# How to deal with non-linear count data?
# Macro-invertebrates in wetlands

In this session we will recognize the advantages of making an effort to better identify the proper error distribution of data and choose the most informative type of model to assess our hypotheses. We use data aimed at understanding the effect of hydrology on species abundance to evaluate government policies encouraging water retention in ranchlands (Bohlen et al., 2014 and Boughton et al., 2019). Researchers 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 (Figure 1). Here, we focus on the abundance of macro-invertebrates. We do not use their whole wealth of data and focus on a subset of relevant variables.

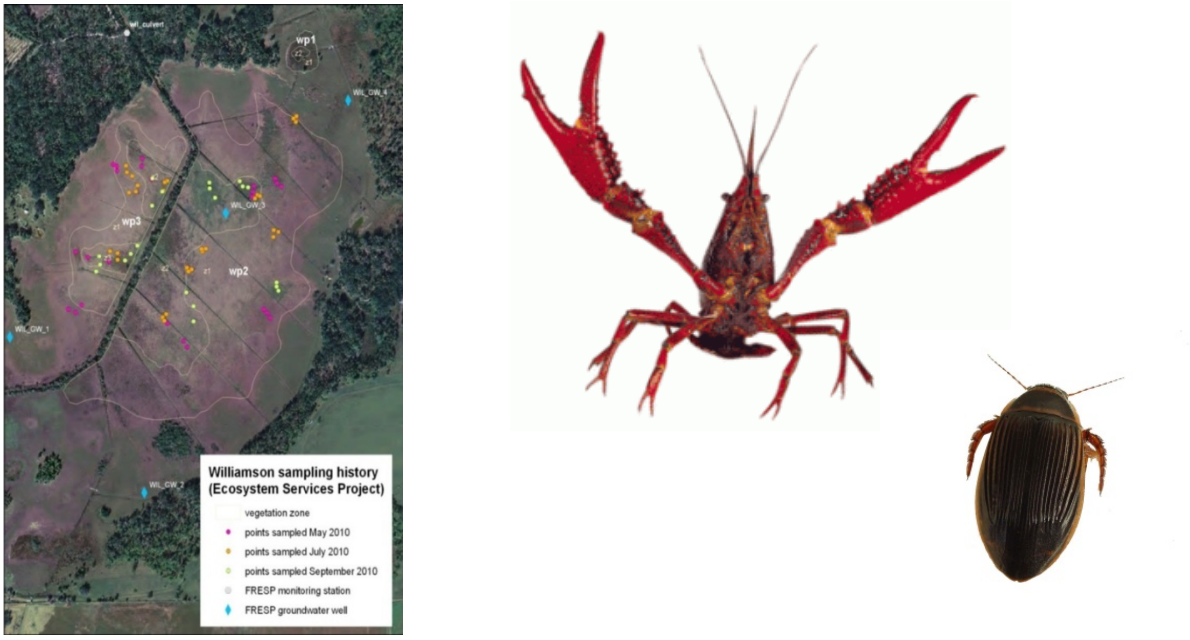


**Figure 1.** Sampling design within a wetland and example organisms

Bohlen et al. (2014) proposed hypotheses on the shape of the responses of organisms to wetland water depth. In particular they expected that macro-invertebrate abundance may increase with water depth to a maximum and then decline (Figure 2). They also predicted that macro-invertebrates abundance may covariate with that of other organisms and that this variation may vary among ranches because of management history and local attributes.

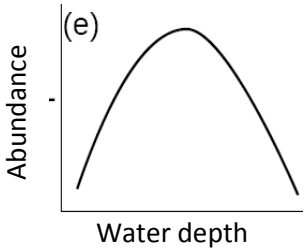


**Figure 2.** Hypothesis of change of macro-invertebrate abundance with water depth

**Model validation: identification of the proper error distribution.**

In this demo we evaluate a model including water depth, abundance of fish, and ranch as explanatory variables of macro-invertebrate abundance. The first two variables are counts and the third is categorical including four ranches. We start with a linear model with Gaussian errors; followed by a generalized linear model with negative binomial errors and then a general additive model with negative binomial errors. These models represent a sample of possible models to assess these data. We emphasize validation of these models as a protocol to compare the quality and amount of information obtained from them. You may wonder how we chose these models. Please reflect on the relevance of the chosen variables for macro-invertebrate abundance in wetlands. Here we focus on the importance of proper likelihood distribution. In this analysis we also do not recognize that data were nested within wetlands by ranch. We will consider this issue in another session.

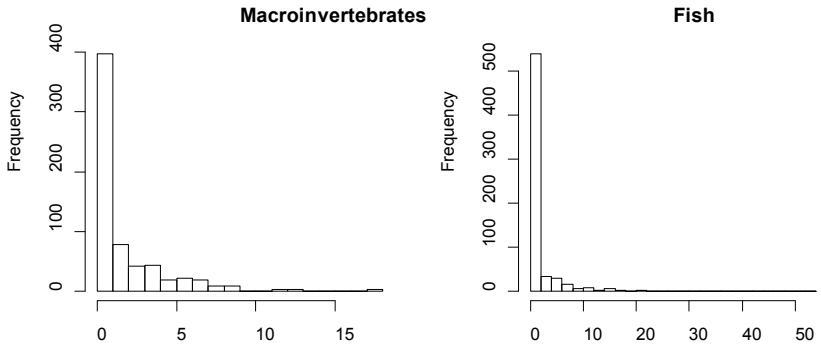


Figure 3. Histogram of Macro-invertebrate and fish abundances

The numbers of macro-invertebrate caught were extremely variable. Most frequently researchers did not catch any, and once they got 18 (Figure 3). Fish were even more variable (Figure 3). For a moment, we ignore the distribution of the data and proceed to evaluate a linear model with Gaussian error. We will see that this is not a good decision but it is, sadly, a relatively common procedure among many ecologists. We use the R platform to retrieve the data and subset them to remove missing values:

```
dataforstats <- read.table("dataforstats_final_BB121613.txt", header=T)
sub1 <-subset(dataforstats,!is.na(upland_elev_m ))
```

We identify the macro-invertebrates as the dependent variable *y*, and depth, fish and ranch as independent variables. We include the variable depth$^2$ to characterize the non-linear nature of the response as hypothesized. We transform fish abundance to its natural logarithm to reduce the leverage of extreme values. We also change the variable that identifies the ranches from strings of names to strings of numbers to facilitate the reading of the output.

```
y <- as.numeric(sub1$macroct)
depth <- sub1$depth
depth2 <- depth^2
fish <- log(sub1$fishct)
rancho <- rep(1,length(sub1$ranch))
rancho[sub1$ranch=="bir"] <- 2
rancho[sub1$ranch=="pal"] <- 3
rancho[sub1$ranch=="wil"] <- 4
rancho <- factor(rancho)
```

Macro-invertebrate abundance was in fact higher in intermediate depths, co-varied with fish and was variable among ranches (Figure 4).

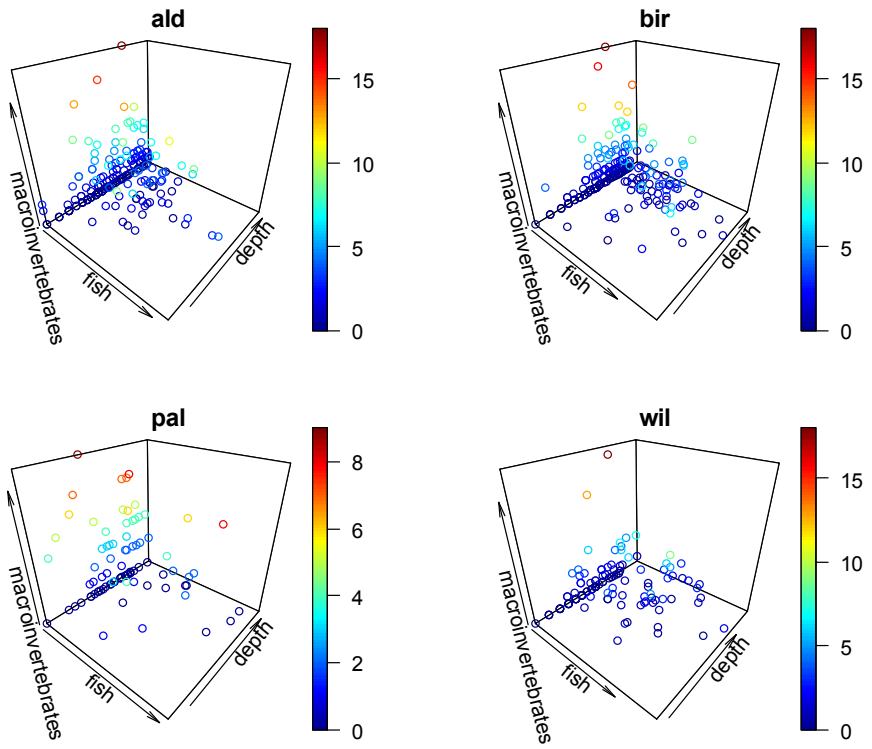


Figure 4. Observed macro-invertebrate abundance with water depth and fish abundance in four ranches in South Florida.

We call the procedure *lm* to obtain a linear model with Gaussian errors that is the default error distribution for the function *lm*.

```
lm(formula = y ~ depth + depth2 + rancho + fish)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7797 -1.7483 -0.9956  0.7213 16.4276

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.1577935  0.3775342   3.067  0.00225 **
depth        0.0955440  0.0341426   2.798  0.00529 **
depth2      -0.0017283  0.0006599  -2.619  0.00903 **
rancho2     -0.4303603  0.2725263  -1.579  0.11479
rancho3     -0.3762514  0.3586106  -1.049  0.29448
rancho4     -0.8806389  0.3227926  -2.728  0.00654 **
fish         0.4980347  0.1336034   3.728  0.00021 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.807 on 651 degrees of freedom
Multiple R-squared: 0.04372,    Adjusted R-squared: 0.03491
F-statistic: 4.961 on 6 and 651 DF,  p-value: 5.592e-05
```

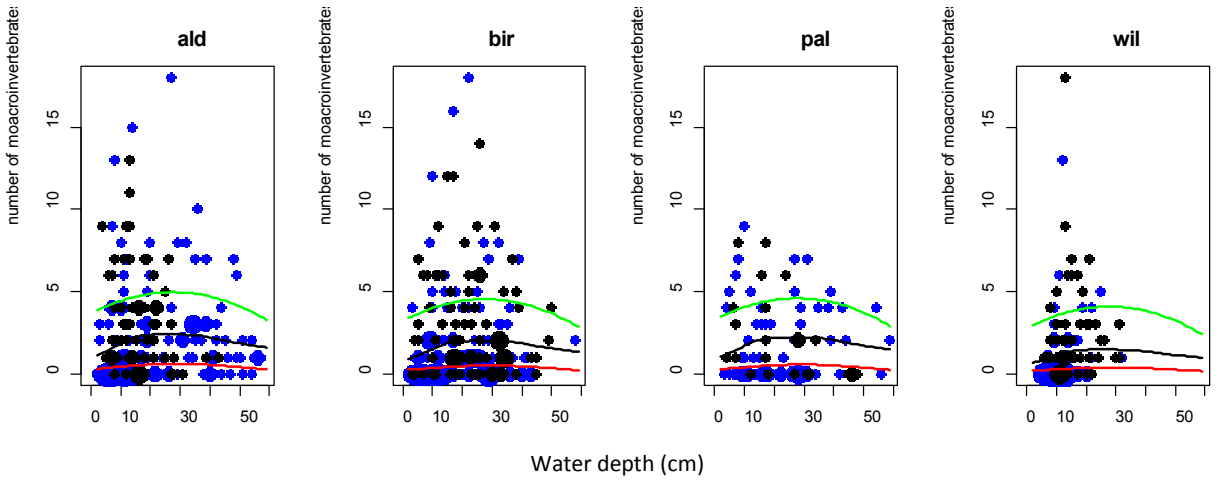


Figure 5. Macro-invertebrate abundance with water depth in four ranches in South Florida. The size of the dots relates to the amount of samples with a given combination of macro-invertebrates and depth. Samples with no fish are in blue, those with fish in black. The models were fit for fish abundance = 5. Linear model with Gaussian errors in green, generalized linear model with negative binomial errors in red and general additive model in black.

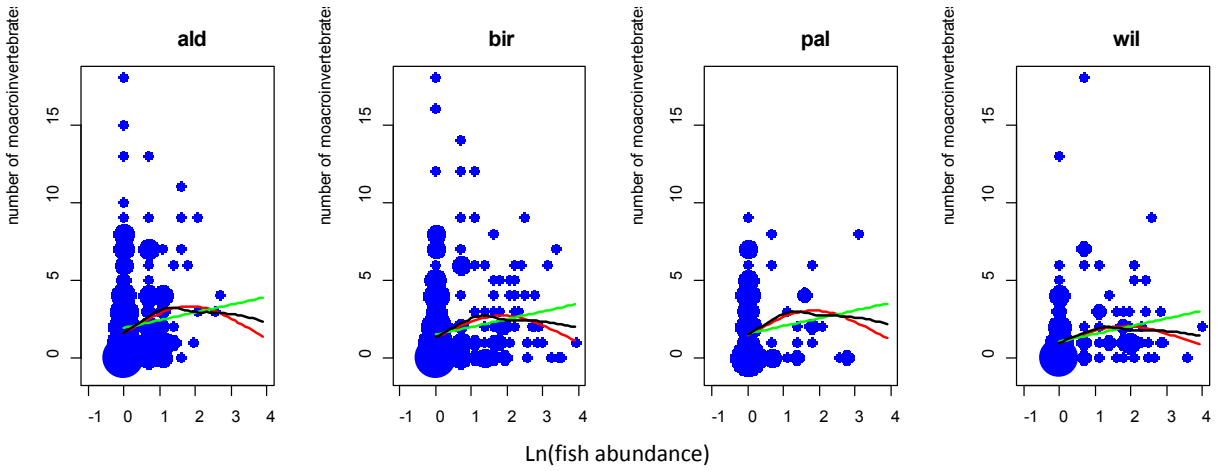


Figure 6. Macro-invertebrate abundance with fish abundance in four ranches in South Florida. The size of the dots relates to the amount of samples with a given combination of macro-invertebrates and fish. The models were fit for depth = 10. Linear model with Gaussian errors in green, generalized linear model with negative binomial errors in red and general additive model in black.

The shape of our linear model with Gaussian errors, represented with a green line in Figures 5 is consistent with the hypothesis of Bohlen et al. (2014). The output of the model allows the rejection of the null hypothesis of no evidence of an association with depth and that of no covariance with fish. It also indicates some differences among ranches. However, it explains less than 3 % of the variance, and its residuals have a strong pattern suggesting increasing positive deviations from predicted values as abundance of macro-invertebrates increases (Figure 7). This model also indicates increasing number of macro-invertebrates with fish that appear not to be consistent with the data. We should not trust this model.

Now we recognize that nature of the data as counts and the nonlinear relationship with fish. We evaluate a generalized linear model with negative binomial errors and allow for non-linearity in the covariance with fish adding the variable fish$^2$. We call the function *glmmadmb* from the package glmmADMB. Notice that we do not include zero Inflation even if the nature of the data may require this procedure.

```
glmmadmb(formula = y ~ depth + depth2 + fish + fish2 + rancho, data = sub1, family =
"nbinom", zeroInflation = FALSE)

AIC: 2446.1

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.028111   0.204410    0.14   0.8906
depth        0.055513   0.018151    3.06   0.0022 **
depth2      -0.000949   0.000348   -2.73   0.0064 **
fish         0.770460   0.187250    4.11  3.9e-05 ***
fish2       -0.207537   0.071678   -2.90   0.0038 **
rancho2     -0.196297   0.139040   -1.41   0.1580
rancho3     -0.085909   0.182590   -0.47   0.6380
rancho4     -0.479517   0.168690   -2.84   0.0045 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Number of observations: total=658
Negative binomial dispersion parameter: 0.66195 (std. err.: 0.059911)

Log-likelihood: -1214.04
```

The generalized linear model with negative binomial errors and quadratic response to fish abundance is more consistent with the data (Figures 6 and 7) and has weaker patterns in the residuals (Figure 7; less marked departures of the predicted values with increasing abundance of macro-invertebrates and fish). However, this departure from the assumptions still does not give strong support to this model.

Finally, we asses a general additive model allowing for more flexible association patterns for the response to depth and fish abundance. This approach allows evaluating nonlinear patterns using a piece by piece fitting across the range of the data. We call the function *gam* from the package mgcv.

```
Family: Negative Binomial(0.669)
Link function: log

Formula:
y ~ s(depth, fx = FALSE, k = -1, bs = "cr") + s(fish, fx = FALSE,
    k = -1, bs = "cr") + rancho

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.78273    0.09806   7.983 1.43e-15 ***
rancho2     -0.18269    0.13997  -1.305  0.19182
rancho3     -0.07754    0.18172  -0.427  0.66961
rancho4     -0.48909    0.17044  -2.870  0.00411 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
```

```
            edf Ref.df Chi.sq  p-value
s(depth) 3.161  3.954  11.52   0.0207 *
s(fish)  2.817  3.361  25.31 2.68e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.0361   Deviance explained = 6.64%
UBRE score = 0.047136  Scale est. = 1         n = 658
```

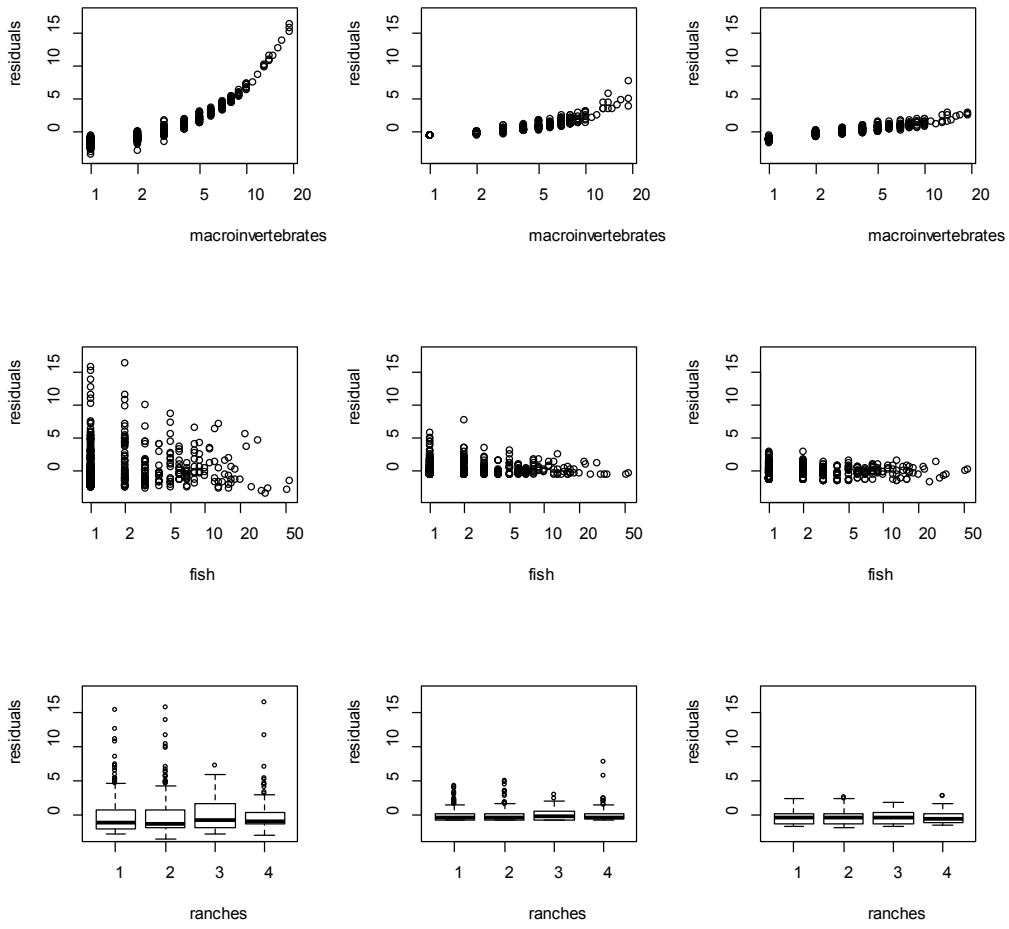


Figure 7. Residuals for the linear model with Gaussian errors (left), Generalized linear model with negative binomial errors (center) and General additive model (right).

The general additive model with negative binomial errors was consistent with the data and has almost removed any pattern in the residuals (Figure 7). This model has the smallest AIC value (lm (8 df) AIC= 3234.416; GLM (9 df) AIC= 2446.080; GAM (edf=9.978163) AIC= 2443.411) and the best association between observed and predicted values (Figure 8). We concur with Bohlen et al. (2014) supporting a non-linear association among water depth and micro-invertebrate abundance which, in the studied wetlands, reached maximum abundances around 20 cm. We also found that fish and macro-invertebrates co-varied and identified significant variation among ranches (Figure 9). The causes of ecological patterns are diverse and rarely follow linear

relationships. We can benefit from assessing models with more realistic and flexible assumptions allowing us to better describe these patterns.

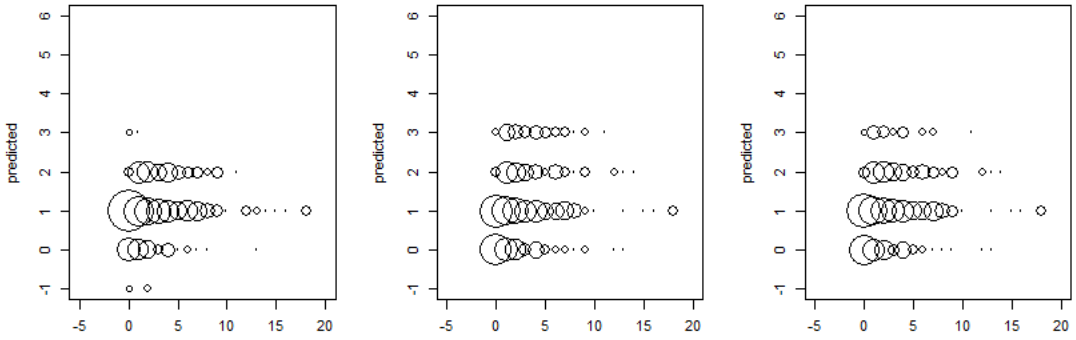


Figure 8. Association among observed and predicted values for the linear model with Gaussian errors (left), Generalized linear model with negative binomial errors (center)and General additive model(right).

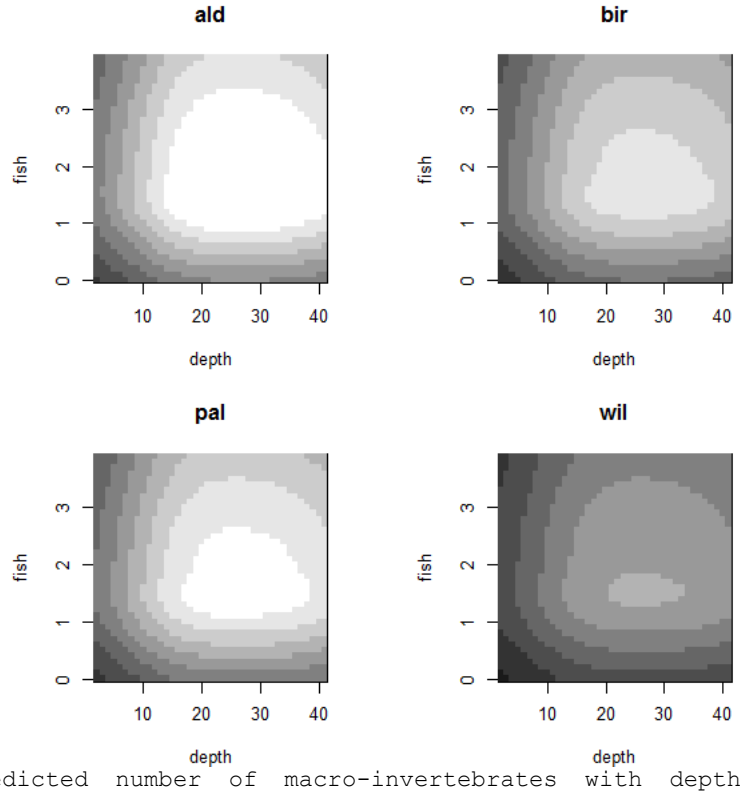


Figure 9. Predicted number of macro-invertebrates with depth and fish (natural logarithmic transformed) after the GAM model with negative binomial errors, range: white= 4 organisms, black 0 organisms.

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

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

## References

Boughton, E. H., P. F. Quintana-Ascencio, D. G. Jenkins, P. J. Bohlen, J. E. Fauth, A. Engel, G. Hendricks, G. Kiker, S. Shukla, & H. M. Swain. 2019. Tradeoffs and synergies in a payment-for-ecosystem services program on ranchlands in the Everglades headwaters. Ecosphere 10(5): e02728. 10.1002/ecs2.2728.

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.