WQP Parameter and Method exploration

Harmonizing disparate data

The data from the water quality portal includes a wide range of methods and characteristic names. For example in the "chlorophyll" this can be chlorophyll a, b, or both and retrieved using a variety of methods. To know which methods and characteristic names to keep and use, we must first get a better understanding of the type of data we have.

Here we are harmonizing the entirety of the water quality portal data even though the vast majority of these sites will not be landsat visible. The computation time to do it for a few extra million samples is not onerous and the intermediate mostly harmonized full dataset will likely be useful for other uses.

We'll start with the easiest first. Secchi depth

Secchi depth

Secchi Table

In many ways, the secchi disk depth measurement is the easiest water quality parameter to harmonize, because there is really only one method for measuring secchi disk depth (it's in the name after all), and there should always be units of depth (m, ft, inches, cm, etc…). So to harmonize secchi depth measurements we simpy drop all units that are not units of depth and convert all units to a single kind with a lookup table.

```
#Read in the raw data from '1_wqdata/out'
secchi <- read_feather('1_wqdata/tmp/wqp/all_raw_secchi.feather') %>%
  wqp.renamer() %>%
  #Remove trailing white space in labels
  mutate(units = trimws(units))

#Summarize by characteristic name and unit code and print
secchi %>%
  group_by(parameter,units) %>%
  summarize(count=n()) %>%
  knitr::kable()
```

parameter	units	count
Depth, Secchi disk depth	cm	116756
Depth, Secchi disk depth	deg C	1
Depth, Secchi disk depth	deg F	6
Depth, Secchi disk depth	ft	386537
Depth, Secchi disk depth	ft/sec	20
Depth, Secchi disk depth	in	55105
Depth, Secchi disk depth	m	1736912
Depth, Secchi disk depth	mg	187
Depth, Secchi disk depth	mi	1
Depth, Secchi disk depth	NA	12043
Depth, Secchi disk depth (choice list)		428
Depth, Secchi disk depth (choice list)	ft	1
Depth, Secchi disk depth (choice list)	m	208
Depth, Secchi disk depth (choice list)	None	13
Depth, Secchi disk depth (choice list)	NA	39281

parameter	units	count
Secchi Reading Condition (choice list)	None	643
Secchi Reading Condition (choice list)	NA	864
Water transparency, Secchi disc	in	559

Secchi disharmony

Now that we can see all the units we have we can drop non-depth units and make a lookup table to convert all units to meters.

```
#Create a lookup table of units and conversion factors that we want to keep
secchi.lookup <- tibble(units=c('cm','ft','in','m','mi'),
                        conversion = c(0.01,.3048,0.0254,1,1609.34))

# Do an anti_join to these units so that all units that aren't kept can be highlighted and displayed
secchi.disharmony <- secchi %>%
  anti_join(secchi.lookup,by='units') %>%
  group_by(units) %>%
  summarize(count=n())


secchi.disharmony %>%
  knitr::kable(.,caption='The following secchi measurements were dropped because the units do not make s
```

Table 2: The following secchi measurements were dropped because the units do not make sense

units	count
	428
deg C	1
deg F	6
ft/sec	20
mg	187
None	656
NA	52188

Secchi harmony in meters

```
#Join secchi by unit name and then multiply by conversion factor to get meters
secchi.harmonized <- secchi %>%
  inner_join(secchi.lookup,by='units') %>%
  mutate(harmonized_parameter = 'secchi',
         harmonized_value=value*conversion,
         harmonized_unit='meters')
```

Next easiest is TSS

TSS

This paper is really useful for exploring this data. In this paper, the USGS directly compares estimates of Suspended Sediment Concentration (SSC) and Total Suspended Solids (TSS). The primary difference between these methods, as laid out in this paper, is that SSC estimates the mass of suspended solids in a sample volume, by drying out the entire sample without subsampling the water volume. TSS methods often involve some form of subsampling of the total water volume. The paper highlights that while many estimates of TSS and SSC are essentially the same, samples with high sand content show systematic bias in TSS estimates. For our purposes, we have no apriori way to distinguish samples with high or low sand, so we have made the choice to assume that measurements of SSC and TSS are, over the bulk of samples, the same. We use the term "TSS" from here on to describe this data that is both SSC and TSS.

```r
#Read in the raw data from '1_wqdata/out'
tss <- read_feather('1_wqdata/out/wqp/all_raw_tss.feather') %>%
  wqp.renamer() %>%
  #Remove trailing white space in labels
  mutate(units = trimws(units))

#Summarize by characteristic name and unit code
tss %>%
  group_by(parameter,units) %>%
  summarize(count=n()) %>%
  knitr::kable()
```

parameter	units	count
Fixed suspended solids	mg/l	220791
Fixed suspended solids	NA	9357
Suspended sediment concentration (SSC)	%	750778
Suspended sediment concentration (SSC)	mg/l	12041
Suspended sediment concentration (SSC)	NA	3496
Suspended Sediment Concentration (SSC)	%	6758
Suspended Sediment Concentration (SSC)	g/l	7
Suspended Sediment Concentration (SSC)	mg/l	1186925
Suspended Sediment Concentration (SSC)	NA	5428
Total suspended solids		35
Total suspended solids	%	5072
Total suspended solids	count	1
Total suspended solids	kg	29
Total suspended solids	mg/l	2855101
Total suspended solids	None	16
Total suspended solids	NTU	1
Total suspended solids	ppm	1680
Total suspended solids	tons/day	529
Total suspended solids	ug/l	478
Total suspended solids	NA	235192

TSS disharmony

As with secchi disk depth, we expect certain units to be associated with total suspended solids or suspended sediment concentration. These include mass per volume measurements like: mg/l, g/l, ug/l and others.

TSS does come with one less obvious parameter which is %. Any sample with a % unit is most commonly a sample where suspended sediments were split into particle size fractions. The relative proportion of clay, silt, and sand can have important impacts on the reflectance properties of water, so this is a useful parameter to keep, though it will require some exploration, using the additional data column that we relabeld as "particle_size."

TSS particle size fractionation

The table below shows all of the various particle fraction categories held within the TSS category. About half of the total observations (760,000) that use "%" as a unit are actually estimating the fraction of particles that are smaller than sand (<0.0625). The rest of the particle fractionation size classes are spread across *29* other particle fractions. This leaves us with a difficult choice. If we kept all of this data, we would widen our final dataset by 29 rows, with very few likely overpasses in a dataset of less than 80k observations per fraction category before checking for sites that are Landsat visible and were collected on relatively cloud free days. If we throw away all of the % data, we use valuable information that may help explain variability between sites with similar TSS but different reflectance values based on the particle size fractionation. Here, we will opt for an intermediate approach and keep only the > 300,000 observations that simply describe the fraction of sand in a sample (<0.0625 mm).

```r
#Select only units for %
tss.p <- tss %>%
  filter(units == '%')

#look at the breakdown of particle sizes
tss.p %>%
  group_by(particle_size) %>%
  summarize(count=n()) %>%
  knitr::kable()
```

particle_size	count
< 0.001 mm	656
< 0.002 mm	33644
< 0.004 mm	45670
< 0.008 mm	25665
< 0.016 mm	44248
< 0.031 mm	24366
< 0.0625 mm	337798
< 0.062mm	172
< 0.063 mm	15
< 0.09 mm	86
< 0.125 mm	81000
< 0.18 mm	86
< 0.25 mm	72412
< 0.355 mm	80
< 0.5 mm	55959
< 0.71 mm	19
< 1 mm	23456
< 1.4 mm	1
< 128 mm	15
< 16 mm	18
< 2 mm	5340
< 256 mm	15
< 3.35 mm	2
< 31.5 mm	1
< 4 mm	180
< 63 mm	15

particle_size	count
< 8 mm	31
sands	1137
silts and clays	1140
NA	9381

```
#Keep only the sand fraction data (~50% of the data)
sand_harmonized  <- tss.p %>%
  filter(particle_size %in%  c('< 0.0625 mm','sands')) %>%
  mutate(conversion=NA,
         harmonized_parameter='p.sand',
         harmonized_value=value,
         harmonized_unit='%')
```

TSS dropping bad units

Now that we have split out the TSS values that had "%" units, we can deal with and drop the more nonsensical or missing units. The table below will also print out the number of "%" observations that we drop, but, remember, we kept about half of these in the above code.

Here we will convert all remaining sediment values to units of mg/L and drop any non mass/volume units.

```
#Make a tss lookup table
tss.lookup <- tibble(units=c('mg/l','g/l','ug/l','ppm'),
                        conversion = c(1,1000,1/1000,1))



tss.disharmony <- tss %>%
  anti_join(tss.lookup,by='units') %>%
  filter(!particle_size %in% c('< 0.0625 mm','sands')) %>%
  group_by(units) %>%
  summarize(count=n())




knitr::kable(tss.disharmony,caption='The following TSS measurements were dropped because the units do n
```

Table 5: The following TSS measurements were dropped because the units do not make sense

units	count
	35
%	423673
count	1
kg	29
None	16
NTU	1
tons/day	529
NA	253418

TSS harmony in mg/l

Now we can convert all TSS measurements to untis of 'mg/l.' We do need to do one final splitting of the data because there is another parameter name called "Fixed suspended solids." Fixed suspended solids are essentialy the inorganic component of a sediment sample that remains after kiln drying at 550°F. We will relable these as a harmonized parameter 'Total inorganic sediment' or tis.

```
#Join to the lookup table and harmonize units

tss.harmonized <- tss %>%
  inner_join(tss.lookup,by='units') %>%
  mutate(harmonized_parameter = 'tss',
         harmonized_value=value*conversion,
         harmonized_unit='mg/l') %>%
  #Change harmonized parameter to tis for parameter "fixed suspended solids"
  mutate(harmonized_parameter = ifelse(parameter == 'Fixed suspended solids','tis',harmonized_parameter)
```

DOC

Didn't keep enough columns to really do this. Need to add resultsampletext and a few others. Otherwise total carbon can include fish biomass. Which is not what we are talking about

Dissolved organic carbon is a much more complex series of parameters, methods, and units. As with TSS we generally expect these to be in units of mass per unit volume, but we have many more possible variations of methods used to extract DOC values.

First let's look at the total counts for parameter unit combinations

```
#Summarize by characteristic name and unit code
doc <- read_feather('1_wqdata/out/wqp/all_raw_doc.feather') %>%
  wqp.renamer() %>%
  #Remove trailing white space in labels
  mutate(units = trimws(units))


doc %>%
  group_by(parameter,units) %>%
  summarize(count=n()) %>%
  knitr::kable(.,caption='Carbon parameter names, units, and observation counts')
```

Table 6: Carbon parameter names, units, and observation counts

parameter	units	count
Non-purgeable Organic Carbon (NPOC)	mg/l	1393
Organic carbon	%	28734
Organic carbon	% by wt	2682
Organic carbon	% recovery	12
Organic carbon	count	1
Organic carbon	g/kg	8145
Organic carbon	mg/g	571
Organic carbon	mg/kg	3436
Organic carbon	mg/l	2028971
Organic carbon	None	762
Organic carbon	ppm	5618
Organic carbon	ug/g	67

parameter	units	count
Organic carbon	ug/kg	2
Organic carbon	ug/l	627
Organic carbon	NA	26879
Total carbon	%	930
Total carbon	% by wt	1457
Total carbon	g/kg	6
Total carbon	g/m2	6
Total carbon	mg/g	14
Total carbon	mg/kg	518
Total carbon	mg/l	14859
Total carbon	NA	28

DOC disharmony

DOC percent values

Once again we have quite a few observations of 'Organic carbon' and 'Total carbon' that are in units of % which is a perplexing unit without some more context. Let's examine these values a little more.

```
doc.p <- doc %>%
  filter(units=='%')
```

Hardest

Chlorophyll

```
#Read in the raw data from '1_wqdata/tmp'
chl <- read_feather('1_wqdata/out/wqp/all_raw_chlorophyll.feather')
```