**Biodiversity project**

**Harriet’s working document**

**Datasets:**

1. Worldfish\_master

* This is a Masterfile from Worldfish listing specie names, specie groups and specie codes.

1. Catch\_cons\_Q7:

* This is the dataset we are interested in. Contains information on data collected on fish caught, consumed, sold, lost, processed.
* Unfortunately, this file DOES NOT contain the variable speciename and speciecode. Only has Specie group.

1. Catch\_cons\_Q7sub:

* Sub-set of the Catch\_cons\_Q7 dataset.
* Only has information on speciename, speciegroup and speciecode

**Steps:**

1. Compared the two datasets “Worldfish\_master” and “Catch\_cons\_Q7sub”. Looked out for discrepancies in speciename, speciegroup and speciecode.
   1. Corrected “wrong” specienames in the “ “Catch\_cons\_Q7sub” file using the “Worldfish\_master” as the masterlist
   2. Corrected wrong speciegroups and speciecodes.
   3. Note that some of these errors had previously been identified and corrections suggested by worldfish (file in the BOX folder).
2. Next, I compared the updated “Catch\_cons\_Q7sub” and the “Catch\_cons\_Q7” datasets.
   1. I made the corrections to the speciegroup (similar to above) to the “Catch\_cons\_Q7” dataset.
   2. I then merged the two corrected datasets by ID, date and speciegroup
   3. This gave us a new dataset (newdata2)with all the information we need – catch and consumption data with plus speciename, speciegroup and speciecode.

COMMENT:

* There were 114 observations () that were in the “Catch\_cons\_Q7” dataset BUT not in the “Catch\_cons\_Q7sub” dataset. These data do not have specienames nor speciecodes. For subsequent analysis, I deleted these 114 observations.
* 3 IDs (169, 176 and 249) need to be double checked
  + Each occurs 2 times (2 rows) with unique info
* ID 383 has 2 CFR names assigned (Trapaing veng (coltime 1-4 2012-2013; and Pur Sdey (coltime 13-nov 2014). Looks like they collected data for first 4 time points, then didn’t see the participant again until the 13th time point.
  + Drop the observation for coltime 13 (nov 2014)…most likely a wrong ID was assigned.

1. Next I worked on the “Q1\_BasicInfo” dataset (this dataset has basic data on province, district, village, commune, cfr, enumerator name, HHID, etc)
   1. I created some new variables (year, month, day, season, collection time, CFR category)
   2. There were some errors in the CFR names, so I made the necessary corrections based on the info from worldfish on BOX
   3. The updated dataset = “Q1\_BasicInfo”.
2. Next I merged the two datasets “newdata2” and “Q1\_BasicInfo” (selected variables)
3. Filtered only those participants who had catch/consumption data [fished for fish (i.e where fishtype = fish)] and those with no missing species code…exported that dataset for Sebastian (“seb\_catch”)
4. Did a few descriptive analysis – for fish catch per household per week (NOTE-recall was over past week), I restructured the data such that we have one row per HH ID per collection time…i.e for each HH, the total amount of fish caught per month will be summed up irrespective of the species.

**Data re-structuring:**

1. I have created two sets of data with different structures (based on the fish catch-consumption
   1. Semi-wide – fish catch, consumed, etc and speciecode as variables [wide\_allmerge]
   2. Long – each specie code is a variable (speciecode 1, 2,3…169). Fish catch, consumed, etc all put together as one variable “Fish\_status”

**Other issues to discuss:**

1. 19 time points of data collection instead of 18 time points
   1. 2012 -Nov
   2. 2013 – Jan, Mar, May, Aug, Sep, Nov
      1. Data collected in Aug instead of Jul cos of national election campaign in Jul 2013
   3. 2014 – Jan, Mar, May, Jul, Sep, Nov
   4. 2015 – Jan, Mar, May, Jul, Sep, Nov
2. HH = 413 instead of 400…check where the extra 13 is coming from (check how many observations we have for each HH
3. Weight of fish caught/sold – some very high values…outliers

Summarized data (total per HH)

* 1. (mean, range) fish caught (8.95, 0.03-850 kg)
  2. (mean, range) fish consumed (3.73, 0-42.28 kg)
  3. (mean, range) fish sold (2.7, 0 - 829 kg)
  4. (mean, range) fish processed (1.81, 0-136 kg)
  5. (mean, range) fish lost (0.52, 0-20.7 kg)
  6. (mean, range) fish other (0.20, 0-59 kg)

Raw values

1. (mean, range) fish caught (1.69, 0-750 kg)
2. (mean, range) fish consumed (0.70, 0-17.82 kg)
3. (mean, range) fish sold (0.51, 0 - 750 kg)
4. (mean, range) fish processed (0.34, 0-90 kg)
5. (mean, range) fish lost (0.10, 0-11 kg)
6. (mean, range) fish other (0.04, 0-55 kg)

Next steps – to do:

1. Summarize the data to calculate biodiversity indices based on the group of interest (CFR, year, season, etc)
2. To calculate biodiversity indices, we’ll need to structure the data such that we have one row per group of interest. Example

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Species 1 | Species 2 | Species 3 | Species 4 | Species 5 |
| Site 1 |  |  |  |  |  |
| Site 2 |  |  |  |  |  |
| Site 3 |  |  |  |  |  |
| Site 4 |  |  |  |  |  |

So we can calculate say specie richness per CFR (total # of species per CFR), specie richness per household fish caught, consumed, etc…but the calculations would need to be done separately for each…

1. Current example dataset answers the question - how many species are caught for each CFR?
2. Based on our objectives to compare diversity in catch, sale and consumption, I am planning on having a dataset with one row per HH and the separate variables for diversity indices for catch, sale, consumption, etc.

Future work - Fish available in the environment

Biomonitoring dataset- Bsub

1. Compare list to world fish master (specie name, specie code)
2. Correct all discrepancies in specie name, specie code
3. Decide on what to include (fish vs non-fish)…figure out how to determine this? NOTE: biomonitoring dataset does not include the “fish type” variable that identifies a species as fish/OAA
   1. Select only the “gillnet” catches (doing this cos gill nets were mostly used most of the times)
   2. Merge the masterlist with the biomonitoring dataset to incorporate the “fish type” variable
4. For each occasion and gear type, fishing gear was set up to 8 times (1-8). Let’s find the average (average or sum? – discuss this with Katie) of all the measured indices (fish no, weight, length, etc) for each occasion.
   1. Sort data by CFRName [