EDA song sparrow data

2023-01-17

Data frame overview

Pedigree

The first object is `d.ped` which contains the pedigree information.

```
summary(d.ped)
```

```
##     ninecode             gendam             gensire
##  Min.   :109137448   Min.   :109137468   Min.   :109137448
##  1st Qu.:146164012   1st Qu.:146130794   1st Qu.:146130313
##  Median :176124850   Median :176124382   Median :176124004
##  Mean   :196520240   Mean   :188116000   Mean   :185463038
##  3rd Qu.:243185045   3rd Qu.:226189260   3rd Qu.:226189228
##  Max.   :999999999   Max.   :999999999   Max.   :266176829
##                      NA's   :59          NA's   :59
```

It has the columns *ninecode*, *gendam*, and *gensire*. The first column cannot be `NA` and is the unique identifier for an individual, whereas `gendam`and `gensire` are references (foreign keys) to the known maternal and paternal link, respectively. Both of these columns have 59 NAs. In fact, these NAs overlap completely since they are the founder population with no defined paternal or maternal link:

```
d.ped[is.na(d.ped$gendam), "gensire"]
```

```
##  [1] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## [26] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## [51] NA NA NA NA NA NA NA NA NA
```

We see that *gensire* is NA for all instances where *gendam* is also NA.

d.Q

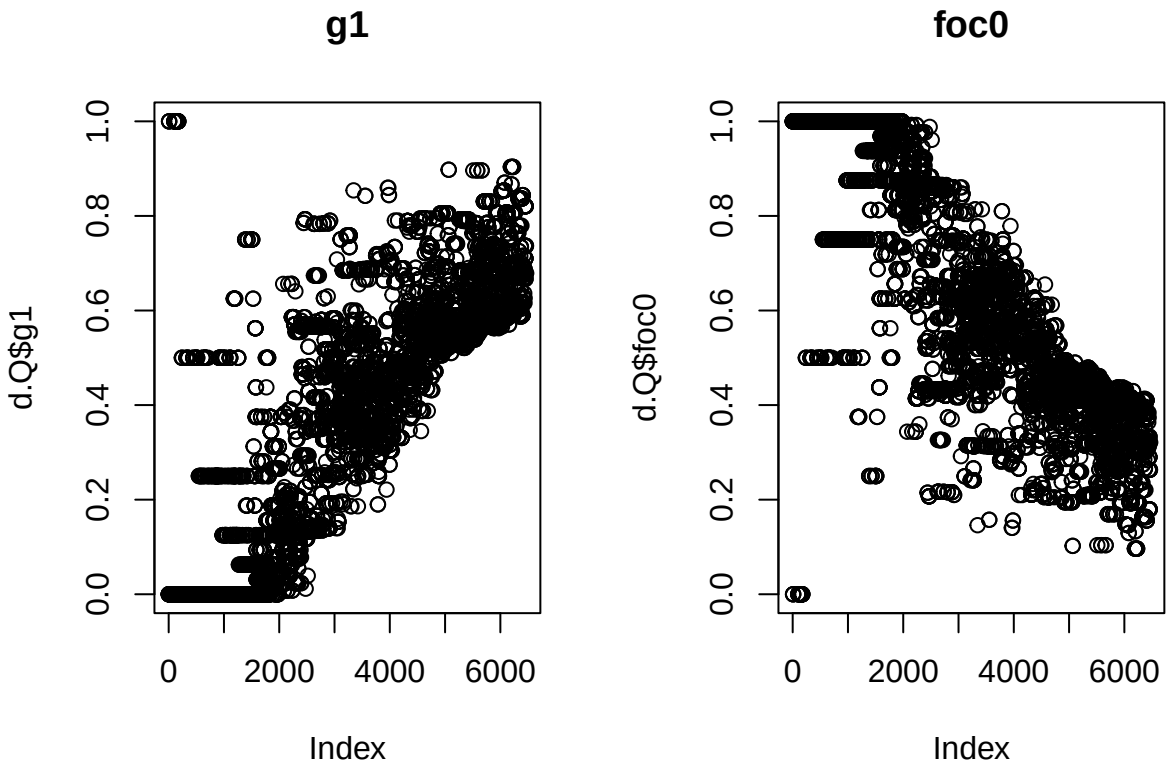
This table has the columns *g1*, *foc0* and *ninecode* (ID).

```
head(d.Q)
```

```
##   foc0 g1  ninecode
## 1    1  0 109137407
## 2    1  0 109137408
## 3    1  0 109137418
## 4    1  0 109137420
## 5    1  0 109137421
## 6    1  0 109137425
```

From the first look, it might seem like these are binary/categorical variables, but plotting the values across indices show that the order of the rows are structured so that they start at 1 and 0 respectively.

```
par(mfrow = c(1, 2))
plot(d.Q$g1, main = "g1")
plot(d.Q$foc0, main = "foc0")
```



We can also look at the correlation between these two values

```
cor(d.Q$foc0, d.Q$g1)
```

```
## [1] -1
```

Very strong negative correlation here. We can also look at the individuals whose ID were in the *founder population*:

```
founder_population.id <- d.ped[is.na(d.ped$gendam), "ninecode"]
table(d.Q[which(d.Q$ninecode %in% founder_population.id), c("foc0", "g1")])
```

```
##     g1
## foc0  0  1
##    0  0 33
##    1 26  0
```

The values seem to be relatively balanced between 0 and 1 in the founder population. This supports the idea that they measure the immigration contribution to the genetic composition of the individuals. All immigrant individuals are completely immigrant, have no pedigree and are thus part of the founder population. The latter are those who are the "inital" natives on the island, meaning that their values must be exactly zero.

ped.prune

This is a pruned pedigree, only considering the 1993-2018 observations but also combining the knowledge of the 1973-1993 observations into them (I think).

qg.data.gg.ind

This object has the following shape:

```
head(qg.data.gg.inds)
```

```
##    ninecode natalyr sex.use nestrec surv.ind.to.ad brood.date sex.use.x1
## 1 111111112    2012       0    3086              0        120          1
## 2 111111121    2015       0    3237              0        141          1
## 3 143173366    1993       1    1838              1         96          1
## 4 143173381    1993       2    1867              1        102          2
## 5 143173382    1993       1    1867              0        102          1
## 6 143173384    1993       1    1851              0        102          1
##       f.coef      foc0        g1 natalyr.no sex
## 1 0.11155218 0.4085679 0.5914321         38   0
## 2 0.04814660 0.3299752 0.6700248         41   0
## 3 0.05108643 0.5283203 0.4716797         19   0
## 4 0.03125000 0.6250000 0.3750000         19   1
## 5 0.03417969 0.4335938 0.5664062         19   0
## 6 0.02148438 0.6328125 0.3671875         19   0
```

The response variable we will use is `surv.ind.to.ad`. Some immediate observations:

```
paste("Earliest year:", min(qg.data.gg.inds$natalyr))
```

```
## [1] "Earliest year: 1993"
```

```
paste(c("Number not survived:", "Number survived:"), table(qg.data.gg.inds$surv.ind.to.ad))
```

```
## [1] "Number not survived: 1817" "Number survived: 661"
```

```
paste("natal year correlation:", cor(qg.data.gg.inds$natalyr, qg.data.gg.inds$natalyr.no))
```

```
## [1] "natal year correlation: 1"
```

```
paste("correlation between sex and sex.x1:", cor(qg.data.gg.inds$sex.use, qg.data.gg.inds$sex.use.x1))
```

```
## [1] "correlation between sex and sex.x1: 0.842997540673555"
```

An overview over the columns:

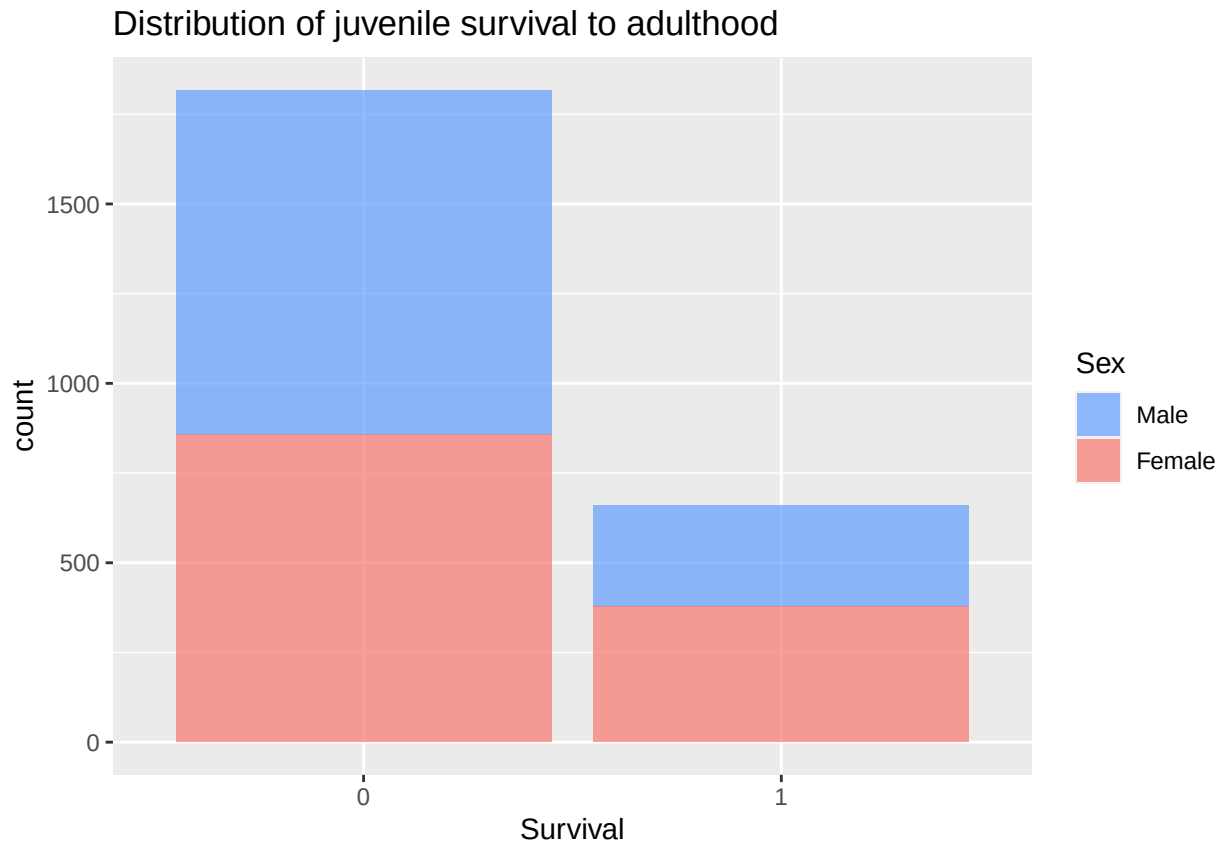
- *ninecode*: Individual ID
- *natalyr*: Year individual was born, e.g. 2015.
- *sex.use*: **Not in use**
- *nestrec*: ID for nest number
- *brood.date*: Day of the year when the first offspring in individuals nest hatched (I think zero is April 1st, it's in Rekkebo thesis)
- *sex.use.x1*: Sex of individual, 1 or 2
- *f.coef*: Inbreeding coefficient
- *foc0*: "How foreign" individual is, related to f.coef
- *g1*: Inverse of *foc0* essentially.
- *natalyr.no*: The same as natal year, starting with 1974 as 0 (2015=41).

Some plotting

```
library(ggplot2)
```

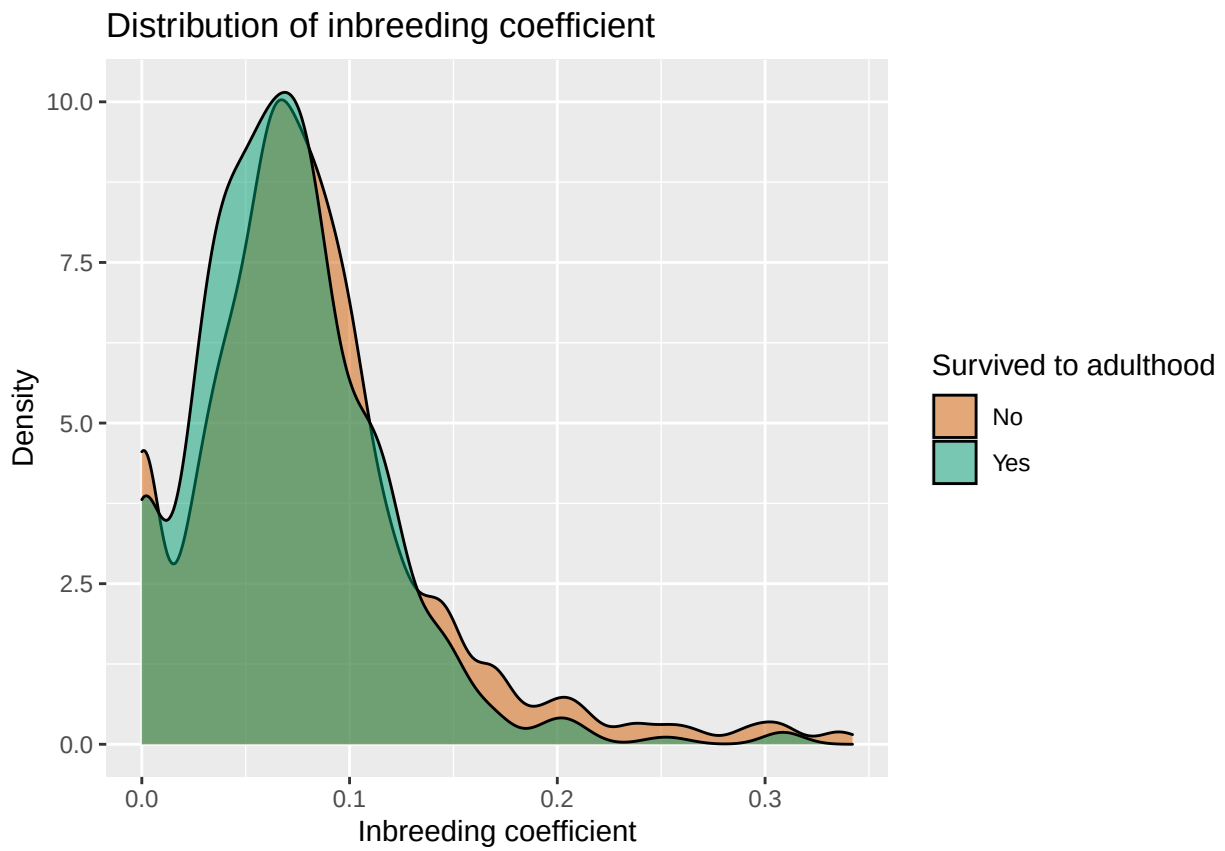
```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

```
ggplot(qg.data.gg.inds, aes(x = factor(surv.ind.to.ad), fill = factor(sex))) +
  ggtitle("Distribution of juvenile survival to adulthood") +
  geom_bar(alpha = 0.7) +
  xlab("Survival") +
  scale_fill_manual(
    name = "Sex",
    labels = c("Male", "Female"),
    values = c("#619CFF", "#F8766D")
  )
```



Seems like the sex in relation to survival is relatively balanced here. Next, we

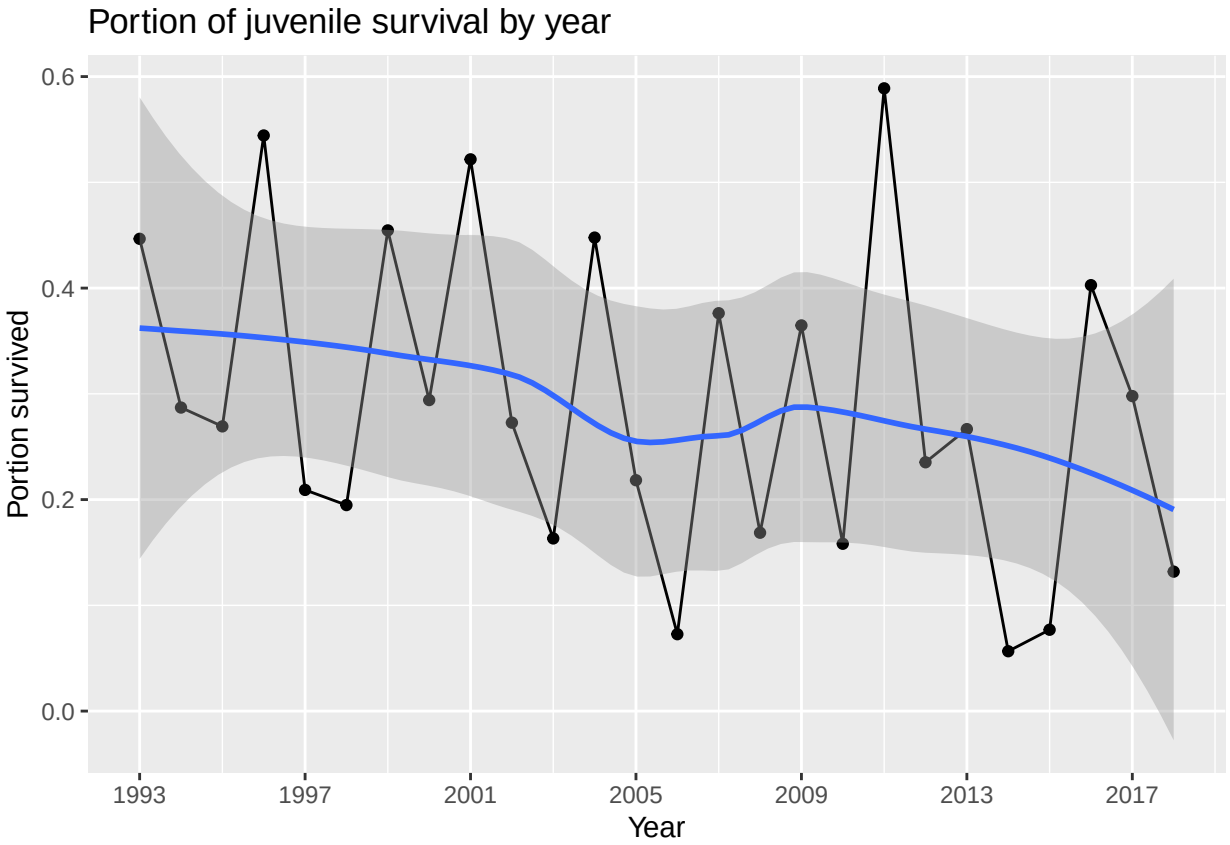
```
ggplot(qg.data.gg.inds, aes(x = f.coef, fill = factor(surv.ind.to.ad))) +
  ggtitle("Distribution of inbreeding coefficient") +
  geom_density(alpha = 0.5) +
  xlab("Inbreeding coefficient") +
  ylab("Density") +
  scale_fill_manual(
    name = "Survived to adulthood",
    label = c("No", "Yes"),
    values = c("#D55E00", "#009E73")
  )
```



Here we see that survival is a bit more skewed towards lower inbreeding coefficients. We can also look at how the proportion of individuals have survived over the years:

```
ggplot(aggregate(surv.ind.to.ad ~ natalyr, qg.data.gg.inds, function(x) {
  sum(x) / length(x)
}), aes(x = natalyr, y = surv.ind.to.ad)) +
  geom_line() +
  geom_point() +
  geom_smooth() +
  ggtitle("Portion of juvenile survival by year") +
  xlab("Year") +
  ylab("Portion survived") +
  scale_x_continuous(breaks = seq(1993, 2018, by = 4))
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



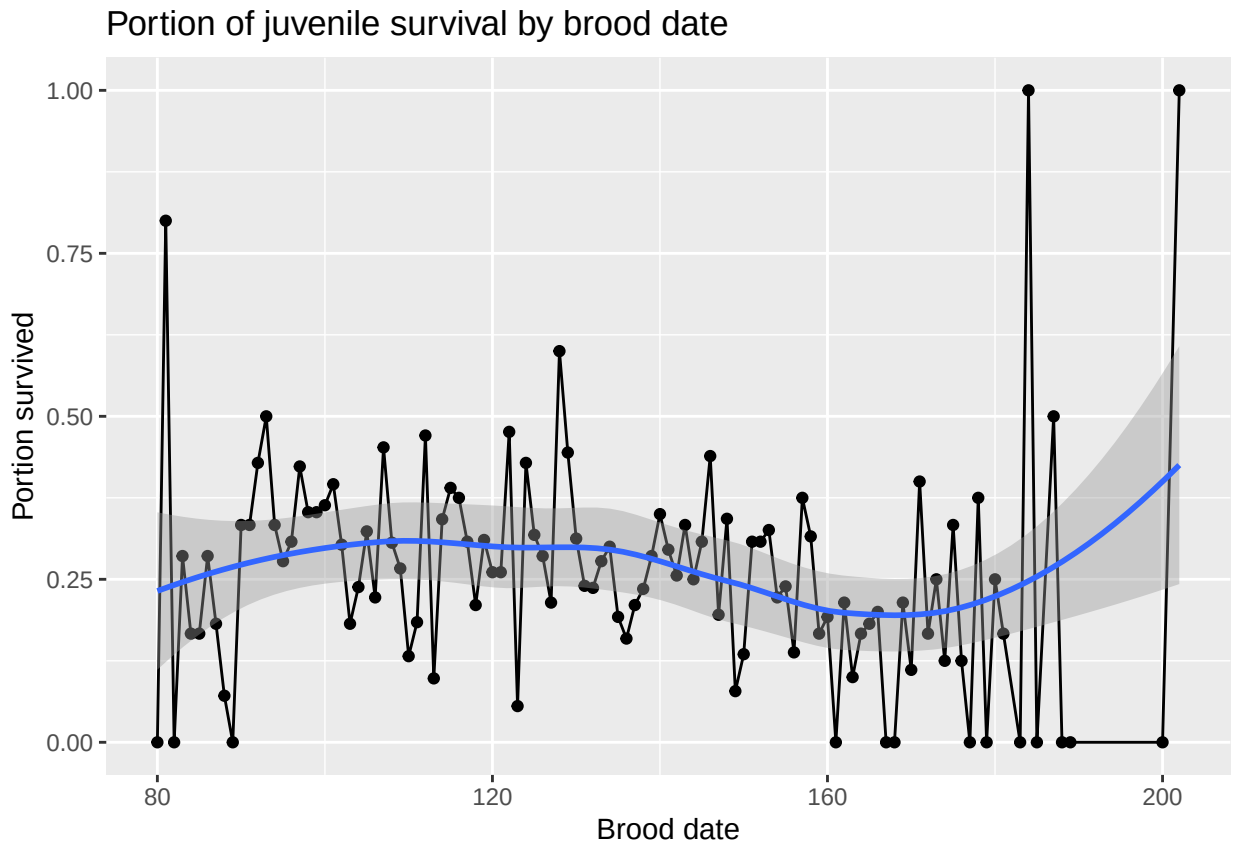
There seem to be very little trending along the years, but possibly a small negative trend. Let's also see if there is some correspondence between genetic group coefficient (`g1`) and juvenile survival.

```
ggplot(qg.data.gg.inds, aes(x = g1, fill = factor(surv.ind.to.ad))) +
  ggtitle("Distribution of genetic group coefficient") +
  geom_density(alpha = 0.5) +
  xlab("Genetic group coefficient") +
  ylab("Density") +
  scale_fill_manual(
    name = "Survived to adulthood",
    label = c("No", "Yes"),
    values = c("#D55E00", "#009E73")
  )
```

This shows a similar result to the inbreeding coefficient, namely a skew towards the right (lower values of coefficient) in the group that survived. We might also look into survival based on brood date:

```
ggplot(aggregate(surv.ind.to.ad ~ brood.date, qg.data.gg.inds, function(x) {
  sum(x) / length(x)
}), aes(x = brood.date, y = surv.ind.to.ad)) +
  geom_line() +
  geom_point() +
  geom_smooth() +
  ggtitle("Portion of juvenile survival by brood date") +
  xlab("Brood date") +
  ylab("Portion survived")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



This last plot seem to indicate that survival is relatively stable and somewhat decreasing for those hatched relatively late. For the largest values of brood date, we get an increasing trend but also much uncertainty since not that many were hatched this late.