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 this analysis we ignore the hierarchical nature of the sampling among wetlands within ranches.

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.

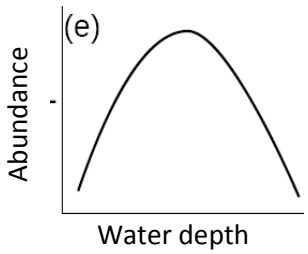

Figure 2. Hypothesis of change of fish abundance with water depth

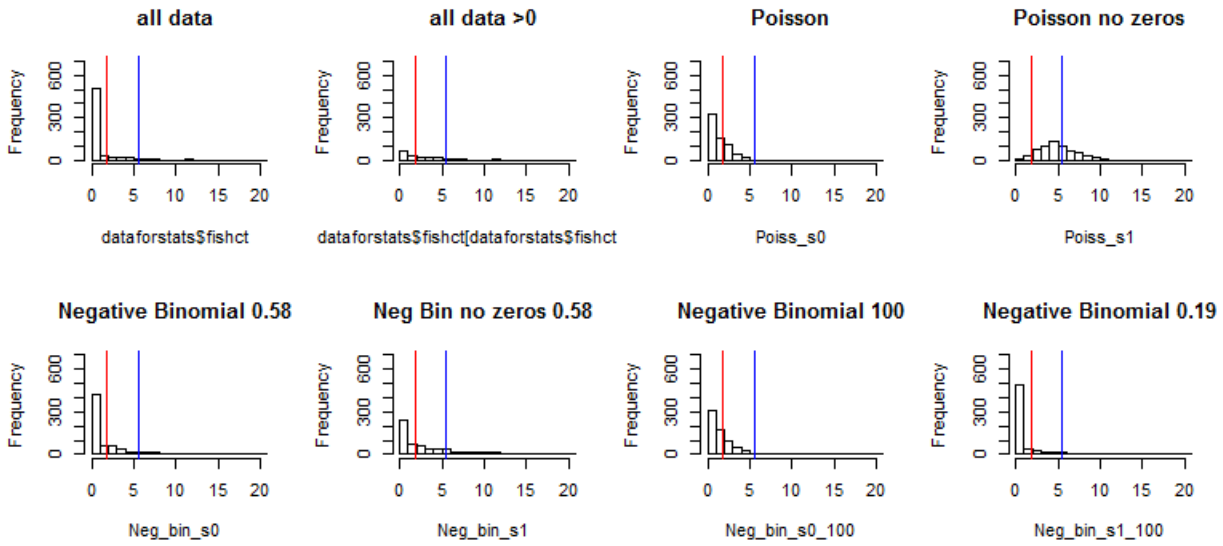

Figure 3. Histograms of fish abundance for (i) the whole sample, (ii) those samples where fish were observed, (iii) expected frequency based on a Poisson distribution based on the overall mean, (iv) expected frequency based on a Poisson distribution based on the mean of samples > 0, (v) expected frequency 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 four possible models to analyze our zero-inflated fish data. The difference between the Poisson and negative binomial is that the second allows for additional over-dispersion in the positive (non-zero) part of the data (Zuur *et al*. 2009). There are also two different ways to consider the zeros: mixture zero-inflated models (known as Zi and Ni models for Poisson and negative binomial distributions, respectively) and two part zero-altered models (Za and Na) (Zuur *et al*. 2009). Zeros in mixed zero-inflated models are modelled as coming from two different processes: the binomial and the count processes, where the binomial GLM estimates the probability of false zeros versus all other type of data (counts and true zeros). In zero-altered models, the non-zero observations are modelled with a truncated Poisson or negative binomial and therefore the count process does not allow for zeros. In both cases it is possible to use different sets of covariates to explain the occurrence and the counts. You can use AIC criteria to select the most informative model, but we agree with Zuur *et al*. (2009) that it is better to use biological knowledge to decide among them. In this case, we only 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 to not be as suitable for fish as intermediate depths.

We start by specifying five possible structures for the effect of depth and ranch on fish count. Anything before the | is the structure of the count model i.e. how many fish, while terms after the | refer to the structure of the binomial model i.e. are there fish or not. If the last part is not specified, the default is to use the same structure for both components (i.e. f1 and f2 below are the same). Formulas 1 to 4 assume an interactive effect of depth (as a quadratic variable) and ranch, but f3 assumes the presence of fish in only affected by depth while f4 assumes it is only affected by ranch. F5 assumes an additive effect of depth and ranch, and f6 assume only depth is important in the count and zero models.

```
f1 <-formula(fishct~depth*ranchn+depth2*ranchn)
f2 <-formula(fishct~depth*ranchn+depth2*ranchn | depth*ranchn+depth2*ranchn)
f3 <-formula(fishct~depth*ranchn+depth2*ranchn | depth)
f4 <-formula(fishct~depth*ranchn+depth2*ranchn | ranch)
f5 <-formula(fishct~depth+depth2+ranchn | depth)
f6 <-formula(fishct~depth+depth2 | depth)
```

Then we decide whether to allocate a Poisson or negative binomial distribution to the count models and end up with the following ten models (notice Zi1 = Zi2 and Ni5 = Ni6).

```
Zi1 <- zeroinfl(f1,dist="poisson", link="logit", data = dataforstats)
Zi2 <- zeroinfl(f2,dist="poisson", link="logit", data = dataforstats)
Zi3 <- zeroinfl(f3,dist="poisson", link="logit", data = dataforstats)
Zi4 <- zeroinfl(f4,dist="poisson", link="logit", data = dataforstats)
Ni5 <- zeroinfl(f1,dist="negbin", link="logit", data = dataforstats)
Ni6 <- zeroinfl(f2,dist="negbin", link="logit", data = dataforstats)
Ni7 <- zeroinfl(f3,dist="negbin", link="logit", data = dataforstats)
Ni8 <- zeroinfl(f4,dist="negbin", link="logit", data = dataforstats)
Ni9 <- zeroinfl(f5,dist="negbin", link="logit", data = dataforstats)
Ni10 <- zeroinfl(f6,dist="negbin", link="logit", data = dataforstats)
```

We identify model Ni5 as the most plausible in our set using AIC (after removing the redundant formulations). The summary of model Ni5 is presented below:

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

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

$$\mu_i = -\alpha + \beta_1 \times depth + \beta_2 \times depth^2 + \beta_{3\,ranch} \times depth + \beta_{4\,ranch} \times depth^2$$

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

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

See Zuur et al (2009) for the details of the negative binomial function $fNB(y)$; Fac = Factorial. Notice that the coefficient for the binomial portion are for P(0) meaning that negative values for the slope represent an increase of the response not a decrease (Thanks to Johnny).

```
AICtab(Zi1,Zi3,Zi4,Ni5,Ni7,Ni8,Ni9,Ni10,weights=TRUE,base=TRUE)
```

```
     AIC    dAIC   df weight
Ni5  1899.3    0.0 25 0.8053
Ni7  1902.4    3.1 15 0.1710
Ni9  1907.3    8.0 9  0.0144
Ni8  1908.2    8.9 17 0.0093
Ni10 1937.7   38.5 6  <0.001
Zi1  2544.3  645.0 24 <0.001
Zi3  2549.7  650.4 14 <0.001
Zi4  2559.0  659.7 16 <0.001
```

FORMULA FOR STATISTICAL MODEL

```
Call:
zeroinfl(formula = f1, data = dataforstats, dist = "negbin", link = "logit")

Pearson residuals:
     Min       1Q   Median       3Q      Max
-0.63917 -0.46581 -0.37015 -0.03652  7.44054
```

Count model coefficients (negbin with log link):

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	0.393599	0.671864	0.586	0.5580	
depth	0.023026	0.058313	0.395	0.6929	
ranchn2	1.809072	0.938194	1.928	0.0538	.
ranchn3	1.050350	1.114154	0.943	0.3458	
ranchn4	-0.454490	1.490929	-0.305	0.7605	
depth2	-0.002257	0.001150	-1.962	0.0497	*
depth:ranchn2	-0.042719	0.086738	-0.493	0.6224	
depth:ranchn3	-0.043722	0.097292	-0.449	0.6531	
depth:ranchn4	0.161196	0.164870	0.978	0.3282	
ranchn2:depth2	0.001941	0.001750	1.109	0.2673	

```
ranchn3:depth2  0.002670   0.001769   1.510   0.1311
ranchn4:depth2 -0.001938   0.004068  -0.476   0.6338
Log(theta)     -0.549258   0.237159  -2.316   0.0206 *
```

```
Zero-inflation model coefficients (binomial with logit link):
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.178401   2.393391   0.492    0.622
depth          -0.011629   0.542144  -0.021    0.983
ranchn2         0.080079   2.468153   0.032    0.974
ranchn3        -0.493727   2.430498  -0.203    0.839
ranchn4         1.702906   2.955669   0.576    0.565
depth2         -0.012371   0.029210  -0.424    0.672
depth:ranchn2  -0.065229   0.546626  -0.119    0.905
depth:ranchn3  -0.001005   0.541265  -0.002    0.999
depth:ranchn4  -0.305828   0.591997  -0.517    0.605
ranchn2:depth2  0.012887   0.029260   0.440    0.660
ranchn3:depth2  0.012466   0.029221   0.427    0.670
ranchn4:depth2  0.017867   0.030065   0.594    0.552
```

```
Theta = 0.5774
Number of iterations in BFGS optimization: 48
Log-likelihood: -924.6 on 25 Df
```

Notice that none of the parameters of the binomial model are statistically different from zero, so we could instead have favored model Ni7 which has a simpler structure (see below). But still, when we compared the mixed zero-inflated model Ni5 against a model without the binomial portion (using both AIC and the Vuong test), we found significant justification for the inclusion of this structure in the model.

```
Call:
zeroinfl(formula = f3, data = dataforstats, dist = "negbin", link = "logit")

Pearson residuals:
     Min       1Q   Median       3Q      Max
-0.52039 -0.43964 -0.35033  0.01414  7.20437
```

```
Count model coefficients (negbin with log link):
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.089421   0.725026  -0.123   0.9018
depth           0.065824   0.068601   0.960   0.3373
ranchn2         1.982829   0.953347   2.080   0.0375 *
ranchn3         1.328865   1.139154   1.167   0.2434
ranchn4        -0.946936   1.499118  -0.632   0.5276
depth2         -0.003004   0.001432  -2.097   0.0360 *
depth:ranchn2  -0.101185   0.088107  -1.148   0.2508
depth:ranchn3  -0.123540   0.100390  -1.231   0.2185
depth:ranchn4   0.196890   0.171093   1.151   0.2498
ranchn2:depth2  0.003065   0.001797   1.706   0.0881 .
ranchn3:depth2  0.003912   0.001938   2.019   0.0435 *
ranchn4:depth2 -0.002904   0.004319  -0.672   0.5013
Log(theta)     -1.202152   0.219717  -5.471 4.47e-08 ***
```

```
Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.1377     0.6245   1.822   0.0685 .
depth        -0.1556     0.0775  -2.008   0.0446 *
```

```
Theta = 0.3005
Number of iterations in BFGS optimization: 32
Log-likelihood: -936.2 on 15 Df
```

```
model1 <- glm.nb(formula = f1, data = dataforstats, init.theta = 0.58, link = log)

vuong(Ni5,model1)
Vuong Non-Nested Hypothesis Test-Statistic:
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
-------------------------------------------------------------
              Vuong z-statistic             H_A    p-value
Raw                    3.265381 model1 > model2 0.00054658
AIC-corrected          1.254439 model1 > model2 0.10484115
BIC-corrected         -3.268451 model2 > model1 0.00054069

AICtab(Ni5,model1,weights=TRUE,base = TRUE)
       AIC    dAIC   df weight
Ni5    1899.3    0.0 25 1
model1 1914.2   15.0 13 <0.001
```

A plot of model Ni5 is presented against the background of the data using the following code (Figure 4). We conclude that the effect of depth on fish counts vary among ranches, since depth affects abundance in a unimodal pattern for all ranches except Pal.

```
x<-seq(min(dataforstats$depth),max(dataforstats$depth),1)
name <-c(" Ald","Bir","Pal","Wil")
lr <- levels(dataforstats$ranchn)
par(mfrow=c (1,4))
for (k in 1:4) {
plot(dataforstats$depth,(dataforstats$fishct+1),type="n",log="y",
ylim=c(1,100),xlim=c(1,60), main=name[k],xlab= "depth (cm)",
      ylab=("number of fish+1"),cex.lab=1.5,cex.main=1.7)
depth_dat <- dataforstats$depth[dataforstats$ranchn==lr[k]]
fish <- dataforstats$fishct[dataforstats$ranchn==lr[k]]
t <- table(fish,depth_dat)
dep <- unique(depth_dat)
pez <- unique(fish)
ord <- order(dep)
dep <- dep[ord]
orf <- order(pez)
pez <- pez[orf]
y11 <- y22 <- rep(0,length(x))
for (i in 1: length(dep)) {
  for (j in 1: length(pez)) {
     points(dep[i],(pez[j]+1),pch=1,cex=log(t[j,i])+1,col="blue")}}
for (i in 1:length(x)){
y11[i]=predict(Ni5,list(depth=x[i],depth2=x[i]^2,ranchn=factor(k,levels=levels(datafor
stats$ranchn))),type="response")}
lines(x,(y11+1),col="black",lwd=2)}
```

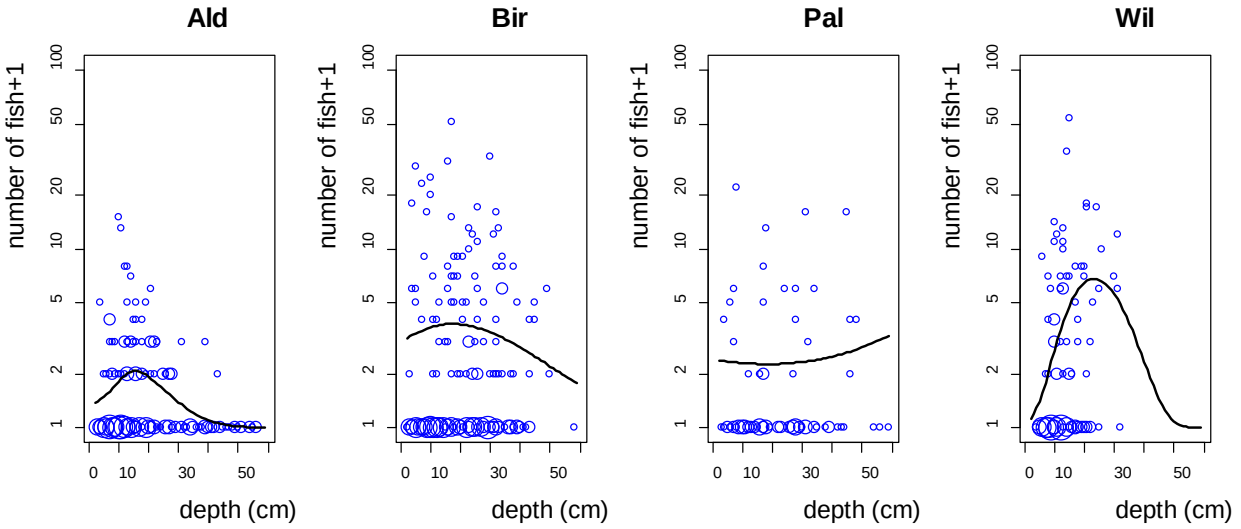

Figure 4. Number of fish +1(in blue) as a function of depth (in cm) by ranch. The size of the symbol is related to its frequency in the sampling. Model Ni5 is depicted in black.

The residuals differ slightly between the zero-inflated model (A) and the one without the correction (B). Both indicate larger residual variation for smaller predicted values and smaller residual variation in Pal.

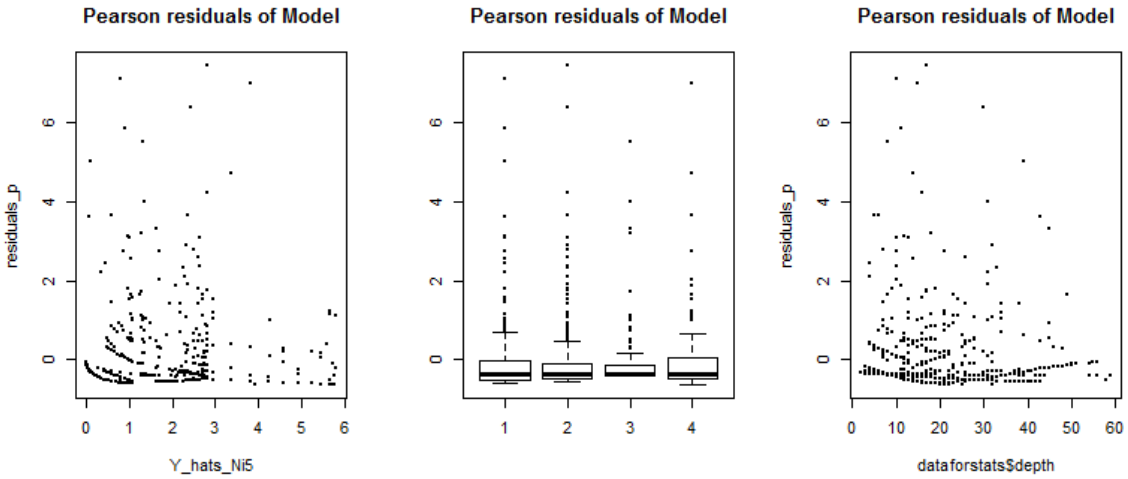

Figure 5 A. Residuals of model Ni5.

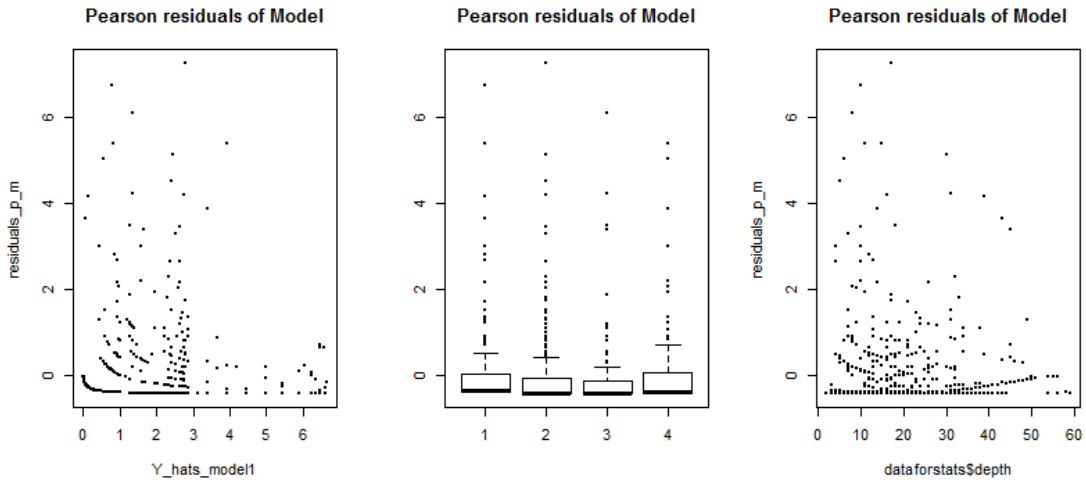

Figure 5 B. Residuals of model1.

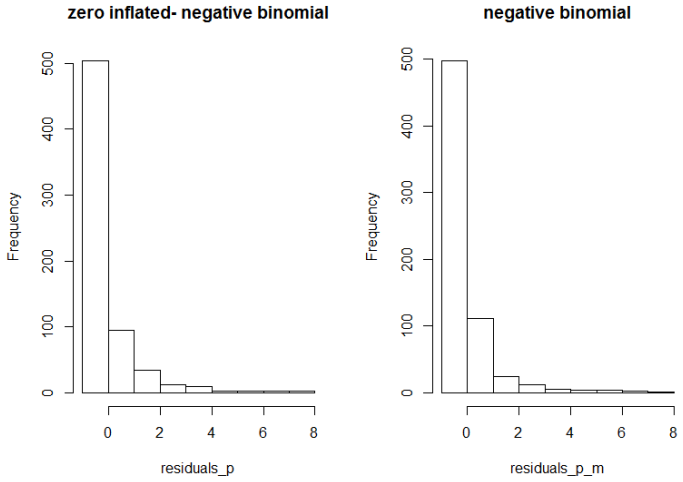

Figure 6. Overall residuals of model Ni5 and model 1.

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.

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.

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.