**Description**

**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.

**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 remove the data whose value is empty or the year is in overlapped period or month exceed May-October. Next, in order to achieve only one data per month, average operation was applied to all 3 lists (Chla, Temperature and TotalP). The following

**Task1-method1: mean value**

The following pseudo code represent 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. If the box is 101, then compute all the missing data using the mean value of the linked months. If the box is 011 or 110, compute one missing data using mean value. Using this algorithm, 101 condition have higher priority so that using more months to computer the mean value.

# 101 represents the case that the missing data has data one month before and after

# 011 represents the case that the missing data has data two months after

# 110 represents the case that the missing data has data two months before

**1** For each year

**2** While (the data for all the month in this year still need to be completed)

**3** If (the year has 101 condition)

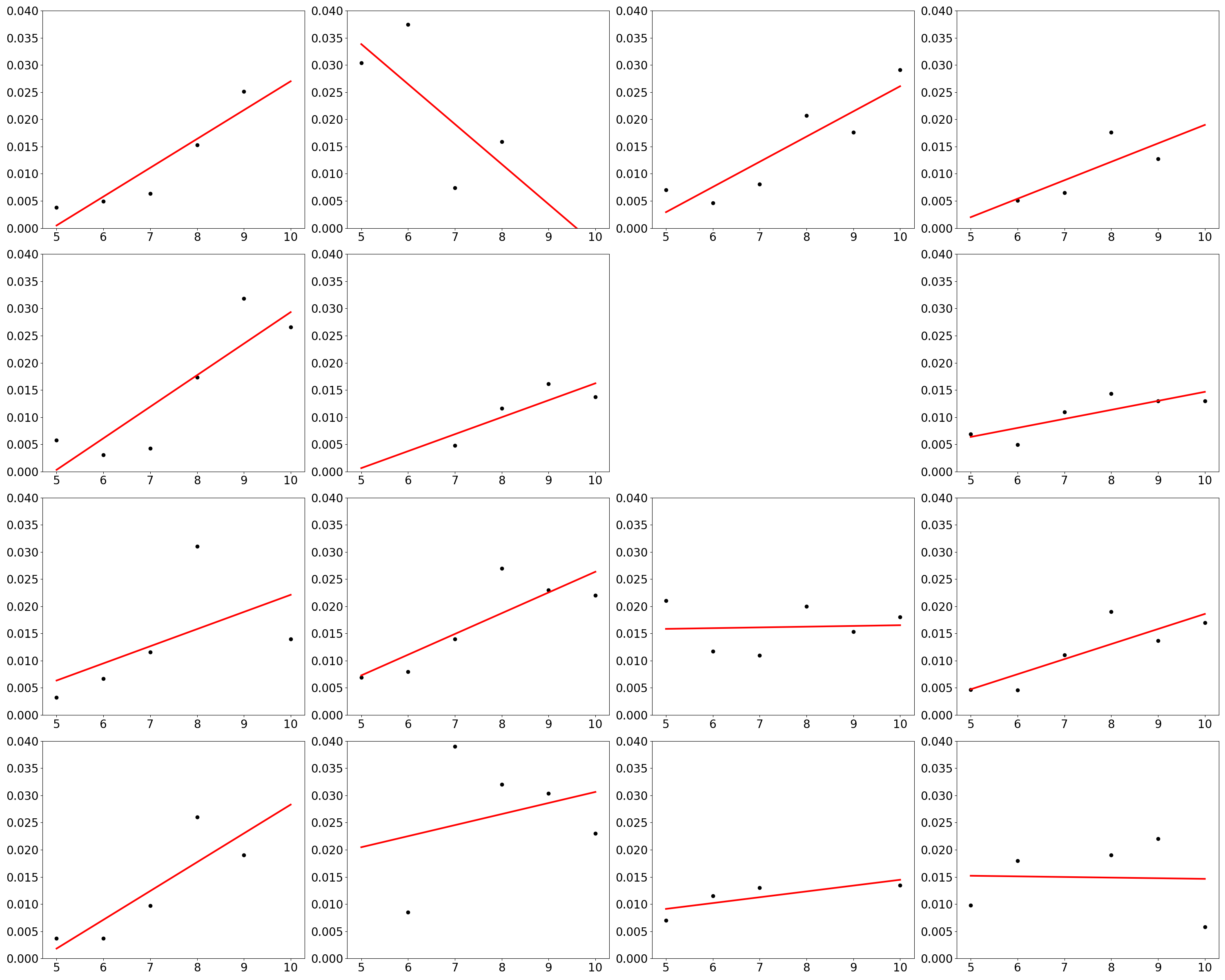
**4** Complete **all** the missing data

**5** Else if (the year has 011 or 110 condition)

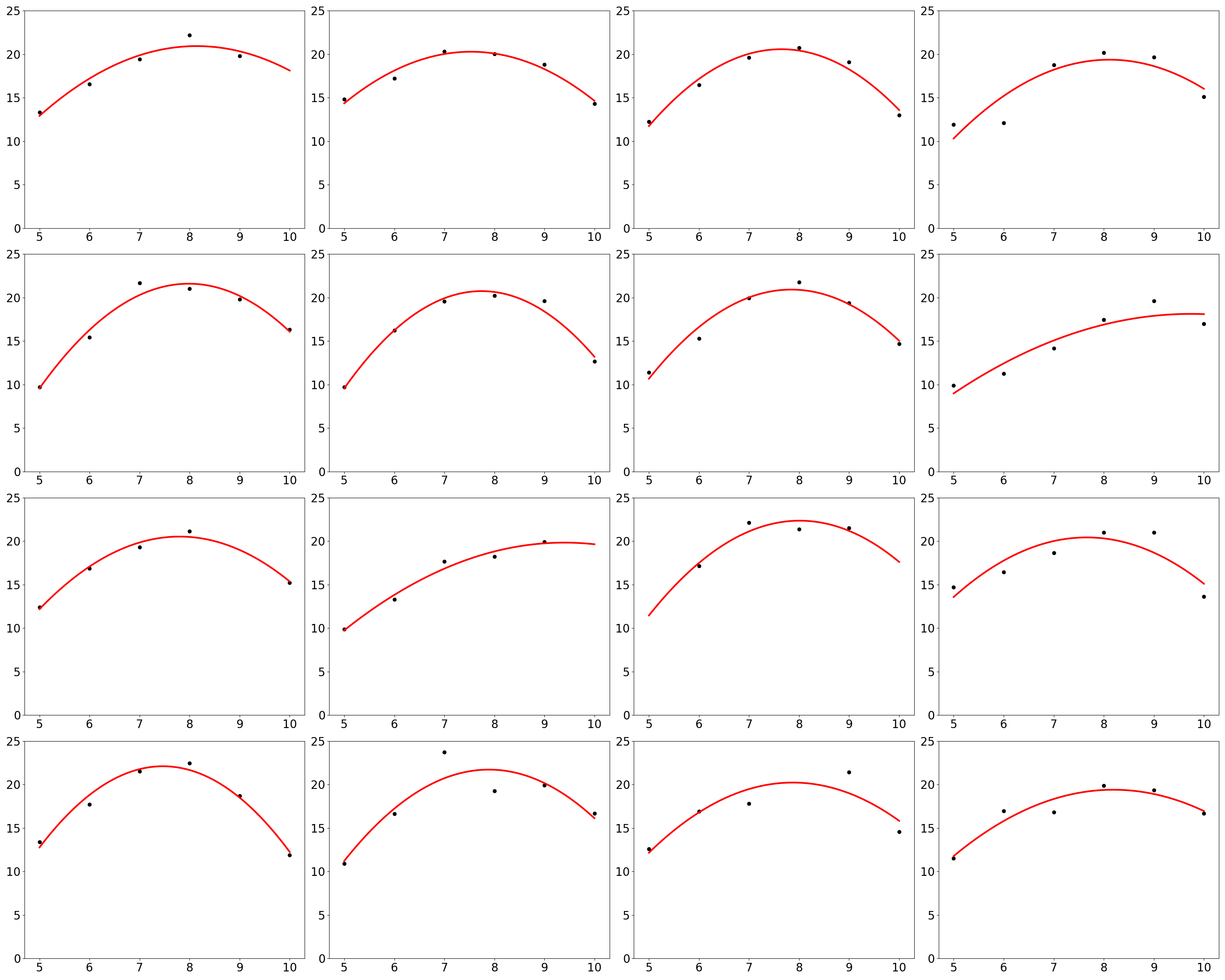
**6** Complete **one** the missing data

**Task1-method2: polynomial regression**

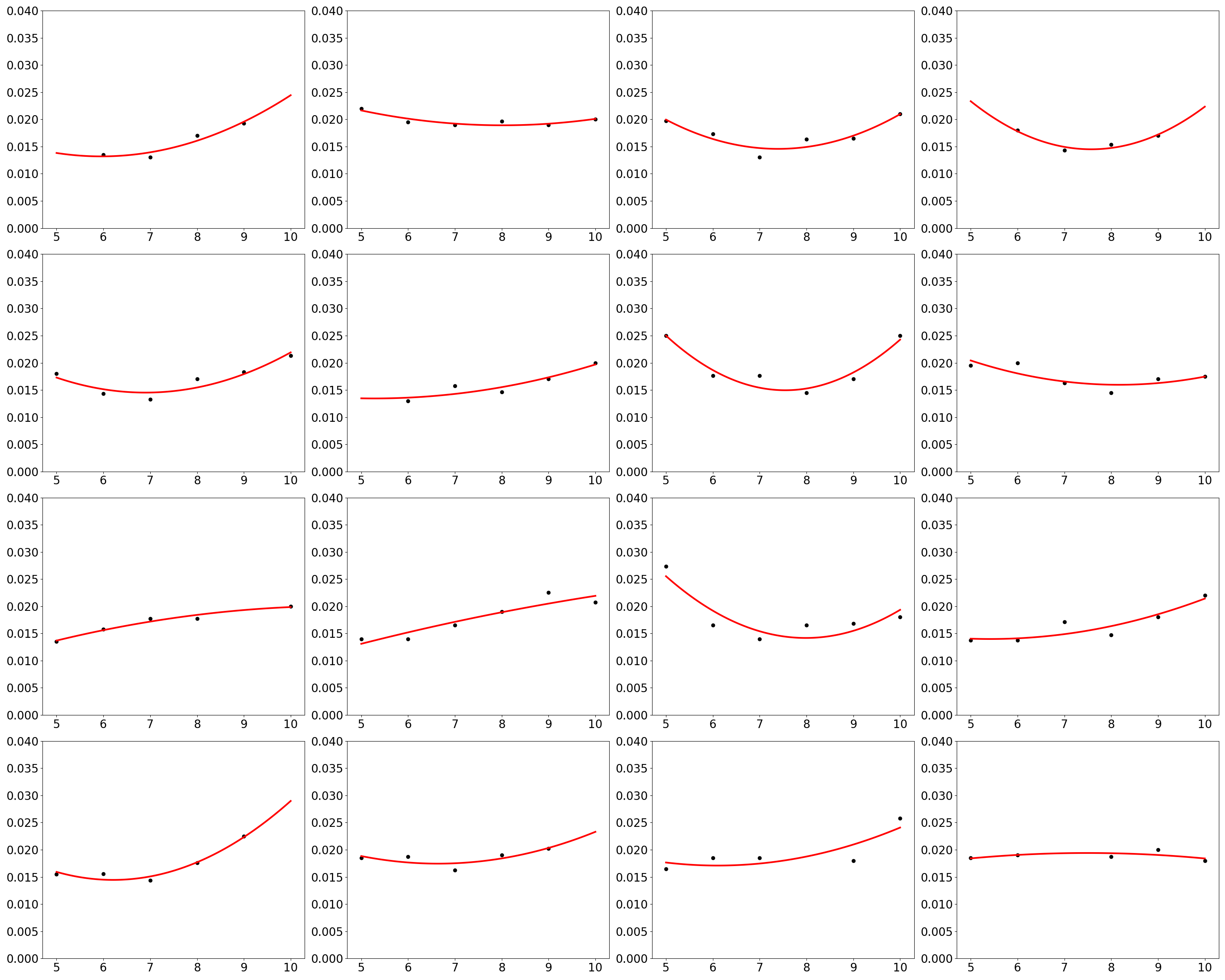
Using polynomial regression to complete the missing data. The input of the polynomial regression is the existing data within a year. There are 15 models for each value (2004 did not have data in the common depth 7). After tuning the degree of the polynomial function, considering overfitting and underfitting 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:**