

LMM_fisheries

Medy Mu

2022-10-13

```
library(dplyr)
library(ggplot2)
library(lme4)
library(broom.mixed)
library(knitr)
library(patchwork)
library(nlme)
library(mgcv)
library(car)
```

1. Input Data; Basic Data Processing

```
x<-read.csv("fullSetWithMask.csv",as.is=TRUE,
             na.strings=c("NA",".",","," "))
dim(x) ## Check the dimension (number of rows and columns) of the data matrix.

## [1] 193439      10

head(x)

##   X     uniqueid year month    cpue logcpue taxa_grouped_weight
## 1 1 00000-BCSCAM 2009     7 24.61905 3.203520          Lutjanidae
## 2 2 00000-BCSCAM 2013     8 19.33333 2.961831          Lutjanidae
## 3 3 00000-BCSCAM 2012     2 32.76190 3.489266          Lutjanidae
## 4 4 00000-BCSCAM 2010     3 10.28571 2.330756         Carangidae
## 5 5 00000-BCSCAM 2013    10  9.52381 2.253795         Carangidae
## 6 6 00000-BCSCAM 2008     8 33.33333 3.506558          Lutjanidae
##                                         agreacion_fed
## 1 Federacion de Sociedades Cooperativas Pesqueras Zona Centro de Baja California Sur
## 2 Federacion de Sociedades Cooperativas Pesqueras Zona Centro de Baja California Sur
## 3 Federacion de Sociedades Cooperativas Pesqueras Zona Centro de Baja California Sur
## 4 Federacion de Sociedades Cooperativas Pesqueras Zona Centro de Baja California Sur
## 5 Federacion de Sociedades Cooperativas Pesqueras Zona Centro de Baja California Sur
## 6 Federacion de Sociedades Cooperativas Pesqueras Zona Centro de Baja California Sur
##   rr coop_edad_10
## 1 3      1.4
## 2 3      1.4
## 3 3      1.4
## 4 3      1.4
```

```

## 5 3          1.4
## 6 3          1.4

## Drop Column of Row Numbers:
x<-x[,-1]
## Rename a Few Variables:
colnames(x)[colnames(x)=="taxa_grouped_weight"]<-"taxa"
colnames(x)[colnames(x)=="agremiacion_fed"]<-"fed"
colnames(x)[colnames(x)=="coop_edad_10"]<-"age"
colnames(x)[colnames(x)=="rr"]<-"region"
x$uniqueid<-as.factor(x$uniqueid)
x$fed<-as.factor(x$fed)
x$taxa<-as.factor(x$taxa)
x$region<-as.factor(x$region)
x$yearFactor<-as.factor(x$year)
x$monthFactor<-as.factor(x$month)

```

2. Basic Summaries

```
table(x$taxa,useNA="always")
```

```

##
##      Carangidae  Centropomidae    Cichlidae    Clupeidae Elasmobranchii
##        13299       11935       11541         291        12759
##      Gerreidae     Haliotidae    Lutjanidae   Mugilidae  Octopodidae
##        8294        3586       16117        5183        4555
## Ommastrephidae    Ostreidae     OTRAS       Palinuridae Pectinidae
##        472         2568       38234        6234        860
##      Penaeidae     Portunidae   Sciaenidae   Scombridae Serranidae
##        18103       6642       12300        3704        16002
## Stichopodidae <NA>
##        760          0

```

```
table(x$region,useNA="always")
```

```

##
##      1      2      3      4      5      6 <NA>
##    4450 10633 62428 50213 22450 43265     0

```

```
length(unique(x$uniqueid))
```

```
## [1] 182
```

```
length(unique(x$fed))
```

```
## [1] 54
```

```

table(x$year)

##
##   2008 2009 2010 2011 2012 2013 2014 2015 2016
## 19573 20485 20413 19346 21198 20111 20515 24172 27626

summary(x$age)

##      Min. 1st Qu. Median     Mean 3rd Qu.     Max.
##      0.100 1.800 2.900 3.706 5.900 8.500

summary(x$age[x$uniqueid=="00000-BCSCAM"])

##      Min. 1st Qu. Median     Mean 3rd Qu.     Max.
##      1.4     1.4     1.4     1.4     1.4     1.4

x$age2<-((10*x$age) + (x$year - 2016))
summary(x$age2)

##      Min. 1st Qu. Median     Mean 3rd Qu.     Max.
##      -6.00 15.00 27.00 33.29 53.00 85.00

table(x$age2<0)

##
##   FALSE    TRUE
## 192186   1253

tbl<-table(as.character(x$uniqueid)[x$age2<0]); tbl

##
## 00549-TAB 00685-BCS 00747-YUC 00748-YUC
##       109        226        212        706

table(x$age2[x$uniqueid %in% names(tbl)],
      as.character(x$uniqueid[x$uniqueid %in% names(tbl)]),
      useNA="always")

##                                     00549-TAB 00685-BCS 00747-YUC 00748-YUC <NA>
## -6                         0         0         3         0         0
## -5                        35         0        35         0         0
## -4                        32         0        68         0         0
## -3                        23         0        28         0         0
## -2                        17        130        35        420         0
## -1                         2         96        43        286         0
##  0                         0        102        31        180         0
##  1                         0        127        24        440         0

```

```

##   2       0     194      53     455      0
##   3     136     200      0     265      0
##   4       0     89      0     438      0
##   5       0    145      0     401      0
##   6       0     51      0     454      0
## <NA>     0       0      0      0      0

## Naive Age Fix (the years associated w/ these samples may be incorrect):
x$age[(x$uniqueid=="00549-TAB")&(x$age2<0)] <- (x$age[(x$uniqueid=="00549-TAB")&(x$age2<0)] + 0.5)
x$age[(x$uniqueid=="00685-BCS")&(x$age2<0)] <- (x$age[(x$uniqueid=="00685-BCS")&(x$age2<0)] + 0.2)
x$age[(x$uniqueid=="00747-YUC")&(x$age2<0)] <- (x$age[(x$uniqueid=="00747-YUC")&(x$age2<0)] + 0.6)
x$age[(x$uniqueid=="00748-YUC")&(x$age2<0)] <- (x$age[(x$uniqueid=="00748-YUC")&(x$age2<0)] + 0.2)
x$age2 <- ((10*x$age) + (x$year - 2016))
summary(x$age2)

```

```

##      Min. 1st Qu. Median   Mean 3rd Qu.   Max.
##      0.00 15.00 27.00 33.31 53.00 85.00

```

2.2 Taxa by coop

```

tbl <- table(x$uniqueid, x$taxa)
dim(tbl)

## [1] 182 21

summary(coopByTaxa <- as.numeric(tbl))

##      Min. 1st Qu. Median   Mean 3rd Qu.   Max.
##      0.00  0.00  0.00 50.61 12.00 6074.00

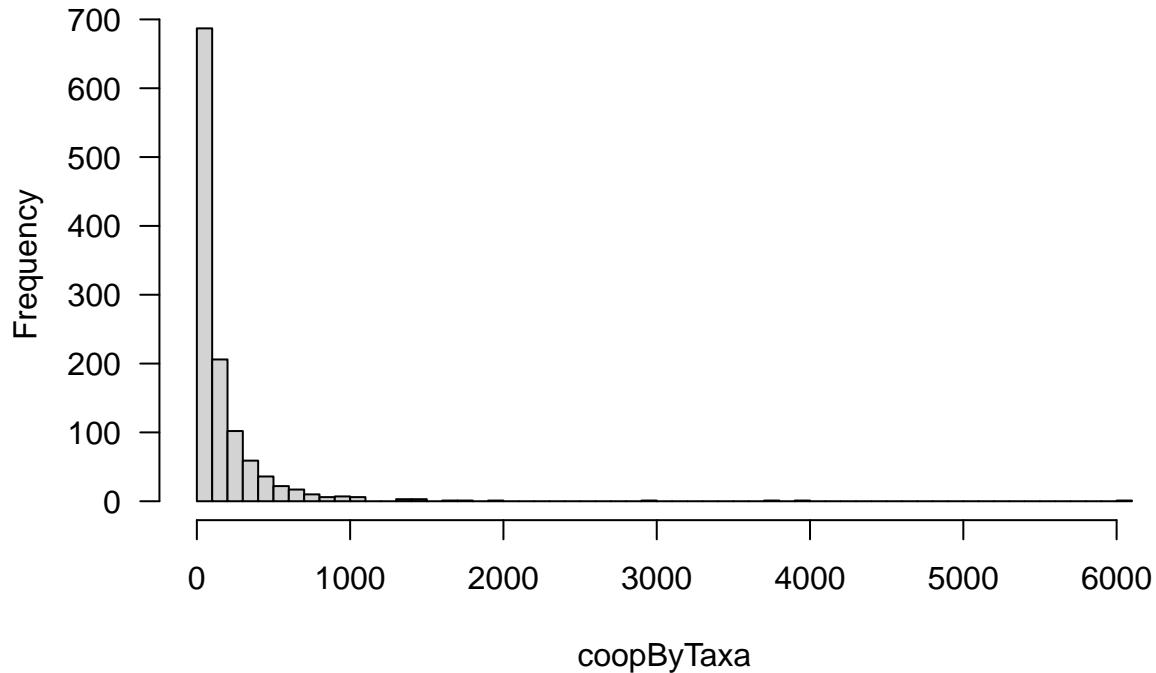
table(coopByTaxa == 0)

##
## FALSE  TRUE
## 1171 2651

coopByTaxa <- coopByTaxa[coopByTaxa > 0]
hist(coopByTaxa, nclass=50, las=1)

```

Histogram of coopByTaxa



3. Client's LME model

Note, from the `nlme::ACF` documentation: “This method function calculates the empirical autocorrelation function for the within-group residuals from an `lme` fit. The autocorrelation values are calculated using pairs of residuals within the innermost group level. The autocorrelation function is useful for investigating serial correlation models for equally spaced data.”

3.1 Full data set

```
keep<-(!is.na(x$logcpue))
lme.out<-nlme::lme(logcpue ~ age + yearFactor + monthFactor + region + taxa,
                     random = ~ 1|fed/uniqueid, data=x,subset=keep)
summary(lme.out)$tTable
```

	Value	Std.Error	DF	t-value	p-value
## (Intercept)	1.403663443	0.41978772	183217	3.34374582	8.267199e-04
## age	-0.195871810	0.05006573	183217	-3.91229336	9.145672e-05
## yearFactor2009	0.027262125	0.01630194	183217	1.67232427	9.446210e-02
## yearFactor2010	-0.109544495	0.01649317	183217	-6.64180891	3.107173e-11
## yearFactor2011	0.023352258	0.01669296	183217	1.39892864	1.618361e-01
## yearFactor2012	-0.073363190	0.01635476	183217	-4.48573939	7.270559e-06
## yearFactor2013	-0.009533991	0.01666463	183217	-0.57210948	5.672485e-01
## yearFactor2014	0.016596349	0.01664188	183217	0.99726401	3.186377e-01

```

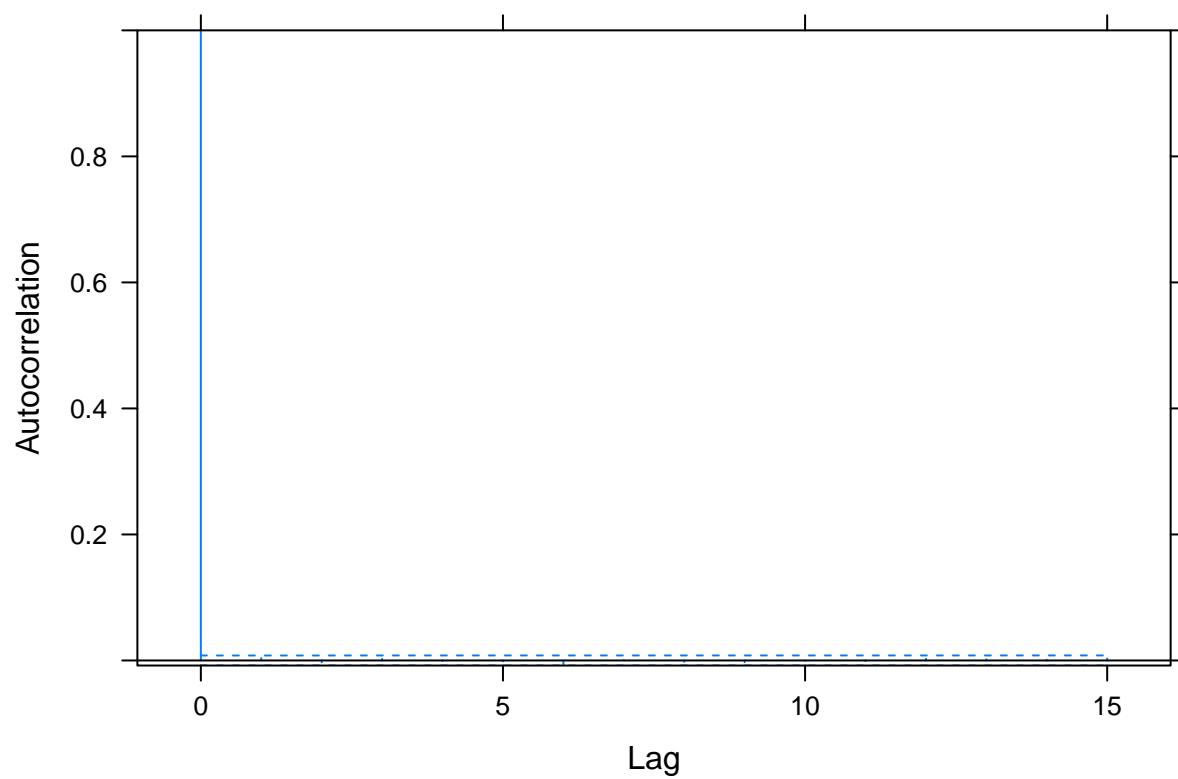
## yearFactor2015      0.005374221 0.01612237 183217    0.33333944 7.388785e-01
## yearFactor2016      0.056012774 0.01570492 183217    3.56657478 3.617698e-04
## monthFactor2        0.035024597 0.01879784 183217    1.86322403 6.243235e-02
## monthFactor3        0.128733929 0.01928850 183217    6.67412803 2.494106e-11
## monthFactor4        0.108429593 0.01939200 183217    5.59145938 2.254885e-08
## monthFactor5        0.153125799 0.01900357 183217    8.05773959 7.818014e-16
## monthFactor6        0.177137033 0.01885260 183217    9.39589312 5.735454e-21
## monthFactor7        0.234956217 0.01801286 183217    13.04380656 7.173641e-39
## monthFactor8        0.190238515 0.01790645 183217    10.62402447 2.345190e-26
## monthFactor9        0.261755659 0.01857591 183217    14.09113312 4.546237e-45
## monthFactor10       0.212403148 0.01807485 183217    11.75130642 7.139438e-32
## monthFactor11       0.167565715 0.01839680 183217    9.10841468 8.440205e-20
## monthFactor12       0.132478801 0.01870975 183217    7.08073436 1.439046e-12
## region2            0.539233765 0.61064720    124    0.88305287 3.789165e-01
## region3            2.078880747 0.46305246     52    4.48951448 3.990735e-05
## region4            1.779437804 0.43295413    124    4.10999157 7.140397e-05
## region5            -0.009984212 0.50885880    124    -0.01962079 9.843774e-01
## region6            0.679820678 0.43890199    124    1.54891226 1.239513e-01
## taxaCentropomidae -0.965678518 0.02264463 183217    -42.64491975 0.000000e+00
## taxaCichlidae      -0.239872411 0.03169818 183217    -7.56738736 3.825719e-14
## taxaClupeidae      2.519489134 0.09733478 183217    25.88477781 1.824149e-147
## taxaElasmobranchii -0.035590802 0.02147244 183217    -1.65751100 9.741792e-02
## taxaGerreidae       -0.748917162 0.02405025 183217    -31.13968742 2.518876e-212
## taxaHaliotidae      1.314033166 0.03325820 183217    39.51005498 0.000000e+00
## taxaLutjanidae      -0.426832776 0.01958706 183217    -21.79156898 3.797143e-105
## taxaMugilidae       0.003277102 0.02865295 183217    0.11437223 9.089429e-01
## taxaOctopodidae     1.240885722 0.03009656 183217    41.23015203 0.000000e+00
## taxaOmmastrephidae  2.599530248 0.08024495 183217    32.39493979 1.449267e-229
## taxaOstreidae       1.952268769 0.03814656 183217    51.17811472 0.000000e+00
## taxaOTRAS            -0.558432769 0.01741086 183217    -32.07382565 4.319535e-225
## taxaPalinuridae     0.663597312 0.02706263 183217    24.52080346 1.452410e-132
## taxaPectinidae      3.039571059 0.06591888 183217    46.11077891 0.000000e+00
## taxaPenaeidae       0.095062030 0.02406377 183217    3.95042196 7.804271e-05
## taxaPortunidae      0.907555138 0.03057724 183217    29.68074025 3.920776e-193
## taxaSciaenidae      -0.170047445 0.02124898 183217    -8.00261698 1.225070e-15
## taxaScombridae       0.047579863 0.03057967 183217    1.55593096 1.197262e-01
## taxaSerranidae       0.556641242 0.02009494 183217    27.70057070 1.534479e-168
## taxaStichopodidae   3.192020724 0.06563502 183217    48.63288706 0.000000e+00

```

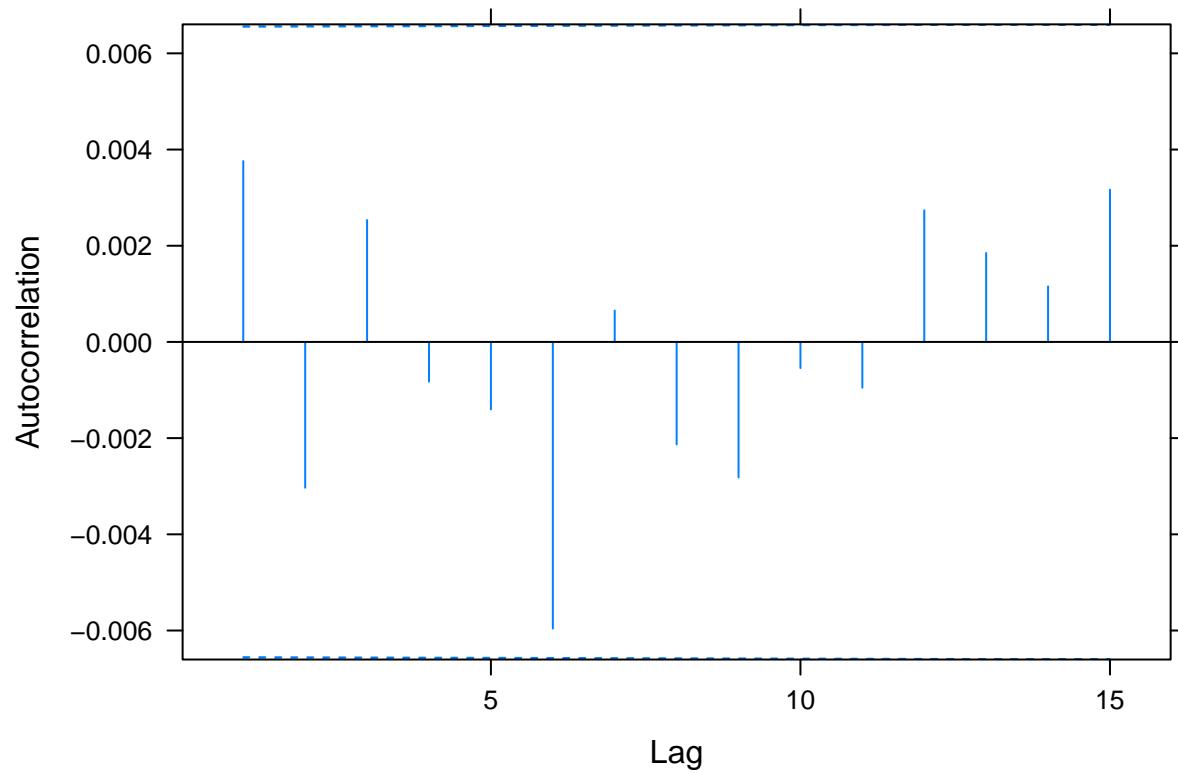
```

acf.lme<-nlme:::ACF(lme.out,maxLag=15)
plot(acf.lme,alpha=0.01/15)

```

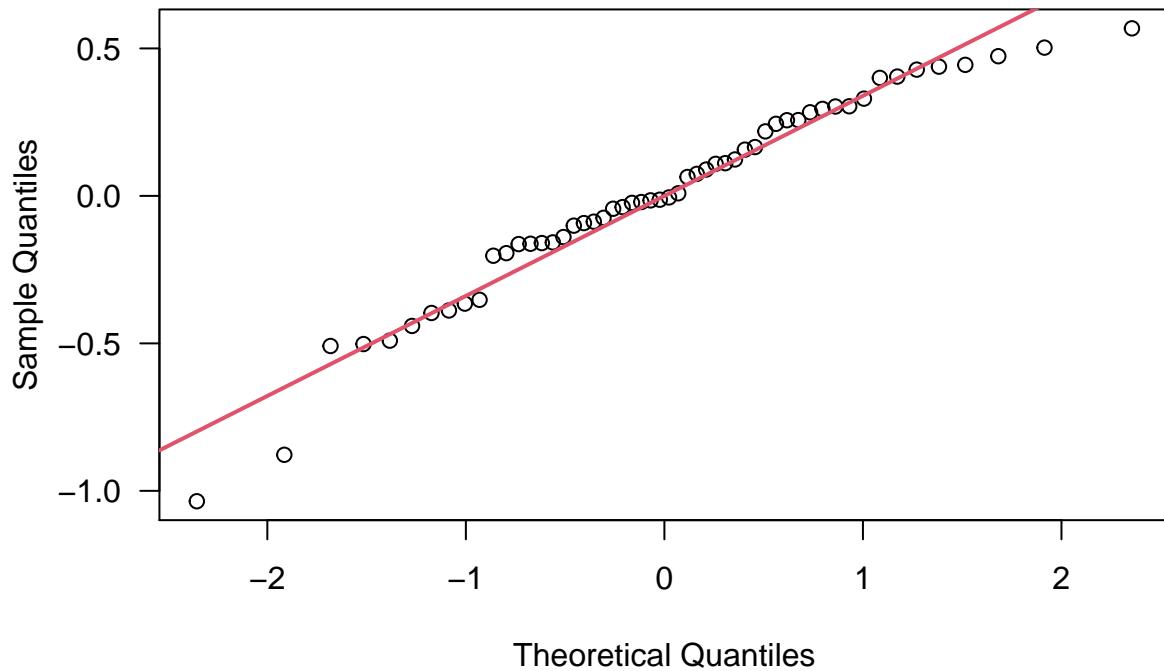


```
acf.lme<-acf.lme[-1,]  
plot(acf.lme,alpha=0.005)
```



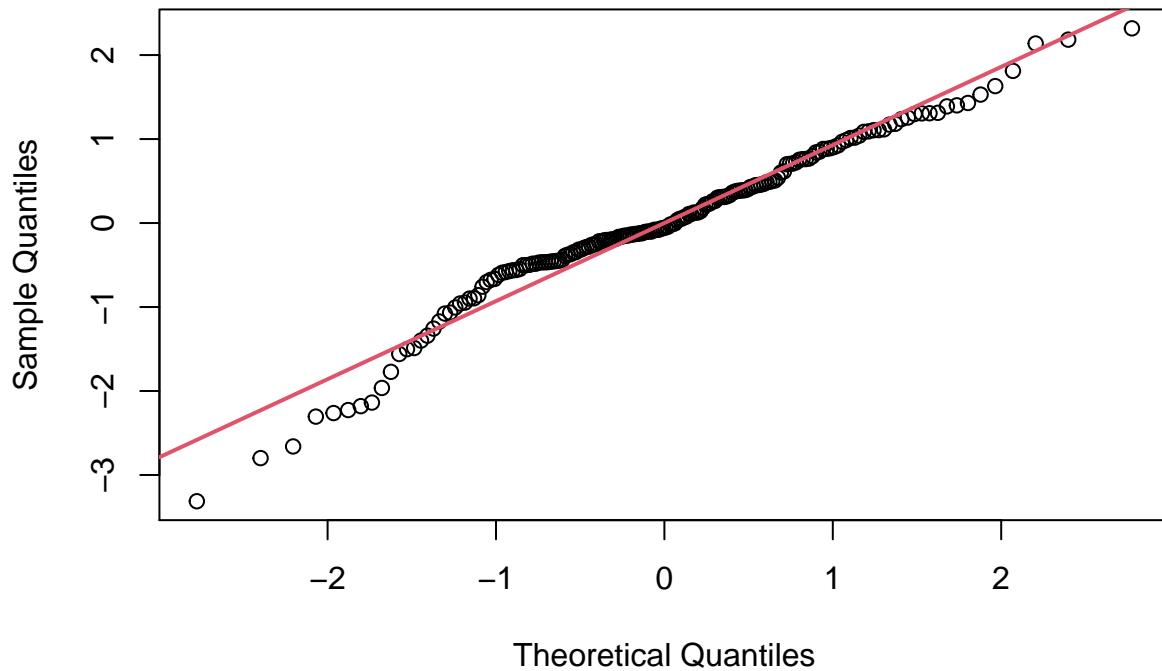
```
qqnorm(fedn.re<-unlist(nlme::ranef(lme.out)$fed),las=1)
abline(a=0,b=sd(fedn.re),lwd=2,col=2)
```

Normal Q-Q Plot



```
qqnorm(uid.re<-unlist(nlme::ranef(lme.out)$uniqueid))  
abline(a=0,b=sd(uid.re),lwd=2,col=2)
```

Normal Q-Q Plot



3.2 Single taxa model

```
keep<-((x$taxa=="OTRAS")&(!is.na(x$logcpue)))
otras.out<-nlme::lme(logcpue ~ age + yearFactor + monthFactor + region,
                      random = ~ 1|fed/uniqueid, data=x,subset=keep)
summary(otras.out)$tTable
```

	Value	Std.Error	DF	t-value	p-value
## (Intercept)	1.08138634	0.84421113	36105	1.28094300	2.002219e-01
## age	-0.28365849	0.08467958	36105	-3.34978610	8.095677e-04
## yearFactor2009	0.04673876	0.03509398	36105	1.33181716	1.829287e-01
## yearFactor2010	-0.09705218	0.03579162	36105	-2.71158925	6.699319e-03
## yearFactor2011	0.13060751	0.03577393	36105	3.65091294	2.616786e-04
## yearFactor2012	0.15795034	0.03464800	36105	4.55871433	5.163641e-06
## yearFactor2013	0.08099962	0.03501745	36105	2.31312171	2.072154e-02
## yearFactor2014	0.09254657	0.03552113	36105	2.60539506	9.180602e-03
## yearFactor2015	0.19715933	0.03424718	36105	5.75695037	8.633969e-09
## yearFactor2016	0.24684516	0.03310370	36105	7.45672392	9.068682e-14
## monthFactor2	0.04320187	0.03999201	36105	1.08026257	2.800325e-01
## monthFactor3	0.03424256	0.03893670	36105	0.87944178	3.791676e-01
## monthFactor4	-0.02316908	0.03917392	36105	-0.59144151	5.542283e-01
## monthFactor5	0.08849087	0.03770718	36105	2.34679076	1.894126e-02
## monthFactor6	0.20005985	0.03774298	36105	5.30058458	1.161094e-07
## monthFactor7	0.13976807	0.03740019	36105	3.73709550	1.864450e-04

```

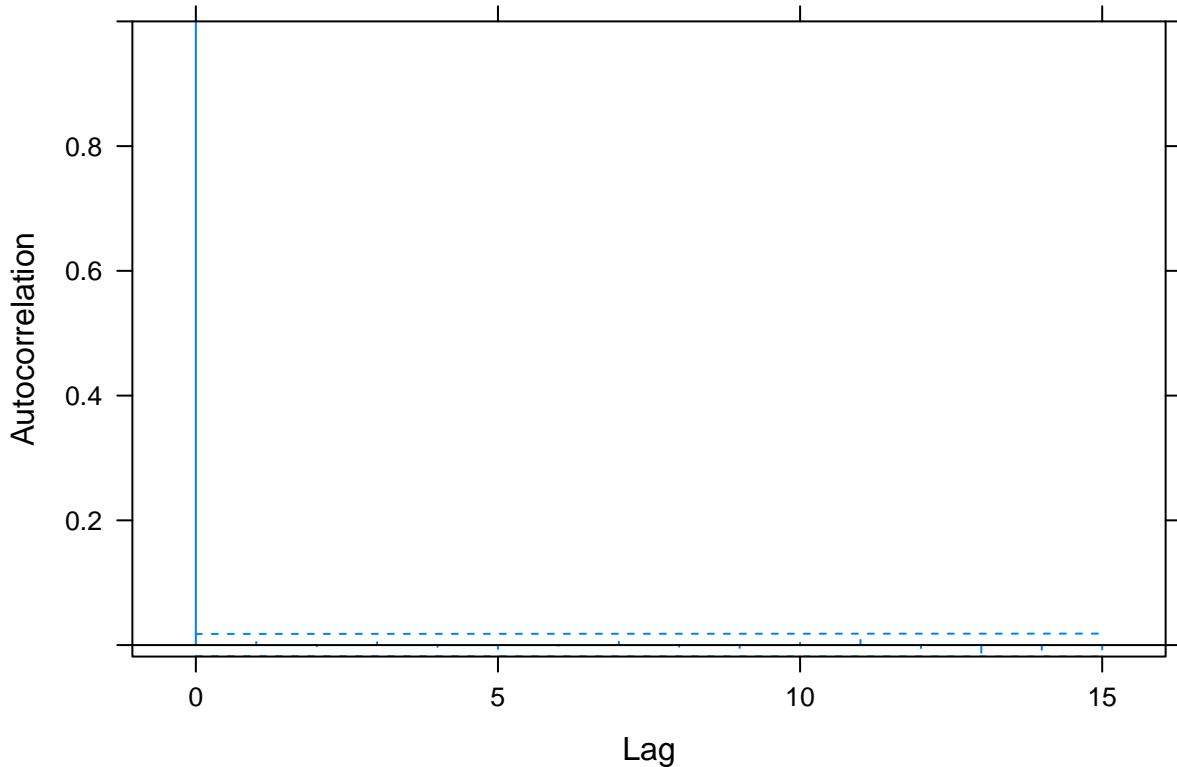
## monthFactor8 -0.04185668 0.03853893 36105 -1.08608830 2.774472e-01
## monthFactor9 -0.15388514 0.04165876 36105 -3.69394395 2.211281e-04
## monthFactor10 -0.04671011 0.04076475 36105 -1.14584556 2.518667e-01
## monthFactor11 -0.00108303 0.04102094 36105 -0.02640188 9.789369e-01
## monthFactor12 0.02164900 0.04099621 36105 0.52807330 5.974517e-01
## region2 0.50800884 1.03291245 87 0.49182178 6.240840e-01
## region3 2.51046462 0.88150159 46 2.84794111 6.557393e-03
## region4 2.13323557 0.87085476 87 2.44958822 1.630636e-02
## region5 0.29610271 0.93507039 87 0.31666355 7.522575e-01
## region6 0.50720661 0.85832077 87 0.59092897 5.561001e-01

```

```

acf.otras<-nlme:::ACF(otras.out,maxLag=15)
plot(acf.otras,alpha=0.01/15)

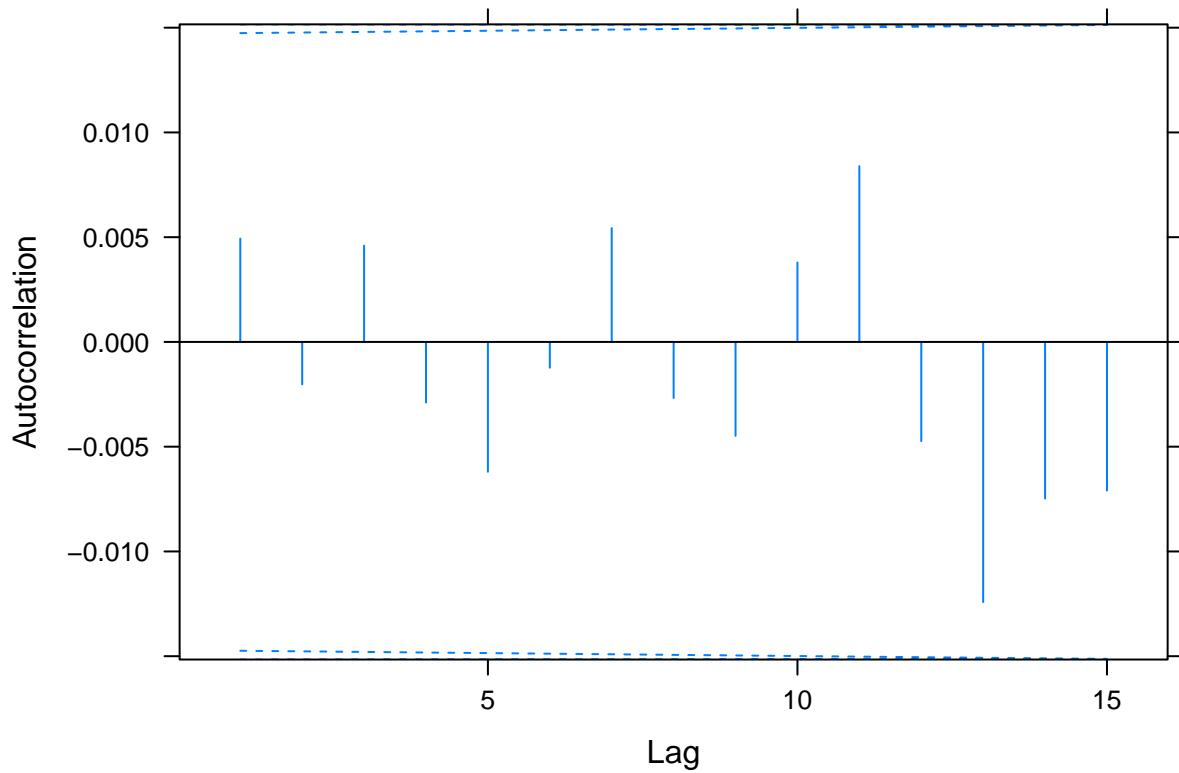
```



```

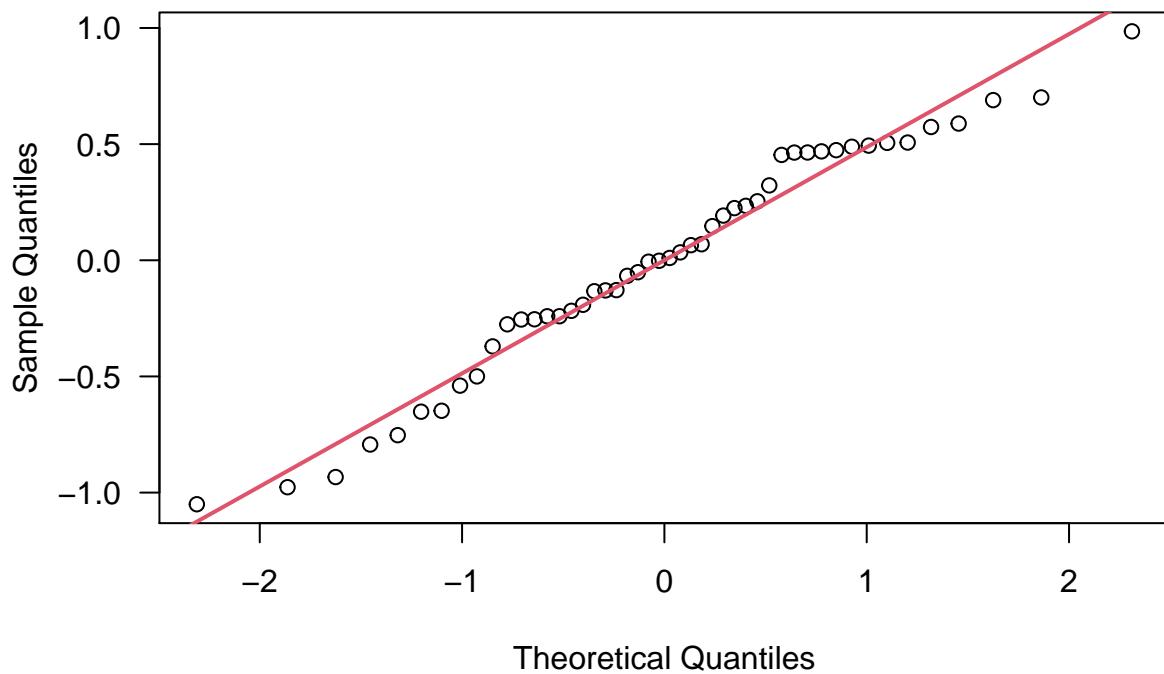
acf.otras<-acf.otras[-1,]
plot(acf.otras,alpha=0.005)

```

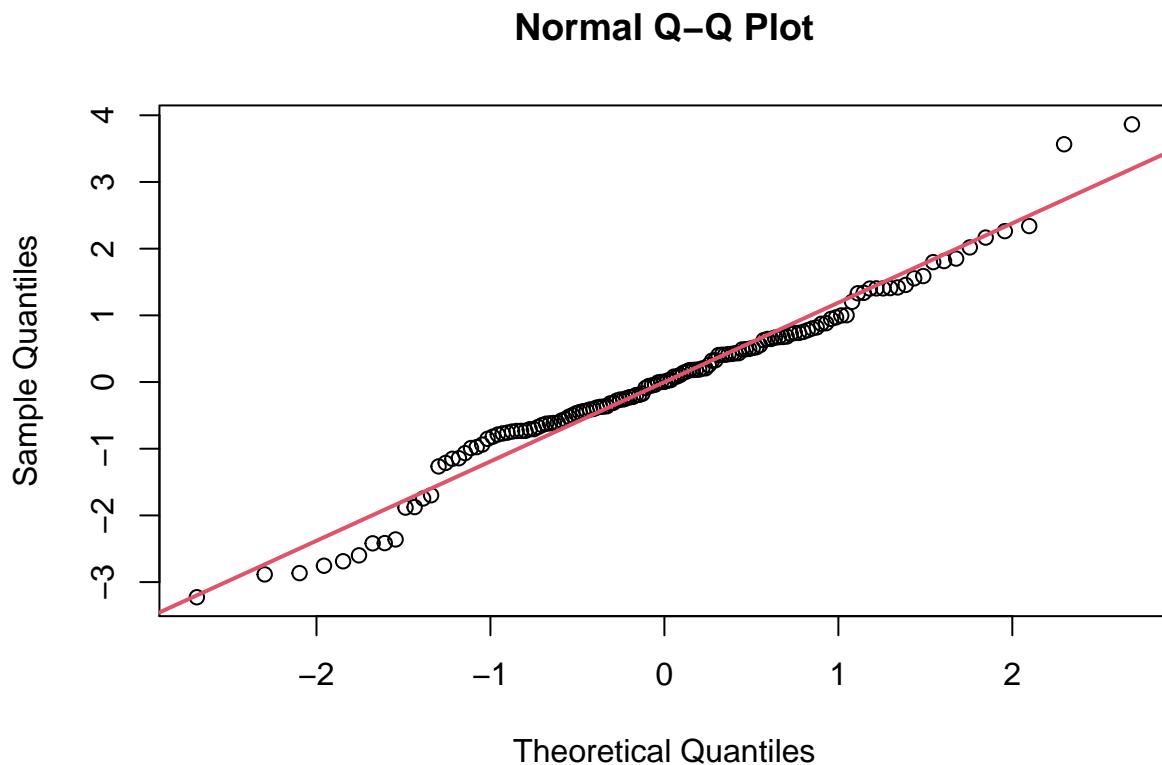


```
qqnorm(fedn.re<-unlist(nlme::ranef(otras.out)$fed),las=1)
abline(a=0,b=sd(fedn.re),lwd=2,col=2)
```

Normal Q-Q Plot



```
qqnorm(uid.re<-unlist(nlme::ranef(otras.out)$uniqueid))
abline(a=0,b=sd(uid.re),lwd=2,col=2)
```



4. Your mgcv Models

EDA

```
response_summary <- x %>%
  summarise(mean = mean(logcpue, na.rm=TRUE),
            median = median(logcpue, na.rm=TRUE),
            variance = var(logcpue, na.rm=TRUE),
            IQR = IQR(logcpue, na.rm=TRUE),
            sd = sd(logcpue, na.rm=TRUE),
            min = min(logcpue, na.rm=TRUE),
            max = max(logcpue, na.rm=TRUE))
knitr::kable(response_summary, caption = "Summary Statistics for \n the logcpue")
```

Table 1: Summary Statistics for the logcpue

mean	median	variance	IQR	sd	min	max
1.740824	1.966113	4.71467	2.888958	2.171329	-7.863267	10.04138

```
ggplot(data = x, aes(x = logcpue)) +
  geom_histogram(fill = "darkgreen", color = "black") +
  labs(x = "logcpue", y = "Count",
```

```

    title = "Distribution of logcpue",
    caption = "Figure 1") +
theme_bw()

```

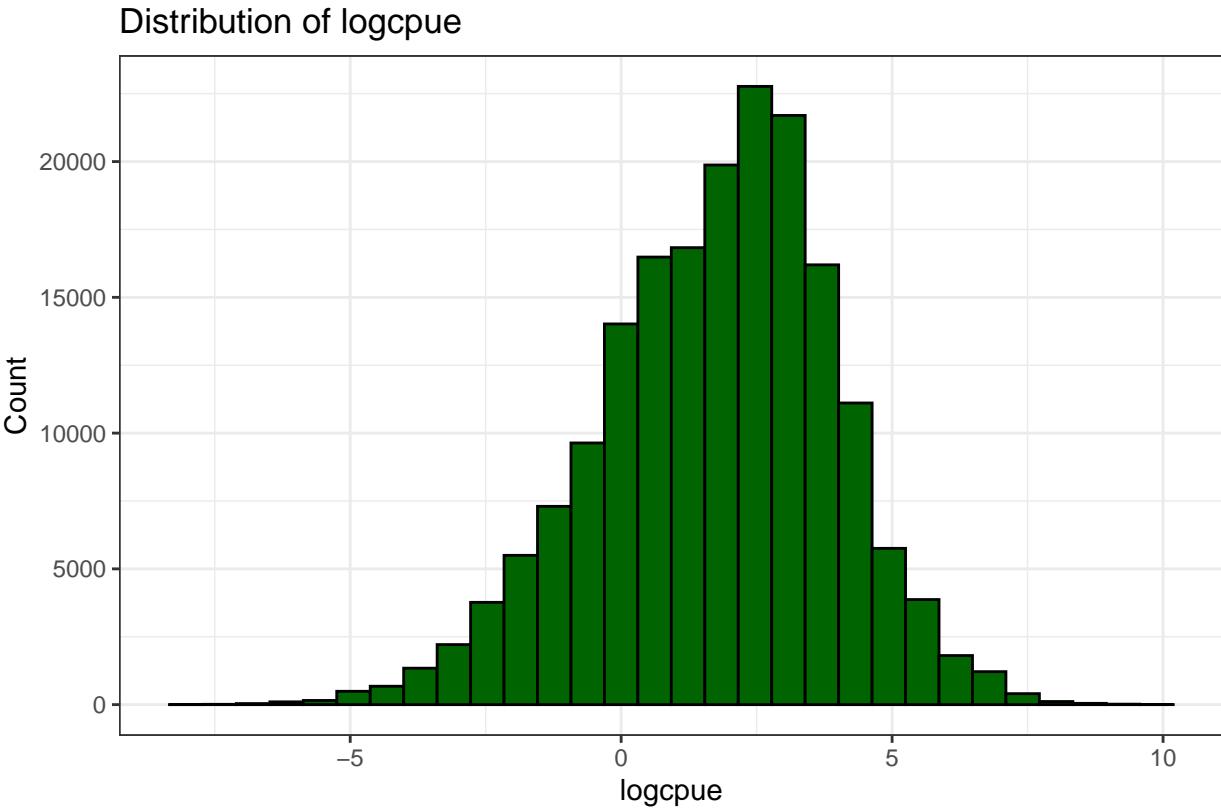


Figure 1

In this dataset, the response variable is `logcpue`, which represents catch per unit of effort, measured during different time points (Figure 1, Table 1). The distribution of this variable is unimodal and slightly skewed to the left. The center, defined by median, is 1.97. The spread, defined by the interquartile range (IQR), is 2.89. There are no evident outliers based on the histogram shown in Figure 1. To better capture the distribution of `logcpue` for individual cooperation, a random sample of 30 cooperation were selected, and their `logcpue` measured at different time points were plotted (Figure 2). Indeed, cooperation exhibit different patterns for the distribution of `logcpue`. Some shows a bimodel distribution, while some shows a unimodel distribution. This suggests that the distribution of `logcpue` differs based on cooperations, so we decided to include it as an individual level in our LMM model.

```

set.seed(031622)
# get sample of 30 cooperations

sample_org <- x %>%
  distinct(uniqueid) %>%
  sample_n(30) %% pull()
# get data for those cooperation
sample_data <- x %>%
  filter(uniqueid %in% sample_org)
# make a histogram of the response for each lemur
ggplot(data = sample_data, aes(x = logcpue)) +
  geom_histogram(fill = "darkgreen", color = "black") +

```

```

facet_wrap(~ uniqueid, scales = "free") +
  labs(x = "logcpue",
       title = "Distribution of logcpue by randomly sampled 30 cooperation",
       caption = "Figure 2") +
  theme_bw()

```

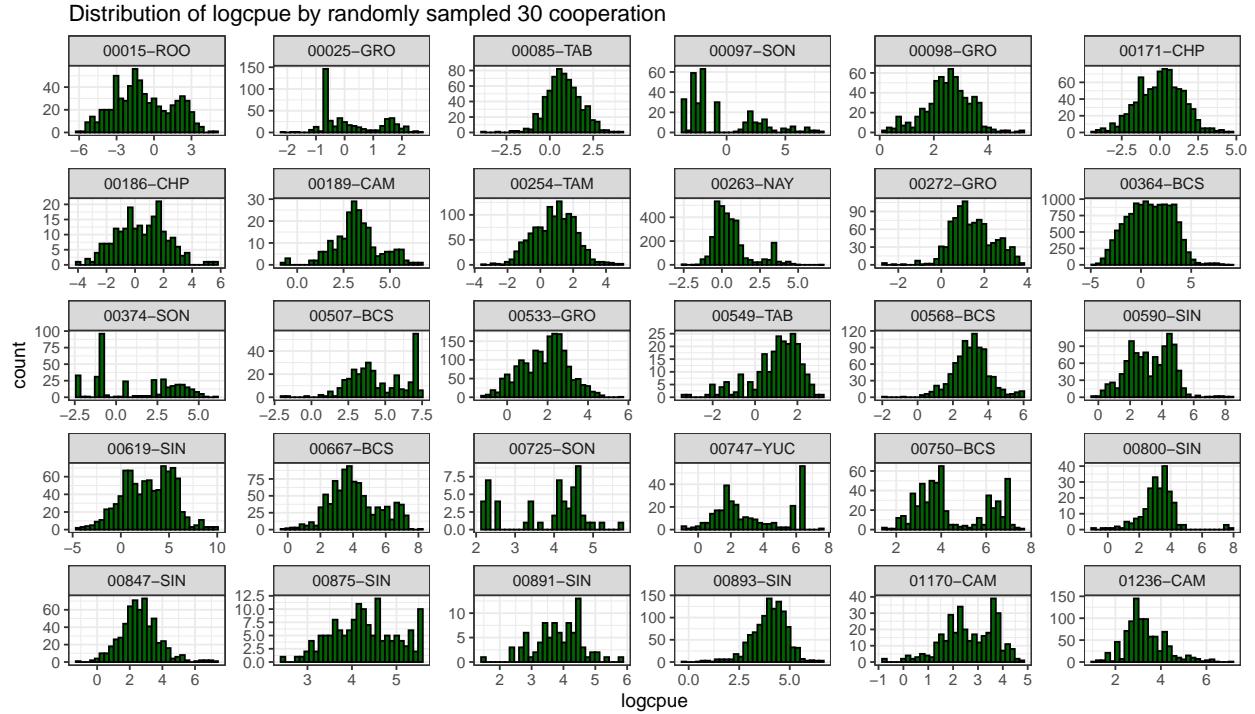


Figure 2

Other variables in the data that are of interests include year and month of each fishing event, region, taxa, and age of the federations. Figure 3 shows the relationships between year and logcpue for 30 randomly selected cooperations. Figure 4 shows the relationships between month and logcpue for 30 randomly selected cooperations. While some cooperations have a constant logcpue value across years or months, some cooperations have varying logcpue based on the year or month the fishery takes place, indicating that the effects of year or month on logcpue varies based on cooperations.

```

ggplot(data = sample_data, aes(x = yearFactor, y = logcpue)) +
  geom_boxplot(fill = "darkgreen", color = "black") +
  facet_wrap(~ uniqueid, scales = "free") +
  labs(x = "Year",
       y = "logcpue",
       title = "logcpue vs year",
       caption = "Figure 3") +
  theme_bw()

```

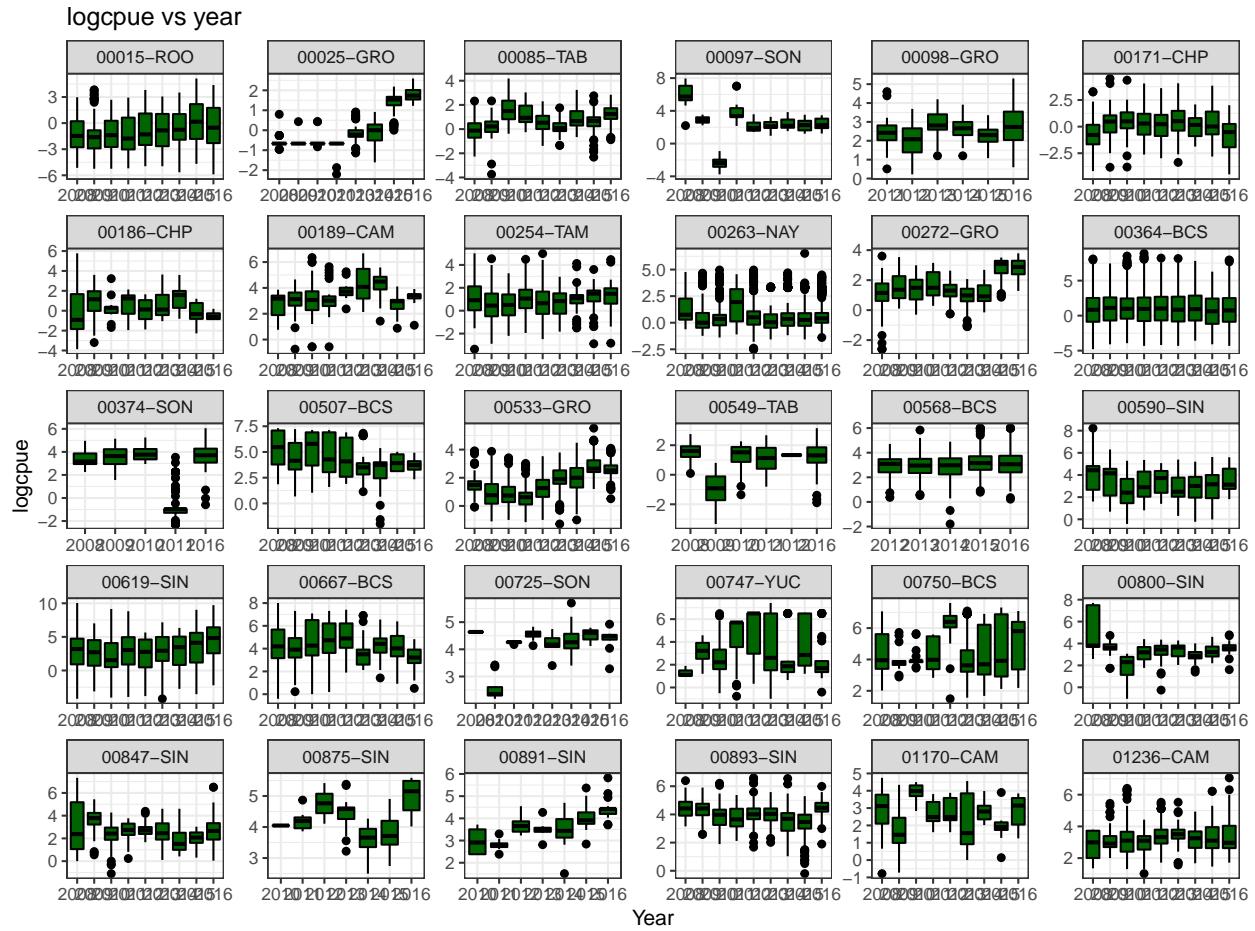


Figure 3

```
ggplot(data = sample_data, aes(x = monthFactor, y = logcpue)) +
  geom_boxplot(fill = "darkgreen", color = "black") +
  facet_wrap(~ uniqueid, scales = "free") +
  labs(x = "Month",
       y = "logcpue",
       title = "logcpue vs month",
       caption = "Figure 4") +
  theme_bw()
```

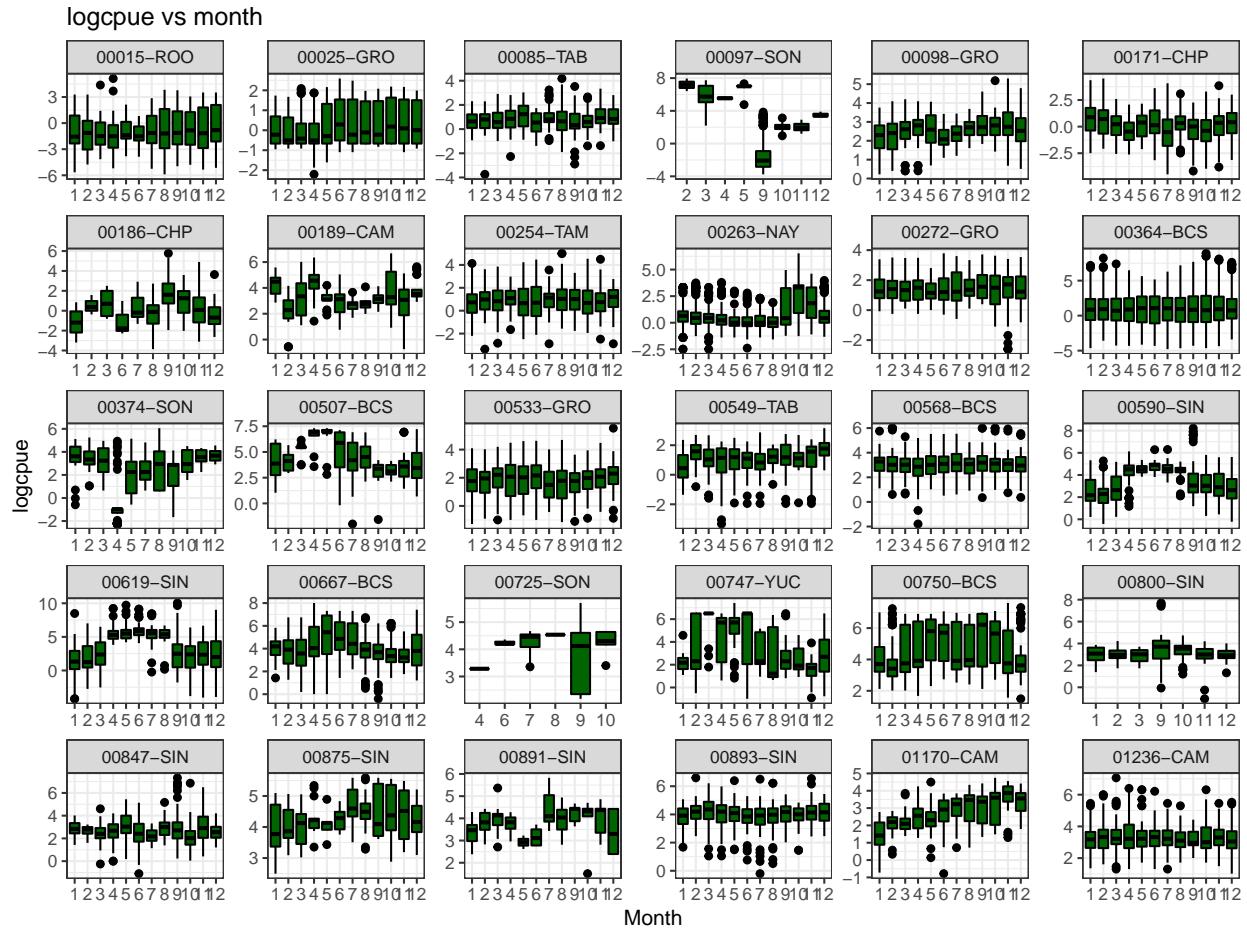


Figure 4

For simplicity purposes, we decided to combine year and month to make a new continuous variable that incorporated both year and month.

```
x$yearMonth <- ((x$year - min(x$year)) *12 + x$month)
```

In addition to year and month, region and taxa also play an important role in explaining logcpue. In particular, logcpue differs based on the region where the fisheries take place and on the taxa of the fish caught (Figure 5,6). Therefore, these two variables are worth including in the final models.

```
ggplot(data = x, aes(x = region, y = logcpue)) +
  geom_boxplot(fill = "darkgreen", color = "black") +
  labs(x = "Region",
       y = "logcpue",
       title = "logcpue vs. region",
       caption = "Figure 5") +
  theme_bw()
```

logcpue vs. region

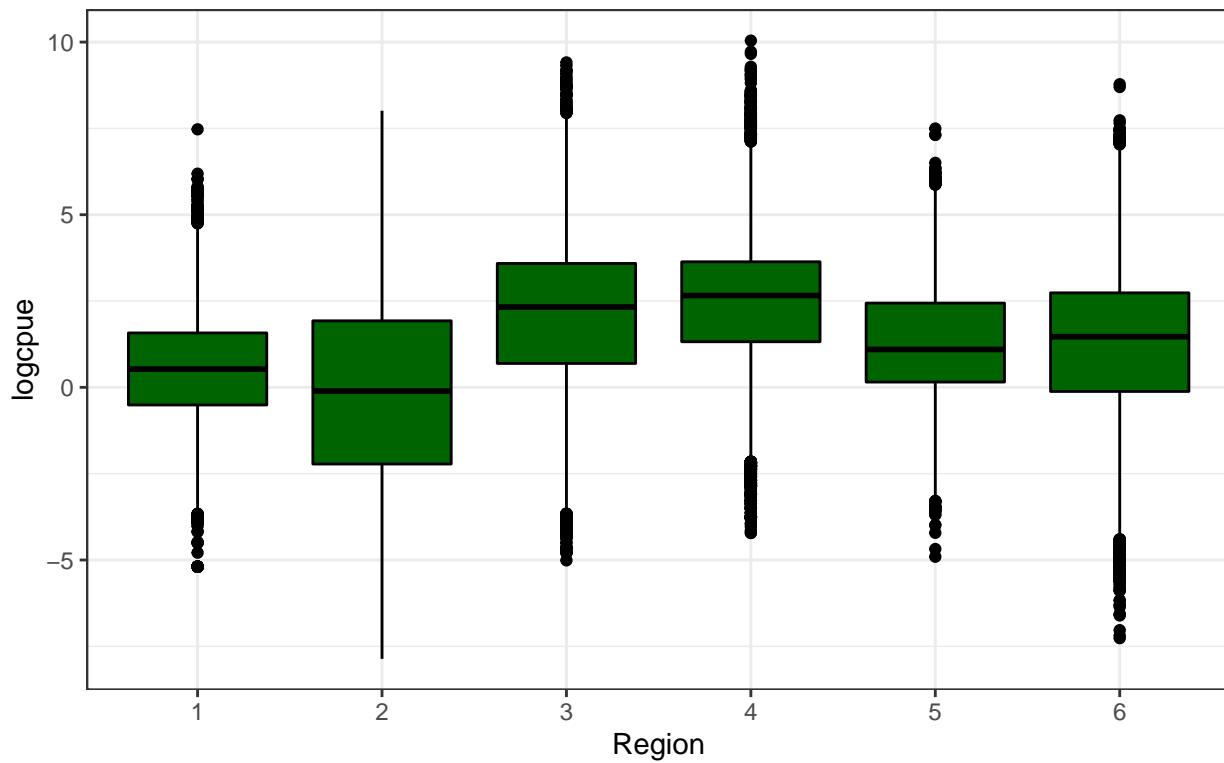


Figure 5

```
ggplot(data = x, aes(x = taxa, y = logcpue)) +  
  geom_boxplot(fill = "darkgreen", color = "black") +  
  labs(x = "Taxa",  
       y = "logcpue",  
       title = "logcpue vs. taxa",  
       caption = "Figure 6") +  
  theme_bw()
```

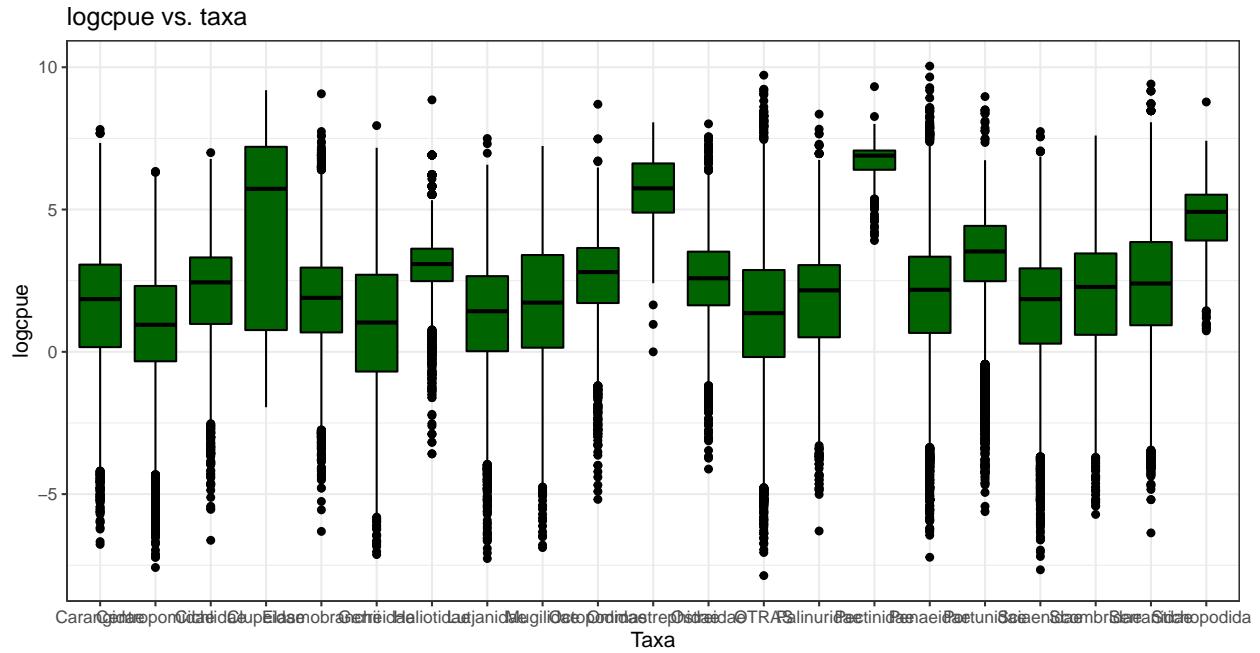


Figure 6

Methods

Both models using mgcv and lme are included in this section.

Since catch per unit of efforts (cpue) were measured at different time points for each cooperation and that cooperations from the same federations tend to have similar distribution of logcpue, a multilevel model analysis was used to model the distribution of logcpue. The multilevel analysis was done at three levels. The first level included time points when cpue were collected; the second level included individual cooperations (`uniqueid`); and the third level included federations (`fed`). To test whether a multilevel model is necessary to capture the distribution of logcpue, an unconditional means model in which there are no predictors at any level was first fitted, and intraclass correlation was calculated to estimate the relative variability between cooperations and between federations. Indeed, the intraclass correlation for cooperations is about 0.22 and for federations is about 0.26, which means that the average correlation between any two responses from the same cooperations or from the same federations is about 0.22 and 0.26 respectively. This suggests that about 48% of the variability in the logcpue is explained by cooperation to cooperation variability and federation to federation variability. Knowing cooperation and federation, therefore, can explain almost half of the variability in the data, providing evidence that the multilevel model structure is useful in this setting.

```
unconditional_means_mgcv <- bam(logcpue ~ 1 + s(fed, bs = "re") + s(uniqueid, bs = "re"),
                                 data = x, na.action = na.omit)
summary(unconditional_means_mgcv)
```

```
##  
## Family: gaussian  
## Link function: identity  
##  
## Formula:  
## logcpue ~ 1 + s(fed, bs = "re") + s(uniqueid, bs = "re")  
##  
## Parametric coefficients:
```

```

##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.2669     0.1898   11.94   <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(fed)    38.5    53 710455 0.00906 **
## s(uniqueid) 140.5   181 238192 0.00028 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.396  Deviance explained = 39.7%
## fREML = 3.5679e+05  Scale est. = 2.8475    n = 183439

unconditional_means <- lme(logcpue ~ 1,
                           list(fed = ~1, uniqueid = ~1),
                           data = x, na.action = na.omit)
summary(unconditional_means)

## Linear mixed-effects model fit by REML
## Data: x
##      AIC      BIC      logLik
## 713580.5 713621 -356786.2
##
## Random effects:
## Formula: ~1 | fed
##           (Intercept)
## StdDev: 1.189196
##
## Formula: ~1 | uniqueid %in% fed
##           (Intercept) Residual
## StdDev: 1.093864 1.687456
##
## Fixed effects: logcpue ~ 1
##                 Value Std.Error DF t-value p-value
## (Intercept) 2.266913 0.189815 183257 11.94275 0
##
## Standardized Within-Group Residuals:
##      Min        Q1        Med        Q3        Max
## -4.25341107 -0.59958742 -0.01052429  0.60605097  5.54494371
##
## Number of Observations: 183439
## Number of Groups:
##          fed uniqueid %in% fed
##            54                182

```

Next, to understand the effect of time before adding other Level One covariates, the model that only includes year and month as Level One predictors was fitted. The Pseudo R^2 estimating the change in within-cooperation variance between the unconditional means and the Model with Time is about 0.42. Therefore, understanding changes of time (i.e. years) accounts for 42% variability in logcpue.

```

model_time <- lme(logcpue ~ yearMonth, random = list(fed = ~yearMonth, uniqueid = ~yearMonth),
                     data = x, na.action = na.omit)
summary(model_time)

## Linear mixed-effects model fit by REML
## Data: x
##      AIC      BIC      logLik
## 704386.3 704477.4 -352184.2
##
## Random effects:
##   Formula: ~yearMonth | fed
##   Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev     Corr
## (Intercept) 1.466296838 (Intr)
## yearMonth   0.008859054 -0.69
##
##   Formula: ~yearMonth | uniqueid %in% fed
##   Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev     Corr
## (Intercept) 1.27828351 (Intr)
## yearMonth   0.01280207 -0.537
## Residual    1.64283974
##
## Fixed effects: logcpue ~ yearMonth
##                 Value Std.Error DF t-value p-value
## (Intercept) 2.2279527 0.23275696 183256 9.572013 0.0000
## yearMonth   0.0004842 0.00166952 183256 0.290035 0.7718
##
## Correlation:
##           (Intr)
## yearMonth -0.632
##
## Standardized Within-Group Residuals:
##      Min       Q1       Med       Q3       Max
## -4.432812192 -0.578993635 -0.006224082  0.588661168  5.862754093
##
## Number of Observations: 183439
## Number of Groups:
##                 fed uniqueid %in% fed
##                  54            182

# Pseudo R^2
sigm_moda <- 1.687456^2
sigm_madb <- 1.64283974
r <- (sigm_moda - sigm_madb)/sigm_moda
r

## [1] 0.4230605

```

Based on the EDA and model with only time as level one predictors, we decided to include `yearMonth`, `region`, and `taxa` as our level one predictors. `age2` was included as level three (federation) predictor since we hypothesized that the age of federation would impact the efficiency of catching fish (`logcpue`) of its subsidiary cooperation. Therefore, two models were fitted. One that includes interactions between age

of federations and region and between age of federations and taxa, and one that does not include any interactions. AIC and BIC of both models were compared. Since both AIC and BIC are lower for model with additional interaction effects, providing evidence that model with interactions is a better model. One thing to note is that no random slopes were included in level two or three in both model because adding them make the model too complex and too computational heavy.

The interaction between age2 and yearMonth was not considered because we believe that the age of federation does not impact the effects of year and month of fishing on logcpue.

```
model_no_int_mgcv <- bam(logcpue ~ yearMonth + region + taxa + age2
                           + s(fed, bs = "re") + s(uniqueid, bs = "re"),
                           data = x, na.action = na.omit)
```

```
model_no_int <- lme(logcpue ~ yearMonth + region + taxa + age2,
                      list(fed = ~1, uniqueid = ~1),
                      data = x, na.action = na.omit)
```

```
model_int_mgcv <- bam(logcpue ~ yearMonth + region + taxa + age2
                        + age2:region + age2:taxa +
                        + s(fed, bs = "re") + s(uniqueid, bs = "re"),
                        data = x, na.action = na.omit)
```

```
model_int <- lme(logcpue ~ yearMonth + region + taxa + age2
                     + age2:region + age2:taxa,
                     list(fed = ~1, uniqueid = ~1),
                     data = x, na.action = na.omit)
```

```
glance(model_int_mgcv) %>%
  select(AIC, BIC) %>%
  kable(digits = 0, caption = "AIC and BIC for Model without Interaction")
```

Table 2: AIC and BIC for Model without Interaction

AIC	BIC
679974	682277

```
glance(model_no_int_mgcv) %>%
  select(AIC, BIC) %>%
  kable(digits = 0, caption = "AIC and BIC for Model with Interactions")
```

Table 3: AIC and BIC for Model with Interactions

AIC	BIC
688412	690458

Moreover, we have also tried adding different interactions in the model, such as interactions effects between yearMonth and region and between yearMonth and taxa as shown below. However, the model selection criterion by AIC and BIC is higher than that for the previous model. Therefore, model with interactions between age2 and region and between age2 and taxa performed better than other interactions models.

```

model_int2_mgcv <- bam(logcpue ~ yearMonth + region + taxa + age2
                         + yearMonth:region + yearMonth:taxa +
                         + s(fed, bs = "re") + s(uniqueid, bs = "re"),
                         data = x, na.action = na.omit)

model_int2<- lme(logcpue ~ yearMonth + region + taxa + age2 + yearMonth:region + yearMonth:taxa,
                   list(fed = ~1, uniqueid = ~1),
                   data = x, na.action = na.omit)

glance(model_int2_mgcv) %>%
  select(AIC, BIC) %>%
  kable(digits = 0, caption = "AIC and BIC for Model with New Interactions")

```

Table 4: AIC and BIC for Model with New Interactions

AIC	BIC
684999	687297

Model in mathematical form

Level 1 Model: Time

$$Logcpue_{ijk} = a_{ij} + b_{ij}yearMonth_{ijk} + c_{ij}region_{ijk} + d_{ij}taxa_{ijk} + \epsilon_{ijk}, \epsilon_{ijk} \sim N(0, \sigma^2)$$

Level 2 Model: Cooperations

$$a_{ij} = a_i + u_{ij}$$

$$b_{ij} = b_i$$

$$c_{ij} = c_i$$

$$d_{ij} = d_i$$

Level 3 Model: Federations

$$a_i = \alpha_0 + \alpha_1 age_i + \tilde{u}_i$$

$$b_i = \beta_0$$

$$c_i = \gamma_0 + \gamma_1 age_i$$

$$d_i = \delta_0 + \delta_1 age_i$$

Composite model

$$\text{Logcpue}_{ijk} = a_i + u_{ij} + b_i \text{yearMonth}_{ijk} + c_i \text{region}_{ijk} + d_i \text{taxa}_{ijk} + \epsilon_{ijk}, \epsilon_{ijk} \sim N(0, \sigma^2)$$

$$\begin{aligned} &= \alpha_0 + \alpha_1 \text{age}_i + \tilde{u}_i + u_{ij} + \beta_0 \text{yearMonth}_{ijk} + \gamma_0 \text{region}_{ijk} + \gamma_1 \text{age}_i \text{region}_{ijk} + \delta_0 \text{taxa}_{ijk} + \delta_1 \text{age}_i \text{taxa}_{ijk} + \epsilon_{ijk} \\ &\quad \epsilon_{ijk} \sim N(0, \sigma^2) \\ &= \alpha_0 + \alpha_1 \text{age}_i + \beta_0 \text{yearMonth}_{ijk} + \gamma_0 \text{region}_{ijk} + \gamma_1 \text{age}_i \text{region}_{ijk} + \delta_0 \text{taxa}_{ijk} + \delta_1 \text{age}_i \text{taxa}_{ijk} \\ &\quad + [\tilde{u}_i + u_{ij} + \epsilon_{ijk}] \\ &\quad \epsilon_{ijk} \sim N(0, \sigma^2), u_{ij} \sim N(0, \sigma_u^2), \tilde{u}_i \sim N(0, \sigma_{\tilde{u}}^2) \end{aligned}$$

Model improvement based on 10/18 class code

Based on the bam code discussed in class on 10/18, we made some improvement to our model with interactions between age2 and region and between age2 and taxa. We decided to add smooth function to age2 since age2 is not strictly linear from the data. This model, indeed, has the lowest AIC and BIC values compared to our model without smooth term. Therefore, this model is our final model. Anova output of this model is provided in model diagnostics section.

```
k_values<-c(3,6,9,12,15)
AIC_ls<- c()

for (k in k_values){
  model_mgcv_sm <- bam(logcpue ~ yearMonth + region + taxa + s(age2, bs="cr", k=k)
    + s(age2,region,bs="fs",k=k)
    + s(age2,taxa,bs="fs",k=k) +
    + s(fed, bs = "re") + s(uniqueid, bs = "re"),
    data = x, na.action = na.omit)
  AIC=AIC(model_mgcv_sm)
  AIC_ls<- append(AIC_ls, AIC)
}
min_idx=which.min(AIC_ls)

k_values[min_idx]

## [1] 15

model_mgcv_sm_final <- bam(logcpue ~ yearMonth + region + taxa + s(age2, bs="cr", k=k)
  + s(age2,region,bs="fs",k=k_values[min_idx])
  + s(age2,taxa,bs="fs",k=k_values[min_idx]) +
  + s(fed, bs = "re") + s(uniqueid, bs = "re"),
  data = x, na.action = na.omit)

glance(model_mgcv_sm_final) %>%
  select(AIC, BIC) %>%
  kable(digits = 0, caption = "AIC and BIC for Model with smooth terms")
```

Table 5: AIC and BIC for Model with smooth terms

AIC	BIC
656082	661175

Sidenote

Shown below are two models we wish to fit that includes cooperation to cooperation difference in slopes for region and taxa. Based on our domain knowledge, the effects of region and taxa of the fish caught varies based on each cooperation since different cooperations may have different fishing strategies, etc. However, running the models below using either `mgcv` or `lme` took hours to run, so we decided to not select the model with random slope due to its computational load.

```
bam(logcpue ~ yearMonth + region + taxa + age2
  + age2:region + age2:taxa
  + s(fed, bs = "re") + s(uniqueid, bs = "re")
  + s(fed, bs = "re") + s(uniqueid, region, bs = "re")
  + s(uniqueid, taxa, bs = "re"),
  data = x, na.action = na.omit)

lme(logcpue ~ yearMonth + region + taxa + age2
  + age2:region + age2:taxa,
  list(fed = ~1, uniqueid = ~region + taxa),
  data = x, na.action = na.omit)
```

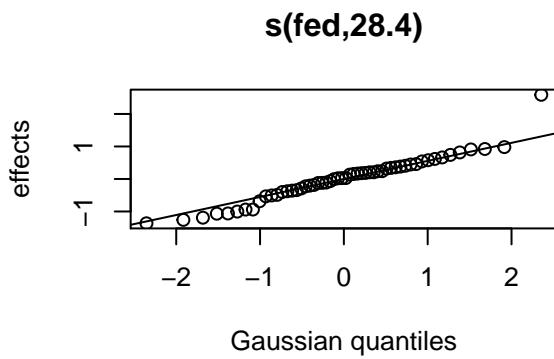
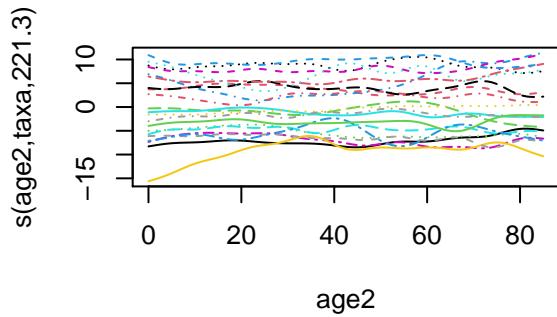
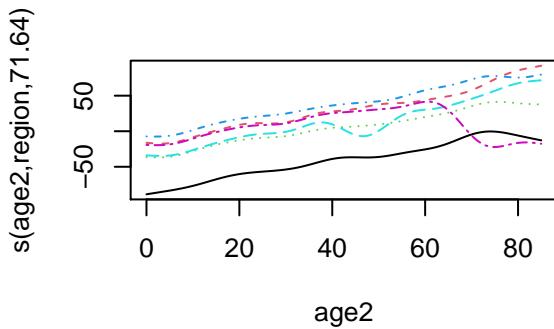
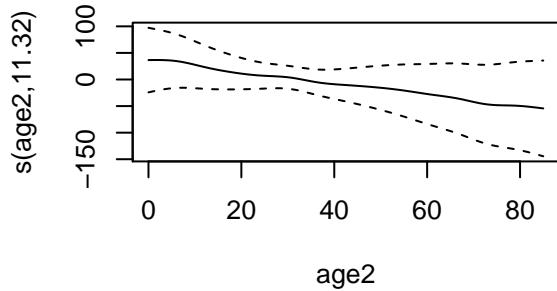
Model Diagnostics

```
anova(model_mgcv_sm_final)
```

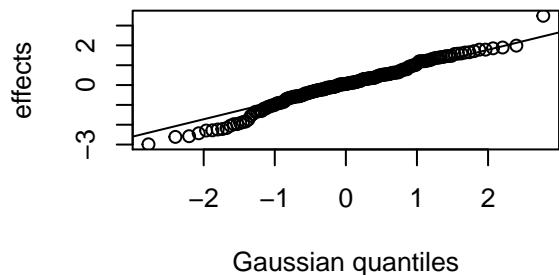
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## logcpue ~ yearMonth + region + taxa + s(age2, bs = "cr", k = k) +
##   s(age2, region, bs = "fs", k = k_values[min_idx]) + s(age2,
##   taxa, bs = "fs", k = k_values[min_idx]) + s(fed, bs = "re") +
##   s(uniqueid, bs = "re")
##
## Parametric Terms:
##           df      F p-value
## yearMonth 1 25.006 5.72e-07
## region    5  3.579  0.00308
## taxa      20 3.971 4.94e-09
##
## Approximate significance of smooth terms:
##             edf Ref.df      F p-value
## s(age2)       11.32  11.81 10.37 < 2e-16
## s(age2,region) 71.64  83.00 7464.86 < 2e-16
## s(age2,taxa)   221.30 291.00 766.59 < 2e-16
```

```
## s(fed)      28.40 51.00 7446.00 0.17878
## s(uniqueid) 140.29 176.00 585.24 0.00293
```

```
par(mfrow=c(2,2))
plot(model_mgcv_sm_final, scale=0)
```

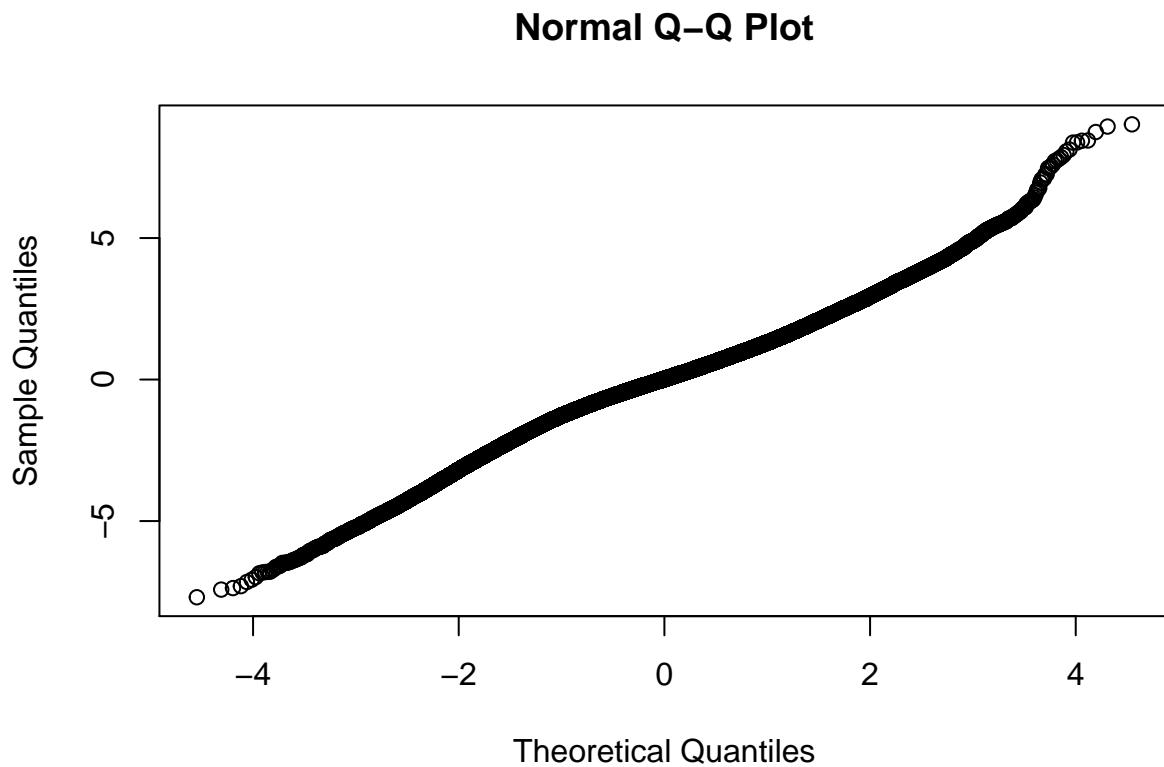


s(uniqueid,140.29)



Normality

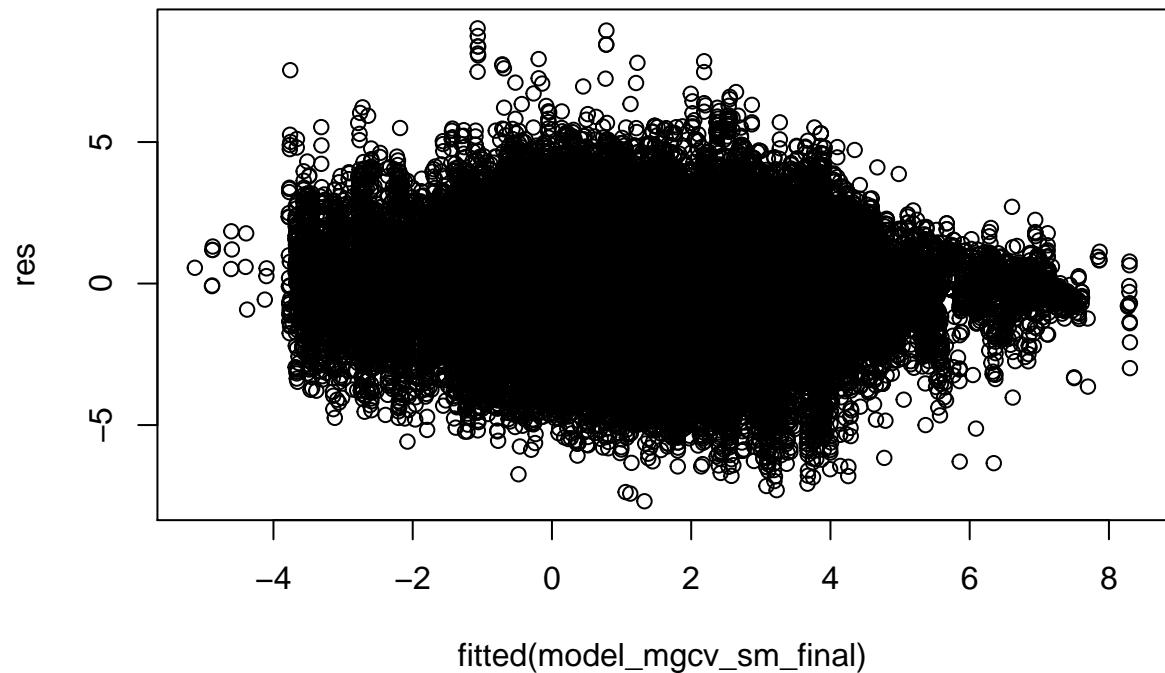
```
qqnorm(residuals(model_mgcv_sm_final))
```



Points mostly fall along a straight diagonal line on the normal quantile plot, so we can safely assume that the data is normally distributed.

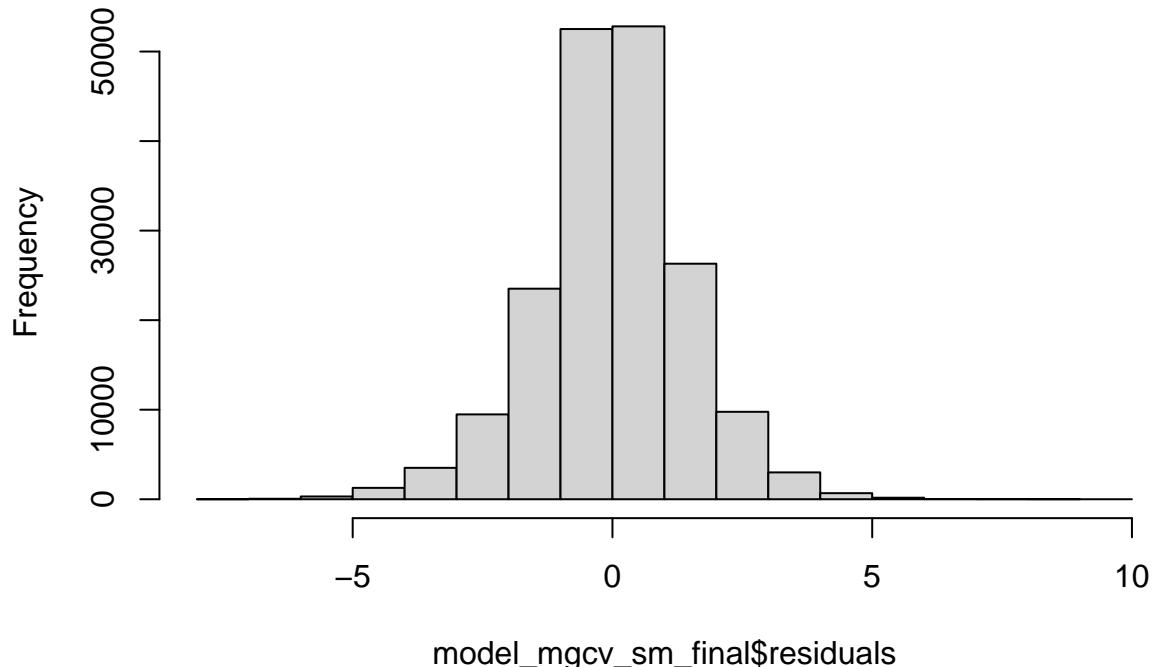
Residual plot

```
res <- resid(model_mgcv_sm_final)
plot(fitted(model_mgcv_sm_final), res)
```



```
hist(model_mgcv_sm_final$residuals)
```

Histogram of model_mgcv_sm_final\$residuals



The vertical spread of the residuals is not constant across the residual plot, suggesting that the constant variance condition is violated for this model. Therefore, some transformation should be done in the future to address this problem.

Prediction

```
# generate prediction file
model <- model_mgcv_sm_final
data <- x %%
  filter(is.na(logcpue))
#####
df <- data %>%
  mutate(prediction = predict(model,data))%>%
  select(c('prediction'))
df <- cbind(index = rownames(df),df)
df$index <- as.numeric(as.character(df$index))
write.csv(df,'predictions.csv',row.names = FALSE) # format: index,prediction
```