Influencing factors of cyanotoxins based on Spatio-Temporal Statistics

Name : Yihang Xing

Studentnumber : B00840889

Professor : Dr. Hong Gu

Data courtesy of : Dr. Barry Husk

Date : 2023/4/10

Content:

1 Background and motivation

We live in a complex world, and clever people are continually coming up with new ways to observe and record increasingly large parts of it so we can comprehend it better . We are squarely in the midst of a “big data” era, and it seems that every day new methodologies and algorithms emerge that are designed to deal with the ever-increasing size of these data streams. Spatio-temporal data are everywhere in science, engineering, business, and industry.In this paper,we describe some basic components of spatio-temporal data structures in R, followed by spatio-temporal visualization and exploratory tools.Then fit simple statistical models to the data to indicate possible patterns and see if assumptions are violated.

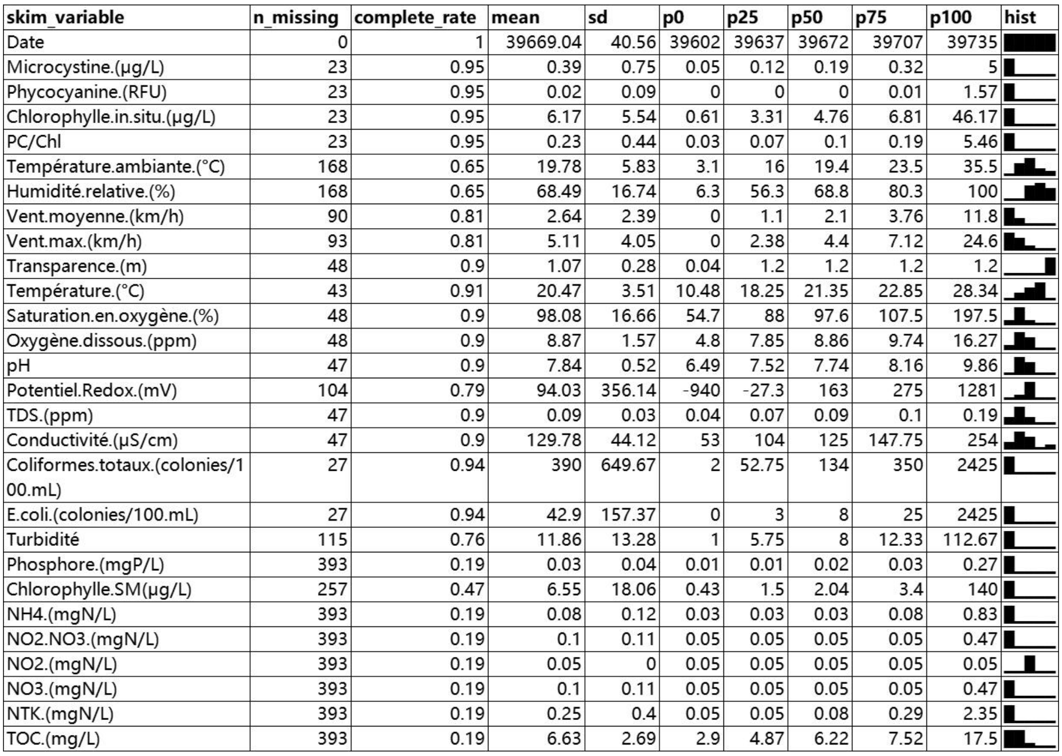
2 Data description and exploration

2.1 Summary

The values of cyanotoxins was monitored by different stations at different time. This paper studies 25 stations at different times .In order to find how the cyanotoxins variables (including Microcystines, Phycocyanines, Chlorophylle,PCChl) are realted to other variables.

The data types of this dataset are divided into numeric variable and categorical variable .There are 485 data samples and 32 variables.And there are 4 factor variables and 28 numeric variables.The descriptive statistics of the data is below.

Table1: Describtion of numeric variable



skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Date	0	1	39669.04	40.56	39602	39637	39672	39707	39735	
Microcystine.(µg/L)	23	0.95	0.39	0.75	0.05	0.12	0.19	0.32	5	
Phycocyanine.(RFU)	23	0.95	0.02	0.09	0	0	0	0.01	1.57	
Chlorophylle.in.situ.(µg/L)	23	0.95	6.17	5.54	0.61	3.31	4.76	6.81	46.17	
PC/Chl	23	0.95	0.23	0.44	0.03	0.07	0.1	0.19	5.46	
Température.ambiante.(°C)	168	0.65	19.78	5.83	3.1	16	19.4	23.5	35.5	
Humidité.relative.(%)	168	0.65	68.49	16.74	6.3	56.3	68.8	80.3	100	
Vent.moyenne.(km/h)	90	0.81	2.64	2.39	0	1.1	2.1	3.76	11.8	
Vent.max.(km/h)	93	0.81	5.11	4.05	0	2.38	4.4	7.12	24.6	
Transparence.(m)	48	0.9	1.07	0.28	0.04	1.2	1.2	1.2	1.2	
Température.(°C)	43	0.91	20.47	3.51	10.48	18.25	21.35	22.85	28.34	
Saturation.en.oxygène.(%)	48	0.9	98.08	16.66	54.7	88	97.6	107.5	197.5	
Oxygène.dissous.(ppm)	48	0.9	8.87	1.57	4.8	7.85	8.86	9.74	16.27	
pH	47	0.9	7.84	0.52	6.49	7.52	7.74	8.16	9.86	
Potentiel.Redox.(mV)	104	0.79	94.03	356.14	-940	-27.3	163	275	1281	
TDS.(ppm)	47	0.9	0.09	0.03	0.04	0.07	0.09	0.1	0.19	
Conductivité.(µS/cm)	47	0.9	129.78	44.12	53	104	125	147.75	254	
Coliformes.totaux.(colonies/100.mL)	27	0.94	390	649.67	2	52.75	134	350	2425	
E.coli.(colonies/100.mL)	27	0.94	42.9	157.37	0	3	8	25	2425	
Turbidité	115	0.76	11.86	13.28	1	5.75	8	12.33	112.67	
Phosphore.(mgP/L)	393	0.19	0.03	0.04	0.01	0.01	0.02	0.03	0.27	
Chlorophylle.SM(µg/L)	257	0.47	6.55	18.06	0.43	1.5	2.04	3.4	140	
NH4.(mgN/L)	393	0.19	0.08	0.12	0.03	0.03	0.03	0.08	0.83	
NO2.NO3.(mgN/L)	393	0.19	0.1	0.11	0.05	0.05	0.05	0.05	0.47	
NO2.(mgN/L)	393	0.19	0.05	0	0.05	0.05	0.05	0.05	0.05	
NO3.(mgN/L)	393	0.19	0.1	0.11	0.05	0.05	0.05	0.05	0.47	
NTK.(mgN/L)	393	0.19	0.25	0.4	0.05	0.05	0.08	0.29	2.35	
TOC.(mg/L)	393	0.19	6.63	2.69	2.9	4.87	6.22	7.52	17.5	

Table2:Describtion of factor variable

skim_variable	n_missing	complete_rate	n_unique	top_counts
Station	0	1	25	ARG: 20, AYL: 20, BPN: 20, BRO: 20
Pluie/Nuage/Soleil	56	0.88	5	Sol: 218, Nua: 133, Plu: 36, sol: 24
Direction.du.vent	213	0.56	2	Ver: 216, Ver: 56
Fleur.d' eau	42	0.91	5	non: 371, Dis: 38, Sou: 24, écu: 9

Tables 1 and 2 show that except Date and Station all variables contain missing data.In this paper, the variables with a large missing proportion are deleted and the following variables are retained:

“Station” ,“Microcystine”,“Phycocyanine”,“Chlorophylle”, “PCChl”,“Ventmoyenne”,“Ventmax”,“Transparence”, “Temp”, “Saturation”,“Oxygene”, “PH”,“TDS”,“ConductivitE”, “Coliformes”, “Ecoli”,

2.2 Outlier handling

From the table above,some variables contain asterisk,interval data and classification data. For these outliers, we deal with them as follows:

1) asterisk converted to NA.

2) ‘<3’converted to 2;‘>2424’ converted to 2425.

3) vent->Mean of each stations

Moyen->Mean of each stations

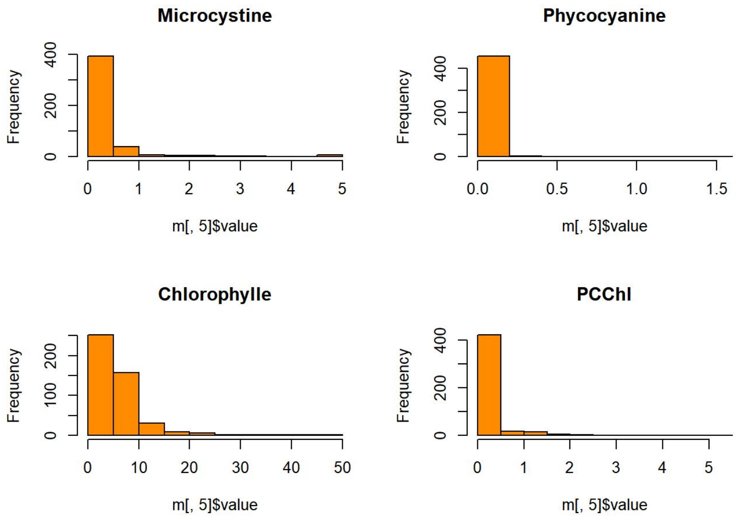
Faible->1st Qu. of each stations

Fort->3rd Qu. of each stations

2.3 Data collation

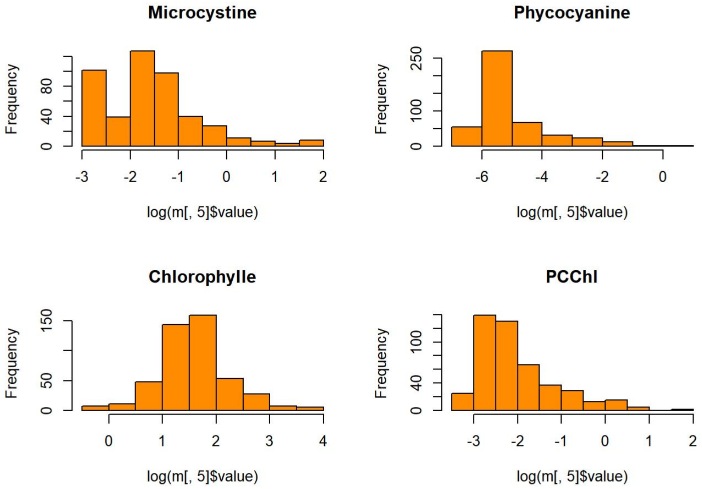
We choose the variables in columns C,D,E,F to analysis.First ,we convert wide data into long data format.Next, let's look at the data distribution through histogram.

2.3.1 Distribution



The histogram of the four variables show skewed distribution, and the data after log transformation shows normal distribution.

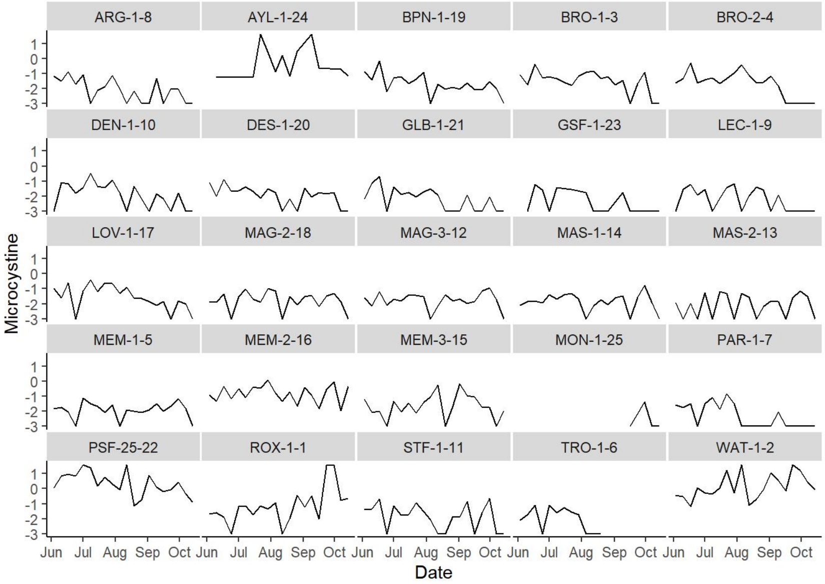
The distribution figure is as follows:



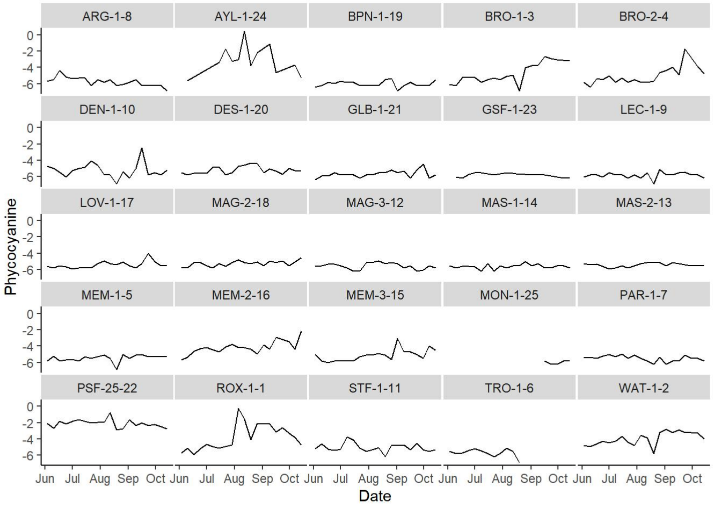
2.4 Time-Series Plots

Next, we look at the time series associated with the columns C,D,E,F in the data set. One can plot the time series at all 25 stations.From the time series, we can see that some stations have less data, such as MON-1-25. And the time series fluctuation of the above four variables is obvious at station AYL-1-24.

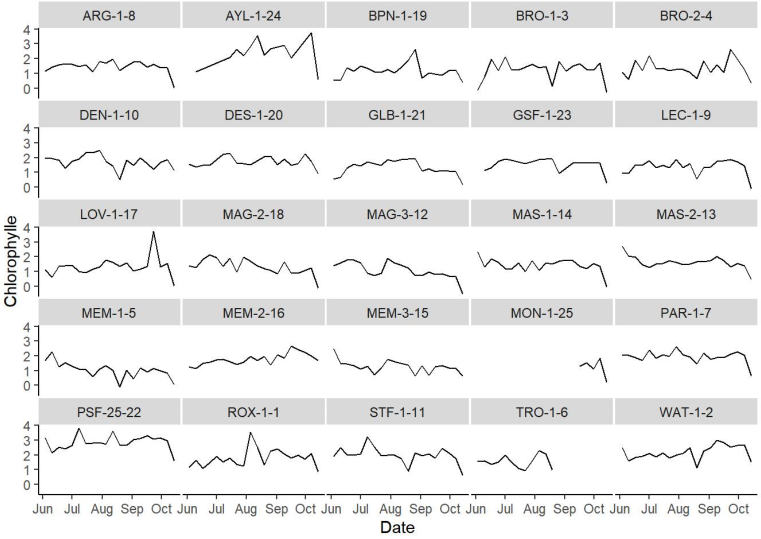
2.4.1 Microcystine



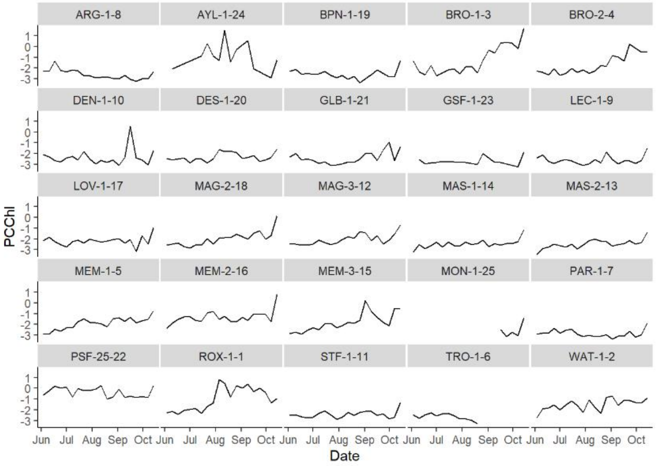
2.4.2 Phycocyanine



2.4.3 Chlorophylle

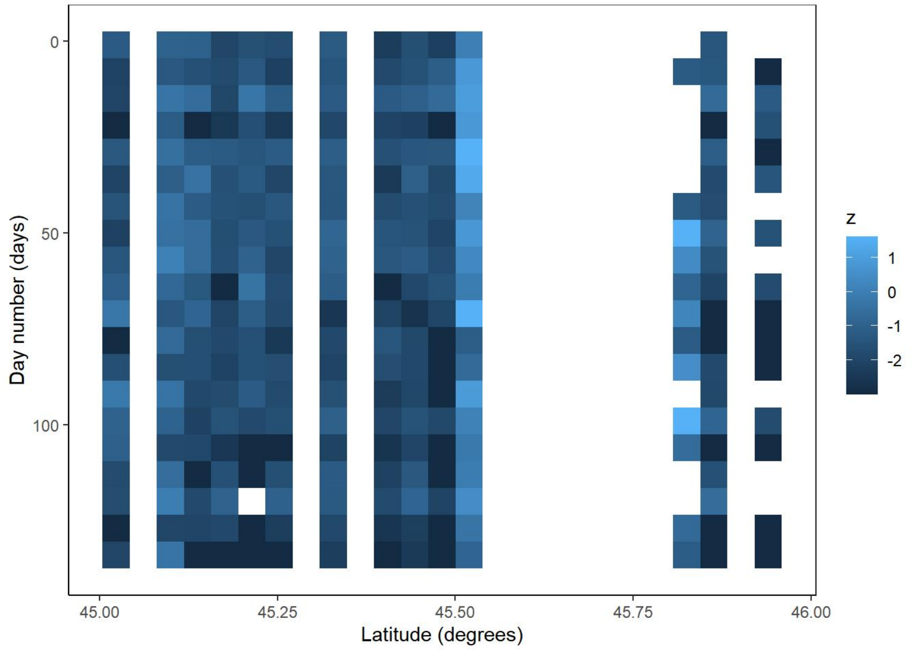


2.4.4 PC/Chl

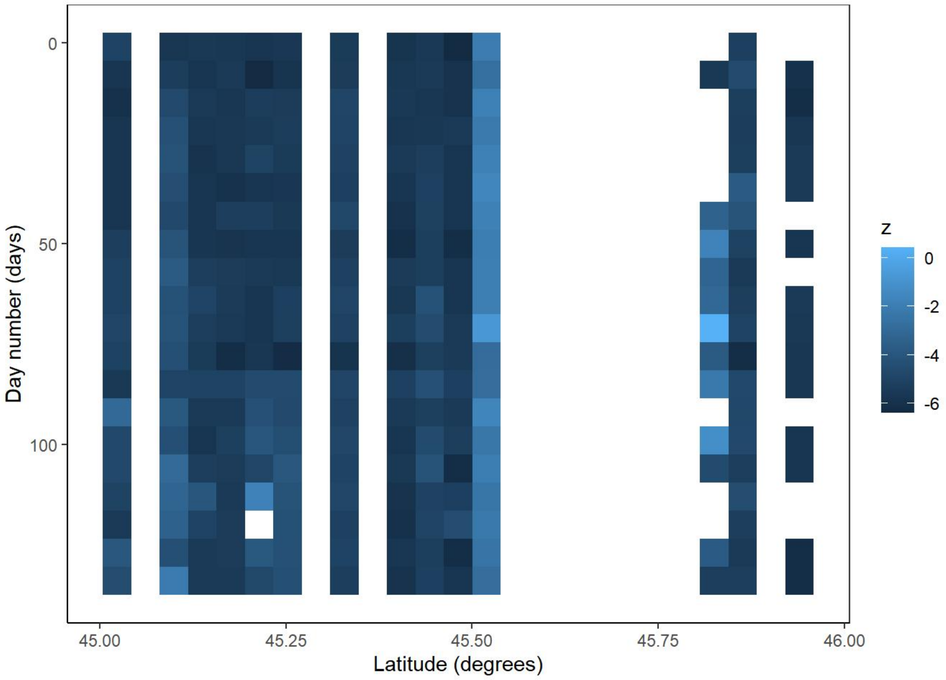


2.5 Hovmoller Plots

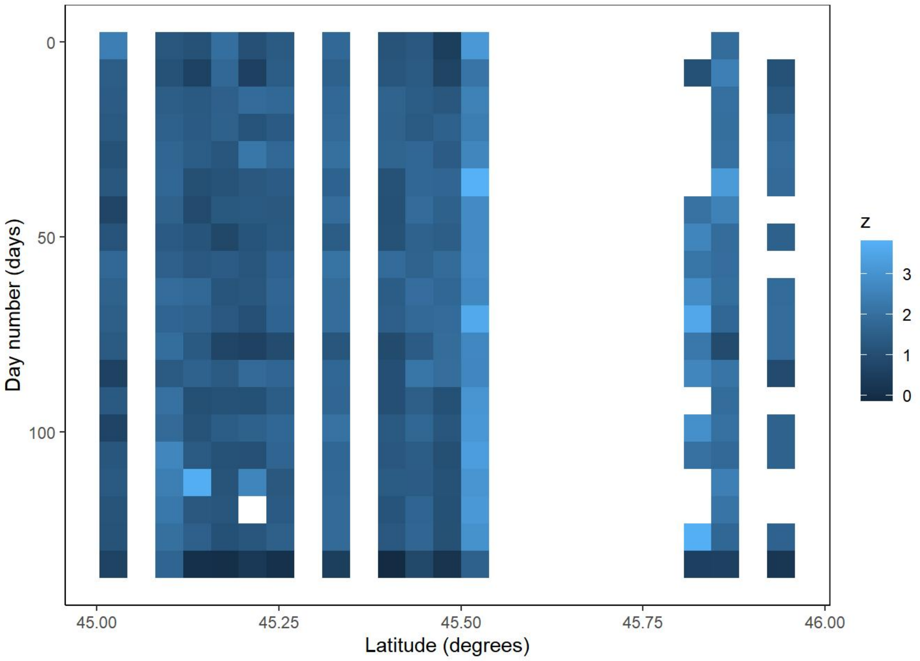
A Hovmöller plot is a two-dimensional space-time visualization, where space is collapsed (projected or averaged) onto one dimension; the second dimension then denotes time.Here,Consider the latitudinal Hovmöller plot. The first step is to generate a regular grid of,say, 25 spatial points and 100 temporal points using the function expand.grid, with limits set to the latitudinal and temporal limits available in the data set. we try to do a Hovmoller Plot for each of four variables.



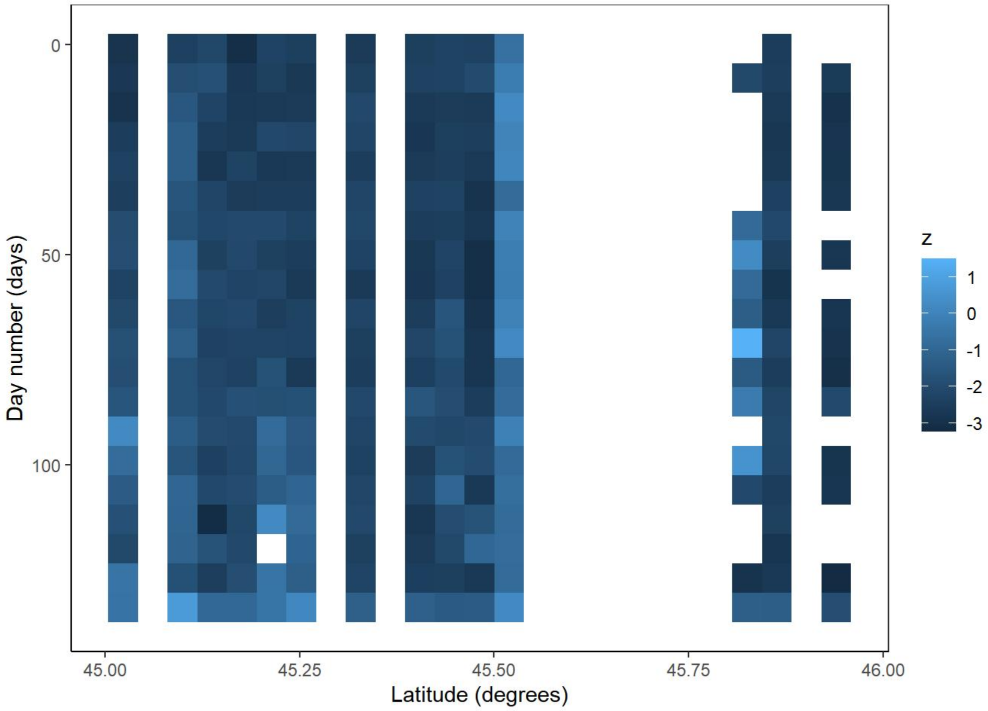
2.5.2 Phycocyanine



2.5.3 Chlorophylle

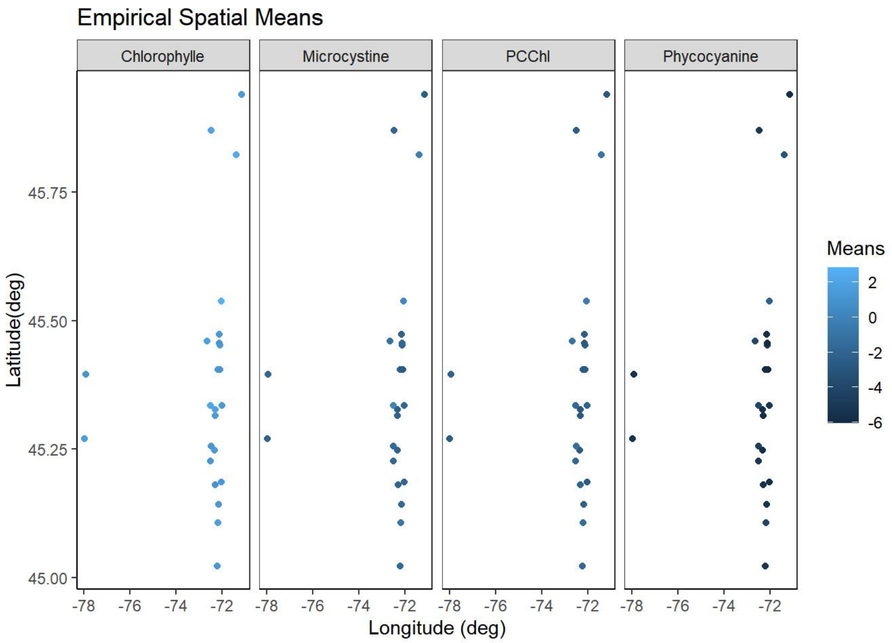


2.5.4 PC/Chl



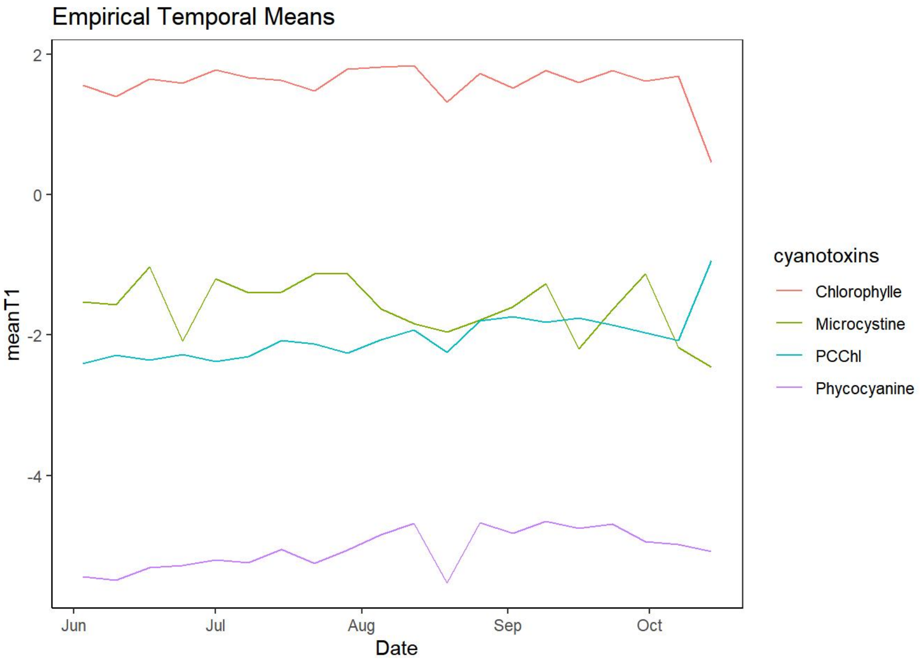
2.6 Empirical Spatial Means

The empirical spatial mean is a spatial quantity that can be stored in a new data frame that contains the spatial locations and the respective average value of each variable at each location.



2.7 Empirical Temporal Means

The empirical temporal mean can be computed easily using the tools of R.first, group the data by time; and second,summarize using the summarise function. From the trend point of view, the fluctuations of the four variables are obviously stationary.



3 Methods for modelling

3.1 Scatter plot

Each station has latitude and longitude. After obtaining latitude and longitude data, then match the two data sets through left_join.The target dataset was named result1.

There are 485 observations and 20 variables in the dataset result1.These variables include:

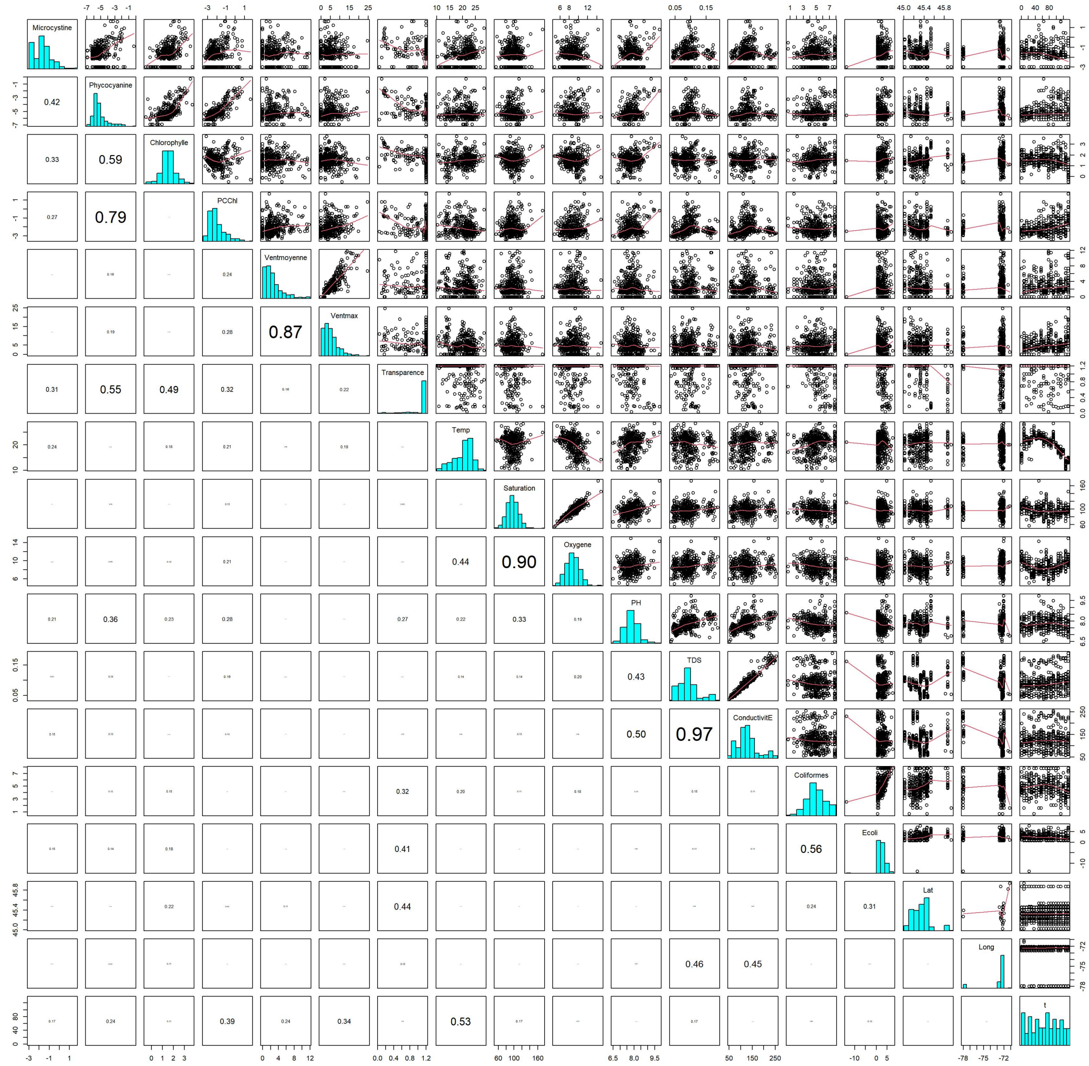
“Date”,“Station” ,“Microcystine”,“Phycocyanine”,“Chlorophylle”, “PCChl”,“Ventmoyenne”,“Ventmax”,“Transparence”, “Temp”, “Saturation”,“Oxygene”, “PH”,“TDS”,“ConductivitE”, “Coliformes”, “Ecoli”,“Lat”,“Long”,“t”.

At the beginning of this paper, the missing situation of each variable is counted.The proportion of missing values in these variables is small, so this paper just omit these observations.

And there are 375 observations and 20 variables in the dataset result2.Next,let’s look at the distribution and correlation coefficient of different variables in result2.

Similarly, do log transformation for variables that do not obey normal distribution .The variables for log transformation are:

"Microcystine","Phycocyanine","Chlorophylle","PCChl","Coliformes","Ecoli"



The correlation coefficient describes the relationship between two variables and the correlation direction. However, the correlation coefficient can not exactly indicate the degree of correlation between the two variables. Its value is between-1 and 1. Correlation coefficient is calculated as follows:

Simple Correlation Coefficient:

$$\mathrm{Cov}(x, y) \; = \; \sum (\mathrm{X_i} - \overline{\mathrm{X}})(\mathrm{Y_i} - \overline{\mathrm{Y}})$$

Pearson Correlation Coefficient:

$$r_{xy} = \frac{Cov(x, y)}{S_x S_y}$$

Partial Correlation Coefficient:

$$r_{xyz} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{(1 - r_{xy}{}^2)}\sqrt{(1 - r_z{}^2)}}$$

Here I choose Pearson Correlation Coefficient.It can be seen from the scatter plot that some variables have high correlation and have collinearity problem. The correlation coefficient between ConductivitE and TDS is 0.97, the correlation coefficient between Ventmax and Ventmoyenne is 0.87, and the correlation coefficient between Oxygene and Nature is 0.9. Therefore, it is necessary to consider the collinearity problem between variables in subsequent modeling.

3.2 Model Introduction

The linear regression results are as follows.In this paper,I use stepwise regression to choose the best regression model. The best model usually has the smallest AIC.In addition, I also use plot function to draw the regression diagnosis diagram.From these four pictures, we can check whether the residuals satisfies Homovariance,normality and independence,and also can check whether there are abnormal observations.

- Residuals vs Fitted: The red line horizontally indicates that there is a good linear relationship.
- QQ plot: We can check whether the residual conforms to the normal distribution.
- Scale-Location plot: Red lines should not have obvious trends.
- Residuals vs Levelage: Large outliers are marked.

For the regression model, this paper is also based on the data set result2.All the variables we used for the regression model are as follows:

"Station" ,"Microcystine","Phycocyanine","Chlorophylle", "PCChl","Ventmoyenne","Ventmax","Transparence", "Temp", "Saturation","Oxygene", "PH","TDS","ConductivitE", "Coliformes", "Ecoli","Lat","Long","t".

3.3 Microcystine

1)model-1

Here, run a linear regression for Microcystine on the predictors,call it model1a.

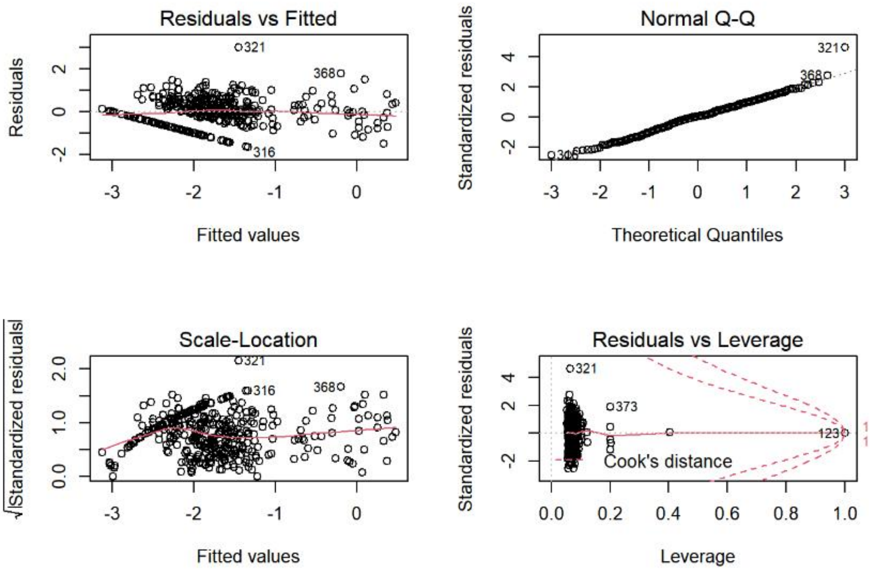
response:Microcystine

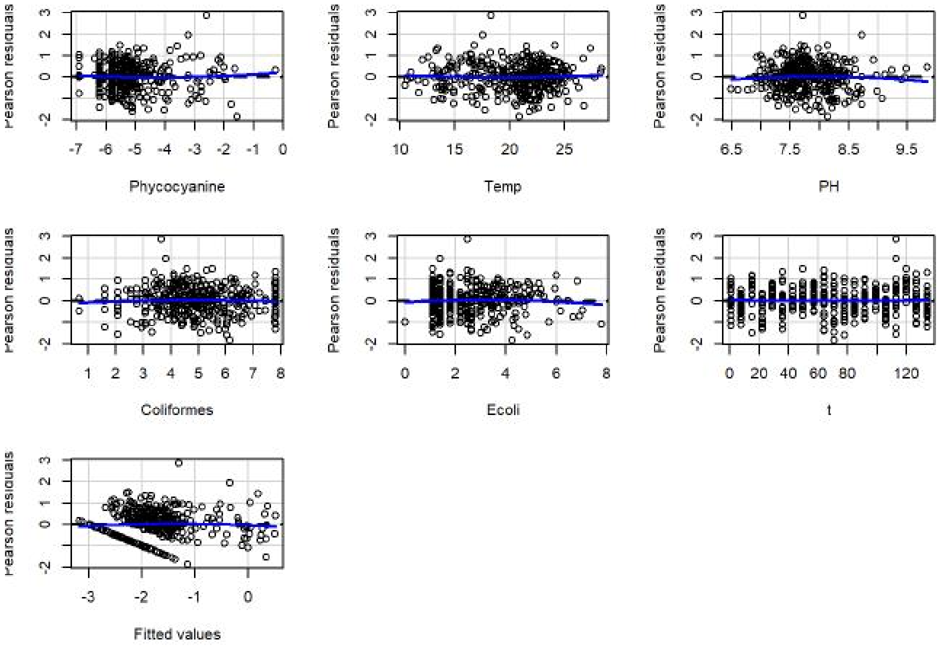
predictors:"Station","Ventmoyenne", "Ventmax","Transparence", "Temp", "Saturation", "Oxygene", "PH","TDS","ConductivitE", "Coliformes","Ecoli","Lat","Long","t"

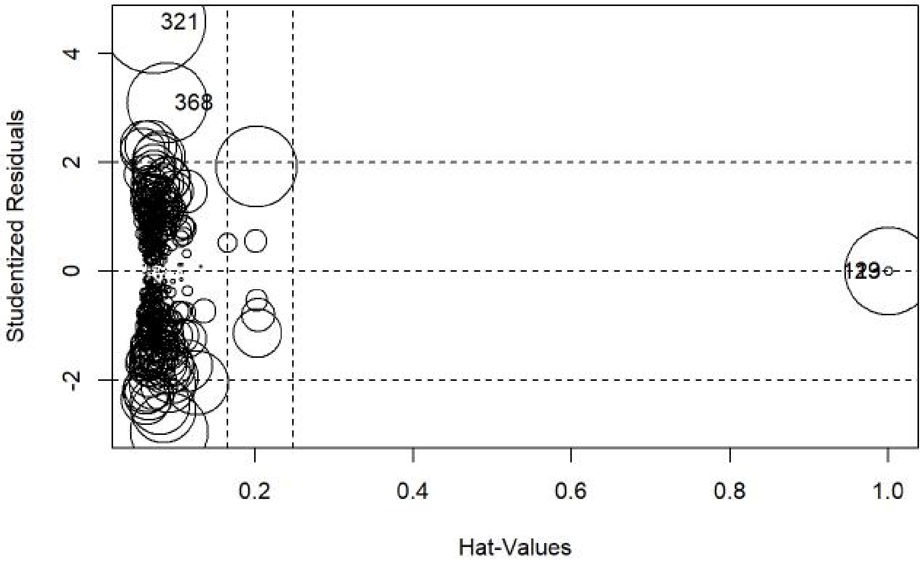
And we use stepwise regression to choose predictors.

the best stepwise regression: lm(formula = Microcystine ~ Station + Phycocyanine + Temp + PH + Coliformes + Ecoli + t, data = resulta)

From the report of model1a,the Adjusted R-squared is 0.4892.

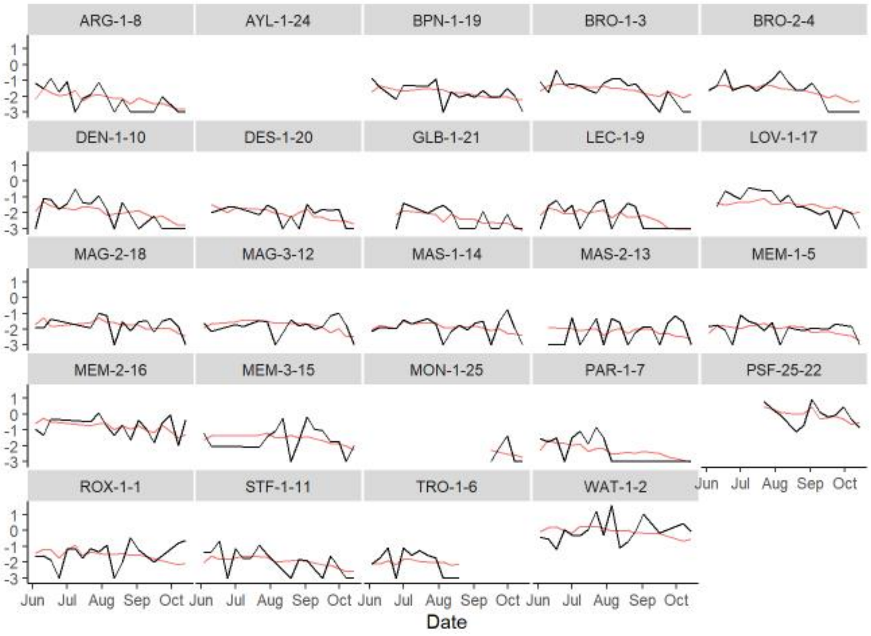


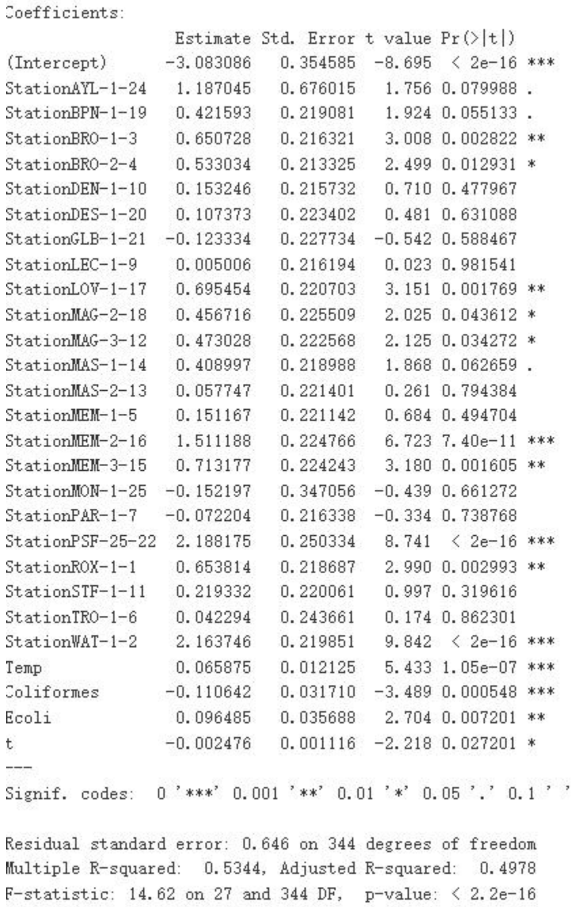




The influencePlot shows that there are some outliers in the model. We remove these outliers and re-run the model call it model1b.The numeric indexes of outliers are 123,321 and 368.

The Adjusted R-squared of model1b is 0.4978.Regression results showed that Temp,t,Ecoli ,Coliformes and part of stations was significant at 0.05 confidence level. Regression coefficients showed that Temp,Coliforms, Ecoli and t had positive effects on Microcystine.





```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -3.083086   0.354585  -8.695  < 2e-16 ***
StationAYL-1-24   1.187045   0.676015   1.756 0.079988 .  
StationBPN-1-19   0.421593   0.219081   1.924 0.055133 .  
StationBRO-1-3    0.650728   0.216321   3.008 0.002822 ** 
StationBRO-2-4    0.533034   0.213325   2.499 0.012931 *  
StationDEN-1-10   0.153246   0.215732   0.710 0.477967    
StationDES-1-20   0.107373   0.223402   0.481 0.631088    
StationGLB-1-21  -0.123334   0.227734  -0.542 0.588467    
StationLEC-1-9    0.005006   0.216194   0.023 0.981541    
StationLOV-1-17   0.695454   0.220703   3.151 0.001769 ** 
StationMAG-2-18   0.456716   0.225509   2.025 0.043612 *  
StationMAG-3-12   0.473028   0.222568   2.125 0.034272 *  
StationMAS-1-14   0.408997   0.218988   1.868 0.062659 .  
StationMAS-2-13   0.057747   0.221401   0.261 0.794384    
StationMEM-1-5    0.151167   0.221142   0.684 0.494704    
StationMEM-2-16   1.511188   0.224766   6.723 7.40e-11 ***
StationMEM-3-15   0.713177   0.224243   3.180 0.001605 ** 
StationMON-1-25  -0.152197   0.347056  -0.439 0.661272    
StationPAR-1-7   -0.072204   0.216338  -0.334 0.738768    
StationPSF-25-22  2.188175   0.250334   8.741  < 2e-16 ***
StationROX-1-1    0.653814   0.218687   2.990 0.002993 ** 
StationSTF-1-11   0.219332   0.220061   0.997 0.319616    
StationTRO-1-6    0.042294   0.243661   0.174 0.862301    
StationWAT-1-2    2.163746   0.219851   9.842  < 2e-16 ***
Temp              0.065875   0.012125   5.433 1.05e-07 ***
Coliformes       -0.110642   0.031710  -3.489 0.000548 ***
Ecoli             0.096485   0.035688   2.704 0.007201 ** 
t                -0.002476   0.001116  -2.218 0.027201 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 

Residual standard error: 0.646 on 344 degrees of freedom
Multiple R-squared:  0.5344, Adjusted R-squared:  0.4978 
F-statistic: 14.62 on 27 and 344 DF,  p-value: < 2.2e-16
```

2)model-2

Here, run a linear regression for Microcystine on the predictors,call it model1c.

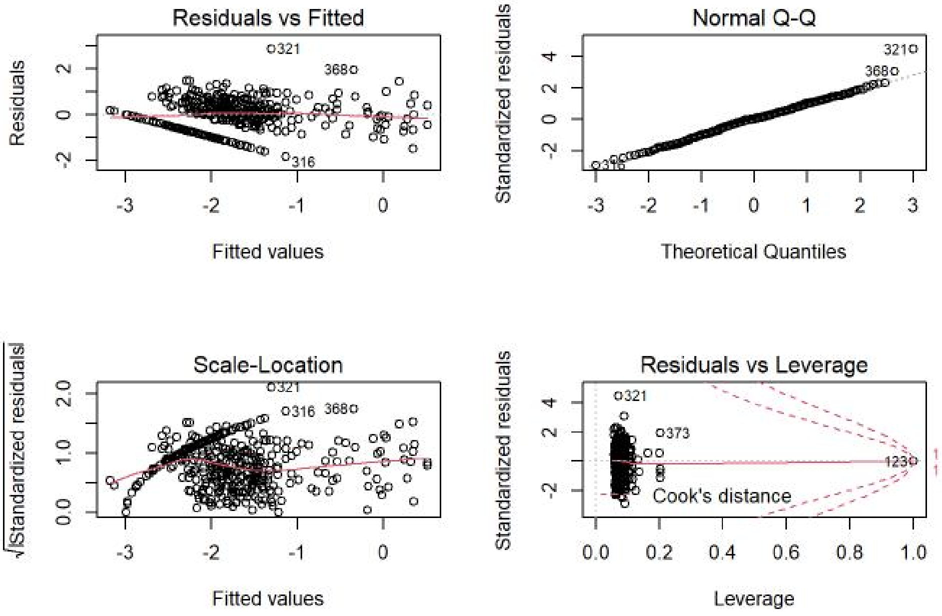
response:Microcystine

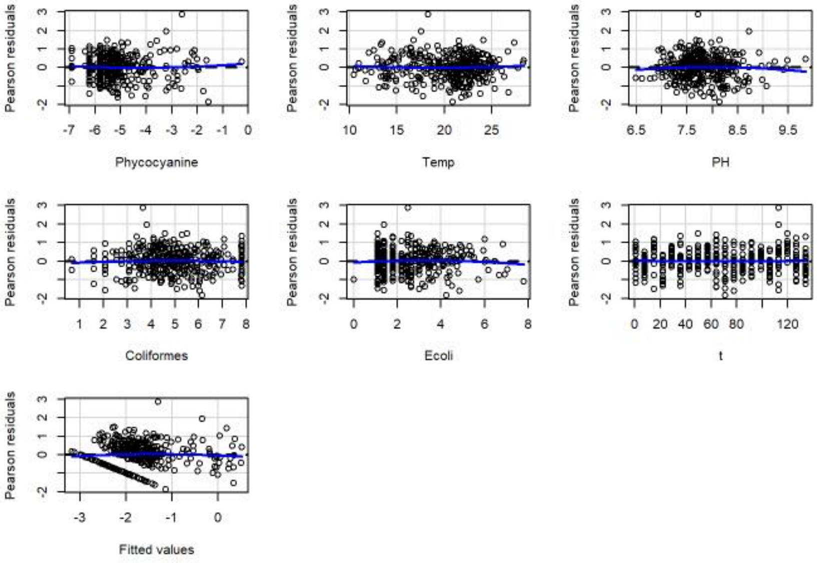
predictors:"Station", "Phycocyanine","Chlorophylle" "PCChl","Ventmoyenne", "Ventmax","Transparence", "Temp", "Saturation", "Oxygene", "PH","TDS","ConductivitE", "Coliformes","Ecoli","Lat","Long","t"

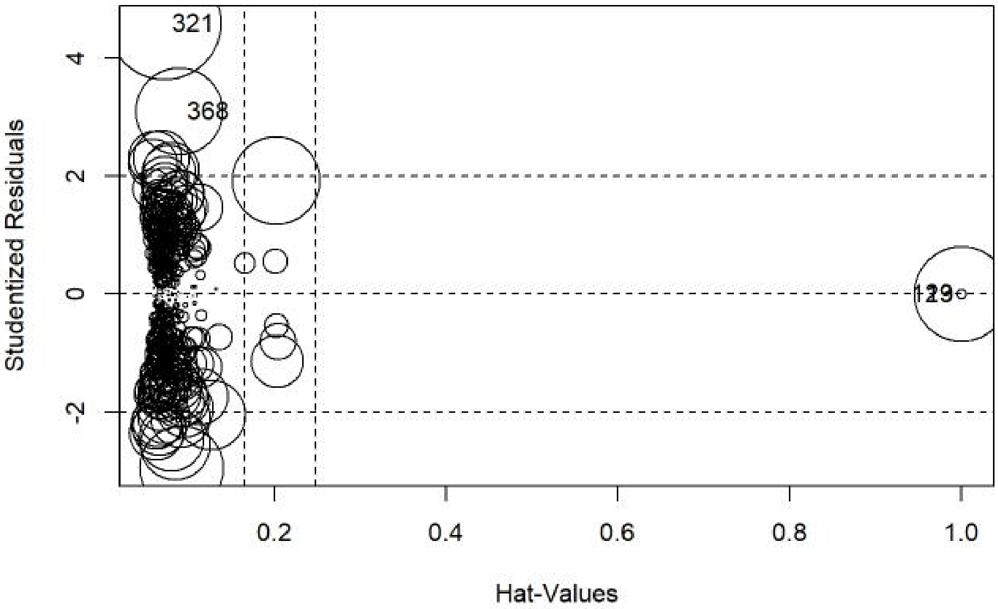
And we use stepwise regression to choose predictors.

the best stepwise regression: lm(formula = Microcystine ~ Station + Phycocyanine + Temp + PH + Coliformes + Ecoli + t, data = resulta)

From the report of model1c,the Adjusted R-squared is 0.4959.

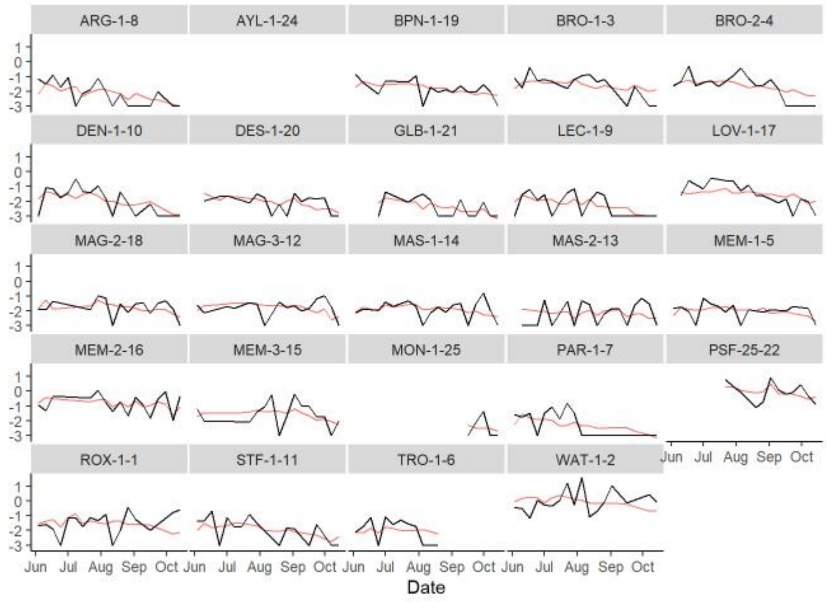

Residuals vs Fitted
Residuals
Fitted values
321
368
316
Normal Q-Q
Standardized residuals
Theoretical Quantiles
321
368
316
Scale-Location
√|Standardized residuals|
Fitted values
321
316
368
Residuals vs Leverage
Standardized residuals
Leverage
321
373
123
Cook's distance
1
1

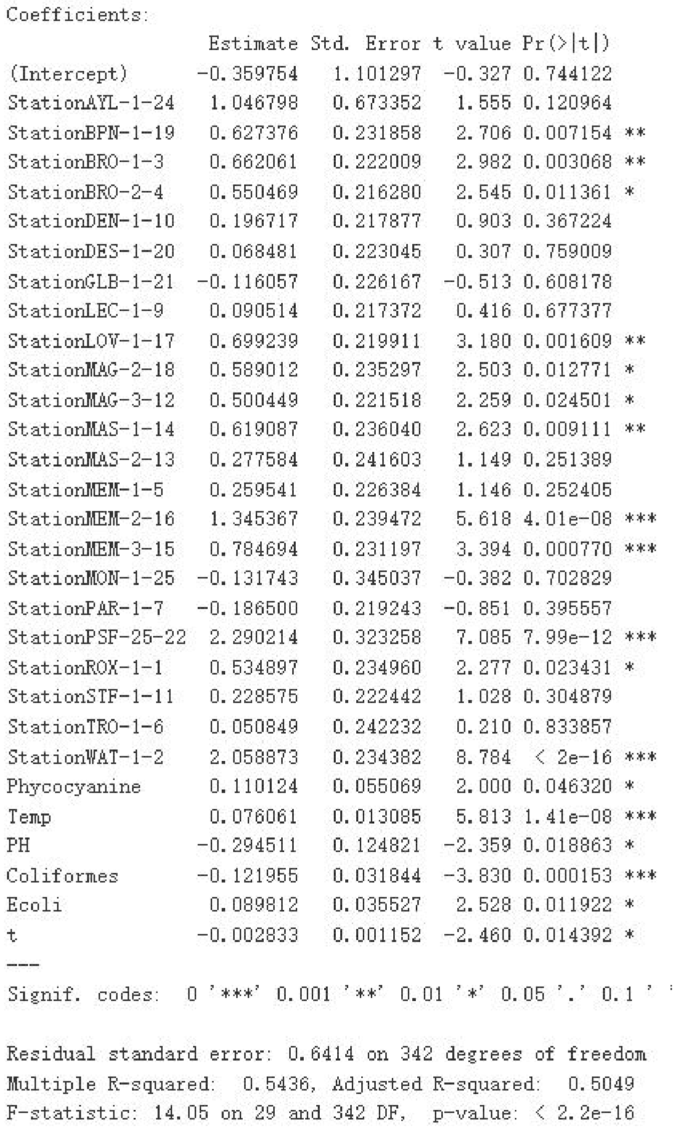

Pearson residuals
Phycocyanine
Temp
PH
Coliformes
Ecoli
t
Fitted values



The influencePlot shows that there are some outliers in the model. We remove these outliers and re-run the model call it model1d.The numeric indexes of outliers are 123,321 and 368.

The Adjusted R-squared of model1d is 0.5049.Regression results showed that PH,Temp,t,Ecoli ,Coliformes and part of stations was significant at 0.05 confidence level. Regression coefficients showed that Temp,Ecoli had positive effects on Microcystine.





```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -0.359754   1.101297  -0.327 0.744122
StationAYL-1-24   1.046798   0.673352   1.555 0.120964
StationBPN-1-19   0.627376   0.231858   2.706 0.007154 **
StationBRO-1-3    0.662061   0.222009   2.982 0.003068 **
StationBRO-2-4    0.550469   0.216280   2.545 0.011361 *
StationDEN-1-10   0.196717   0.217877   0.903 0.367224
StationDES-1-20   0.068481   0.223045   0.307 0.759009
StationGLB-1-21  -0.116057   0.226167  -0.513 0.608178
StationLEC-1-9    0.090514   0.217372   0.416 0.677377
StationLOV-1-17   0.699239   0.219911   3.180 0.001609 **
StationMAG-2-18   0.589012   0.235297   2.503 0.012771 *
StationMAG-3-12   0.500449   0.221518   2.259 0.024501 *
StationMAS-1-14   0.619087   0.236040   2.623 0.009111 **
StationMAS-2-13   0.277584   0.241603   1.149 0.251389
StationMEM-1-5    0.259541   0.226384   1.146 0.252405
StationMEM-2-16   1.345367   0.239472   5.618 4.01e-08 ***
StationMEM-3-15   0.784694   0.231197   3.394 0.000770 ***
StationMON-1-25  -0.131743   0.345037  -0.382 0.702829
StationPAR-1-7   -0.186500   0.219243  -0.851 0.395557
StationPSF-25-22  2.290214   0.323258   7.085 7.99e-12 ***
StationROX-1-1    0.534897   0.234960   2.277 0.023431 *
StationSTF-1-11   0.228575   0.222442   1.028 0.304879
StationTRO-1-6    0.050849   0.242232   0.210 0.833857
StationWAT-1-2    2.058873   0.234382   8.784  < 2e-16 ***
Phycocyanine      0.110124   0.055069   2.000 0.046320 *
Temp              0.076061   0.013085   5.813 1.41e-08 ***
PH               -0.294511   0.124821  -2.359 0.018863 *
Coliformes       -0.121955   0.031844  -3.830 0.000153 ***
Ecoli             0.089812   0.035527   2.528 0.011922 *
t                -0.002833   0.001152  -2.460 0.014392 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 0.6414 on 342 degrees of freedom
Multiple R-squared:  0.5436, Adjusted R-squared:  0.5049
F-statistic: 14.05 on 29 and 342 DF,  p-value: < 2.2e-16
```

3.4 Phycocyanine

1) model-1

Next, run a linear regression for Phycocyanine ,call it model2a.

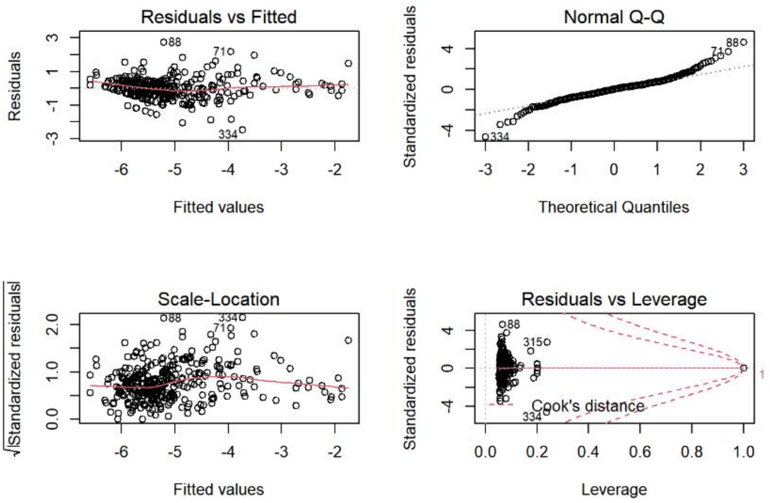
Response:Phycocyanine

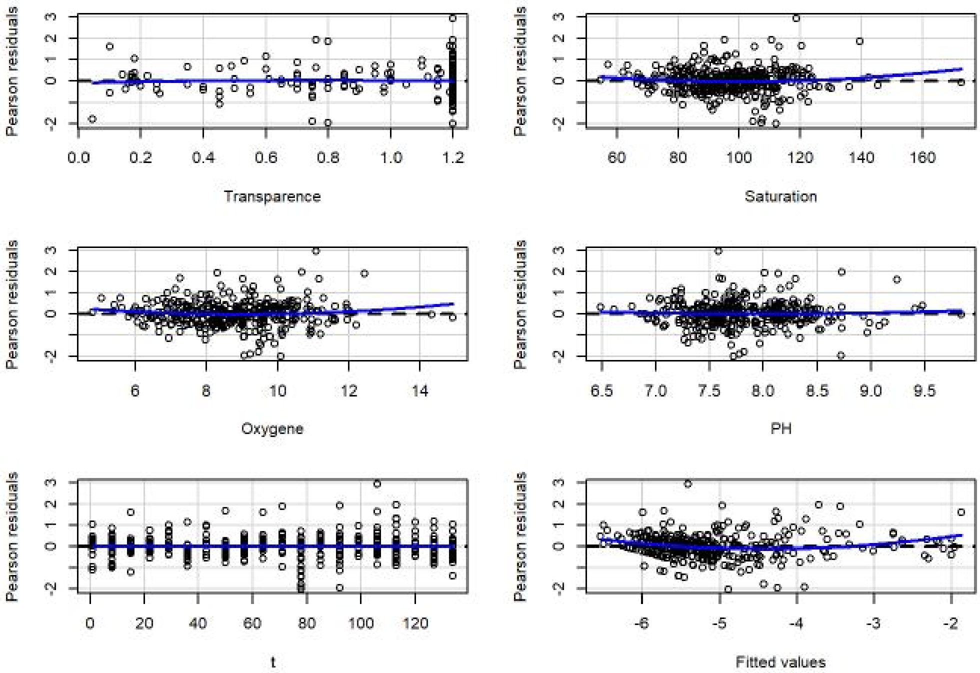
Predictors:

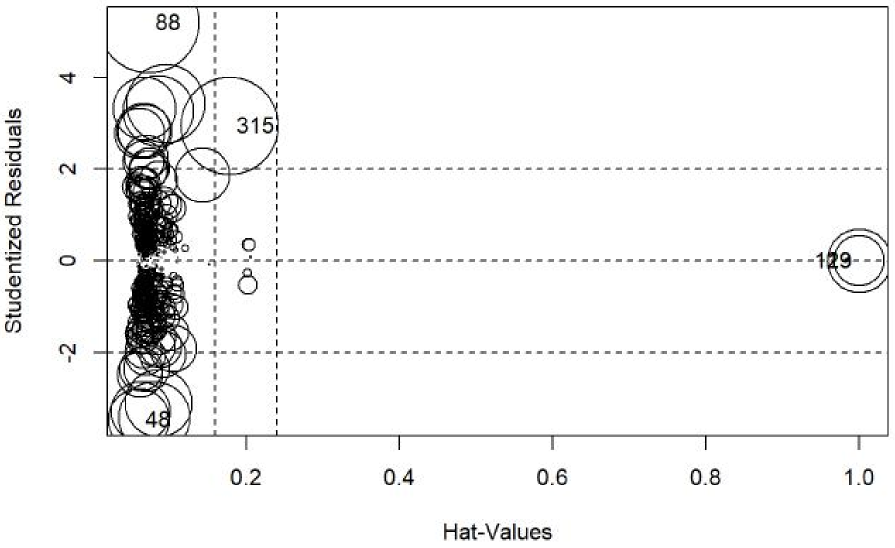
"Station" ,"Ventmoyenne","Ventmax","Transparence", "Temp", "Saturation","Oxygene", "PH","TDS","ConductivitE", "Coliformes", "Ecoli","Lat","Long","t".

From the report,the Adjusted R-squared of model2a is 0.6532.

The best stepwise regression:lm(formula = Phycocyanine ~ Station + Transparence + Saturation + Oxygene + PH + t, data = resultb)

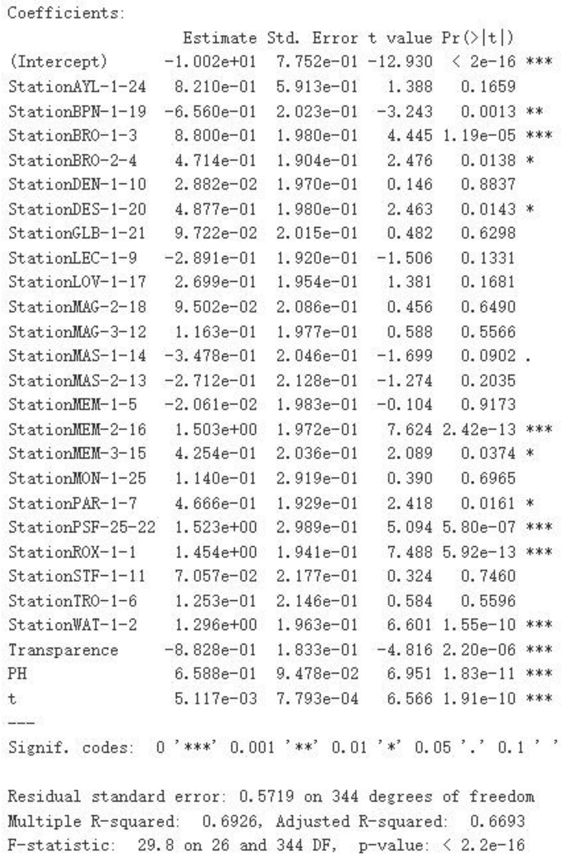






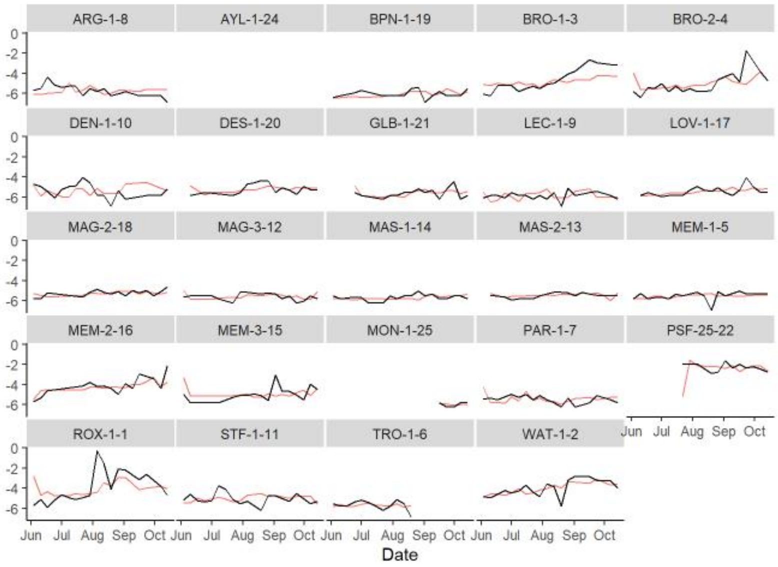
The influencePlot shows that there are some outliers in the model. We remove these outliers and re-run the model call it model2b.The numeric indexes of outliers are 48,88,123 and 315.

From the report ,the Adjusted R-squared is 0.6693.Regression results showed that Transparencet,t,PH,and part of stations were significant at 0.05 confidence level.



```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.002e+01  7.752e-01 -12.930  < 2e-16 ***
StationAYL-1-24   8.210e-01  5.913e-01   1.388   0.1659
StationBPN-1-19  -6.560e-01  2.023e-01  -3.243   0.0013 **
StationBRO-1-3    8.800e-01  1.980e-01   4.445 1.19e-05 ***
StationBRO-2-4    4.714e-01  1.904e-01   2.476   0.0138 *
StationDEN-1-10   2.882e-02  1.970e-01   0.146   0.8837
StationDES-1-20   4.877e-01  1.980e-01   2.463   0.0143 *
StationGLB-1-21   9.722e-02  2.015e-01   0.482   0.6298
StationLEC-1-9   -2.891e-01  1.920e-01  -1.506   0.1331
StationLOV-1-17   2.699e-01  1.954e-01   1.381   0.1681
StationMAG-2-18   9.502e-02  2.086e-01   0.456   0.6490
StationMAG-3-12   1.163e-01  1.977e-01   0.588   0.5566
StationMAS-1-14  -3.478e-01  2.046e-01  -1.699   0.0902 .
StationMAS-2-13  -2.712e-01  2.128e-01  -1.274   0.2035
StationMEM-1-5   -2.061e-02  1.983e-01  -0.104   0.9173
StationMEM-2-16   1.503e+00  1.972e-01   7.624 2.42e-13 ***
StationMEM-3-15   4.254e-01  2.036e-01   2.089   0.0374 *
StationMON-1-25   1.140e-01  2.919e-01   0.390   0.6965
StationPAR-1-7    4.666e-01  1.929e-01   2.418   0.0161 *
StationPSF-25-22  1.523e+00  2.989e-01   5.094 5.80e-07 ***
StationROX-1-1    1.454e+00  1.941e-01   7.488 5.92e-13 ***
StationSTF-1-11   7.057e-02  2.177e-01   0.324   0.7460
StationTRO-1-6    1.253e-01  2.146e-01   0.584   0.5596
StationWAT-1-2    1.296e+00  1.963e-01   6.601 1.55e-10 ***
Transparence     -8.828e-01  1.833e-01  -4.816 2.20e-06 ***
PH                6.588e-01  9.478e-02   6.951 1.83e-11 ***
t                 5.117e-03  7.793e-04   6.566 1.91e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 0.5719 on 344 degrees of freedom
Multiple R-squared:  0.6926, Adjusted R-squared:  0.6693
F-statistic:  29.8 on 26 and 344 DF,  p-value: < 2.2e-16
```



2) model-2

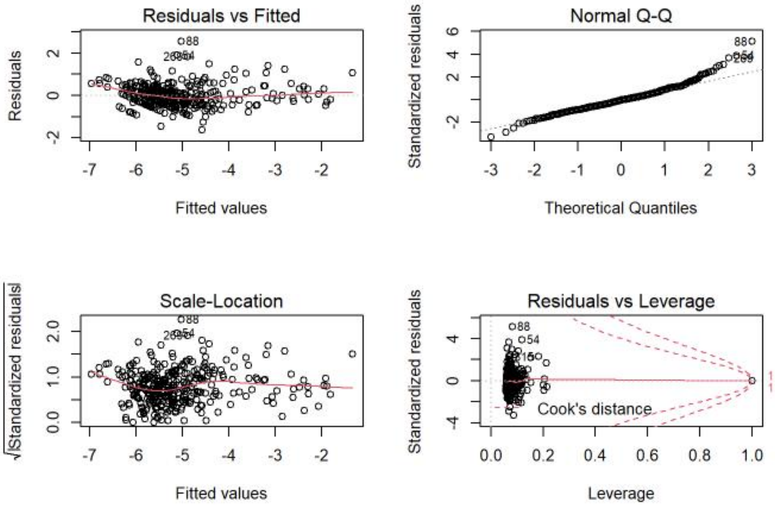
Next, run a linear regression for Phycocyanine ,call it model2c.

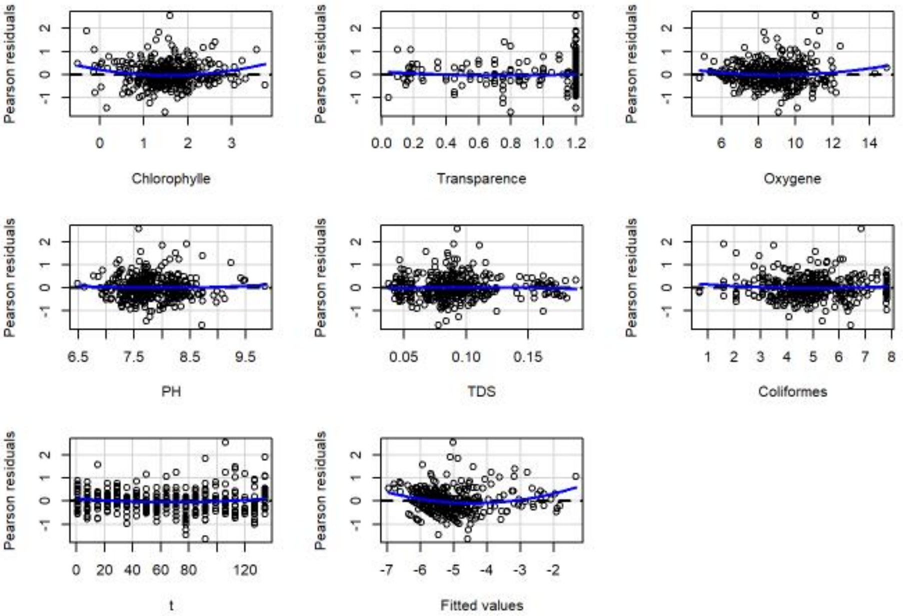
Response:Phycocyanine

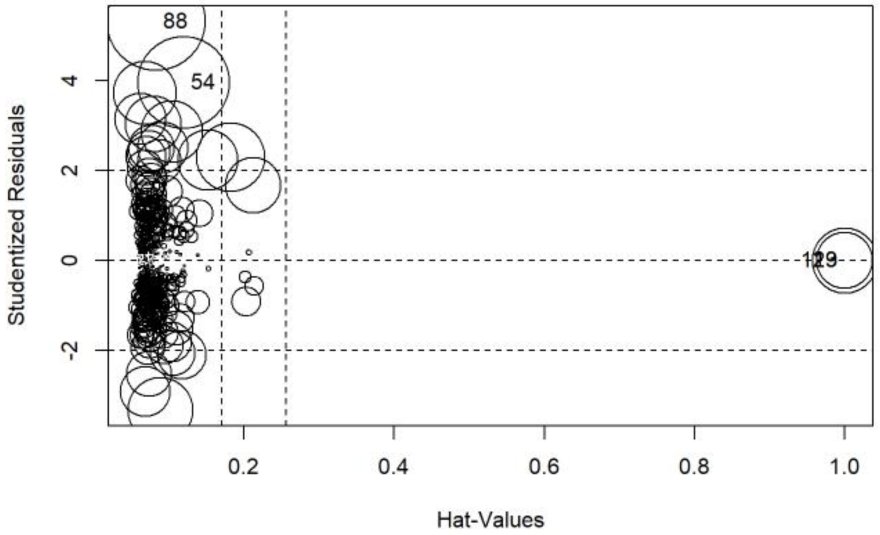
predictors:"Station","Microcystine" ,"Chlorophylle" "PCChl","Ventmoyenne", "Ventmax","Transparence", "Temp", "Saturation", "Oxygene", "PH","TDS","ConductivitE", "Coliformes","Ecoli","Lat","Long","t"

From the report,the Adjusted R-squared of model2c is 0.7467.

The best stepwise regression: lm(formula = Phycocyanine ~ Station + Chlorophylle + Transparence + Oxygene + PH + TDS + Coliformes + t, data = resultb)

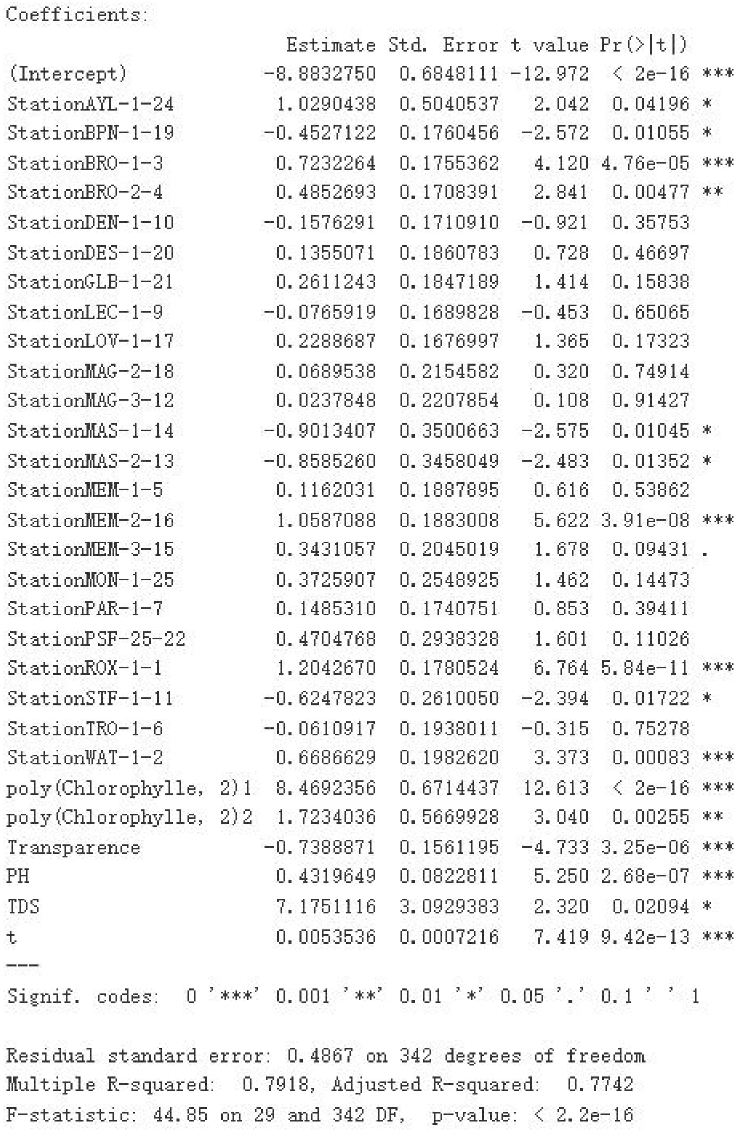






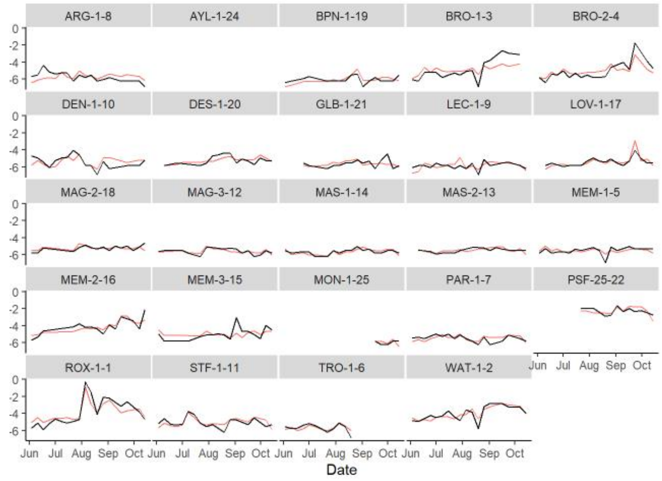
The influencePlot shows that there are some outliers in the model. We remove these outliers and re-run the model call it model2d.The numeric indexes of outliers are 54,88 and 123.

The Adjusted R-squared of model2d is 0.7742.Regression results showed that Chlorophylle,Transparence,PH,TDS,t and part of stations was significant at 0.05 confidence level.



```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)           -8.8832750  0.6848111 -12.972  < 2e-16 ***
StationAYL-1-24        1.0290438  0.5040537   2.042  0.04196 *  
StationBPN-1-19       -0.4527122  0.1760456  -2.572  0.01055 *  
StationBRO-1-3         0.7232264  0.1755362   4.120 4.76e-05 ***
StationBRO-2-4         0.4852693  0.1708391   2.841  0.00477 ** 
StationDEN-1-10       -0.1576291  0.1710910  -0.921  0.35753    
StationDES-1-20        0.1355071  0.1860783   0.728  0.46697    
StationGLB-1-21        0.2611243  0.1847189   1.414  0.15838    
StationLEC-1-9        -0.0765919  0.1689828  -0.453  0.65065    
StationLOV-1-17        0.2288687  0.1676997   1.365  0.17323    
StationMAG-2-18        0.0689538  0.2154582   0.320  0.74914    
StationMAG-3-12        0.0237848  0.2207854   0.108  0.91427    
StationMAS-1-14       -0.9013407  0.3500663  -2.575  0.01045 *  
StationMAS-2-13       -0.8585260  0.3458049  -2.483  0.01352 *  
StationMEM-1-5         0.1162031  0.1887895   0.616  0.53862    
StationMEM-2-16        1.0587088  0.1883008   5.622 3.91e-08 ***
StationMEM-3-15        0.3431057  0.2045019   1.678  0.09431 .  
StationMON-1-25        0.3725907  0.2548925   1.462  0.14473    
StationPAR-1-7         0.1485310  0.1740751   0.853  0.39411    
StationPSF-25-22       0.4704768  0.2938328   1.601  0.11026    
StationROX-1-1         1.2042670  0.1780524   6.764 5.84e-11 ***
StationSTF-1-11       -0.6247823  0.2610050  -2.394  0.01722 *  
StationTRO-1-6        -0.0610917  0.1938011  -0.315  0.75278    
StationWAT-1-2         0.6686629  0.1982620   3.373  0.00083 ***
poly(Chlorophylle, 2)1 8.4692356  0.6714437  12.613  < 2e-16 ***
poly(Chlorophylle, 2)2 1.7234036  0.5669928   3.040  0.00255 ** 
Transparence          -0.7388871  0.1561195  -4.733 3.25e-06 ***
PH                     0.4319649  0.0822811   5.250 2.68e-07 ***
TDS                    7.1751116  3.0929383   2.320  0.02094 *  
t                      0.0053536  0.0007216   7.419 9.42e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4867 on 342 degrees of freedom
Multiple R-squared:  0.7918, Adjusted R-squared:  0.7742 
F-statistic: 44.85 on 29 and 342 DF,  p-value: < 2.2e-16
```



3.5 Chlorophylle

1)model-1

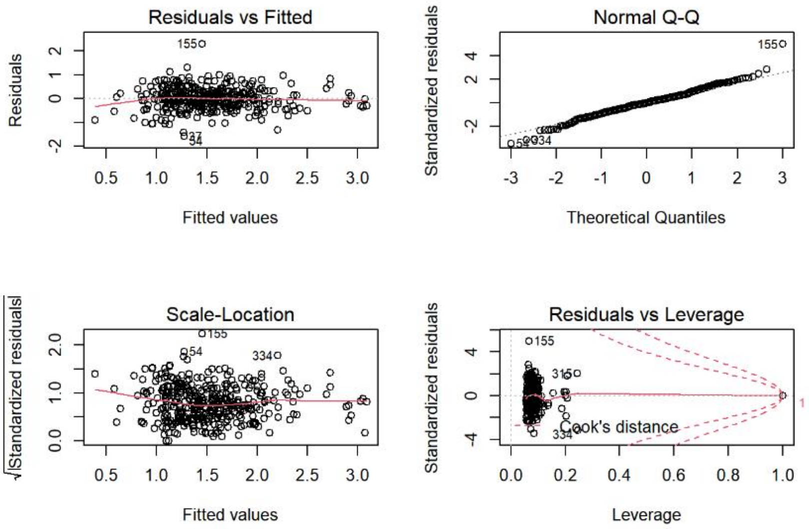
Next, run a linear regression for Chlorophylle,call it model3a.

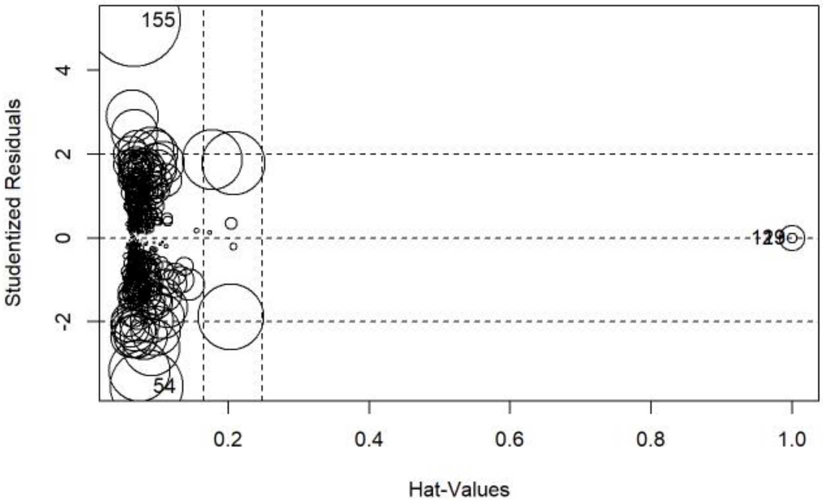
response： Chlorophylle

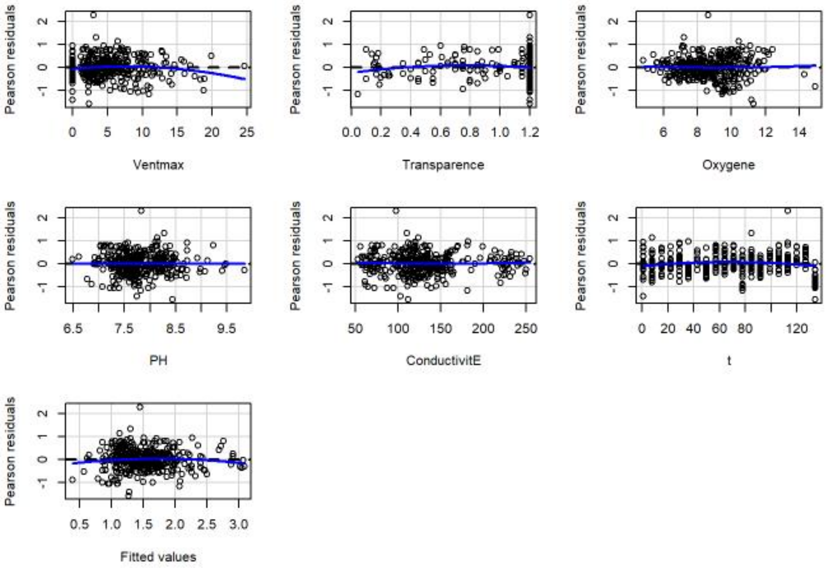
Predictors:“Station” ,“Chlorophylle”, “Ventmoyenne”,“Ventmax”,“Transparence”, “Temp”, “Saturation”,“Oxygene”, “PH”,“TDS”,“ConductivitE”, “Coliformes”, “Ecoli”,“Lat”,“Long”,“t”.

The report shows that the Adjusted R-squared is 0.4352.

The best sepwise regression:lm(formula = Chlorophylle ~ Station + Ventmax + Transparence + Oxygene + PH + ConductivitE + t, data = resultc)

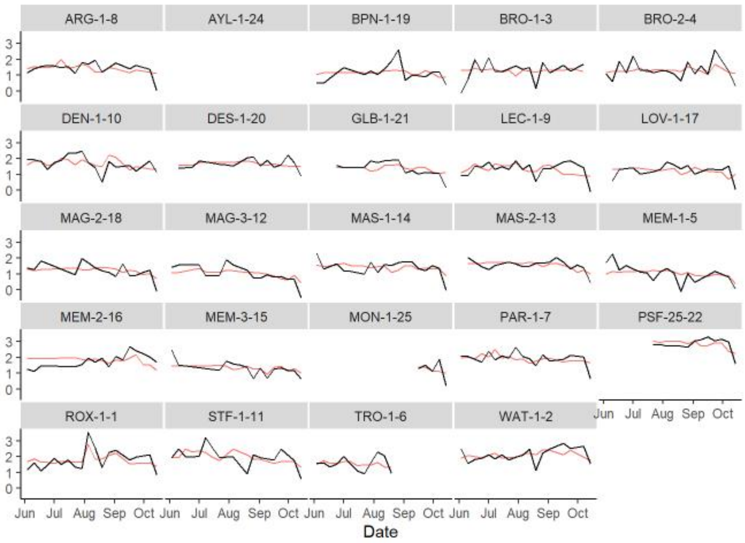


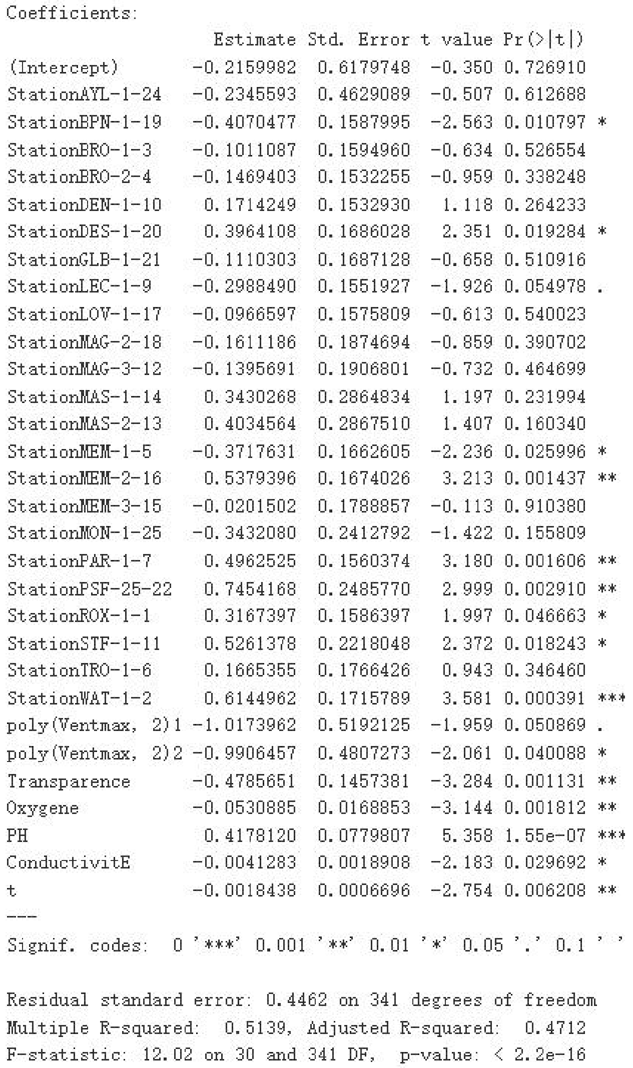




Residual plot show that Ventmax show non-linear relations in the residual plots.This paper re-run the linear model with polynomial order 2 and I also remove these outliers.The numeric indexes of outliers are 54,123,155 and 334 .We remove these outliers and re-run the model call it mode3b.

From the report ,the Adjusted R-squared is 0.471.Regression results showed that poly(Ventmax,2),Transparence,t,PH,Oxygene,ConductivitE and part of stations were significant at 0.05 confidence level.The regression coefficients show that the effects of ploy(Ventmax,2),Transparency, Oxygene, ConductivitE, t on Chlorophylle are negative and PH is positive.





```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -0.2159982  0.6179748  -0.350 0.726910    
StationAYL-1-24   -0.2345593  0.4629089  -0.507 0.612688    
StationBPN-1-19   -0.4070477  0.1587995  -2.563 0.010797 *  
StationBRO-1-3    -0.1011087  0.1594960  -0.634 0.526554    
StationBRO-2-4    -0.1469403  0.1532255  -0.959 0.338248    
StationDEN-1-10    0.1714249  0.1532930   1.118 0.264233    
StationDES-1-20    0.3964108  0.1686028   2.351 0.019284 *  
StationGLB-1-21   -0.1110303  0.1687128  -0.658 0.510916    
StationLEC-1-9    -0.2988490  0.1551927  -1.926 0.054978 .  
StationLOV-1-17   -0.0966597  0.1575809  -0.613 0.540023    
StationMAG-2-18   -0.1611186  0.1874694  -0.859 0.390702    
StationMAG-3-12   -0.1395691  0.1906801  -0.732 0.464699    
StationMAS-1-14    0.3430268  0.2864834   1.197 0.231994    
StationMAS-2-13    0.4034564  0.2867510   1.407 0.160340    
StationMEM-1-5    -0.3717631  0.1662605  -2.236 0.025996 *  
StationMEM-2-16    0.5379396  0.1674026   3.213 0.001437 ** 
StationMEM-3-15   -0.0201502  0.1788857  -0.113 0.910380    
StationMON-1-25   -0.3432080  0.2412792  -1.422 0.155809    
StationPAR-1-7     0.4962525  0.1560374   3.180 0.001606 ** 
StationPSF-25-22   0.7454168  0.2485770   2.999 0.002910 ** 
StationROX-1-1     0.3167397  0.1586397   1.997 0.046663 *  
StationSTF-1-11    0.5261378  0.2218048   2.372 0.018243 *  
StationTRO-1-6     0.1665355  0.1766426   0.943 0.346460    
StationWAT-1-2     0.6144962  0.1715789   3.581 0.000391 ***
poly(Ventmax, 2)1 -1.0173962  0.5192125  -1.959 0.050869 .  
poly(Ventmax, 2)2 -0.9906457  0.4807273  -2.061 0.040088 *  
Transparence      -0.4785651  0.1457381  -3.284 0.001131 ** 
Oxygene           -0.0530885  0.0168853  -3.144 0.001812 ** 
PH                 0.4178120  0.0779807   5.358 1.55e-07 ***
ConductivitE      -0.0041283  0.0018908  -2.183 0.029692 *  
t                 -0.0018438  0.0006696  -2.754 0.006208 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 0.4462 on 341 degrees of freedom
Multiple R-squared:  0.5139, Adjusted R-squared:  0.4712 
F-statistic: 12.02 on 30 and 341 DF,  p-value: < 2.2e-16
```

2)model-2

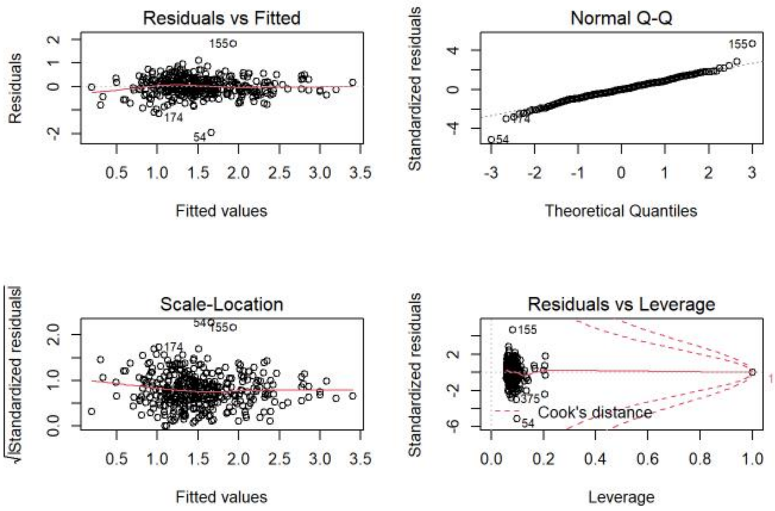
Next, run a linear regression for Chlorophylle,call it model3c.

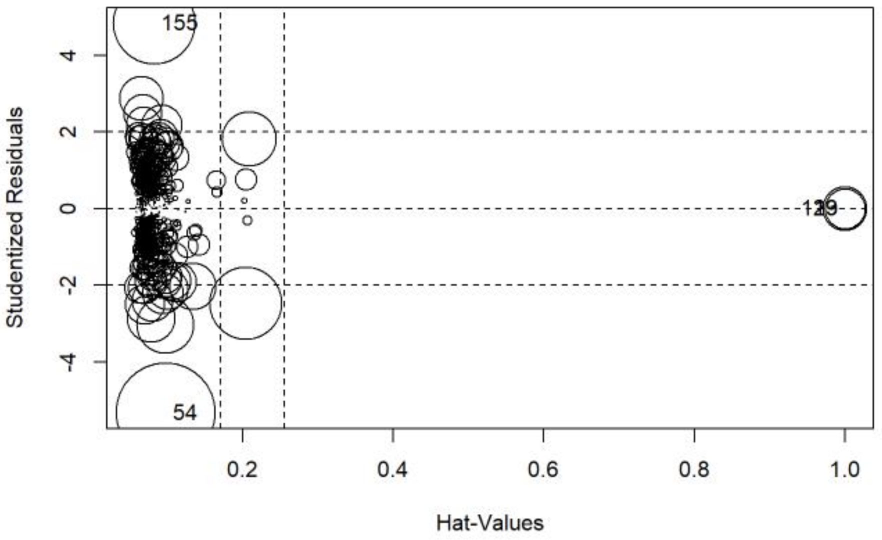
response： Chlorophylle

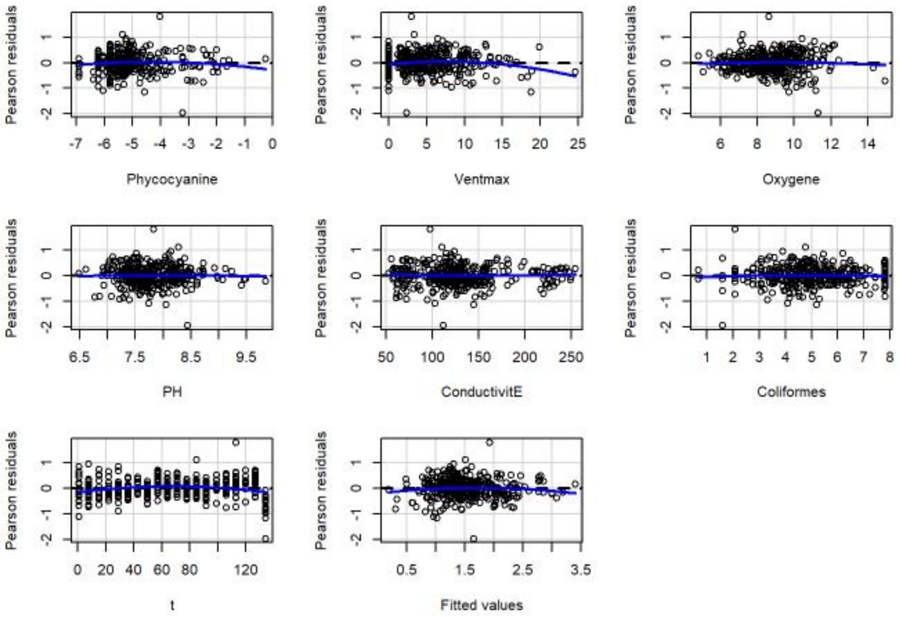
predictors:"Station","Microcystine" "Phycocyanine","PCChl","Ventmoyenne", "Ventmax","Transparence", "Temp", "Saturation", "Oxygene", "PH","TDS","ConductivitE", "Coliformes","Ecoli","Lat","Long","t"

The report shows that the Adjusted R-squared is 0.5892.

> The best sepwise regression:(lm(formula = Chlorophylle ~ Station + Phycocyanine + Ventmax + Oxygene + PH + ConductivitE + Coliformes + t, data = result c)

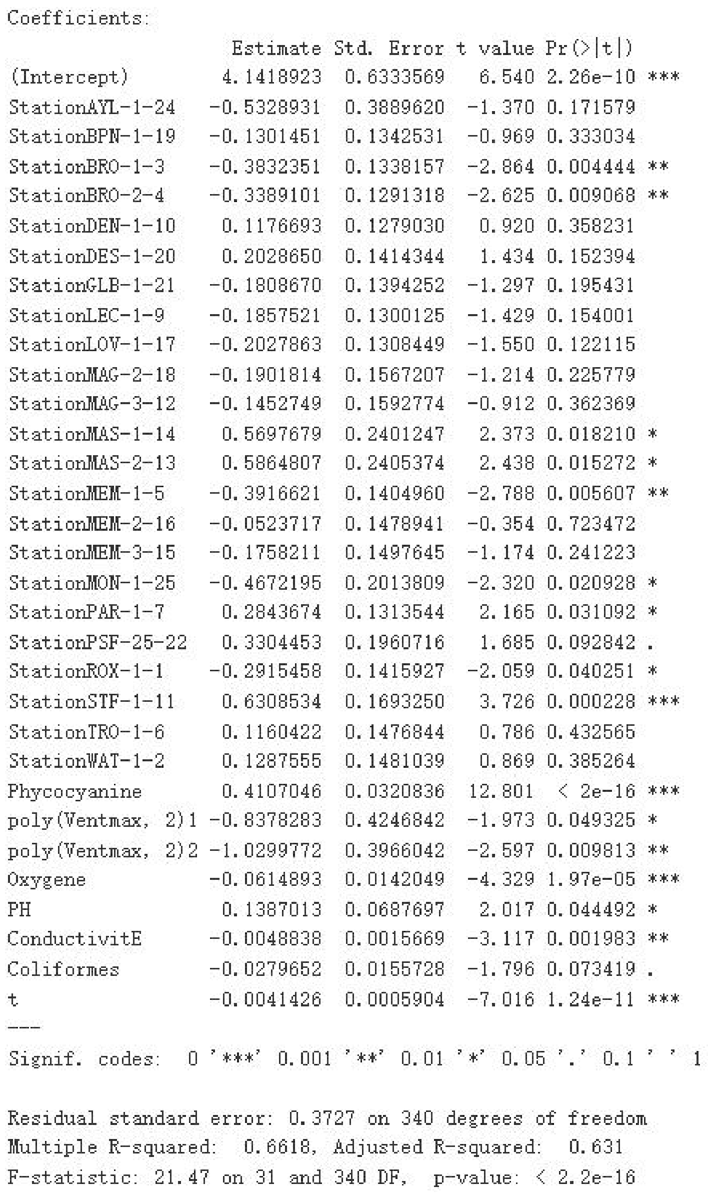

Residuals vs Fitted
Residuals
Fitted values
155
174
54
Normal Q-Q
Standardized residuals
Theoretical Quantiles
155
174
54
Scale-Location
√|Standardized residuals|
Fitted values
54
155
174
Residuals vs Leverage
Standardized residuals
Leverage
155
375
54
Cook's distance
1


Studentized Residuals
Hat-Values
155
54
129
0.2
0.4
0.6
0.8
1.0


Pearson residuals
Phycocyanine
Ventmax
Oxygene
PH
ConductivitE
Coliformes
t
Fitted values

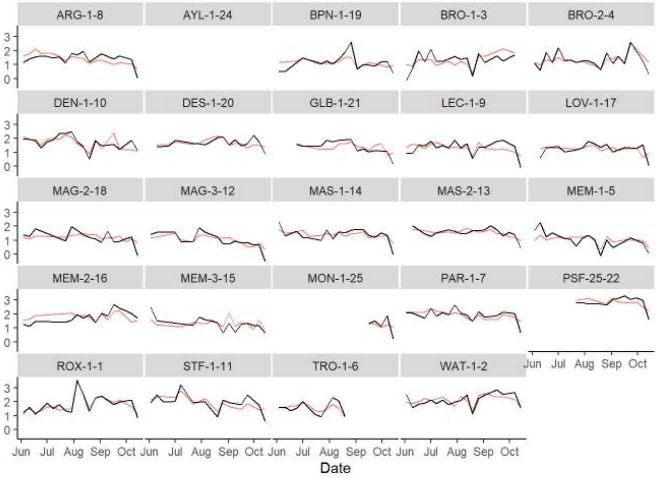
Residual plot show that Ventmax show non-linear relations in the residual plots.This paper re-run the linear model with polynomial order 2 and I also remove these outliers.The numeric indexes of outliers are 54,123,155 .We remove these outliers and re-run the model call it mode3c.

From the report ,the Adjusted R-squared is 0.631.Regression results showed that poly(Ventmax,2),t,PH,,ConductivitE,Coliformes and part of stations were significant at 0.05 confidence level.



```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)        4.1418923  0.6333569   6.540 2.26e-10 ***
StationAYL-1-24   -0.5328931  0.3889620  -1.370 0.171579    
StationBPN-1-19   -0.1301451  0.1342531  -0.969 0.333034    
StationBRO-1-3    -0.3832351  0.1338157  -2.864 0.004444 ** 
StationBRO-2-4    -0.3389101  0.1291318  -2.625 0.009068 ** 
StationDEN-1-10    0.1176693  0.1279030   0.920 0.358231    
StationDES-1-20    0.2028650  0.1414344   1.434 0.152394    
StationGLB-1-21   -0.1808670  0.1394252  -1.297 0.195431    
StationLEC-1-9    -0.1857521  0.1300125  -1.429 0.154001    
StationLOV-1-17   -0.2027863  0.1308449  -1.550 0.122115    
StationMAG-2-18   -0.1901814  0.1567207  -1.214 0.225779    
StationMAG-3-12   -0.1452749  0.1592774  -0.912 0.362369    
StationMAS-1-14    0.5697679  0.2401247   2.373 0.018210 *  
StationMAS-2-13    0.5864807  0.2405374   2.438 0.015272 *  
StationMEM-1-5    -0.3916621  0.1404960  -2.788 0.005607 ** 
StationMEM-2-16   -0.0523717  0.1478941  -0.354 0.723472    
StationMEM-3-15   -0.1758211  0.1497645  -1.174 0.241223    
StationMON-1-25   -0.4672195  0.2013809  -2.320 0.020928 *  
StationPAR-1-7     0.2843674  0.1313544   2.165 0.031092 *  
StationPSF-25-22   0.3304453  0.1960716   1.685 0.092842 .  
StationROX-1-1    -0.2915458  0.1415927  -2.059 0.040251 *  
StationSTF-1-11    0.6308534  0.1693250   3.726 0.000228 ***
StationTRO-1-6     0.1160422  0.1476844   0.786 0.432565    
StationWAT-1-2     0.1287555  0.1481039   0.869 0.385264    
Phycocyanine       0.4107046  0.0320836  12.801  < 2e-16 ***
poly(Ventmax, 2)1 -0.8378283  0.4246842  -1.973 0.049325 *  
poly(Ventmax, 2)2 -1.0299772  0.3966042  -2.597 0.009813 ** 
Oxygene           -0.0614893  0.0142049  -4.329 1.97e-05 ***
PH                 0.1387013  0.0687697   2.017 0.044492 *  
ConductivitE      -0.0048838  0.0015669  -3.117 0.001983 ** 
Coliformes        -0.0279652  0.0155728  -1.796 0.073419 .  
t                 -0.0041426  0.0005904  -7.016 1.24e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3727 on 340 degrees of freedom
Multiple R-squared:  0.6618,	Adjusted R-squared:  0.631 
F-statistic: 21.47 on 31 and 340 DF,  p-value: < 2.2e-16
```



3.6 PCChl

1) Model-1

Here, run a linear regression for PCChl on all the predictors,call it model4a.

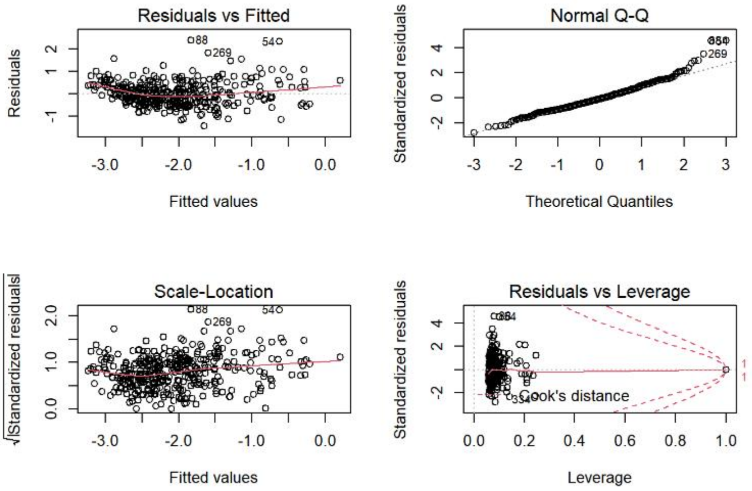
Response:PCChl

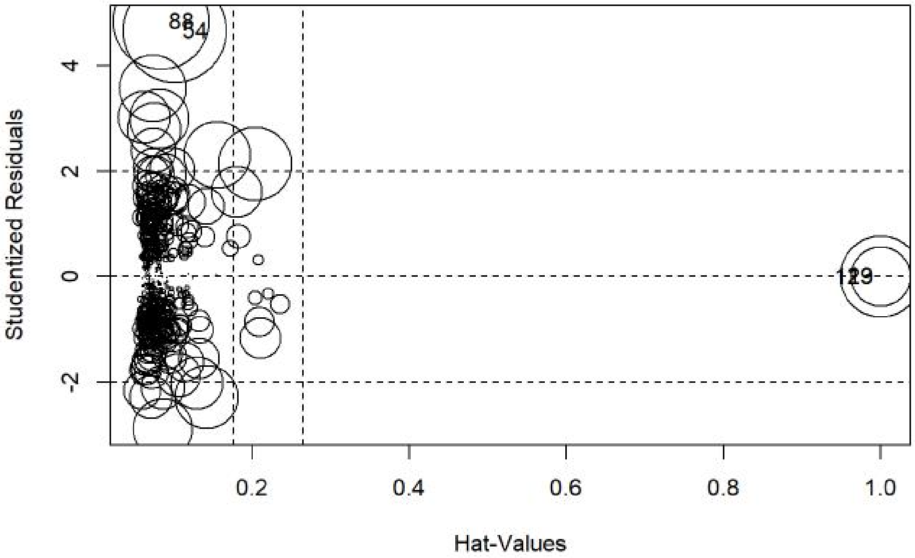
Predictors:

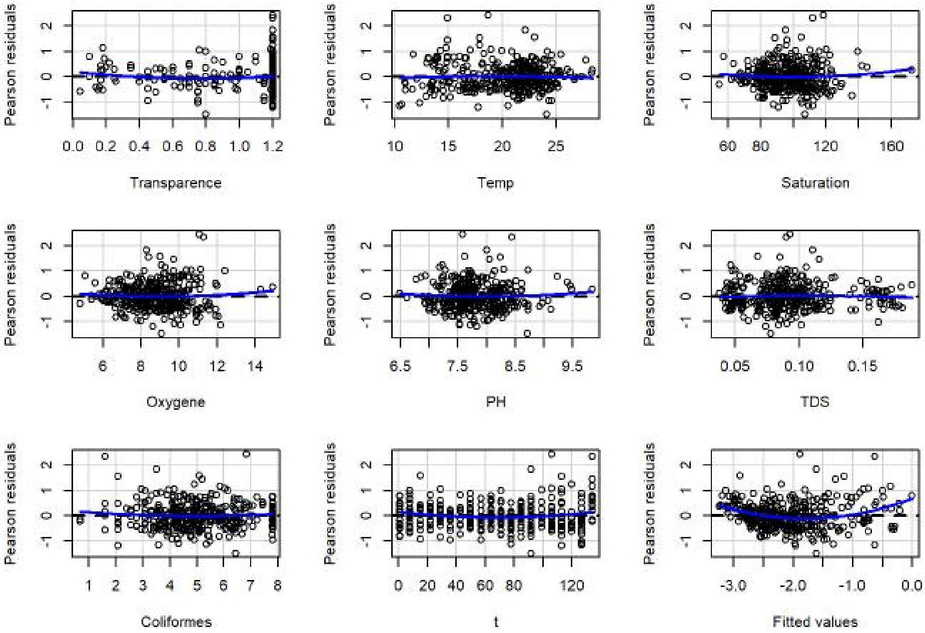
“Station” ,“Ventmoyenne”,“Ventmax”,“Transparence”, “Temp”, “Saturation”,“Oxygene”, “PH”,“TDS”,“ConductivitE”, “Coliformes”, “Ecoli”,“Lat”,“Long”,“t”.

the Adjusted R-squared of model4 is 0.5768.

The best stepwise regression:lm(formula = PCChl ~ Station + Transparence + Temp + Saturation + Oxygene + PH + TDS + Coliformes + t, data = resultd)

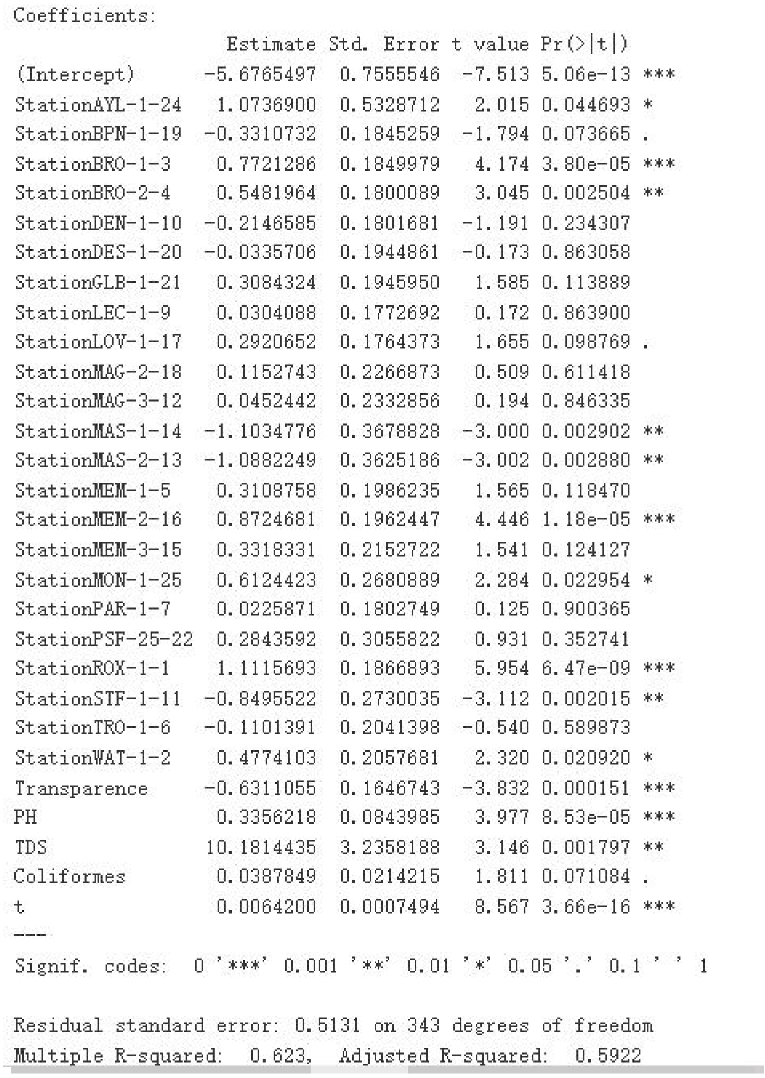






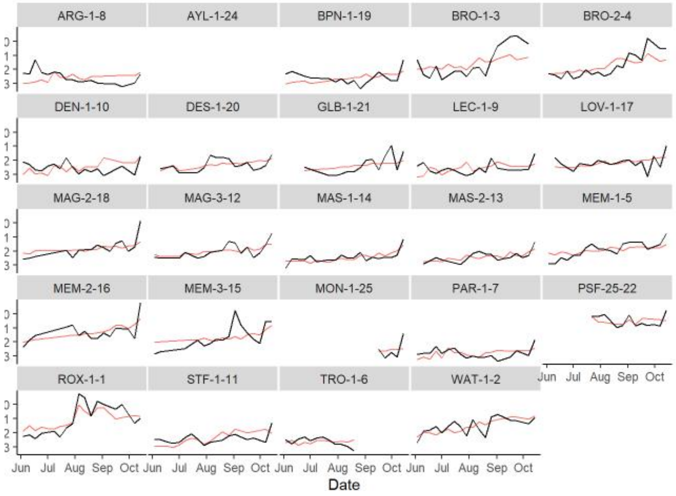
It can be seen that Transparence shows non-linear relations in the residual plots.I re-run the linear model with polynomial order 2. I also remove these outliers.The numeric indexes of outliers are 54,88 and 123.We remove these outliers and re-run the model call it mode4b.

From the report ,the Adjusted R-squared is 0.5922.From the results of regression model,t,Transparence,TDS,PH,Coliformes and part of stations were significant at 0.05 confidence level.The regression coefficients show that t,Coliformes,TDS,PH have a positive effect on PCChl.



```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -5.6765497  0.7555546  -7.513 5.06e-13 ***
StationAYL-1-24   1.0736900  0.5328712   2.015 0.044693 *  
StationBPN-1-19  -0.3310732  0.1845259  -1.794 0.073665 .  
StationBRO-1-3    0.7721286  0.1849979   4.174 3.80e-05 ***
StationBRO-2-4    0.5481964  0.1800089   3.045 0.002504 ** 
StationDEN-1-10  -0.2146585  0.1801681  -1.191 0.234307    
StationDES-1-20  -0.0335706  0.1944861  -0.173 0.863058    
StationGLB-1-21   0.3084324  0.1945950   1.585 0.113889    
StationLEC-1-9    0.0304088  0.1772692   0.172 0.863900    
StationLOV-1-17   0.2920652  0.1764373   1.655 0.098769 .  
StationMAG-2-18   0.1152743  0.2266873   0.509 0.611418    
StationMAG-3-12   0.0452442  0.2332856   0.194 0.846335    
StationMAS-1-14  -1.1034776  0.3678828  -3.000 0.002902 ** 
StationMAS-2-13  -1.0882249  0.3625186  -3.002 0.002880 ** 
StationMEM-1-5    0.3108758  0.1986235   1.565 0.118470    
StationMEM-2-16   0.8724681  0.1962447   4.446 1.18e-05 ***
StationMEM-3-15   0.3318331  0.2152722   1.541 0.124127    
StationMON-1-25   0.6124423  0.2680889   2.284 0.022954 *  
StationPAR-1-7    0.0225871  0.1802749   0.125 0.900365    
StationPSF-25-22  0.2843592  0.3055822   0.931 0.352741    
StationROX-1-1    1.1115693  0.1866893   5.954 6.47e-09 ***
StationSTF-1-11  -0.8495522  0.2730035  -3.112 0.002015 ** 
StationTRO-1-6   -0.1101391  0.2041398  -0.540 0.589873    
StationWAT-1-2    0.4774103  0.2057681   2.320 0.020920 *  
Transparence     -0.6311055  0.1646743  -3.832 0.000151 ***
PH                0.3356218  0.0843985   3.977 8.53e-05 ***
TDS              10.1814435  3.2358188   3.146 0.001797 ** 
Coliformes        0.0387849  0.0214215   1.811 0.071084 .  
t                 0.0064200  0.0007494   8.567 3.66e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5131 on 343 degrees of freedom
Multiple R-squared:  0.623,	Adjusted R-squared:  0.5922
```



2) model-2

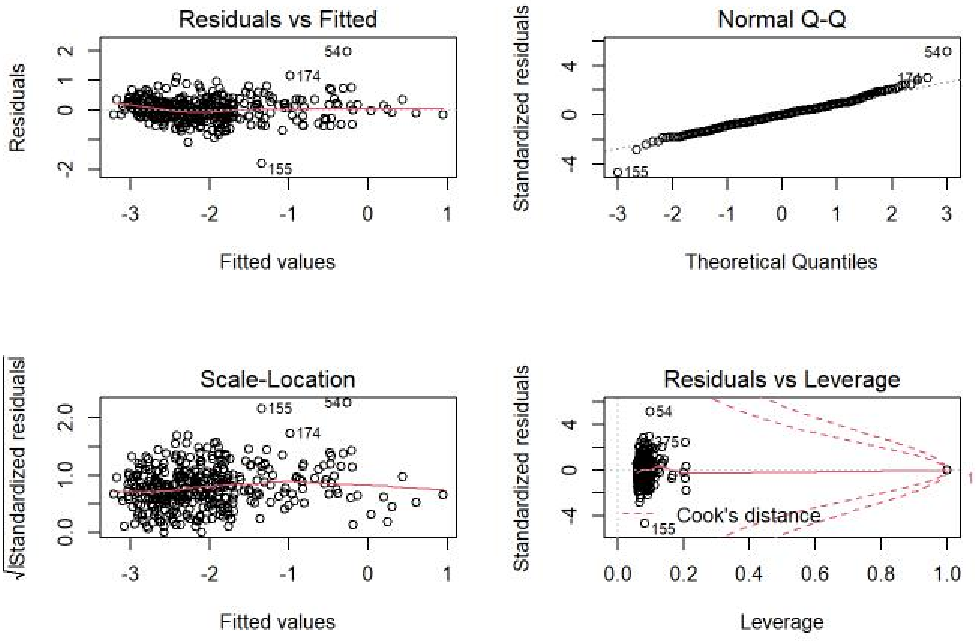
Next, run a linear regression for PCChl,call it model4c.

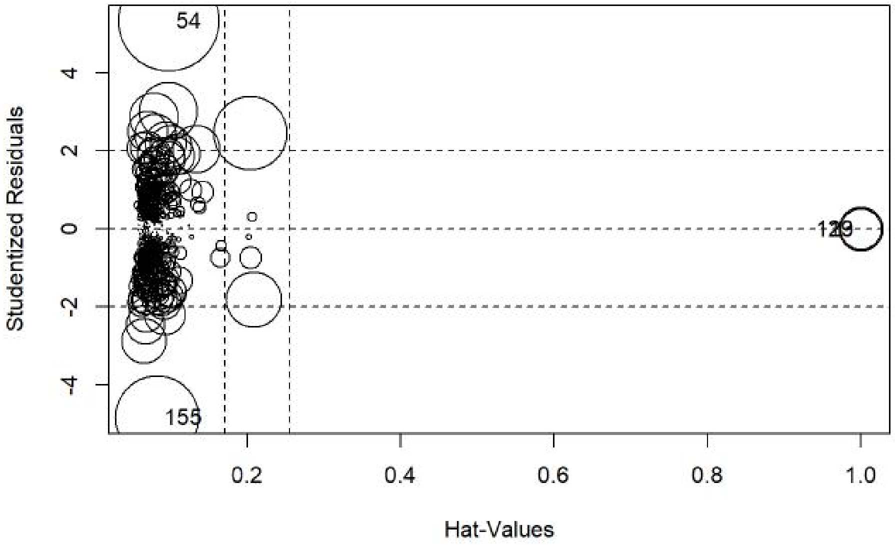
response： PCChl

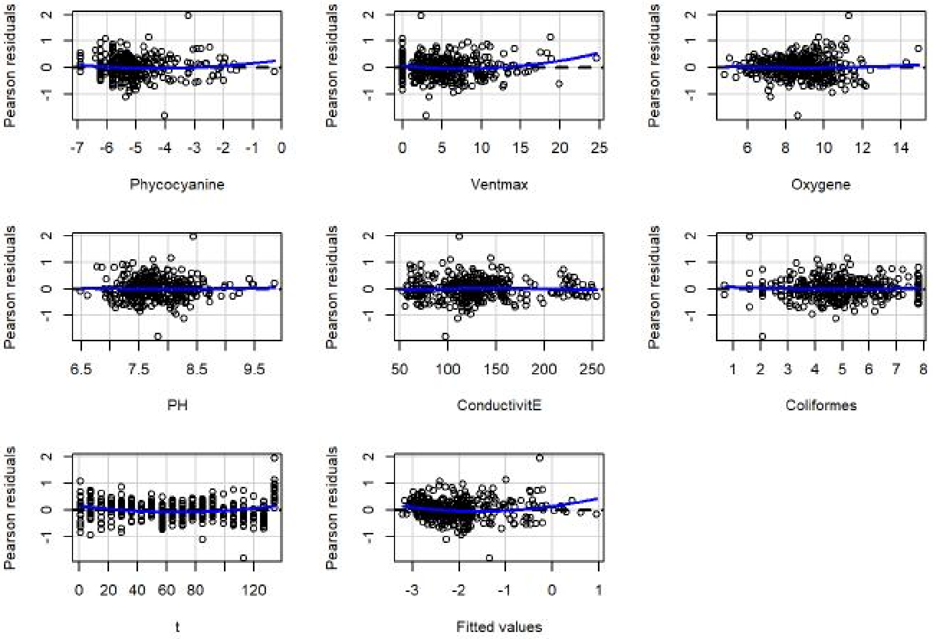
predictors:"Station","Microcystine" "Phycocyanine","Chlorophylle","Ventmoyenne", "Ventmax","Transparence", "Temp", "Saturation",   "Oxygene", "PH","TDS","ConductivitE", "Coliformes","Ecoli","Lat","Long","t"

The report shows that the Adjusted R-squared is 0.7668.

```
The best sepwise regression:lm(formula = PCChl ~ Station + Phycocyanine + Ve
ntmax + Oxygene + PH + ConductivitE + Coliformes + t, data = resultd)
```

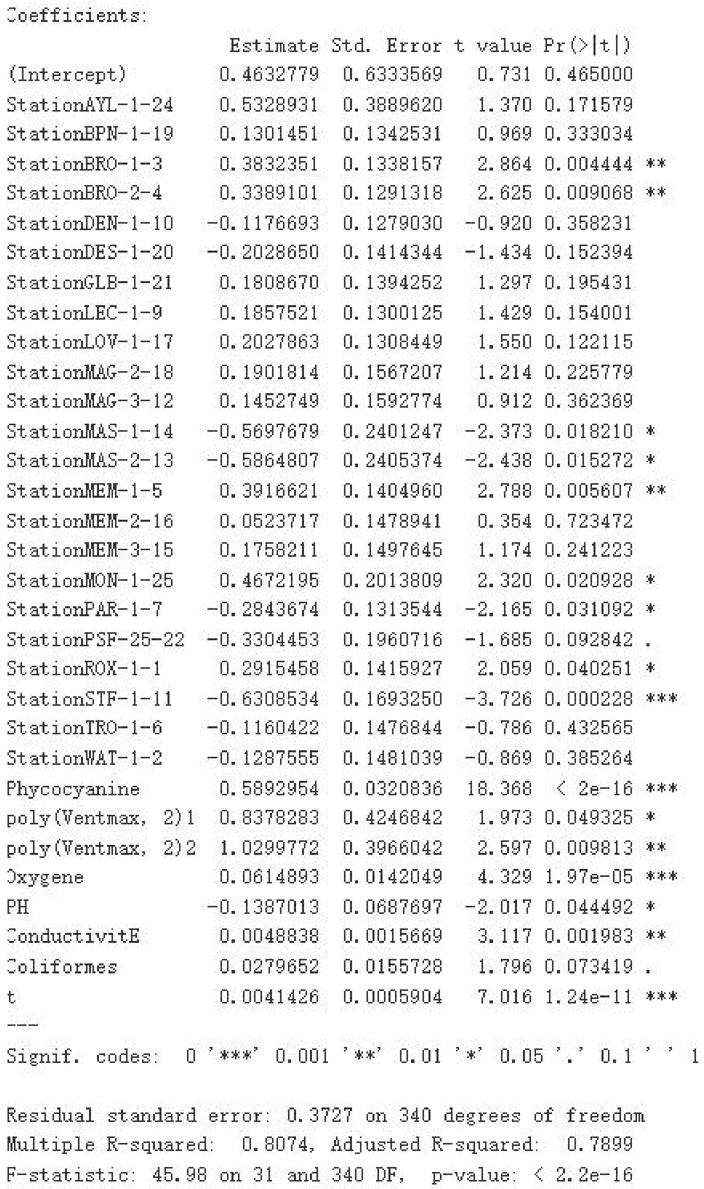






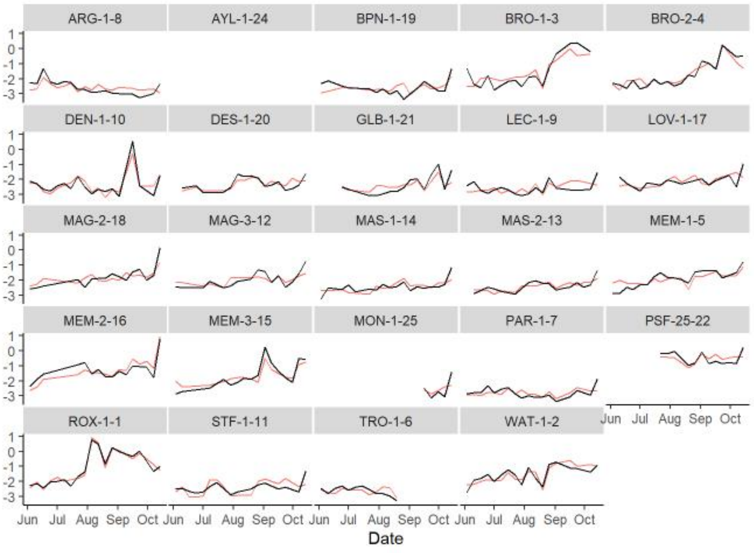
It can be seen that Transparence shows non-linear relations in the residual plots.I re-run the linear model with polynomial order 2. I also remove these outliers.The numeric indexes of outliers are 55,155 and 123.We remove these outliers and re-run the model call it model4d.

From the report ,the Adjusted R-squared is 0.7899.From the results of regression model,t,Coliformes,Oxygene ,ConductivitE ,poly(Ventmax,2) and part of stations were significant at 0.05 confidence level.



```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.4632779  0.6333569   0.731 0.465000
StationAYL-1-24    0.5328931  0.3889620   1.370 0.171579
StationBPN-1-19    0.1301451  0.1342531   0.969 0.333034
StationBRO-1-3     0.3832351  0.1338157   2.864 0.004444 **
StationBRO-2-4     0.3389101  0.1291318   2.625 0.009068 **
StationDEN-1-10   -0.1176693  0.1279030  -0.920 0.358231
StationDES-1-20   -0.2028650  0.1414344  -1.434 0.152394
StationGLB-1-21    0.1808670  0.1394252   1.297 0.195431
StationLEC-1-9     0.1857521  0.1300125   1.429 0.154001
StationLOV-1-17    0.2027863  0.1308449   1.550 0.122115
StationMAG-2-18    0.1901814  0.1567207   1.214 0.225779
StationMAG-3-12    0.1452749  0.1592774   0.912 0.362369
StationMAS-1-14   -0.5697679  0.2401247  -2.373 0.018210 *
StationMAS-2-13   -0.5864807  0.2405374  -2.438 0.015272 *
StationMEM-1-5     0.3916621  0.1404960   2.788 0.005607 **
StationMEM-2-16    0.0523717  0.1478941   0.354 0.723472
StationMEM-3-15    0.1758211  0.1497645   1.174 0.241223
StationMON-1-25    0.4672195  0.2013809   2.320 0.020928 *
StationPAR-1-7    -0.2843674  0.1313544  -2.165 0.031092 *
StationPSF-25-22  -0.3304453  0.1960716  -1.685 0.092842 .
StationROX-1-1     0.2915458  0.1415927   2.059 0.040251 *
StationSTF-1-11   -0.6308534  0.1693250  -3.726 0.000228 ***
StationTRO-1-6    -0.1160422  0.1476844  -0.786 0.432565
StationWAT-1-2    -0.1287555  0.1481039  -0.869 0.385264
Phycocyanine       0.5892954  0.0320836  18.368  < 2e-16 ***
poly(Ventmax, 2)1  0.8378283  0.4246842   1.973 0.049325 *
poly(Ventmax, 2)2  1.0299772  0.3966042   2.597 0.009813 **
Oxygene            0.0614893  0.0142049   4.329 1.97e-05 ***
PH                -0.1387013  0.0687697  -2.017 0.044492 *
ConductivitE       0.0048838  0.0015669   3.117 0.001983 **
Coliformes         0.0279652  0.0155728   1.796 0.073419 .
t                  0.0041426  0.0005904   7.016 1.24e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3727 on 340 degrees of freedom
Multiple R-squared:  0.8074, Adjusted R-squared:  0.7899
F-statistic: 45.98 on 31 and 340 DF,  p-value: < 2.2e-16
```



4 Results and discussion

In this paper, the data of cyanotoxins at 25 stations were studied. Firstly, the data is preprocessed and then the data distribution of the four main variables is examined. Finally, the factors affecting cyanotoxins were analyzed by nonlinear model.

1)without including the other three toxins

The final four models are as follows:

```
model1 <-lm(formula = Microcystine ~ Station + Phycocyanine + Temp + PH
+ Coliformes + Ecoli + t, data = resulta)

Model2 =:lm(formula = Phycocyanine ~ Station + Transparence + Saturation +
Oxygene + PH + t, data = resultb)

Model3 =lm(formula = Chlorophylle ~ Station + Ventmax + Transparence + Ox
ygene + PH + ConductivitE + t, data = resultc)

Model4 =lm(formula = PCChl ~ Station + Transparence + Temp + Saturation
+ Oxygene + PH + TDS + Coliformes + t, data = resultd)
```

2) including the other three toxins

```
Model5<-lm(formula = Microcystine ~ Station + Phycocyanine + Temp + PH +
 Coliformes + Ecoli + t, data = resulta)

Model6<-lm(formula = Phycocyanine ~ Station + Chlorophylle + Transparence +
  Oxygene + PH + TDS

+ Coliformes + t, data = resultb)

Model7<-lm(formula = Chlorophylle ~ Station + Phycocyanine + Ventmax + Ox
ygene + PH + ConductivitE + Coliformes + t, data = resultc)

Model8<-lm(formula = PCChl ~ Station + Phycocyanine + Ventmax + Oxygene
 + PH + ConductivitE + Coliformes + t, data = resultd)
```

R-squared of the model

	including	without
Microcystine	0.4978	0.5049
Phycocyanine	0.6693	0.7742
Chlorophylle	0.4712	0.631
PCChl	0.5922	0.7899

From the results, the model with three variables performs better than the model without them.The Adjusted R-squared of model1d is 0.5049.Regression results showed that PH,Temp,t,Ecoli ,Coliformes and part of stations was significant at 0.05 confidence level. Regression coefficients showed that Phycocyanine ,Temp,Ecoli had positive effects on Microcystine.PH and t had Negative effects on Microcystine.The Adjusted R-squared of model2d is 0.7742.Regression results showed that Chlorophylle,Oxygene,Transparence,PH,TDS,t and part of stations was significant at 0.05 confidence level. Regression coefficients showed that Chlorophylle,

Xxygene,PH,Coliformes,TDS,t had positive effects on Phycocyanine.

From the report ,the Adjusted R-squared of model3d is 0.631.Regression results showed that poly(Ventmax,2),t,PH,,ConductivitE,Coliformes and part of stations were significant at 0.05 confidence level.From the report ,the Adjusted R-squared of model4d is 0.7899.From the results of regression model,t,Coliformes,Oxygene ,ConductivitE ,

poly(Ventmax,2) and part of stations were significant at 0.05 confidence level.

5 Code and appendix

Data description and exploration:

1 summary

```
for (i in 1:32){
   df[,i] <- gsub("[*]","",df[,i])
}
for (i in 1:32){
   df[,i] <- gsub("<3","2",df[,i])
   df[,i] <- gsub(">2424","2425",df[,i])
}
stations = unique(df$Station)
fs <- function(df,n){
   result = data.frame()
   for ( s in stations){
      ds1<-subset(df,Station==s)
      values= as.vector(ds1[,n])
      v = values[!(values %in% c("vent","moyen","faible","fort"))]
      #v[is.na(v)] = 0
      #v1 = v
      v1 = as.numeric(v)
      v2= t(summary(v1))
      ds1[,n] <- gsub("vent",v2[4],ds1[,n])
      ds1[,n] <- gsub("moyen",v2[4],ds1[,n])
      ds1[,n] <- gsub("faible",v2[2],ds1[,n])
      ds1[,n] <- gsub("fort",v2[5],ds1[,n])
      result = rbind(result,ds1)
   }
   return(result)
}
df_a = fs(df,9)
df_b = fs(df_a,10)
```

```
nums <-c(1,3:10,13:20,22:32)
df_c <- df_b
for (i in nums) {
    df_c[,i] <- as.numeric(df_c[,i])
}
numss <- c(2,11,12,21)
for (i in numss) {
    df_c[,i] <- as.factor(df_c[,i])
}
skim(df_c)
```

2 Data collation

```
df_c$date<-as.Date(df_c$Date,origin='1900-1-1')-2
data = df_c[,c(33,1:6)]
names(data) =
c("Date","id","Station","Microcystine","Phycocyanine","Chlorophylle","PCChl")
data_long = data %>%
    pivot_longer(-c(Date, Station,id), names_to = "cyanotoxins", values_to =
"value",values_drop_na = T)
head(data_long)
```

3 Time-Series Plots

```
f <- function(ca){
  data_sub = data_long %>%
    filter(cyanotoxins == ca)
TS <- ggplot(data_sub) +
    geom_line(aes(x = Date, y = value)) +
    facet_wrap(~Station, ncol = 5) +
    xlab("Date") +
    ylab(ca) +
    theme(panel.spacing = unit(0.1, "lines"))+
    theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
panel.background = element_blank(), axis.line = element_line(colour = "black"))
print(TS)
}
```

```
f("Microcystine")
```

4 Hovmoller Plots

```
dt_a = dt[,c(1,4:5)]
names(dt_a) = c("Station","Lat","Long")
data_join = left_join(data_long, dt_a , by = "Station")
f0 <- function(ca){
T1 <- filter(data_join,
                    cyanotoxins    == ca)
T1$t = T1$id - 39601
lim_lat <- range(T1$Lat) # latitude range
lim_t <- range(T1$t) # time range
lat_axis <- seq(lim_lat[1], lim_lat[2],length=25)
t_axis <- seq(lim_t[1], lim_t[2],length=100)
lat_t_grid <- expand.grid(lat = lat_axis,
t = t_axis)
T1_grid <- T1
dists <- abs(outer(T1$Lat, lat_axis, "-"))
T1_grid$Lat <- lat_axis[apply(dists, 1, which.min)]
T1_lat_Hov <- ddply(T1_grid, .(Lat, t) ,summarize,z = mean(value))
Hovmoller_lat <- ggplot(T1_lat_Hov) + # take data
geom_tile(aes(x = Lat, y = t, fill = z)) +
scale_y_reverse() + # rev y scale
ylab("Day number (days)") + # add y label
xlab("Latitude (degrees)") + # add x label
theme_bw() +
theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
panel.background = element_blank(), axis.line = element_line(colour = "black"))
Hovmoller_lat
}
f0("Microcystine")
```

5 Empirical Spatial Means

```
T1 <- data_join
```

```
T1$t = T1$id - 39601
spat_av <- ddply(T1,.(Lat,Long,cyanotoxins),summarize,Means = mean(value))
ggplot(spat_av) +
geom_point(aes(Long,Lat, colour = Means)) +
xlab("Longitude (deg)") +
ylab("Latitude(deg)") + theme_bw()+ggtitle("Empirical Spatial Means")+
facet_wrap(~cyanotoxins, ncol = 5)+
theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
panel.background = element_blank(), axis.line = element_line(colour = "black"))
```

6 Empirical Temporal Means

```
T1 <-data_join
T1$t = T1$id - 39601
T1_av <- ddply(T1,.(Date,t,cyanotoxins),summarize,meanT1 = mean(value))
gTmaxav <-ggplot(T1_av) + geom_line(aes(x = Date, y = meanT1,group=
cyanotoxins,colour = cyanotoxins)) +theme_bw()+
   ggtitle("Empirical Temporal Means")+
   theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
panel.background = element_blank(), axis.line = element_line(colour = "black"))
gTmaxav
```

7 scatter plots

```
result2 = na.omit(result1)
los <- c("Microcystine","Phycocyanine","Chlorophylle","PCChl","Coliformes")
for (l in los){
    result2[l] = log(result2[l])
}
result2["Ecoli"] = log(result2["Ecoli"]+1)
panel.hist <- function(x, ...)
{
     usr <- par("usr"); on.exit(par(usr))
     par(usr = c(usr[1:2], 0, 1.5) )
     h <- hist(x, plot = FALSE)
     breaks <- h$breaks; nB <- length(breaks)
```

```
        y <- h$counts; y <- y/max(y)
        rect(breaks[-nB], 0, breaks[-1], y, col = "cyan", ...)
}
panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...)
{
        usr <- par("usr"); on.exit(par(usr))
        par(usr = c(0, 1, 0, 1))
        r <- abs(cor(x, y))
        txt <- format(c(r, 0.123456789), digits = digits)[1]
        txt <- paste0(prefix, txt)
        if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
        text(0.5, 0.5, txt, cex = cex.cor * r)
}
pairs(result2[,c(3:20)],panel = panel.cor,upper.panel = panel.smooth,
            diag.panel = panel.hist,
            data=result2[,c(3:20)],main="")
```

8 modeling

1 Microcystine

1) without

```
resulta    = result[,c(1:2,6:19)]
fit <- lm(Microcystine ~., data = resulta)
step.fit <- stepAIC(fit,direction = "both",trace = F)
summary(step.fit)
model1 = lm(formula = Microcystine ~ Station + Temp + Coliformes + Ecoli +
        t, data = resulta)
par(mfrow = c(2, 2))
plot(model1)

result4    = resulta[-c(368,321,123),]
model11 <- lm(formula = Microcystine ~ Station + Temp + Coliformes + Ecoli +
        t, data = result4)
summary(model11)
```

```
r1 = re[-c(368,321,123),]
r1['fit'] = model11$fitted.values
r1$date<-as.Date(r1$Date,origin='1900-1-1')-2
data = r1[,c(23,3,4,22)]
names(data) = c("Date","Station","Microcystine","fit")
TS <- ggplot(data) +
   geom_line(aes(x = Date, y = fit,color = "red")) +
     geom_line(aes(x = Date, y = Microcystine))+
   facet_wrap(~Station, ncol = 5) +
   xlab("Date") +
   ylab("") +
   theme(panel.spacing = unit(0.1, "lines"))+
   theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
panel.background = element_blank(), axis.line = element_line(colour = "black")) +
theme(legend.position="none")
print(TS)
```

2) Including

```
resulta    = result

fit <- lm(Microcystine ~., data = resulta)

step.fit <- stepAIC(fit,direction = "both",trace = F)
summary(step.fit)

model1 = lm(formula = Microcystine ~ Station + Phycocyanine + Temp + PH +
      Coliformes + Ecoli + t, data = resulta)
par(mfrow = c(2, 2))
plot(model1)
result4    = resulta[-c(321,368,123),]
model11 <-    lm(formula = Microcystine ~ Station + Phycocyanine + Temp + PH +
      Coliformes + Ecoli + t, data = result4)
summary(model11)
```

```
r1 = re[-c(368,321,123),]
r1['fit'] = model11$fitted.values
r1$date<-as.Date(r1$Date,origin='1900-1-1')-2
data = r1[,c(23,3,4,22)]
names(data) = c("Date","Station","Microcystine","fit")
TS <- ggplot(data) +
  geom_line(aes(x = Date, y = fit,color = "red")) +
   geom_line(aes(x = Date, y = Microcystine))+
  facet_wrap(~Station, ncol = 5) +
  xlab("Date") +
  ylab("") +
  theme(panel.spacing = unit(0.1, "lines"))+
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
panel.background = element_blank(), axis.line = element_line(colour = "black")) +
theme(legend.position="none")
print(TS)
```

2 Phycocyanine

1) Without

```
resultb = result[,c(1,3,6:19)]
fit <- lm(Phycocyanine ~., data = resultb)
step.fit <- stepAIC(fit,direction = "both",trace = F)
summary(step.fit)
model3 <- lm(formula = Phycocyanine ~ Station + Transparence + Saturation +
    Oxygene + PH + t, data = resultb)


par(mfrow = c(2, 2))
plot(model3)
result5    = resultb[-c(48,88,315,123),]
model21<-    lm(formula = Phycocyanine ~ Station + Transparence    + PH + t, data =
result5)
summary(model21)
r1 = re[-c(48,88,123,315),]
r1['fit'] = model21$fitted.values
```

```
r1$date<-as.Date(r1$Date,origin='1900-1-1')-2
data = r1[,c(23,3,5,22)]
names(data) = c("Date","Station","Phycocyanine","fit")
TS <- ggplot(data) +
   geom_line(aes(x = Date, y = fit,color = "red")) +
     geom_line(aes(x = Date, y = Phycocyanine))+
   facet_wrap(~Station, ncol = 5) +
   xlab("Date") +
   ylab("") +
   theme(panel.spacing = unit(0.1, "lines"))+
   theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
panel.background = element_blank(), axis.line = element_line(colour = "black")) +
theme(legend.position="none")
print(TS)
```

2) Including

```
resultb =    result[,c(1,2,3,4,6:19)]
fit <- lm(Phycocyanine ~., data = resultb)
step.fit <- stepAIC(fit,direction = "both",trace = F)
summary(step.fit)
model3 <- lm(formula = Phycocyanine ~ Station + Chlorophylle + Transparence +
       Oxygene + PH + TDS + Coliformes + t, data = resultb)
par(mfrow = c(2, 2))
plot(model3)
result5    = resultb[-c(88,54,123),]
model21<-    lm(formula = Phycocyanine ~ Station + poly(Chlorophylle,2) +
Transparence + PH + TDS + t, data = result5)
summary(model21)
r1 = re[-c(54,88,123),]
r1['fit'] = model21$fitted.values
r1$date<-as.Date(r1$Date,origin='1900-1-1')-2
data = r1[,c(23,3,5,22)]
names(data) = c("Date","Station","Phycocyanine","fit")
```

```
TS <- ggplot(data) +
   geom_line(aes(x = Date, y = fit,color = "red")) +
     geom_line(aes(x = Date, y = Phycocyanine))+
   facet_wrap(~Station, ncol = 5) +
   xlab("Date") +
   ylab("") +
   theme(panel.spacing = unit(0.1, "lines"))+
   theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
panel.background = element_blank(), axis.line = element_line(colour = "black")) +
theme(legend.position="none")
print(TS)
```

3 Chlorophylle

1) Without

```
resultc    = result[,c(1,4,6:19)]
fit <- lm(Chlorophylle ~., data = resultc)
step.fit <- stepAIC(fit,direction = "both",trace = F)
summary(step.fit)
model4 <- lm(formula = Chlorophylle ~ Station + Ventmax + Transparence +
       Oxygene + PH + ConductivitE + t, data = resultc)
par(mfrow = c(2, 2))
plot(model4)
result6    = resultc[-c(54,155,123),]
model31 <-      lm(formula = Chlorophylle ~ Station + poly(Ventmax,2) + Transparence
+ Oxygene + PH + ConductivitE + t, data = result6)
summary(model31)
r1 = re[-c(54,155,123),]
r1['fit'] = model31$fitted.values
r1$date<-as.Date(r1$Date,origin='1900-1-1')-2
data = r1[,c(23,3,6,22)]
names(data) = c("Date","Station","Chlorophylle","fit")
TS <- ggplot(data) +
   geom_line(aes(x = Date, y = fit,color = "red")) +
     geom_line(aes(x = Date, y = Chlorophylle))+
   facet_wrap(~Station, ncol = 5) +
   xlab("Date") +
   ylab("") +
   theme(panel.spacing = unit(0.1, "lines"))+
   theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
panel.background = element_blank(), axis.line = element_line(colour = "black")) +
theme(legend.position="none")
print(TS)
```

2) Including

```
resultc =   result[,c(1,2,3,4,6:19)]
fit <- lm(Chlorophylle ~., data = resultc)
step.fit <- stepAIC(fit,direction = "both",trace = F)
summary(step.fit)
model4 <- lm(formula = Chlorophylle ~ Station + Phycocyanine + Ventmax +
      Oxygene + PH + ConductivitE + Coliformes + t, data = resultc)
par(mfrow = c(2, 2))
plot(model4)
result6   = resultc[-c(54,155,123),]
model31 <-   lm(formula = Chlorophylle ~ Station + Phycocyanine + poly(Ventmax,2)
+ Oxygene + PH + ConductivitE + Coliformes + t, data = result6)
summary(model31)
r1 = re[-c(54,155,123),]
r1['fit'] = model31$fitted.values
r1$date<-as.Date(r1$Date,origin='1900-1-1')-2
data = r1[,c(23,3,6,22)]
names(data) = c("Date","Station","Chlorophylle","fit")
TS <- ggplot(data) +
  geom_line(aes(x = Date, y = fit,color = "red")) +
   geom_line(aes(x = Date, y = Chlorophylle))+
  facet_wrap(~Station, ncol = 5) +
  xlab("Date") +
  ylab("") +
  theme(panel.spacing = unit(0.1, "lines"))+
   theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
panel.background = element_blank(), axis.line = element_line(colour = "black")) +
theme(legend.position="none")
print(TS)
```

4 **PCChl**

1) Without

```
resultd    = result[,c(1,5,6:19)]
fit <- lm(PCChl ~., data =    resultd)
step.fit <- stepAIC(fit,direction = "both",trace = F)
summary(step.fit)
model5 <- lm(formula = PCChl ~ Station + Transparence + Temp + Saturation +
        Oxygene + PH + TDS + Coliformes + t, data = resultd)
par(mfrow = c(2, 2))
plot(model5)
result5    = resultd[-c(54,88,123),]
model41 <- lm(formula = PCChl ~ Station + Transparence+ PH + TDS + Coliformes
+ t, data = result5)
summary(model41)
r1 = re[-c(54,88,123),]
r1['fit'] = model41$fitted.values
r1$date<-as.Date(r1$Date,origin='1900-1-1')-2
data = r1[,c(23,3,7,22)]
names(data) = c("Date","Station","PCChl","fit")
TS <- ggplot(data) +
   geom_line(aes(x = Date, y = fit,color = "red")) +
     geom_line(aes(x = Date, y = PCChl))+
   facet_wrap(~Station, ncol = 5) +
   xlab("Date") +
   ylab("") +
   theme(panel.spacing = unit(0.1, "lines"))+
   theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
panel.background = element_blank(), axis.line = element_line(colour = "black")) +
theme(legend.position="none")
```

```
print(TS)
```

2) Including

```
resultd    = result[,c(1:3,5,6:19)]
fit <- lm(PCChl ~., data =    resultd)
step.fit <- stepAIC(fit,direction = "both",trace = F)
summary(step.fit)
model5 <- lm(formula = PCChl ~ Station + Phycocyanine + Ventmax + Oxygene +
        PH + ConductivitE + Coliformes + t, data = resultd)
par(mfrow = c(2, 2))
plot(model5)
result5    = resultd[-c(54,155,123),]
model41 <- lm(formula = PCChl ~ Station + Phycocyanine + poly(Ventmax,2) +
Oxygene + PH + ConductivitE + Coliformes + t, data = result5)
summary(model41)
r1 = re[-c(54,155,123),]
r1['fit'] = model41$fitted.values
r1$date<-as.Date(r1$Date,origin='1900-1-1')-2
data = r1[,c(23,3,7,22)]
names(data) = c("Date","Station","PCChl","fit")
TS <- ggplot(data) +
    geom_line(aes(x = Date, y = fit,color = "red")) +
      geom_line(aes(x = Date, y = PCChl))+
    facet_wrap(~Station, ncol = 5) +
    xlab("Date") +
    ylab("") +
    theme(panel.spacing = unit(0.1, "lines"))+
    theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
panel.background = element_blank(), axis.line = element_line(colour = "black")) +
theme(legend.position="none")
print(TS)
```