Analysis of aerial and shipboard surveys of fin whale data with corrections for $g(0)$ and availability

David L Miller, David A Fifield, Ewan Wakefield and Douglas B Sigourney

Introduction

These data consist of observations of fin whales as part of NOAA's Atlantic Marine Assessment Program for Protected Species.

The analysis here is based on that in *Developing and assessing a density surface model in a Bayesian hierarchical framework with a focus on uncertainty: insights from simulations and an application to fin whales (Balaenoptera physalus)* available at PeerJ.

Elements of this analysis

- 2 detection functions
  - Aerial survey: multiple covariate distance sampling (MCDS) with fixed $g(0)$ correction from literature
  - Shipboard survey: double observer survey (mark-recapture distance sampling; MRDS)
- Density surface model combining these two sources
- Observations of fin, sei and fin/sei uncertain species identifications were used to fit the detection function, but only certain fin detections were then used in the density surface model.
- Availability correction for aerial surveys (0.374, CV=0.34)
- Comparison of using average group size versus observed group size

The analysis of these data in Sigourney et al., uses a fully Bayesian model, incorporating uncertainty in group size and availability. Here we do not address these sources of uncertainty. The most appropriate comparison is to results provided in their Table 3 where availability was treated as constant (since uncertainty in mean group size contributed a negligible amount to uncertainty). This estimate is 4345 fin whales with a CV=0.21 (further comparisons are provided at the end of this document).

Preliminaries

Data has already been processed and is stored in `findata.RData`.

```
# modelling
library(patchwork)
library(mrds)
```

```
## This is mrds 2.2.4
## Built: R 4.0.2; ; 2020-11-30 17:31:53 UTC; unix
```

```
library(dsm)
```

```
## Loading required package: mgcv
```

```
## Loading required package: nlme
```

```
## This is mgcv 1.8-34. For overview type 'help("mgcv-package")'.
```

```
## Loading required package: numDeriv
```

```
## This is dsm 2.3.1
## Built: R 4.0.2; ; 2021-03-23 21:16:18 UTC; unix
```

```
# for plotting predictions
library(ggplot2)
library(rnaturalearth)
library(rnaturalearthdata)
library(sf)
```

```
## Linking to GEOS 3.8.1, GDAL 3.1.4, PROJ 6.3.1
```

```
# data, pre-processed
load("findata.RData")
```

Note that `dsm` package version 2.3.1 is required for multiple detection function functionality and `mgcv` version 1.8-34 is required for Metropolis-Hastings sampling.

Model fitting

We now fit the detection functions, followed by the density surface model. Note that we'd normally do full model selection at each stage, but since we are simply duplicating the analysis from the paper we just use the models selected there.

Data used to fit the detection functions (`ship_detections` and `plane_detections`) includes detections of fin whales, sei whales and detections that could only be classified to "fin or sei whale". This improves the fit of the detection functions as more detections are included. For fitting the density surface model, we only want to include detections of fin whales (since that is the species we want an abundance estimate for). We can exclude the non-fin whale observations by including only the `object` IDs that we want in the observation `data.frame fin_obs`. See the data setup script (available at `https://github.com/densitymodelling/nefsc_fin_mrds_dsm`) for how this was done.

We first setup the truncations for each detection function model:

```
w_ship <- 6
w_plane <- 0.9
```

and then need to set up a vector of availabilities (taken from supplementary code from the paper) for the DSM. This just needs to be the same length as the segment data (if we have a count model). We have a slight misuse of arguments here as we include the $g(0)$ estimate for the plane (as computed from a different analysis) here as fixed to match what's in the paper. The estimate of $g(0) = 0.67$ (CV $= 0.16$) is from Palka et al. (2017).

```
a_plane <- 0.374*0.67
a_ship <- 1

avail <- c(a_ship, a_plane)[fin_segs$ddfobj]
```

Ship surveys - MRDS

We can fit an independent observer model to the double observer data for the ship.

```
Ship_mrds <- ddf(dsmodel = ~mcds(key = "hr", formula = ~beaufort + SUBJ_WAVG),
                 mrmodel = ~glm(~distance),
                 data = ship_detections, method = "io",
                 meta.data = list(width = w_ship))
summary(Ship_mrds)
```

```
##
## Summary for io.fi object
## Number of observations   :  144
## Number seen by primary   :  102
## Number seen by secondary :  75
## Number seen by both      :  33
## AIC                      :  312.0793
##
##
## Conditional detection function parameters:
##               estimate        se
## (Intercept) -0.3242831 0.2894210
## distance    -0.1275751 0.1441381
##
##                          Estimate         SE        CV
## Average primary p(0)   0.4196323 0.07048589 0.1679706
## Average secondary p(0) 0.4196323 0.07048589 0.1679706
## Average combined p(0)  0.6631733 0.08181548 0.1233697
##
##
## Summary for ds object
## Number of observations :  144
## Distance range         :  0  -  6
## AIC                    :  406.9032
##
## Detection function:
##  Hazard-rate key function
##
## Detection function parameters
## Scale coefficient(s):
##               estimate        se
## (Intercept) -3.8983160 1.7744903
## beaufort     0.1286412 0.2003129
## SUBJ_WAVG    1.4114861 0.4864454
##
## Shape coefficient(s):
##              estimate        se
## (Intercept) 0.5165719 0.1541891
##
##            Estimate         SE        CV
## Average p 0.2952727 0.04569428 0.1547528
##
##
## Summary for io object
## Total AIC value :  718.9825
##
##                         Estimate          SE        CV
## Average p              0.1958169   0.0388336 0.1983158
## N in covered region 735.3806999 156.3960939 0.2126736
```

Aerial surveys - MCDS

The aerial survey is a regular multiple covariate detection function:

```
Plane_ds <- ddf(dsmodel = ~mcds(key = "hr", formula = ~beaufort),
                data = plane_detections, meta.data = list(width = w_plane))
summary(Plane_ds)
```

```
##
## Summary for ds object
## Number of observations :  36
## Distance range         :  0  -  0.9
## AIC                    :  -32.54007
##
## Detection function:
##  Hazard-rate key function
##
## Detection function parameters
## Scale coefficient(s):
##               estimate        se
## (Intercept)  0.4500577 0.8426747
## beaufort    -0.5460440 0.2778157
##
## Shape coefficient(s):
##             estimate        se
## (Intercept) 1.103263 0.3227304
##
##                         Estimate          SE        CV
## Average p             0.3980921  0.07995532 0.2008463
## N in covered region 90.4313451 21.96646110 0.2429076
```

Density surface model

We can then fit the spatial model using the below code. Note a list of detection function objects is provided to `ddf.obj`, their ordering matters and corresponds to the `ddfobj` column in the segment data (`fin_segs`). The availability correction is supplied via the `availability=` argument.

```
fin_b <- dsm(count ~ s(DIST125, bs="ts", k=5) +
                     s(DEPTH, bs="ts", k=5) +
                     s(DIST2SHORE, bs="ts", k=5) +
                     s(SST, bs="ts", k=5),
             ddf.obj=list(Ship_mrds, Plane_ds),
             family=tw(),
             availability=avail,
             segment.data=fin_segs,
             observation.data=fin_obs)
summary(fin_b)
```

```
##
## Family: Tweedie(p=1.274)
## Link function: log
##
## Formula:
## count ~ s(DIST125, bs = "ts", k = 5) + s(DEPTH, bs = "ts", k = 5) +
##     s(DIST2SHORE, bs = "ts", k = 5) + s(SST, bs = "ts", k = 5) +
##     offset(off.set)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  -6.1033     0.2689  -22.69   <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(DIST125)    1.0191      4 7.458  < 2e-16 ***
## s(DEPTH)      0.9667      4 3.529 8.96e-05 ***
## s(DIST2SHORE) 1.0576      4 6.987 5.20e-07 ***
## s(SST)        0.7909      4 0.871    0.033 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.102   Deviance explained = 35.3%
## -REML = 439.27  Scale est. = 5.1877    n = 3794
```

We can review the summary output above and verify that we have given the smooth terms sufficient degrees of freedom.

We can then propagate the variance from the detection functions through to predictions over the study area from the DSM:

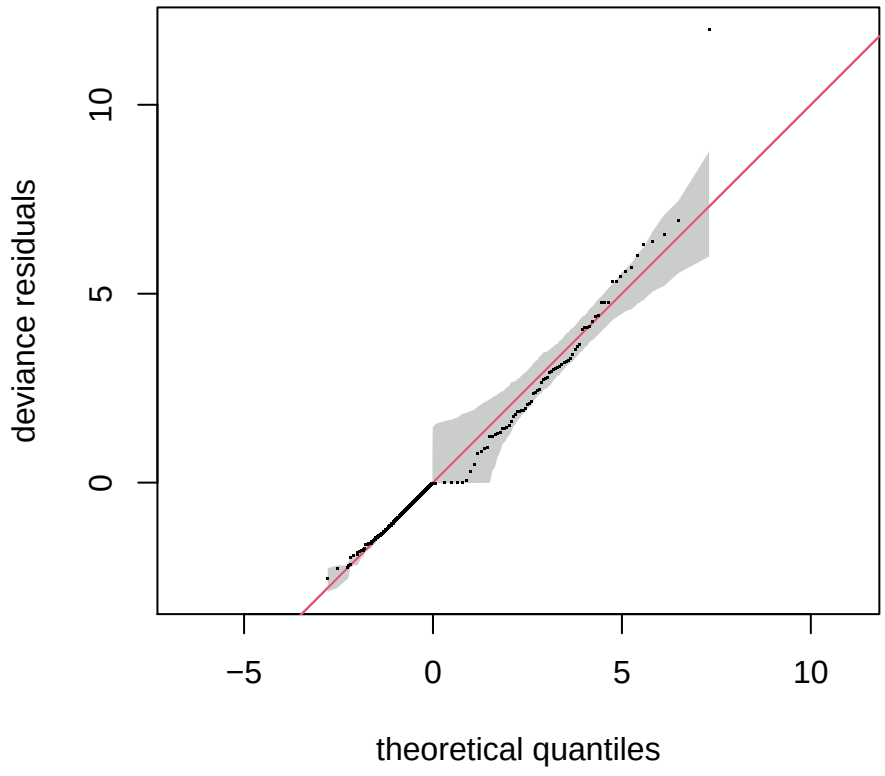
```
vp <- dsm_varprop(fin_b, predgrid)
vp
```

```
## Summary of uncertainty in a density surface model calculated
##  by variance propagation.
##
## Probability of detection in fitted model and variance model
## Detection function 1
##   SUBJ_WAVG beaufort Fitted.model Fitted.model.se Refitted.model
## 5     2.000   2.6070   0.10191161      0.03527256      0.1235225
## 7     2.018   1.1536   0.08793953      0.04869427      0.1312472
## 2     2.279   2.6070   0.14479315      0.03573346      0.1600423
## 6     2.297   4.4400   0.18129926      0.05037092      0.1541893
## 4     2.653   1.1536   0.19387079      0.06055648      0.2320331
## 3     2.718   4.4400   0.29204300      0.09528551      0.2241965
## 9     2.754   4.4400   0.30330924      0.10029464      0.2311932
## 8     2.866   2.6070   0.28618688      0.05184305      0.2665722
## 1     3.000   1.1536   0.28665693      0.06691199      0.3078034
## Detection function 2
##   beaufort Fitted.model Fitted.model.se Refitted.model
## 1  1.86475    0.7334362      0.19007870      0.7332436
## 2  3.15350    0.4062064      0.06996712      0.4351019
## 3  4.58500    0.1912136      0.08636975      0.2247847
##
## Approximate asymptotic confidence interval:
##     2.5%     Mean    97.5%
## 2767.085 3935.350 5596.857
## (Using log-Normal approximation)
##
## Detection function CV          : 0.1983, 0.2008
##
## Point estimate                 : 3935.35
## Standard error                 : 712.93
## Coefficient of variation       : 0.1812
```

The `vp$refit` object is now the model we will use for predictions and further checking as this has the refitted model following the methods of Bravington et al. (2021).

We can also look at a quantile-quantile plot to verify the response distribution is reasonable (grey reference bands allow us to assess the degree of departure from their ideal values):

```
qq.gam(vp$refit, rep=200, asp=1)
```



We can compare observed and expected number of groups by Beaufort chunk:

```
obs_exp(vp$refit, "beaufort", 0:5)
```

```
##              (0,1]    (1,2]    (2,3]    (3,4]    (4,5]       <NA>
## Observed  2.00000 49.00000 97.00000 42.00000 16.00000 0.00000000
## Expected 10.10313 35.19818 97.59719 46.50723 15.02086 0.05461915
```

We can use the `predict` function to get the estimate of abundance (using the model with the variance propagated using the correct component of `dsm_varprop`):

```
# add extra columns to predgrid to make predict work
# the variance propagation procedure includes a new random effect covariate
# called "XX", we need this to predict (though we ignore its effect by setting
# the value to zero)
predgrid$XX <- matrix(0, nrow(predgrid), ncol(vp$refit$data$XX))
predN <- predict(vp$refit, newdata=predgrid, off.set=predgrid$off.set)
```

We can now fit same model but using observed rather than estimated group sizes

The only difference between this model and `dsm_avg_group` is now we don't set `group=TRUE`. We can see that we get pretty similar results:

But we now don't need to multiply by the average group size:

Rather than use the analytic approach, it might be preferable to use a Metropolis-Hastings sampler to get a posterior sample for the model. We need to use tools from `https://github.com/dill/GAMsampling` to ensure that the sampling works for the variance-propagated DSM.

```
# load additional code
source("likelihood_tools.R")
source("gam.mh_fix.R")
source("ttools.R")
# do the sampling
bs <- gam.mh(vp$refit, burn=2000, thin=10)
```

Now we need to construct a prediction matrix:

```
# create the matrix that maps model coefficients to the linear predictor
Xp <- predict(vp$refit, predgrid, type="lpmatrix")
# generate predictions
preds_mh <- predgrid$off.set * exp(Xp %*% t(bs$bs))
```

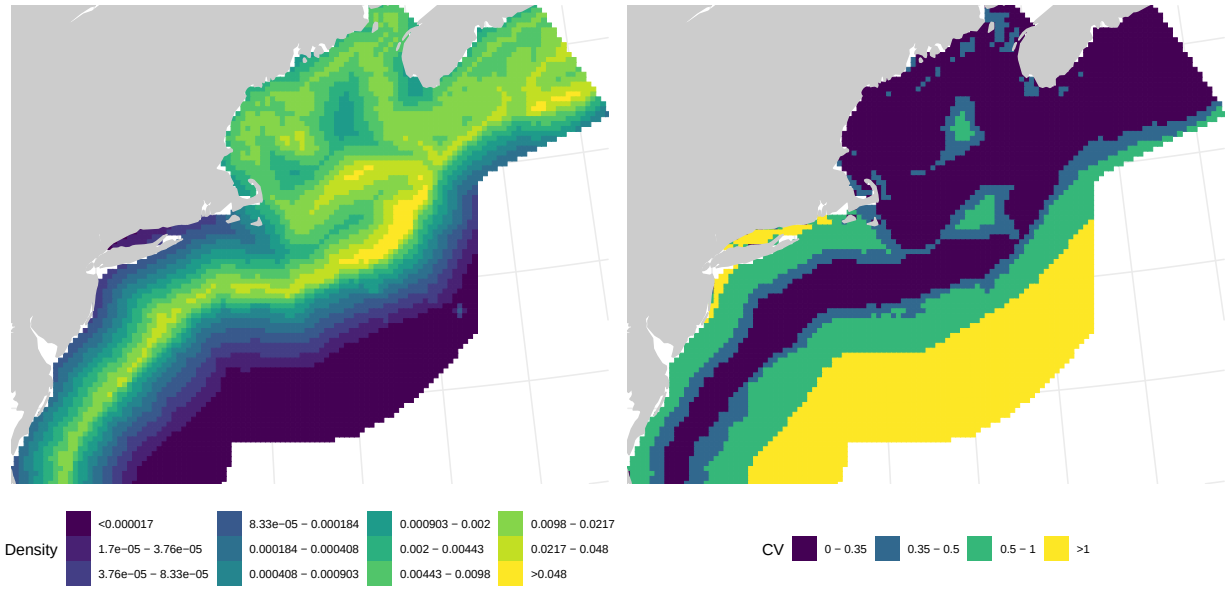
Now calculate our posterior statistics per grid cell:

```
# mean density estimate per cell from simulation
# results as density in animals/km^2
predgrid$Density <- rowMeans(preds_mh/predgrid$off.set)
# per cell CV
predgrid$CV <- apply(preds_mh/10^2, 1, sd)/predgrid$Density
```

Finally, plotting these:

```
# plot theme
this_theme <- theme(legend.position="bottom",
                    axis.title.y=element_blank(),
                    axis.text.y=element_blank(),
                    axis.title.x=element_blank(),
                    axis.text.x=element_blank(),
                    legend.text=element_text(size=8))


# get scale for predictions as in paper, thanks to Beth Josephson for this code
mind <- 0.000017
maxd <- 0.048
# geometrical breaks
n <- 10
k <- (maxd/mind)^(1/n)
brks <- c(0, mind, (mind*(k^seq(n))), max(predgrid$Density))
predgrid$pred_d <- cut(predgrid$Density, brks)
# nicer break labels (fiddly!)
levels(predgrid$pred_d) <- sub(",", " - ", levels(predgrid$pred_d))
levels(predgrid$pred_d) <- sub("\\(", "", levels(predgrid$pred_d))
levels(predgrid$pred_d) <- sub("]", "", levels(predgrid$pred_d))
levels(predgrid$pred_d)[1] <- "<0.000017"
levels(predgrid$pred_d)[length(levels(predgrid$pred_d))] <- ">0.048"
```

```r
levels(predgrid$pred_d) <- unlist(lapply(levels(predgrid$pred_d), function(x){
  if(grepl("e", x)){
    x <- as.character(as.numeric(strsplit(x, " - ")[[1]]))
    x <- paste0(x[1], " - ", x[2])
  }
  x
}))

# and the breaks for the CV (as in paper)
cvbrks <- c(0, 0.35, .5, 1, ceiling(max(predgrid$CV)))
predgrid$CV_d <- cut(predgrid$CV, cvbrks)
# nicer break labels
levels(predgrid$CV_d) <- sub(",", " - ", levels(predgrid$CV_d))
levels(predgrid$CV_d) <- sub("\\(", "", levels(predgrid$CV_d))
levels(predgrid$CV_d) <- sub("]", "", levels(predgrid$CV_d))
levels(predgrid$CV_d)[4] <- ">1"

# land outline
na <- ne_countries(continent="North America", returnclass="sf", scale=50)
na <- st_transform(na, "+proj=omerc +lat_0=35 +lonc=-75 +alpha=40 +k=0.9996 +x_0=0 +y_0=0 +gamma=40 +da

predplot <- ggplot(predgrid) +
  geom_tile(aes(x=x, y=y, fill=pred_d, colour=pred_d, width=10000, height=10000)) +
  geom_sf(data=na, colour="grey80", fill="grey80")+
  theme_minimal() +
  scale_fill_viridis_d() +
  scale_colour_viridis_d(guide=FALSE) +
  labs(fill="Density") +
  this_theme +
  coord_sf(xlim=range(predgrid$x), ylim=range(predgrid$y), expand=FALSE)

CVplot <- ggplot(predgrid) +
  geom_tile(aes(x=x, y=y, fill=CV_d, colour=CV_d, width=10000, height=10000)) +
  geom_sf(data=na, colour="grey80", fill="grey80")+
  theme_minimal() +
  scale_fill_viridis_d() +
  scale_colour_viridis_d(guide=FALSE) +
  labs(fill="CV") +
  this_theme +
  coord_sf(xlim=range(predgrid$x), ylim=range(predgrid$y), expand=FALSE)

# print the plots side-by-side
predplot + CVplot
```



We can also calculate an overall summary for comparison with results in the paper:

```
# abundance estimate
sum(vp$pred)
```

```
## [1] 3935.35
```

```
# standard error
sqrt(var(colSums(preds_mh)))
```

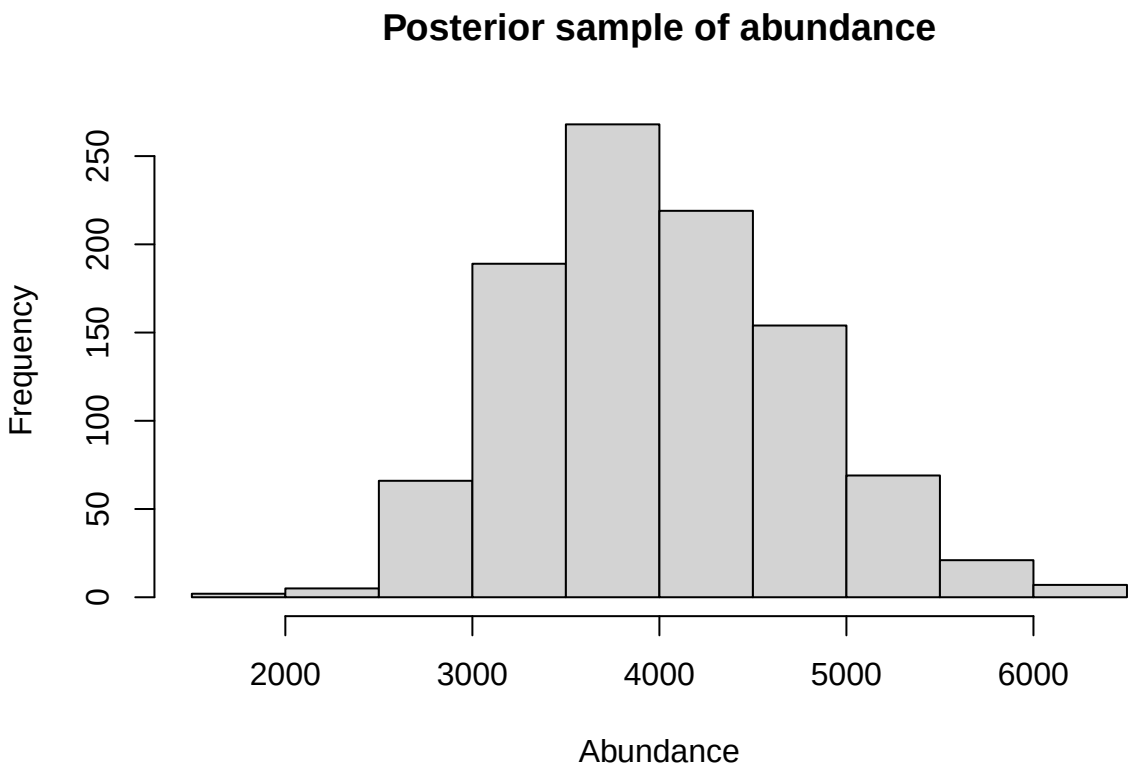
```
## [1] 717.3893
```

```
# coefficient of variation for DSM
dsm_cv <- sqrt(var(colSums(preds_mh)))/(sum(vp$pred))
dsm_cv
```

```
## [1] 0.1822937
```

We may also want to plot a histogram of the posterior distribution of abundance:

```
hist(colSums(preds_mh), xlab="Abundance", main="Posterior sample of abundance")
```



Adding availability (0.34) and $g(0)$ CV (0.16) for the aerial survey to get a "total" CV:

```
sqrt(dsm_cv^2 + 0.34^2 + 0.16^2)
```

```
## [1] 0.4176493
```

Although we cannot make an exact comparison from Sigourney et al. (2020), the authors make estimates fixing the group size to observed values ($\hat{N}$=4013; CV=0.31) and fixing the availability to its mean as we do here ($\hat{N}$=4345; CV=0.21) while leaving the rest of the model as-is. Here we assume that availability and $g(0)$ are independent and sum squared CVs to get the total CV which is higher than that from the original paper.

References

- Palka DL, Chavez-Rosales S, Josephson E, Cholewiak D, Haas HL, Garrison L, Jones M, Sigourney D, Waring G, Jech M, Broughton E, Soldevilla M, Davis G, DeAngelis A, Sasso CR, Winton MV, Smolowitz RJ, Fay G, LaBrecque E, Leiness JB, Dettloff Warden M, Murray K, Orphanides C. 2017. Atlantic marine assessment program for protected species: 2010-2014, OCS Study BOEM 2017-071. Washington: US Dept. of the Interior, Bureau of Ocean Energy Management, Atlantic OCS Region. 211
- Sigourney DB, Chavez-Rosales S, Conn PB, Garrison L, Josephson E, Palka D. 2020. Developing and assessing a density surface model in a Bayesian hierarchical framework with a focus on uncertainty: insights from simulations and an application to fin whales (Balaenoptera physalus) PeerJ 8:e8226 https://doi.org/10.7717/peerj.8226