Milestone 4

Group 6

2022-09-08

```
riverquality<-read.csv("../Data/ms4cleaned.csv")
riverquality<-riverquality[,-1]
rq08<-read.csv("../Data/rq08.csv")
rq90<-read.csv("../Data/rq90.csv")
rq98<-read.csv("../Data/rq98.csv")

num.time<-riverquality[,c("period","Phosphorus","Nitrogen","E.Coli","Turbidity")]
num.time$period <- as.factor(num.time$period)

num.melt <- melt(data=num.time,
id.vars = "period",
variable.name = "Compound")


class.new <- num.time[1500,1]
x.new <- num.time[1500,-1]

num.time <- num.time[-45, ]
num.melt <- melt(data=num.time,
id.vars = "period",
variable.name = "Compound")
```

Mahalanobis distance

The distance between x1500 and the compound model is 2.250291, and the dataset has a probability of 0.69 falling into this distance (Fig. 1).

```
est.mu <- colMeans(num.time[, -1])
est.covar <- var(num.time[, -1])

(d.new <- mahalanobis(x.new, center = est.mu, cov = est.covar))
```

```
##     1500
## 2.250291
```

```
pchisq(d.new, df = 4, lower.tail = FALSE)
```

```
##      1500
## 0.6898334
```

```
dM <- mahalanobis(num.time[,-1], center = est.mu, cov = est.covar)
upper.quantiles <- qchisq(c(.9, .95, .99), df=4)
density.at.quantiles <- dchisq(x=upper.quantiles, df=4)
cut.points <- data.frame(upper.quantiles, density.at.quantiles)
ggplot(data.frame(dM), aes(x=dM)) +
geom_histogram(aes(y=..density..), bins=nclass.FD(dM),
fill="white", col="black") +
stat_function(fun="dchisq", args = list(df=4),
col="blue", size=2, alpha=.7, xlim=c(0,65)) +
geom_point(aes(x=d.new, y=0), size=3, col="red") +
geom_segment(data=cut.points,
aes(x=upper.quantiles, xend=upper.quantiles,
y=rep(0,3), yend=density.at.quantiles),
col="red", size=1) +xlab("Mahalanobis distances and cut points") +
ylab("Histogram and density")
```

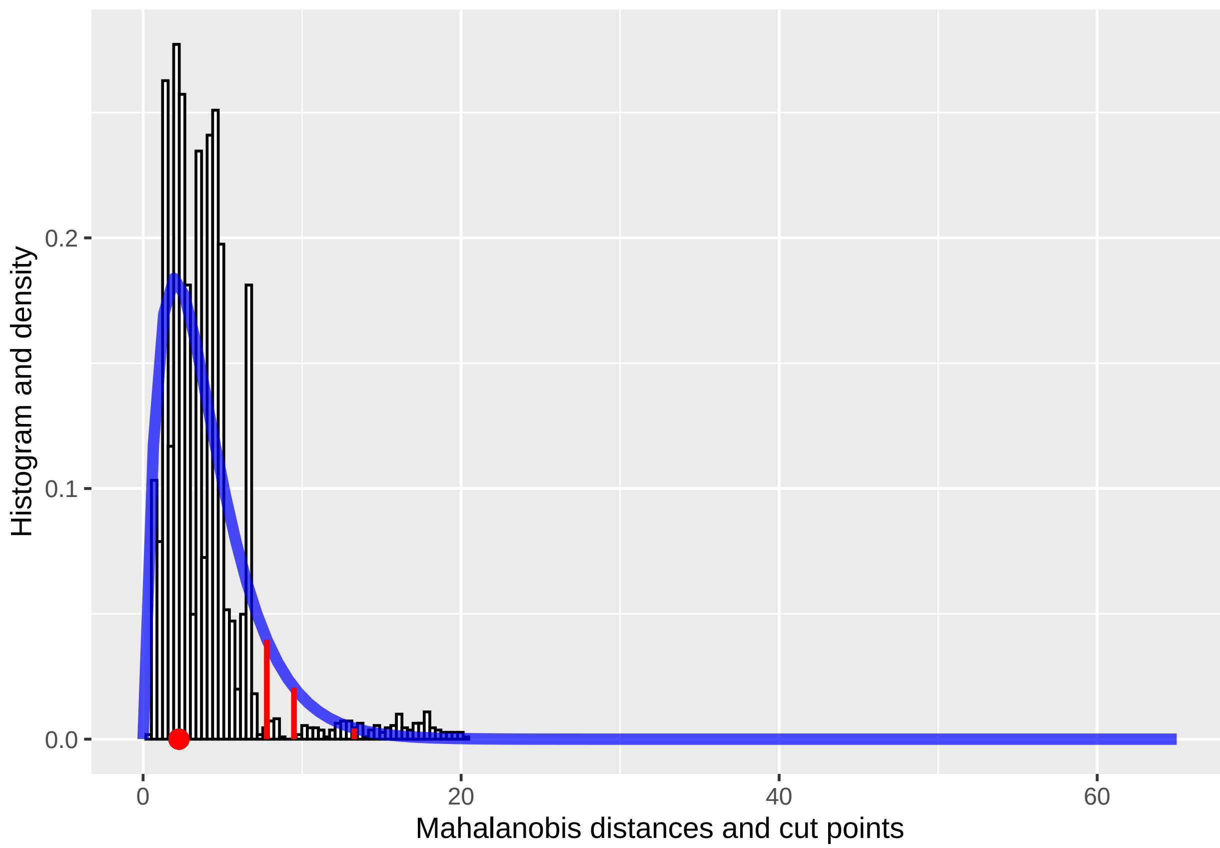

Figure 1: Distribution of the Mahalanobis distance of river observations to the aggregated model

The new measure is consistent with both the first and second time periods; 1990-2017 and 1998-2017, and closest to 1998-2017.

```
est.mu1 <- colMeans(subset(num.time, period=="1990-2017")[,-1])
est.covar1 <- var(subset(num.time, period=="1990-2017")[,-1])
est.mu2 <- colMeans(subset(num.time, period=="1998-2017")[,-1])
est.covar2 <- var(subset(num.time, period=="1998-2017")[,-1])
```

```r
est.mu3 <- colMeans(subset(num.time, period=="2008-2017"
)[,-1])
est.covar3 <- var(subset(num.time, period=="2008-2017"
)[,-1])

(d.new1 <- mahalanobis(x.new, center = est.mu1, cov = est.covar1))
```

```
##     1500
## 2.246352
```

```r
pchisq(d.new1, df = 4, lower.tail = FALSE)
```

```
##      1500
## 0.6905527
```

```r
(d.new2 <- mahalanobis(x.new, center = est.mu2, cov = est.covar2))
```

```
##     1500
## 2.102957
```

```r
pchisq(d.new2, df = 4, lower.tail = FALSE)
```

```
##      1500
## 0.7168292
```

```r
(d.new3 <- mahalanobis(x.new, center = est.mu3, cov = est.covar3))
```

```
##     1500
## 43.36518
```

```r
pchisq(d.new3, df = 4, lower.tail = FALSE)
```

```
##         1500
## 8.690891e-09
```

Cleaned data (overall)

We can see that all river health indicators; E.Coli, Phosphorus, Nitrogen and Turbidity peak at a value of around 110 (Fig. 2). Phosphorus, Nitrogen and turbidity also show smaller peaks at 250 and 340 (Fig. 2).

```
##  waterq.num_id    waterq.E.Coli   waterq.Phosphorus waterq.Nitrogen
##  Min.   :   1.0   Min.   : 36.0   Min.   : 36       Min.   : 36.0
##  1st Qu.: 789.2   1st Qu.:109.0   1st Qu.:119       1st Qu.:119.0
##  Median :1577.5   Median :119.0   Median :216       Median :120.0
##  Mean   :1577.5   Mean   :109.2   Mean   :193       Mean   :181.3
##  3rd Qu.:2365.8   3rd Qu.:120.0   3rd Qu.:252       3rd Qu.:251.0
##  Max.   :3154.0   Max.   :247.0   Max.   :336       Max.   :324.0
```

```
## waterq.Longitude   waterq.lat     waterq.Turbidity
## Min.   :167.5    Min.   :-46.57   Min.   : 36.0
## 1st Qu.:171.1    1st Qu.:-44.17   1st Qu.:118.0
## Median :174.3    Median :-40.02   Median :120.0
## Mean   :173.5    Mean   :-40.94   Mean   :178.6
## 3rd Qu.:175.8    3rd Qu.:-38.20   3rd Qu.:252.0
## Max.   :177.9    Max.   :-35.04   Max.   :336.0
```

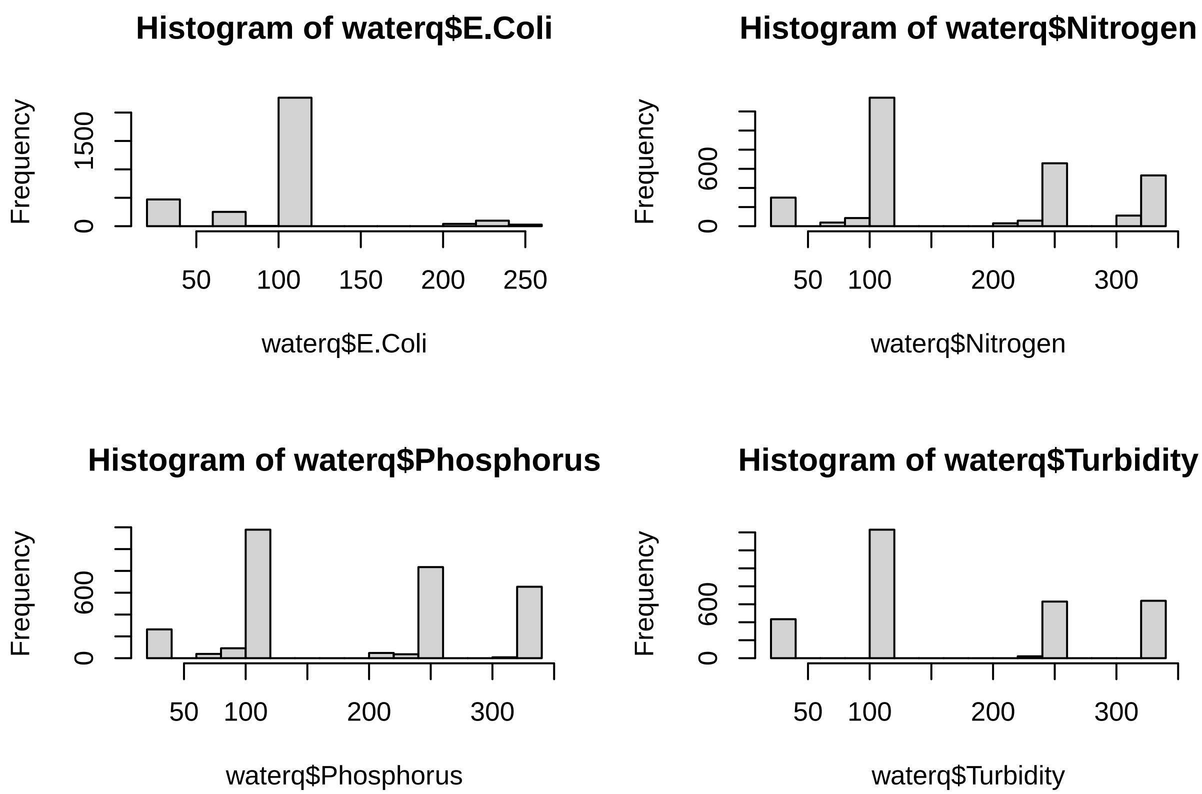

Figure 2: Histograms for river health indicators; E.Coli, Nitrogen, Phosphorus and Turbidity in NZ rivers

```
## Using s_id, dominant_landcover, period, Trend, percent_annual_change as id variables
```

In this case, we fit the river health indicators; "E.Coli" with "Turbidity" to assign a plot of Fitted vs Residual values.

Residual vs fitted values are not normally distributed about zero, with differing 'heights' (Fig. 4). This suggests non-constant variance. The Normal Q-Q plot shows that the data is not normally distributed, as it does not show a straight line (Fig. 4).

Levene Test: H_0 : equal variance H_1 : not all variance are equal

```
leveneTest(waterq$E.Coli ~ as.factor(waterq$Turbidity), data = waterq)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
```

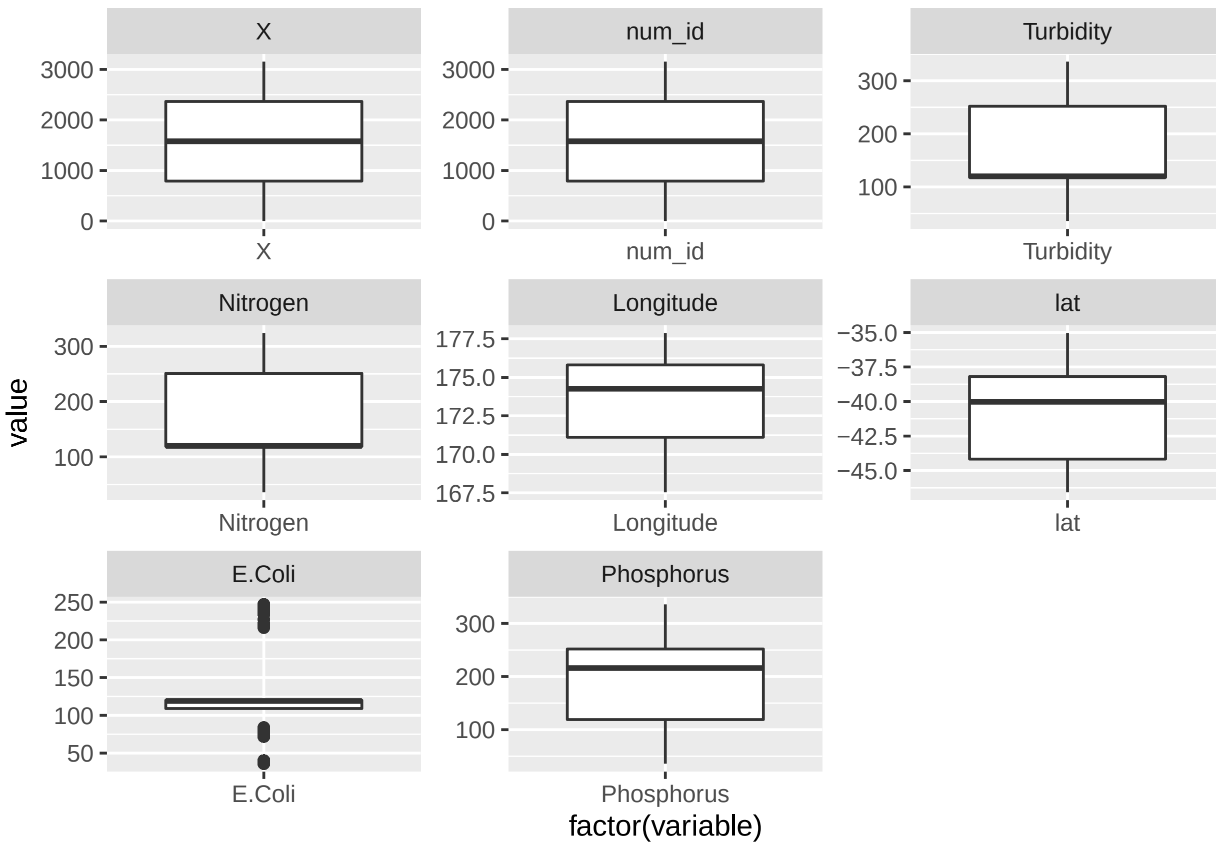

Figure 3: Boxplots showing the distributions of NZ river health and information variables

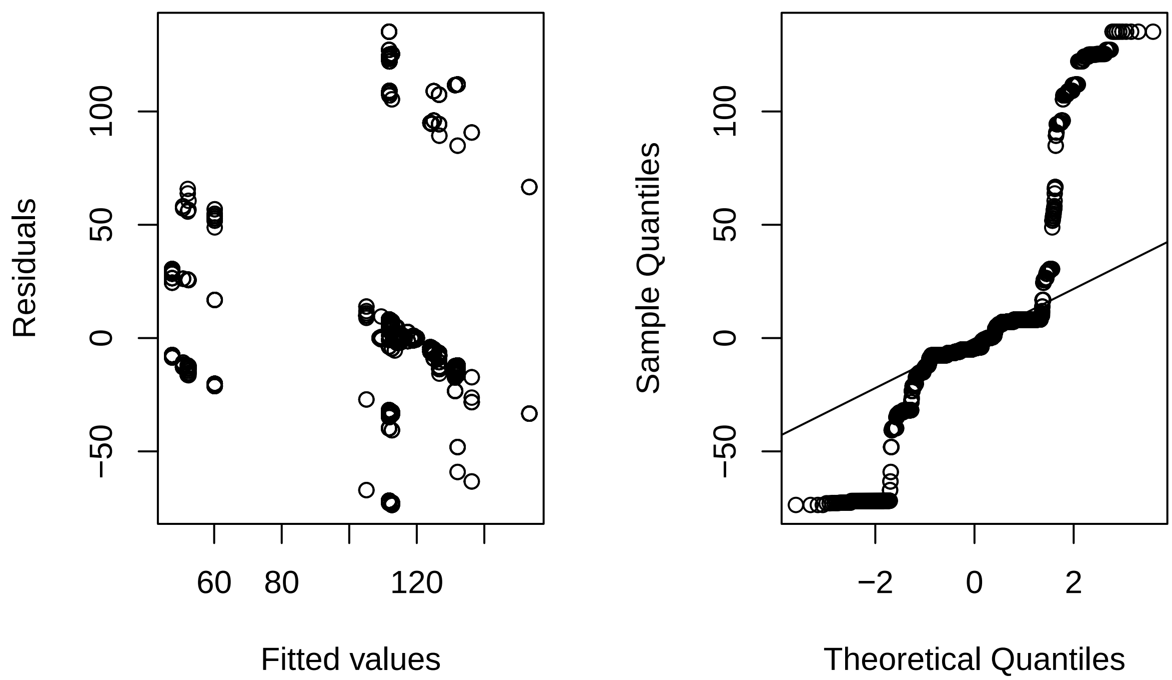

Figure 4: Plot of residuals vs. fitted values and normal Q-Q plot for E.Coli and Turbidity factors in NZ rivers

```
## group   35  9.0905 < 2.2e-16 ***
##       3118
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Levene's test produces the F-statistic = 9.0905, and p-value = 2.2e-16, which provides strong evidence to reject the null hypothesis of equal variance.

We have the fitted data that has no constant variance and is not distributed normally. Therefore, we use permutation test instead of ANOVA.

Permutation: H_0 : There is no difference in E.Coli across Turbidity H_1 : E.Coli differs across Turbidity

```r
Fobs<-anova(mod1)$F[1]
Fnull<-rep(NA,2000)
for(t in 1:2000)
{
reorder<-sample(waterq$Turbidity)
Fnull[t]<-(anova(lm(waterq$E.Coli~as.factor(reorder))))$F[1]
}
p<-sum(Fnull>=Fobs)/2000
p
```

```
## [1] 0
```

p-value = 0, therefore there is evidence that E.Coli differs across Turbidity.

Period 1 (1990-2017)

E.Coli, Nitrogen and Turbidity show peaks at values around 110, whereas Phosphorus peaks at a high value of 350 (Fig. 5). This concentration of the E.Coli and Phosphorus data towards these value is evident in the boxplots (Fig. 6). We can also see a range of measurements of Nitrogen and Turbidity in NZ rivers during this period (1990-2017) (Fig. 5, Fig, 6).

```
##  waterq90.num_id waterq90.E.Coli waterq90.Phosphorus waterq90.Nitrogen
##  Min.   :   1    Min.   : 39.0   Min.   :101.0       Min.   : 37
##  1st Qu.:1253    1st Qu.:118.0   1st Qu.:331.8       1st Qu.:120
##  Median :1806    Median :120.0   Median :334.0       Median :250
##  Mean   :1752    Mean   :123.3   Mean   :309.7       Mean   :216
##  3rd Qu.:2360    3rd Qu.:120.0   3rd Qu.:335.0       3rd Qu.:319
##  Max.   :3027    Max.   :247.0   Max.   :336.0       Max.   :324
##  waterq90.Longitude  waterq90.lat    waterq90.Turbidity
##  Min.   :167.5      Min.   :-46.39   Min.   : 36.0
##  1st Qu.:170.9      1st Qu.:-44.27   1st Qu.:120.0
##  Median :174.3      Median :-40.24   Median :251.0
##  Mean   :173.4      Mean   :-41.03   Mean   :215.7
##  3rd Qu.:176.1      3rd Qu.:-38.27   3rd Qu.:332.0
##  Max.   :177.9      Max.   :-35.27   Max.   :336.0
```

```
## Using s_id, dominant_landcover, period, Trend, percent_annual_change as id variables
```

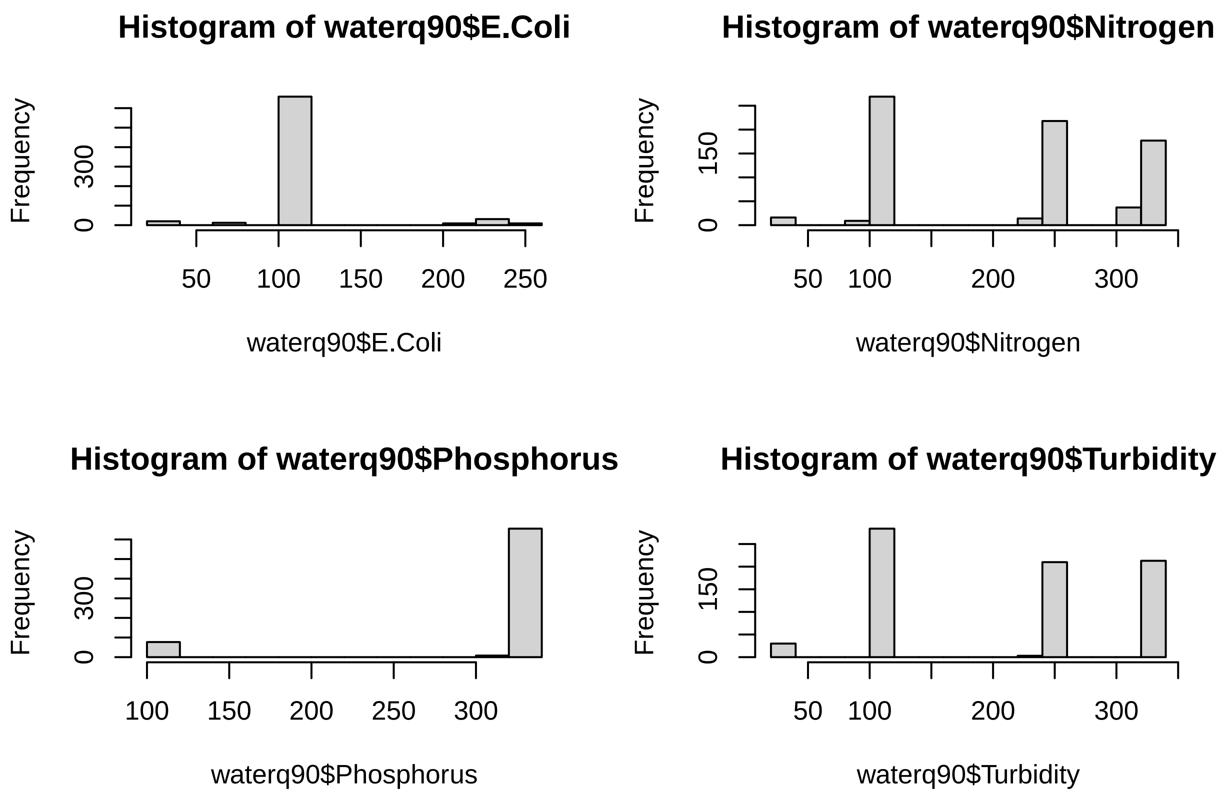

Figure 5: Histograms showing NZ river health indicators; E.Coli, Nitrogen, Phosphorus and Turbidity from 1990-2017

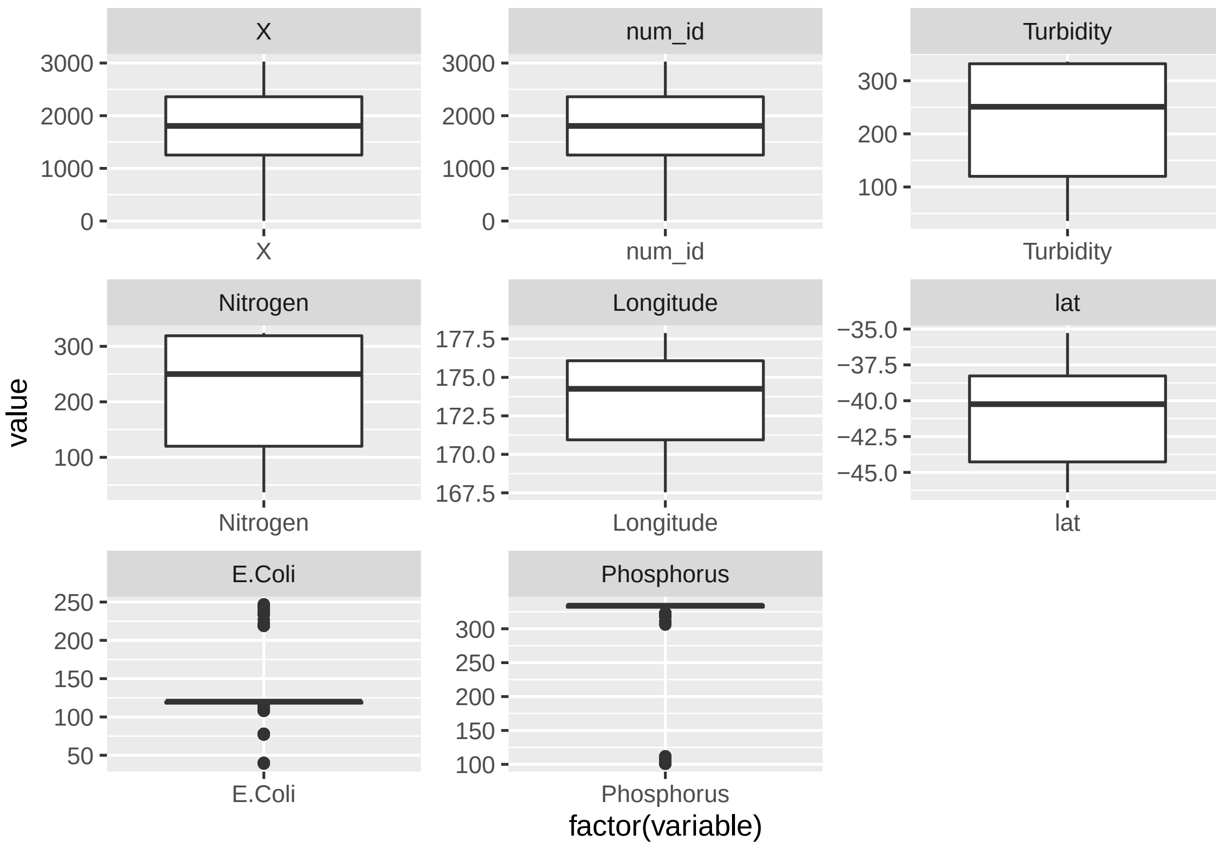

Figure 6: Boxplots showing the distributions of NZ river health and information variables from 1990-2017

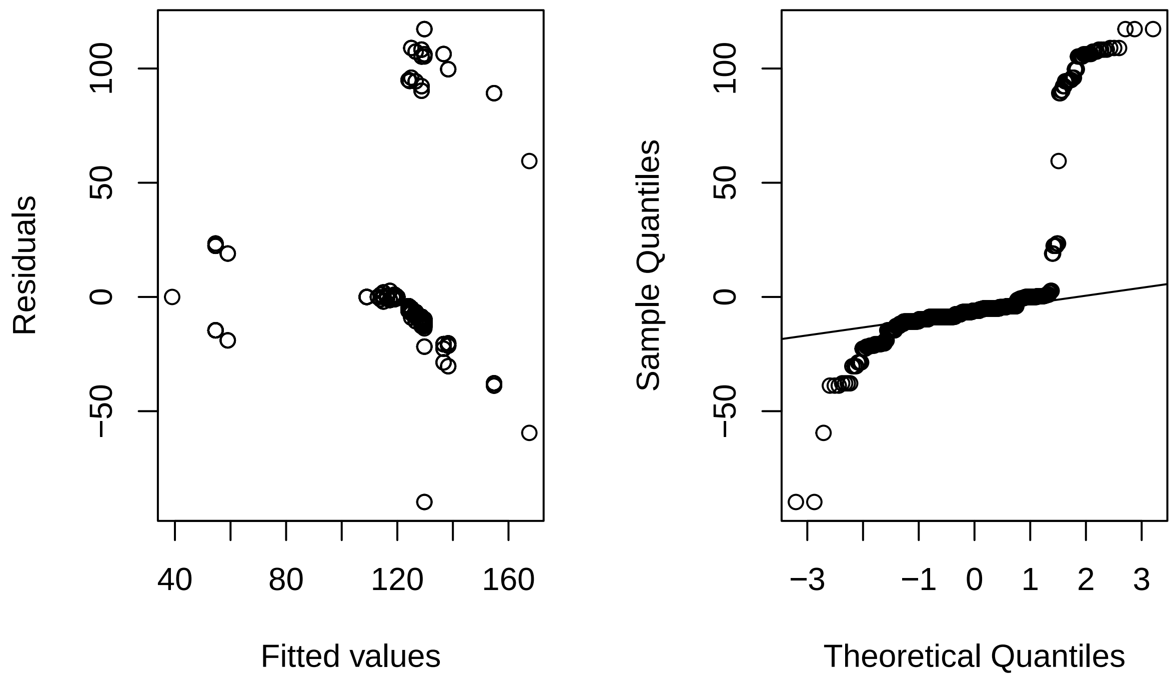

Figure 7: Plot of residuals vs. fitted values and normal Q-Q plot for E.Coli and Turbidity factors in NZ rivers; 1990-2017

Fit E.Coli with Turbidity to analyse the difference in all 3 periods.

Residual vs fitted values are not normally distributed about zero, with differing 'heights'. This suggests non-constant variance. The Normal Q-Q plot shows that the data is not normally distributed, as it does not show a straight line.

Levene Test: H_0 : equal variance H_1 : not all variance are equal

```
leveneTest(waterq90$E.Coli ~ as.factor(waterq90$Turbidity), data = waterq90)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value   Pr(>F)
## group  28  1.7753 0.008519 **
##       711
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-statistic = 1.7753, and p-value = 0.008519 of Levene's test provides strong evidence to reject the null hypothesis of equal variance.

We have the fitted data that has no constant variance and is not distributed normally. Therefore, we use permutation test instead of ANOVA to test the equality of means.

Permutation: H_0 : There is no difference in E.Coli across Turbidity H_1 : E.Coli differs across Turbidity

```
## [1] 0
```

p-value = 0, therefore, there is evidence that E.Coli differs across turbidity in NZ rivers, from 1990-2017. # Period 2 (1998-2017) We can see that there tends to be similar trends in river health indicators to those from 1990-2017. However, E.Coli shows a greater range of measurements and frequencies of all indicators are higher (Fig. 8, Fig. 9).

```
##  waterq98.num_id waterq98.E.Coli waterq98.Phosphorus waterq98.Nitrogen
##  Min.   :   2    Min.   : 38.0   Min.   : 72.0       Min.   : 37.0
##  1st Qu.: 730    1st Qu.: 80.0   1st Qu.:246.0       1st Qu.:119.0
##  Median :1629    Median :119.0   Median :251.0       Median :120.0
##  Mean   :1614    Mean   :107.9   Mean   :227.2       Mean   :183.9
##  3rd Qu.:2417    3rd Qu.:120.0   3rd Qu.:251.0       3rd Qu.:251.0
##  Max.   :3153    Max.   :247.0   Max.   :252.0       Max.   :324.0
##  waterq98.Longitude  waterq98.lat    waterq98.Turbidity
##  Min.   :167.5      Min.   :-46.57   Min.   : 36.0
##  1st Qu.:171.1      1st Qu.:-44.19   1st Qu.:119.0
##  Median :174.3      Median :-39.72   Median :120.0
##  Mean   :173.5      Mean   :-40.86   Mean   :179.9
##  3rd Qu.:175.8      3rd Qu.:-38.09   3rd Qu.:252.0
##  Max.   :177.9      Max.   :-35.11   Max.   :336.0
```

```
## Using s_id, dominant_landcover, period, Trend, percent_annual_change as id variables
```

Fit E.Coli with Turbidity to analyse the difference in all 3 periods.

Residual vs fitted values are not normally distributed about zero, with differing 'heights'. This suggests non-constant variance (Fig. 10). The Normal Q-Q plot shows that the data is not normally distributed, as it does not show a straight line (Fig. 10).

Levene Test: H_0 : equal variance H_1 : not all variance are equal

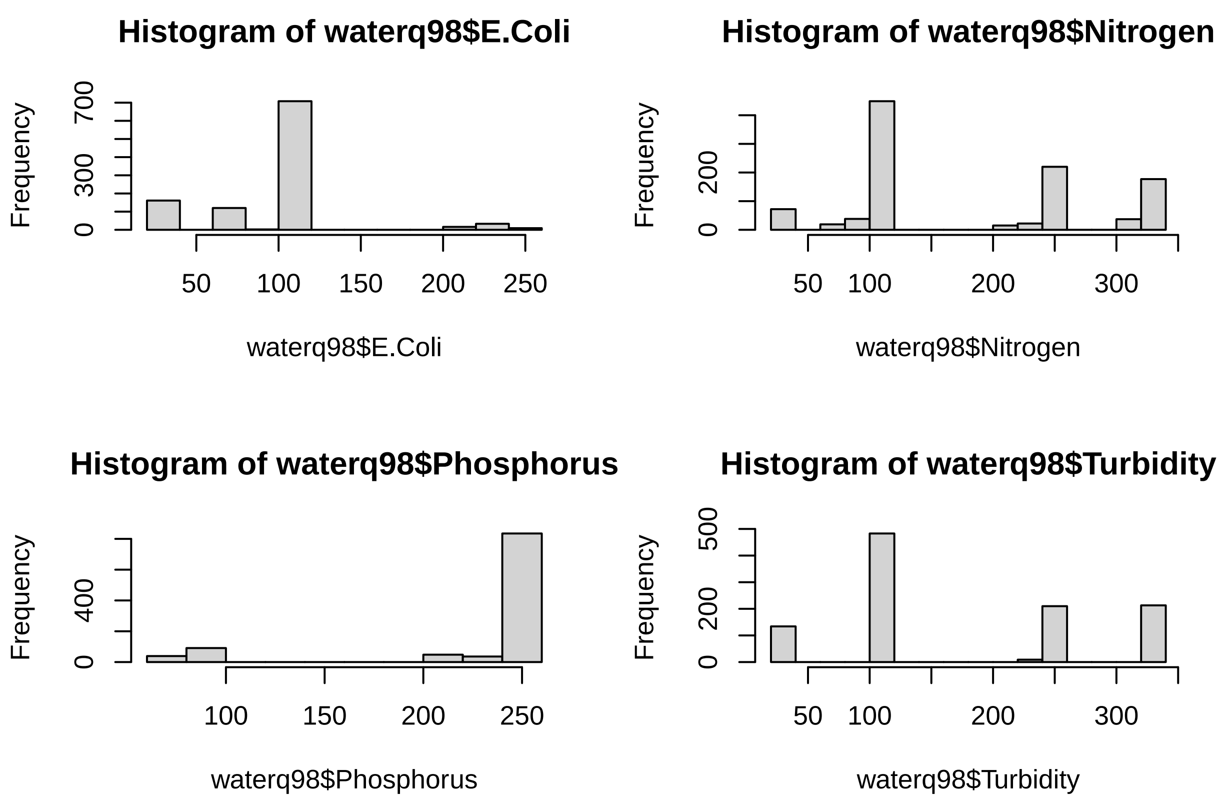

Figure 8: Histograms for river health indicators; E.Coli, Nitrogen, Phosphorus and Turbidity in NZ rivers from 1998-2017

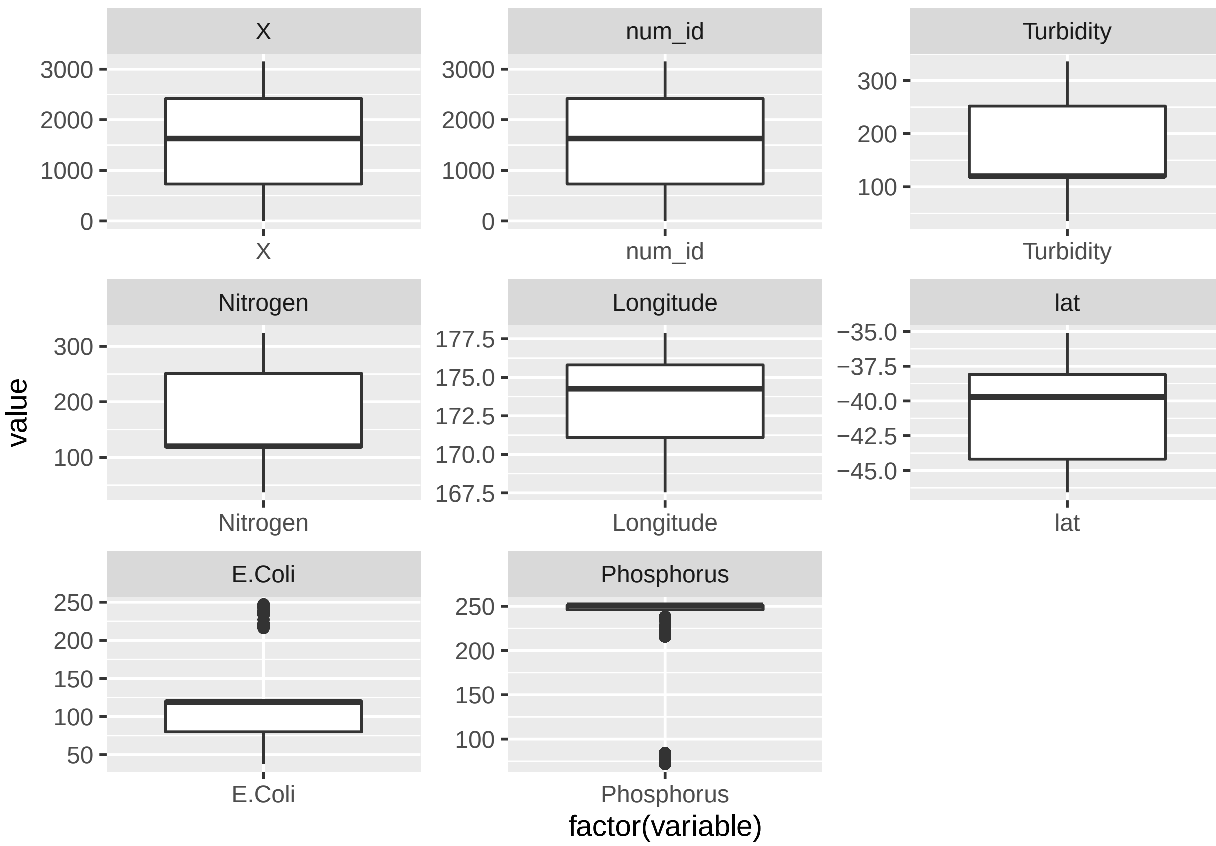

Figure 9: Boxplots of the distributions of NZ river health and information variables from 1998-2017

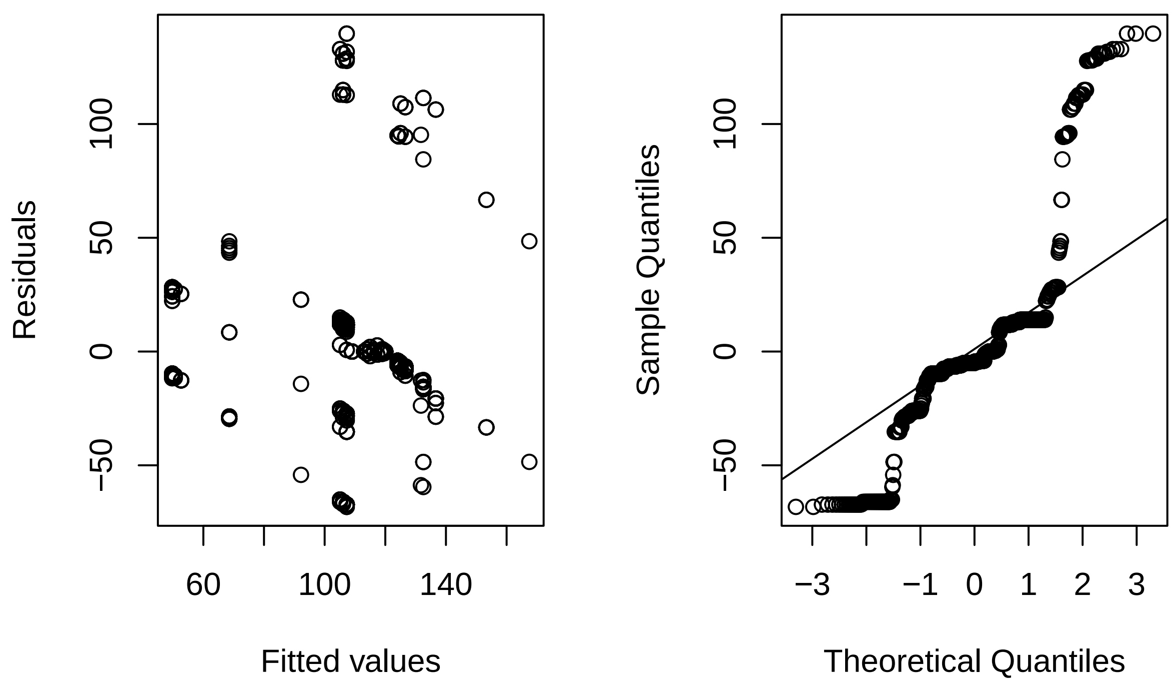

Figure 10: Plot of residuals vs. fitted values and normal Q-Q plot for E.Coli and Turbidity factors in NZ rivers; 1998-2017

```r
leveneTest(waterq98$E.Coli ~ as.factor(waterq98$Turbidity), data = waterq98)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
## group   32  5.1587 < 2.2e-16 ***
##       1016
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-statistic = 5.1587, and p-value = 2.2e-16, Levene's test shows strong evidence to reject the null hypothesis of equal variance.

We have the Fitted data that has no constant variance and is not distributed normally. Therefore we use permutation test instead of anova.

Permutation: H_0 : There is no difference in E.Coli across Turbidity H_1 : E.Coli differs across Turbidity

```r
Fobs<-anova(mod3)$F[1]
Fnull<-rep(NA,2000)
for(t in 1:2000)
{
reorder<-sample(waterq98$Turbidity)
Fnull[t]<-(anova(lm(waterq98$E.Coli~as.factor(reorder))))$F[1]
}
p<-sum(Fnull>=Fobs)/2000
p
```

```
## [1] 0
```

p-value = 0, therefore there is evidence that E.Coli differs across turbidity in NZ rivers, throughout the period of 1998-2017.

Period 3 (2008-2017)

This period tends to show lower frequencies of all river health indicators (Fig. 11, Fig. 12)

```
##  waterq08.num_id waterq08.E.Coli waterq08.Phosphorus waterq08.Nitrogen
##  Min.   :   3    Min.   : 36.0   Min.   : 36.0       Min.   : 36.0
##  1st Qu.: 617    1st Qu.: 78.0   1st Qu.:115.0       1st Qu.:114.0
##  Median :1297    Median :118.0   Median :119.0       Median :120.0
##  Mean   :1455    Mean   :102.6   Mean   :103.4       Mean   :160.5
##  3rd Qu.:2320    3rd Qu.:120.0   3rd Qu.:120.0       3rd Qu.:251.0
##  Max.   :3154    Max.   :247.0   Max.   :120.0       Max.   :324.0
##  waterq08.Longitude  waterq08.lat    waterq08.Turbidity
##  Min.   :167.5      Min.   :-46.57   Min.   : 36.0
##  1st Qu.:171.3      1st Qu.:-44.05   1st Qu.:115.0
##  Median :174.3      Median :-40.24   Median :120.0
##  Mean   :173.6      Mean   :-40.96   Mean   :157.5
##  3rd Qu.:175.7      3rd Qu.:-38.26   3rd Qu.:251.0
##  Max.   :177.9      Max.   :-35.04   Max.   :336.0
```

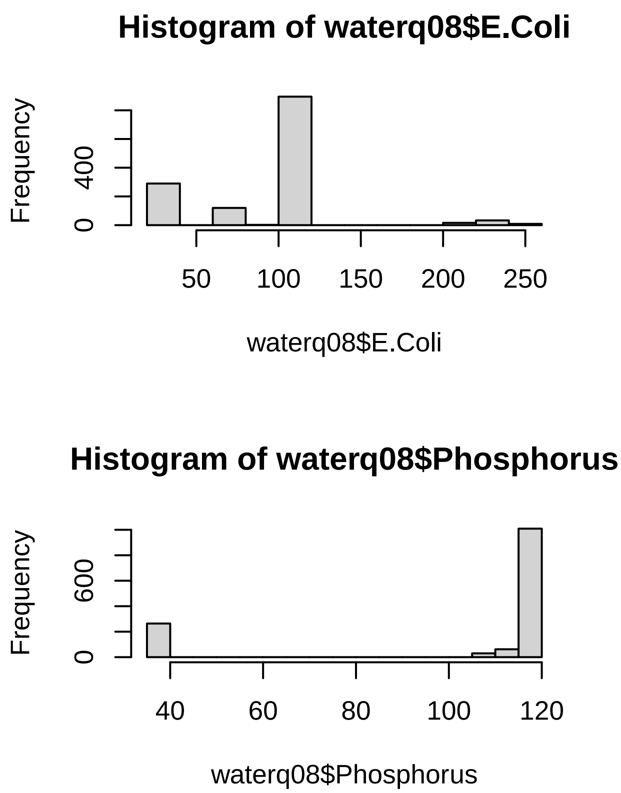

Figure 11: Histograms for river health indicators; E.Coli, Nitrogen, Phosphorus and Turbidity in NZ rivers from 2008-2017

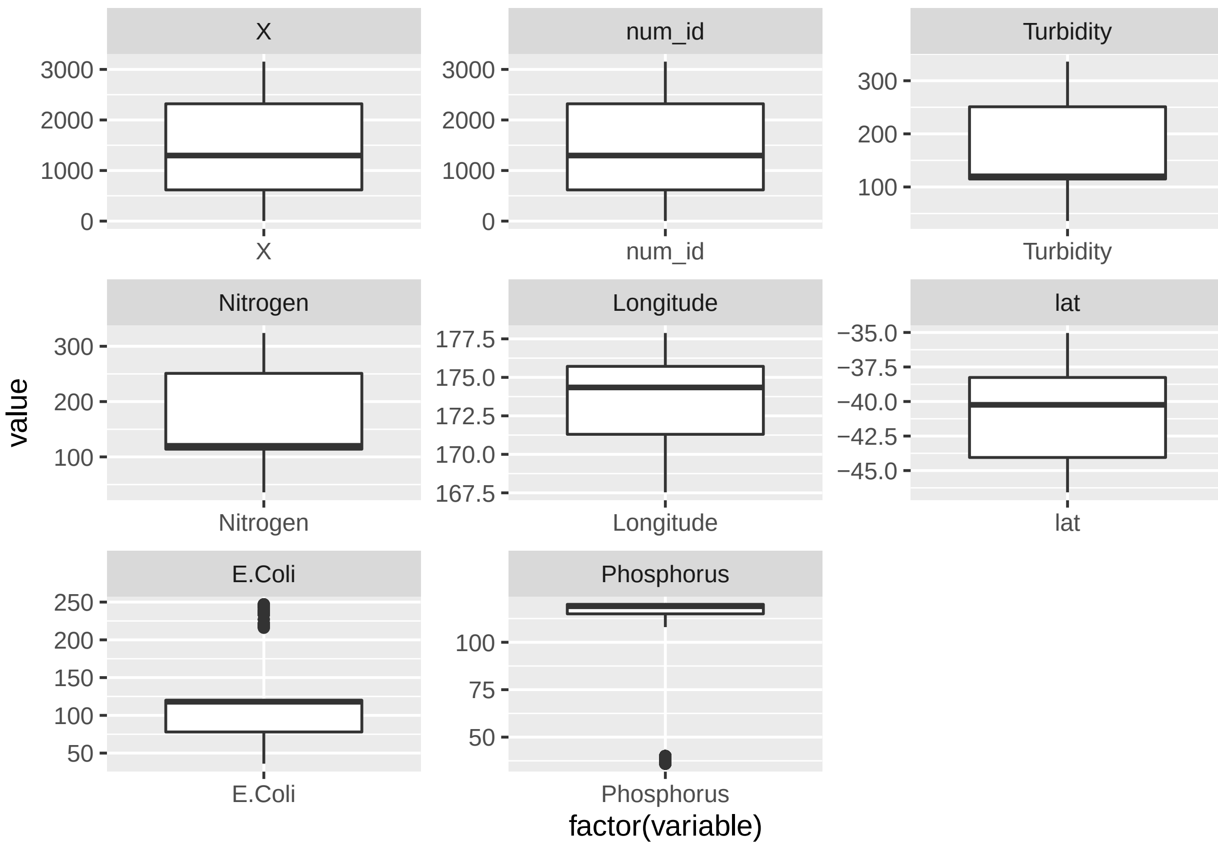

Figure 12: Boxplots of the distributions of NZ river health and information variables from 2008-2017

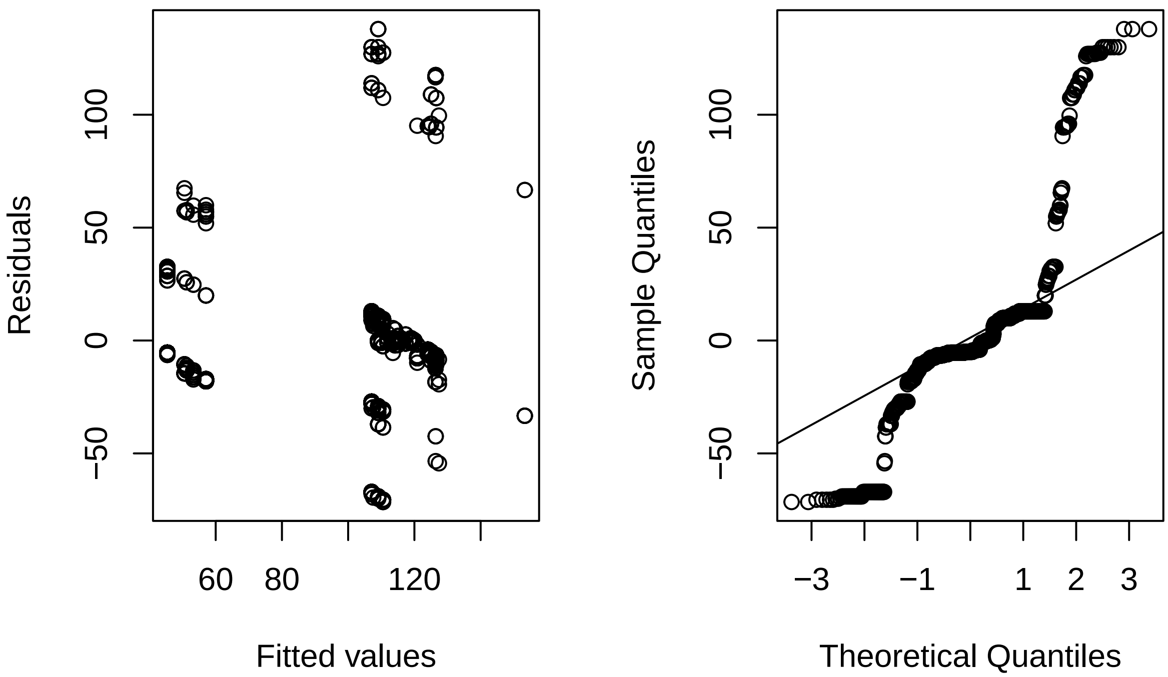

Figure 13: Plot of residuals vs. fitted values and normal Q-Q plot for E.Coli and Turbidity factors in NZ rivers; 2008-2017

```
## Using s_id, dominant_landcover, period, Trend, percent_annual_change as id variables
```

Fit E.Coli with Turbidity to analysis the difference in all 3 periods.

Residual vs fitted values are not normally distributed about zero, with differing 'heights'. This suggests non-constant variance (Fig. 13). The Normal Q-Q plot shows that the data is not normally distributed, as it does not show a straight line (Fig. 13).

Levene Test: H_0 : equal variance H_1 : not all variance are equal

```
leveneTest(waterq08$E.Coli ~ as.factor(waterq08$Turbidity), data = waterq08)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
## group   35  4.0892 4.218e-14 ***
##       1329
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-statistic = 4.0892, and p-value = 4.218e-14, Levene's test shows strong evidence to reject the null hypothesis of equal variance.

Fitted data shows no homogeneity of variances and is not distributed normally. Therefore, for equality testing, we use permutation test instead of ANOVA.

Permutation: H_0 : There is no difference in E.Coli across Turbidity H_1 : E.Coli differs across Turbidity

```
Fobs<-anova(mod4)$F[1]
Fnull<-rep(NA,2000)
for(t in 1:2000)
{
reorder<-sample(waterq08$Turbidity)
Fnull[t]<-(anova(lm(waterq08$E.Coli~as.factor(reorder))))$F[1]
}
p<-sum(Fnull>=Fobs)/2000
p
```

```
## [1] 0
```

p-value = 0, therefore, there is evidence that E.Coli differs across turbidity, between 2008 and 2017.

Thus, we found evidence that E.Coli differs across turbidity in NZ rivers, across all three periods of time; 1990-2017, 1998-2017, 2008-2017.