1 Introduction

This document describes data and provides code to support “Estimating group size from acoustical footprint to improve Blainville’s beaked whale abundance estimation” by Patrícia A. Jorge, Tiago A. Marques, Helena Mouriño, Len Thomas, David J. Moretti, Karin Dolan, Diane Claridge, Charlotte Dunn and submitted to Applied Acoustics.

This data and code package includes this file with code itself and 4 data files, provided as txt files. These are

1. The modelling dataset - modeddata.txt
2. The density estimation dataset without false positives - dedata.txt
3. The full density estimation dataset including false positives - dedatauncut.txt
4. The bootstrap results - bootdata.txt

Below we provide code to read in the files and use them along the lines of what was done in the paper. The bootstrap results file is provided primarily to ensure that figure 2 in the paper can be easily recreated.

We begin by reading in the available datasets and describing the data in each dataset.

We also include a couple of images which, although would not fit nicely into the paper, were mentioned in discussions with the original reviewers of the manuscript and hence might be useful for those trying to understand the data in detail.

2 The modelling dataset

The model data, corresponding to the acoustic footprint of 51 groups with verified group size

```r
modeldata=read.table("modeddata.txt")
```

- gID - a group ID
- cs - the group size
- conf - the confidence in the group size (1-best; 2-medium;3-low)
- maxcount - the maximum number of clicks in a single hydrophone over the hydrophones the group was detected
- meancount - the mean number of clicks detected per hydrophone
- nhyd - the number of hydrophones the group as detected on
- cdur - the duration of the vocal period
- nclicks - the total number of clicks detected
- crate - the total detected click rate - nclicks/cdur
- wisk - whether a Wiskey hydrophone was involved in the detection
- direction - whether a bidirectional phone was involved in the detection
This was the final model used for inference

\[
\text{mp7} \leftarrow \text{vglm}(\text{cs} \sim \text{crate}, \text{family} = \text{pospoisson}, \text{data} = \text{modeldata})
\]

Below we also present two additional very close competing models we chose to ignore given very close AIC values, the variables not being significant (at the 5% significance level) and density estimation results being insensitive to model choice, hence opting for the simpler model with just click rate

\[
\text{library(VGAM)}
\]

\[
\text{mpT1} \leftarrow \text{vglm}(\text{cs} \sim \text{crate} + \text{nhyd} + \text{cdur}, \text{family} = \text{pospoisson}, \text{data} = \text{modeldata})
\]

\[
\text{mpT2} \leftarrow \text{vglm}(\text{cs} \sim \text{crate} + \text{nhyd}, \text{family} = \text{pospoisson}, \text{data} = \text{modeldata})
\]

We output here the values used in table 1 in the paper

% latex table generated in R 3.4.3 by xtable 1.8-2 package % Tue Sep 25 19:38:16 2018

<table>
<thead>
<tr>
<th>Model name</th>
<th>Variables</th>
<th>Coefficient</th>
<th>P-value</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>click duration</td>
<td>0.0103</td>
<td>0.0113</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>M3 number of hydrophones</td>
<td>-0.0926</td>
<td>0.0713</td>
<td>150.49</td>
</tr>
<tr>
<td>3</td>
<td>click rate</td>
<td>0.0037</td>
<td>0.1374</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>M2 number of hydrophones</td>
<td>-0.0557</td>
<td>0.0205</td>
<td>150.71</td>
</tr>
<tr>
<td>5</td>
<td>click rate</td>
<td>0.0033</td>
<td>0.1931</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>M1 click rate</td>
<td>0.0018</td>
<td>0.0298</td>
<td>150.55</td>
</tr>
</tbody>
</table>

3 The density estimation dataset

This is the dataset with the acoustical footprint of each group for which the group size was not known and hence had to be estimated, so that then density estimates could be obtained via the proposed modified dive counting procedure.

Note that false positives have already been excluded as described in the paper. However, the unfiltered data is also provided as a fourth datafile, “dedatauncut.txt”, in case one might wish to do another kind of false positive exclusion procedure.

\[\text{dedata=read.table("dedata.txt",sep="\t")}\]

The relevant available variables include all those available in the modelling dataset, with the difference that the variable cs represents the estimated group size using the best model, i.e. a truncated poisson GLM with click rate as a covariate. Then the columns named 1 to 999 actually represent the estimated group size at each of the 999 bootstrap resamples.

4 The bootstrap results

This data file includes the estimated density per day for each of the bootstrap resamples. Therefore it includes 109 rows (one row per day) and 1000 columns (999 resamples)

\[\text{bootdata=read.table(file="bootdata.txt")}\]

These results were obtained by running the following bootstrap code. The code is not evaluated here but provided for completeness.

\[\text{set.seed(12397)}\]
\[\text{B } \leftarrow \text{999}\]
res <- numeric(B)
grupospredboot <- matrix(NA, nrow=nrow(dados.filtered), ncol=999)
tableboot <- dados.filtered
tab99 <- matrix(nrow=8271, ncol=999)
tableboot <- cbind(tableboot, tab99)
for (i in 1:B){
  index = sample(1:51, 51, replace=TRUE)
dados4boot = d4reg[index,]
mp7boot = vglm(formula = cs ~ crate, family = pospoisson, data = dados4boot)
preds = predict(mp7boot, dados.filtered, type="response")
tableboot[,i+19] = preds
}
# Obtaining the cs mean for each day (mean within bootstraps)
tablebootday <- matrix(nrow=109, ncol=999)
B <- 999
for (i in 20:(20+B-1)){
  # Cluster size mean for each day
  tablebootday[,i-19] = tapply(X=tableboot[,i], INDEX=tableboot$day, FUN=mean)
}
# Density bootstrap
rdboot <- rnorm(999, 0.36, 0.04)
tableboottot <- matrix(nrow=109, ncol=999)
tablebootfix <- as.data.frame(tablebootday)
ttime <- dados.eachday$ttime
ngroups <- dados.eachday$groups
tablebootday <- rbind(tablebootday, rdboot)
# Function which calculates density for each cell
fdens <- function(x, y) { ((ngroups[y]) * tablebootday[y,][x]) / ((tablebootday[110,][x]) * ttime[y]) / 1291 * 1000 }
for (i in 1:999){
  for (j in 1:109){
    tableboottot[j,][i] = fdens(i, j) }
  }
tableboottot = as.data.frame(tableboottot, header=F)

5 Creating the paper figures

Here we present the required code to recreate the paper figures

5.1 Figure 1

Note that this figure is not outputed in this document given an apparent incompatibility between ggplot2 and knitr. This was unexpected as these were supposed to know each other very well!

library(boot)
library(ggplot2)
newdata <- expand.grid(modeldata$crate)
newdata$yhat <- predict(mp7, modedata, type = "response")
names(newdata)[1]<-"crate"
## function to return predicted values

```r
define (function (data, i, newdata) {
  require(VGAM)
  m <- vglm(formula = cs ~ crate, family = pospoisson(),
             data = data[i, ], coefstart = c(0.447, 0.002))
  predict (m, newdata, type = "response")
})
```

## set seed and run bootstrap with 999 draws

```r
set.seed(10)
respred <- boot(modeldata, fpred, R = 999, newdata = newdata,
                parallel = "snow", ncpus = 4)
```

## get the bootstrapped percentile CIs

```r
yhat <- t (sapply (1:nrow (newdata), function(i) {
  out <- boot.ci (respred, index = i, type = c("perc"))
  with(out, c(Est = t0, pLL = percent[4], pUL = percent[5]))
}))
```

## merge CIs with predicted values

```r
newdata <- cbind (newdata, yhat)
obsval = modeldata$cs
newdata <- cbind (newdata, obsval)
```

## graph with CIs

```r
ggplot (newdata, aes (x = crate, y = obsval)) +
  geom_ribbon (aes (ymin = pLL, ymax = pUL), alpha = .25) +
  geom_point () +
  geom_line ()
```

## graph used in the Paper with fit line

```r
ggplot (newdata, aes (x = crate, y = obsval)) +
  geom_ribbon (aes (ymin = pLL, ymax = pUL), alpha = .25) +
  geom_point () +
  geom_line (data = newdata, aes (x = crate, y = yhat), color = 'red') +
  xlab("Click rate") +
  ylab("Observed Group size") +
  ggtitle("Click Rate vs Observed Group size, with best model and 95% CI") +
  theme (plot.title = element_text (hjust = 0.5))
```

### 5.2 Figure 2

# make a general plot

#(1) 61 days from the 28th of April to the 27th of June,
#(2) 18 days from the 20th of October to the 6th of November, and
#(3) 30 days from the 2nd to the 31st of December.

# January 1st is Julian day 0

```r
tmp1 <- as.POSIXlt ("28April", format = "%d%b%y")
```
tmp1 <- as.POSIXlt("27Jun11", format = "%d%b%y")
tmp2 <- tmp1$yday + 1
tmp3 <- as.POSIXlt("20Oct11", format = "%d%b%y")
tmp4 <- tmp3$yday + 1
tmp5 <- as.POSIXlt("06Nov11", format = "%d%b%y")
tmp6 <- tmp5$yday + 1
tmp7 <- as.POSIXlt("02Dec11", format = "%d%b%y")
tmp8 <- tmp7$yday + 1
tmp9 <- as.POSIXlt("31Dec11", format = "%d%b%y")
tmp10 <- tmp9$yday + 1

daysTAM <- c(tmp1:tmp2, tmp3:tmp4, tmp5:tmp6)

par(mfrow=c(1,1),mar=c(4,4,0.2,0.2))

plot(daysTAM, apply(bootdata,1,mean), ylab="Density (animals/1000 Km2)",
     xlab="Julian day (in 2011)", ylim=c(0,45))

limsup <- apply(bootdata,1,quantile, probs=0.975)
liminf <- apply(bootdata,1,quantile, probs=0.025)

#points(daysTAM, liminf)
#points(daysTAM, limsup)
for(i in 1:length(daysTAM)){
  segments(x0=daysTAM[i], y0=limsup[i], x1 = daysTAM[i], y1 =liminf[i] )
}
6 Calculating density and group size per time period

Just as examples, below we show how one could recreate the values for the estimates for density and mean group size for each of the 3 time periods involved

```r
# Get density and mean group size per survey period (and corresponding variances)

# DENSITY

given days 1 to 109 (i.e. not Julian day but day in sample)
Day 1 to 61 - period 1
Day 62 to 79 - period 2
Day 80 to 109 - period 3

# get the right per period average density variance estimates
# 1st period
Dsperiod1=bootdata[1:61,]
meanDperiod1=apply(Dsperiod1,2,mean)
quantile(meanDperiod1,c(0.025,0.975))
## 2.5%  97.5%
## 13.61399 21.87861
# 2nd period
Dsperiod2=bootdata[62:79,]
meanDperiod2=apply(Dsperiod2,2,mean)
quantile(meanDperiod2,c(0.025,0.975))
## 2.5%  97.5%
## 13.80851 22.37260
# 3rd period
Dsperiod3=bootdata[80:109,]
meanDperiod3=apply(Dsperiod3,2,mean)
quantile(meanDperiod3,c(0.025,0.975))
## 2.5%  97.5%
## 13.19567 21.22106

# Get density and mean group size per survey period (and corresponding variances)

# group size

# From the paper, we have that
detections 1:4562 are period 1
detections 4563:4563+1439-1 are period 2
detections 4563+1439:4563+1439+2270-1 are from period 3
gs1end=4562
gs2end=gs1end+1439
gs3end=gs2end+2270

# Now we just need to get the right averages for each period
```
#1st period
\texttt{mean(dedata$cs[1:gs1end])}

## [1] 2.355688

#2nd period
\texttt{mean(dedata$cs[(gs1end+1):gs2end])}

## [1] 2.29761

#3rd period
\texttt{mean(dedata$cs[(gs2end+1):gs3end])}

## [1] 2.331028

#now get average group sizes variances

#remove the first few columns, which are just
#the predictors and other meta.data
tablebootTAM=dedata[,20:1018]

#1st period
GSsperiod1=tablebootTAM[1:gs1end,]
meanGSsperiod1=apply(GSsperiod1,2,mean)
quantile(meanGSsperiod1,c(0.025,0.975))

## 2.5% 97.5%
## 2.149069 2.596826

#2nd period
GSsperiod2=tablebootTAM[(gs1end+1):gs2end,]
meanGSsperiod2=apply(GSsperiod2,2,mean)
quantile(meanGSsperiod2,c(0.025,0.975))

## 2.5% 97.5%
## 2.080316 2.557527

#3rd period
GSsperiod3=tablebootTAM[(gs2end+1):gs3end,]
meanGSsperiod3=apply(GSsperiod3,2,mean)
quantile(meanGSsperiod3,c(0.025,0.975))

## 2.5% 97.5%
## 2.118437 2.579719

7 A couple additional figures

7.1 The distribution of click rate in the density estimation dataset

A reviewer noted that “Given a huge sample size, the sole predictive variable (detected click rate) might be expected to have multiple modes corresponding to the most frequency group sizes (1, 2 and possibly 3). It would be good to plot a histogram of this variable for the second, larger data set.”.

Indeed this is an interesting remark, and the reason we have not presented the figure in the paper is because it turned out that the expected pattern was not verified. We did add to the discussion stating that “Given that click rate was the single predictor in the model used for predicting group size, click rate might present a multi-modal distribution in the density estimation data set, where each group size would be visible. However,
we plotted the variable and unfortunately the multi-mode pattern was not present, instead a unimodal right heavy tailed distribution appeared, so we do not present the plot. This implies that while click rate can be used to predict group size, and might work quite well to predict average group sizes over say a day, it might not be enough to accurately predict each group’s group size.”

The relevant figure is presented here

```r
hist(dedata$crate,breaks=seq(0,850,by=20),
xlab="Click rate (in clicks per minute)",freq=FALSE,main="")
```

7.2 The distribution of vocal periods in the density estimation data set

The same reviewer noted that “Please provide descriptive stats of this distribution (detected dive duration). Did you find unusual long dives (dive vocal phases) that might correspond to diving groups partially overlapping in space? This is interesting for your results and also from a foraging ecology perspective.”

Again, while this figure would be a bit out of scope in a paper about a method regarding density estimation, we felt that it is a very interesting image and hence added it here:

```r
hist(dedata$cdur,breaks=seq(0,120,by=5),xlab="Click duration",freq=FALSE,main="")
```
It is certainly possible that the longer durations do not correspond to a group of beaked whales, either because:

- Two or more different groups were considered a single group
- It was not a group of beaked whales at all