Report

Name: Tianlang Tan
Id: 20028268

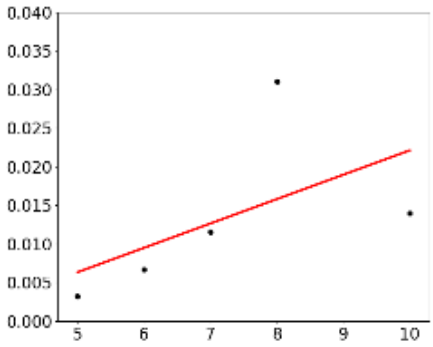
This report compares and analyses different methods in task1 and task2 based on their corresponding performance.

Task1:

This section compares mean value completion and polynomial regression completion.

The first method used mean value in continuous three months in one year to complete the data. One of the advantages of this method is that it considers the survival time of chlorophyll. Basically, the survival time of chlorophyll will not last longer than 3 months. So, this complete method is reliable. Also, this method is easy to implement with qualified data. However, one significant drawback is that based on this method, some of the missing data will be calculated as negative numbers. Although they were replaced by 0, it is impossible for a lake that the chlorophyll and total P are zero (temperature can be negative) which may affect the authenticity of the data. Although there are zero data for this completing method, the number of the zero numbers is relatively small which will not have great impact on the data. Hence, this method is effective.

The second method used polynomial regression to predict the regression function. After that, the missing data are completed with the predicted value calculated from the regression function. The advantage of this method is that it tried to find the hidden pattern behind the data and used the pattern to fit in the missing data. In most of the years, the regression fit the data well without overfitting. The degrees of the polynomial functions were chosen by comparing the figure of different degree. However, in some of the years such as the figure shown below, the data distribution is dispersed so that the regression effect is not good enough. To summarize, although some of the year is not suitable for polynomial regression, polynomial regression method to complete the missing data is good enough in most of the years.



After comparison, if the data is not very dispersed, polynomial regression method gives better completion. If the data is dispersed, mean value method is more effective.

Task2

The data after mean completion are used to calculate the correlation between CHLA and Temperature & Total P. The rankings are also provided.

1. **Covariance**

   Covariance measures the degree of the linear relationship between variables. A large positive value indicates positively correlated data. The absolute magnitude of the covariance measures the degree of redundancy. Here is the result:

   Temperature (0.01476022097378277)

   TotalP (1.2624319600499372e-05)

   It seems that Temperature has higher relationship with CHLA. However, because the variable is not in same range so that the magnitude of the variable largely affects the covariance. This correlation is not objective enough.

2. **Pearson correlation coefficient**

   Pearson correlation coefficient is an updated version that eliminates the disadvantage of Covariance. To reduce the effect caused by the magnitude of the variable. Pearson correlation coefficient is the result that divides the covariance of two variables by the product of the corresponding standard deviation. Here is the result:

   Temperature (0.42840534669449903)

   TotalP (0.38783071711886474)

   Temperature have higher correlation with CHLA. This method also has its drawbacks. It only finds the linear relationship between variables. Also, the data should be normal distribution.

3. **Spearman Correlation Coefficient**

   Spearman Correlation Coefficient measure of rank correlation between the variable. Spearman correlation coefficient is appropriate for both continuous and discrete ordinal variables. Here is the result:

   Temperature (0.4351123352974459)

   TotalP (0.496517180869418)

   Total P has higher correlation with CHLA

4. **Distance correlation**

   Distance correlation is a measure of dependence between two paired random vectors of arbitrary. Distance correlation can measure both linear and nonlinear association between two random variables or random vectors. Here is the result:

   Temperature (0.571594653305501)

   TotalP (0.6121692828811354)

   TotalP has higher correlation with CHLA

5. **Maximal information coefficient (MIC)**

   Maximal information coefficient is a robust method for data correlation calculation. It normalizes mutual information into a range from 0 to 1 to represent the importance rank for variable. Here is the result:

Temperature (0.3673695513789908)

TotalP (0.3052521179120473)

Temperature has higher correlation with CHLA. MIC can measure the correlation no matter the variable is linear or non-linear related.

To summarize, according to the 5 methods, TotalP have higher correlation with CHLA. Covariance and Pearson correlation coefficient may not be very suitable for this project because the magnitude of the variable largely affects the covariance and Pearson correlation coefficient is only suitable for linear relationship. The rest 3 methods can be applied to measure the correlation between CHLA and Temperature & Total P