Filtering water quality dataset

Amber Lee

6/15/2021

Unfinished

Is it okay if, after filtering/processing the original data, we only have 105,267 rows? We are only keeping about half of the original data…

Summary

One important goal before we interpolate missing values is to filter the data, to decide on the scope of our analysis. Filtering the data to what is necessary will also reduce the amount of interpolation needed. We will see that removing duplicate rows will greatly reduce the missing values (that aren't actually missing), thus making the interpolation step significantly easier.

In cleaning the LTRM Water Quality dataset, we encounter the following questions:

- Which variables (columns) are the most important for us to keep (out of the 133 total variables)?
- Why do duplicate rows happen, and how do we deal with them?
- Which samples (rows) are high enough quality for us to keep? (This involves the QF codes.)

The LTRM Water Quality dataset has duplicate rows for the same `SHEETBAR`, which is problematic because `SHEETBAR` is a unique identifier for a water data sheet (sample at a date, time, and location).

```r
library(tidyverse)
library(ggplot2)
library(lubridate)
library(corrplot)
library(RColorBrewer)
library(kableExtra)
```

```r
# set working directory to source file location
# setwd("~/Documents/GitHub/UMR-TDA-2021")

water20 <- read.csv(file = "pools EDA/pool data/ltrm_water_data_lat_long.csv")
```

Important variables

```r
water_var <- c('TN','TP','TEMP','DO','TURB',
               'COND','VEL','SS','WDP','CHLcal','SECCHI')

waterQF_var <- paste(water_var, "QF", sep = "")

identifier_var <- c('SHEETBAR', 'DATE', 'LATITUDE', 'LONGITUDE', 'FLDNUM', 'STRATUM', 'LOCATCD')
```

We decided that the 11 continuous variables of importance were: total nitrogen, total phosphorous, temperature, dissolved oxygen, turbidity, water condition, velocity, suspended solids, water depth, chlorophyll-a, and Secchi distance.

In addition, we will want to include the QA/QC codes, along with identifier variables like `SHEETBAR` and date.

Lastly, we manually edit these variable strings because:

- The water depth variable is `WDP`, but the corresponding quality factor is `ZMAXQF` rather than `WDPQF`.
- `CHLcal`, calibrated fluorometric chlorophyll a, does not have a corresponding quality factor code. According to the metadata: "`CHLcal` is generated by calibration of fluorometric chlorophyll readings (`CHLF`) to season and year specific measurements of spectrophotometric chlorophyll (`CHLS`). Data from sites where CHLS and CHLF are both collected are used to build river-specific calibration curves for these data. Values are corrected for pheophytin. Units are micrograms per liter.“

```
waterQF_var <- waterQF_var[waterQF_var != "WDPQF" &
                             waterQF_var != "CHLcalQF"]

waterQF_var <- c(waterQF_var, "ZMAXQF")
```

Duplicate rows

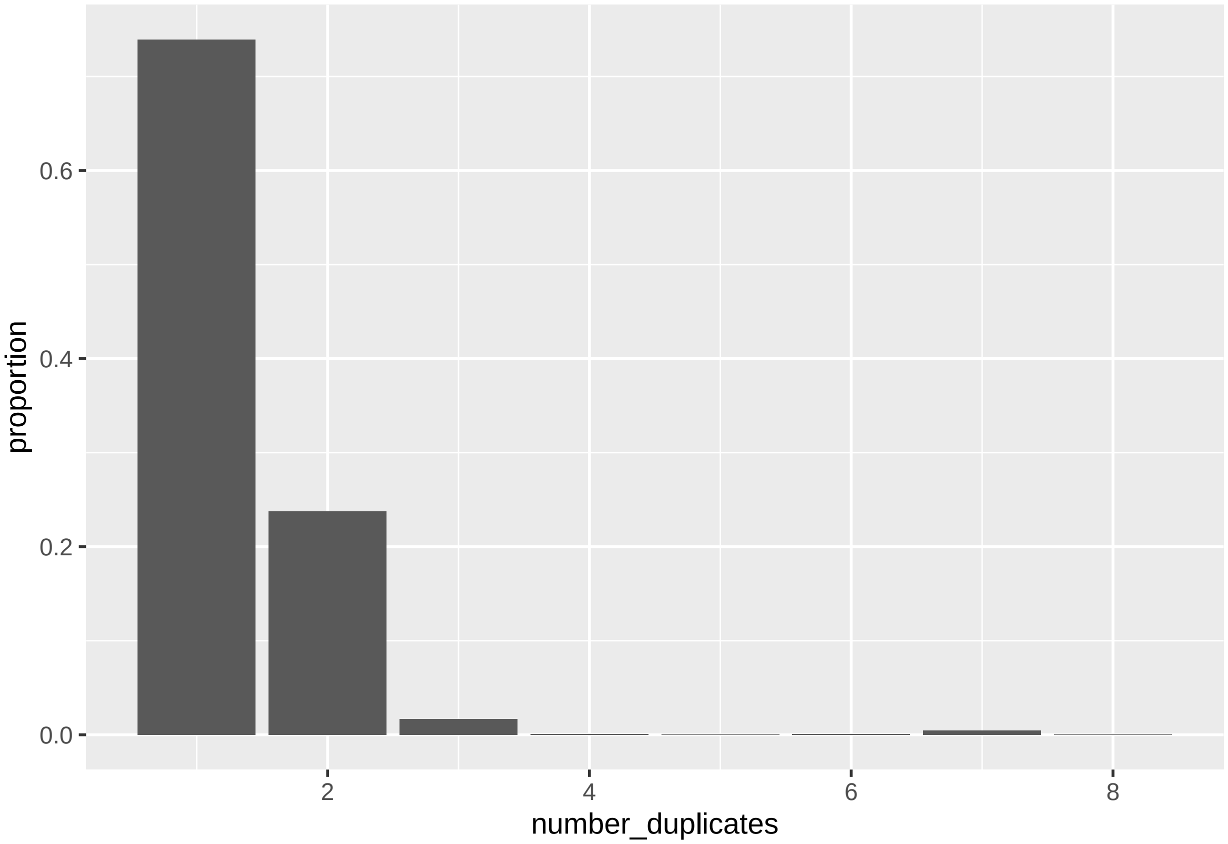
There are 204305 total rows in the LTRM water quality dataset. Of these rows, there are 156474 distinct `SHEETBAR` codes.

We visualize duplicates as follows. To identify the duplicate rows, we count the number of occurences of each unique `SHEETBAR` value in the dataset. Then, we can calculate and plot the distribution of `SHEETBAR` duplicates.

```
duplicates <- water20 %>%
  select(SHEETBAR) %>%
  group_by(SHEETBAR) %>%
  summarize(count = n())

duplicates %>% head()
```

```
## # A tibble: 6 x 2
##   SHEETBAR count
##      <int> <int>
## 1 -4604348     1
## 2 -4604347     2
## 3 -4604346     2
## 4 -4604345     2
## 5 -4604344     1
## 6 -4604343     2
```

```
count_n_duplicates <- function(n, df) {
  return((df %>% filter(count == n) %>% dim())[1]/156474) #156k distinct sheetbars
}

count_duplicates <- data.frame(proportion = sapply(1:8, count_n_duplicates,
                                                   duplicates),
                               number_duplicates = 1:8)

ggplot(count_duplicates, aes(x = number_duplicates, y = proportion)) +
  geom_bar(stat = "identity")
```



```
count_duplicates %>% kbl(booktabs = T)
```

proportion	number_duplicates
0.7393497	1
0.2377072	2
0.0166481	3
0.0006199	4
0.0002365	5
0.0007605	6
0.0046781	7
0.0000000	8

The proportion of SHEETBARs with at least one duplicated row is0.2606503 (representing about 47,000 rows). When do duplicated rows occur?

We look at two SHEETBARs with duplicated rows.

```
water20 %>%
  filter(SHEETBAR == -4604347) %>%
  select(SHEETBAR, Z, CALCZCD, DO, TP, TN) %>%
  kbl(booktabs = T) %>%
  kable_styling(latex_options = "striped")
```

```
water20 %>%
  filter(SHEETBAR == 41015929    ) %>%
  select(SHEETBAR, Z, CALCZCD, DO, TP, TN) %>%
  kbl(booktabs = T) %>%
```

SHEETBAR	Z	CALCZCD	DO	TP	TN
-4604347	0.2	SF	7.4	0.265	7.49
-4604347	4.8	BT	7.5	NA	NA

SHEETBAR	Z	CALCZCD	DO	TP	TN
41015929	0.2	SF	14.4	0.075	2.557
41015929	1.0	OT	14.8	NA	NA
41015929	2.0	OT	14.9	NA	NA
41015929	3.0	OT	15.0	NA	NA
41015929	4.0	OT	15.2	NA	NA
41015929	4.6	BT	15.2	NA	NA

```
  kable_styling(latex_options = "striped")
```

Here, we see that `TP` and `TN`, total phosphorous and total nitrogen, are measured only at the surface level (when `CALCZCD == "SF"`). The variable `CALCZCD` is a categorical variable with levels surface, middle, bottom, and other. It is calculated with the sample depth and the total water depth (of the river site).

In contrast, dissolved oxygen `DO` is measured at various depths (denoted by `Z`) because different parts of the water column have different levels of `DO`. It would be inappropriate to average the dissolved oxygen levels because they were taken at different sample depths.

Thus, this missing values of `TP` and `TN` are occuring at different sample depths at the same sampling site. These missing values aren't *really* missing values; they would be redundant to interpolate.

We can reasonably keep only the samples taken at the surface level

We decided to filter for rows that were labelled as surface level,`CALCZCD == "SF"`. Implicitly, this filtering step removes samples for which the sample depth is missing.

```
table(water20$CALCZCD)
```

```
##
##              BT     MD     OT     SF
##    8436   34991   2971  10736 147171
```

More than 70% of the samples were taken on the surface level. Of these measurements taken at the surface level, the eleven important continuous variables (in `water_var`) were recorded with a recording rate of at least 50%. This is a good sanity check because `TN` and `TP` are never recorded in the middle and bottom water depths. we checked to see the recording rate of the eleven important variables and found recording rates greater than 50%.

```
filterwater <- water20 %>%
  filter(CALCZCD == "SF") %>%
  select(all_of(water_var))

sapply(filterwater, function(x) sum(is.na(x)/length(x)))
```

```
##         TN         TP       TEMP         DO       TURB       COND        VEL
## 0.49246115 0.50078480 0.01157837 0.01390220 0.01440501 0.01510488 0.39597475
##         SS        WDP     CHLcal     SECCHI
## 0.17570717 0.07648925 0.26196058 0.09751921
```

What if CALCZCD is missing?

```
(water20 %>%
  filter(CALCZCD == "") %>%
  dim())[1]
```

```
## [1] 8436
```

There are are about 8000 samples with missing `CALCZCD`.

But this is okay because when `CALCZCD` is missing, nearly all of our continuous variables are missing too.

```
sapply((water20 %>%
  filter(CALCZCD == "") %>%
  select(all_of(water_var))), function(x) sum(is.na(x)/length(x)))
```

```
##        TN        TP      TEMP        DO      TURB      COND       VEL        SS
## 0.9998815 0.9998815 0.9992888 0.9992888 0.9996444 0.9992888 0.9997629 0.9996444
##       WDP    CHLcal    SECCHI
## 0.8488620 0.9996444 0.9665718
```

When `CALCZCD` isn't recorded, the rest of the variables aren't recorded. Since these samples represent 0.041 percent of the original data, we decide that it is okay to exclude these observations.

Water column variables

What is the difference between WDP, ZMAX, CALCZCD, SAMPZCD? (Could be answered later because not all these variables will be used in our analysis)

Filter for QA/QC

Combine

```
filterwater <- water20 %>%
  filter(CALCZCD == "SF") %>%
  select(all_of(c(identifier_var, water_var, waterQF_var))) %>%
  filter_at(vars(c("TURBQF","TEMPQF","DOQF","VELQF",
                 "ZMAXQF","SECCHIQF","CONDQF")),
          any_vars(. != "A" | . != 0)) %>%
  # forrest, edit this
  filter_at(vars(c("TNQF","TPQF","SSQF")),
          any_vars(. != 8 | . != 64))

filterwater %>% head() %>%
  kbl(booktabs = T) %>%
  kable_styling(latex_options = "striped")
```

```
dim(filterwater)
```

```
## [1] 105267     28
```

use this tutorial about `filter_at` https://suzan.rbind.io/2018/02/dplyr-tutorial-3/

```
write.csv(filterwater, "water_data_filtered.csv", row.names = FALSE)
```

SHEETBAR	DATE	LATITUDE	LONGITUDE	FLDNUM	STRATUM	LOCATCD	TN	TP	TEM
-4203181	12/03/1991	43.84134	-91.25233	2	NA	BK01.0N	1.82	0.083	-0
-4103818	02/13/1993	44.39877	-92.05263	1	NA	M761.5E	3.20	0.123	0
-4103809	02/12/1993	44.42327	-92.13400	1	NA	M766.0I	3.92	0.152	0
-4102507	02/06/1992	44.32732	-91.94685	1	5	M753.2S	5.27	0.114	0
-4102504	02/06/1992	44.39194	-91.98051	1	3	M758.6Y	4.85	0.101	0
-4102501	02/06/1992	44.39877	-92.05263	1	1	M761.5E	5.36	0.134	0