Models for data with too many zeros: Fish in Florida ranches' wetlands

Species counts are particularly difficult to analyze because commonly zero inflation results from having far more zeros that what would be expected for Poisson or Negative Binomial distributions. We use data collected by Bohlen *et al.* (2014) aimed at understanding the effect of hydrology on species abundance to evaluate government policies encouraging water retention. They used a stratified random sampling method to gather data on abundance of several organisms in wetlands within four ranches in Highlands and Okeechobee Counties in Florida, USA. Here, we focus on the abundance of fish (Figure 1). For the first analysis we ignore the hierarchical nature of the sampling among wetlands within ranches, then we incorporate wetland variation.

Figure 1. Above: View of one of the sampled wetlands. Below: Female and male Mosquito fish (*Gambusia affinis*), one of the species found in our samples.

Bohlen *et al.* (2014) proposed hypotheses on the shape of the responses of organisms to wetland water depth; in particular they expected a unimodal distribution for fish abundance (Figure 2). They also predicted that fish abundance will vary among ranches because of management history and local attributes.

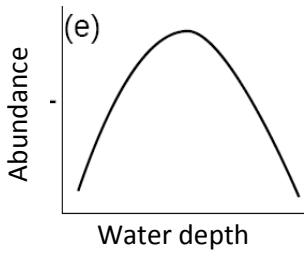


Figure 2. Hypothesis of change of fish abundance with water depth

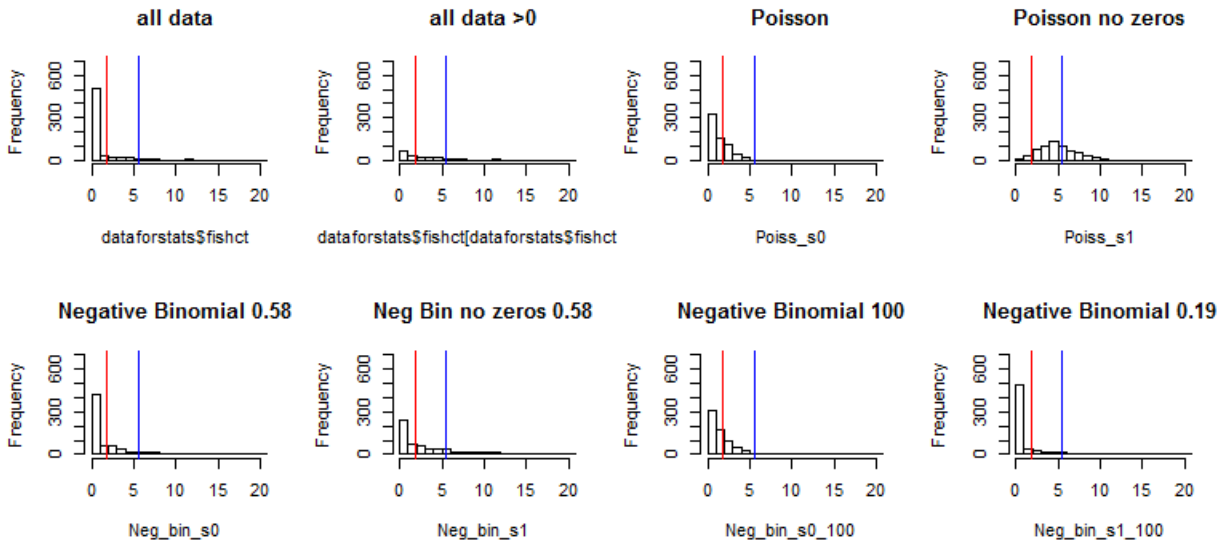


Figure 3. Histograms of fish abundance for (i) the whole sample, (ii) those samples where fish were observed, (iii) expected frequency based on a Poisson distribution based on the overall mean, (iv) expected frequency based on a Poisson distribution based on the mean of samples > 0, (v) expected frequency based on a negative binomial distribution based on the overall mean and dispersal parameter = 0.58, (vi) expected frequency based on a negative binomial distribution based on the mean of samples > 0 and dispersal parameter = 0.58, (vii) expected frequency based on a negative binomial distribution based on the overall mean and dispersal parameter = 100, (viii) expected frequency based on a negative binomial distribution based on the overall mean and dispersal parameter = 0.19. In red the location of the overall mean, in blue that of the data > 0. *Data is truncated to occurrences < 20 fish per sample.*

The origin of zeros

The numbers of fish caught per sample was extremely variable. Most frequently no fish were caught, but 70 fish were caught at once in one occasion. A Poisson distribution with the same mean as the one observed in this study expects 297 zeros, not close to the 509 zeros observed. The prediction of 450 from the Negative Binomial with dispersal parameter 0.58 is closer but Zuur *et al.* (2009) caution that ignoring zero inflation biases the standard errors and causes over-dispersion. There are techniques that deal with this excess of zeros, but they require understanding the nature of the zeros. We use the classification described in Martin *et al.* (2005) to classify those encountered in our study.

1. Structural zeros: Fish were not present because the habitat was not suitable for them.
2. Design errors: Fish were not found because of poor experimental design or sampling practices.
3. Observer error: Fish were there but they were not seen.
4. Organism error: The habitat was suitable but fish were not there.
5. Bad zeros: Sampling outside the species range, for example fish out of water.

Zeros due to design, observer and impossible species range are called false zeros or false negatives and we should do our best to avoid those (Zuur *et al*. 2009). Researchers have little control of organismal error but it can be minimized with better designs. Structural zeros are called positive or true zeros, but these definitions are open to discussion (Martin *et al.* 2005). We recognize that

our study probably includes false negatives, for example, the methods we used did not sample large fish well. We did try to minimize design and observer error by having experienced biologists collect and retrieve the samples.

There are several possible models to analyze our fish data. Here, we will use three alternative formulations. We can be optimistic and assume that there were no zero-inflation in our data. The difference between the Poisson and negative binomial is that the second allows for additional over-dispersion in the data (Zuur *et al*. 2009). We will also evaluate a zero-inflation formulation. Zeros in mixed zero-inflated models are modeled as coming from two different processes: the binomial and the count processes, where the binomial portion estimates the probability of false zeros versus all other type of data (counts and true zeros). In this formulation it is possible to use different sets of covariates to explain the occurrence and the counts. You can use information criteria to select the most informative model, but we agree with Zuur *et al.* (2009) that it is better to include biological knowledge to decide among them. We use the mixed zero-inflated type of models because we are convinced on the existence of genuine structural zeros. We expect the shallower and the deeper areas in the wetlands not be as suitable for fish as intermediate depths.

We specify possible model structures for the effect of depth and ranch on fish count. Models assume an effect of depth (as a quadratic variable) and ranch on fish counts, but *m1* and *m2* assume that there was no zero-inflation. Model *m3* evaluates the possibility of zero-inflation, evaluating count variation as in the other models but the occurrence of extra zeros as a function of depth.

```
m1 <- map2stan(
  alist(
    fishct ~ dpois(lambda),
    log(lambda) <- a + b*depth + c*depth2 + d*r2 + e*r3 + f*r4,
    c(a,b,c,d,e,f) ~ dnorm(0,10)
          ),
  data = dataf, chains =3, start=list(a=0,b=0,c=0,d=0,e=0,f=0),
)

m2 <- map2stan(
  alist(
    fishct ~ dgampois(pbar,scale),
    log(pbar) <- a + b*depth + c*depth2 + d*r2 + e*r3 + f*r4,
    c(a,b,c,d,e,f) ~ dnorm(0,10),
    scale ~ dcauchy(0,2)
  ),
  data = dataf,
  constraints =list(theta ="lower=0"),
  start=list(a=0, b=0,c=0,d=0,e=0,f=0,scale=3),
  iter=4000, warmup=1000, chains =3
)

m3 <- map2stan(
  alist(
    fishct ~ dzipois(pbar,lambda),
    logit(pbar) <- z1 + z2*depth,
    log(lambda) <- a + b*depth + c*depth2 + d*r2 + e*r3 + f*r4,
    c(a,b,c,d,e,f) ~ dnorm(0,10),
    c(z1,z2) ~ dnorm(0,10)
  ),
  data = dataf, start=list(a=0,b=0,c=0,d=0,e=0,f=0,z1=0.1,z2=0.1),
  iter=4000, warmup=1000, chains =3
)
```

```
> compare(m1,m2,m3)
     WAIC pWAIC  dWAIC weight     SE    dSE
m3 2634.6  44.2    0.0      1 229.46     NA
m2 3966.0 299.0 1331.4      0 298.10 132.13
m1 4348.8  63.7 1714.2      0 412.09 224.27
```

We identify model *m3* as the most plausible in our set using WAIC. The summary of model *m3* is presented below (Figure 4).

$$P(y_i = 0) = \pi_i + (1 - \pi_i) \times \left(\frac{k}{\mu_i + k}\right)^k$$

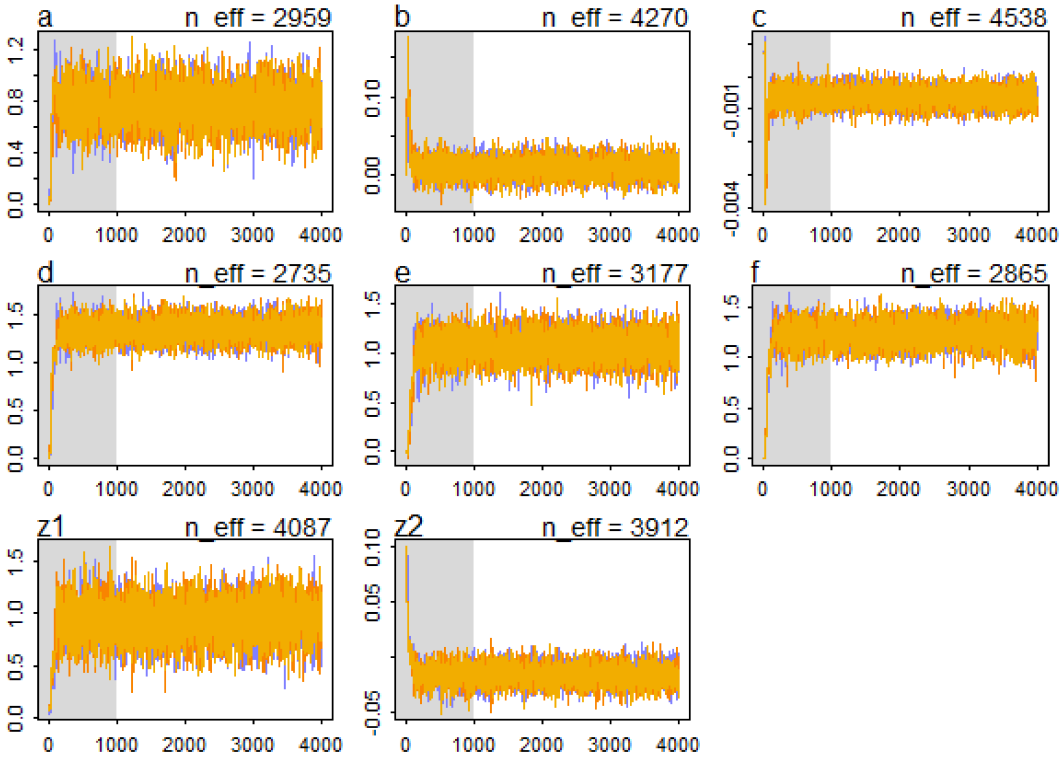
$$P(y_i > 0) = (1 - \pi_i) \times fNB(y)$$

$$\mu_i = \alpha + \beta_1 \times depth$$

$$\pi = \frac{e^{-\gamma_i}}{1 + e^{-\gamma_i}}$$

$$\gamma_i = -\alpha + \beta_1 \times depth + \beta_2 \times depth^2 + \beta_{i\,ranch}$$

$$fNB(y) = P(y_i; k, \mu_i | y_i \geq 0) = \frac{Fac(y_i + k)}{Fac(k) \times Fac(y_i + 1)} \times \left(\frac{k}{\mu_i + k}\right)^k \times \left(1 - \frac{k}{\mu_i + k}\right)^{y_i}$$



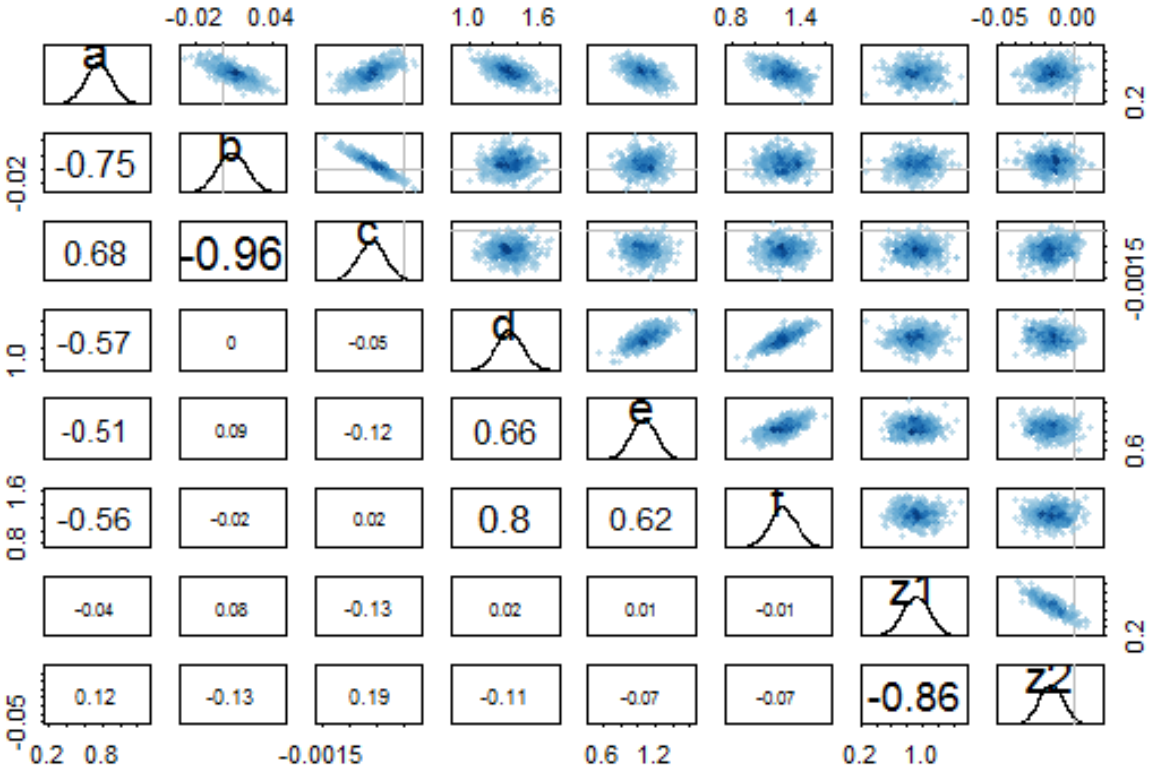


Figure 4. On top the panel shows the sequence of the generating chains and at the bottom the distribution and the correlation of the parameters of the model *m3*.

See Zuur et al (2009) for the details of the negative binomial function $fNB(y)$; Fac = Factorial.

Count model coefficients

	Mean	StdDev	lower 0.89	upper 0.89	n_eff	Rhat
a	0.763	0.149	0.523	0.999	2959	1.000
b	0.008	0.012	-0.011	0.027	4270	1.000
c	-0.001	0.000	-0.001	0.000	4538	1.000
d	1.338	0.107	1.174	1.517	2735	1.000
e	1.064	0.138	0.846	1.281	3177	1.000
f	1.240	0.111	1.066	1.422	2865	1.000

Zero-inflation model coefficients (binomial with logit link):

z1	0.918	0.174	0.647	1.201	4087	1.001
z2	-0.016	0.008	-0.029	-0.003	3912	1.001

A plot of model *m3* is presented against the background of the data using the code in the demo (Figure 5). We conclude that the effect of depth on fish counts vary among ranches and depth affects abundance in a unimodal pattern.

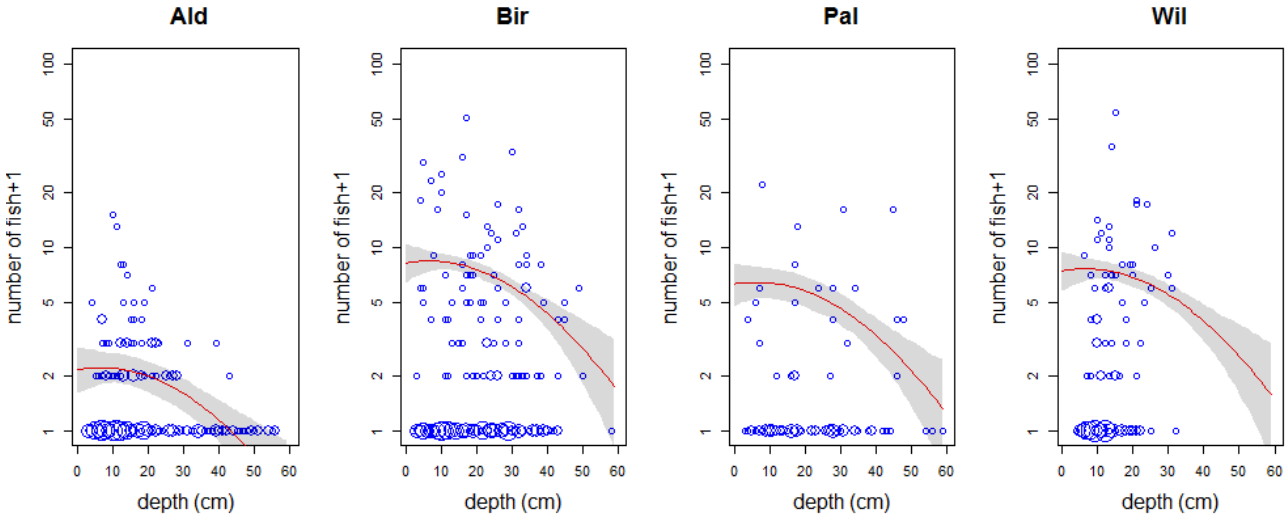


Figure 5. Number of fish +1(in blue) as a function of depth (in cm) by ranch. The size of the symbol is related to its frequency in the sampling. Model m3 is depicted in red.

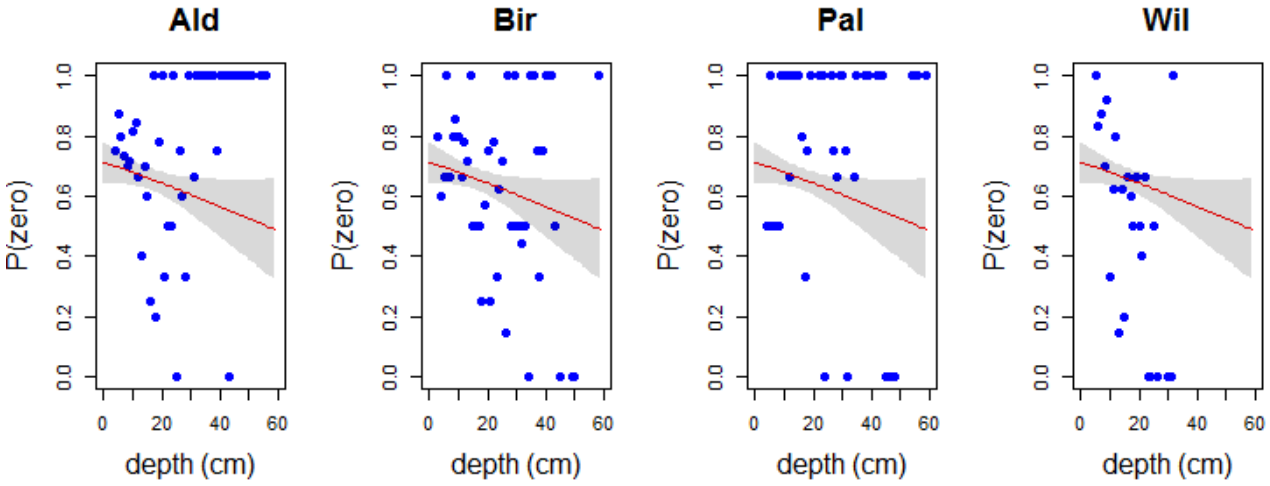


Figure 5. Probability of not finding fish as a function of depth (in cm) by ranch. Model m3 is depicted in red.

NOTE: all the materials for this demo can be found at:

https://sciences.ucf.edu/biology/d4lab/methods-2/

References

Martin, T.G., B. A. Wintle, J.R. Rhodes, P.M. Kuhnert, S.A. Field, S.J. Low-Choy, A.J. Tyre, and H. P. Possingham (2005) Zero tolerance ecology: improvement ecological inference by modeling the source of zero observation. Ecology Letters 8: 1235-11246.

McElreath, R.M. 2016. Statistical Rethinking: a Bayesian course with examples in R and Stan. Chapman and Hall.

Patrick J. Bohlen, Elizabeth Boughton, John E. Fauth, David Jenkins, Greg Kiker, Pedro F. Quintana-Ascencio, Sanjay Shukla, and Hilary M. Swain. 2014. Assessing Trade-Offs among Ecosystem Services in a Payment-for-Water Services Program on Florida Ranchlands Final Report. USA Environmental Protection Agency.

Sujit.K.G., P. Mukhopadhyay, J-C Lu. 2006. Bayesian analysis of zero-inflated regression models. Journal of statistical planning and inference 136: 1360-1375.

Zuur, A.F., E.N. Ieno, N.J. Walker, A. Savaliev, G.M. Smith. 2009. Mixed effects models and extensions in Ecology with R. Springer.