missed_flag_ratio_analysis

Marlin

2022-11-09

```
library(DSIWastewater)
library(ggplot2)
```

```
## Warning: package ’ggplot2’ was built under R version 4.2.1
```

```
library(dplyr)
```

```
## Warning: package ’dplyr’ was built under R version 4.2.1
```

```
##
## Attaching package: ’dplyr’
```

```
## The following objects are masked from ’package:stats’:
##
##     filter, lag
```

```
## The following objects are masked from ’package:base’:
##
##     intersect, setdiff, setequal, union
```

```
load_Dataset <- function(){
  #load WasteWater_data into the environment
  data(WasteWater_data, package = "DSIWastewater")
  baseWaste_DF <-  buildWasteAnalysisDF(WasteWater_data)
  data(Case_data, package = "DSIWastewater")
  Case_DF <- Case_data

  Flag_DF <- read.csv("Temp/DHSFlagingMethodOutput.csv")%>%
            mutate(date = as.Date(date))%>%
   select(-X)
  return(Flag_DF)

}
Flag_DF <- load_Dataset()
date_Flag_DF <- DF_date_vector(Flag_DF, "date",
               names(Flag_DF)[3:68])
baseWaste_DF <-  buildWasteAnalysisDF(WasteWater_data)
baseWaste_DF$site <- ifelse(baseWaste_DF$site == "Madison MSD WWTF",
                            "Madison", baseWaste_DF$site)
```

```
#"case_flag_Cases"                         "case_flag_7DayCases"
#"case_flag_plus_comm.threshold_Cases"     "case_flag_plus_comm.threshold_7DayCases"
#"slope_switch_flag_Cases"                 "slope_switch_flag_7DayCases"
dep_flags <- names(Flag_DF)[9:68]
edgeThresh <- 21
CaseFlag <- "slope_switch_flag_Cases"
rawDateDistDF <- date_Flag_DF%>%
  date_distance_calc(CaseFlag, dep_flags)%>%
  select(site, date, all_of(dep_flags))%>%
  tidyr::pivot_longer(cols = dep_flags,
                      names_to = c("FlagType","window", "quant"),
                      values_to = "FlagError",
                      names_sep = "_")%>%
  mutate(window = as.numeric(window), quant = as.numeric(quant))
```

```
## Warning: Using an external vector in selections was deprecated in tidyselect 1.1.0.
## i Please use ‘all_of()‘ or ‘any_of()‘ instead.
##   # Was:
##   data %>% select(dep_flags)
##
##   # Now:
##   data %>% select(all_of(dep_flags))
##
## See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
```

```
library(lubridate)
```

```
## Warning: package ’lubridate’ was built under R version 4.2.1
```

```
##
## Attaching package: ’lubridate’
```

```
## The following objects are masked from ’package:base’:
##
##     date, intersect, setdiff, union
```

```
city_data <- baseWaste_DF%>%
  group_by(site, week(date), year(date))%>%
  summarise(n = n(), pop = mean(population_served))%>%
  group_by(site)%>%
  summarise(sampleRate = round(mean(n)), pop = mean(pop))%>%
  mutate(pop = ntile(pop, 3))
```

```
## ‘summarise()‘ has grouped output by ’site’, ’week(date)’. You can override
## using the ‘.groups‘ argument.
```

```
#flaging method
DistSummaryMainSite <- rawDateDistDF%>%
  #filter(window > 30)%>%
  group_by(window, quant, FlagType)%>%
  summarise(Mean = mean(FlagError, na.rm = TRUE),
            Var = var(FlagError, na.rm = TRUE),
            num_flags = sum(!is.na(FlagError)),
            missed_percent = mean(abs(FlagError)>edgeThresh, na.rm = TRUE),
            MeanErrorSquard = mean(
                          ifelse(abs(FlagError)>edgeThresh,
                            NA,FlagError)^2, na.rm = TRUE))%>%
  filter(num_flags != 0)
```

```
## ‘summarise()‘ has grouped output by ’window’, ’quant’. You can override using
## the ‘.groups‘ argument.
```

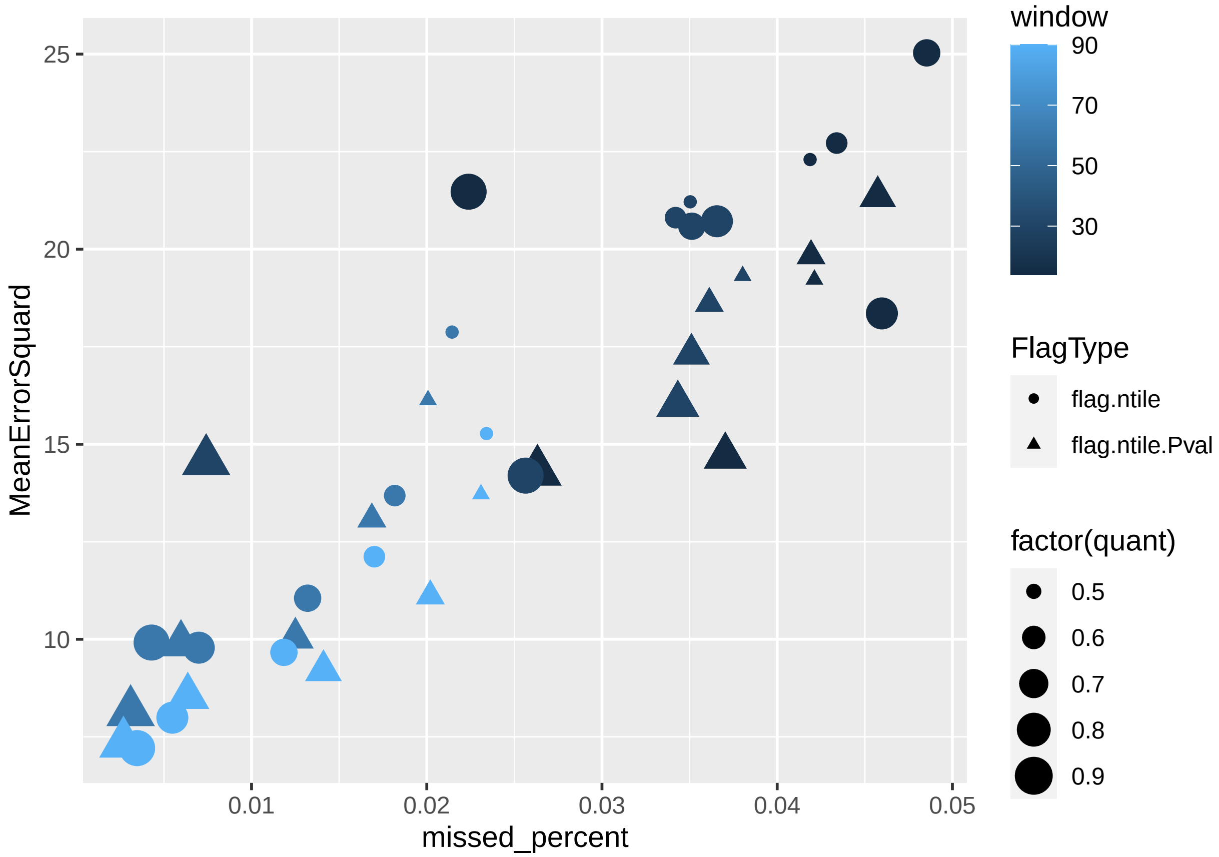
```
DistSummaryMainSite <- DistSummaryMainSite%>%
  filter(FlagType != "cdc.flag")

DistSummaryMainSite%>%
  lm(missed_percent~MeanErrorSquard + window + quant + FlagType,data = .)%>%
  summary()
```

```
##
## Call:
## lm(formula = missed_percent ~ MeanErrorSquard + window + quant +
##     FlagType, data = .)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.015953 -0.003097  0.000634  0.003445  0.014235
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.637e-02  1.800e-02   0.910  0.36921
## MeanErrorSquard         1.711e-03  5.155e-04   3.319  0.00212 **
## window                 -1.430e-04  7.813e-05  -1.831  0.07567 .
## quant                  -1.754e-02  1.038e-02  -1.689  0.10006
## FlagTypeflag.ntile.Pval  2.139e-03  2.175e-03   0.983  0.33213
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.006075 on 35 degrees of freedom
## Multiple R-squared:  0.8401, Adjusted R-squared:  0.8218
## F-statistic: 45.96 on 4 and 35 DF,  p-value: 1.84e-13
```

```
DistSummaryMainSite%>%
  #filter(missed_percent != 0)%>%
  ggplot(aes(x = missed_percent, y = MeanErrorSquard,
             color = window, size = factor(quant), shape = FlagType))+
  geom_point()
```

```
## Warning: Using size for a discrete variable is not advised.
```



```
#flaging method
DistSummarySite <- rawDateDistDF%>%
  left_join(city_data)%>%
  #filter(window > 30)%>%
  group_by(window, quant, FlagType, sampleRate, pop)%>%
  summarise(Mean = mean(FlagError, na.rm = TRUE),
            Var = var(FlagError, na.rm = TRUE),
            num_flags = sum(!is.na(FlagError)),
            missed_percent = mean(abs(FlagError)>edgeThresh, na.rm = TRUE),
            MeanErrorSquard = mean(
                          ifelse(abs(FlagError)>edgeThresh,
                            NA,FlagError)^2, na.rm = TRUE))%>%
  filter(num_flags != 0)
```

```
## Joining, by = "site"
## `summarise()` has grouped output by 'window', 'quant', 'FlagType',
## 'sampleRate'. You can override using the `.groups` argument.
```

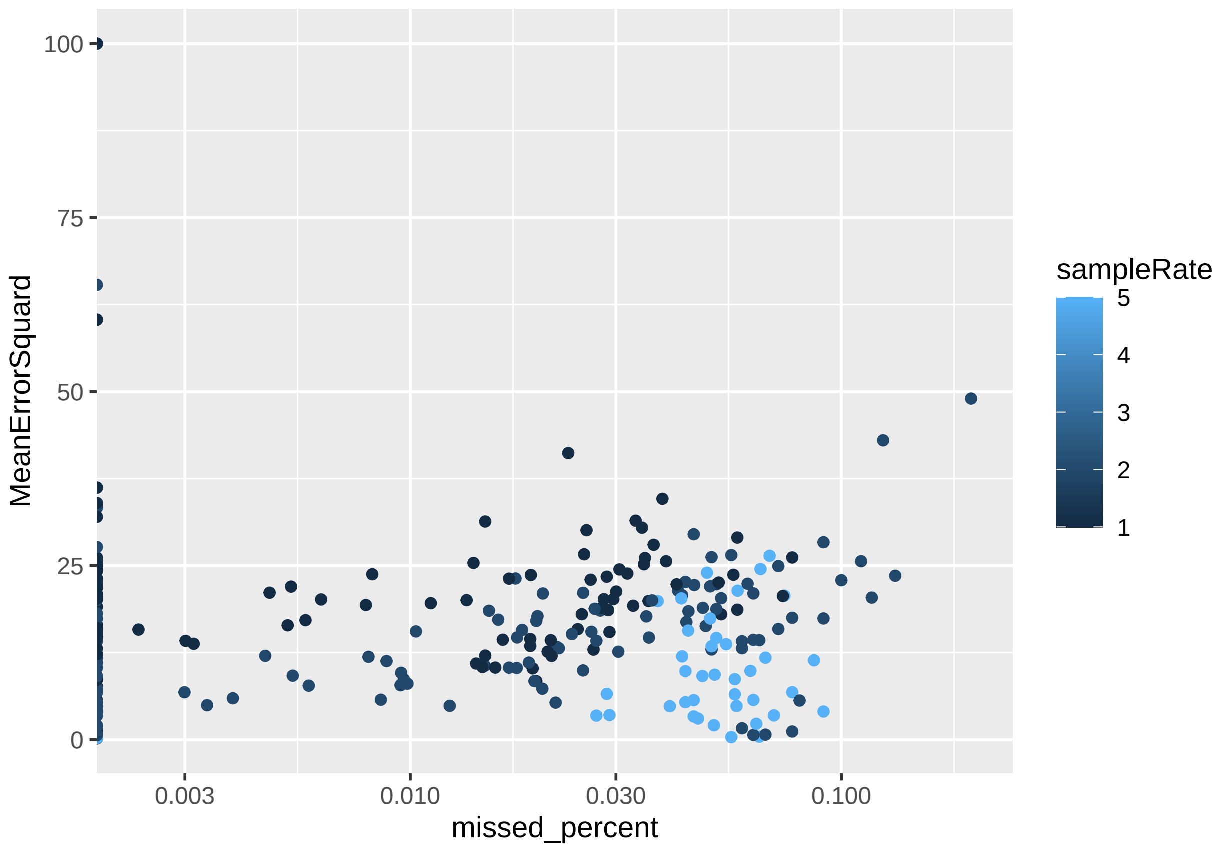
```
QuantDistSummarySite <- DistSummarySite%>%
  filter(FlagType != "cdc.flag")

QuantDistSummarySite%>%
  lm(missed_percent ~ window + quant + FlagType + sampleRate + pop, data = .)%>%
  summary()
```

```
##
## Call:
## lm(formula = missed_percent ~ window + quant + FlagType + sampleRate +
##     pop, data = .)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.055479 -0.014843 -0.001901  0.010442  0.160906
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             6.950e-02  8.753e-03   7.941 5.86e-14 ***
## window                 -2.584e-04  4.945e-05  -5.225 3.55e-07 ***
## quant                  -3.926e-02  1.026e-02  -3.825 0.000163 ***
## FlagTypeflag.ntile.Pval -2.809e-03  2.868e-03  -0.979 0.328277
## sampleRate              1.310e-02  1.205e-03  10.876  < 2e-16 ***
## pop                    -1.353e-02  1.923e-03  -7.036 1.72e-11 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.02351 on 263 degrees of freedom
## Multiple R-squared:  0.3758, Adjusted R-squared:  0.3639
## F-statistic: 31.66 on 5 and 263 DF,  p-value: < 2.2e-16
```

```
QuantDistSummarySite%>%
  #filter(missed_percent != 0)%>%
  ggplot(aes(x = missed_percent, y = MeanErrorSquard, color = sampleRate))+
  geom_point()+
  scale_x_log10()
```

```
## Warning: Transformation introduced infinite values in continuous x-axis
```

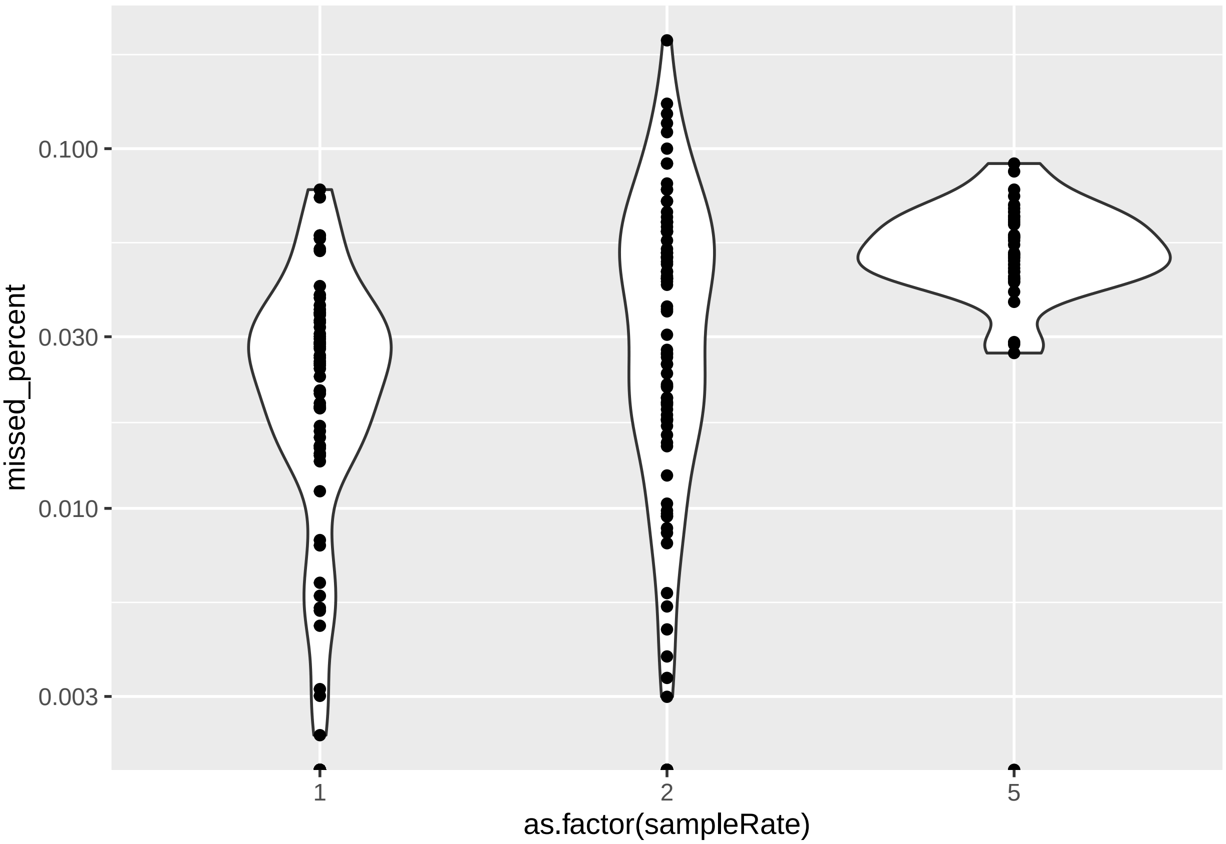


```
QuantDistSummarySite%>%
  ggplot(aes(x = as.factor(sampleRate), y = missed_percent))+
  geom_violin()+
  geom_point()+
  scale_y_log10()
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Removed 88 rows containing non-finite values (stat_ydensity).
```


missed_percent
0.100
0.030
0.010
0.003
1
2
5
as.factor(sampleRate)