Non-linear patterns?
Abundance with elevation and time-since-fire

Estimation of changes in abundance as a function of environmental variables is central information to develop ecological theory and design management for conservation. In this demonstration you will be introduced to General Additive Models in R which will allow you to calculate these estimates for intricate non-linear responses. Lack of linear response does not always mean that variables do not have an effect on the response variable. There are many biological phenomena that have predictable patterns that are not linear. We use data on *Hypericum cumulicola* abundance (Quintana-Ascencio *et al.* 2003). In this case, we want to characterize the variation in abundance of *Hypericum cumulicola* as a function of time-since-fire and site elevation. These two variables are surrogates of habitat suitability since they change resource distribution in time and space. The abundance of aboveground plants of this species was measured in replicated quadrats within sites at different elevation and across different time-since-fire.

Figure 1. *Hypericum cumulicola* plants

Getting ready. For this demo, you will need to download two files from the course website:
1) The *Hypericum cumulicola* individual abundance data (`density 2013.txt`).
2) The main R script for the analyses (`GAM 2017.R`).

Part I. Preparing the data

Open the R script. The first line in the program `rm(list=ls())`allows you to clear R's memory. The following lines call different libraries that we will use. The function `read.table("file`

`name.txt", header=T)` obtains the data from the *txt* files saved in your directory.

Since 1994, Quintana-Ascencio et al. (2003) have collected demographic data of *Hypericum cumulicola* in 14 populations at Archbold Biological Station, Florida USA. The line:

```
hc_density <- reshape(datas,varying=list(names(datas)[7:26],names(datas)[27:46]),
              direction="long", v.names=c("yearlings","plants"),times=2013:1994)
```

changes the arrangement of the data to accommodate them to the format required for analysis and allocates the original data to the new data frame "hc_density". We use the code in the box to identify the burn year by site and calculate time-since-fire

```
hc_density$burn.year <- NA
hc_density$burn.year[hc_density$site==1] <- 1966
hc_density$burn.year[hc_density$site==1 & hc_density$time>1998 &
hc_density$gap>5] <- 1999
hc_density$burn.year[hc_density$site==29 | hc_density$site==32] <- 1985
hc_density$burn.year[(hc_density$site==29 | hc_density$site==32) &
hc_density$time>1996] <- 1997
hc_density$burn.year[hc_density$site==42 | hc_density$site==50 |
hc_density$site==57] <- 1993
hc_density$burn.year[(hc_density$site==42 | hc_density$site==50 |
hc_density$site==57) & hc_density$time>2009] <- 2010
hc_density$burn.year[hc_density$site==59] <- 1968
hc_density$burn.year[hc_density$site==59 & hc_density$time>2009] <- 2010
hc_density$burn.year[hc_density$site==62] <- 1967
hc_density$burn.year[hc_density$site==62 & hc_density$time>2003] <- 2004
hc_density$burn.year[hc_density$site==62 & hc_density$time>2009] <- 2010
hc_density$burn.year[hc_density$site==67 | hc_density$site==87] <- 1986
hc_density$burn.year[(hc_density$site==67 | hc_density$site==87) &
hc_density$time>2007] <- 2008
hc_density$burn.year[hc_density$site==88 | hc_density$site==91] <- 1986
hc_density$burn.year[hc_density$site==93] <- 1972

hc_density$tsf<- hc_density$time-hc_density$burn.year
```

It is useful to check the distribution of the data. We use the function `hist()` to create a histogram of the number of aboveground plants and the function `abline()` to add in red the location of the mean. Your plot should look like Figure 2, where the data are highly skewed to the left.

```
par(mfrow=c(1,1))
hist(log(hc_density$plants+1),breaks=seq(0,5,0.1),main="plants per quadrat",
        xlab="log(plants+1)",col="gray")
abline(v =mean(log(hc_density$plants+1)), col="red",lwd=2)
```

Part II. Evaluating the models

We use the R program ***mgcv*** to obtain the general additive models. This program uses cubic regression splines as smoothing functions (Zuur et al. 2009). There are many smoothing

techniques. You can check many important details of these techniques in Hastie and Tibshirani (1990).

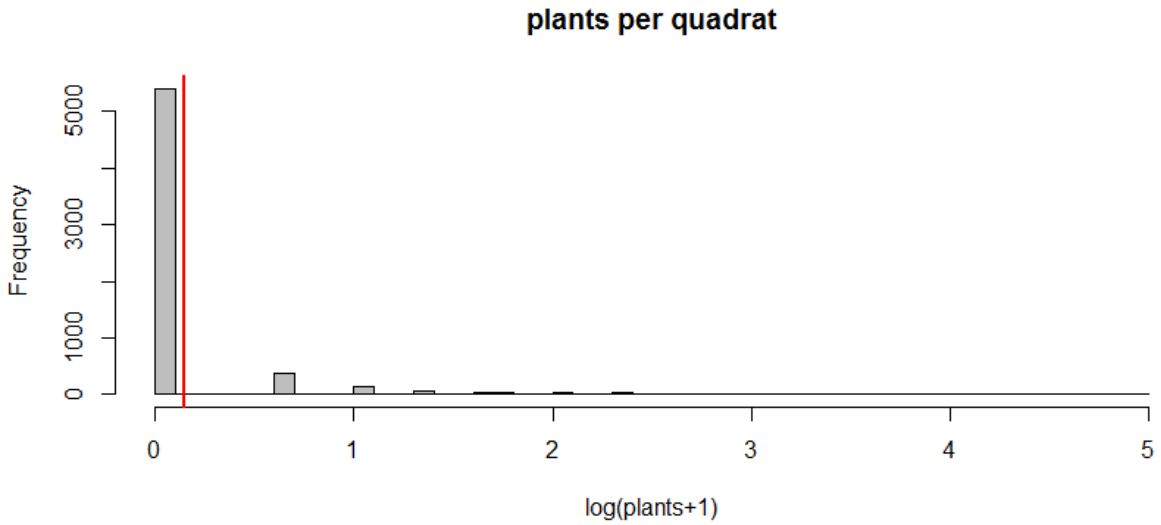

Figure 2. Histogram of adult plants per quadrat (logarithmic transformed data). Notice the highly aggregated pattern with most plots not having any plants.

The general gam function is (the error distribution can change accordingly with the data)

$$Y_i = \alpha + f(X_i) + \varepsilon_i \epsilon\, N(0, \sigma^2)$$

The format of this ***gam*** function is similar to that of linear regression. Note that the only difference with prior linear models is the replacement of $\boldsymbol{\beta\, x\, X_i}$ by the smoothing curve $f(X_i)$. **P**rior models gave us a formula for the relationship between Y and X. In a GAM, instead of this equation, we have a smoother (Zuur et al. 2009). The additive models fit a smoothing curve through the data. We can still use this model for prediction but not with a simple equation. The expression `Y ~ s(X)` in the function ***gam*** (see below) means that a smoothing function is used for the explanatory variable *X*. The `fx=FALSE, k=-1` bit means that the amount of smoothing is not fixed to a default value; hence, cross validation is used to estimate the optimal amount of smoothing (Zuur et al. 2009). The `bs="cr"` code tells R that a cubic regression spline is used (Zuur et al. 2009). We use a a negative binomial distribution since these are count data (`family=nb`). The basic idea is that the gradient is divided into a certain number of intervals and for each segment a cubic polynomial is fitted and then all the segments are glued together (Zuur et al. 2009). The main advantage of the function ***gam*** in the ***mgcv*** package is that it allows for cross-validation which automatically determine the optimal amount of smoothing (Zuur et al. 2009). Zuur et al. (2009) advise not to follow blindly the results of the cross-validation. It is wise to verify the cross-validation results with smoothers selected manually. Finding the optimal span of the smoother is a matter of bias-variance trade-off and another option is to use Akaike Information Criterion (Zuur et al. 2009).

We use this approach and evaluate three models for the data of the abundance of *H. cumullicola*.

```
M0 <- gam(plants ~ s(tsf, fx=FALSE, k=-1, bs="cr")+
s(RelativeElevation,fx=FALSE,k=-1,bs="cr"),data = hc_density,family=nb(c(1)))
M3030 <- gam(plants ~ s(tsf, fx=TRUE, k=30, bs="cr")+
s(RelativeElevation,fx=TRUE,k=30,bs="cr"),data = hc_density,family=nb(c(1)))
M3010 <- gam(plants ~ s(tsf, fx=TRUE, k=30, bs="cr")+
s(RelativeElevation,fx=TRUE,k=10,bs="cr"),data = hc_density,family=nb(c(1)))

AICctab(M0,M3030,M3010,weights=T)
```

-1,

```
> summary(M3030)

Family: Negative Binomial(1)
Link function: log

Formula:
plants ~ s(tsf, fx = TRUE, k = 30, bs = "cr") + s(RelativeElevation,
    fx = TRUE, k = 30, bs = "cr")

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.55017    0.03876  -39.99   <2e-16 ***
---
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(tsf)              29     29  431.2  <2e-16 ***
s(RelativeElevation)  29     29  714.8  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.0654   Deviance explained =   24%
-REML = 4275.8  Scale est. = 1         n = 6180
```

The three models were consistent with observed data (Figure 3) and each other (Figures 4-6). Model 30 30, with the lowest AIC, has an R^2 of 0.06 explaining a limited amount of the variation. All models identified higher number of individuals shortly after fire and in long-unburned conditions. The lowest numbers were associated with intermediate time-since-fire. Additionally, there was more uncertainty for the intermediate and the long-unburned conditions. Residuals are more evenly distributed in the model identified with AIC (Figure 7).

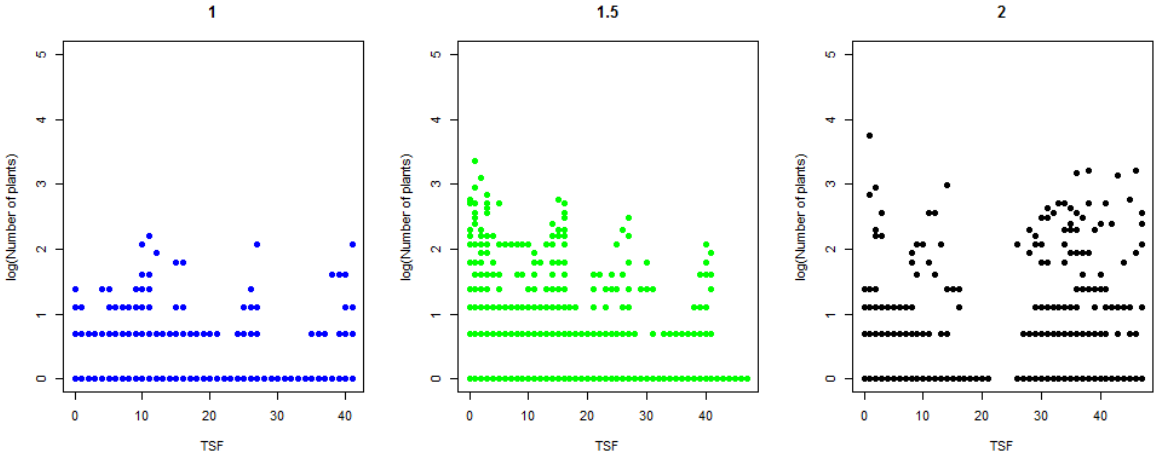

Figure 3. Plot of *H. cumulicola* abundance with time-since-fire and elevation: 1 (blue), 1.5 (green) and 2 (black) meters above the wetlands. The observed abundances were logarithmic transformed after adding one.

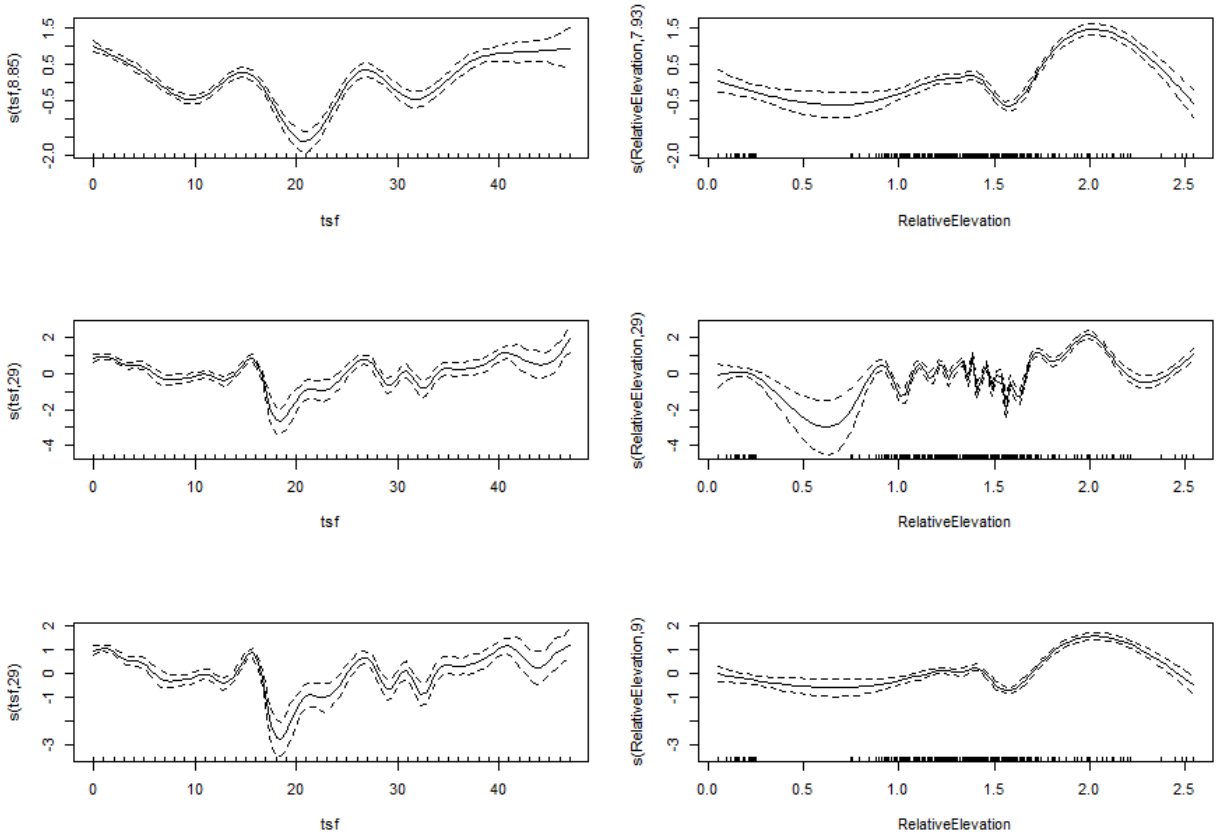

Figure 4. Plots of three models evaluated.

The cross-validation function identified the following model M -1 -1

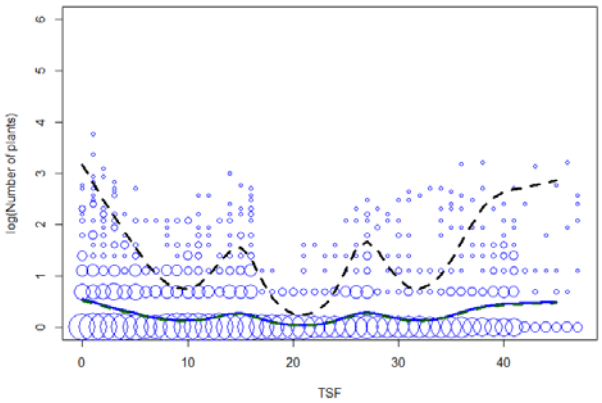

Figure 5. Plot of *H. cumulicola* abundance and predicted smoothing model as a function of time-since-fire and elevation: 1 (blue), 1.5 (green) and 2 (black) meters above the wetlands. This is the model with k= -1 for both functions. The blue circles are the observed abundances (logarithmic transformed after adding one). The size of the circle increases with sample size.

Using AIC we identify M 30 30 as the most plausible model. It has the best residuals

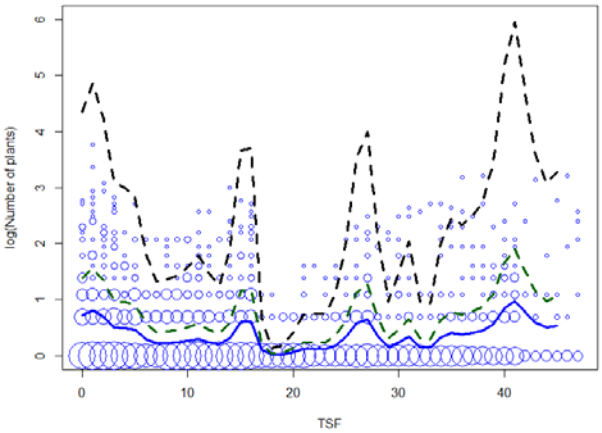

Figure 6. Plot of *H. cumulicola* abundance and predicted smoothing model as a function of time-since-fire and elevation: 1 (blue), 1.5 (green) and 2 (black) meters above the wetlands. This is a model with k= 30 for both functions. In blue circles there are the observed abundances (logarithmic transformed after adding one). The size of the circle increases with sample size.

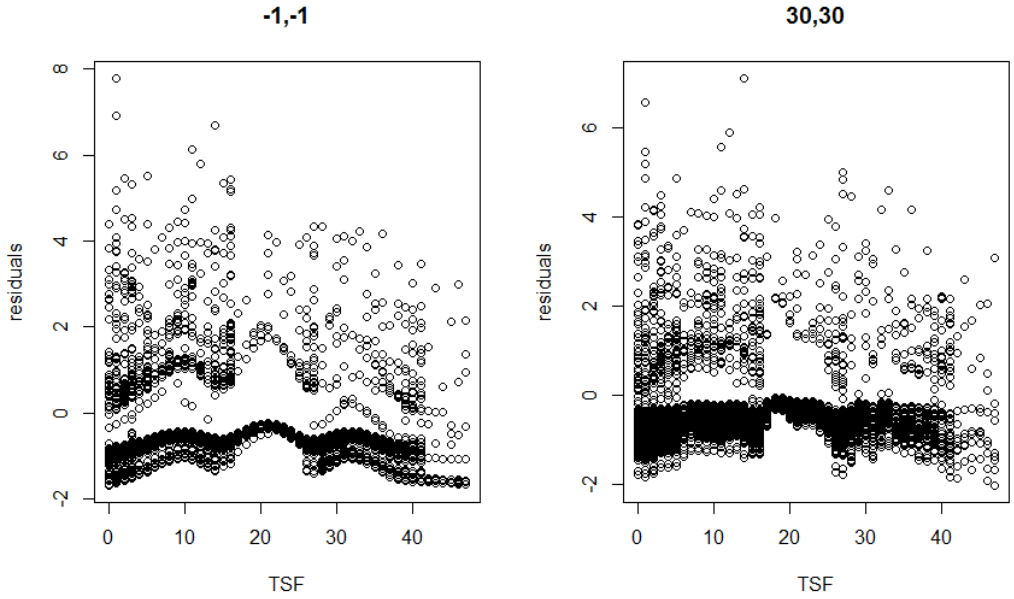

Figure 7. Residuals of the two chosen models.

References.

Quintana-Ascencio, P. F., E. S. Menges, and C. Weekley. 2003. A fire-explicit population viability analysis of *Hypericum cumulicola* in Florida rosemary scrub. Conservation Biology 17: 433-449.

Hastie, T., R. Tibshirani. 1990. Generalized Additive Models. Chapman and Hall. London.

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.