Data Screening

PAZ

06/04/2017

Introduction

This Data Screenining notebook follows the GUide to STatistical Analysis in Microbial Ecology (GUSTA ME). The purpose is to inspect the variables that we'll be using to test for hypotheses later on, and check whether they follow typical assumptions made in parametric tests such as normality, freedom from heteroskedasticity (difference in variability btw. two+ variables) and outliers.

Reference:

https://sites.google.com/site/mb3gustame/home Buttigieg PL, Ramette A (2014) A Guide to Statistical Analysis in Microbial Ecology: a community-focused, living review of multivariate data analyses. FEMS Microbiol Ecol. 90: 543-550.

Packages

```
library(sm)
library(vioplot)

library(dplyr)
library(ggplot2)
library("ggrepel")
```

Missing values

1. Missing chemical and isotope data due to machine failure or automatic sampling servicing program.

These have been considered to be Values Missing Completely at Random (MCAR) as they are associated to the end of the automatic sampler's capacity for a certain number of events where servicing was inadequate for the discharge amounts seen during a sampling week. Here the values' missingess is not related to any other value in the data set.

2. Isotope data for both soil and water samples due to concentration value being below the limit of detection.

These values must be considered to be Missing at Random (MAR) as the missing value has no relation to the value that 'should' be there, but does depend on other variables in the data set. Thus, other variables must be taken into account for MAR data to be considered random (i.e. missing data is "conditioned by" other data in the data set).

Import soils

Convert to single time observation for merging with water observation.

```r
# Soils
soils = read.csv2("Data/MassBalance_R.csv",
                        na.strings=c('#DIV/0!', '', 'NA'), header = TRUE)
colnames(soils)[colnames(soils) == "ti"] <- "Date.ti"
soils$Date.ti <- as.POSIXct(strptime(soils$Date.ti,
                                             "%Y-%m-%d %H:%M", tz="EST")) # csv typos, option 1
sum(is.na(soils$Date.ti)) == 0
```

```
## [1] TRUE
```

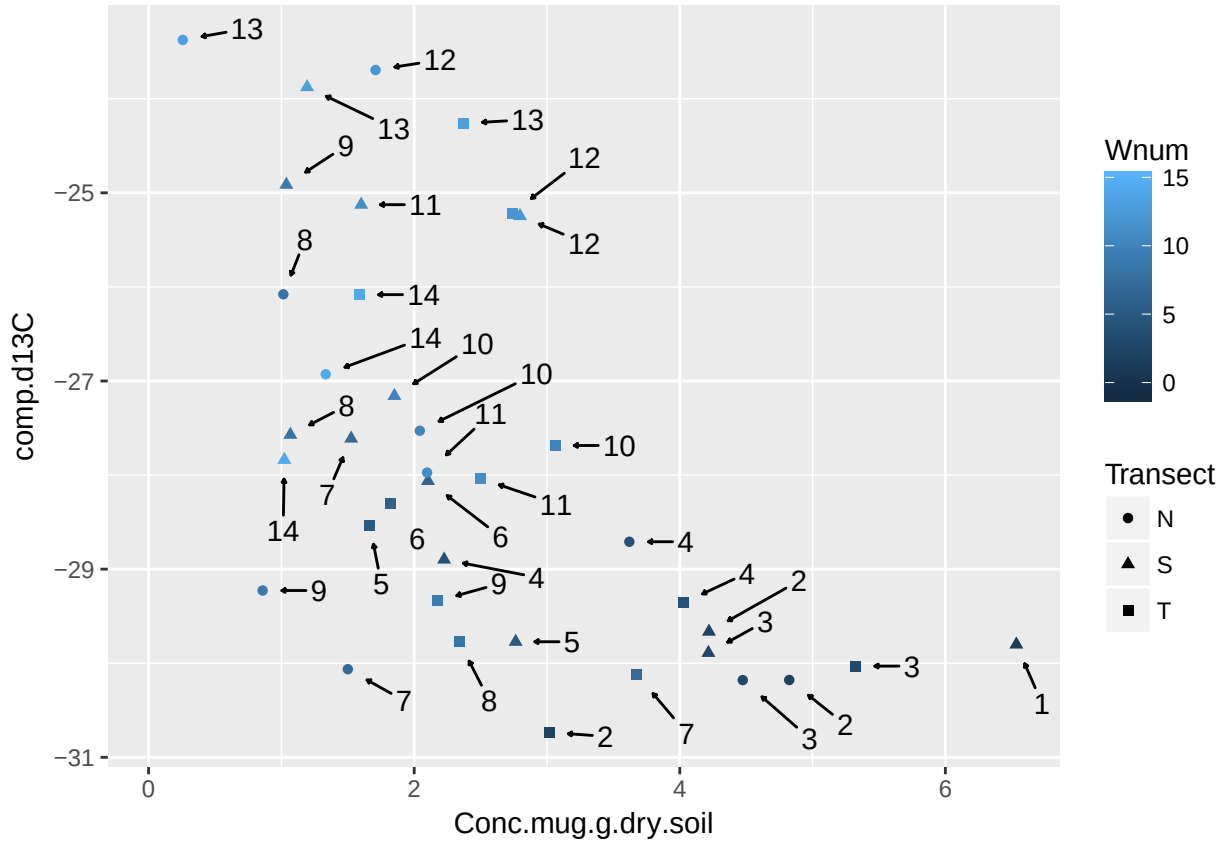
```r
dropSoil <- c("WeekSubWeek", "Event",
              "B.diss", "B.filt", "CumOutDiss.g", "CumOutFilt.g", "CumOutAppMass.g",  "CumOutMELsm.g",
              "CumAppMass.g",
              "ID.N", "ID.T",  "Area.N", "Area.T", "Area.S",
              "comp.d13C.SE.North", "comp.d13C.SE.Talweg", "comp.d13C.SE.South",
              "f.max.comp", "f.mean.comp", "f.min.comp", "ngC.SD","ngC.SE", "N_compsoil", "N_ngC")
soils <- soils[ , !(names(soils) %in% dropSoil)]


# Quasi-Molten SOILS
soilGroups = read.csv2("Data/WeeklySoils_Rng.csv",
                        na.strings=c('#DIV/0!', '', 'NA'), header = TRUE)
soilGroups$Date.ti <- as.POSIXct(strptime(soilGroups$Date.ti,
                                             "%Y-%m-%d %H:%M", tz="EST")) # csv typos, option 1
sum(is.na(soilGroups$Date.ti)) == 0
```

```
## [1] TRUE
```

```r
str(soils)
```

```
## 'data.frame':    52 obs. of  23 variables:
##  $ Date.ti            : POSIXct, format: "2016-03-25 00:04:00" "2016-03-25 12:04:00" ...
##  $ B.mean.comp.North  : num  NA NA NA NA NA ...
##  $ B.max.comp.North   : num  NA NA NA NA NA ...
##  $ B.min.comp.North   : num  NA NA NA NA NA ...
##  $ MassSoil.g.North   : num  12.6 NA NA 613.1 NA ...
##  $ comp.d13C.North    : num  NA NA NA NA NA ...
##  $ comp.d13C.SD.North : num  NA NA NA NA NA ...
##  $ B.mean.comp.Talweg : num  NA NA NA NA NA ...
##  $ B.max.comp.Talweg  : num  NA NA NA NA NA ...
##  $ B.min.comp.Talweg  : num  NA NA NA NA NA ...
##  $ MassSoil.g.Talweg  : num  4.44 NA NA 173.27 NA ...
##  $ comp.d13C.Talweg   : num  NA NA NA NA NA ...
##  $ comp.d13C.SD.Talweg: num  NA NA NA NA NA ...
##  $ B.mean.comp.South  : num  NA NA NA NA NA ...
##  $ B.max.comp.South   : num  NA NA NA NA NA ...
##  $ B.min.comp.South   : num  NA NA NA NA NA ...
##  $ MassSoil.g.South   : num  18.8 NA NA 2112.1 NA ...
##  $ comp.d13C.South    : num  NA NA NA NA NA ...
##  $ comp.d13C.SD.South : num  NA NA NA NA NA ...
##  $ ID.S               : Factor w/ 17 levels "AW-S-0","AW-S-0x",..: 2 NA NA 1 NA NA 3 NA NA 10 ...
##  $ CatchMassSoil.g    : num  35.8 NA NA 2898.5 NA ...
##  $ BulkMass.g         : num  14.1 NA NA 1183.7 NA ...
##  $ BulkCatch.d13      : num  NA NA NA NA NA ...
```

```
ggplot(soilGroups, aes(x=Conc.mug.g.dry.soil, y=comp.d13C))+
  geom_point(aes(group = Transect, colour = Wnum, shape = Transect))+
  geom_text_repel(aes(label=Wnum),
                  arrow = arrow(length = unit(0.005, 'npc'), type = "closed"),
                  force = 1,
                  point.padding = unit(1.0, 'lines'),
                  max.iter = 2e3,
                  nudge_x = .2)
```




```
  #stat_smooth(method = "lm", formula = y ~ poly(x, 2)) +
  #stat_smooth(method = "lm", formula = y~x, se=F)
```

Correlation Soils

```
cor.test(soilGroups$comp.d13C, soilGroups$Conc.mug.g.dry.soil)
```

```
##
##  Pearson's product-moment correlation
##
## data:  soilGroups$comp.d13C and soilGroups$Conc.mug.g.dry.soil
## t = -4.2655, df = 36, p-value = 0.0001379
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7585899 -0.3187871
## sample estimates:
```

```
##        cor
## -0.5794211
```

Soils to confirm in IRMS:

North: 13, 12, 10, 9, 11, 7 Talweg: 13, 10, 9, 8, 7 South: 13

No isotopes: Talweg: 1, 15 North: 1, 5, 6, 15 South: 15

Repeat: T-1, T-13, T-15 N-1, N-13, N-12, N-6, N-5 S-13, S-9

Import water

```
waters = read.csv2("Data/WeeklyHydroContam_R.csv")
waters$ti <- as.POSIXct(strptime(waters$ti, "%Y-%m-%d %H:%M", tz="EST"))
colnames(waters)[colnames(waters) == "ti"] <- "Date.ti"
waters$Events <- factor(waters$Events, levels = unique(waters$Events))
waters$Event <- factor(waters$Event, levels = unique(waters$Event))

dropWater <- c("N.x", "N.y",
               "Markers" , "TimeDiff",
               "se.d13C", "MES.mg.L", "MES.sd", "MO.mg.L", "filt.se.d13C", "f.diss", "f.filt",
               "Appl.Mass.g",
               "DissSmeto.mg", "DissSmeto.mg.SD",
               "DissOXA.mg", "DissOXA.mg.SD",
               "DissESA.mg", "DissESA.mg.SD",
               "FiltSmeto.mg", "DissSmeto.mg.SD",
               "TotSMout.mg", "TotSMout.mg.SD",
               "FracDiss", "FracFilt")
waters <- waters[ , !(names(waters) %in% dropWater)]

# Date conversion correct:
sum(is.na(waters$Date.ti)) == 0
```

```
## [1] TRUE
```

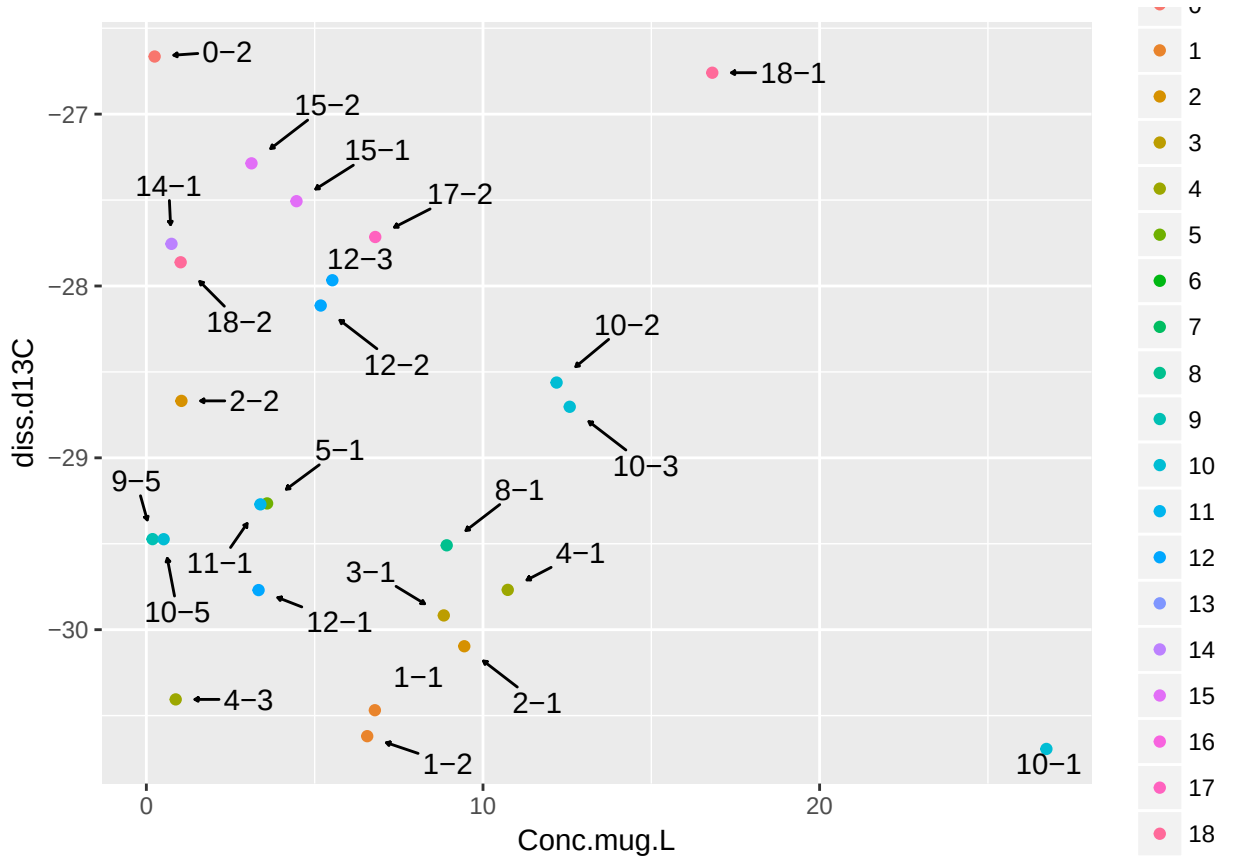
```
str(waters)
```

```
## 'data.frame':    51 obs. of  64 variables:
##  $ Date.ti               : POSIXct, format: "2016-03-25 00:04:00" "2016-03-25 12:04:00" ...
##  $ WeekSubWeek           : Factor w/ 51 levels "W0-0x","W0-1",..: 1 2 3 4 5 6 26 27 28 29 ...
##  $ tf                    : Factor w/ 51 levels "2016-03-25 12:02:00",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ iflux                 : num  1.25 1.12 1.31 1.46 16.33 ...
##  $ fflux                 : num  1.13 1.31 1.46 16.45 15.18 ...
##  $ changeflux            : num  -0.119 0.189 0.148 14.989 -1.15 ...
##  $ maxQ                  : num  1.25 1.38 1.64 38.4 18.67 ...
##  $ minQ                  : num  1.118 1.082 0.929 1.449 13.201 ...
##  $ Duration.Hrs          : num  12 82.5 37.6 27.3 23.1 ...
##  $ chExtreme             : num  -0.13 0.256 0.33 36.944 -3.133 ...
##  $ Peak                  : int  NA NA NA 1 NA NA 2 NA NA 3 ...
##  $ AveDischarge.m3.h     : num  1.2 1.21 1.28 14.32 15.53 ...
##  $ Volume.m3             : num  14.4 100.2 48.3 390.4 359.2 ...
##  $ Sampled.Hrs           : num  12 82.5 37.6 27.3 23.1 ...
##  $ Sampled               : Factor w/ 2 levels "Not Sampled",..: 1 2 1 2 2 1 2 2 1 2 ...
##  $ Conc.mug.L            : num  0.246 0.246 3.517 6.788 6.561 ...
##  $ Conc.SD               : num  0.0193 0.0193 0.1544 0.2894 0.1906 ...
##  $ OXA_mean              : num  4.82 4.82 17.68 30.53 32.49 ...
##  $ OXA_SD                : num  1.141 1.141 5.663 10.185 0.243 ...
```

```
##  $ ESA_mean              : num  18.1 18.1 32 46 41.3 ...
##  $ ESA_SD                : num  3.497 3.497 3.267 3.037 0.853 ...
##  $ diss.d13C             : num  NA -26.7 NA -30.5 -30.6 ...
##  $ SD.d13C               : num  NA 0.936 NA 0.106 0.151 ...
##  $ Conc.Solids.mug.gMES  : num  0.645 0.645 0.385 0.126 0.436 ...
##  $ Conc.Solids.ug.gMES.SD: num  0.0232 0.0232 0.0252 0.0271 0.1232 ...
##  $ filt.d13C             : num  NA NA NA NA NA ...
##  $ filt.SD.d13C          : num  NA NA NA NA NA ...
##  $ DD13C.diss            : num  NA 4.545 NA 0.741 0.59 ...
##  $ DD13C.filt            : num  NA NA NA NA NA ...
##  $ B.diss                : num  NA 93.1 NA 35.4 29.4 ...
##  $ B.filt                : num  NA NA NA NA NA ...
##  $ NH4.mM                : num  NA NA NA 0.05 NA NA NA NA NA NA ...
##  $ TIC.ppm.filt          : num  NA NA NA 51.8 44.8 NA 66.7 52.1 NA 69.4 ...
##  $ Cl.mM                 : num  NA NA NA 1.48 1574 ...
##  $ NO3...mM              : num  NA NA NA 616 778 ...
##  $ PO4..mM               : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ NPOC.ppm              : num  NA NA NA 4 4.4 NA 5.8 3.4 NA 9.1 ...
##  $ TIC.ppm.unfilt        : num  NA NA NA 44.8 26.4 NA 39 32.3 NA 54.8 ...
##  $ TOC.ppm.unfilt        : num  NA NA NA 4.7 5.4 NA 2.7 3.8 NA 3.9 ...
##  $ ExpMES.Kg             : num  5.35 5.35 14.88 24.4 8.08 ...
##  $ CumAppMass.g          : num  6369 6369 6369 6369 6369 ...
##  $ DissSmeto.g           : num  0.00354 0.0246 0.17004 2.64991 2.357 ...
##  $ DissSmeto.g.SD        : num  0.000278 0.001934 0.007463 0.11298 0.068486 ...
##  $ DissOXA.g             : num  0.0695 0.4832 0.8547 11.9184 11.6727 ...
##  $ DissOXA.g.SD          : num  0.0165 0.1143 0.2738 3.976 0.0873 ...
##  $ DissESA.g             : num  0.26 1.81 1.55 17.95 14.83 ...
##  $ DissESA.g.SD          : num  0.0504 0.3503 0.158 1.1855 0.3066 ...
##  $ FiltSmeto.mg.SD       : num  0.124 0.124 0.374 0.66 0.996 ...
##  $ FiltSmeto.g           : num  0.00345 0.00345 0.00573 0.00307 0.00352 ...
##  $ FiltSmeto.g.SD        : num  0.000124 0.000124 0.000374 0.00066 0.000996 ...
##  $ TotSMout.g            : num  0.00699 0.02806 0.17577 2.65298 2.36052 ...
##  $ TotSMout.g.SD         : num  0.000216 0.00137 0.005284 0.07989 0.048432 ...
##  $ MELsm.g               : num  0.302 2.078 2.379 30.241 27.008 ...
##  $ MELsm.g.SD            : num  0.0269 0.1868 0.1789 2.4062 0.1634 ...
##  $ CumOutDiss.g          : num  0.00354 0.02815 0.19818 2.84809 5.2051 ...
##  $ CumOutFilt.g          : num  0.00345 0.0069 0.01263 0.01571 0.01923 ...
##  $ CumOutSmeto.g         : num  0.00699 0.03505 0.21082 2.8638 5.22432 ...
##  $ CumOutMELsm.g         : num  0.302 2.38 4.76 35.001 62.009 ...
##  $ BalMassDisch.g        : num  6369 6367 6365 6334 6307 ...
##  $ prctMassOut           : num  4.98e-05 2.00e-04 1.25e-03 1.89e-02 1.68e-02 ...
##  $ FracDeltaOut          : num  0 -0.00533 0 -0.57576 -0.51483 ...
##  $ Events                : Factor w/ 51 levels "0-1","0-2","0-3",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Weeks                 : Factor w/ 16 levels "W0","W1","W10",..: 1 1 1 2 2 2 9 9 9 10 ...
##  $ Event                 : Factor w/ 19 levels "0","1","2","3",..: 1 1 1 2 2 2 3 3 3 4 ...
```

```r
# Conc.mug.L
# TotSMout.g
# MELsm.g
ggplot(waters, aes(x=Conc.mug.L, y=diss.d13C))+
  geom_point(aes(group = Event, colour = Event))+
  geom_text_repel(aes(label=Events),
                  arrow = arrow(length = unit(0.005, 'npc'), type = "closed"),
                  force = 1,
```

```
                point.padding = unit(1.0, 'lines'),
                max.iter = 2e3,
                nudge_x = .2)
```



Correlations Waters

```
cor.test(waters$Conc.mug.L, waters$diss.d13C)
```

```
##
##  Pearson's product-moment correlation
##
## data:  waters$Conc.mug.L and waters$diss.d13C
## t = -1.0794, df = 23, p-value = 0.2916
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5656317  0.1922336
## sample estimates:
##        cor
## -0.2195757
```

```
#cor.test(waters$TotSMout.g, waters$diss.d13C)

#esaoxa <- waters$MELsm.g-waters$TotSMout.g
# cor.test(esaoxa, waters$diss.d13C)
```

Merge Soil and Water data frames

```
WaterSoils <- merge(waters, soils, by = "Date.ti", all = F)
str(WaterSoils)
```

```
## 'data.frame':    51 obs. of  86 variables:
##  $ Date.ti               : POSIXct, format: "2016-03-25 00:04:00" "2016-03-25 12:04:00" ...
##  $ WeekSubWeek           : Factor w/ 51 levels "W0-0x","W0-1",..: 1 2 3 4 5 6 26 27 28 29 ...
##  $ tf                    : Factor w/ 51 levels "2016-03-25 12:02:00",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ iflux                 : num  1.25 1.12 1.31 1.46 16.33 ...
##  $ fflux                 : num  1.13 1.31 1.46 16.45 15.18 ...
##  $ changeflux            : num  -0.119 0.189 0.148 14.989 -1.15 ...
##  $ maxQ                  : num  1.25 1.38 1.64 38.4 18.67 ...
##  $ minQ                  : num  1.118 1.082 0.929 1.449 13.201 ...
##  $ Duration.Hrs          : num  12 82.5 37.6 27.3 23.1 ...
##  $ chExtreme             : num  -0.13 0.256 0.33 36.944 -3.133 ...
##  $ Peak                  : int  NA NA NA 1 NA NA 2 NA NA 3 ...
##  $ AveDischarge.m3.h     : num  1.2 1.21 1.28 14.32 15.53 ...
##  $ Volume.m3             : num  14.4 100.2 48.3 390.4 359.2 ...
##  $ Sampled.Hrs           : num  12 82.5 37.6 27.3 23.1 ...
##  $ Sampled               : Factor w/ 2 levels "Not Sampled",..: 1 2 1 2 2 1 2 2 1 2 ...
##  $ Conc.mug.L            : num  0.246 0.246 3.517 6.788 6.561 ...
##  $ Conc.SD               : num  0.0193 0.0193 0.1544 0.2894 0.1906 ...
##  $ OXA_mean              : num  4.82 4.82 17.68 30.53 32.49 ...
##  $ OXA_SD                : num  1.141 1.141 5.663 10.185 0.243 ...
##  $ ESA_mean              : num  18.1 18.1 32 46 41.3 ...
##  $ ESA_SD                : num  3.497 3.497 3.267 3.037 0.853 ...
##  $ diss.d13C             : num  NA -26.7 NA -30.5 -30.6 ...
##  $ SD.d13C               : num  NA 0.936 NA 0.106 0.151 ...
##  $ Conc.Solids.mug.gMES  : num  0.645 0.645 0.385 0.126 0.436 ...
##  $ Conc.Solids.ug.gMES.SD: num  0.0232 0.0232 0.0252 0.0271 0.1232 ...
##  $ filt.d13C             : num  NA NA NA NA NA ...
##  $ filt.SD.d13C          : num  NA NA NA NA NA ...
##  $ DD13C.diss            : num  NA 4.545 NA 0.741 0.59 ...
##  $ DD13C.filt            : num  NA NA NA NA NA ...
##  $ B.diss                : num  NA 93.1 NA 35.4 29.4 ...
##  $ B.filt                : num  NA NA NA NA NA ...
##  $ NH4.mM                : num  NA NA NA 0.05 NA NA NA NA NA NA ...
##  $ TIC.ppm.filt          : num  NA NA NA 51.8 44.8 NA 66.7 52.1 NA 69.4 ...
##  $ Cl.mM                 : num  NA NA NA 1.48 1574 ...
##  $ NO3...mM              : num  NA NA NA 616 778 ...
##  $ PO4..mM               : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ NPOC.ppm              : num  NA NA NA 4 4.4 NA 5.8 3.4 NA 9.1 ...
##  $ TIC.ppm.unfilt        : num  NA NA NA 44.8 26.4 NA 39 32.3 NA 54.8 ...
##  $ TOC.ppm.unfilt        : num  NA NA NA 4.7 5.4 NA 2.7 3.8 NA 3.9 ...
##  $ ExpMES.Kg             : num  5.35 5.35 14.88 24.4 8.08 ...
##  $ CumAppMass.g          : num  6369 6369 6369 6369 6369 ...
##  $ DissSmeto.g           : num  0.00354 0.0246 0.17004 2.64991 2.357 ...
##  $ DissSmeto.g.SD        : num  0.000278 0.001934 0.007463 0.11298 0.068486 ...
##  $ DissOXA.g             : num  0.0695 0.4832 0.8547 11.9184 11.6727 ...
##  $ DissOXA.g.SD          : num  0.0165 0.1143 0.2738 3.976 0.0873 ...
##  $ DissESA.g             : num  0.26 1.81 1.55 17.95 14.83 ...
##  $ DissESA.g.SD          : num  0.0504 0.3503 0.158 1.1855 0.3066 ...
```

```
##  $ FiltSmeto.mg.SD       : num  0.124 0.124 0.374 0.66 0.996 ...
##  $ FiltSmeto.g           : num  0.00345 0.00345 0.00573 0.00307 0.00352 ...
##  $ FiltSmeto.g.SD        : num  0.000124 0.000124 0.000374 0.00066 0.000996 ...
##  $ TotSMout.g            : num  0.00699 0.02806 0.17577 2.65298 2.36052 ...
##  $ TotSMout.g.SD         : num  0.000216 0.00137 0.005284 0.07989 0.048432 ...
##  $ MELsm.g               : num  0.302 2.078 2.379 30.241 27.008 ...
##  $ MELsm.g.SD            : num  0.0269 0.1868 0.1789 2.4062 0.1634 ...
##  $ CumOutDiss.g          : num  0.00354 0.02815 0.19818 2.84809 5.2051 ...
##  $ CumOutFilt.g          : num  0.00345 0.0069 0.01263 0.01571 0.01923 ...
##  $ CumOutSmeto.g         : num  0.00699 0.03505 0.21082 2.8638 5.22432 ...
##  $ CumOutMELsm.g         : num  0.302 2.38 4.76 35.001 62.009 ...
##  $ BalMassDisch.g        : num  6369 6367 6365 6334 6307 ...
##  $ prctMassOut           : num  4.98e-05 2.00e-04 1.25e-03 1.89e-02 1.68e-02 ...
##  $ FracDeltaOut          : num  0 -0.00533 0 -0.57576 -0.51483 ...
##  $ Events                : Factor w/ 51 levels "0-1","0-2","0-3",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Weeks                 : Factor w/ 16 levels "W0","W1","W10",..: 1 1 1 2 2 2 9 9 9 10 ...
##  $ Event                 : Factor w/ 19 levels "0","1","2","3",..: 1 1 1 2 2 2 3 3 3 4 ...
##  $ B.mean.comp.North     : num  NA NA NA NA NA ...
##  $ B.max.comp.North      : num  NA NA NA NA NA ...
##  $ B.min.comp.North      : num  NA NA NA NA NA ...
##  $ MassSoil.g.North      : num  12.6 NA NA 613.1 NA ...
##  $ comp.d13C.North       : num  NA NA NA NA NA ...
##  $ comp.d13C.SD.North    : num  NA NA NA NA NA ...
##  $ B.mean.comp.Talweg    : num  NA NA NA NA NA ...
##  $ B.max.comp.Talweg     : num  NA NA NA NA NA ...
##  $ B.min.comp.Talweg     : num  NA NA NA NA NA ...
##  $ MassSoil.g.Talweg     : num  4.44 NA NA 173.27 NA ...
##  $ comp.d13C.Talweg      : num  NA NA NA NA NA ...
##  $ comp.d13C.SD.Talweg   : num  NA NA NA NA NA ...
##  $ B.mean.comp.South     : num  NA NA NA NA NA ...
##  $ B.max.comp.South      : num  NA NA NA NA NA ...
##  $ B.min.comp.South      : num  NA NA NA NA NA ...
##  $ MassSoil.g.South      : num  18.8 NA NA 2112.1 NA ...
##  $ comp.d13C.South       : num  NA NA NA NA NA ...
##  $ comp.d13C.SD.South    : num  NA NA NA NA NA ...
##  $ ID.S                  : Factor w/ 17 levels "AW-S-0","AW-S-0x",..: 2 NA NA 1 NA NA 3 NA NA 10 ...
##  $ CatchMassSoil.g       : num  35.8 NA NA 2898.5 NA ...
##  $ BulkMass.g            : num  14.1 NA NA 1183.7 NA ...
##  $ BulkCatch.d13         : num  NA NA NA NA NA ...
```

Outliers

```r
# Test function
g_param = 1.5
# g_param = 2.2  #  (Hoaglin et al.,1986; Hoaglin & Iglewicz, 1987)
is_outlier <- function(x) {
  return(x < quantile(x, 0.25) - g_param * IQR(x) | x > quantile(x, 0.75) + g_param * IQR(x))
}
```

Soil concentrations

Correlation will be made after variable transformation. Options tested:

a) Z-scoring transformation by translation and expansion is done to create unit-free variables with means of zero and standard deviations of one. Standardised values differ from one another in units of standard deviation. The mean of each variable is subtracted from the original values and the difference divided by the variable's standard deviation and is given by:

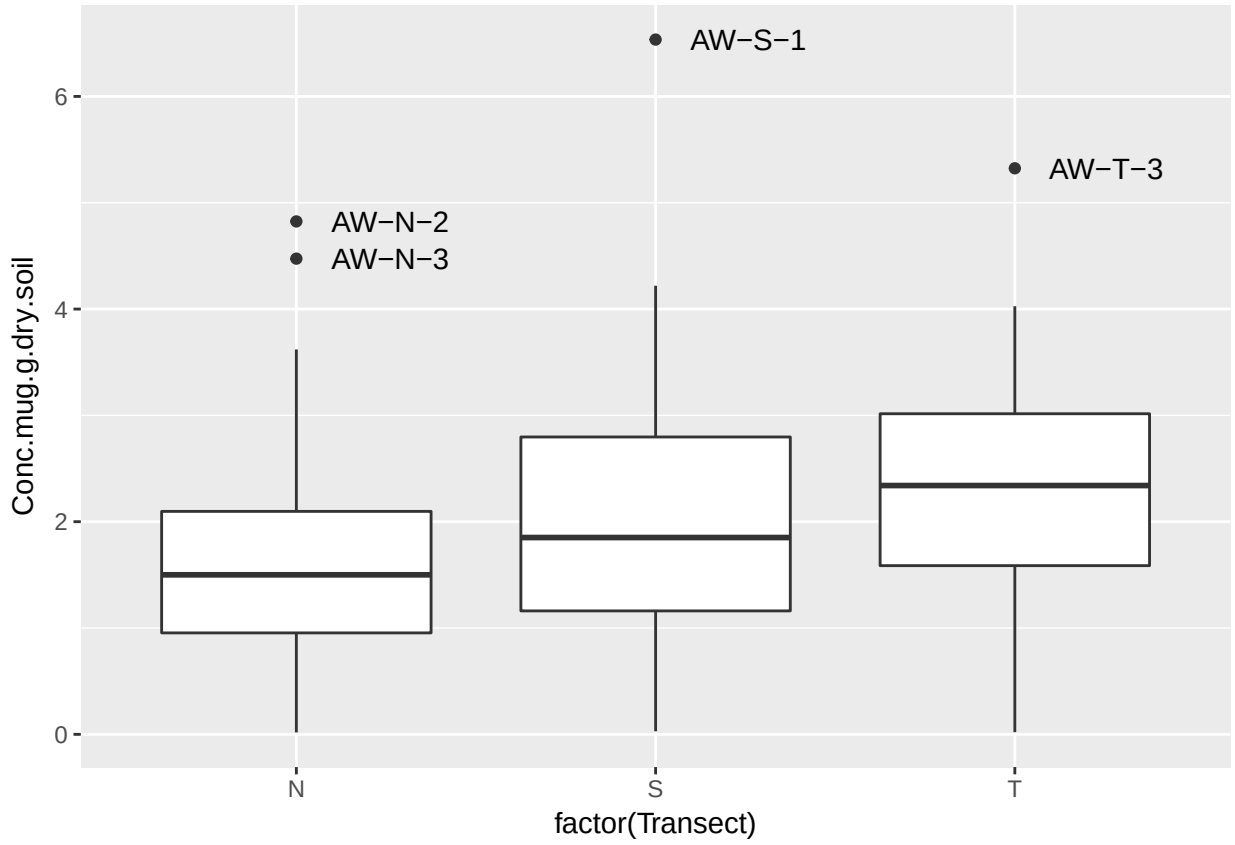
$$z_i = \frac{y_i - \bar{y}}{s_y}$$

Z-scoring did not change correlation results, nor outlier reduction.

b) Scaling by expansion where all values are divided by the maximum observation.

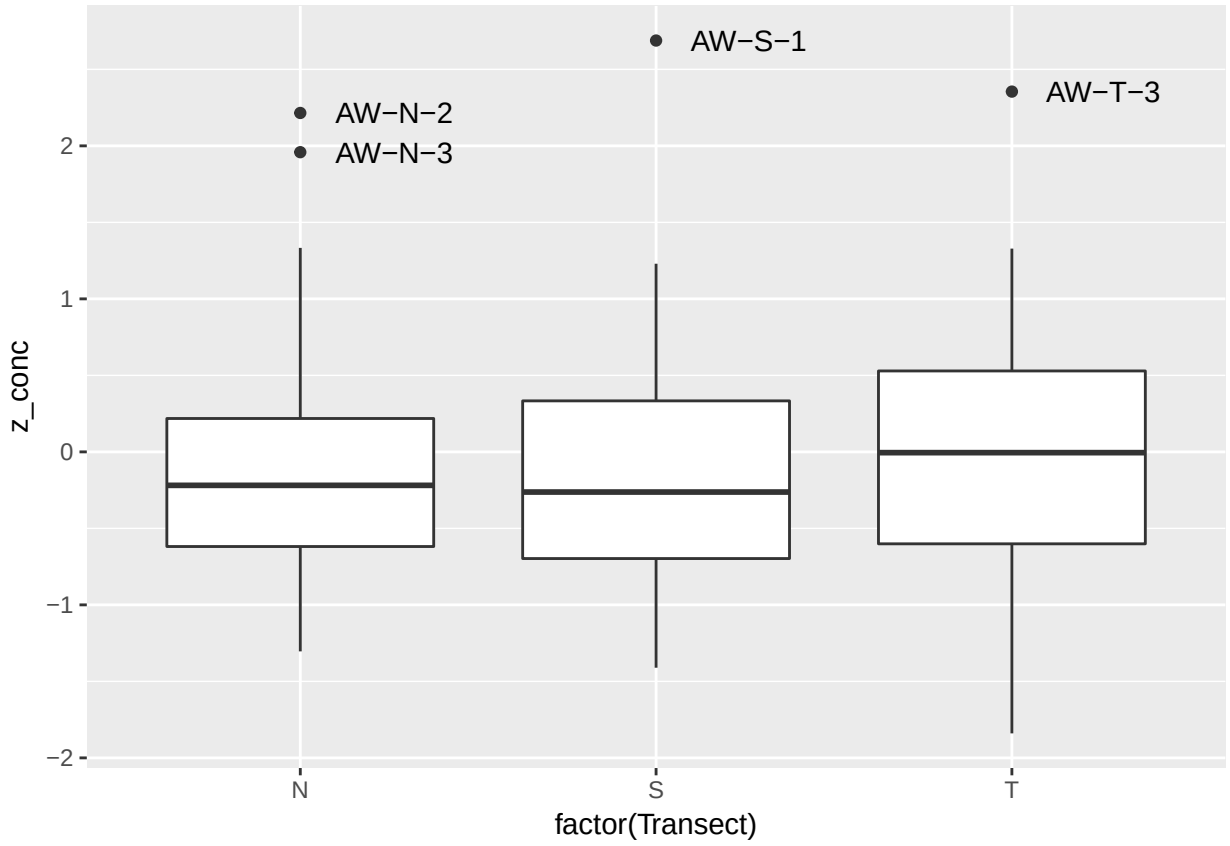
Outliers before transformation

```
# Concentrations
soilGroups %>%
  group_by(Transect) %>%
  mutate(outlier = ifelse(is_outlier(Conc.mug.g.dry.soil), as.character(ID), NA)) %>%
  ggplot(., aes(x = factor(Transect), y = Conc.mug.g.dry.soil)) +
    geom_boxplot() +
    geom_text(aes(label = outlier), na.rm = TRUE, hjust = -0.3)
```



Outliers after transformation

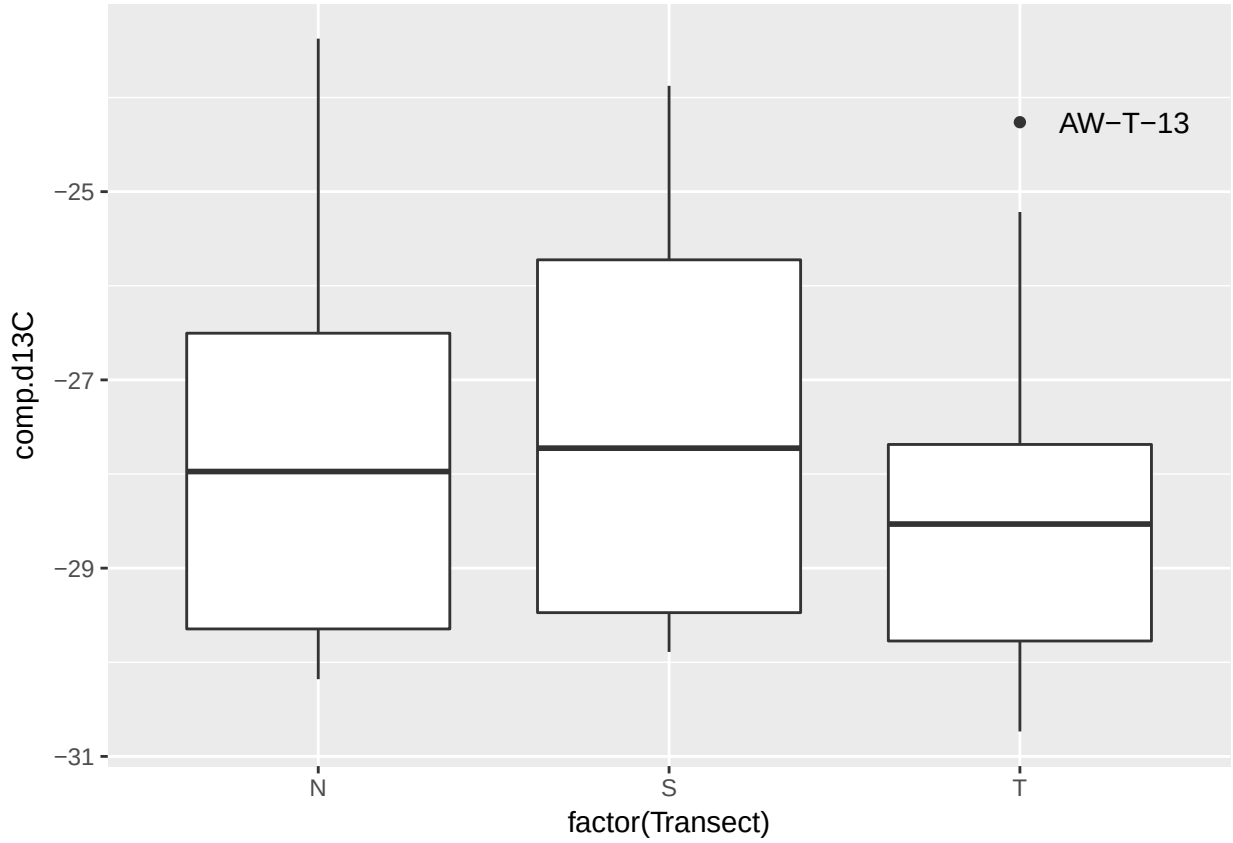
```
soilGroups <- soilGroups %>%
  group_by(Transect) %>%
  mutate(z_conc = (Conc.mug.g.dry.soil-mean(Conc.mug.g.dry.soil))/sd(Conc.mug.g.dry.soil))


soilGroups %>%
  group_by(Transect) %>%
  mutate(outlier = ifelse(is_outlier(z_conc), as.character(ID), NA)) %>%
  ggplot(., aes(x = factor(Transect), y = z_conc)) +
    geom_boxplot() +
    geom_text(aes(label = outlier), na.rm = TRUE, hjust = -0.3)
```



Soil Isotopes

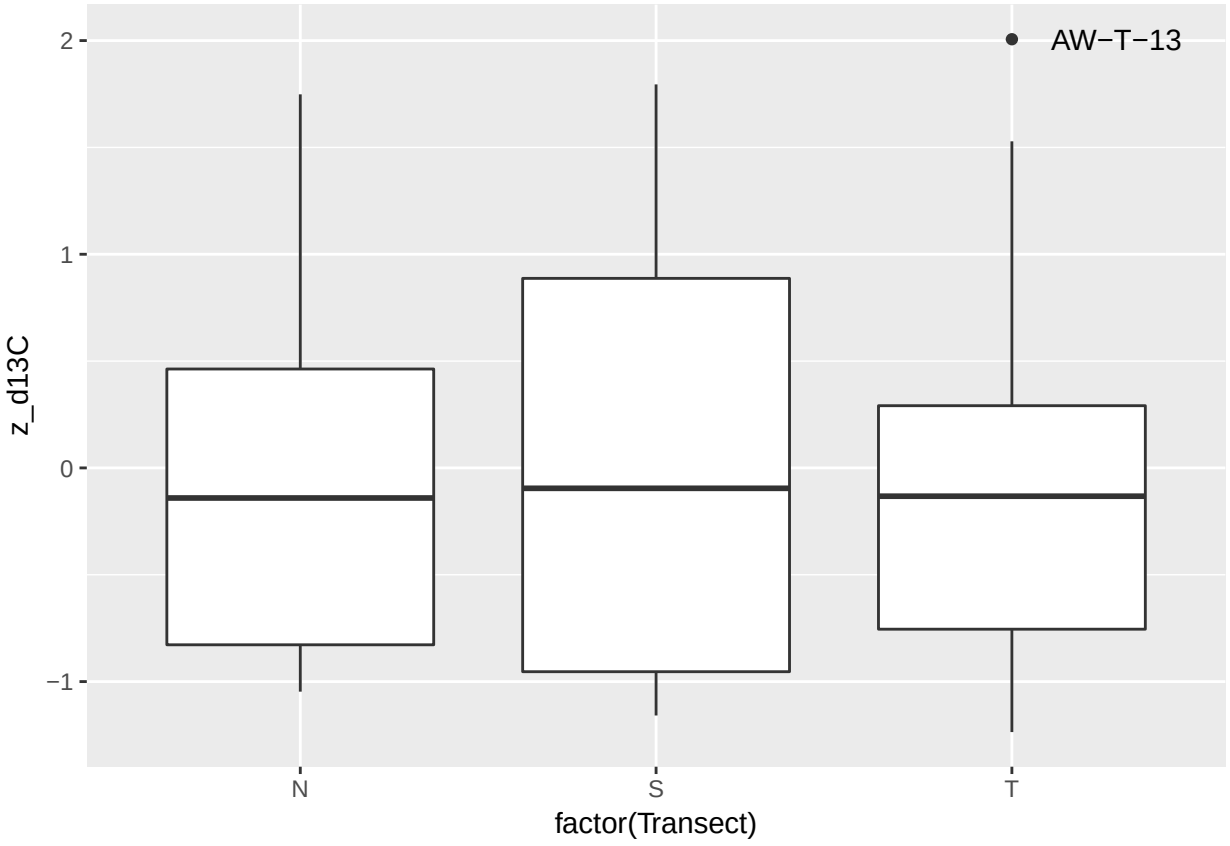
```
# Isotopes

temp <- na.omit(soilGroups)

temp %>%
  group_by(Transect) %>%
  mutate(outlier = ifelse(is_outlier(comp.d13C), as.character(ID), NA)) %>%
  ggplot(., aes(x = factor(Transect), y = comp.d13C)) +
```

```
    geom_boxplot() +
    geom_text(aes(label = outlier), na.rm = TRUE, hjust = -0.3)
```



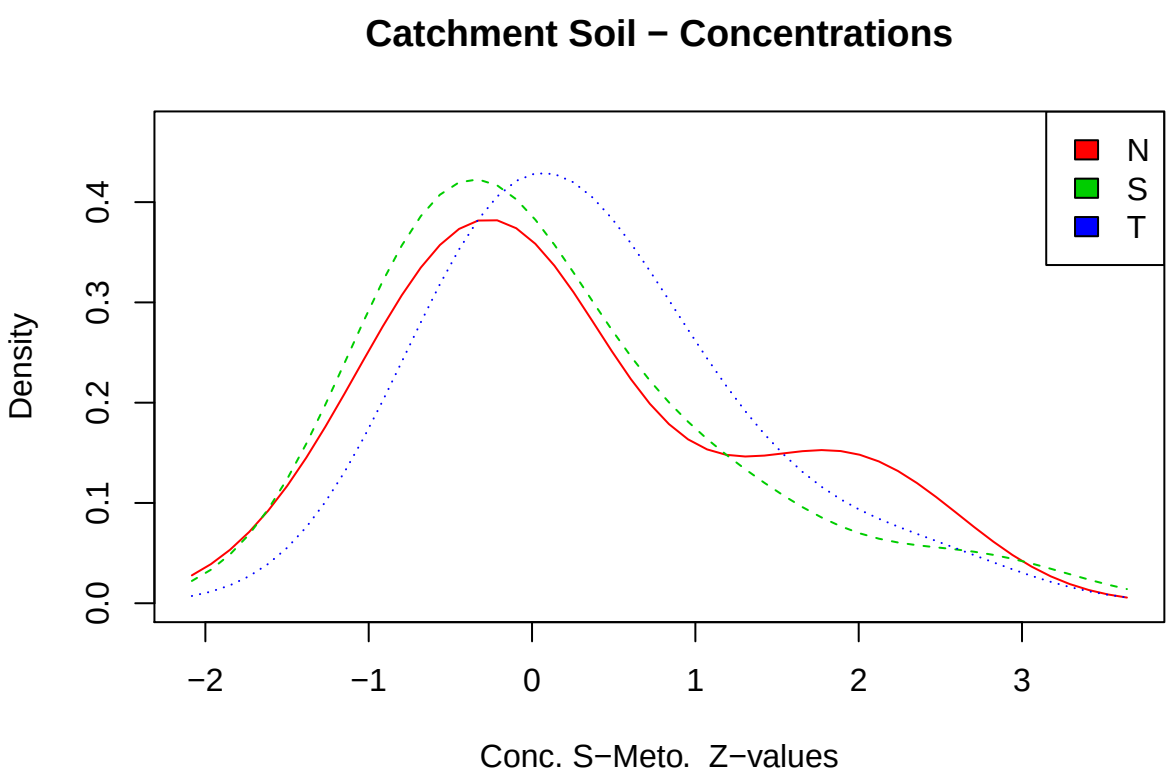
Looks like 7 potential outliers in concentrations and 1 for isotopes. Removing NA's for isotopes and re-computing outliers, reduces the number of outliers to 2 in concentrations and 1 for isotopes.

```
temp <- temp %>%
  group_by(Transect) %>%
  mutate(z_d13C = (comp.d13C-mean(comp.d13C))/sd(comp.d13C))

temp %>%
  group_by(Transect) %>%
  mutate(outlier = ifelse(is_outlier(z_d13C), as.character(ID), NA)) %>%
  ggplot(., aes(x = factor(Transect), y = z_d13C)) +
    geom_boxplot() +
    geom_text(aes(label = outlier), na.rm = TRUE, hjust = -0.3)
```



Distribution of z values (same as non-transformed)

```
# plot densities
#sm.density.compare(temp$z_conc, temp$Transect, xlab=expression(paste("Conc. S-Meto.  ", {({mu}*g / g.s
sm.density.compare(temp$z_conc, temp$Transect, xlab=expression(paste("Conc. S-Meto.  Z-values")))
title(main="Catchment Soil - Concentrations")
legend("topright", levels( soilGroups$Transect), fill=2+(0:nlevels(soilGroups$Transect)))
```

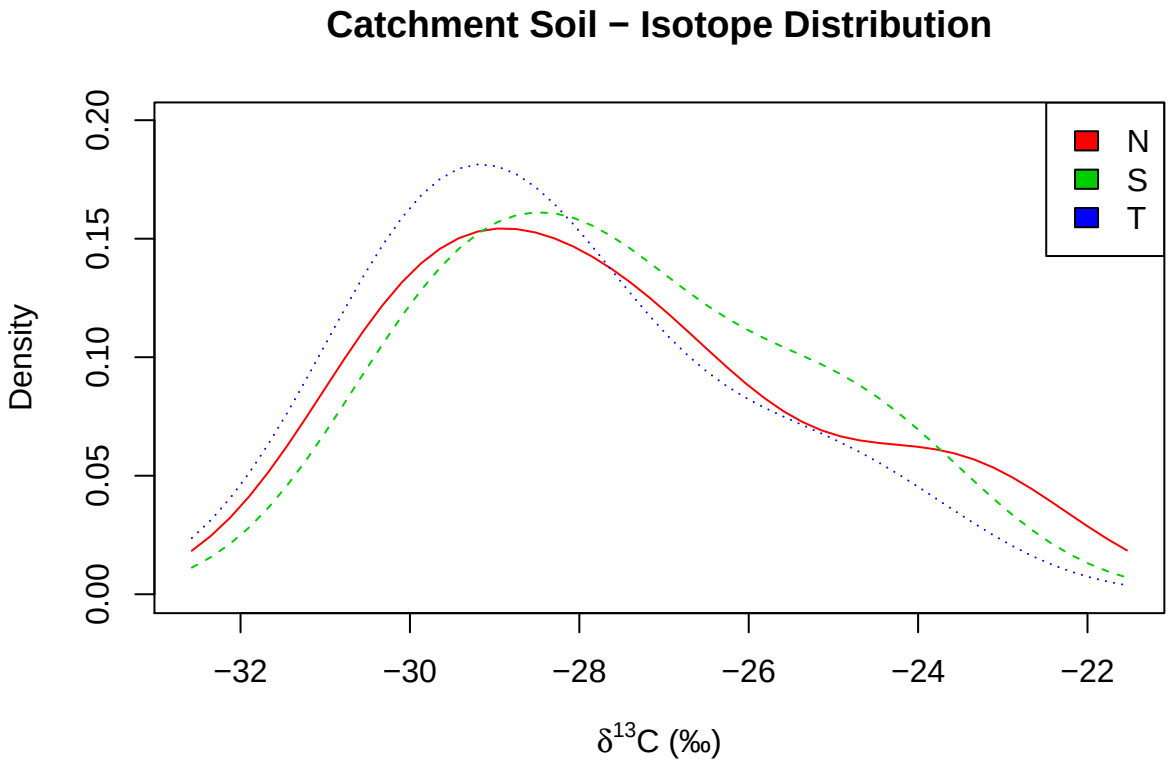



```
#vioplot(soilGroups$Conc.mug.g.dry.soil,  names = "Catchment")
#title(expression(paste("Conc. S-Meto.  ", {({mu}*g / g.soil.dry)})))
```

Soil Isotopes

```
#vioplot(na.omit(soilGroups$comp.d13C),  names = "Catchment")
#title(expression(paste({delta}^"13","C", ' (\u2030)')))
```

```
temp <- na.omit(soilGroups)
sm.density.compare(temp$comp.d13C, temp$Transect,
                   xlab=expression(paste({delta}^"13","C", ' (\u2030)')))
title(main="Catchment Soil - Isotope Distribution")
legend("topright", levels( soilGroups$Transect), fill=2+(0:nlevels(soilGroups$Transect)))
```


Catchment Soil − Isotope Distribution
Density
0.00
0.05
0.10
0.15
0.20
−32
−30
−28
−26
−24
−22
δ13C (‰)
N
S
T