Independent Project Report

Wynne Moss

November 28, 2018

The data

SUMMARY OF DATASET: I quantified parasite communities across 10 different pond sites in the East Bay of California; each site was visited 4-6 times within the 2017 summer. At each visit I collected 10-12 individuals from 2 amphibian species. Individuals were measured and parasite infection was quantified.

GROUPING VARIABLES:

- Site (10 ponds) - this is a random variable.
- Population (2 species at each site) - this is a random variable

PREDICTOR VARIABLES:

- Species (2 species) - I think this is a better designation for species
- Body size (snout-vent-length) - A continuous variable (fixed effect) at the individual level
- Developmental stage - A continuous variable at the individual level. . . problematic since development is measured differently in newts vs. frogs
- Sex - A factor variable at the individual level
- Visit - A continuous variable (though could be treated as a factor if the relationship si non linear)

RESPONSE VARIABLE:

- The number of parasites found within an individual. For this analysis, just count number of Echinostoma parasites. Can be modeled as Poisson distributed with a negative binomial to account for aggregation. Infection status (1 or 0) could also be modeled as a binary response variable using logistic regression.

Questions

1) Does the impact of species and body size change over the course of the summer (interact with visit)?
2) How much of the variation in overall parasite load is explained by visit-level, species-level, site-level, or individual-level variation?

Display the structure of the data

```
dis <- read.csv("diss.data.2017.csv")
# colnames(dis)
# str(dis)
```

Add some covariates

```
dis$Pop <- paste(dis$SiteCode, dis$SpeciesCode, sep = "_")
dis$fsample <- factor(dis$visit) # for random effect
# each population is sampled multiple time so the observations are nested within pop
```

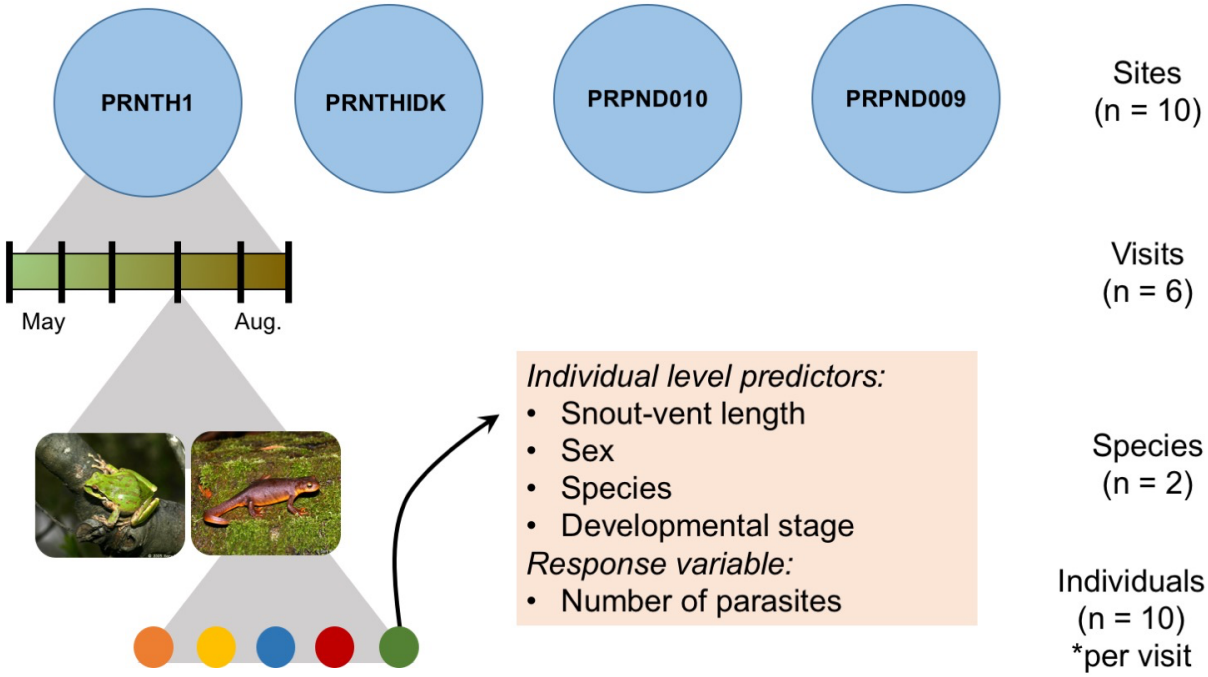


Figure 1: Experimental Design

```
# let's get a scaled SVL (snout-vent length)
# in this case we want negatives to be lower than average at that time, and positives to be higher than
meanSVLVisitSpec <- dis %>% group_by (visit, SpeciesCode) %>% summarise(meanVSVL = mean(SVL), sdVSVL = s
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

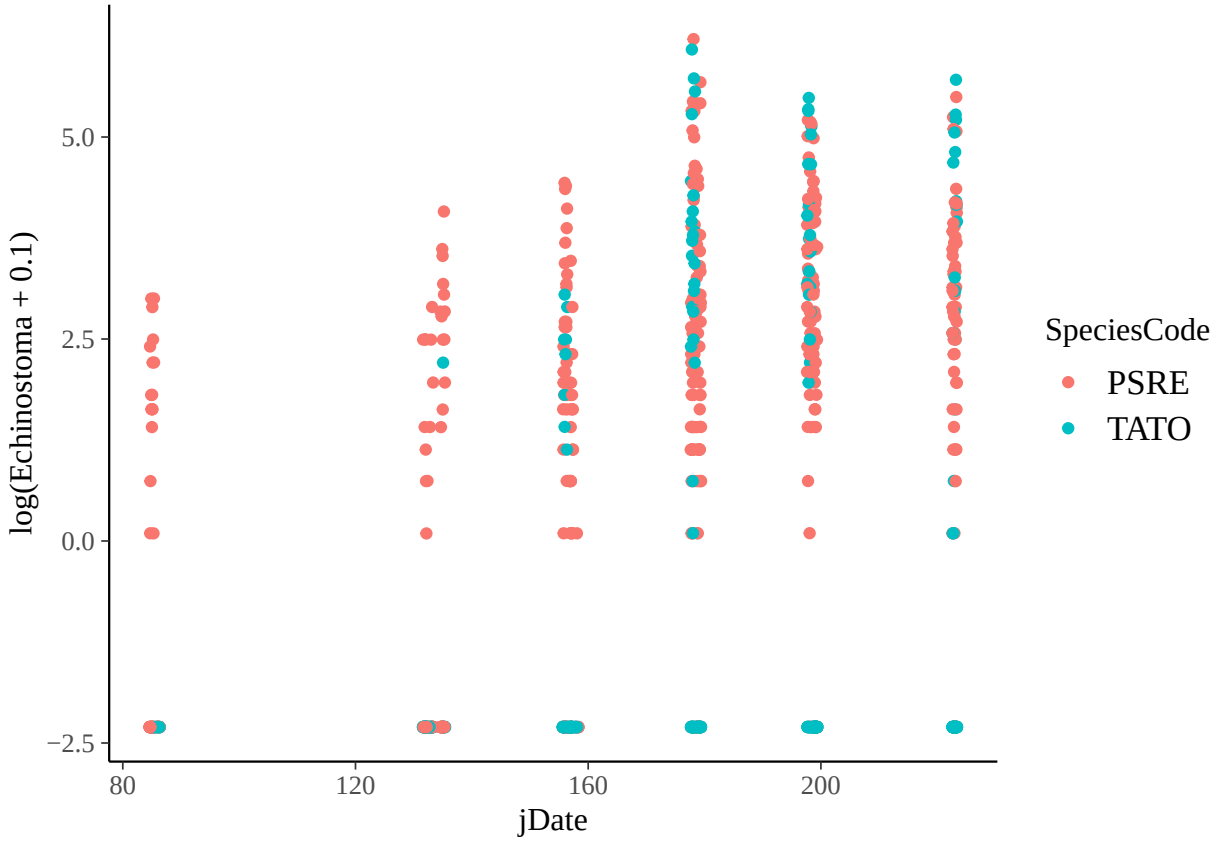
```
dis <- left_join(dis, meanSVLVisitSpec, by = c("visit", "SpeciesCode"))
dis <- dis %>% mutate(visitScaledSVL = (SVL-meanVSVL)/sdVSVL)
dis$meanVSVL =NULL
dis$sdVSVL = NULL

#add in lat and long
sitedat <- read.csv("data/CoreSites.csv"
)
dis <- left_join(dis, sitedat, by = "SiteCode")

# format visit as a date
# first need to split out the part after the underscore (date of sample)
spl <- strsplit(as.character(dis$CollectionCode), "_")
# then format as date
dis$Dated <- sapply(spl, function(x) x[2]) %>% as.Date("%Y%m%d")

# get julian date
dates <- as.POSIXlt(dis$Dated, format = "%Y%m%d")
dis$jDate <-dates$yday
```

Exploratory data analysis

Look at the distribution of echinostoma over time by site. SVL vs. load.

How does parasite load change over time?

```
ggplot(data = dis) +
  geom_jitter(aes(x=jDate, y = log(Echinostoma+.1), color = SpeciesCode))
```

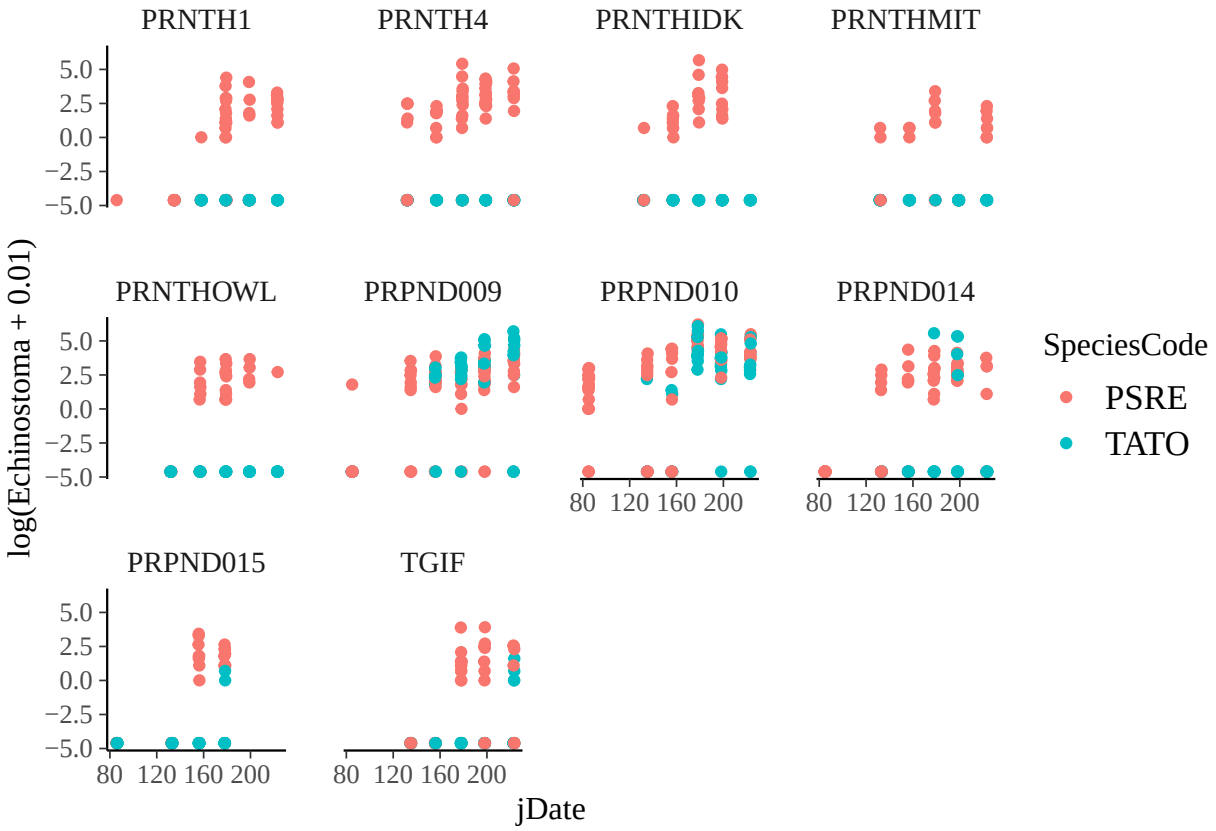



```
# ggplot(data = dis) +
#   geom_jitter(aes(x=visit, y = tot.para, color = SpeciesCode))
```

It kind of looks like Echinostoma load increases over time and that TATO have fewer, but this relationship isn't super obvious.

Let's try separating out by site to reduce some of this noise

```
ggplot(data = dis) +
  geom_jitter(aes(x=jDate, y = log(Echinostoma+.01), color = SpeciesCode))+
  facet_wrap(facets = ~SiteCode)
```

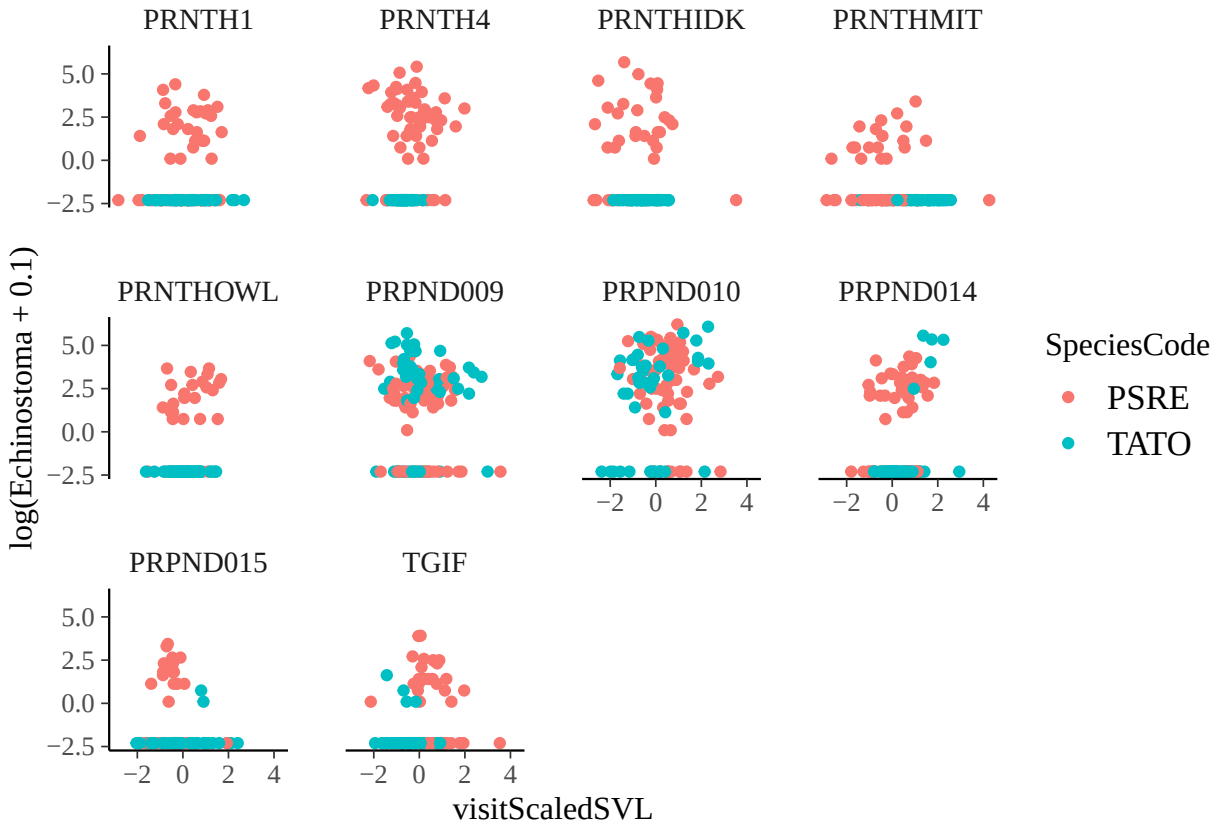


It looks like newts (TATO) don't seem to get Echinostoma at some of the sites (weird!)

In some sites, it seems like Echinostoma count goes up with time. So there might be a site X species interaction. And there might be a site X visit interaction. It also seems like the difference between species changes over time, but that only happens in some sites (e.g. PRPND010). Yikes... that seems like a 3 way interaction. Or the interaction between visit*species depends on site (random slope)?

Do individuals' relative size (are they late bloomers or early bloomers) impact their parasite counts?

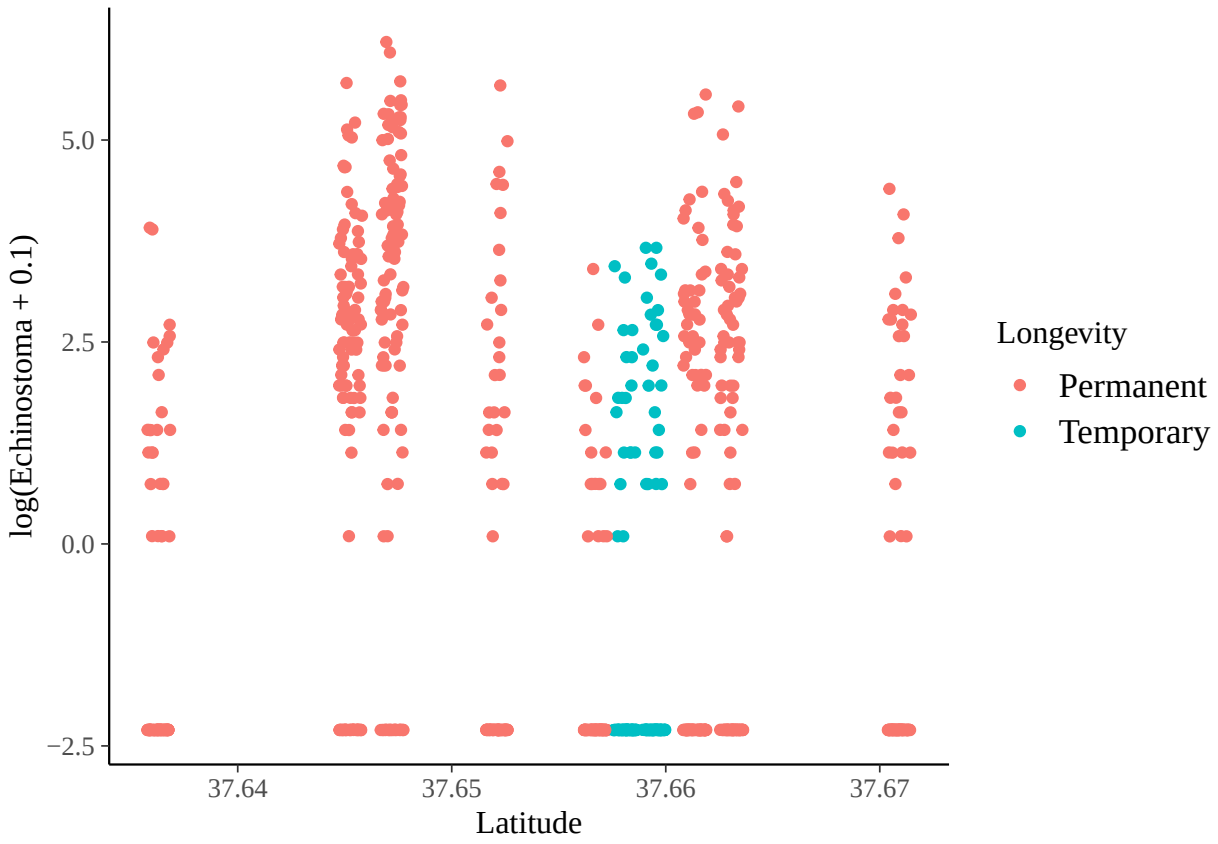
```
ggplot(data = dis)+
  geom_jitter(aes(x = visitScaledSVL, y = log(Echinostoma + .1), color = SpeciesCode))+
  facet_wrap(facets= ~SiteCode)
```



Doesn't seem like it.

Does location of a pond or its longevity matter?

```
ggplot(data=dis)+
  geom_jitter(aes(x=Latitude, y = log(Echinostoma+.1), color = Longevity))
```



Model formulation

For now, I'll focus on question 1 with the response variable being `Echinostoma` (number of Echinostoma parasites found within that individual) *I think that site should be nested within visit because I think that this is a higher level; e.g. ALL sites will be higher at visit 3, etc. But I'm not totally sure about this.*

Individual level

At the individual level, parasite load (y_i) is Poisson distributed with an expected mean of μ_i. Technically the data are negative-binomially (or overdispersed Poisson) - distributed, but I deal with that below by adding an extra error term, ϵ.

Equation 1: $y_i \sim \text{Poisson}(\mu_i)$

The log(expected mean parasite level) is predicted by individual-level covariates: date, species, snout vent length, and interaction, with each individual deviating from expected by ϵ_i.

Equation 2: $log(\mu_i) = \alpha_{j[i]} + \beta_{j[i]} \times date + \beta_2 \times species + \beta_3 \times SVL + \beta_4 \times visit \times species + \epsilon_i$

Those deviations follow a normal distribution centered around 0 with variance of σ^2_ϵ. This is the overdispersion parameter.

Equation 3: $\epsilon_i \sim \text{Normal}(0, \sigma^2_\epsilon)$

Sampling event level

Each sampling event mean (α_j) can be predicted from a linear relationship with date, but varies by a random amount. The relationship with date is site specific (random slope)

Equation 4: $\alpha_{j[i]} \sim \text{Normal}(\mu_{\alpha[j]}, \sigma_j^2)$

Equation 5: $\mu_{\alpha[j]} = \gamma_{k[i]} + \beta_j[i] \times date$

Site level

Each site varies randomly from the average in the mean parasite load. The mean for a particular site is drawn from a distribution centered at the mean among sites ($\bar{\gamma}$):

Equation 6: $\gamma_k \sim \text{Normal}(\bar{\omega}, \sigma_\omega^2)$

The effect of visit depends on site, thus we have a random slope as well, at the site-level. I don't know what greek symbols to use here...

Equation 7 $\beta_j \sim \text{Normal}(\bar{\beta}, \sigma_\beta^2)$.

If there were site level covariates (in the future), $\bar{\gamma}$ could be predicted with another linear model with site-level covariates (e.g. snail density, size, location).

Stan formulation

Model 1: This model includes interactions between time and species and time and SVL. I think that the differences between species or the importance of SVL could vary over the course of the summer.

```
stan.fit <- stan_glmer(Echinostoma ~ visit*SpeciesCode + visit*scale(SVL) + (1|SiteCode) + (1|Collectio
stan.fit1 <- stan.fit
summary(stan.fit1)
# launch_shinystan(stan.fit)
plot(stan.fit1)
stan.samp1 <- sample(stan.fit1)
save(stan.samp1, file = "stan.fit.1.RData")
```

The interactions aren't significant (overlap with zero) so I will remove and re-fit. I'm also going to add a random slope model because I think the effect of species depends on site. The two species are very similar at some sites and very different at other sites.

Can I have two random slopes for Site, if so how to specify? jDate|SiteCode + SpeciesCode|SiteCode doesn't work.

```
stan.fit2 <- stan_glmer(Echinostoma ~ visitScaledSVL + (jDate|SiteCode) + SpeciesCode*jDate + (1|Collec

summary(stan.fit2)
# launch_shinystan(stan.fit2)
```

Try a third model; get rid of longevity and latitude; they're not significant. Let's add that interaction.

Technically this isn't the best way of model comparisons since I'm not doing this in sequence with one term dropped or added at a time. But these models take 20 mins to run so I am rowing impatient.

```
stan.fit3 <- stan_glmer(Echinostoma ~ visit + SpeciesCode + visitScaledSVL + (SpeciesCode*visit|SiteCod

summary(stan.fit3)
launch_shinystan(stan.fit3)
```

Exploring model fit

```
# variance partitioning
```

Posterior predictive checks. Simulating the model

Visualization

What are the best ways to represent these data? Should I do residual plots?

Conclusions

Next steps: Perhaps use a random slope model as well (visit|SiteCode). Developmental stages could perhaps be included with an interaction with species, which would allow TATO to have some levels and PSRE to have other levels. e.g. I need to put them in the same column. Will autocorrelation of fixed and random effects (e.g. Species, Population) become an issue?