Reanalysis of GAMS with Significant Chlorophyll Predictors

Curtis C. Bohlen, Casco Bay Estuary Partnership

6/27/2023

Contents

Introduction

This notebook provides analyses using GAMs, of a couple of models that included a single, high chlorophyll value. The intent is to run reach analysis omitting that high Chlorophyll point and confirm that the interpretation of the data does not change (much).

Load Libraries

```
library(tidyverse)
#> -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
#> v dplyr     1.1.1     v readr     2.1.4
#> v forcats   1.0.0     v stringr   1.5.0
#> v ggplot2   3.4.1     v tibble    3.2.1
#> v lubridate 1.9.2     v tidyr     1.3.0
#> v purrr     1.0.1
#> -- Conflicts ------------------------------------------ tidyverse_conflicts() --
#> x dplyr::filter() masks stats::filter()
#> x dplyr::lag()    masks stats::lag()
#> i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
library(readxl)
library(mgcv)      # for GAM models
#> Loading required package: nlme
#>
#> Attaching package: 'nlme'
#>
#> The following object is masked from 'package:dplyr':
#>
#>     collapse
#>
#> This is mgcv 1.8-42. For overview type 'help("mgcv-package")'.
library(emmeans)   # For extracting useful "marginal" model summaries

theme_set(theme_classic())
```

Input Data

Folder References

```
data_folder <- "Original_Data"
```

Load Data

Data preparation follows the same steps as before. I simplify discussion a bit just to save typing / save space.

```
filename.in <- "penob.station.data EA 3.12.20.xlsx"
file_path <- file.path(data_folder, filename.in)
station_data <- read_excel(file_path,
                           sheet="Final", col_types = c("skip", "date",
                                                        "numeric", "text", "numeric",
                                                        "text", "skip", "skip",
                                                        "skip",
                                                        rep("numeric", 10),
                                                        "text",
                                                        rep("numeric", 47),
                                                        "text",
                                                        rep("numeric", 12))) %>%
```

```
  rename_with(~ gsub(" ", "_", .x)) %>%
  rename_with(~ gsub("\\.", "_", .x)) %>%
  rename_with(~ gsub("\\?", "", .x)) %>%
  rename_with(~ gsub("%", "pct", .x)) %>%
  rename_with(~ gsub("_Abundance", "", .x)) %>%
  filter(! is.na(date))
#> New names:
#> * `` -> `...61`

names(station_data)[10:12]
#> [1] "discharge_week_cftpersec" "discharg_day"
#> [3] "discharge_week_max"
names(station_data)[10:12] <- c('disch_wk', 'disch_day', 'disch_max')

station_data <- station_data %>%
  mutate(station = factor(as.numeric(factor(station))))
head(station_data)
#> # A tibble: 6 x 76
#>   date                year month month_num season riv_km station station_num
#>   <dttm>              <dbl> <chr>     <dbl> <chr>   <dbl> <fct>         <dbl>
#> 1 2013-05-28 00:00:00  2013 May           5 Spring  22.6  1                 1
#> 2 2013-05-28 00:00:00  2013 May           5 Spring  13.9  2                 2
#> 3 2013-05-28 00:00:00  2013 May           5 Spring   8.12 3                 3
#> 4 2013-05-28 00:00:00  2013 May           5 Spring   2.78 4                 4
#> 5 2013-07-25 00:00:00  2013 July          7 Summer  22.6  1                 1
#> 6 2013-07-25 00:00:00  2013 July          7 Summer  13.9  2                 2
#> # i 68 more variables: depth <dbl>, disch_wk <dbl>, disch_day <dbl>,
#> #   disch_max <dbl>, tide_height <dbl>, Full_Moon <dbl>, Abs_Moon <dbl>,
#> #   Spring_or_Neap <chr>, ave_temp_c <dbl>, ave_sal_psu <dbl>,
#> #   ave_turb_ntu <dbl>, ave_do_mgperl <dbl>, ave_DO_Saturation <dbl>,
#> #   ave_chl_microgperl <dbl>, sur_temp <dbl>, sur_sal <dbl>, sur_turb <dbl>,
#> #   sur_do <dbl>, sur_chl <dbl>, bot_temp <dbl>, bot_sal <dbl>, bot_turb <dbl>,
#> #   bot_do <dbl>, bot_chl <dbl>, max_temp <dbl>, max_sal <dbl>, ...
```

Subsetting to Desired Data Columns

```
base_data <- station_data %>%
  rename(Date = date,
         Station = station,
         Year = year) %>%
  select(-c(month, month_num)) %>%
  mutate(Month = factor(as.numeric(format(Date, format = '%m')),
                                                levels = 1:12,
                                                labels = month.abb),
         DOY = as.numeric(format(Date,format = '%j')),
         season = factor(season, levels = c('Spring', 'Summer', 'Fall')),
         is_sp_up = season == 'Spring' & Station == 1,
         Yearf = factor(Year)) %>%
  rename(Season = season,
         Density = combined_density,
         Temp = ave_temp_c,
         Sal = ave_sal_psu,
```

```
          Turb = sur_turb,
          AvgTurb = ave_turb_ntu,
          DOsat = ave_DO_Saturation,
          Chl = ave_chl_microgperl,
          Fish = `___61`,
          RH = Herring
          ) %>%
  select(Date, Station, Year, Yearf, Month, Season, is_sp_up, DOY, riv_km,
          disch_wk, disch_day, disch_max,
          Temp, Sal, Turb, AvgTurb, DOsat, Chl,
          Fish, RH,
          Density, H, SEI,
          Acartia, Balanus, Eurytemora, Polychaete, Pseudocal, Temora) %>%
  arrange(Date, Station)

rm(station_data)
```

Complete Cases

This drops two samples, one for missing Zooplankton data, one for missing fish data. We needed a "reduced"complete cases" data set to run The `step()` function in earlier analysis steps. It makes little sense to try stepwise model selection if each time you add or remove a variable, the sample you are studying changes.

```
complete_data <- base_data %>%
  select(Season, Station, Yearf,
          is_sp_up, Temp, Sal, Turb, Chl, Fish, RH,
          Density, H,
          Acartia, Balanus, Eurytemora, Polychaete, Pseudocal, Temora) %>%
  filter(complete.cases(.))
```

Reduced Data

The low salinity spring samples are doing something rather different, and they complicate model fitting. Models are better behaved if we exclude a few extreme samples.

```
drop_low <- complete_data %>%
  filter(Sal > 10)    # Pulls three samples, including one fall upstream sample
                      # a fourth low salinity sample lacks zooplankton data
```

And finally, we generate a data set that omits the high Chlorophyll sample. We will compare (informally) the results of running the same models on these last two data sets.

Here's the high Chlorophyll sample (just to see when and where it occurred).

```
drop_low %>%
  filter(Chl >= 15)
#> # A tibble: 1 x 18
#>   Season Station Yearf is_sp_up  Temp   Sal  Turb   Chl  Fish    RH Density
#>   <fct>  <fct>   <fct> <lgl>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl>
#> 1 Summer 4       2017  FALSE     12.8  26.3  4.28  17.8  87.3     0   4679.
#> # i 7 more variables: H <dbl>, Acartia <dbl>, Balanus <dbl>, Eurytemora <dbl>,
#> #   Polychaete <dbl>, Pseudocal <dbl>, Temora <dbl>
```

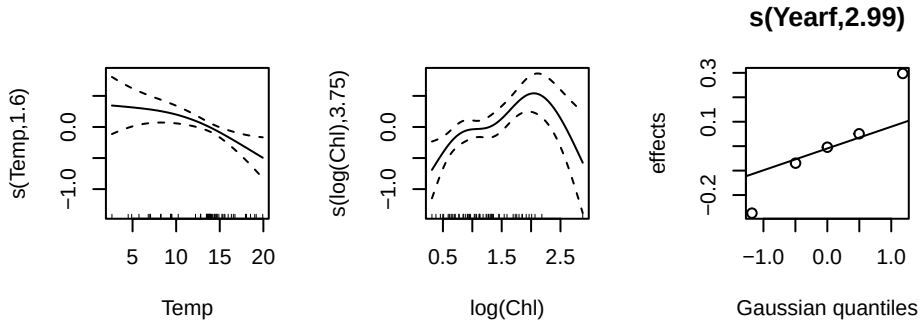
```r
drop_chl <- drop_low %>%
  filter(Chl <15)
```

Shannon Diversity

With High Chlorophyll Sample

```r
shannon_gam_no_low <- gam(H ~
                      s(Temp, bs="ts", k = 5) +
                      s(log(Chl), bs="ts", k = 5) +
                      s(Yearf, bs = 're'),
                    data = drop_low, family = 'gaussian')
summary(shannon_gam_no_low)
#>
#> Family: gaussian
#> Link function: identity
#>
#> Formula:
#> H ~ s(Temp, bs = "ts", k = 5) + s(log(Chl), bs = "ts", k = 5) +
#>     s(Yearf, bs = "re")
#>
#> Parametric coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)   1.3313     0.1183   11.25 9.35e-15 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Approximate significance of smooth terms:
#>               edf Ref.df      F  p-value
#> s(Temp)     1.596      4  4.136 0.003328 **
#> s(log(Chl)) 3.752      4 11.026 0.000294 ***
#> s(Yearf)    2.994      4  2.938 0.007168 **
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> R-sq.(adj) =   0.41   Deviance explained = 50.1%
#> GCV = 0.20093  Scale est. = 0.1668    n = 55
```

```r
oldpar <- par(mfrow = c(2,3))
plot(shannon_gam_no_low)
par(oldpar)
```



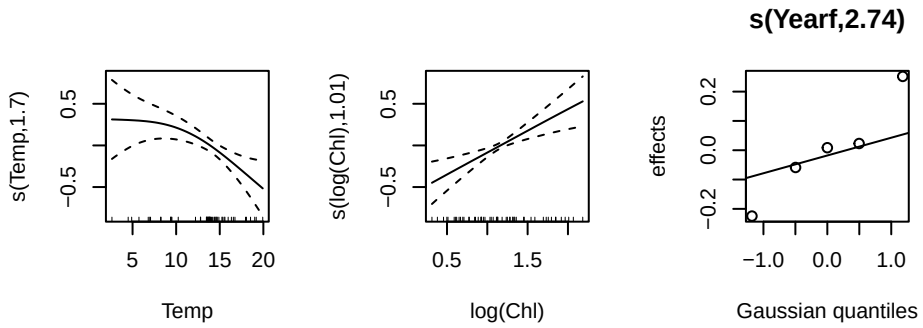
Without High Chlorophyll Sample

```
shannon_gam_no_low_2 <- gam(H ~
                      s(Temp, bs="ts", k = 5) +
                      s(log(Chl), bs="ts", k = 5) +
                      s(Yearf, bs = 're'),
                    data = drop_chl, family = 'gaussian')
summary(shannon_gam_no_low_2)
#>
#> Family: gaussian
#> Link function: identity
#>
#> Formula:
#> H ~ s(Temp, bs = "ts", k = 5) + s(log(Chl), bs = "ts", k = 5) +
#>     s(Yearf, bs = "re")
#>
#> Parametric coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)   1.3394     0.1071    12.5   <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Approximate significance of smooth terms:
#>                edf Ref.df     F  p-value
#> s(Temp)      1.701      4 3.969 0.002320 **
#> s(log(Chl)) 1.010      4 6.833 0.000311 ***
#> s(Yearf)     2.738      4 2.254 0.016662 *
```

```
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> R-sq.(adj) =  0.373   Deviance explained = 43.7%
#> GCV = 0.20224  Scale est. = 0.17809    n = 54
```

We are left with an effectively linear relationship between chlorophyll and diversity. As we suspected by looking at the previous graphics, all the curvature comes about by trying to fit that one high chlorophyll sample.

```
oldpar <- par(mfrow = c(2,3))
plot(shannon_gam_no_low_2)
par(oldpar)
```



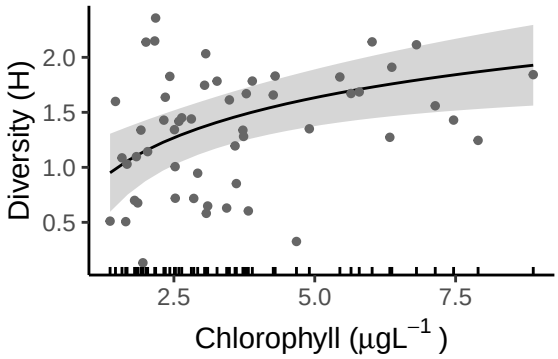
Related Graphic

Here's a graphic that you can compare to the graphics I prepared earlier in the week.

```
r = range(drop_chl$Chl)
stops = log(seq(r[1], r[2], length.out = 25))
chl_emms <- emmeans(shannon_gam_no_low_2, "log(Chl)",
                    at = list('log(Chl)' = stops),
                    type = 'response')

chl_emms_2 <- as_tibble(chl_emms) %>%
  mutate(Chl = exp(`log(Chl)`)) %>%
  relocate(Chl)
```

```
ggplot(chl_emms_2, aes(Chl, emmean)) +
  geom_ribbon(aes(ymin = lower.CL, ymax = upper.CL), alpha = 0.20) +
  geom_line() +
  #geom_point() +
  geom_point(data = drop_chl, mapping = aes(x = Chl, y = H),
             size = 1, color = "gray40") +
  geom_rug(data = drop_chl, mapping = aes(x = Chl, y = NULL)) +
  xlab(expression("Chlorophyll (" * mu * g * L ^-1 ~")")) +
  ylab("Diversity (H)")
```



Balanus

We are only interested in Balanus, so rather than repeat the automated species by species analysis I used before, I've just run the analysis on Balanus directly.

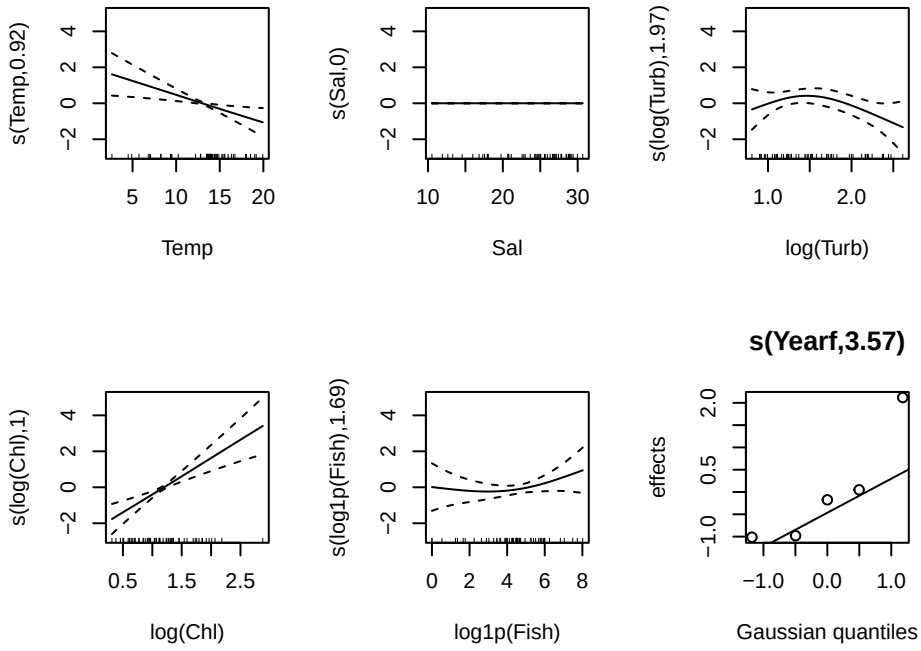
With High Chlorophyll Sample

```
spp_data <- drop_low %>%
  select(Yearf, Season, Station, Temp,
         Sal, Turb, Chl, Fish, Balanus)
```

```
balanus_gam <- gam(log1p(Balanus) ~
        s(Temp, bs="ts", k = 5) +
        s(Sal, bs="ts", k = 5) +
        s(log(Turb), bs="ts", k = 5) +
        s(log(Chl), bs="ts", k = 5) +
        s(log1p(Fish),bs="ts", k = 5) +
        s(Yearf, bs = 're'),
      data = spp_data, family = "gaussian")
summary(balanus_gam)
#>
#> Family: gaussian
#> Link function: identity
#>
#> Formula:
#> log1p(Balanus) ~ s(Temp, bs = "ts", k = 5) + s(Sal, bs = "ts",
```

```
#>     k = 5) + s(log(Turb), bs = "ts", k = 5) + s(log(Chl), bs = "ts",
#>     k = 5) + s(log1p(Fish), bs = "ts", k = 5) + s(Yearf, bs = "re")
#>
#> Parametric coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)   3.6930     0.6478   5.701 8.74e-07 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Approximate significance of smooth terms:
#>                       edf Ref.df      F  p-value
#> s(Temp)         9.192e-01      4  2.998  0.00414 **
#> s(Sal)          1.782e-10      4  0.000  0.52552
#> s(log(Turb))    1.967e+00      4  1.779  0.06016 .
#> s(log(Chl))     1.004e+00      4 14.125 2.07e-05 ***
#> s(log1p(Fish)) 1.686e+00      4  0.691  0.22444
#> s(Yearf)        3.568e+00      4  7.912 1.75e-05 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> R-sq.(adj) =  0.581   Deviance explained = 65.2%
#> GCV = 2.7021  Scale est. = 2.2038    n = 55
```

```
oldpar <- par(mfrow = c(2,3))
plot(balanus_gam)
```



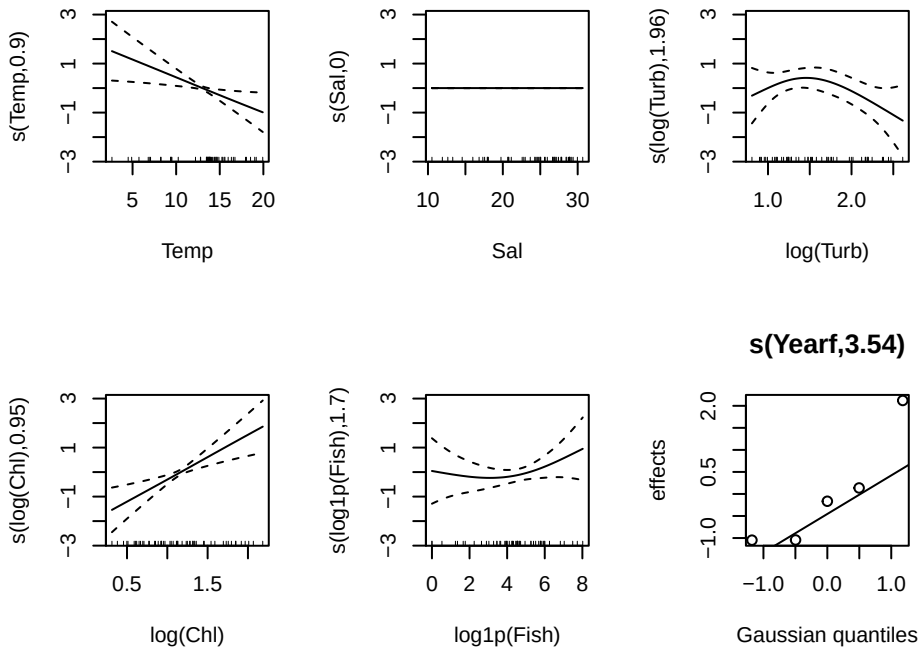
```
par(oldpar)
```

Without High Chlorophyll Sample

```
spp_data_2 <- drop_chl %>%
  select(Yearf, Season, Station, Temp,
         Sal, Turb, Chl, Fish, Balanus)
```

```
balanus_gam_2 <- gam(log1p(Balanus) ~
        s(Temp, bs="ts", k = 5) +
        s(Sal, bs="ts", k = 5) +
        s(log(Turb), bs="ts", k = 5) +
        s(log(Chl), bs="ts", k = 5) +
        s(log1p(Fish),bs="ts", k = 5) +
        s(Yearf, bs = 're'),
      data = spp_data_2, family = "gaussian")
summary(balanus_gam_2)
#>
#> Family: gaussian
#> Link function: identity
#>
#> Formula:
#> log1p(Balanus) ~ s(Temp, bs = "ts", k = 5) + s(Sal, bs = "ts",
#>     k = 5) + s(log(Turb), bs = "ts", k = 5) + s(log(Chl), bs = "ts",
#>     k = 5) + s(log1p(Fish), bs = "ts", k = 5) + s(Yearf, bs = "re")
#>
#> Parametric coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)   3.6071     0.6382   5.652  1.1e-06 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Approximate significance of smooth terms:
#>                       edf Ref.df     F  p-value
#> s(Temp)         9.023e-01      4 2.600 0.007533 **
#> s(Sal)          1.009e-10      4 0.000 0.632842
#> s(log(Turb))    1.957e+00      4 1.616 0.073666 .
#> s(log(Chl))     9.514e-01      4 9.978 0.000411 ***
#> s(log1p(Fish)) 1.695e+00      4 0.703 0.220716
#> s(Yearf)        3.536e+00      4 7.898 1.73e-05 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> R-sq.(adj) =  0.561   Deviance explained = 63.6%
#> GCV = 2.7469  Scale est. = 2.2361    n = 54
```

```
oldpar <- par(mfrow = c(2,3))
plot(balanus_gam_2)
```



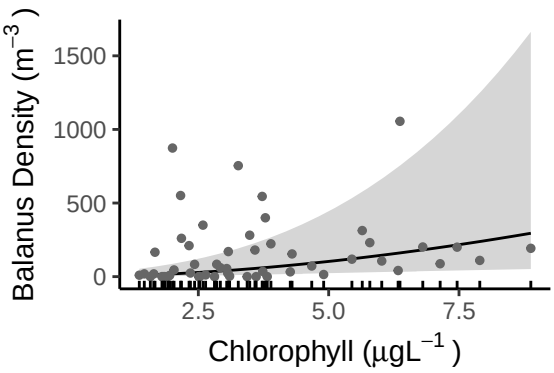

```
par(oldpar)
```

Related Graphic

```
r = range(drop_chl$Chl)
stops = log(seq(r[1], r[2], length.out = 25))
chl_emms <- emmeans(balanus_gam_2, "log(Chl)",
                    at = list('log(Chl)' = stops),
                    type = 'response')

chl_emms_2 <- as_tibble(chl_emms) %>%
  mutate(Chl = exp(`log(Chl)`)) %>%
  relocate(Chl)
#chl_emms_2
```

```
ggplot(chl_emms_2, aes(Chl, response)) +
  geom_ribbon(aes(ymin = lower.CL, ymax = upper.CL), alpha = 0.20) +
  geom_line() +
  geom_point(data = drop_chl, mapping = aes(x = Chl, y = Balanus),
             size = 1, color = "gray40") +
  geom_rug(data = drop_chl, mapping = aes(x = Chl, y = NULL)) +
  xlab(expression("Chlorophyll (" * mu * g * L ^-1 ~")")) +
  ylab(expression("Balanus Density (" * m ^-3 ~ ")" ))
```



I'm actually surprised by that. The primary effect of the high Chloride number is to slightly increase the slope of the regression line and greatly expand the width of the error band. In other words, no real change in qualitative behavior.