Regional estimation - Big Sur example

AUTHOR
David A. Boughton

PUBLISHED
February 26, 2024

Regional estimation example

This memo walks the reader through an example of regional estimation of rainbow trout abundance and O. mykiss density, using the approach laid out by Boughton, Nelson and Lacey 2022 (henceforth referred to as FB182). Data was provided by Matt Michie of CDFW. This memo was generated in the programming software R, as a Quarto file which produces a mixture of descriptive text and programming calculations.

Set up the computing environment

The embedded computer code below sets up the R computing environment by adding some useful R packages:

```
library(tidyverse) # various useful Hadley-Wickham resources
library(readxl)    # for importing excel files
library(FSA)       # Simple Fisheries Stock Assessment Methods (for depletion estimator)
```

Target of estimation

The target of estimation is the Big Sur Coast BPG, which Matt reports to have 141 km of accessible habitat, divided into a sample frame of 691 short reaches.

```
sumL1_m <- 141000   # total length of all short reaches, in meters
N <- 691            # number of short reaches
L1_bar <- sumL1_m/N # mean length of short reach
L1_bar
```

```
 [1] 204.0521
```

The mean reach length `L1_bar` is 204 meters. Notice that `L1` is the total length of a short reach, as measured in the GIS of the sampling frame. Below we will also use `L2`, the measured length of stream sampled between the two block nets, which is usually less than `L1`.

Reach-level abundances

A subset of the 691 short reaches were randomly selected and sampled for *O. mykiss* abundance. On page 99 of FB182, equations 6 & 7 give formulas for estimating reach-level abundances from mark-recapture data. However, here a depletion sampling approach was used rather than mark-recapture, so a different estimation procedure is required. Here I use the Carle-Strub method, as implemented in the `removal` function in the [FSA package](#) by Derek Ogle. There are a couple different statistical models for estimating abundance from depletion samples, but this one seems to perform well even at small sample sizes.

Matt provided an excel file `R4 Yr 1 Low-Flow Survey E-fishing Data_QAQC 1-5-24.xlsx` with two sheets. The first sheet has one row for each pass of the electrofisher at each site; the second sheet has one row for each fish captured. Below I read in these tables and use them to compute the number of fish captured at each pass at each site. Note that zeros are important here: a pass that captured zero fish is part of the depletion, and a site that captured zero fish across all passes is part of the dataset and should not be omitted.

```r
fn <- "data-raw/R4 Yr 1 Low-Flow Survey E-fishing Data_QAQC 1-5-24.xlsx"  # file name
# read_xlsx function from package readxl
fish <- read_xlsx(path=fn,                    # file name
                  sheet="Fish Sampled",   # sheet with the data
                  skip=1,                      # skip first row
                  col_names=c("Date", "SiteID", "Pass", "SpCode", "FL_mm", "Wt_g", "Notes"))
pass <- read_xlsx(path=fn,                      # file name
                  sheet="Pass Information",  # sheet with the data
                  range="B3:L51",              # subset of columns with necessary information
                  col_names=c("StreamName", "Surveyors", "SiteID", "L2_m", "Weather", "GPSstart", "G
                              "AirTemp_C", "WaterTemp_C", "Q_cfs", "Pass"))

# count up number of fish captured per pass
tmp <- fish |> group_by(SiteID, Pass) |> summarise(catch=n())
```

`summarise()` has grouped output by 'SiteID'. You can override using the
`.groups` argument.

```r
# join with pass file
pass <- left_join(pass, tmp, by=c("SiteID", "Pass"))
rm(tmp)
# insert zero catches into pass file
pass <- pass |> mutate(catch=ifelse(is.na(catch), 0, catch))
# create site file (one row per site)
site <- select(pass, StreamName, SiteID, L2_m) |> distinct() # condense to one row per site
# number of sample sites
n <- nrow(site)
# sample lengths between block nets
L2_m <- site$L2_m
```

So there were a total `n` of 15 reaches sampled out of the 691 of the sample frame. All the formulas in FB182 include
the possibility that some sampled reaches were dry. These data provide information on drought refugia and are
summarised as `fw`, the fraction of wet reaches. In this dataset, no dry reaches were reported and thus `fw` is simply
1.0.

```r
fw <- n/n    # fraction of wet reaches
```

Now I loop through the sites and compute a depletion estimate for each site, using the `removal` function in package
`FSA

```r
# create empty vectors
M_hat <- V_M_hat <- rep(NA, n)
# loop through sites
for(i in 1:n) {
  id <- site$SiteID[i]
  dat <- filter(pass, SiteID==id)
  est <- removal(dat$catch)  # from the FSA package
  M_hat[i] <- est$est[1]      # estimate
  V_M_hat[i] <- est$est[2]^2  # exponent converts SE to sample variance
}
# if zero catch, removal reports sample variance as NaN (not a number)
```

```
# assign it zero, similar to if one fish was caught
V_M_hat[M_hat==0] <- 0
```

Here are the estimates and standard errors:

```
         StreamName SiteID L2_m M_hat    SE_M_hat
1     Malpaso Creek    533   99    15   3.9760618
2       Rocky Creek     51   95    42   1.8101474
3       Bixby Creek      7   96    24   1.0291681
4   Partington Creek   672   56    19   0.6203348
5     Big Sur River      9   91   178  41.7543158
6         Garrapata     14   92    47   3.8545372
7        Mill Creek     21   93     0   0.0000000
8      Salmon Creek      8   63   105   8.6902576
9         Big Creek     18   81   126  78.4699334
10     Devil's Creek    210   82   120  10.5814141
11     Prewitt Creek     17   96    77   2.8134004
12     Prewitt Creek      1   67    47   3.8908042
13       Bixby Creek     19  100    74   7.9771370
14     San Jose Creek    28   82   113  10.3004969
15     Limekiln Creek     76   97     1   0.0000000
```

Make the regional estimate of density

Now I'll make a regional estimate of *O. mykiss* density as on page 99 of FB182. I start with the reach-level estimates of abundance and its sampling variance, *aka* `M_hat` and `V_M_hat`, calculated above.

To get average density for the BPG we divide these by the wetted area, but notice that there are two ways one could do this:

1. Compute the fish density for each reach, and **then** take the average across reaches. This seems (to me at least) like the most intuitive way to do it, but in fact it is **not** the best way.

2. Add up all the abundances across the reaches, and add up all the wetted areas across the reaches, and **then** divide the summed abundances by the summed areas to get the average density for the whole BPG.

This second way to do it in fact gives a better estimate, and is called a "ratio estimator" by Thompson (2012). The reason it gives a better estimate is that all the individual estimates of abundance, and all the individual estimates of wetted area have small unknown errors of estimation - some positive, some negative, but averaging zero overall. These distort the estimate of density obtained by division (fish per area), but when you first add up the samples before doing the division, the errors tend to cancel each other out (the positive errors cancel the negative errors) and the distortion is smaller.

So all the various estimators in the appendix of FB182 use this trick to increase accuracy, with the cost that the equations are sometimes less intuitive.

I can't compute 2D fish density (fish per meter square) with the data provided, but I can compute 1D fish density (fish per meter of channel). It uses the same equations, except that everywhere there is an estimate of wetted area (`A`) we substitute a length (`L2`). Here are the calculations using the ratio estimators:

```
Dw_hat <- sum(M_hat)/sum(L2_m)    # estimated density, FB182 equation 8
Dw_hat
```

```
[1] 0.7658915
```

```
# error calculations
Vr <- (1/(n-1))*sum((M_hat-Dw_hat*L2_m)^2)  # ratio variance, FB182 equation 9
Vm <- mean(V_M_hat)                         # site-level variance, FB182 equation 10
L2_bar <- sum(L2_m)/n                       # mean length of sampled portion of reaches

# and now, the sample variance for density, FB182 equation 11
Vdw <- (1/L2_bar^2)*(Vr*(((N*fw)-n)/(n*N*fw))+(Vm/(N*fw)))
sqrt(Vdw)  # standard error is the square root of this variance
```

[1] 0.165921

```
# finally, the 95% confidence intervals from equation 12, using the t-distribution
Dw_l95 <- Dw_hat + sqrt(Vdw)*qt(0.5-95/200, n-1)
Dw_u95 <- Dw_hat + sqrt(Vdw)*qt(0.5+95/200, n-1)
c(Dw_hat, Dw_l95, Dw_u95)
```

[1] 0.7658915 0.4100264 1.1217565

So, in the Big Sur region, the average density of *O. mykiss* (per meter of stream channel) was 0.766 fish per meter, with a standard error of 0.166 and 95% confidence limits of {0.41, 1.122}. The coefficient of variation (the SE scaled as percent of the mean) is 21.7%, which is not bad for a sample size of 15 reaches.

Make the regional estimate of rainbow trout abundance

Now I'll use these data to make a regional estimate as in "Abundance of Rainbow Trout" on page 104 of FB182. This is a little more complicated, but I'm still using ratio estimators. There are actually two ratio estimators: one for ratio of rainbow trout to total *O. mykiss*, and one for the total *O. mykiss* per meter of stream channel. These are multiplied by each other, and then by the total length of stream channel in the sample frame, to get an estimate of total number of rainbow trout:

(RBT/fish) x (fish/meter) x (total meters of channel) = total number of rainbow trout

Following through on the instructions on page 104, we already have the results from Eqns. 6 & 7, `M_hat` and `V_M_hat` from our calculations above. Now we need the counts of captured fish as in Eqns. 27 and 28. The formulas given are for mark-recapture, but really these are just the total number of fish handled at each site (Eqn. 27), and the number handled that meet some criteria for adulthood (Eqn. 28). On page 46 of FB182 we developed a simple scheme for adulthood (adult rainbow trout) based on Fork Length:

Fork Length < 150 mm	Mostly immature fish, some precocious males
Fork Length 150 - 200 mm	Enigmatic; some adults, some presmolts, some large parr
Fork Length > 200 mm	Mostly adult rainbow trout; some presmolts

So the counts for Eqn. 28 are simply fish with FL > 200 mm. For completeness, I will also compute estimates for the juvenile fish and the enigmatic fish. For this I will use the second spreadsheet provided by Matt, which gives fork lengths for all the fish handled at each site during the depletion sampling.

```
# classify fish as juveniles, enigmatic, adult trout, and tally up the counts
counts <- fish |> group_by(SiteID) |>             # make counts by site
        summarise(m=n(),                          # total count
                a=sum(FL_mm>200),                 # number of adults
                en=sum(FL_mm>=150 & FL_mm<=200),  # number of enigmatic
```

```r
                     j=sum(FL_mm<150))                      # number of juveniles
# arrange to have same order as site
counts <- left_join(site, counts, by=c("SiteID"))
# replace NAs with zeros for sites where no fish were caught
counts <- counts |> mutate(m= ifelse(is.na(m),  0, m),
                           a= ifelse(is.na(a),  0, a),
                           en=ifelse(is.na(en), 0, en),
                           j= ifelse(is.na(j),  0, j))
m <- counts$m; a <- counts$a; en <- counts$en; j <- counts$j # pop to global environment

print(counts)
```

```
# A tibble: 15 × 7
   StreamName        SiteID  L2_m      m      a     en      j
   <chr>              <dbl> <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
 1 Malpaso Creek        533    99     13      0      0     13
 2 Rocky Creek           51    95     41      0      5     36
 3 Bixby Creek            7    96     24      1      3     20
 4 Partington Creek     672    56     19      0      4     15
 5 Big Sur River          9    91    109      0      0    109
 6 Garrapata             14    92     43      0      0     43
 7 Mill Creek            21    93      0      0      0      0
 8 Salmon Creek           8    63     90      2     14     74
 9 Big Creek             18    81     57      0      9     48
10 Devil's Creek        210    82    101      2     16     83
11 Prewitt Creek         17    96     74      0      6     68
12 Prewitt Creek          1    67     43      0      1     42
13 Bixby Creek           19   100     63      1      7     55
14 San Jose Creek        28    82     95      0      3     92
15 Limekiln Creek        76    97      1      1      0      0
```

As you can see, most of the sites have zero adults. One of the nice things about the ratio estimators are that all these zeros do not produce a large distortion in the estimates.

So, now I estimate the 1D density of adults (rainbow trout per meter of stream channel), using Eqn. 29 in FB182, and then multiply it by total wetted channel length to get the estimate for total adult rainbow trout in the Big Sur

```r
Dr_hat <- sum(a)*sum(M_hat)/(sum(m)*sum(L2_m))   # estimated density, FB182 equation 29
Dr_hat
```

```
[1] 0.006935628
```

```r
Tr_hat <- N*fw*L1_bar*Dr_hat                        # estimated total, FB182 equation 30
Tr_hat
```

```
[1] 977.9235
```

And now the calculations for standard error:

```r
# error calculations
Vr3 <- (1/(n-1))*sum((M_hat*a-Dr_hat*L2_m*m)^2)  # ratio variance, FB182 equation 31
m_bar=mean(m)                                      # mean catch

# sample variance for adult linear density, FB182 equation 32
```

```
Vdr <- (1/(L2_bar*m_bar)^2)*(Vr3*(((N*fw)-n)/(n*N*fw))+(Vm/(N*fw)))
sqrt(Vdr)  # standard error is the square root of this variance
```

`[1] 0.004266098`

```
# sample variance for total adult abundance, FB182 equation 33
Vtr <- Vdr*(N*fw*L1_bar)^2
sqrt(Vtr)  # standard error is the square root of this variance
```

`[1] 601.5198`

```
# finally, the 95% confidence intervals from equation 12, using the t-distribution
Tr_l95 <- Tr_hat + sqrt(Vtr)*qt(0.5-95/200, n-1)
Tr_u95 <- Tr_hat + sqrt(Vtr)*qt(0.5+95/200, n-1)
c(Tr_hat, Tr_l95, Tr_u95)
```

`[1]  977.9235 -312.2082 2268.0553`

So, in the Big Sur region, the total abundance of Rainbow Trout was 978 adult residents, with a standard error of 602 and 95% confidence limits of {-312, 2268}. Note that the lower limit is negative, which is impossible and reflects that the formula for confidence intervals is approximate. Of course, we can sum up the `a`'s and determine that the minimum abundance is 7.

The coefficient of variation (the SE scaled as percent of the mean) is 61.5%, which is significantly higher than the coeffient of variation for density, 21.7%. This is a typical and inevitable feature of estimating totals versus densities, arising from equation 33 in FB182.

References

Boughton, D. A., J. Nelson and M. K. Lacy (2022). "Integration of steelhead viability monitoring, recovery plans and fisheries management in the southern coastal area." Fish Bulletin 182. State of California, Department of Fish and Game. Available from https://nrm.dfg.ca.gov/FileHandler.ashx?DocumentID=199225

Thompson, S. K. (2012). Sampling, third edition. Hoboken, NJ, USA, John Wiley & Sons, Inc.