Description

Name: Tianlang Tan
Id: 20028268

Introduction

China lake was selected as the data in this coursework. The following section will discuss the procedural of task1 and task2. All the operation is done by coding without the usage of Excel.

Language and Library

Python 3.7
openpyxl 3.0.5
matplotlib 3.2.2
scipy 1.5.0
minepy 1.2.5
sklearn 0.0

Preprocessing

Firstly, the China lake file was read by using openpyxl library and found out the overlapped month and year. After that, 3 different 3-dimensional lists were constructed to store the Chla, Temperature and TotalP. The first dimension represented a year, the second dimension represented a month, and the third dimension represented all the records in one month. Then, data cleaning operation was applied that removes the data whose value is empty or the year is not in overlapped years or month exceed May-October or the depth is not the majority one. Next, in order to achieve only one data per month, average operation was applied to all 3 lists (Chla, Temperature and TotalP). After that, to prioritize best match, three new lists for corresponding best match values were created. The values of the best match lists also applied average operations, so the value of best match is averaged. If the best match also had multiple data for one month, the average is calculated in one day first then one month (E.g.: In 1999, there are 8/4, 8/17, 8/30, 8/30 four data for August, the average for August in 1999 is: mean (8/4, 8/17, mean (8/30))) Then, the previous 3 lists (Chla, Temperature and TotalP) are compared with the best match lists. If the data is different, the update the best match value to the previous lists. Finally, the following two methods are the implementations to complete the data.

Task1-method1: mean value

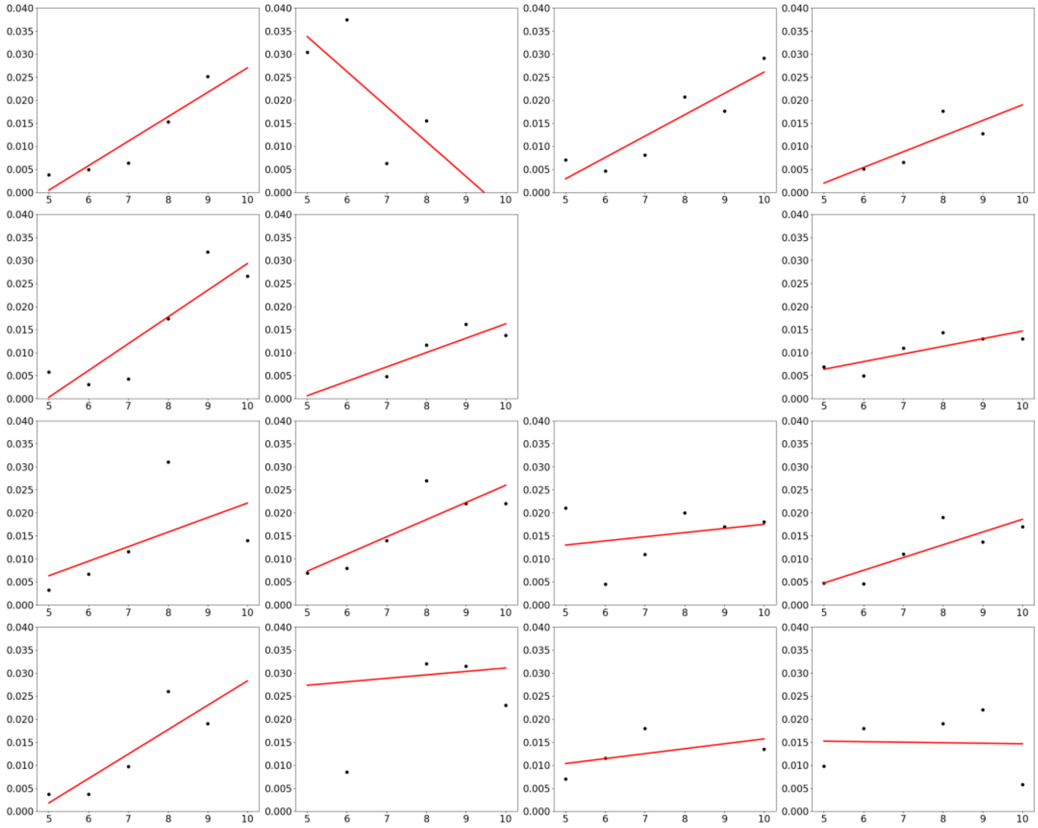
The following pseudo-code represents the mean value method to complete missing data, the basic idea is to keep move a box contain 3 continuous months in one year from 5,6,7 to 8,9,10 so that the data can only be completed within 3 continuous months. If the box is **x0y**, then compute all the missing data using the mean value of the adjacent months. If the box is **0xy** or **yx0**, compute one missing data using mean value (miss = 2x-y). Using this algorithm, **x0y** condition has higher priority so that using more months to computer the mean value.

# **x0y** There are one month data before and after the missing month data
# **0xy** There are two continuous months data after the missing month data
# **yx0** There are two continuous months data before the missing month data

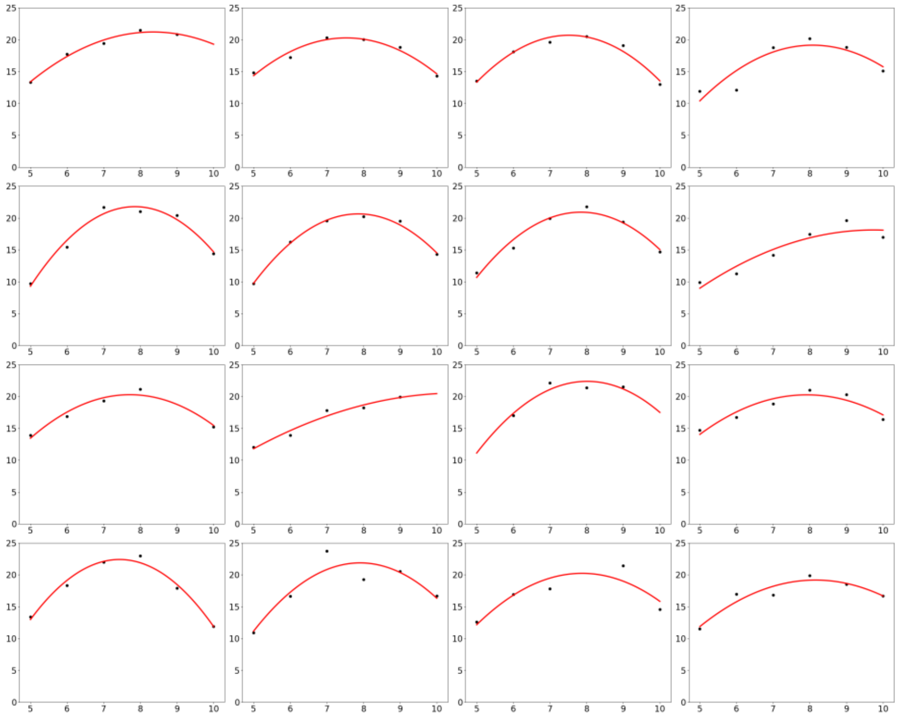
```
For each year
    While (the data for all the month in this year still need to be completed)
        If (the year has x0y condition)
            Complete all the missing data
        If (the year has 0xy or yx0 condition)
            Complete one the missing data
```

Task1-method2: polynomial regression

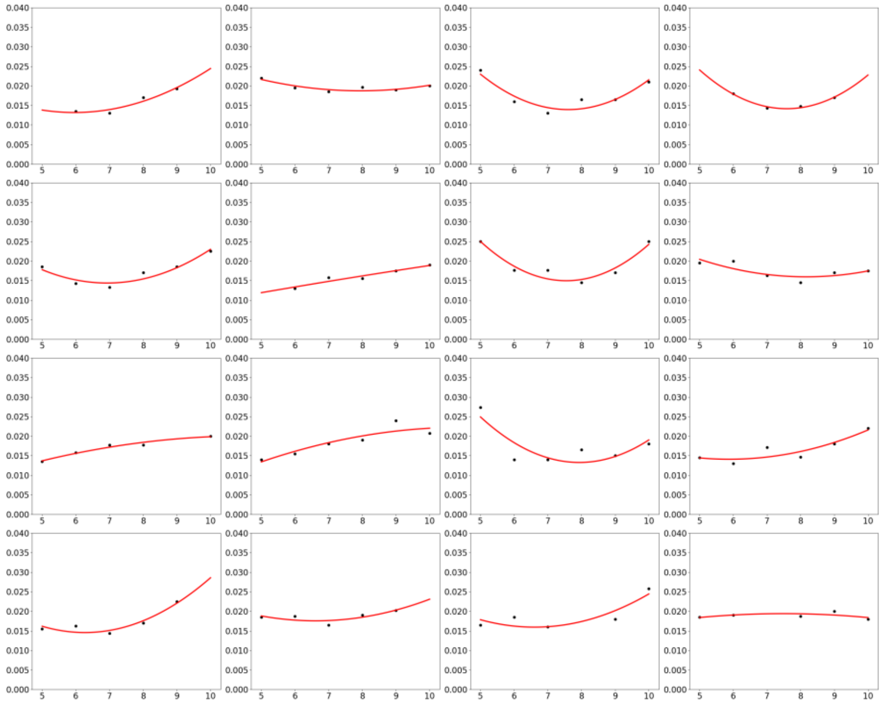
Using polynomial regression to complete the missing data. The input of the polynomial regression is the existing data within a year. There are 15 models for each value (2004 did not have data in the common depth 7). After tuning the degree of the polynomial function, considering overfitting and underfitting condition, Chla, Temperature, TotalP used 1 degree, 2 degree and 2 degree, respectively. The following figure shows the polynomial regression result.



Chla: Degree = 1



Temperature: Degree = 2



TotalP: Degree = 2

After generating the function, the missing data were completed by feeding into the polynomial function and computer the result.

Task2:

Here are the five methods implemented to calculate the correlation between CHLA and Temperature & Total P. For each method, if the result is greater, the associated factor is more important.

1. **Covariance**
   Use np.cov() to calculate the correlation between CHLA and Temperature & TotalP
2. **Pearson correlation coefficient**
   Use pearsonr() from scipy.stats to calculate the correlation between CHLA and Temperature & TotalP
3. **Spearman correlation coefficient**
   Use spearmanr() from scipy.stats to calculate the correlation between CHLA and Temperature & TotalP
4. **Distance correlation**
   Use correlation() from scipy.spatial.distance to calculate the correlation between CHLA and Temperature & TotalP
5. **Maximal information coefficient (MIC)**
   Use MINE() from minepy to calculate the correlation between CHLA and Temperature & TotalP