COMP3055 Machine Learning
Coursework Report

Chen Liao 20030694

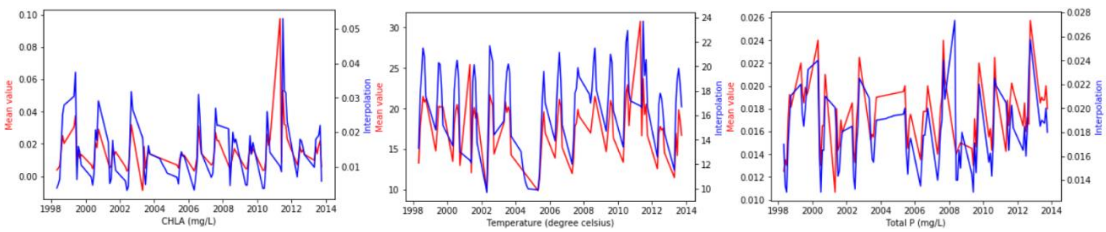
1. Introduction

The data chosen in this coursework comes from the China Lake. Rudimentary data preprocessing is leveraged with Pandas library. Firstly, by filtering data with their features, we found that most data are collected at depth of 7, station 1. Therefore, data does not match these criteria is deprecated. Secondly, after filtering, valid data of CHLA, Temperature, and Total P apparently overlaps in the period from 1998 to 2013. Therefore, only data within this period is kept for further analysis.

2. Filling in missing data

Two methods are deployed to deal with missing data, one is the mean value method, the other is linear interpolation. The basic idea of the mean value method is to use the average of the former two values as an estimation of the third. Therefore, a relationship among these three values is established and we could thus derive any one of them if the other two are known. The second method involves linear interpolation which regards all missing values between two valid data points sit exactly on the shortest line strip defined by the two pints.

Priorities and drawbacks intertwine for both methods. The mean value method involves human understanding of the data. For example, water resource data alters periodically and the most possible values for missing data of a month comes from the former two months of its nearest. However, this also causes data waste because at least two complete tuples each year are necessary to derive a full dataset. People cannot use data from other years to infer data this year. On contrary, linear interpolation has no requirement for minimum number of complete tuples, this maximizes the utilization of the known data. However, as a pure mathematical strategy, linear interpolation deprecates human intuition but only relies on relationships among values. Figure 1 delivers a more intuitive comparison between the plots of two methods.




```
Mean value method average: 0.015
Mean value method range: 0.106
Mean value method standard deviation: 0.013
Interpolation method average: 0.015
Interpolation method range: 0.05
Interpolation method standard deviation: 0.009

Mean value method average: 17.918
Mean value method range: 21.05
Mean value method standard deviation: 3.496
Interpolation method average: 17.002
Interpolation method range: 14.0
Interpolation method standard deviation: 3.301

Mean value method average: 0.017
Mean value method range: 0.015
Mean value method standard deviation: 0.003
Interpolation method average: 0.018
Interpolation method range: 0.014
Interpolation method standard deviation: 0.003
```

Figure 1: plots and statistical data for two methods of three features

As shown in the figure and the statistical data, both methods deliver similar average values, maintaining the bias of data at similar magnitude. However, the range and the standard deviation of mean value method is generally greater than of interpolation method, which means the former method generates a sparser data within a broader range. This may be because linear interpolation restricts possible values of missing data into the range of non-missing data based on its mathematical nature, whereas the mean value method does not have this constraint.

3. Correlations

In this section five different correlation measurements are introduced, compared, and analyzed based on their math nature, scope of use, and performance in experiments. For a concise overview, we refer readers to table 1.

	Scope of Use	Standarization	Complexity	Robustness
Pearson	linear	Yes	Low	Low
Spearman	linear, simple non-linear	Yes	Low	Medium
Kendall	linear, simple non-linear	Yes	Low	Medium
Biweighted Midcorrelation	linear, non-linear	Yes	Medium	Medium
Maximal Information Coefficient	linear, non-linear	Yes	Low	High

Table 1: Overview of five methods on some primary characteristics

3.1. Pearson correlation factor

Mathematically, Pearson correlation factor (PCF) is the quotient between the covariance of two features and their standard deviation (see equation 1). Figure 2 demonstrates correlation matrices and their rankings of relevance. Comparing to traditional cosine similarity measurement, PCF centers the data by subtracting its average before measurement. Therefore, it generally delivers a more reliable correlation result. However, this method does not suit non-continuous features or sparse features. It only performs well when the data distribution resembles Gaussian normal distribution.

$$P(x,y) = \frac{cov(x,y)}{\sigma x \sigma y} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \quad \textbf{(Equation 1)}$$

Pearson correlation matrix:

	CHLA	Temperature	Total P		
CHLA	1.000000	0.516167	0.217029	CHLA	1.0
Temperature	0.516167	1.000000	-0.242729	Temperature	2.0
Total P	0.217029	-0.242729	1.000000	Total P	3.0

Pearson correlation matrix:

	CHLA	Temperature	Total P		
CHLA	1.000000	0.416073	0.35509	CHLA	1.0
Temperature	0.416073	1.000000	-0.23112	Temperature	2.0
Total P	0.355090	-0.231120	1.00000	Total P	3.0

Figure 2: Pearson correlation matrices for mean value method (upper-left), for interpolation method (bottom-left), and their corresponding correlation rankings (upper-right and bottom right).

3.2. Spearman correlation factor

Spearman correlation factor (SCF) is a non-parametric measure of correlation focusing on the ranks of data (see equation 2). Figure 3 demonstrates correlation matrices and their rankings of relevance. the SCF between two variables equals the PCF between the rank scores of those two variables. However, PCF assesses linear relationships, whereas SCF assesses monotonic relationships.

$$S = 1 - \frac{n\sum_{i=1}^{n}(rank_{1i} - rank_{2i})}{n(n-1)} \quad \textbf{(Equation 2)}$$

Spearman correlation matrix:

	CHLA	Temperature	Total P		
CHLA	1.000000	0.419223	0.491585	CHLA	1.0
Temperature	0.419223	1.000000	-0.173602	Temperature	3.0
Total P	0.491585	-0.173602	1.000000	Total P	2.0

Spearman correlation matrix:

	CHLA	Temperature	Total P		
CHLA	1.000000	0.411919	0.459396	CHLA	1.0
Temperature	0.411919	1.000000	-0.289360	Temperature	3.0
Total P	0.459396	-0.289360	1.000000	Total P	2.0

Figure 3: Spearman correlation matrices for mean value method (upper-left), for interpolation method (bottom-left), and their corresponding correlation rankings (upper-right and bottom right).

3.3. Kendall correlation factor

Kendall factor (KCF) is defined as the quotient between the difference in number of concordant pairs and discordant pairs and the number of total pairs (see equation 3). Figure 4 demonstrates correlation matrices and their rankings of relevance. Although sophisticated and complex, KCF is normally preferred to SCF because it increases robustness and efficiency with a smaller gross error sensitivity and a smaller asymptotic variance.

$$K = \frac{P - (\frac{n(n-1)}{2} - P)}{\frac{n(n-1)}{2}} = \frac{4P}{n(n-1)} - 1 \quad \textbf{(Equation 3)}$$

Kendall correlation matrix:

	CHLA	Temperature	Total P
CHLA	1.000000	0.293359	0.333799
Temperature	0.293359	1.000000	-0.122243
Total P	0.333799	-0.122243	1.000000

CHLA	1.0
Temperature	3.0
Total P	2.0

Kendall correlation matrix:

	CHLA	Temperature	Total P
CHLA	1.000000	0.286908	0.308648
Temperature	0.286908	1.000000	-0.199053
Total P	0.308648	-0.199053	1.000000

CHLA	1.0
Temperature	3.0
Total P	2.0

Figure 4: Kendall correlation matrices for mean value method (upper-left), for interpolation method (bottom-left), and their corresponding correlation rankings (upper-right and bottom right).

3.4 Biweight midcorrelation

Different from PCF, biweight midcorrelation (BM) leverages the median rather than the mean to evaluate correlations (see equation 4). Therefore, it is less sensitive to outliers, resulting in a more robust correlation result (see figure 5).

$$bicor(x,y) = \sum_{i=1}^{m} \tilde{x}_i \tilde{y}_i \quad \textbf{(Equation 4)}$$

$$\tilde{x}_i = \frac{(x_i - med(x)) w_i^{(x)}}{\sqrt{\sum_{j=1}^{m} \left[\left(x_j - med(x)\right) w_j^{(x)} \right]^2}} \quad \textbf{(Equation 5)}$$

$$\tilde{y}_i = \frac{(y_i - med(y)) w_i^{(y)}}{\sqrt{\sum_{j=1}^{m} [(y_j - med(y)) w_j^{(y)}]^2}} \quad \textbf{(Equation 6)}$$

Biweighted midcorrelation matrix:

	CHLA	Temperature	Total P
CHLA	1.0	0.347632	0.504578

Biweighted midcorrelation ranking:

CHLA	1.0
Temperature	3.0
Total P	2.0

Name: CHLA, dtype: float64

Biweighted midcorrelation matrix:

	CHLA	Temperature	Total P
CHLA	1.0	0.373172	0.451775

Biweighted midcorrelation ranking:

CHLA	1.0
Temperature	3.0
Total P	2.0

Name: CHLA, dtype: float64

Figure 5: Biweighted midcorrelation matrices for mean value method (upper-left), for interpolation method (bottom-left), and their corresponding correlation rankings (upper-right and bottom right).

3.5. Maximal information coefficient

Mutual information measures to what extent a variable encrypts another. Maximal information coefficient (MIC) leverages joint probabilities of mutual information of two variables to evaluate the correlation between them (see equation 5), which is quite different from other methods. Figure 6 demonstrates correlation matrices and their rankings of relevance.

$$MIC[x;y] = max_{|X||Y|<B} \frac{I[X;Y]}{\log_2(\min(|X|,|Y|))} \quad \text{(Equation 5)}$$

Maximal information coefficient matrix:

	CHLA	Temperature	Total P
CHLA	1.0	0.349548	0.477061

Maximal information coefficient ranking:

CHLA	1.0
Temperature	3.0
Total P	2.0

Name: CHLA, dtype: float64

Maximal information coefficient matrix:

	CHLA	Temperature	Total P
CHLA	1.0	0.354342	0.39469

Maximal information coefficient ranking:

CHLA	1.0
Temperature	3.0
Total P	2.0

Name: CHLA, dtype: float64

Figure 6: Maximal information coefficient matrices for mean value method (upper-left), for interpolation method (bottom-left), and their corresponding correlation rankings (upper-right and bottom right).

3.6. Analysis and comparison

As shown in the data, all measurements indicate positive correlations between CHLA and the other two factors. Most of the measurements assign Total P with a higher correlation value, ranking it as a more relevant factor with CHLA, but only the ranking of PCF contrasts with other correlation measurements, indicating Temperature is more relevant to the accumulation of CHLA. This is because the distribution of data does not resemble single-peak normal distribution, which undermines the validity of PCF. For the degree of relevance, most measurements lands in the range of medium extent ($0.4 \leq \rho \leq 0.6\ or\ -0.6 \leq \rho \leq -0.4$) for both variables. However, the results of KCF drops significantly into a weak correlation range ($0.2 \leq \rho \leq 0.4\ or\ -0.4 \leq \rho \leq -0.2$) for both factors. This may be because the number of arbitrary pairs which are concluded as concordant pairs are low, as the complex nature of the data. To draw a conclusion, both Temperature and Total P have a positive correlation with CHLA with a medium intensity in this dataset. Although only nuances exist, Total P still ranks as the most relevant factor with CHLA.

(Word count: 834)