Stom Dataset Checks

2024-07-08

This document is to check why the prey_ind_weight column isn't correct in some cases, or if I have misinterpretted something. (the problem is that prey individual weight is only equal to prey weight / prey count in approximately 50% of the cases)

I have found the document that describes the columns. (https://www.cefas.co.uk/media/glyfpvo1/dapstom-phase-4-report-2014-dlm.pdf) Howere this has no description of the prey_weight_g column, but does describe prey_ind_weight_g as the individual prey weight. It might be the stomach_weight column that has been changed (Total Stomach Weight – total weight (in grams) of all contents within the individual predator stomach.)

(side note, n_stomachs "indicates how many individuals the PRED ID represents. This has a value of '1' where the data has been collected at the individual animal level, however it may be higher for pooled data", should we have been using this for the weighting?)

Will first separate this data into dapstom and ices (the two data sources)

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
load("C:/Users/lucab/Downloads/stomach_dataset.Rdata")

dapstom <- stom_df%>%filter(data=="Dapstom")
ices <- stom_df%>%filter(data=="ICES_YOTS")
```

So the DAPSTOM data does have a prey_weight_g column, but this isn't discussed in the document. Strange.

Now I am going to check where prey_weight_g / n_preyperpred, in either the dapstom or ices datasets

```
Ydapstom <- dapstom[dapstom$prey_ind_weight_g==dapstom$prey_weight_g/dapstom$nprey_perpred,]
Ndapstom <- dapstom[dapstom$prey_ind_weight_g!=dapstom$prey_weight_g/dapstom$nprey_perpred,]

Yices <- ices[ices$prey_ind_weight_g==ices$prey_weight_g/ices$nprey_perpred,]
Nices <- ices[ices$prey_ind_weight_g!=ices$prey_weight_g/ices$nprey_perpred,]
```

So Ydapstom and Yices are correct, or they operate how I think they should, where prey_weight_g/n_preyperpred=prey_ind
50,146/190,295 in the ices data is correct. 97,653/110,243 in the dapstom data is correct.

In the dataset, we have columns prey_count, prey_weight, prey_ind_weight, gprey_perpred and n_preyperpred.

It seems based on the naming, that gprey_perpred and n_preyperpred are from the same datasources, and prey_count and prey_weight are from different sources.

Therefore, prey_ind_weight may be from prey_weight/prey_count or from gprey_perpred/n_preyperpred.

From looking at Ndapstom, gprey_perpred / n_preyperpred gets prey_ind_weight, lets check.

```
checkedNdapstom <- Ndapstom[Ndapstom$prey_ind_weight_g==
                              Ndapstom$gprey_perpred/Ndapstom$nprey_perpred,]
countNdapstom <- Ndapstom[Ndapstom$prey_ind_weight_g==
                              Ndapstom$prey_weight_g/Ndapstom$prey_count,]
```

Yes, the prey_ind_weight column is equal to gprey_perpred / nprey_perpred, so it was just that in some cases, gprey_perpred = prey_weight, but these are from different datasrouces.

(However, for only some datapoints is prey_weight_g/prey_count = prey_ind_weight)

Now going to check for the ices datasets.

```
checkedNices <- Nices[Nices$prey_ind_weight_g==
                        Nices$gprey_perpred/Nices$nprey_perpred,]
```

Yep, it is the same.

So prey_ind_weight_g is calculated by gprey_perpred/nprey_perpred, but sometimes gprey_perpred = prey_weight, so it worked in the calculation.

Going to check their PPMR column to see what they have done there as well.

```
ppmr <- stom_df[stom_df$ppmr==stom_df$pred_weight_g/stom_df$prey_ind_weight_g,]
```

Yes, this works, they have calculated it using prey_ind_weight_g and pred_weight_g

However, now I have noticed the column n_stomach, which as said before is the number of predators with the same prey weight?

(Number of Stomachs – indicates how many individuals the PRED ID represents. This has a value of '1' where the data has been collected at the individual animal level, however it may be higher for pooled data.)

We need to figure out if this should factor into the weighting, from this definition, it should. It should factor against the prey weighting column, prey_count, as if across a given PRED ID there are multiple prey counts, this should be scaled by the number of stomachs that these prey counts were taken from, so I think we should do prey_count/stomach count.

Except that this also wouldn't work.

I think number of stomachs means the number of predators that has been sampled, which have the same weight and were sampled at the same time?

Or does it mean that they take 10 predators, put all the stomach contents into one sample, then take from that sample? Then they take the average pred weight across the 10 predators for the predation column?

If it means the latter, then I am not sure how to factor it into the weighting, prey_count / number of stomachs may work. As this would control for the number of predators that are aggregated into a single sample, and also for the number of the same prey that are taken from this sample.(thinking about it this has now confused me as to what prey_count means)