Author: Do Thanh Dat LE, Gan WANG
Binomial's number: 18
Dataset's name: Ocean

I. Datasets preparation

There are some changes compared to the first report. First, about the time-series, we aim to equalize the distance between all periods. The formula to transform the "5days" into "sin.5days" and "cos.5day" which are defined as below:

$$sin.5days = \sin\left(2\pi\left(\frac{5days}{73}\right)\right)$$

$$cos.5days = \cos\left(2\pi\left(\frac{5days}{73}\right)\right)$$

We know that the raw data is the observations in 9 locations according to latitude and longitude including the BATS (32N -64W) and these observations are obtained by time-series (5-days period and year). Therefore, we create 3 data: BATS data, spatial data, and temporal data.

1. BATS data

The BATS data contains all observations at the BATS (32N -64W). In this data, we have 10 variables: **SSH, CS, WS, SR, THERM 1, CHL 1, year, log10CHL4, sin.5days, cos.5days**. And we choose the **log10CHL4** as the target variables and the others are covariables (except the **year)** so we have 8 covariables.

2. Spatial data

We consider the BATS and all the points around the BATS according to their latitudes and longitudes. Then we totally have 9 points (9 locations including the BATS). We have 58 variables and 1241 observations in this data. For each location, we have **SSH_i, CS_i, WS_i, SR_i, THERM 1_i, CHL 1_i** (i is the index of the location so i is from 1 to 9). We also have 4 others variables **year, logCHL4, sin.5days, cos.5days.** We also choose the log10CHL4 as the target variables and the others are covariables (except the **year)** so we have 56 covariables.

3. Temporal data

We have 34 variables including **SSH, CS, WS, SR, THERM 1, CHL 1** and also these variables in 5days, 10 days before and after. We also have 4 others variables which are **year, logCHL4, sin.5days, cos.5days.** We also choose the log10CHL4 as the target variables and the others are covariables (except the **year)** so we have 32 covariables.

II. Model Selection

Firstly, we divide each dataset (BATS data, spatial data, and temporal data) into training set, the test set and the validation set with the proportion approximately 60%,20%,20% according to the **year**. We have the observations in 17 years (1992-2008). Then we sort the data by putting the value of **year** into increasing order. The training set contains the first 11 years (1992-2002), the test set contains the next 3 years (2003-2005), the last 3 years are contained in the validation set (2006-2008).

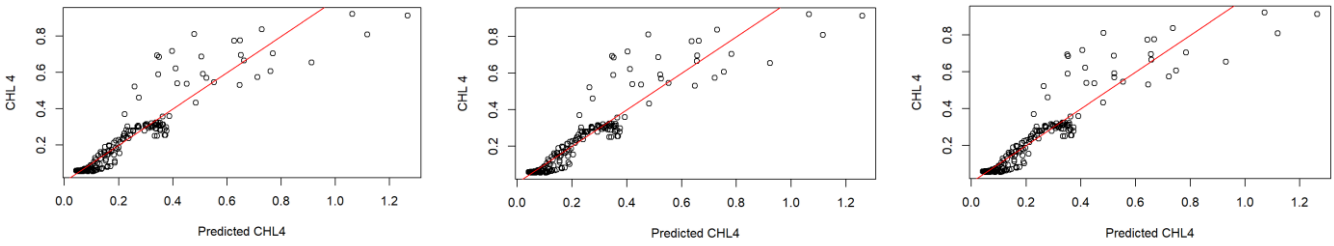

Figure 1. Graph (predicted CHL4, CHL4) Stepwise model (left), Ridge model (center) and Lasso model (right) for the test set of BATS data

According to the figure 1, we observe that based on the BATS data, the 3 models cannot predict well if the **CHL4**'s value is greater than 0.4. But they predict well if **CHL4**'s value is smaller than 0.4. Besides, we find that there are more points above the line than those below the line. So, we probably underestimate the real value.

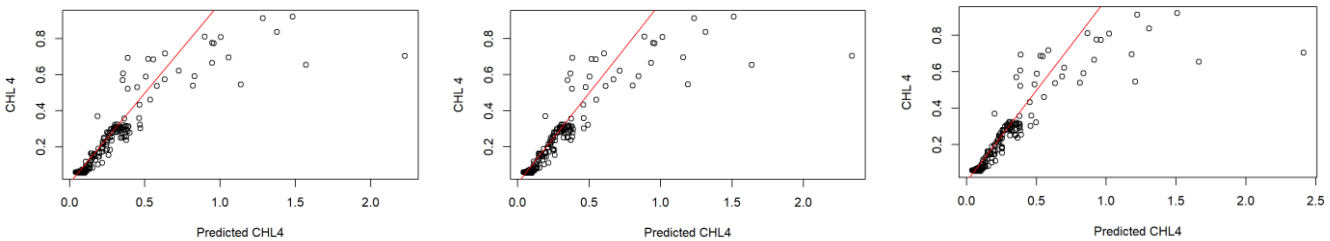

Figure 2. Graph (predicted CHL4, CHL4) of Stepwise model (left), Ridge model (center) and Lasso model (right) for the test set of spatial data

The Figure 2 based on the spatial data shows the same thing as the figure 1. But there are more points below the line so that we probably overestimate the real value. It also shows we may predict some extreme values.

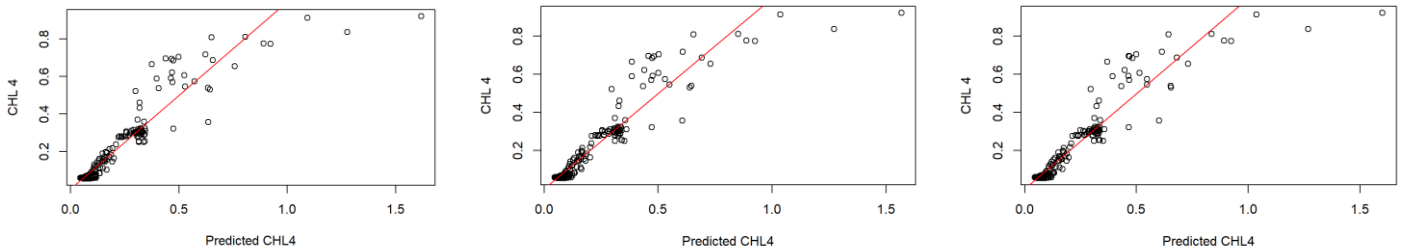

Figure 3. Graph (predicted CHL4, CHL4) Stepwise model (left), Ridge model (center) and Lasso model (right) for the test set of temporal data

The Figure 3 based on the temporal data shows the thing as Figure 1 and 2. But The distribution of the point above and below the line is balanced.

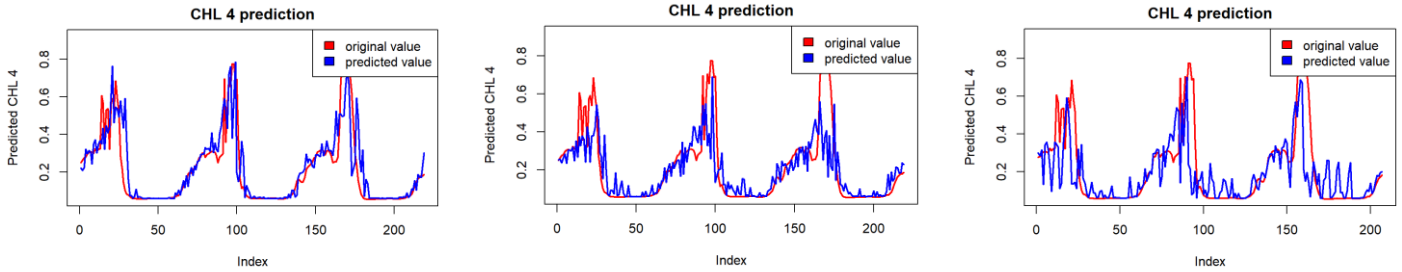

Figure 4. Graph of predicted CHL 4 and original CHL 4 of KNN model based on the BATS data (Left), spatial data (Center) and temporal data (Right)

We observe in Figure 4 that the KNN model based on the BATS data predicts the best because the prediction line (blue line) fits well the original line (red line).

The table below shows the RMSE, R^2 and adjusted R^2 for all our models:

Model with train and test set	RMSE	R^2	Adjusted R^2
Model 1 : Linear regression with stepwise (BATS data)	0.0794309	0.8960	0.8955
Model 2 : Linear regression L1 (Lasso) (BATS data)	0.07847807	0.8963328	0.89542
Model 3 : Linear regression L2 (Ridge) (BATS data)	0.07869614	0.8928186	0.8918749
Model 4 : Linear regression with spatial data	0.1566687	0.9439	0.941
Model 5 : Linear regression with temporal data	0.085525	0.9309	0.9297
Model 6 : Model 4 + L1 (Lasso)	0.1685647	0.9421853	0.9379286
Model 7 : Model 4 + L2 (Ridge)	0.164497	0.9421266	0.9378655
Model 8 : Model 5 + L1 (Lasso)	0.0834542	0.9321883	0.9292968
model 9 : Model 5 + L2 (Ridge)	0.08251172	0.9263664	0.9232266
Model 10 : KNN with BATS data	0.1127356	0.8755815	0.874486
Model 11: KNN with spatial data	0.137742	0.6635478	0.6387756
Model 12: KNN with temporal data	0.163511	0.6344531	0.6344531

Table 1. Result of all models trained by training set and predict on the test set

As we can see by the table, models based on the BATS data shows the smallest RMSE which is followed by the models based on the temporal data. However, we can keep more information about our initial data by using models based on temporal data because it shows bigger R^2 and adjusted R^2.

In conclusion, if we just consider the normal values of **CHL4**, we can choose the model 2 (Lasso model based on the BATS data). But if we need to consider about the extreme values, it's better to choose model 8 (Lasso model based on the temporal data) because it contains more information of **CHL4**.

We use model 2 and model 8 to predict CHL4 on the validation set. We obtain the RMSE of the model 2 (Lasso model based on BATS data) is 0.1077295 and the RMSE of the model 8 (Lasso model based on temporal data) is 0.09879187. Since the RMSE of model is smaller than model 2 and RMSE of model 8 is not too different when predicting on test set (0.0834542) and validation set (0.09879187), model 8 is better and more stable than model 2. Therefore, we will choose the model 8 (Lasso model based on temporal data) is the model to predict CHL 4.