Water Availability of Water-Springs

1. Introduction

1.1 Background Information

Water level in a waterbody (water spring, lake, river, or aquifer) is dynamic and changes along with the seasons. During fall and winter, water bodies are refilled, but during spring and summer, they start to drain. It is important to forecast the water level in a waterbody (water spring, lake, river, or aquifer) to handle daily consumption. Acea Group, one of the leading Italian multiutility operators in the water services sector, supplying 9 million inhabitants in Lazio, Tuscany, Umbria, Molise, Campania, operates a competition to predict the most efficient water availability, in terms of level and water flow for each day of the year.
The main goal of this project is to predict the hydrometry of the Arno river, the second largest river in peninsular Italy and the main source of water supply of the metropolitan area of Florence-Prato-Pistoia. The availability of water for this waterbody is evaluated by checking the hydrometric level of the river at the section of Nave di Rosano.

1.2 Data

The dataset (River_Arno.csv), containing the Arno river's data for rainfall, temperature, and hydrometry from January 1998 to the end of June 2020, is part of the database from Acea Smart Water Analytics Competition. The raw dataset has 8217 observations. The data for rainfall and temperature don't always go alongside the date and need to be cleaned later.

2. Basic Analysis

2.0 Data Processing and Cleansing

I can observe that the raw dataset includes 17 columns and 8217 rows, containing information about the Arno River. However, there are several problems in the dataset. First of all, there exists a large amount of null values distributed across the dataset. Especially for the

data points which are recorded before 2004, there is no information about rainfall in different regions and temperatures at all. Besides, I can also discover that there are 14 columns recording rainfall in various regions along the Arno River. Unfortunately, the data for rainfall is not complete as well. I can observe that for some days, there are only records for rainfalls in some of those 14 regions.

I first tried to solve the problem of different rainfalls. Our solution is that I take the mean value of the number of rains of all regions recorded for all the dates that there is some data for the rainfall across regions. This means that, for example, if, for one row, there are only 6 data points of precipitation among all 14 regions, I take the mean value of that six data points as the rainfall data I are going to use. After this operation, I were able to combine the 14 columns and provide a new dataframe with 4 columns in total.

Then I dealt with the problem of missingness. By observing the new data frame, I discovered that most of the rows that contain missingness are data points recorded before January 1st, 2004. I chose to drop all rows that contain any null values. As a result, 4942 observations remained in our new dataset.

The last problem I solved is the format of the dates. The dates recorded in the dataset are given in the "day/month/year" format. I changed the order of the day, month, and year and changed the type of this column from characters to date objects. With such transformation, I can plot other columns in the order of time.

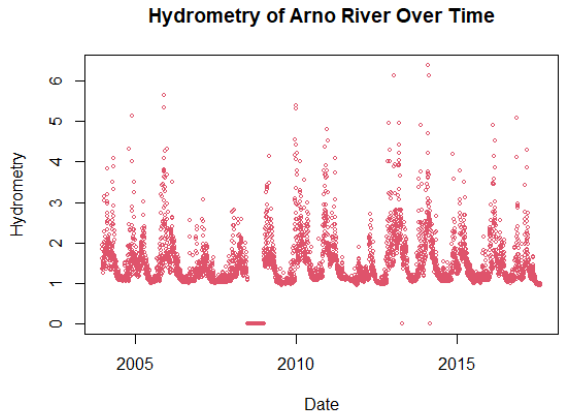
After all the previous data cleansing, the new data frame columns are Date, Temperature, Rainfall, and Hydrometry. Now I are finally ready to proceed with our data analysis.

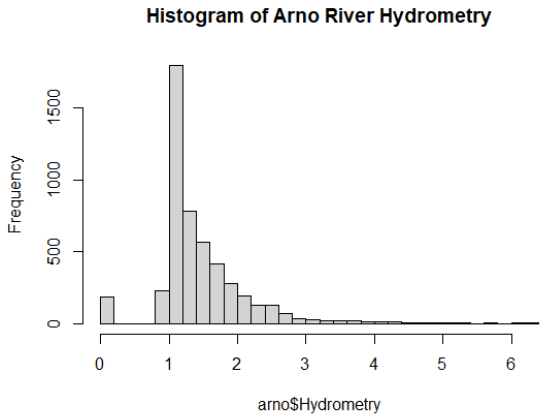
2.1. Distribution of Hydrometry, Temperature, and Rainfall overtime

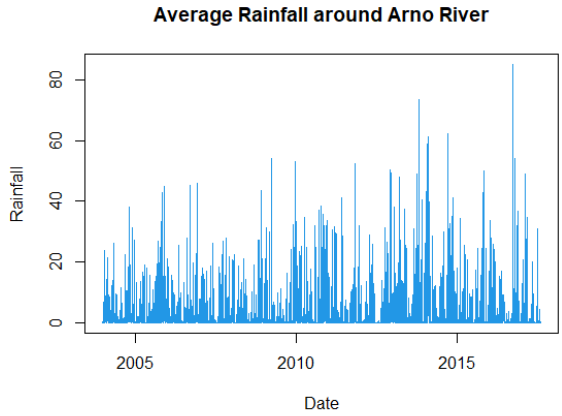
Method

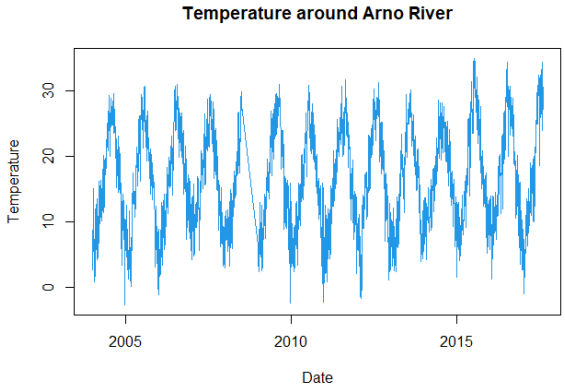
After the data cleaning, I wish to observe the distributions of hydrometry, rainfall, and temperature over time. The most efficient method is using the graphical method. I plotted these three columns with dates as the x-axis. Besides, I also plotted a histogram of the hydrometry to observe the distribution and discover abnormality and outliers.

Analysis









The first thing I observed is that both hydrometry and temperature have obvious periodic characteristics. This is reasonable, especially for the temperature due to the geographical location of the Arno River. Meanwhile, for the average rainfall, I did not observe any obvious trend considering that the column contains many zeros. However, I can still see that the average amount of rainfall seems to be increasing over the years.

One crucial issue I noticed is that in the plot of the hydrometry over time, I observed a period of abnormal zeros during 2008. The distribution of the hydrometry also showed the section of zeros. However, when I looked back at the temperature and rainfall data during that period in the original dataset, there is no obvious difference between this period and other dates. Hence, I can reasonably assume that some other reasons cause such an abnormal period of zero data points. Considering the fact that outliers can significantly influence the prediction model, I decided to remove these data points.

Conclusion

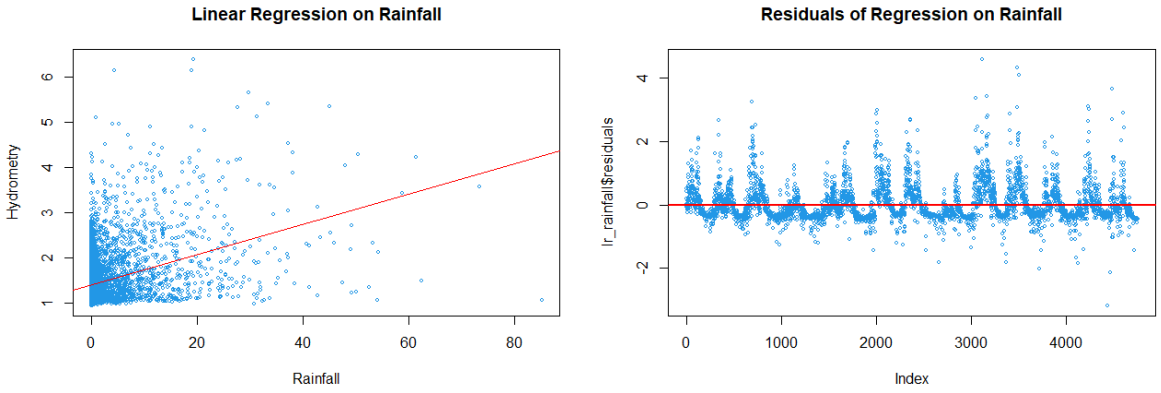
The distribution of hydrometry is roughly periodic over time. Meanwhile, the temperature is also periodic due to natural facts, which together might indicate the correlation between temperature and the measure of hydrometry. Besides, there exists a group of strange outliers of hydrometry that equals zero. After observing the original data, this group of outliers was removed from the dataset for further analysis.

2.2. Linear Regression with One Variable

Method

For this question, I tried to fit a linear regression model with one variable to the dataset. I used hydrometry measurements as our response variable and tried both rainfall and temperature as our independent variables. I also plotted the residual plots for each regression model. Then I calculated the mean squared error and observed the r-squared value to justify the models statistically.

Analysis

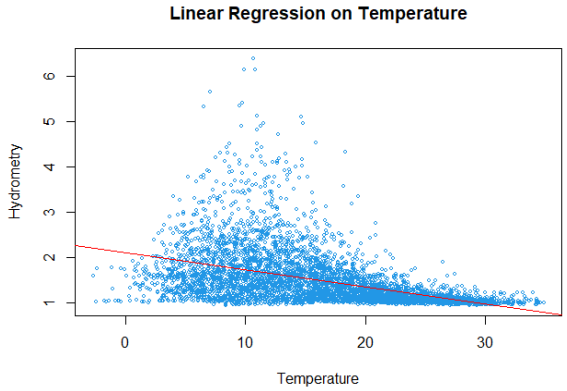


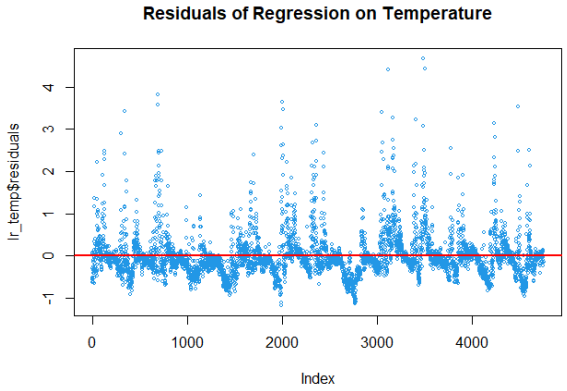
For the linear regression model on rainfall, I have the following statistics:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.402447   0.008823   159.0   <2e-16 ***
Rainfall    0.033541   0.001229    27.3   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5601 on 4754 degrees of freedom
Multiple R-squared:  0.1355,    Adjusted R-squared:  0.1353
F-statistic:   745 on 1 and 4754 DF,  p-value: < 2.2e-16
```

With a mean squared error of 0.3135. By looking at the scatter plot against the linear model, I can see that the linear regression model is not a good fit. The r-squared value also showed that: only 0.1353 of the variance in hydrometry was explained by the rainfall measurement.





For linear regression model on temperature, I have the following statistics:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.109502   0.018691  112.86   <2e-16 ***
Temperature -0.037226   0.001033  -36.03   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5339 on 4754 degrees of freedom
Multiple R-squared:  0.2145,    Adjusted R-squared:  0.2143
F-statistic:  1298 on 1 and 4754 DF,  p-value: < 2.2e-16
```

With a mean squared error of 0.2849. Similar to the rainfall, temperature alone is also not a very good independent variable. I can observe that from both the scatter plot with regression line and the residual plot. However, I can still see that temperature itself seems to be a more efficient dependent variable since more variance in hydrometry can be explained by temperature. The mean squared error is also smaller than the one of the rainfall regression model.

Besides, I can also observe the periodicity of residuals of both regression models. This periodicity is caused by the fact that the data points are sorted by order of time. Such a clear pattern also indicates that a linear regression model with one variable is probably not our best choice.

However, I did notice that for both regression models, the probability of the correlation caused by randomness is extremely small. This means that the two independent variables have consulting value when predicting the hydrometry, but I need a better selection of the model.

Conclusion

Linear regression model with a single variable can explain some variances of the hydrometry measurements of the Arno River. However, this model is not our best choice due to the relatively low r-squared value and unstable residuals.

2.3. Multiple Regression Model Predicting Hydrometry

Method

With previous information, I decided to combine the two independent variables, rainfall, and temperature to perform a multiple regression to predict the hydrometry measurements of the Arno River. Once again, I tried to justify the efficiency of the model using mean squared error and r-squared value as our test statistic. Then, I compared the multiple regression model with previous regression models and observed how much I improved.

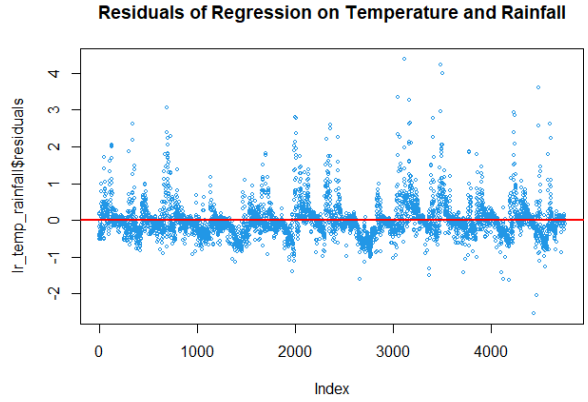
Analysis

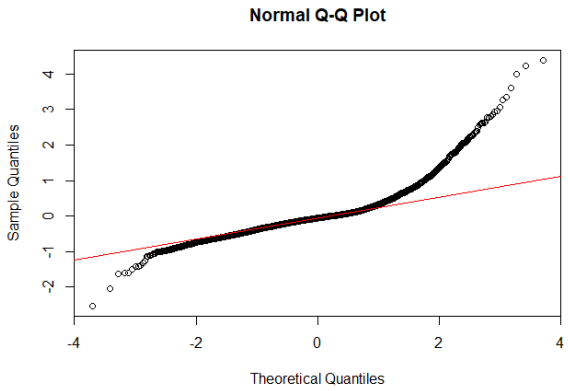
```
Residuals:
    Min      1Q  Median      3Q     Max
-2.5327 -0.2650 -0.0599  0.1315  4.3840

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.9768738  0.0182194  108.50   <2e-16 ***
Rainfall     0.0287229  0.0011048   26.00   <2e-16 ***
Temperature -0.0340653  0.0009745  -34.96   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4996 on 4753 degrees of freedom
Multiple R-squared:  0.3123,    Adjusted R-squared:  0.312
F-statistic:  1079 on 2 and 4753 DF,  p-value: < 2.2e-16
```

Since the multiple regression model is hard to plot, I performed our analysis only based on the statistics of the regression fit and the residual plot. The mean squared error of this regression model is approximately 0.2494. I can easily observe that the mean squared error has decreased significantly compared with the two other models. Meanwhile, the adjusted r-squared value also increased, indicating that this multiple regression model is a better choice than the simple linear regression.





The above plot is the residual plot of the multiple regression. Comparing this plot to the previous residuals, I can observe that the periodic pattern is relatively flattened. However, there are still many abnormally high residuals. This is also reflected in the Q-Q plot. I can see that the residuals are not quite distributed normally. Although the distribution is roughly normal for the residuals around 0, there is a non-negligible amount of data points that have large residuals. This means that the multiple regression model is not performing well when the data has relatively large hydrometry measurements.

Conclusion

Multiple regression is a better choice compared with simple linear regression. However, the residuals indicate that a linear model is still not a good choice due to the inconsistent performance of the model when the hydrometry is large.

2.4. Model Selection - Logistic Transformation

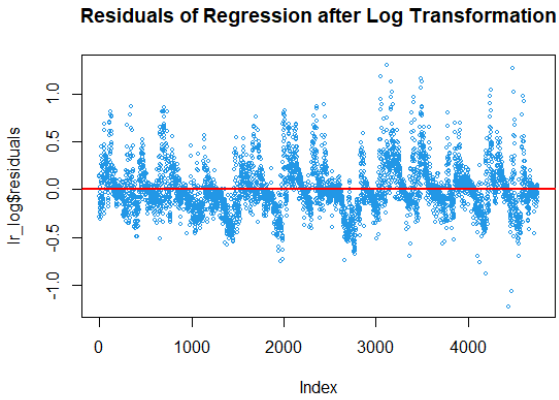
Method

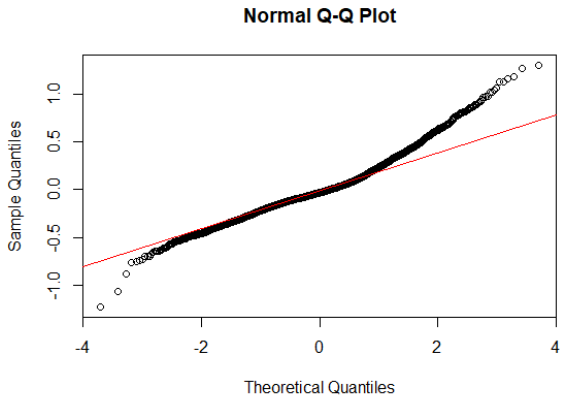
After all the analysis I finished before, I found that I need to make the correlation between hydrometry measurements and other variables more "linear." I noticed in section 2.3 that the regression model performs well when the hydrometry is relatively low but provides larger error as the hydrometry increases. This is the characteristic of an exponential distribution. As a result, I decided to apply the logistic transformation to the hydrometry and perform another multiple regression to the new data.

Analysis

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.6632083  0.0093243   71.13   <2e-16 ***
Rainfall     0.0136530  0.0005654   24.15   <2e-16 ***
Temperature -0.0217026  0.0004987  -43.52   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2557 on 4753 degrees of freedom
Multiple R-squared:  0.3692,    Adjusted R-squared:  0.369
F-statistic:  1391 on 2 and 4753 DF,  p-value: < 2.2e-16
```





From the statistical analysis, I observe that the adjusted r-squared did increase, although not for a significant amount. However, According to the residual plot, I found that the extremely high residuals in previous models seem to be less extreme here. To further verify that, I draw the normal Q-Q plot for the residuals and found that the distribution of the residuals after the log transformation is much closer to the normal distribution than the previous models. This means that the linear regression model is much more convincing when predicting the log of the hydrometry of the Arno River.

Conclusion

The multiple regression model on the log-transformed hydrometry is the most proper model I have found so far. Under this model, I are able to explain 0.369 of the variance in the log of hydrometry measurements of the Arno River.

3. Advanced Analysis

Method

From the very beginning, I have noticed that the hydrometry seems to be periodic each year. As a result, in the advance analysis, I also wish to include the time information as an independent variable to predict the hydrometry of the Arno River. However, since the date is hard to convert to a continuous variable, I decided to categorize the time by month. After that, I used one-hot encoding to turn the categorical variable month into a binary matrix to apply a linear regression model.

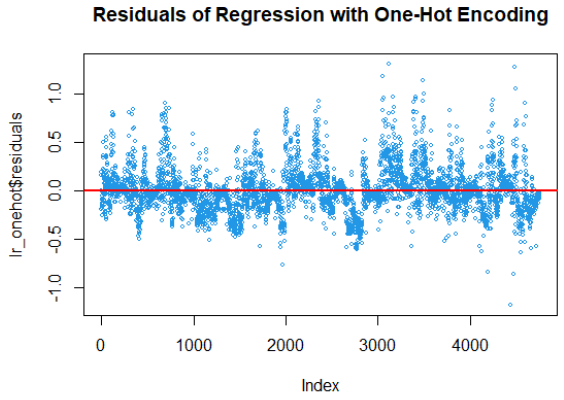
Analysis

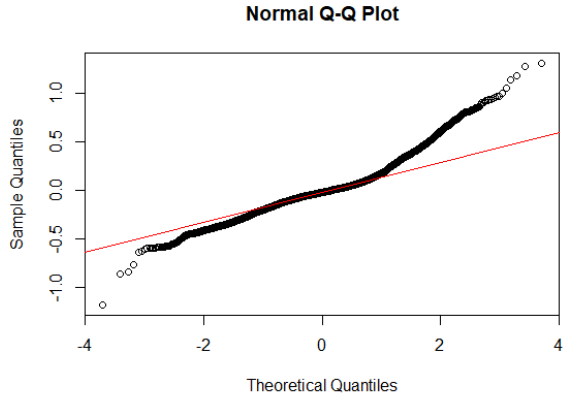
```
Call:
lm(formula = Log_hydrometry ~ Rainfall + Temperature + month_01 +
    month_02 + month_03 + month_04 + month_05 + month_06 + month_07 +
    month_08 + month_09 + month_10 + month_11 + month_12, data = oneHot_arno)

Residuals:
     Min       1Q   Median       3Q      Max
-1.17762 -0.12362 -0.01762  0.08472  1.31103

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.4819755  0.0150669  31.989  < 2e-16 ***
Rainfall     0.0142215  0.0005183  27.440  < 2e-16 ***
Temperature -0.0072643  0.0010702  -6.788 1.28e-11 ***
month_01     0.0327869  0.0165088   1.986 0.047088 *
month_02     0.1376530  0.0168575   8.166 4.06e-16 ***
month_03     0.1784214  0.0168756  10.573  < 2e-16 ***
month_04     0.0641707  0.0184836   3.472 0.000522 ***
month_05    -0.0161533  0.0204048  -0.792 0.428608
month_06    -0.1449979  0.0238361  -6.083 1.27e-09 ***
month_07    -0.2390203  0.0262675  -9.099  < 2e-16 ***
month_08    -0.2514351  0.0253533  -9.917  < 2e-16 ***
month_09    -0.2654054  0.0225777 -11.755  < 2e-16 ***
month_10    -0.2313766  0.0195218 -11.852  < 2e-16 ***
month_11    -0.0279763  0.0178279  -1.569 0.116659
month_12            NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2332 on 4742 degrees of freedom
Multiple R-squared:  0.4767,    Adjusted R-squared:  0.4752
F-statistic: 332.2 on 13 and 4742 DF,  p-value: < 2.2e-16
```





The mean squared error of this model is 0.054, which decreased from the mean squared error of the model in section 2.4, 0.065. Note that this MSE is not comparable to the MSEs in sections 2.2 and 2.3 since I applied log transformation to the hydrometry. I can also observe that the adjusted r-squared value of the regression model increased from 0.369 in section 2.4 to 0.475 here. If I look into the probability of each variable explained by randomness, I can see that most of them are extremely small, except for the indicator of May. I might need more professional information on the Arno River to explore the reason behind such a phenomenon. However, despite that small factor, the time information overall is strongly correlated to the hydrometry of the Arno River. All this statistical information shows that including the time information can increase the efficiency of our model significantly. However, I should also notice that the Q-Q plot of this regression indicates that the residuals of this regression model are less normal than the regression model without time information.

Conclusion

Including time information can increase the efficiency of the regression model significantly. However, the normality of the residuals is decreased due to the high-hydrometry data points.

4. Discussion and Conclusion

From the graphical analysis, I found that both hydrometry and temperature have obvious periodic characteristics, which together might indicate the correlation between temperature and the measure of hydrometry. After removing the outliers of hydrometry, the residual plots for hydrometry measurements with the independent variable of rainfall and temperature respectively show that the linear regression model with a single variable can explain some variances of the hydrometry measurements but not the best choice due to the relatively low r-squared value and unstable residuals. Combining two independent variables, I found that multiple regression is better compared with simple linear regression. However, there still exists inconsistent performance of the model when the hydrometry is large. Furthermore, I made a logistic transformation to the hydrometry. I performed another multiple regression to the transformed data. The multiple regression model on the log-transformed hydrometry turns out to be the most proper model I have found so far. In the advanced analysis, I included the time information as an independent variable to predict the hydrometry, and the included time information indeed increased the efficiency of the regression model significantly. However, the one-hot encoding of the month might be collinear with temperature since dates and months also influence the temperature.

For further improvement, I could verify the existence of potential overfitting of the training data by separating the data into training and testing groups multiple times. Besides, I can also do cross-validation to testify our model.