Outlying Labs Analysis

S. Eanes

12/23/2020

Setup data

```r
coral <- read.csv("~/School/Fossil Coral/data/coral_3weighted.csv")

#omit data with no response
coral <- coral[!is.na(coral$U238),]


#select three largest species
genus_trim <- c("Acropora", "Porites")
#remove coral with age > 10
coral <- coral[which(coral$Genus %in% genus_trim),]  %>% dplyr::filter(Age < 10)
coral <- coral[(coral$Calcite <= 1 | is.na(coral$Calcite)),]

#clean up a nice dataframe
coral.df <- coral %>% mutate(Temperature = Temp) %>%
  data.frame()

coral.df$U238 <- coral.df$U238 * 0.421
nrow(coral.df)
```

```
## [1] 700
```

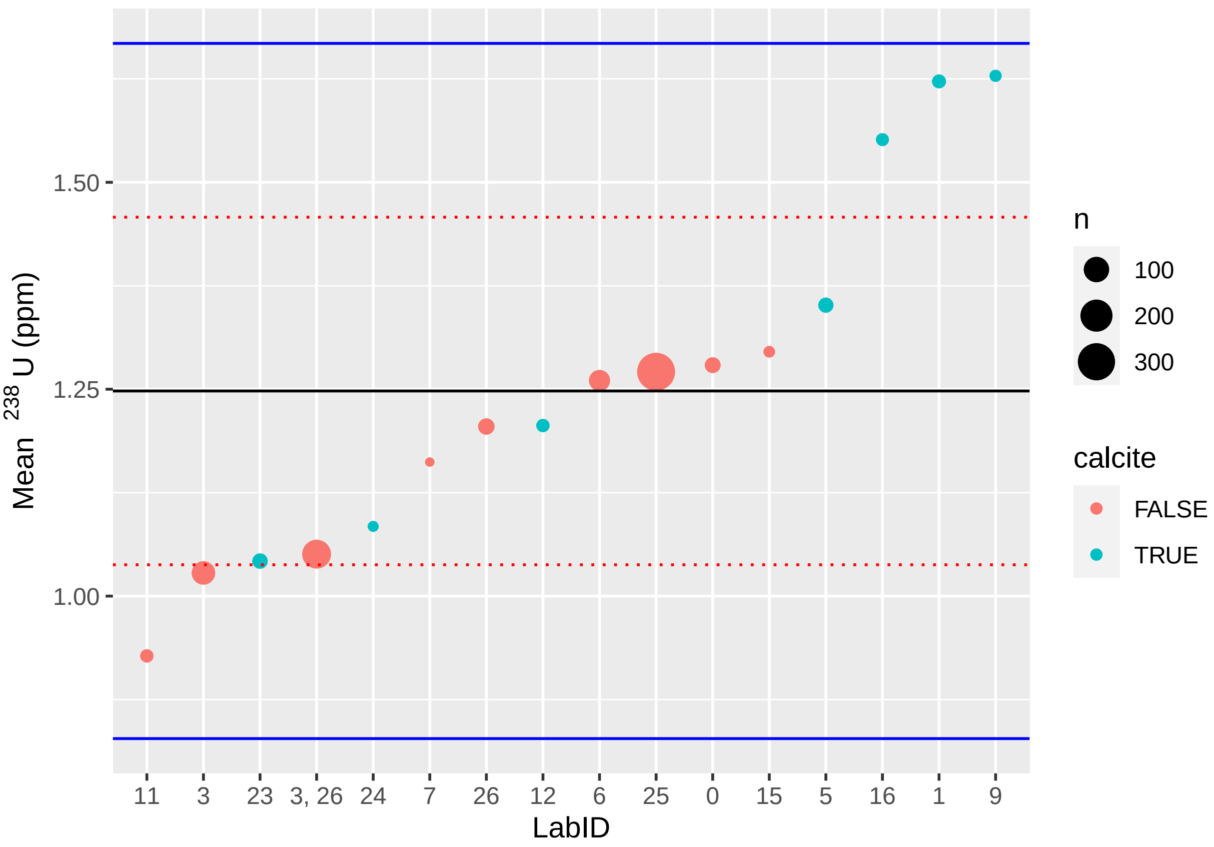
Outlying Labs Figure

Labs 1, 3, 9, 11, and 16 produced data that was more than one standard deviation from the mean.

```r
sumtib <- coral.df %>% dplyr::group_by(LabID) %>% dplyr::summarise(n = n(), mean_u238 = mean(U238),sd=s

sumtib$LabID <- factor(sumtib$LabID, levels = sumtib$LabID[order(sumtib$mean_u238)])

sumtib %>%
  ggplot(aes(x=LabID,y=mean_u238,size=n,color=calcite))+
  geom_point()+
  geom_hline(yintercept = mean(sumtib$mean_u238))+
  geom_hline(yintercept = mean(sumtib$mean_u238)-2*sd(sumtib$mean_u238),color="blue")+
  geom_hline(yintercept = mean(sumtib$mean_u238)+2*sd(sumtib$mean_u238),color="blue")+
  geom_hline(yintercept = mean(sumtib$mean_u238)-sd(sumtib$mean_u238),color="red",linetype="dotted")+
  geom_hline(yintercept = mean(sumtib$mean_u238)+sd(sumtib$mean_u238),color="red",linetype="dotted")+
  ylab(bquote('Mean  ' ^{"238"}~'U (ppm)'))
```



Modeling

We build a model on all the data except from those labs mentioned above, and then predict on the labs excluded to see if there data is abnormal.

```
exclude_labs <- c(1,3,9,11,16)
train_rows <- !(coral.df$LabID %in% exclude_labs)
train.dat <- coral.df[train_rows,]
test.dat <- coral.df[!train_rows,]
nrow(train.dat)+nrow(test.dat) == nrow(coral.df)
```

```
## [1] TRUE
```

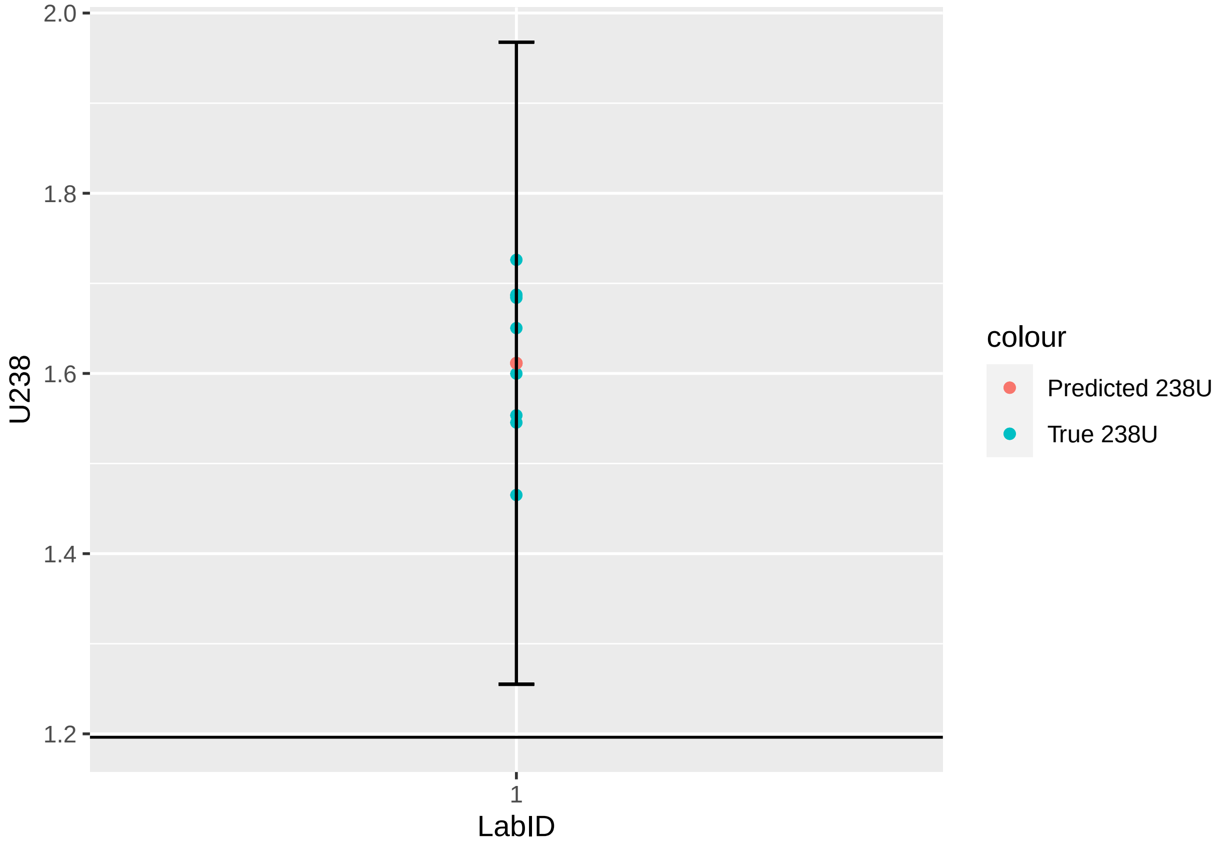
```
model <- lm(U238 ~ Temperature + Salinity + pH + OmegaA, data=coral.df)

prediction <- predict(model,test.dat,interval = "prediction",level = .95)
test.dat <- cbind(test.dat,prediction)
```

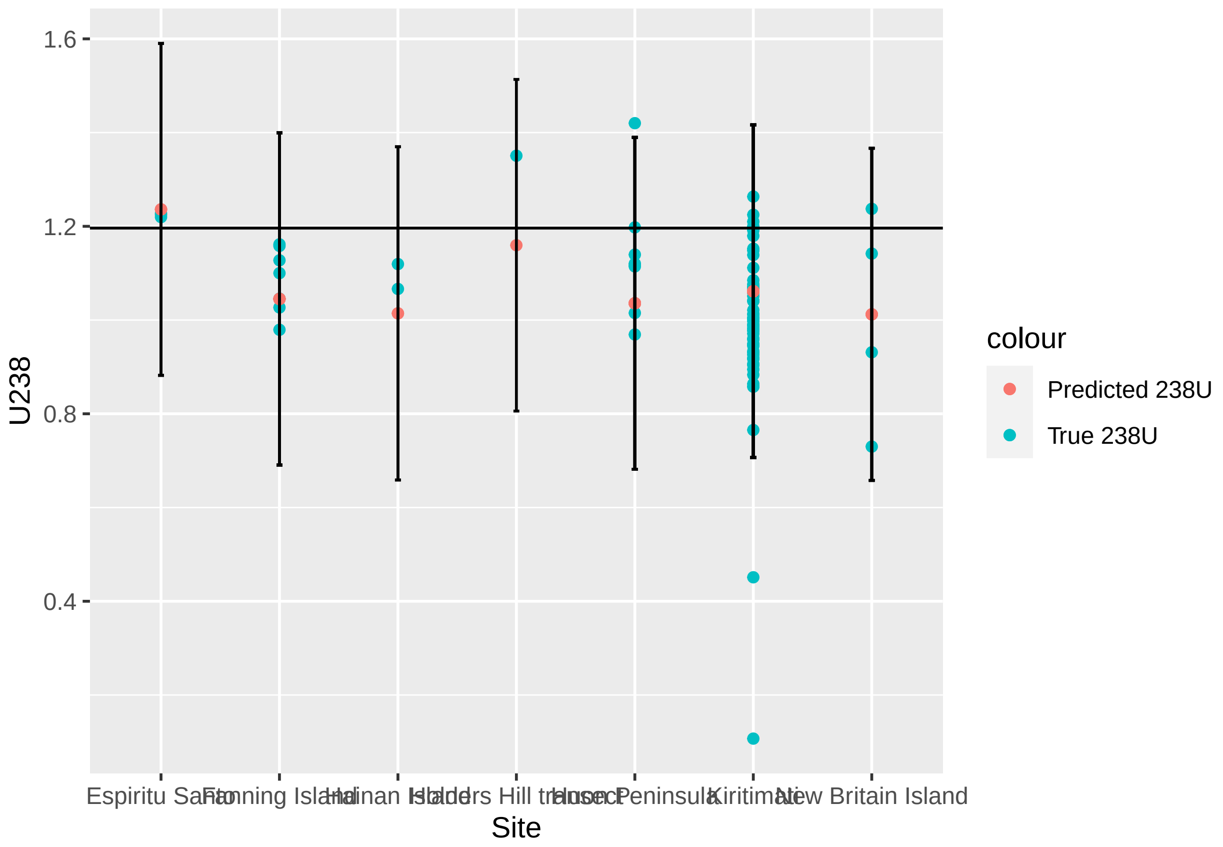
Graph It

Most to all observations for each location/site fall into the 95% prediction interval. Thus, when controlling for environmental factors, none of the sites observations are unusual, and all can be included in the analysis.

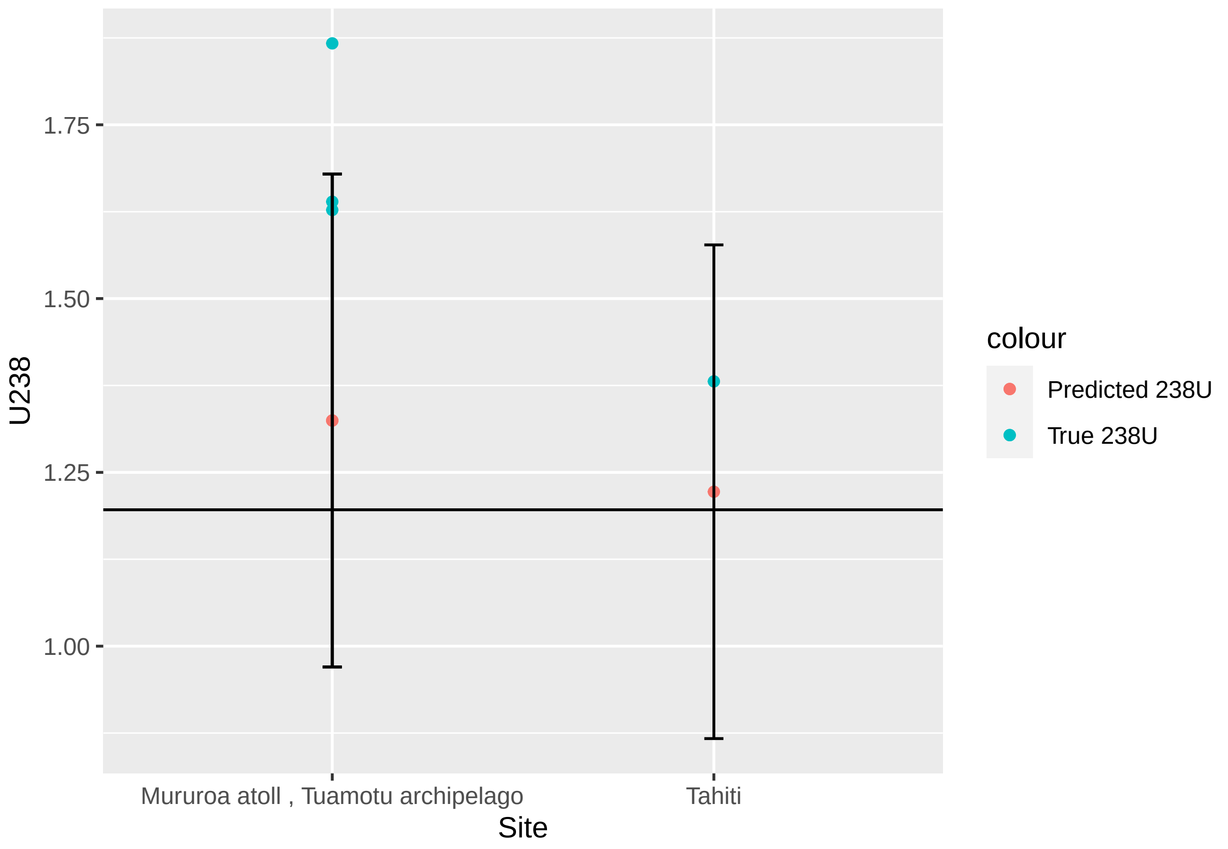
```
test.dat[test.dat$LabID==1,] %>%
  ggplot(aes(LabID))+
  geom_point(aes(y=U238, color="True 238U"))+
  geom_point(aes(y=fit, color="Predicted 238U"))+
  geom_errorbar(aes(ymin=lwr, ymax=upr), width=.05)+
  geom_hline(yintercept = mean(coral.df$U238))
```



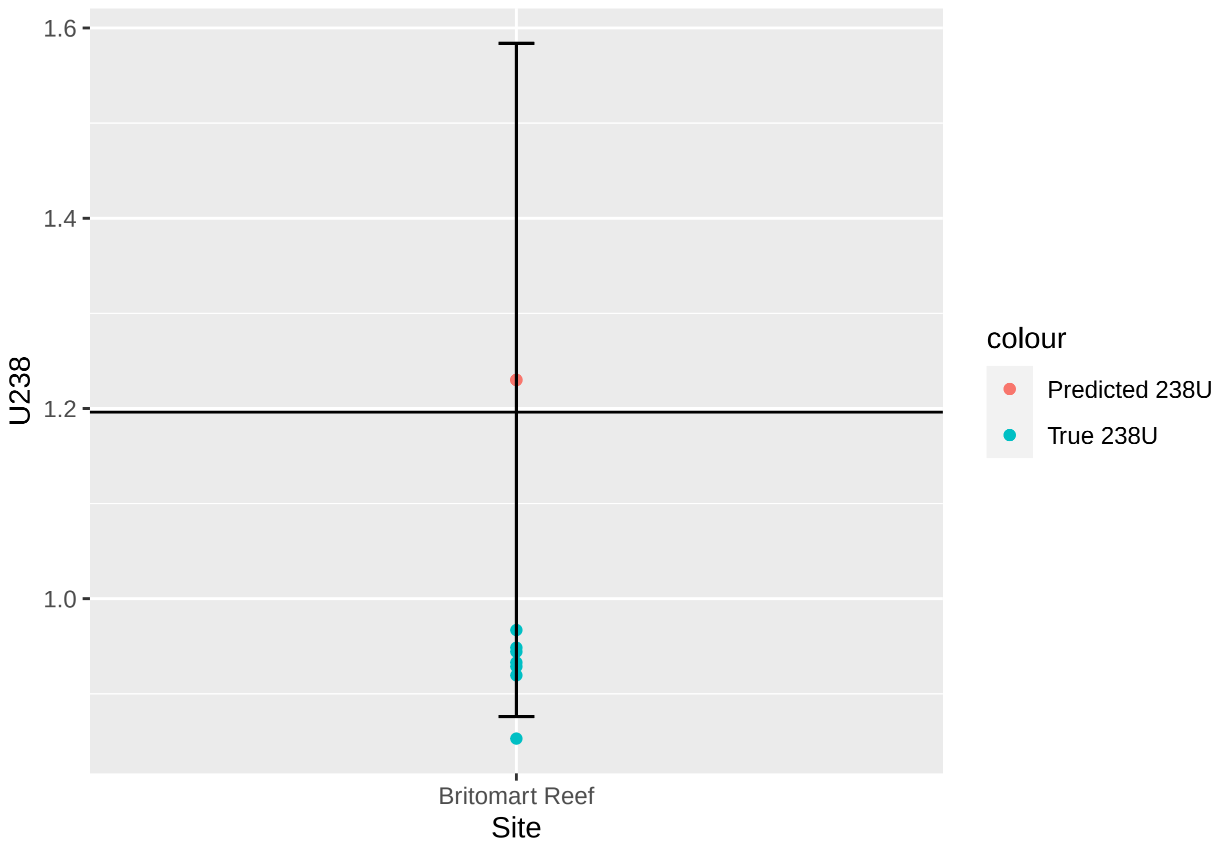
```
test.dat[test.dat$LabID==3,] %>%
  ggplot(aes(Site))+
  geom_point(aes(y=U238, color="True 238U"))+
  geom_point(aes(y=fit, color="Predicted 238U"))+
  geom_errorbar(aes(ymin=lwr, ymax=upr), width=.05)+
  geom_hline(yintercept = mean(coral.df$U238))
```



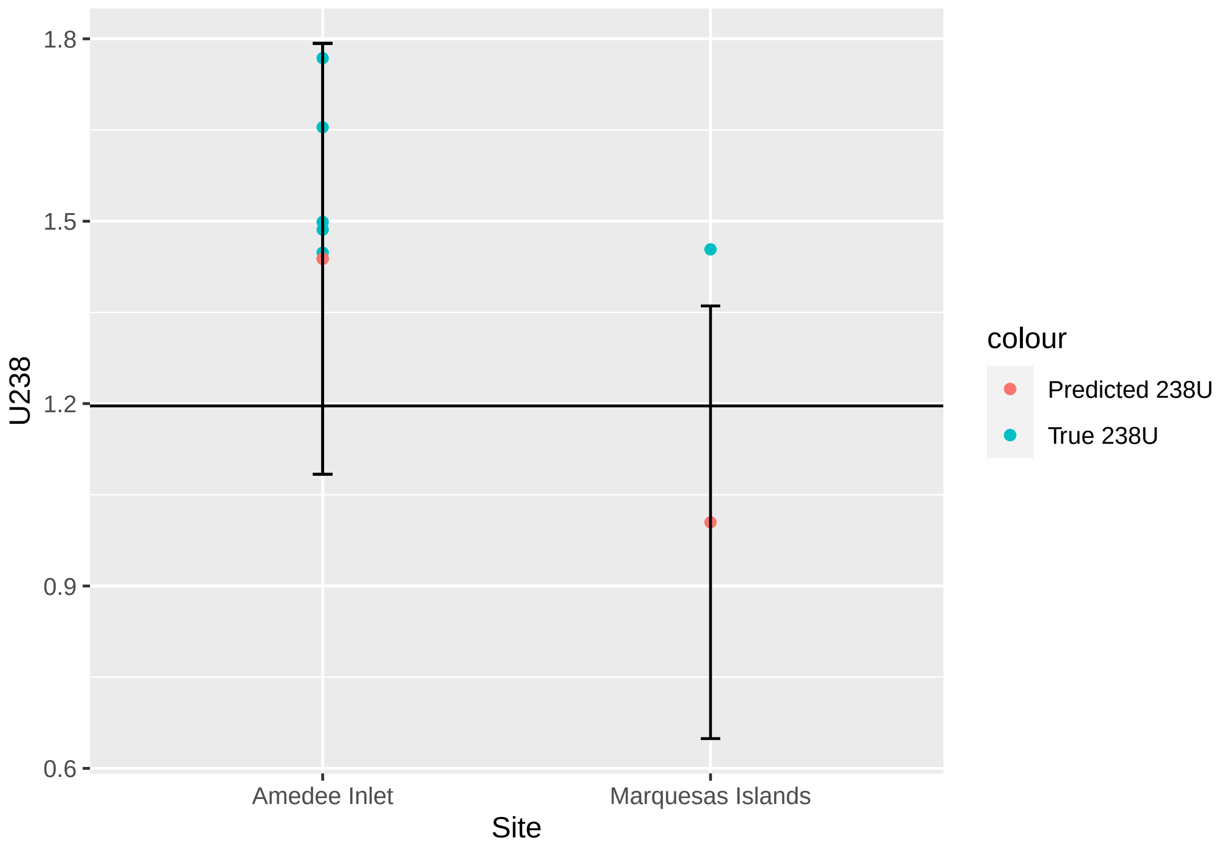
```r
test.dat[test.dat$LabID==9,] %>%
  ggplot(aes(Site))+
  geom_point(aes(y=U238, color="True 238U"))+
  geom_point(aes(y=fit, color="Predicted 238U"))+
  geom_errorbar(aes(ymin=lwr, ymax=upr), width=.05)+
  geom_hline(yintercept = mean(coral.df$U238))
```



```
test.dat[test.dat$LabID==11,] %>%
  ggplot(aes(Site))+
  geom_point(aes(y=U238, color="True 238U"))+
  geom_point(aes(y=fit, color="Predicted 238U"))+
  geom_errorbar(aes(ymin=lwr, ymax=upr), width=.05)+
  geom_hline(yintercept = mean(coral.df$U238))
```



```
test.dat[test.dat$LabID==16,] %>%
  ggplot(aes(Site))+
  geom_point(aes(y=U238, color="True 238U"))+
  geom_point(aes(y=fit, color="Predicted 238U"))+
  geom_errorbar(aes(ymin=lwr, ymax=upr), width=.05)+
  geom_hline(yintercept = mean(coral.df$U238))
```



U238
1.8
1.5
1.2
0.9
0.6
Amedee Inlet
Marquesas Islands
Site
colour
Predicted 238U
True 238U