Data Prep

Ben Buzzee, Biometrician, ADFG

July 30, 2018

This file is dedicated to loading, cleaning and preparing the raw data for analysis. Three datasets will be output. One for length composition, one for population estimate, and one for diet analysis.

Data for Population Estimate

First we will start by reading in the mark event data. This is stored in mark_data_raw.csv. The data we are interested is in the first 7 columns and first 563 rows (including column headers). We'll also rename the columns for easier manipulation and add a mark-event indicator column that we will use later.

```
mark <- read.csv("../data/mark_data_raw.csv",
                 header = TRUE,
                 stringsAsFactors = FALSE,
                 nrows = 562,
                 skip = 1)


# keep only the first seven columns
# No need to keep pike # or samplers (for now)
mark <- mark[,c(1,4:7)]


names(mark) <- c("gear", "section", "date", "fl", "tag")

mark$tag <- as.character(mark$tag)

#check for NA's

# apply(X = mark,
#       MAR = 2,
#       FUN = function(x){any(is.na(x))}
#       )
#
#
# sum(is.na(mark$fl))


# add mark event indicator
# filter removes the 2 missing fl values

mark_mr <- mark %>% mutate(mark = 1)
```

Next we'll load the recap data. We'll get this from the dissection excel document. Since so few tags are lost, and we can probably assume there was no systemic reason for tag losses, we will remove those observations. We will also remove non-pike observations.

```r
recap <- read.csv("../data/dissection_data_raw.csv",
                  header = TRUE,
                  stringsAsFactors = FALSE)


#sum(recap$Tag.. == "TL")
# remove tag losses and non-pike observations


recap <- recap %>% filter(Tag.. != "TL", Species == "NP")
```

Now we will piece together the dataset we will use to estimate the population size. We will remove fish caught in Fyke nets, and join the datasets from the mark event and recapture events on tag number.

```r
# I'm going to remove fyke net captures since they violate the ideal of uniform effort
# but so few pike were caught in them it doesn't really influence our estimate

fyke_tags_mark <- mark$tag[mark$gear == 'Fyke']

recap_mr <- recap  %>% select(Length, Tag.., Section.., Gear ) %>% mutate(recap = 1)

# mark recapture dataframe
# filter by fish lengths here
mr_df <- full_join(mark_mr, recap_mr, by = c("tag" = "Tag..")) %>% select(gear_mark = gear,
                                                                          section_mark = section,
                                                                          mark_len = fl,
                                                                          tag,
                                                                          section_recap = Section..,
                                                                          recap_len = Length,
                                                                          gear_recap = Gear,
                                                                          mark = mark,
                                                                          recap = recap)


mr_df <- mr_df %>% mutate(both_events = ifelse(mark == 1 & recap == 1, yes = 1, no = 0))



mr_df <- mr_df %>% filter((is.na(gear_mark) | gear_mark != "Fyke") & (is.na(gear_recap) | gear_recap !=

mr_df <- mr_df %>% filter()

# sum((mr_df$tag %in% fyke_tags_mark))
# unique(mr_df$gear_mark)
# which(mr_df$gear_mark == "")
# mr_df[97,]


# sum(!is.na(mr_df$mark))
# sum(!is.na(mr_df$recap))
# sum(!is.na(mr_df$both_events))
```

```
mr_df$mark  <- ifelse(is.na(mr_df$mark), yes = 0, no = 1)
mr_df$recap  <- ifelse(is.na(mr_df$recap), yes = 0, no = 1)
mr_df$both_events <- ifelse(is.na(mr_df$both_events), yes = 0, no = 1)
mr_df <- mr_df %>% filter(is.na(recap_len) | recap_len >= 300)

write_csv(mr_df, path = "../data/mr_data.csv")
```

The final dataset we will use in our analysis looks like the following:

```
as_tibble(mr_df)
```

```
## # A tibble: 924 x 10
##    gear_mark section_mark mark_len tag   section_recap recap_len gear_recap
##    <chr>     <chr>           <int> <chr> <chr>             <int> <chr>
##  1 Gillnets  1                 379 1000  <NA>                 NA <NA>
##  2 Gillnets  1                 417 1003  <NA>                 NA <NA>
##  3 Gillnets  1                 476 1004  1                   477 Gill Net
##  4 Gillnets  1                 355 1005  1                   351 Gill Net
##  5 Gillnets  1                 336 1006  2                   343 Gill Net
##  6 Gillnets  1                 413 1007  1                   418 Gill Net
##  7 Gillnets  1                 380 1008  2                   392 Gill Net
##  8 Gillnets  1                 333 1009  2                   354 Gill Net
##  9 Angling   4                 379 1025  <NA>                 NA <NA>
## 10 Angling   4                 397 1026  4                   418 Gill Net
## # ... with 914 more rows, and 3 more variables: mark <dbl>, recap <dbl>,
## #   both_events <dbl>
```