sampling effort (number of tows, if available)?
- If we are getting more data:
  - How do we account for multiple sampling seasons (years)?
  - How do the assumptions of occupancy models change when we have several sampling years versus only 1

Additional questions, not directly related to the simulation:

- How do we account for this multilevel (for lack of a better word) testing process? For example, sometimes (?) when they find them in the microscope, then they test them using polymerase chain reaction (PCR) to confirm positive identification, and sometimes gene sequencing (? –

looks like it was only used on one observation in the sample data), or scanning electron microscopy (SEM).

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


# load eDNA data
eDNA <- read.csv(
  "C:/Users/mwind/OneDrive/Writing Project_EXTRA/eDNA.csv")
length(levels(eDNA$Lake))
levels(eDNA$Lake) <- c("BOM", "LG", "MG")
names(eDNA)
eDNA <- eDNA %>%
  filter(Site != "tb")


# number of samples where there were (0, 3] positive droplets
eDNA %>%
  count(Positive.Droplets <= 3 & Positive.Droplets > 0)

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


# 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 = "10 sample rows of the eDNA data.",
             align = 'c', row.names = F) %>%
  kable_styling(latex_options =
                  c("scale_down", "hold_position"))


# summary of positive droplets by lake
tapply(eDNA$Positive.Droplets, eDNA$Lake, summary)


# summary of water temperature by lake
tapply(eDNA$Water.Temp, eDNA$Lake, summary)


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

BOM <- eDNA.df %>%
  filter(Lake == "BOM") %>%
  ggplot(aes(x = Site,
             y = Positive.Droplets,
             colour = Date.Collected)) +
  labs(title = 'Positive ddPCR replicates at Lake BOM by Site',
       x = 'Site',
       y = 'Positive Droplets',
       color = 'Date') +
  geom_point(alpha = 0.7,
             position = position_jitter(width = 0.25,
                                        height = 1,
```

```
                                          seed = 03172020)) +
  theme_bw() +
  theme(text = element_text(size = 8),
        panel.border = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "black"))


LG <- eDNA.df %>%
  filter(Lake == "LG") %>%
  ggplot(aes(x = Site,
             y = Positive.Droplets,
             colour = Date.Collected)) +
  labs(title = 'Positive ddPCR replicates at Lake LG by Site',
       x = 'Site',
       y = 'Positive Droplets',
       color = 'Date') +
  ylim(c(-0.01, 3)) +
  geom_point(alpha = 0.7,
             position = position_jitter(width = 0.25,
                                        height = 0.001,
                                        seed = 03172020)) +
  theme_bw() +
  theme(text = element_text(size = 8),
        panel.border = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "black"))


MG <- eDNA.df %>%
  filter(Lake == "MG") %>%
  ggplot(aes(x = Site,
             y = Positive.Droplets,
             colour = Date.Collected)) +
  labs(title = 'Positive ddPCR replicates at Lake MG by Site',
```

```
       x = 'Site',
       y = 'Positive Droplets',
       color = 'Date') +
  ylim(c(-0.01, 3)) +
  geom_point(alpha = 0.7,
             position = position_jitter(width = 0.25,
                                        height = 0.001,
                                        seed = 03172020)) +
  theme_bw() +
  theme(text = element_text(size = 8),
        panel.border = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "black"))

temp <- eDNA %>%
  ggplot(aes(x = Water.Temp,
             y = Positive.Droplets)) +
  geom_point(alpha = 0.7) +
  labs(title = 'Positive ddPCR replicates by Water Temperature',
       x = expression(paste('Water Temperature (',
                            degree, 'C)')),
       y = 'Positive Droplets') +
  theme_bw() +
  theme(text = element_text(size = 8),
        panel.border = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "black"))

grid.arrange(BOM, LG, MG, temp, nrow = 2)
```