HAB Modeling

Table of Contents

[Variable Selection for Predictive Model 2](#_Toc515412791)

[MC Sum from grab sample as response 2](#_Toc515412792)

[Stepwise Regression 6](#_Toc515412793)

[Univariate Scatterplots 8](#_Toc515412794)

[Forward Stepwise Regression 18](#_Toc515412795)

[Final Model 20](#_Toc515412796)

[MC Sum from SPATTs as Response Variable 21](#_Toc515412797)

[Subset Analysis 21](#_Toc515412798)

[Stepwise Regression 24](#_Toc515412799)

[Univariate Scatterplots 26](#_Toc515412800)

[Final Model 28](#_Toc515412801)

[mcyE gene as Response 29](#_Toc515412802)

[Subset Analysis 29](#_Toc515412803)

[Step-wise Regression 30](#_Toc515412804)

[Conclusion from Subset Regression 32](#_Toc515412805)

[K-fold Model Validation 32](#_Toc515412806)

[Sucessful Models 33](#_Toc515412807)

[SUM ~ MCYE + OP + Medium\_impervious 33](#_Toc515412808)

[SUM ~ MCYE 36](#_Toc515412809)

[SPATTS SUM ~ MCYE + Herbaceous 39](#_Toc515412810)

[SPATTS SUM ~ Herbaceous 41](#_Toc515412811)

[Failed Models 44](#_Toc515412812)

[SUM ~ Medium\_impervious 44](#_Toc515412813)

[SUM ~ OP 46](#_Toc515412814)

[SUM ~ OP + Turbidity 49](#_Toc515412815)

[SPATTS SUM ~ MCYE 51](#_Toc515412816)

[Linear Mixed Model 54](#_Toc515412817)

[Ecoregions 60](#_Toc515412818)

[Present/Absent of Zebra Mussels 63](#_Toc515412819)

[SPATTS vs SUM 64](#_Toc515412820)

[R Packages 64](#_Toc515412821)

# Variable Selection for Predictive Model

We will search for the best possible set of variables that explains our model. AN exhuastive search will be done by subset analysis for best canidate variables. Subset analysis is done with regsubset() from *leaps* package in R.

The MC congeners and categorical variables are not observed.

With the best variables from the regression subset, a forward and backward step-wise regression will be preformed to further search the best fit model. First, variables will be backwardly selected based on Anova type II Sum of Square test. Type II SS tests each variable after accounting all the other variable. This assumes no interactions between predictor variables.

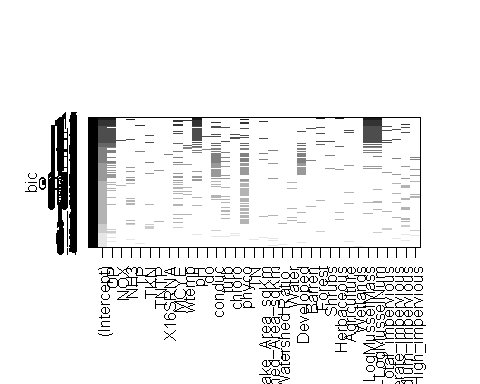
Our best model will be chosen by the lowest Bayesian Information Criterion (BIC). The goal is to have a model that is simple and is robust.

The chosen model(s) will be further validated by k-fold repeated cross-validation.

### MC Sum from grab sample as response

Here all analysis is done on an averaged dataset based on each lake. This results with observations. Each response variable is assessed seperatly.

MCSUMresponseAverage <- LOGTransformed %>% #Average our data  
 group\_by(LK\_CODE) %>%  
 summarise\_all(funs(mean(., na.rm=TRUE))) %>%  
 select(-1:-7, -16:-28, -30:-47, -56, -57, -75:-83)  
   
zFull <- regsubsets(SUM ~ .,  
 nvmax=6  
 nbest=100,   
 really.big=   
 method = "exhaustive")

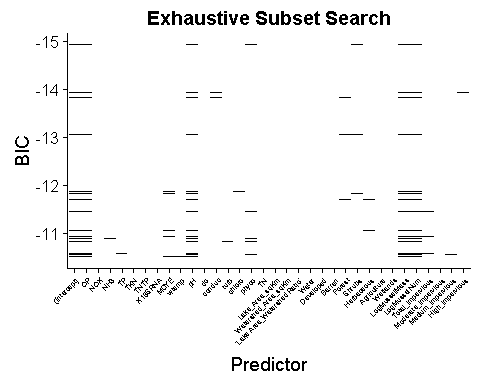


Subset Analysis: Exhaustive Search

Each row is a model. Each shaded rectangles in each column signify being included. The frequency of the variables being included can be important. The presence of one variable in the model and the effect of another variables presence/absence can also be important.

We see OP, *mcyE*, turb, phyco, pH, conduc, forest, shrubs, mussle mass/number and medium impervious areas frequently included in the lowest BIC values (Higher up on the y-axis). They may signify some imporntance.

*There is a clearer plot in the next page*



Subset Analysis for Predicting MC Sum: Top 15 Models from Exhaustive Search

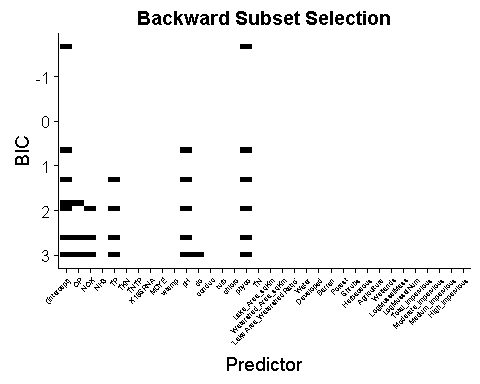
*A plot of a portion of the full subset. BIC range of -10 to -15 is shown here.As the BIC is lower (higher on the y-axis), the better the model.*

The variables in the top 3 models includes orthophosphate, *mcyE*, pH, mussel mass/number and area of impervious surfaces.

We can also run a forward and a backward subset regression.

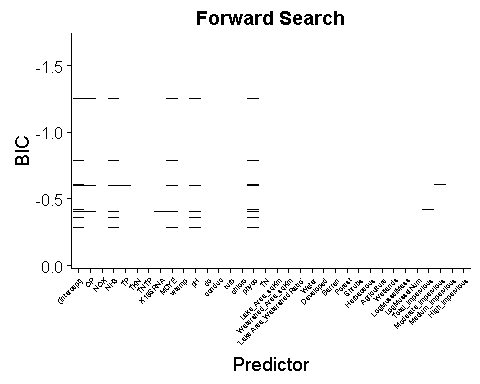
R Code

backward.set <-  
 regsubsets(SUM~.,  
 nvmax=6,  
 nbest=100,  
 data=MCSUMresponseAverage,  
 method="backward")  
  
forward.set <-  
 regsubsets(SUM~.,  
 nvmax=6,  
 nbest=100,  
 data=MCSUMresponseAverage,  
 method="forward")



Subset Analysis for Predicting MC Sum: Backward Search

The backward selection analysis starts with orthophosphate, nitrate+nitrite, total phosphorus, temperature, dissolved oxygen and pH. As the model selection is improved, the best model with phycocyanin by itself had the lowest BIC. Other models with slightly higher BIC included pH, total phosphorus and orthophosphate.



Subset Analysis on Predicting MC Sum: Forward Selection

Orthophosphate, ammonia, *mcyE*, pH, phyco, and Lake:Watershed ratio was found to be the best model.

Orthophosphate and *mcyE* both seem to be the best variables as they appear in all subset analysis. Forward and backward subset analysis don’t converge to the same set of models.

#### Stepwise Regression

Take all the best variables from regression subset analysis and remove variables. We will backwardly remove each variable by F-test when

zamod1 <- lm(SUM~  
 OP +  
 MCYE +  
 turb +  
 Herbaceous +  
 Forest +  
 pH +  
 phyco +  
 LogMusselMass +  
 LogMusselNum +  
 Total\_impervious +  
 Medium\_impervious +  
 Shrubs,  
 data=MCSUMresponseAverage)  
  
Anova(zamod1)

## Anova Table (Type II tests)  
##   
## Response: SUM  
## Sum Sq Df F value Pr(>F)   
## OP 0.85562 1 11.3316 0.003930 \*\*  
## MCYE 0.57577 1 7.6254 0.013904 \*   
## turb 0.03985 1 0.5278 0.478042   
## Herbaceous 0.03208 1 0.4249 0.523756   
## Forest 0.08756 1 1.1596 0.297517   
## pH 0.84194 1 11.1504 0.004161 \*\*  
## phyco 0.19585 1 2.5939 0.126828   
## LogMusselMass 0.89545 1 11.8592 0.003338 \*\*  
## LogMusselNum 0.76461 1 10.1263 0.005790 \*\*  
## Total\_impervious 0.05274 1 0.6985 0.415589   
## Medium\_impervious 0.04533 1 0.6003 0.449771   
## Shrubs 0.06385 1 0.8456 0.371443   
## Residuals 1.20811 16   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

We will drop one variable with the highest value. Herbaceous is dropped in this instance. This is repeated until the variables are all signifigant.

## Anova Table (Type II tests)  
##   
## Response: SUM  
## Sum Sq Df F value Pr(>F)   
## OP 0.95563 1 12.7721 0.002338 \*\*  
## MCYE 0.57924 1 7.7416 0.012770 \*   
## turb 0.04997 1 0.6678 0.425112   
## Herbaceous 0.11692 1 1.5627 0.228207   
## Forest 0.10326 1 1.3801 0.256275   
## pH 0.94411 1 12.6182 0.002450 \*\*  
## phyco 0.17607 1 2.3532 0.143426   
## LogMusselMass 0.83521 1 11.1627 0.003871 \*\*  
## LogMusselNum 0.71299 1 9.5292 0.006689 \*\*  
## Total\_impervious 0.05705 1 0.7625 0.394719   
## Medium\_impervious 0.05341 1 0.7138 0.409929   
## Residuals 1.27197 17   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Next subsequents steps are not shown.

## Anova Table (Type II tests)  
##   
## Response: SUM  
## Sum Sq Df F value Pr(>F)   
## OP 0.23621 1 1.8251 0.18930   
## MCYE 0.60397 1 4.6667 0.04096 \*  
## turb 0.35107 1 2.7126 0.11259   
## Medium\_impervious 0.20716 1 1.6007 0.21795   
## Residuals 3.10612 24   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Our last four variables are Orthophosphate, *mcyE*, turbidity, and area of medium impervious surfaces.

Medium impervious surface gets dropped

## Anova Table (Type II tests)  
##   
## Response: SUM  
## Sum Sq Df F value Pr(>F)   
## OP 0.6061 1 4.5736 0.042426 \*   
## MCYE 1.1897 1 8.9768 0.006095 \*\*  
## turb 0.2594 1 1.9572 0.174097   
## Residuals 3.3133 25   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Turbidity gets dropped

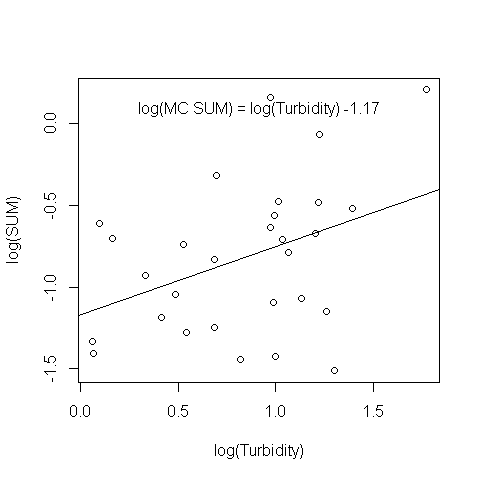
## Anova Table (Type II tests)  
##   
## Response: SUM  
## Sum Sq Df F value Pr(>F)   
## OP 1.1093 1 8.0728 0.008621 \*\*  
## MCYE 1.2875 1 9.3694 0.005073 \*\*  
## Residuals 3.5727 26   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

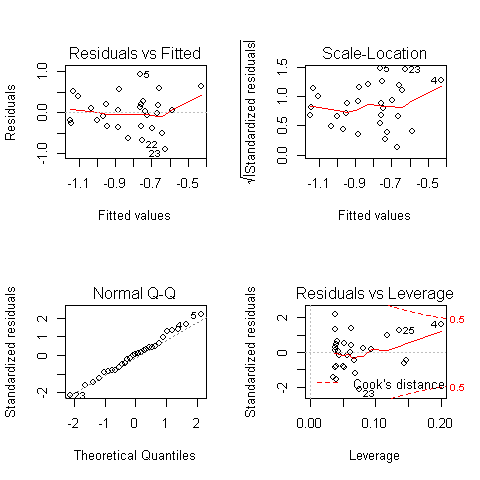
At this point, *mcyE*, turbidity, orthophosphate and area of medium impervious surface will be our pool of variables for a forward stepwise regression.

Lets assess each variable and see which one is the most signifigant by value.

#### Univariate Scatterplots

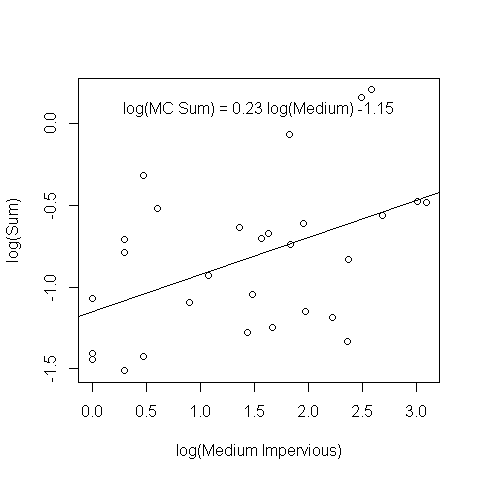
Scatterplots of turbidity, area of impervious surfaces, *mcyE* and orthophosphate scatterplot and residual are shown here.





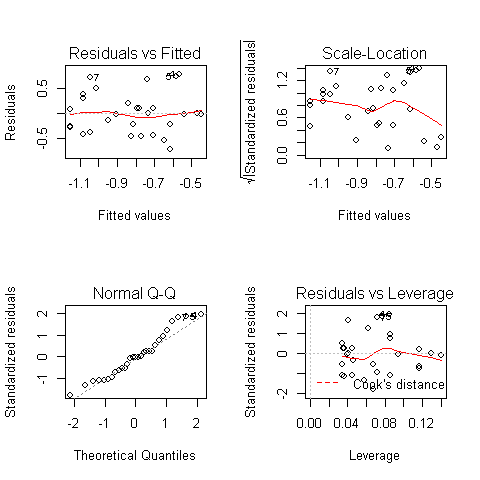
Residual Analysis of log(SUM) ~ log(Turbidity)

**The residuals for turbidity explaining MC Sum is fairly normal. There are a few outliers but does not seem to leverage the correlation signifigantly. However they do seem to pull the correlation upward**



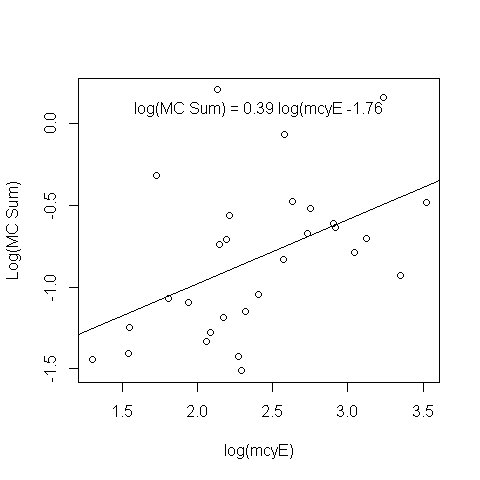
MC Sum as a Function of Impervious Surfaces

`



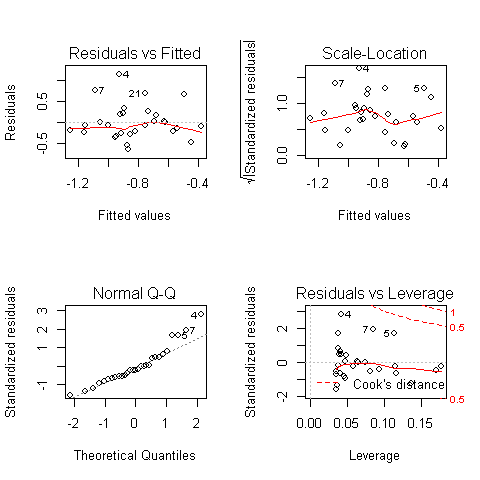
Residual Analysis of SUM ~ Medium Impervious surface

**Residuals are normally scattered for area of impervious surfaces. No leverage points.**



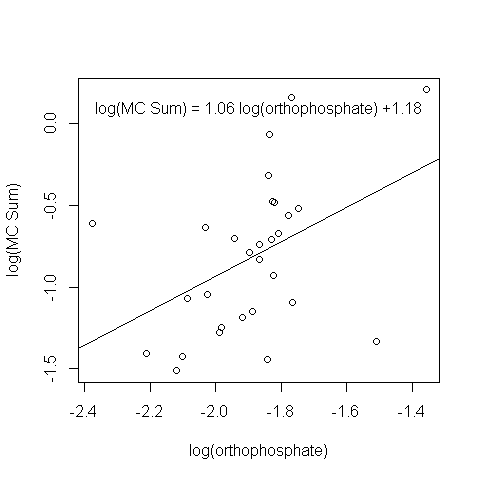
MC Sum as a function of mcyE

`



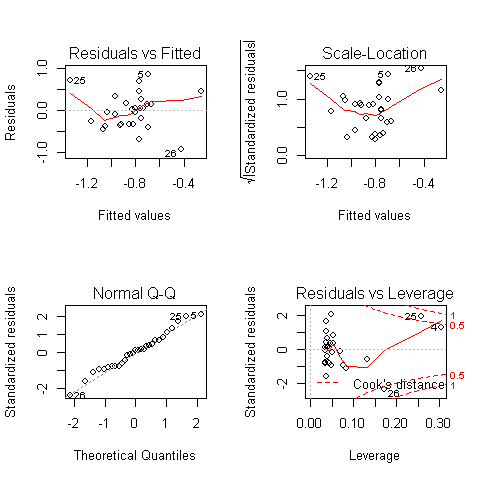
Residual Analysis of SUM ~ mcyE

**Residual of *mcyE* explaining MC SUM is normal and seems fairly random. Variation is not scattered at the low and high end.**



MC Sum as a function of Orthophosphate

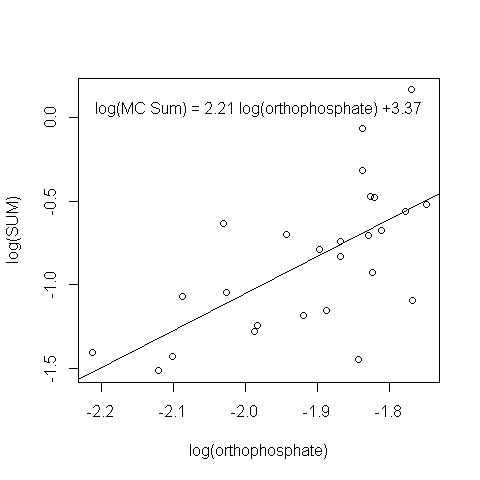
`



Residual Analysis of SUM~Orthophosphate

**Orthophosphate is driven by three signifigant outliers.**  Lets see the effect of removing the outliers and evaluate the relationship of orthophosphate to MC sum.

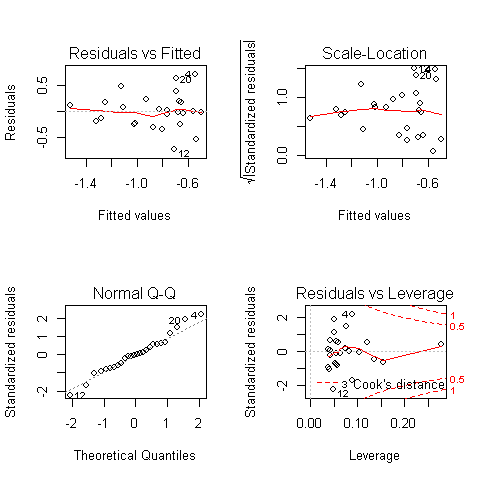
Although, removing all three is not a good idea.



MC Sum as a function of Orthophosphate with omitted outliers

`

##   
## Call:  
## lm(formula = SUM ~ OP, data = omitted, na.action = na.omit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.73774 -0.22009 -0.01207 0.17599 0.70645   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.373 1.044 3.230 0.003571 \*\*   
## OP 2.213 0.546 4.053 0.000461 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3394 on 24 degrees of freedom  
## Multiple R-squared: 0.4064, Adjusted R-squared: 0.3816   
## F-statistic: 16.43 on 1 and 24 DF, p-value: 0.0004606



Residual Analysis of MC Sum ~ Orthophosphate with outliers removed

## Anova Table (Type II tests)  
##   
## Response: SUM  
## Sum Sq Df F value Pr(>F)   
## OP 1.8926 1 16.429 0.0004606 \*\*\*  
## Residuals 2.7648 24   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**With Brighton, Silver, and Stony creek ommited, we see orthophosphate signifigant effect on MC sum. However, the residual does not have a uniform variation. I has a cone shape**  ..

#### Forward Stepwise Regression

## Anova Table (Type II tests)  
##   
## Response: SUM  
## Sum Sq Df F value Pr(>F)   
## MCYE 1.3621 1 7.855 0.009265 \*\*  
## Residuals 4.6820 27   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Anova Table (Type II tests)  
##   
## Response: SUM  
## Sum Sq Df F value Pr(>F)   
## OP 1.1839 1 6.5773 0.01621 \*  
## Residuals 4.8601 27   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Anova Table (Type II tests)  
##   
## Response: SUM  
## Sum Sq Df F value Pr(>F)   
## turb 0.9390 1 4.9662 0.03437 \*  
## Residuals 5.1051 27   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Anova Table (Type II tests)  
##   
## Response: SUM  
## Sum Sq Df F value Pr(>F)   
## Medium\_impervious 1.3258 1 7.5871 0.01039 \*  
## Residuals 4.7182 27   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

From here, our starting variable for a forward selection is *mcyE*

## Anova Table (Type II tests)  
##   
## Response: SUM  
## Sum Sq Df F value Pr(>F)   
## MCYE 1.2875 1 9.3694 0.005073 \*\*  
## OP 1.1093 1 8.0728 0.008621 \*\*  
## Residuals 3.5727 26   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Anova Table (Type II tests)  
##   
## Response: SUM  
## Sum Sq Df F value Pr(>F)   
## MCYE 1.1856 1 7.8651 0.009409 \*\*  
## turb 0.7625 1 5.0584 0.033201 \*   
## Residuals 3.9194 26   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Anova Table (Type II tests)  
##   
## Response: SUM  
## Sum Sq Df F value Pr(>F)   
## MCYE 0.5927 1 3.7351 0.06424 .  
## Medium\_impervious 0.5564 1 3.5065 0.07242 .  
## Residuals 4.1256 26   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Orthophosphate is the second most signifigant

## Anova Table (Type II tests)  
##   
## Response: SUM  
## Sum Sq Df F value Pr(>F)   
## MCYE 0.7817 1 5.6528 0.02539 \*  
## OP 0.6684 1 4.8332 0.03739 \*  
## Medium\_impervious 0.1155 1 0.8351 0.36954   
## Residuals 3.4572 25   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Anova Table (Type II tests)  
##   
## Response: SUM  
## Sum Sq Df F value Pr(>F)   
## MCYE 1.1897 1 8.9768 0.006095 \*\*  
## OP 0.6061 1 4.5736 0.042426 \*   
## turb 0.2594 1 1.9572 0.174097   
## Residuals 3.3133 25   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Adding a third variable does not improve our model. SUM~OP+MCYE is most likely the best model.

#### Final Model

Lets compare each model by AIC. All possible combination of models are listed here:

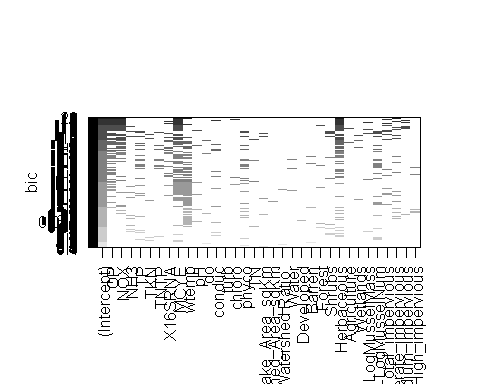
* FullModel = SUM ~ MCYE + turb + Medium\_impervious + OP
* Model 1 = SUM ~ turb + Medium\_impervious + OP
* Model 2 = SUM ~ MCYE + Medium\_impervious + turb
* Model 3 = SUM ~ turb + MCYE + OP
* Model 4 = SUM ~ Medium\_impervious + MCYE + OP
* Model 5 = SUM ~ MCYE + OP
* Model 6 = SUM ~ MCYE + turb
* Model 7 = SUM ~ MCYE + Medium\_impervious
* Model 8 = SUM ~ Medium\_impervious + OP
* Model 9 = SUM ~ Medium\_impervious + turb
* Model 10 = SUM ~ Medium\_impervious
* Model 11 = SUM ~ OP
* Model 12 = SUM ~ turb
* SUM ~ *mcyE* + orthophosphate is our best model.
* SUM ~ turbidity + orthophosphate + *mcyE* is second best
* SUM ~ *mcyE* + Medium Impervious Area + turbidity is our third best.

These models will be validated later on with K-fold cross validation.

### MC Sum from SPATTs as Response Variable

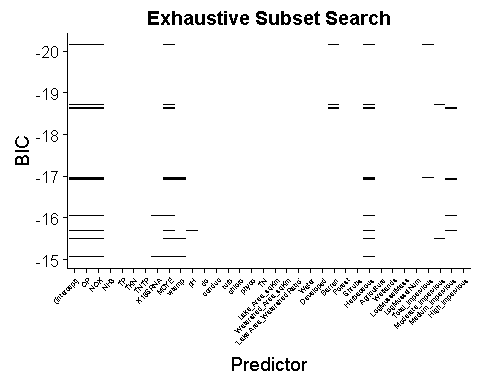
#### Subset Analysis

#Setup   
SPATTSresponseAverage <- LOGTransformed %>%   
 group\_by(LK\_CODE) %>%  
 summarise\_all(funs(mean(., na.rm=TRUE))) %>%  
 select(-1:-7, -16:-29, -30:-45,-47, -56, -57, -75:-83)  
   
zsFull <- regsubsets(S\_SUM ~ .,  
 nvmax=6,   
 nbest=100,   
 really.big=TRUE,  
 SPATTSresponseAverage,   
 method = "exhaustive")

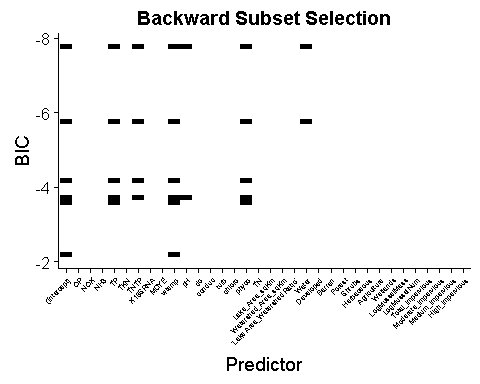


Subset Analysis for Predicting MC from SPATT: Exhaustive Search

Orthophosphate, nitrate+nitrite, *16s rRNA*, *mcyE*, water temprature, Herbaceous, Mussel Mass, and area of medium impervious surface are selected as candidate variables.

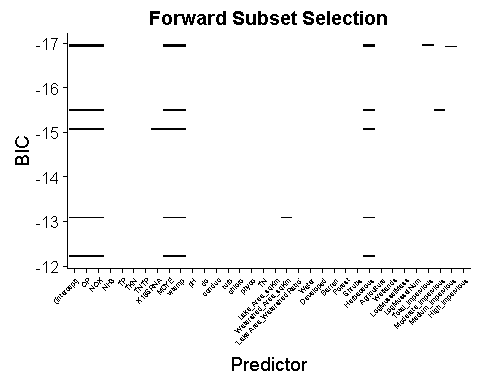


Subset Analysis for Predicting MC from SPATT:Top 15 Models Exhaustive Search



Subset Analysis for Predicting MC Sum from SPATT: Backward Search

From the backward selection analysis, total phosphorus, TN:TP ratio, water temperature, phycocyanin and water land use may be signifigant variables.



Subset Analysis for Predicting MC Sum from SPATT: Forward Search

From the forward subset analysis, orthophosphate, nitrate+nitrite, *16s rRNA*, *mcyE*, water temperature, Herbaceous, Mussel mass/number, area of impervious surfaces may also have an influence on MC Sum.

#### Stepwise Regression

model1 <- lm(S\_SUM~ #Starting out all variables, then backward selection  
 OP +  
 NOX +  
 X16SRNA +  
 MCYE +  
 wtemp +  
 Herbaceous +  
 LogMusselMass +  
 LogMusselNum +  
 Medium\_impervious ,  
 data=SPATTSresponseAverage)  
Anova(model1)

## Anova Table (Type II tests)  
##   
## Response: S\_SUM  
## Sum Sq Df F value Pr(>F)   
## OP 0.5118 1 1.7630 0.19998   
## NOX 0.0043 1 0.0148 0.90453   
## X16SRNA 0.4327 1 1.4903 0.23710   
## MCYE 1.0293 1 3.5452 0.07512 .  
## wtemp 0.1210 1 0.4167 0.52633   
## Herbaceous 0.5394 1 1.8579 0.18881   
## LogMusselMass 0.0031 1 0.0105 0.91931   
## LogMusselNum 0.0677 1 0.2331 0.63472   
## Medium\_impervious 0.0054 1 0.0184 0.89342   
## Residuals 5.5162 19   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

model1 <- lm(S\_SUM~   
 OP +  
 MCYE +  
 Herbaceous ,  
 data=SPATTSresponseAverage)  
Anova(model1)

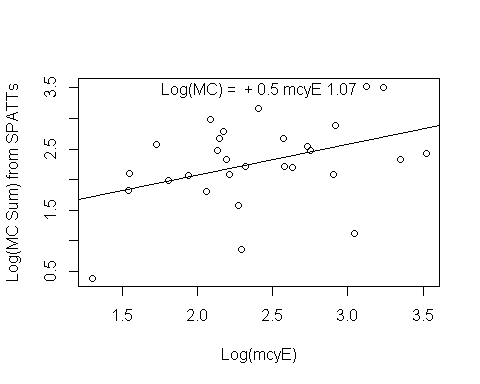
## Anova Table (Type II tests)  
##   
## Response: S\_SUM  
## Sum Sq Df F value Pr(>F)   
## OP 0.9866 1 3.7967 0.0626637 .   
## MCYE 2.0175 1 7.7636 0.0100261 \*   
## Herbaceous 4.7509 1 18.2817 0.0002434 \*\*\*  
## Residuals 6.4968 25   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Anova Table (Type II tests)  
##   
## Response: S\_SUM  
## Sum Sq Df F value Pr(>F)   
## MCYE 2.1252 1 7.3838 0.011553 \*   
## Herbaceous 3.9385 1 13.6834 0.001019 \*\*  
## Residuals 7.4835 26   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

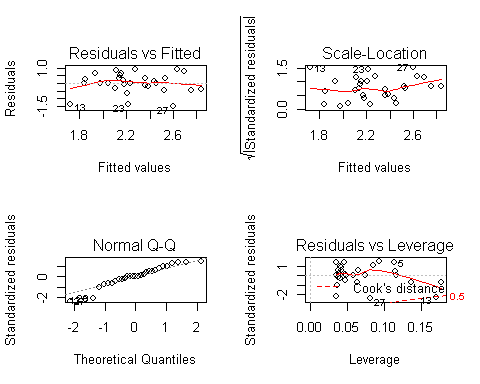
## Anova Table (Type II tests)  
##   
## Response: S\_SUM  
## Sum Sq Df F value Pr(>F)   
## Herbaceous 4.0479 1 11.374 0.002263 \*\*  
## Residuals 9.6087 27   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Anova Table (Type II tests)  
##   
## Response: S\_SUM  
## Sum Sq Df F value Pr(>F)   
## MCYE 2.2347 1 5.2825 0.02951 \*  
## Residuals 11.4219 27   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

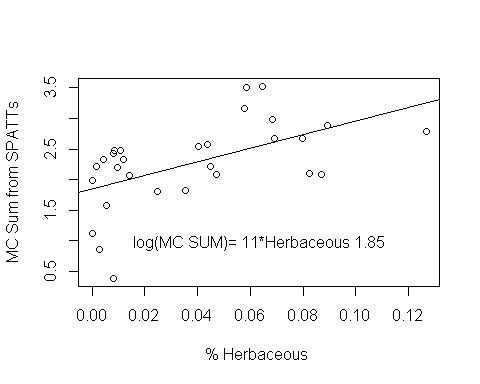
#### Univariate Scatterplots

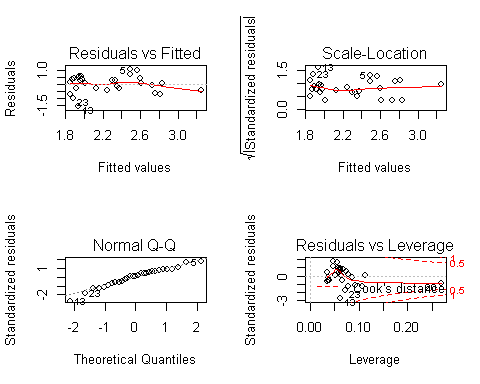


MC from Spatts ~ mcyE



Residual Analysis of MC Sum ~ mcyE





Residual analysis of MC Sum ~ Herbaceous

#### Final Model

* FullModel = S\_SUM ~ MCYE + Herbaceous
* Model 1 = S\_SUM ~ MCYE
* Model 2 = S\_SUM ~ Herbaceous
* Model 3 = S\_SUM ~ Agriculture
* Model 4 = S\_SUM ~ Forest
* Model 5 = S\_SUM ~ Shrubs

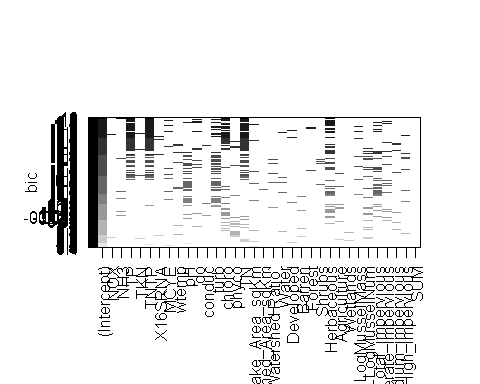
|  |  |  |
| --- | --- | --- |
| * df |  | * AIC |
| * FullModel | * 4 | * 51.01511 |
| * Model1 | * 3 | * 61.27740 |
| * Model2 | * 3 | * 56.26433 |
| * Model3 | * 3 | * 64.66383 |
| * Model4 | * 3 | * 63.98582 |
| * Model5 | * 3 | * 62.41383 |

The best predictor model for MC from SPATTs is *mcyE* and Herbaceous land use.

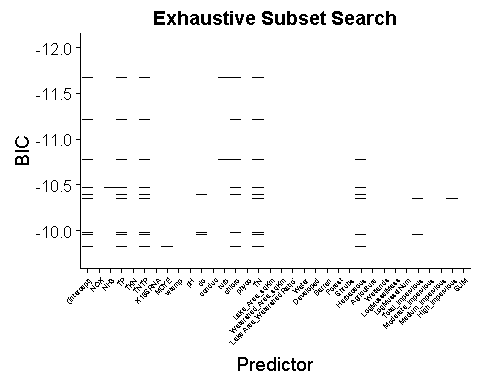
### mcyE gene as Response

#### Subset Analysis

zFull1 <- regsubsets(OP ~ .,  
 nvmax=6,   
 nbest=100,   
 force.out = "SUM",  
 really.big=TRUE,  
 MCSUMresponseAverage,   
 method = "exhaustive")



Subset Analysis Predicting mcyE: Exhaustive Search



Subset Analysis for Predicting mcyE:Top 15 Models Exhaustive Search

#### Step-wise Regression

model1 <- lm(MCYE~  
 NOX +  
 wtemp +  
 do +  
 conduc +  
 Herbaceous +  
 LogMusselNum +  
 High\_impervious +  
 Total\_impervious,  
 data=MCSUMresponseAverage)  
Anova(model1)

## Anova Table (Type II tests)  
##   
## Response: MCYE  
## Sum Sq Df F value Pr(>F)   
## NOX 0.09832 1 0.7485 0.3997488   
## wtemp 1.14356 1 8.7049 0.0094024 \*\*   
## do 1.49517 1 11.3815 0.0038695 \*\*   
## conduc 1.39297 1 10.6035 0.0049548 \*\*   
## Herbaceous 0.04294 1 0.3269 0.5754467   
## LogMusselNum 1.52396 1 11.6006 0.0036144 \*\*   
## High\_impervious 2.55227 1 19.4283 0.0004401 \*\*\*  
## Total\_impervious 0.05951 1 0.4530 0.5105321   
## Residuals 2.10190 16   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Anova Table (Type II tests)  
##   
## Response: MCYE  
## Sum Sq Df F value Pr(>F)   
## wtemp 0.0016 1 0.0074 0.93211   
## LogMusselNum 0.6593 1 3.1160 0.09207 .  
## High\_impervious 0.8687 1 4.1057 0.05563 .  
## Residuals 4.4432 21   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Anova Table (Type II tests)  
##   
## Response: MCYE  
## Sum Sq Df F value Pr(>F)   
## LogMusselNum 0.6860 1 3.3955 0.07889 .  
## High\_impervious 1.2407 1 6.1410 0.02136 \*  
## Residuals 4.4448 22   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Anova Table (Type II tests)  
##   
## Response: MCYE  
## Sum Sq Df F value Pr(>F)   
## High\_impervious 0.8612 1 3.8605 0.06163 .  
## Residuals 5.1308 23   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##   
## Call:  
## lm(formula = MCYE ~ High\_impervious, data = MCSUMresponseAverage)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.85334 -0.30985 -0.09367 0.33147 0.95961   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.3867541 0.1042812 22.888 <2e-16 \*\*\*  
## High\_impervious 0.0014612 0.0007437 1.965 0.0616 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4723 on 23 degrees of freedom  
## (4 observations deleted due to missingness)  
## Multiple R-squared: 0.1437, Adjusted R-squared: 0.1065   
## F-statistic: 3.861 on 1 and 23 DF, p-value: 0.06163

## Anova Table (Type II tests)  
##   
## Response: MCYE  
## Sum Sq Df F value Pr(>F)  
## LogMusselNum 0.2599 1 0.8183 0.3737  
## Residuals 8.5745 27

## Anova Table (Type II tests)  
##   
## Response: MCYE  
## Sum Sq Df F value Pr(>F)  
## wtemp 0.2515 1 0.7912 0.3816  
## Residuals 8.5828 27

None of our variables seem to explain a signifigant relationship. Area of impervious surface has a slight positive effect,

## Conclusion from Subset Regression

MC Sum from the grab samples are best predicted by *mcyE* + Turbidity + area of medium impervious surfaces. MC from SPATTs are best predicted with a model including *mcyE* and herbaceous.

# K-fold Model Validation

Here we will verify each model by K-fold cross-validation. First, the dataset will be divided into “K” folds. One fold is held away for validation, while the remaining are used to train the model. Next the trained model is validated against the validation fold. Each validation, a skill score is recorded. Here the RMSE is assessed along with its standard deviation.

Next the model is tested by calculating the sum of square errors of fitted vs actual on the validation set. This is repeated “K” times. For all “K” time, the prediction errors between the training and testing set are averaged.

A K-fold repeated cross-validation is effective with small datasets. The usual way of evaluating a model by testing the effective predictability.

Using the library(caret) package.

Seed is set to 629. This was chosen by a random generator online:

<https://registrar.uoregon.edu/faculty-staff/random-number-generator>

Here each model that was found to be effective from previous section is tested.

### Sucessful Models

#### SUM ~ MCYE + OP + Medium\_impervious

set.seed(629)   
zdata <- LOGTransformed %>% # Only select variables of interest  
 select(SUM, MCYE, OP, Medium\_impervious) %>%  
 na.omit() # Omit all NA's  
  
Dataframe <- as.data.frame(zdata) # Convert it as a dataframe  
  
# Split our Data if we wanted to test it later. However not neccesary. This seems to be the old way.   
  
ind = createDataPartition(Dataframe$SUM, p=2/3, list=FALSE)   
traindata <- Dataframe[ind,] # Training dataset is 66%  
testdata <- Dataframe[-ind,] # Testing dataset is 33%  
  
#I was playing around with random forest. Ignore this.  
#parameterGrid <- expand.grid(mtry=c(1,2,3))   
  
  
# This is a stored setting variable   
# which sets up how the training resamples the data.   
train\_control <- trainControl(method="repeatedcv",   
 number =10, # The K fold  
 repeats=20 # Repeated times  
   
 # , savePredictions = TRUE  
 )  
  
model <- train(SUM~.,  
 data=traindata,  
 method="lm", # glm, rf, but no lme or mixed model support  
 trControl = train\_control  
 #tuneGrid = parameterGrid  
 )

Model Performance:

## Linear Regression   
##   
## 62 samples  
## 3 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 20 times)   
## Summary of sample sizes: 55, 55, 54, 55, 56, 56, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 0.5397455 0.3803679 0.4599782  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

Our model: SUM ~ MCYE + OP + Medium\_impervious

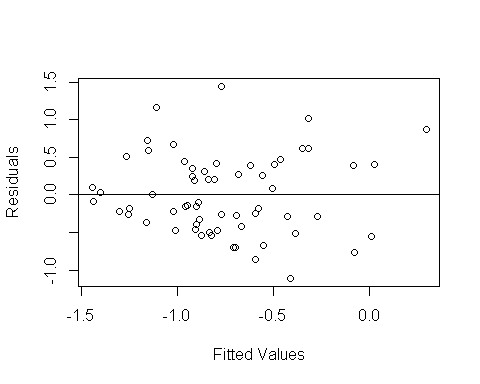
Across all 10-fold cross-validation, the average of the root mean square deviation (RMSE) = 0.5397455 and deviates = 0.14.

Average = 0.38 with = 0.27

Here is the trained model summary statistic:

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.1096 -0.3826 -0.1217 0.3968 1.4436   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.12836 0.51663 0.248 0.80466   
## MCYE 0.16060 0.06374 2.519 0.01453 \*   
## OP 0.81210 0.24754 3.281 0.00176 \*\*  
## Medium\_impervious 0.17989 0.08496 2.117 0.03852 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5355 on 58 degrees of freedom  
## Multiple R-squared: 0.3391, Adjusted R-squared: 0.3049   
## F-statistic: 9.919 on 3 and 58 DF, p-value: 2.25e-05

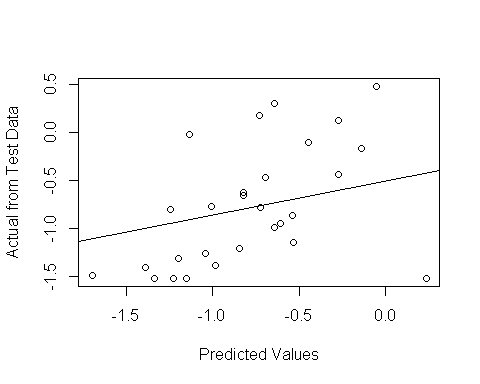
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 0.1283579 | 0.5166343 | 0.2484502 | 0.8046637 |
| MCYE | 0.1606045 | 0.0637448 | 2.5194936 | 0.0145294 |
| OP | 0.8120976 | 0.2475380 | 3.2806983 | 0.0017552 |
| Medium\_impervious | 0.1798908 | 0.0849561 | 2.1174555 | 0.0385197 |



Residual Plot of Trained Model

Using the trained model from k-folds, lets further validate this model and check how well it predicts 20% of the data.

predictionz <- predict(model,testdata)  
  
  
plot(testdata$SUM~predictionz, xlab="Predicted Values", ylab="Actual from Test Data")  
abline(lm(predictionz~testdata$SUM))



summary <- lm(testdata$SUM~predictionz)  
summary(summary)

##   
## Call:  
## lm(formula = testdata$SUM ~ predictionz)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.46023 -0.32420 -0.05463 0.25974 1.01250   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.2290 0.2145 -1.068 0.29548   
## predictionz 0.7022 0.2386 2.942 0.00676 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5528 on 26 degrees of freedom  
## Multiple R-squared: 0.2498, Adjusted R-squared: 0.221   
## F-statistic: 8.658 on 1 and 26 DF, p-value: 0.006764

#### SUM ~ MCYE

## Linear Regression   
##   
## 63 samples  
## 1 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 20 times)   
## Summary of sample sizes: 57, 55, 56, 57, 56, 56, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 0.5616086 0.3226707 0.4762965  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

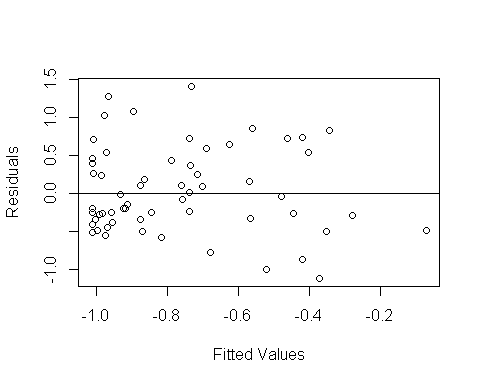
Our model: SUM ~ MCYE

Across all 10-fold cross-validation, the average of the root mean square deviation (RMSE) = 0.5616086 = 0.12 with an average of = 0.32 with = 0.29

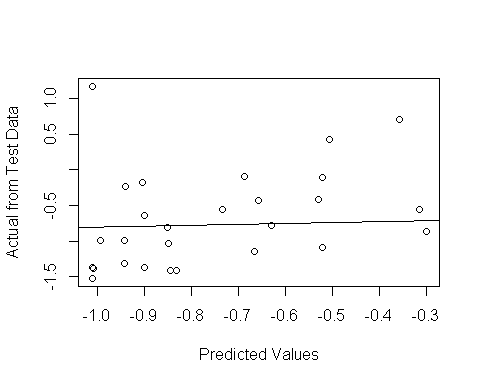
Here is the trained model summary statistic:

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.1148 -0.4284 -0.1951 0.4129 1.4030   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.29235 0.17018 -7.594 2.2e-10 \*\*\*  
## MCYE 0.21564 0.06519 3.308 0.00158 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5672 on 61 degrees of freedom  
## Multiple R-squared: 0.1521, Adjusted R-squared: 0.1382   
## F-statistic: 10.94 on 1 and 61 DF, p-value: 0.001579

## \begin{table}[ht]  
## \centering  
## \begin{tabular}{rrrrr}  
## \hline  
## & Estimate & Std. Error & t value & Pr($>$$|$t$|$) \\   
## \hline  
## (Intercept) & -1.2924 & 0.1702 & -7.59 & 0.0000 \\   
## MCYE & 0.2156 & 0.0652 & 3.31 & 0.0016 \\   
## \hline  
## \end{tabular}  
## \end{table}



Using the trained model from k-folds, lets further validate this model and check how well it predicts the 20% of the data.



##   
## Call:  
## lm(formula = testdata$SUM ~ predictionz)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.6702 -0.5097 -0.1542 0.2525 2.1595   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.1436 0.4406 0.326 0.7471   
## predictionz 1.1171 0.5538 2.017 0.0541 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.656 on 26 degrees of freedom  
## Multiple R-squared: 0.1353, Adjusted R-squared: 0.1021   
## F-statistic: 4.069 on 1 and 26 DF, p-value: 0.0541

#### SPATTS SUM ~ MCYE + Herbaceous

## Linear Regression   
##   
## 55 samples  
## 2 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 20 times)   
## Summary of sample sizes: 49, 51, 49, 50, 50, 49, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 0.6846797 0.4042762 0.5672053  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

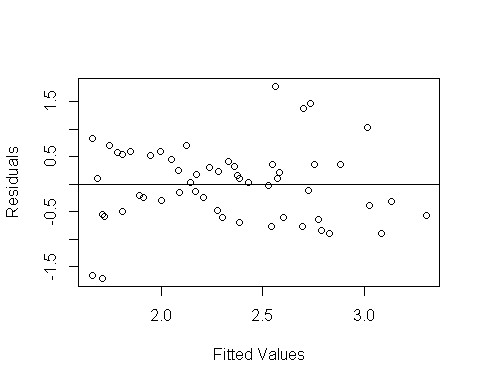
Our model: S\_SUM ~ MCYE + Herbaceous

Across all 10-fold cross-validation, the average of the root mean square deviation (RMSE) = 0.6846797 = 0.22 with an average of = 0.4 with = 0.26

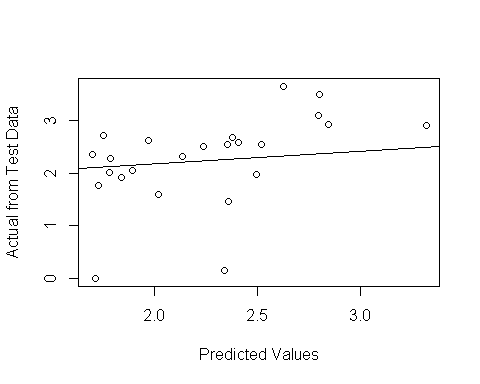
Here is the trained model summary statistic:

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.70911 -0.51704 0.02959 0.39291 1.78182   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.34299 0.26115 5.143 4.17e-06 \*\*\*  
## MCYE 0.21652 0.08631 2.509 0.015277 \*   
## Herbaceous 10.63776 2.74831 3.871 0.000305 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7014 on 52 degrees of freedom  
## Multiple R-squared: 0.2865, Adjusted R-squared: 0.2591   
## F-statistic: 10.44 on 2 and 52 DF, p-value: 0.0001541

## \begin{table}[ht]  
## \centering  
## \begin{tabular}{rrrrr}  
## \hline  
## & Estimate & Std. Error & t value & Pr($>$$|$t$|$) \\   
## \hline  
## (Intercept) & 1.3430 & 0.2611 & 5.14 & 0.0000 \\   
## MCYE & 0.2165 & 0.0863 & 2.51 & 0.0153 \\   
## Herbaceous & 10.6378 & 2.7483 & 3.87 & 0.0003 \\   
## \hline  
## \end{tabular}  
## \end{table}



Using the trained model from k-folds, lets further validate this model and check how well it predicts the 20% of the data.



##   
## Call:  
## lm(formula = testdata$S\_SUM ~ predictionz)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.2080 -0.1015 0.1678 0.3518 1.0321   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.1419 0.8367 0.170 0.8668   
## predictionz 0.9448 0.3669 2.575 0.0173 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7736 on 22 degrees of freedom  
## Multiple R-squared: 0.2316, Adjusted R-squared: 0.1967   
## F-statistic: 6.632 on 1 and 22 DF, p-value: 0.01726

#### SPATTS SUM ~ Herbaceous

## Linear Regression   
##   
## 56 samples  
## 1 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 20 times)   
## Summary of sample sizes: 50, 49, 51, 52, 48, 52, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 0.7277864 0.3400115 0.5759204  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

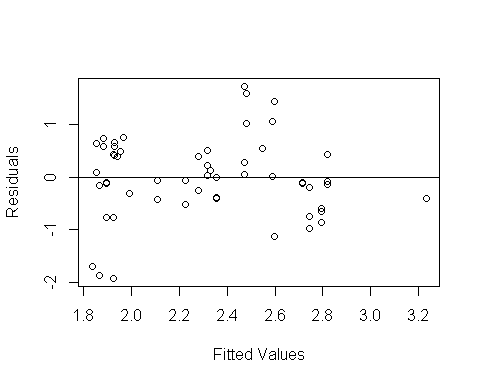
Our model: S\_SUM ~ Herbaceous

Across all 10-fold cross-validation, the average of the root mean square deviation (RMSE) = 0.7277864 = 0.24 with an average of = 0.34 with = 0.25

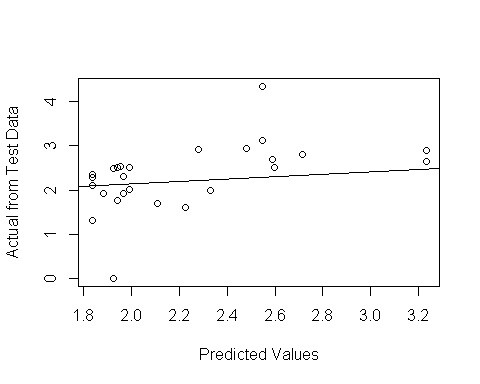
Here is the trained model summary statistic:

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.92416 -0.40531 -0.03112 0.49698 1.72997   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.8368 0.1657 11.084 1.58e-15 \*\*\*  
## Herbaceous 11.0092 3.1197 3.529 0.000861 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7614 on 54 degrees of freedom  
## Multiple R-squared: 0.1874, Adjusted R-squared: 0.1724   
## F-statistic: 12.45 on 1 and 54 DF, p-value: 0.0008613

## \begin{table}[ht]  
## \centering  
## \begin{tabular}{rrrrr}  
## \hline  
## & Estimate & Std. Error & t value & Pr($>$$|$t$|$) \\   
## \hline  
## (Intercept) & 1.8368 & 0.1657 & 11.08 & 0.0000 \\   
## Herbaceous & 11.0092 & 3.1197 & 3.53 & 0.0009 \\   
## \hline  
## \end{tabular}  
## \end{table}



Using the trained model from k-folds, lets further validate this model and check how well it predicts the 20% of the data.



##   
## Call:  
## lm(formula = testdata$S\_SUM ~ predictionz)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.0393 -0.3373 0.0306 0.4062 1.7263   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.2481 0.7449 0.333 0.74202   
## predictionz 0.9309 0.3302 2.820 0.00949 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6772 on 24 degrees of freedom  
## Multiple R-squared: 0.2488, Adjusted R-squared: 0.2175   
## F-statistic: 7.95 on 1 and 24 DF, p-value: 0.009486

### Failed Models

#### SUM ~ Medium\_impervious

## Linear Regression   
##   
## 114 samples  
## 1 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 20 times)   
## Summary of sample sizes: 102, 102, 102, 102, 104, 102, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 0.5521589 0.2039828 0.4555411  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

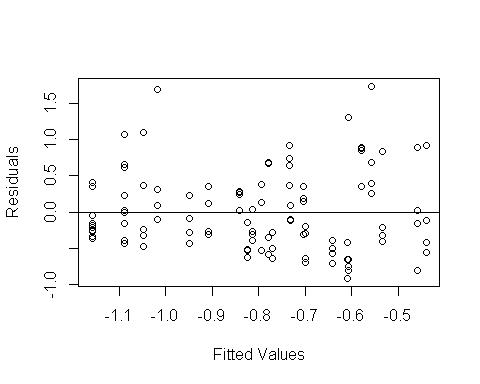
Our model: SUM ~ Medium\_impervious

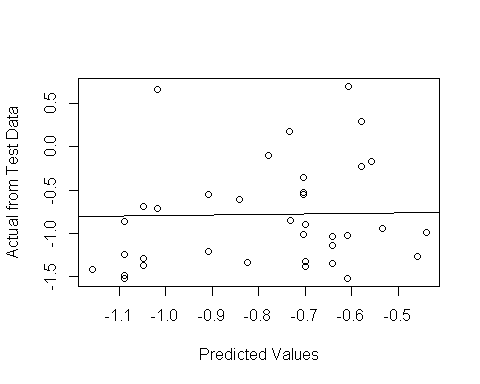
Across all 10-fold cross-validation, the average of the root mean square deviation (RMSE) = 0.5521589 = 0.1 with an average of = 0.2 with = 0.18

Here is the trained model summary statistic:

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.9133 -0.4078 -0.1533 0.3536 1.7309   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.15788 0.09636 -12.016 < 2e-16 \*\*\*  
## Medium\_impervious 0.23271 0.05568 4.179 5.83e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5579 on 112 degrees of freedom  
## Multiple R-squared: 0.1349, Adjusted R-squared: 0.1272   
## F-statistic: 17.46 on 1 and 112 DF, p-value: 5.825e-05

## \begin{table}[ht]  
## \centering  
## \begin{tabular}{rrrrr}  
## \hline  
## & Estimate & Std. Error & t value & Pr($>$$|$t$|$) \\   
## \hline  
## (Intercept) & -1.1579 & 0.0964 & -12.02 & 0.0000 \\   
## Medium\\_impervious & 0.2327 & 0.0557 & 4.18 & 0.0001 \\   
## \hline  
## \end{tabular}  
## \end{table}





##   
## Call:  
## lm(formula = testdata$SUM ~ predictionz)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.8199 -0.4054 -0.2551 0.2972 1.6235   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -0.3290 0.3994 -0.824 0.416  
## predictionz 0.6134 0.4933 1.244 0.222  
##   
## Residual standard error: 0.6018 on 34 degrees of freedom  
## Multiple R-squared: 0.04351, Adjusted R-squared: 0.01537   
## F-statistic: 1.546 on 1 and 34 DF, p-value: 0.2222

#### SUM ~ OP

## Linear Regression   
##   
## 113 samples  
## 1 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 20 times)   
## Summary of sample sizes: 101, 101, 101, 101, 103, 103, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 0.5756312 0.1697646 0.4725588  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

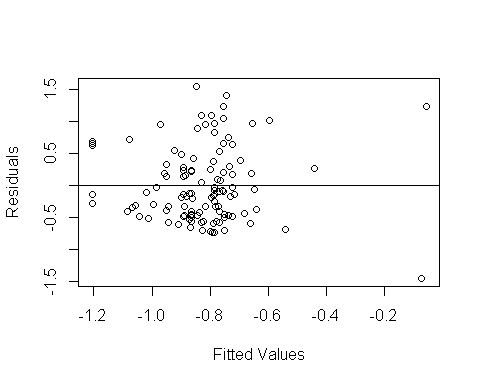
Our model: SUM ~ OP

Across all 10-fold cross-validation, the average of the root mean square deviation (RMSE) = 0.5756312 = 0.1 with an average of = 0.17 with = 0.18

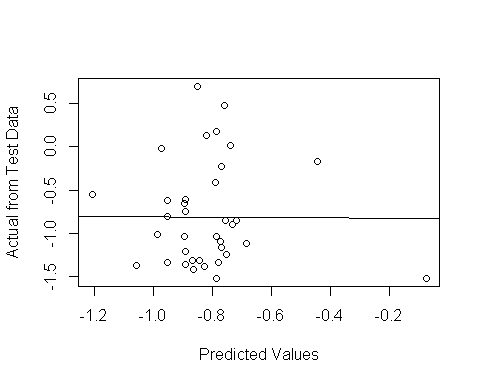
Here is the trained model summary statistic:

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.4492 -0.4553 -0.1345 0.2982 1.5489   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.3158 0.3736 0.845 0.39973   
## OP 0.6034 0.1954 3.088 0.00254 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5745 on 111 degrees of freedom  
## Multiple R-squared: 0.07913, Adjusted R-squared: 0.07083   
## F-statistic: 9.538 on 1 and 111 DF, p-value: 0.002543

## \begin{table}[ht]  
## \centering  
## \begin{tabular}{rrrrr}  
## \hline  
## & Estimate & Std. Error & t value & Pr($>$$|$t$|$) \\   
## \hline  
## (Intercept) & 0.3158 & 0.3736 & 0.85 & 0.3997 \\   
## OP & 0.6034 & 0.1954 & 3.09 & 0.0025 \\   
## \hline  
## \end{tabular}  
## \end{table}



Using the trained model from k-folds, lets further validate this model and check how well it predicts the 20% of the data.



##   
## Call:  
## lm(formula = testdata$SUM ~ predictionz)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.7196 -0.5229 -0.1608 0.2346 1.4951   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.9220 0.4779 -1.929 0.0621 .  
## predictionz -0.1513 0.5729 -0.264 0.7934   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6042 on 34 degrees of freedom  
## Multiple R-squared: 0.002046, Adjusted R-squared: -0.02731   
## F-statistic: 0.0697 on 1 and 34 DF, p-value: 0.7934

#### SUM ~ OP + Turbidity

## Linear Regression   
##   
## 112 samples  
## 2 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 20 times)   
## Summary of sample sizes: 100, 100, 100, 101, 101, 100, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 0.5623782 0.1876965 0.4611664  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

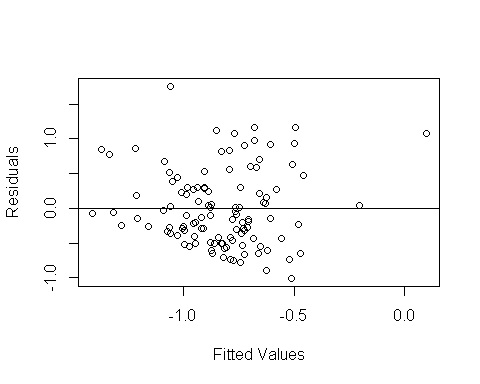
Our model: SUM ~ OP + turb

Across all 10-fold cross-validation, the average of the root mean square deviation (RMSE) = 0.5623782 = 0.1 with an average of = 0.19 with = 0.19

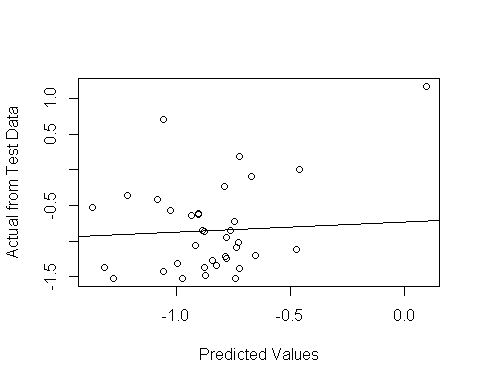
Here is the trained model summary statistic:

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.0099 -0.4251 -0.1192 0.2973 1.7577   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.1786 0.4013 -0.445 0.6572   
## OP 0.4759 0.1950 2.440 0.0163 \*   
## turb 0.2975 0.1023 2.908 0.0044 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5566 on 109 degrees of freedom  
## Multiple R-squared: 0.1473, Adjusted R-squared: 0.1317   
## F-statistic: 9.417 on 2 and 109 DF, p-value: 0.0001689

## \begin{table}[ht]  
## \centering  
## \begin{tabular}{rrrrr}  
## \hline  
## & Estimate & Std. Error & t value & Pr($>$$|$t$|$) \\   
## \hline  
## (Intercept) & -0.1786 & 0.4013 & -0.44 & 0.6572 \\   
## OP & 0.4759 & 0.1950 & 2.44 & 0.0163 \\   
## turb & 0.2975 & 0.1023 & 2.91 & 0.0044 \\   
## \hline  
## \end{tabular}  
## \end{table}



Using the trained model from k-folds, lets further validate this model and check how well it predicts the 20% of the data.



##   
## Call:  
## lm(formula = testdata$SUM ~ predictionz)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.7943 -0.4594 -0.1557 0.4259 1.7060   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.07868 0.34628 -0.227 0.8216   
## predictionz 0.87662 0.38920 2.252 0.0309 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6034 on 34 degrees of freedom  
## Multiple R-squared: 0.1298, Adjusted R-squared: 0.1042   
## F-statistic: 5.073 on 1 and 34 DF, p-value: 0.03087

#### SPATTS SUM ~ MCYE

## Linear Regression   
##   
## 79 samples  
## 1 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 20 times)   
## Summary of sample sizes: 71, 72, 71, 71, 71, 71, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 0.774129 0.2286159 0.5840406  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

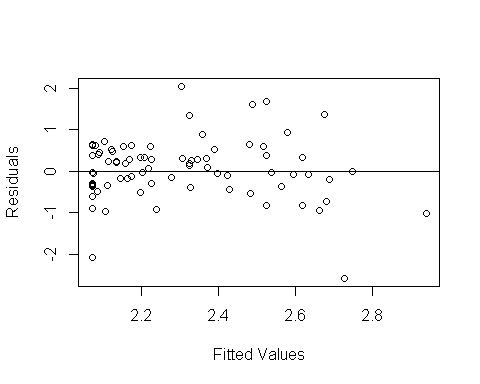
Our model: S\_SUM ~ MCYE

Across all 10-fold cross-validation, the average of the root mean square deviation (RMSE) = 0.774129 = 0.24 with an average of = 0.23 with = 0.2

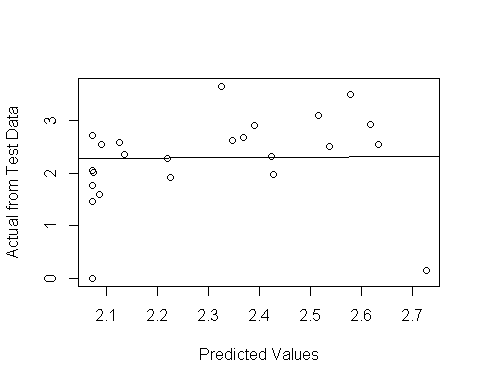
Here is the trained model summary statistic:

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.58324 -0.36148 0.05479 0.43081 2.04074   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.81457 0.22211 8.170 4.79e-12 \*\*\*  
## MCYE 0.19837 0.08267 2.399 0.0188 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8007 on 77 degrees of freedom  
## Multiple R-squared: 0.06957, Adjusted R-squared: 0.05748   
## F-statistic: 5.757 on 1 and 77 DF, p-value: 0.01884

## \begin{table}[ht]  
## \centering  
## \begin{tabular}{rrrrr}  
## \hline  
## & Estimate & Std. Error & t value & Pr($>$$|$t$|$) \\   
## \hline  
## (Intercept) & 1.8146 & 0.2221 & 8.17 & 0.0000 \\   
## MCYE & 0.1984 & 0.0827 & 2.40 & 0.0188 \\   
## \hline  
## \end{tabular}  
## \end{table}



Using the trained model from k-folds, lets further validate this model and check how well it predicts the 20% of the data.



##   
## Call:  
## lm(formula = testdata$S\_SUM ~ predictionz)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.48233 -0.27784 0.06374 0.47566 1.37607   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 0.2703 1.9053 0.142 0.888  
## predictionz 0.8639 0.8245 1.048 0.306  
##   
## Residual standard error: 0.8613 on 22 degrees of freedom  
## Multiple R-squared: 0.04753, Adjusted R-squared: 0.004232   
## F-statistic: 1.098 on 1 and 22 DF, p-value: 0.3061

## Linear Mixed Model

Model1 <- lmer(SUM~  
 MCYE +   
 OP +   
 Medium\_impervious +  
 (1|LK\_CODE),  
 data=LOGTransformed)  
  
Model2 <- lmer(SUM~  
 MCYE +  
 OP +  
 (1|LK\_CODE),  
 data=LOGTransformed)  
  
Model3 <- lmer(SUM~  
 OP +  
 (1|LK\_CODE),  
 data=LOGTransformed)  
  
Model4 <- lmer(SUM~  
 Medium\_impervious +  
 (1|LK\_CODE),  
 data=LOGTransformed,  
 )  
  
Model5 <- lmer(SUM~  
 MCYE +   
 OP +   
 Medium\_impervious +  
 I(doy^2) +   
 (1|LK\_CODE),  
 data=LOGTransformed)  
  
Model6 <- lmer(SUM~  
 OP +   
 Medium\_impervious +  
 I(doy-mean(doy)) +   
 I((doy-mean(doy))^2) +  
 (1|LK\_CODE),  
 data=LOGTransformed)  
  
Model7 <- lmer(SUM~  
 MCYE +  
 I(doy-mean(doy)) +   
 I((doy-mean(doy))^2) +  
 (1|LK\_CODE),  
 data=LOGTransformed)  
Model8 <- lm(SUM~  
 MCYE +  
 OP +   
 Medium\_impervious,  
 data=LOGTransformed)  
  
summary(Model1)

## Linear mixed model fit by REML ['lmerMod']  
## Formula: SUM ~ MCYE + OP + Medium\_impervious + (1 | LK\_CODE)  
## Data: LOGTransformed  
##   
## REML criterion at convergence: 141  
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -1.58369 -0.55760 -0.08809 0.32993 2.90642   
##   
## Random effects:  
## Groups Name Variance Std.Dev.  
## LK\_CODE (Intercept) 0.09763 0.3125   
## Residual 0.18797 0.4336   
## Number of obs: 90, groups: LK\_CODE, 29  
##   
## Fixed effects:  
## Estimate Std. Error t value  
## (Intercept) -0.64561 0.41681 -1.549  
## MCYE 0.15360 0.04648 3.305  
## OP 0.41028 0.19144 2.143  
## Medium\_impervious 0.19325 0.08162 2.368  
##   
## Correlation of Fixed Effects:  
## (Intr) MCYE OP   
## MCYE -0.146   
## OP 0.917 0.088   
## Medim\_mprvs -0.441 -0.155 -0.224

bics <- BIC(Model1,Model2, Model3, Model4, Model5, Model6, Model7, Model8)  
kable(bics)

|  |  |  |
| --- | --- | --- |
|  | df | BIC |
| Model1 | 6 | 167.9841 |
| Model2 | 5 | 165.6203 |
| Model3 | 4 | 194.7763 |
| Model4 | 4 | 195.5273 |
| Model5 | 7 | 191.1848 |
| Model6 | 7 | 227.7431 |
| Model7 | 6 | 199.8022 |
| Model8 | 5 | 159.7109 |

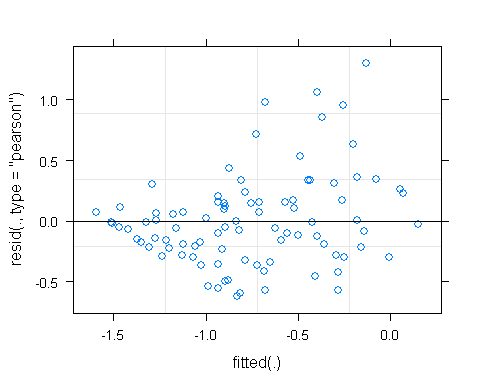
zlme1<-lmer(SUM ~ I(doy-mean(doy)) + I((doy-mean(doy))^2) + Medium\_impervious + turb + (1|LK\_CODE) ,data=LOGTransformed)  
  
zlme2 <- lmer(SUM~MCYE + (1|LK\_CODE), data=LOGTransformed)  
  
  
BIC(zlme1,zlme2)

## df BIC  
## zlme1 7 225.3180  
## zlme2 4 167.8859

Anova(zlme1,test="F")

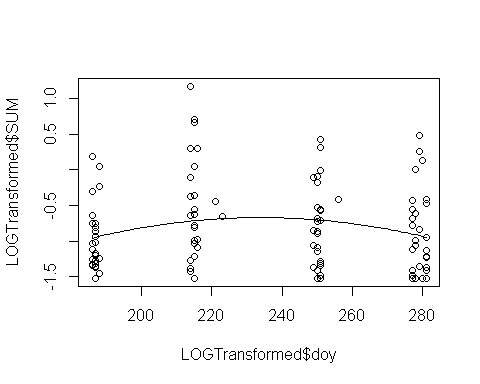
## Analysis of Deviance Table (Type II Wald F tests with Kenward-Roger df)  
##   
## Response: SUM  
## F Df Df.res Pr(>F)   
## I(doy - mean(doy)) 0.2077 1 82.192 0.649783   
## I((doy - mean(doy))^2) 8.2986 1 81.937 0.005062 \*\*  
## Medium\_impervious 9.1279 1 26.147 0.005569 \*\*  
## turb 7.9937 1 99.656 0.005674 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

plot(Model7)



Anova(Model3, test="F")

## Analysis of Deviance Table (Type II Wald F tests with Kenward-Roger df)  
##   
## Response: SUM  
## F Df Df.res Pr(>F)   
## OP 3.393 1 108.98 0.06819 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1



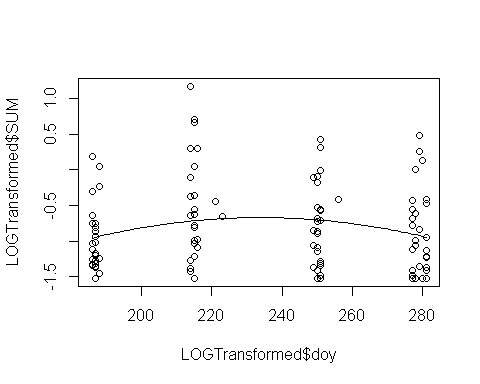
## Analysis of Deviance Table (Type II Wald chisquare tests)  
##   
## Response: SUM  
## Chisq Df Pr(>Chisq)   
## turb 7.1759 1 0.007389 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## [1] 183.132

## [1] 200.0006

## df BIC  
## model3 5 225.6609  
## model2 4 210.9454

## Data: LOGTransformed  
## Models:  
## model2: SUM ~ doy + (1 | LK\_CODE)  
## model3: SUM ~ doy + I(doy^2) + (1 | LK\_CODE)  
## Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)   
## model2 4 185.29 196.24 -88.648 177.29   
## model3 5 178.92 192.60 -84.461 168.92 8.3722 1 0.00381 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1



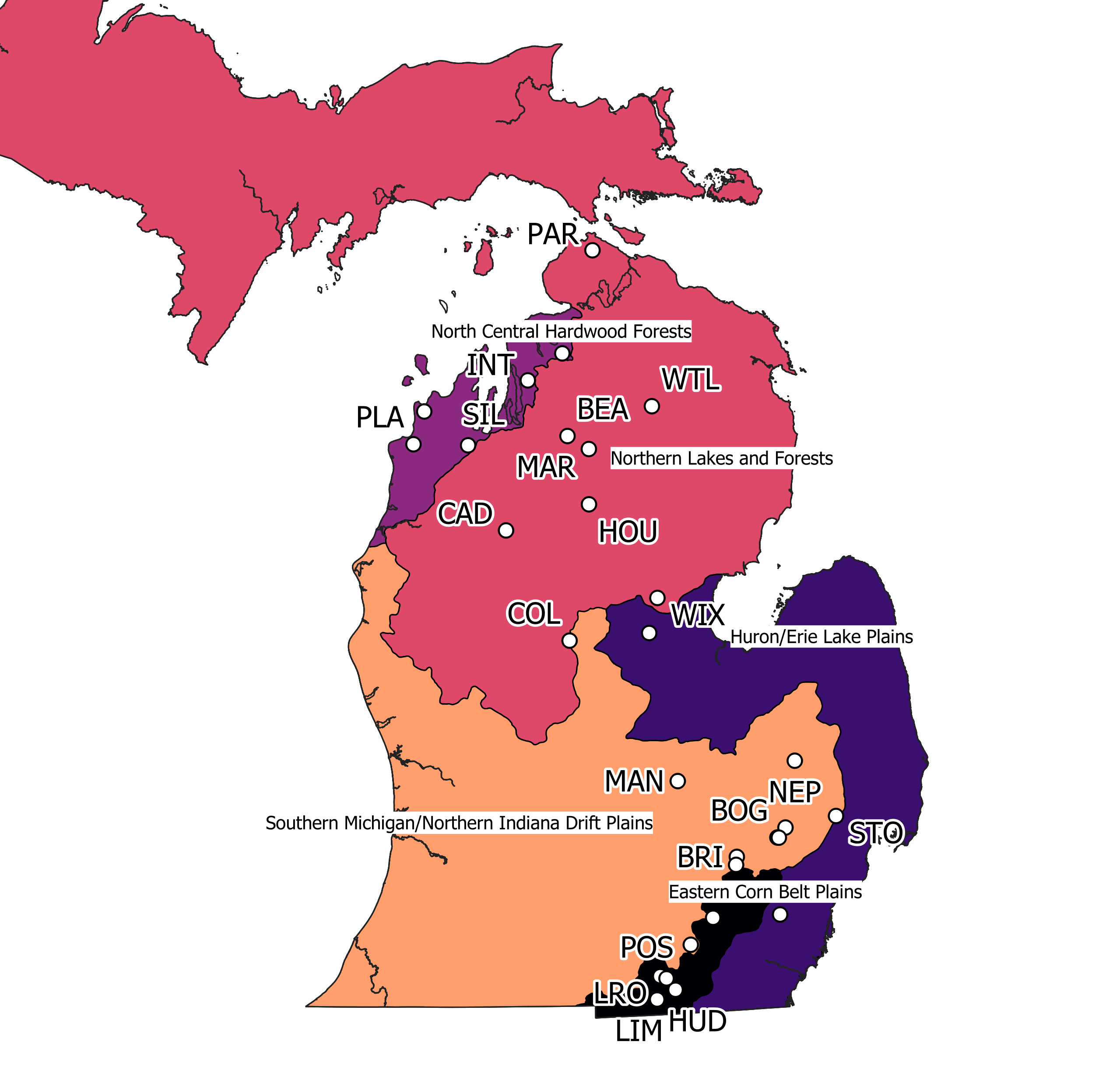
## Analysis of Deviance Table (Type II Wald F tests with Kenward-Roger df)  
##   
## Response: SUM  
## F Df Df.res Pr(>F)   
## I(doy - mean(doy)) 0.2077 1 82.192 0.649783   
## I((doy - mean(doy))^2) 8.2986 1 81.937 0.005062 \*\*  
## Medium\_impervious 9.1279 1 26.147 0.005569 \*\*  
## turb 7.9937 1 99.656 0.005674 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Analysis of Deviance Table (Type II Wald F tests with Kenward-Roger df)  
##   
## Response: SUM  
## F Df Df.res Pr(>F)   
## I(doy - mean(doy)) 0.1189 1 104.844 0.7309715   
## I((doy - mean(doy))^2) 5.3676 1 104.695 0.0224609 \*   
## Medium\_impervious 16.8017 1 33.290 0.0002508 \*\*\*  
## turb 11.6319 1 50.594 0.0012817 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

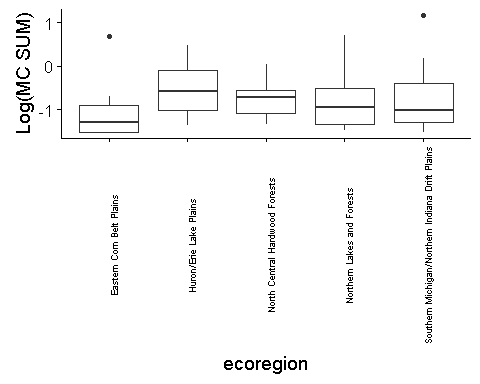
## $ecoregion  
## (Intercept)  
## Eastern Corn Belt Plains -1.301974  
## Huron/Erie Lake Plains -1.301974  
## North Central Hardwood Forests -1.301974  
## Northern Lakes and Forests -1.301974  
## Southern Michigan/Northern Indiana Drift Plains -1.301974  
## I(doy - mean(doy))  
## Eastern Corn Belt Plains 0.0004867134  
## Huron/Erie Lake Plains 0.0004867134  
## North Central Hardwood Forests 0.0004867134  
## Northern Lakes and Forests 0.0004867134  
## Southern Michigan/Northern Indiana Drift Plains 0.0004867134  
## I((doy - mean(doy))^2)  
## Eastern Corn Belt Plains -0.0001243482  
## Huron/Erie Lake Plains -0.0001243482  
## North Central Hardwood Forests -0.0001243482  
## Northern Lakes and Forests -0.0001243482  
## Southern Michigan/Northern Indiana Drift Plains -0.0001243482  
## Medium\_impervious  
## Eastern Corn Belt Plains 0.2298468  
## Huron/Erie Lake Plains 0.2298468  
## North Central Hardwood Forests 0.2298468  
## Northern Lakes and Forests 0.2298468  
## Southern Michigan/Northern Indiana Drift Plains 0.2298468  
## turb  
## Eastern Corn Belt Plains 0.3517558  
## Huron/Erie Lake Plains 0.3517558  
## North Central Hardwood Forests 0.3517558  
## Northern Lakes and Forests 0.3517558  
## Southern Michigan/Northern Indiana Drift Plains 0.3517558  
##   
## attr(,"class")  
## [1] "coef.mer"

## Ecoregions

Our sampled lakes are spread across 5 ecoregions in Michigan. Lets observe if the averages of MC and mcyE is signifigantly different.

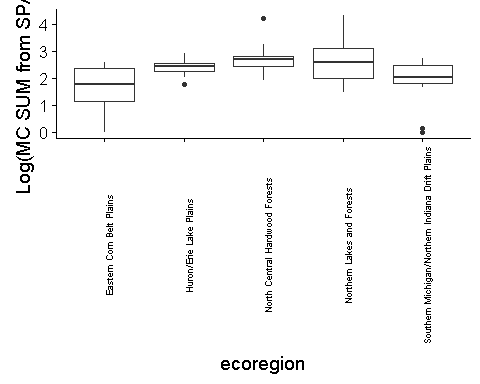


EPA Level III Ecoregions in Michigan



MC Sum Averages in Each Level III Ecoregion

|  |  |
| --- | --- |
| Ecoregions | Mean MC Sum |
| Eastern Corn Belt Plains | 0.332 |
| Huron/Erie Lake Plains | 0.687 |
| North Central Hardwood Forests | 0.238 |
| Northern Lakes and Forests | 0.500 |
| Southern Michigan/Northern Indiana Drift Plains | 0.679 |
|  |  |



Average of MC from SPATTS (ng MC/g of resin)

|  |  |
| --- | --- |
| Ecoregions | Mean MC Sum from SPATTs |
| Eastern Corn Belt Plains | 125.4500 |
| Huron/Erie Lake Plains | 315.2250 |
| North Central Hardwood Forests | 1547.0850 |
| Northern Lakes and Forests | 2515.4378 |
| Southern Michigan/Northern Indiana Drift Plains | 180.6021 |

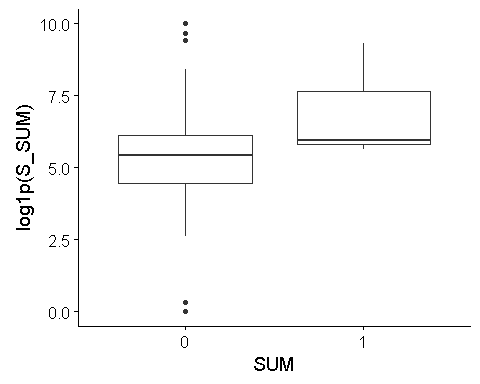
## Present/Absent of Zebra Mussels

|  |  |
| --- | --- |
| Present\_Absent\_Zebra | Mean MC Sum from LC-MS/MS |
| Absent | 0.3045975 |
| Present | 0.5677606 |

Welch Two Sample t-test

data: present and absent t = 1.1441, df = 107.27, p-value = 0.1276 alternative hypothesis: true difference in means is greater than 0 95 percent confidence interval: -0.1184792 Inf sample estimates: mean of x mean of y 0.5677606 0.3045975

# SPATTS vs SUM



##   
## Welch Two Sample t-test  
##   
## data: above and below  
## t = 0.8065, df = 2.0433, p-value = 0.2515  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## -7482.975 Inf  
## sample estimates:  
## mean of x mean of y   
## 3925.435 1012.424

# R Packages

#install.packages("RPostgreSQL")  
#install.packages("tidyverse")  
#install.packages("GGally")  
#install.packages("viridis")  
#install.packages("ggthemes")  
#install.packages("xtable")  
#install.packages("RColorBrewer")  
#install.packages("stargazer")  
#install.packages("lme4")  
#install.packages("leaps")  
#install.packages("knitr")  
#install.packages("rmarkdown")  
#install.packages("cowplot")  
#install.packages("FactoMineR")  
#install.packages("factoextra")  
#install.packages("missMDA")  
#install.packages("texreg")  
#install.packages("caret")  
#install.packages("DAAG")  
#install.packages("regclass")  
  
library(RPostgreSQL)  
library(tidyverse)  
library(GGally)  
library(viridis)  
library(RColorBrewer)  
library(ggthemes)  
library(xtable)  
library(readxl)  
library(openxlsx)  
library(stargazer)  
library(lme4)  
library(leaps)  
library(rmarkdown)  
library(kableExtra)  
library(leaps)  
library(cowplot)  
library(FactoMineR)  
library(factoextra)  
library(lubridate)  
library(lavaan)  
library(semPlot)  
require(missMDA)  
library(texreg)  
library(car)  
library(caret)  
library(DAAG)  
library(reshape2)