Oxygen QC

This notebook describes the pre processing done on Oxygen data in order to parse historical data before QC procedures. There have been consistent issues with the quality of historical oxygen data at BIO and this project for Catherine Johnson requires high quality, reliable oxygen data for analyzing nutrient and oxygen trends on the scotian shelf. The following stes were taken to ensure the removal of any unreliable data from the historical set.

Data was pulled from BIOCHEM database by Shelley Bond, where possible, data was pulled from an updated set known as the BIOCHEM reboot. This updated set has been quality controlled by Gordana Lazin and is already high quality, reliable data from 1990-2013. There were significant issues in the pre 1990 data set from BIOCHEM. These issues have been described in multiple reports by Bond and Lazin. The main issues that were identified and that this pre processing attempts to remove were as follows:

1. The data set included a large chunk of data collected by external organizations and groups outside DFO which were considered unreliable as there is often little informations about the source methods used on collection.
2. The data set needed to be isolated to a relevant geographical area for the area of concern to this particular study
3. There were many input errors in the data set from cruises being duplicated to incorrect metadata including start/end dates, lat/lon information and mission descriptor and sample ID information
4. The units in BIOCHEM for oxygen were all listed as mmol/m3 however this is incorrect and was the default setting depsite data being entered in various units including ml/l and % saturation

The data set pulled by Bond was according to the following SQL script

alter session set nls\_date\_format = 'DD/MM/YYYY';  
  
call the file below Emily\_Winkler  
  
SELECT   
NULL dis\_data\_num,  
 name mission,  
 DESCRIPTOR mission\_descriptor,  
-- LEADER,  
-- MISSION\_START,  
-- MISSION\_END,  
-- INSTITUTE,  
 COLLECTOR\_EVENT\_ID event\_collector\_Event\_id,  
 COLLECTOR\_STATION\_NAME event\_collector\_stn\_name,  
 HEADER\_START\_DEPTH Dis\_header\_start\_depth,  
 HEADER\_END\_DEPTH dis\_header\_end\_depth,  
 HEADER\_START\_LAT dis\_header\_slat,  
 HEADER\_START\_LON dis\_header\_slon,  
 HEADER\_START dis\_header\_sdate,  
 HEADER\_START\_TIME dis\_header\_stime,  
 DATA\_TYPE\_SEQ dis\_data\_type\_seq,  
 METHOD Data\_type\_method,  
 DATA\_VALUE dis\_detail\_data\_val,   
 DATA\_QC\_CODE dis\_detail\_data\_qc\_code,  
 null Dis\_detail\_detection\_limit,  
 COLLECTOR Dis\_detail\_detail\_collector,  
 COLLECTOR\_SAMPLE\_ID dis\_detail\_collector\_samp\_id,  
 'Jay Bugden' created\_by,  
 null created\_date,  
 DATA\_CENTER\_CODE,  
 Institute,  
 'NR' Process\_flag,  
 null batch\_seq,  
 discrete\_detail\_seq dis\_sample\_key\_value  
FROM DISCRETE\_DATA  
WHERE upper(method) like '%WINKLER%'  
and institute not in ('Ministerio de la Ind','IOS','DALHOUSIE UNIVERSITY',  
'DAL','DalhousieU','Private','PINRO','US DOC NOAA NMFS (WO','DREP')  
and descriptor NOT IN  
(select distinct mission\_descriptor from gordana\_winkler)  
UNION ALL  
select \* from gordana\_winkler  
;  
  
The next file should be called Emily\_all\_dupe\_sampleids  
  
select mission, mission\_descriptor, w.dis\_detail\_collector\_samp\_id,  
dis\_header\_sdate, dis\_header\_stime,  
Dis\_header\_start\_depth,  
dis\_detail\_data\_val  
from emily\_winkler w, (  
select dis\_detail\_collector\_samp\_id, count(1)  
from emily\_winkler   
group by dis\_detail\_collector\_samp\_id  
having count(1) > 1) e  
where  
w.dis\_detail\_collector\_samp\_id = e.dis\_detail\_collector\_samp\_id  
--and w.dis\_detail\_collector\_samp\_id not like '-%'  
--and data\_center\_code != 30  
order by dis\_detail\_collector\_samp\_id  
;  
  
The next file Emily\_dupe\_sampids\_metadata  
  
select e.mission\_descriptor, e.dis\_detail\_collector\_samp\_id,  
e.dis\_header\_sdate, e.dis\_header\_stime,  
e.Dis\_header\_start\_depth, e.dis\_detail\_data\_val  
from emily\_winkler e, (  
select mission\_descriptor, dis\_detail\_collector\_samp\_id,  
dis\_header\_sdate, dis\_header\_stime,  
Dis\_header\_start\_depth, count(1)  
from emily\_winkler   
group by mission\_descriptor, dis\_detail\_collector\_samp\_id,   
dis\_header\_sdate, dis\_header\_stime, Dis\_header\_start\_depth  
having count(1) > 1) d  
where e.mission\_descriptor = d.mission\_descriptor  
and e.dis\_detail\_collector\_samp\_id = d.dis\_detail\_collector\_samp\_id  
and e.dis\_header\_sdate = d.dis\_header\_sdate  
and e.dis\_header\_stime = d.dis\_header\_stime  
and e.Dis\_header\_start\_depth = d.Dis\_header\_start\_depth;

This script pulled all relevant Winkler oxygen data from BIOCHEM after it was previously established that electrode and probe data were too unreliable to be included.

The first step was to determine the actual units of the data within the historical set. This was done using Lazin’s ranges for oxygen units.

0-14 is assumed to be ml/l 50-105 is assumed to be % stauration 105-400 is assumed to be mmol/m3 14-90 is considered suspect as it is possible these values could be recorded in multiple units Values less than 0 were thrown out due to them being outside of possible data range

Data was grouped by potential unit.

mll <- length(data$DIS\_DETAIL\_DATA\_VAL[data$DIS\_DETAIL\_DATA\_VAL < 14])  
mmolm <- length(data$DIS\_DETAIL\_DATA\_VAL[data$DIS\_DETAIL\_DATA\_VAL > 105])  
unkn <- length(data$DIS\_DETAIL\_DATA\_VAL[data$DIS\_DETAIL\_DATA\_VAL >= 14 & data$DIS\_DETAIL\_DATA\_VAL <= 105])  
err <- length(data$DIS\_DETAIL\_DATA\_VAL[data$DIS\_DETAIL\_DATA\_VAL < 0])

Data was then checked for duplicate using a file pulled by Bond from SQL to identify duplicates in metadata and flags were placed in data set if data values were exactly duplicated.

dupes <- read\_xlsx('D:/DATA/Dec27-Winkler/Emily\_dupe\_sampids\_metadata.xlsx')  
  
#flag duplicated data values  
tf <- duplicated(dupes$DIS\_DETAIL\_DATA\_VAL)  
for (i in 1:length(dupes$DIS\_DETAIL\_DATA\_VAL)){  
  
if (tf[i] == T){  
 dupes$flag[i] <- 4  
}  
  
}

Geographical limitations were then set on data based on area of interest to the specific study. Records outside the boundaries of -72:-48 Latitude and 37.5:48 Longitude were then flagged “5” in the data set.

geolim <- data$DIS\_DETAIL\_DATA\_VAL[data$DIS\_HEADER\_SLON > -72 & data$DIS\_HEADER\_SLON < -48 & data$DIS\_HEADER\_SLAT > 37.5 & data$DIS\_HEADER\_SLAT < 48]  
  
  
for (i in 1:length(data$DIS\_DATA\_NUM)){  
 if (!(data$DIS\_HEADER\_SLON[i] > -72 & data$DIS\_HEADER\_SLON[i] < -48 & data$DIS\_HEADER\_SLAT[i] > 37.5 & data$DIS\_HEADER\_SLAT[i] < 48)){  
 data$DIS\_DETAIL\_DATA\_QC\_CODE[i] <- 5 #outside geographical limits  
   
 }  
}

Check for start and end dates within correct boundaries.

#check start/end dates  
#master mission list dates are in weird format, difficult to compare (YYYYMMDD - as numeric)  
#convert to readable date  
#dd <- as.character(master$startdate)  
#as.Date(dd, "%Y%m%d")  
  
  
#sink(file = 'date\_check\_3.txt')  
  
mismatch <- list()  
for (i in 1:length(data$MISSION\_DESCRIPTOR)){  
 g <- grep(master$CR\_NUMBER, pattern = data$MISSION\_DESCRIPTOR[i])  
 if (length(g) > 0){  
 sdate<- as.character(master$START\_DATE[g])  
 sdate <- as.Date(sdate, "%Y%m%d")  
 edate<- as.character(master$END\_DATE[g])  
 edate <- as.Date(edate, "%Y%m%d")  
 if (!(sdate <= as.Date(data$DIS\_HEADER\_SDATE[i]) & as.Date(data$DIS\_HEADER\_SDATE[i]) <= edate)){  
mismatch[i] <- data$MISSION\_DESCRIPTOR[i]  
   
 }  
   
 }  
}  
  
  
length(unique(unlist(mismatch)))  
  
  
#sink()