Regression analysis on oceanographic dataset

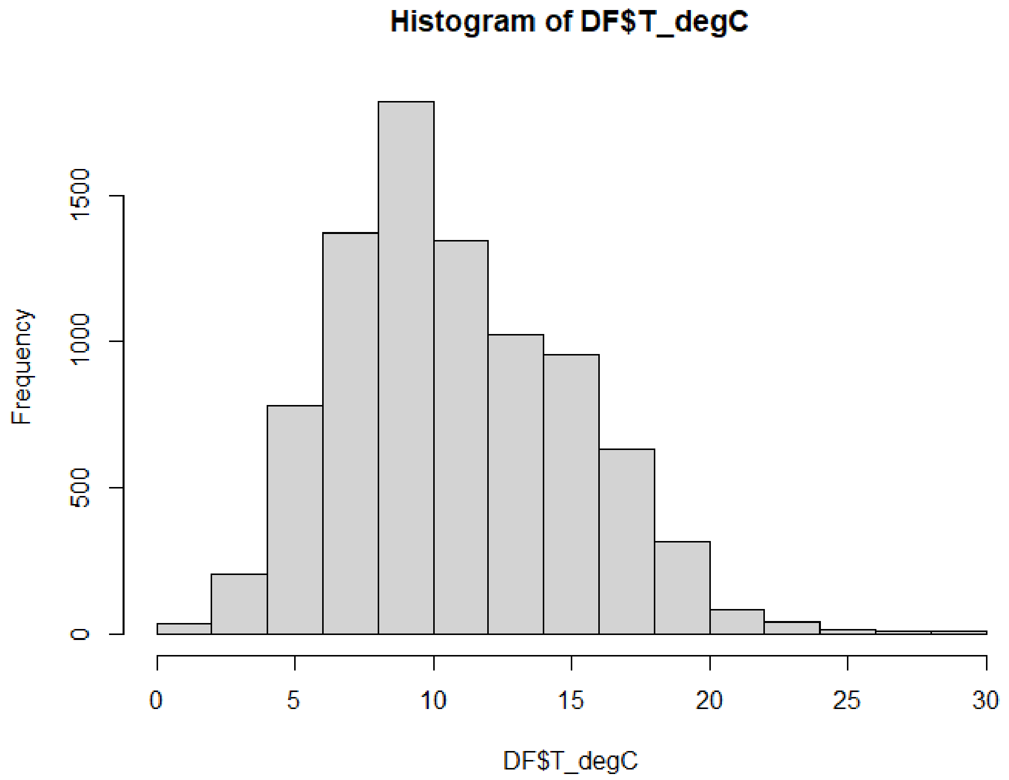
Group members: Diana Gorshechnikova and Tatiana Parashina

STA 9890

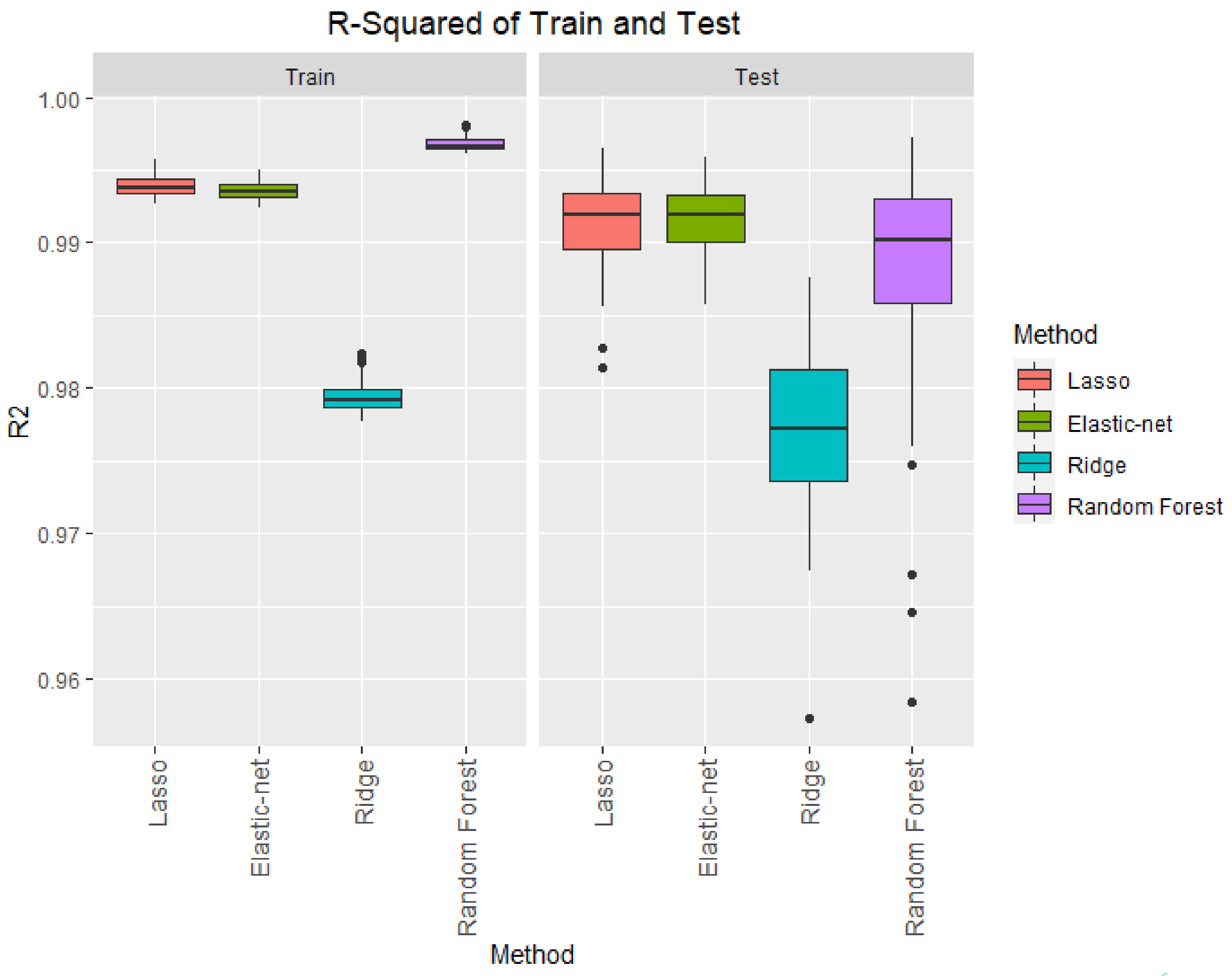
The California Cooperative Oceanic Fisheries Investigations (CalCOFI) dataset

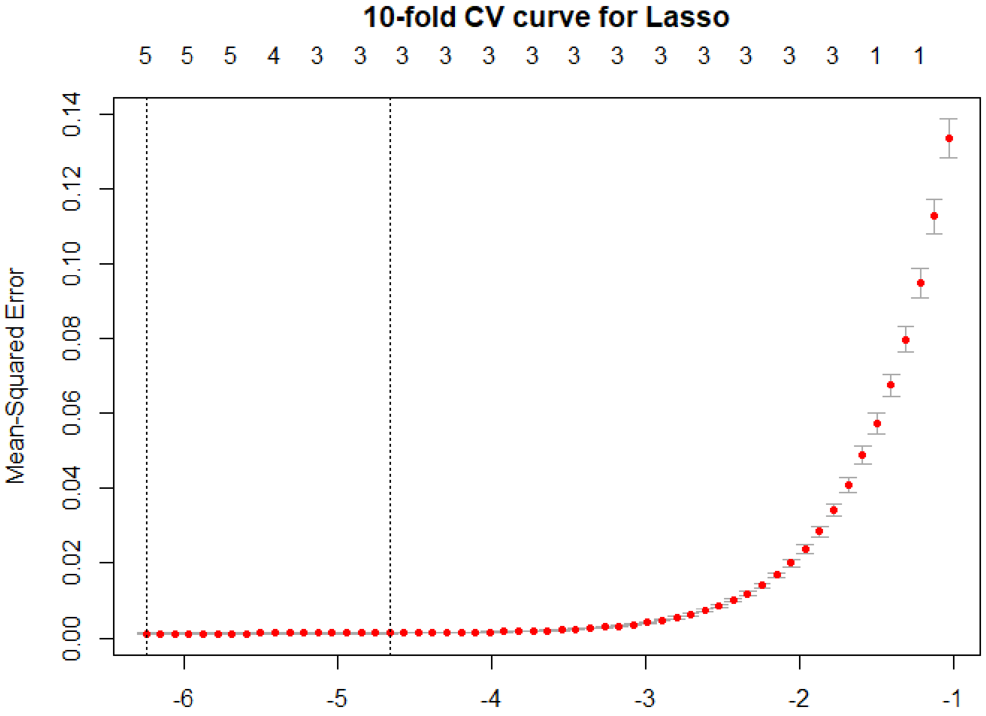
- Original dataset has 74 columns and 864863 observations;
- Response variable is water temperature;
- Increasing ocean temperatures severely affect marine species and ecosystems;
- Rising temperatures can contribute to coral bleaching and the loss of breeding grounds for marine fishes and mammals;
- Machine learning can be useful to predict what contributes to water temperature increase and to mitigate the rising temperatures in a timely fashion.

Minimum t	Maximum t	Mean t
1.44 C	31.14 C	10.8 C

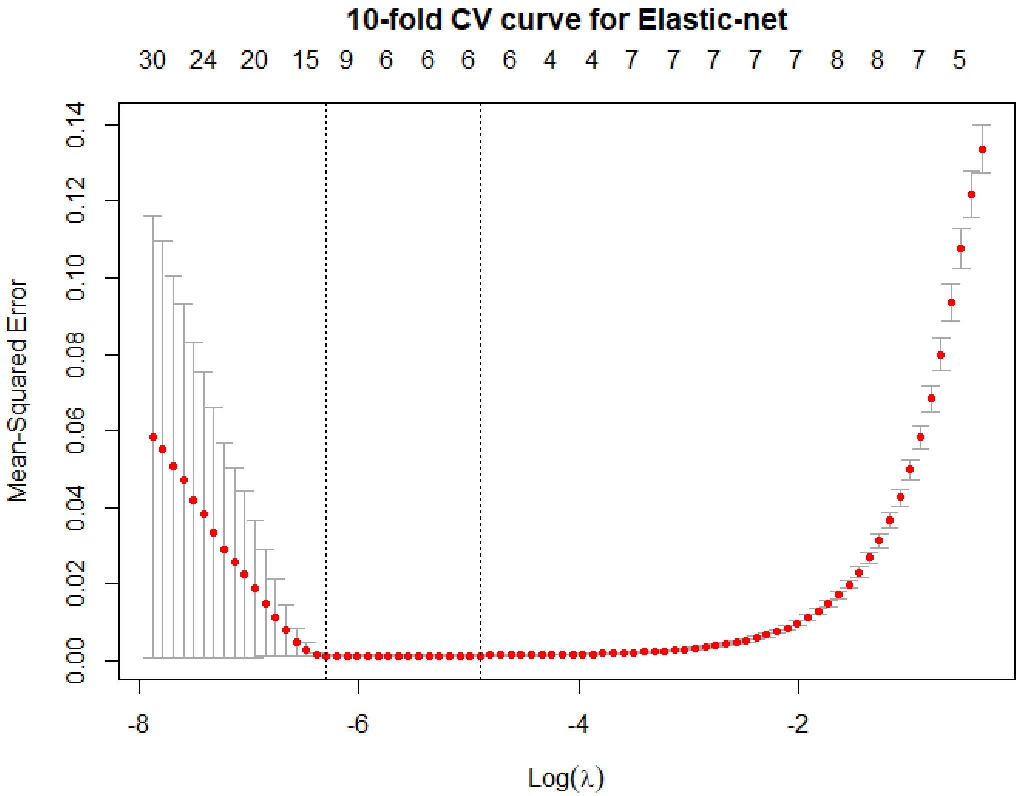


- The predictors are: salinity, oxygen, phosphate, silicate, nitrate and nitrite, chlorophyll, transmissometer, PAR, C14 primary productivity, phytoplankton biodiversity, zooplankton biomass, zooplankton biodiversity, etc.
- N = 1300
- P = 40

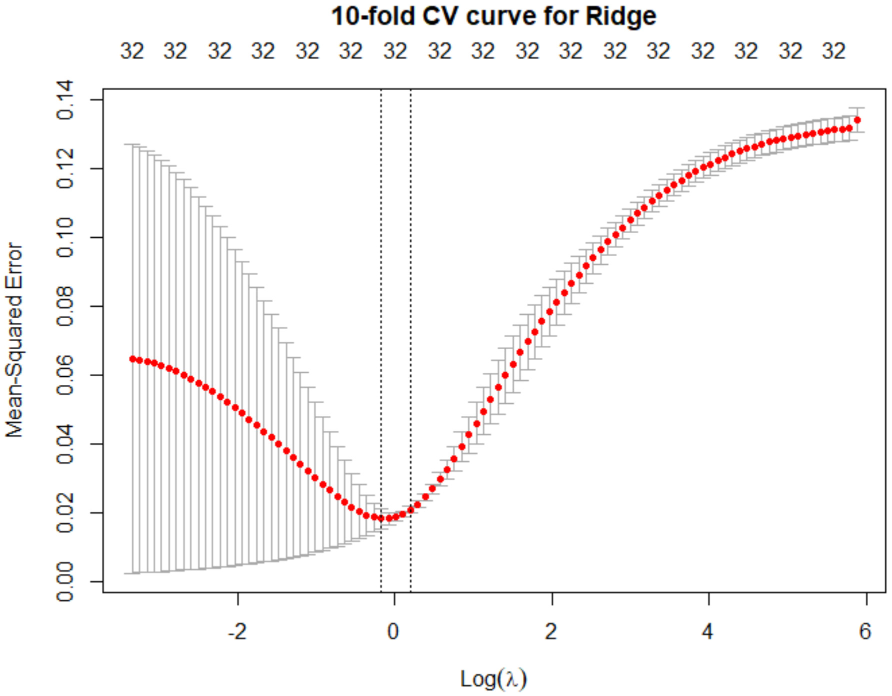

R-Squared of Train and Test
Train
Test
R2
1.00
0.99
0.98
0.97
0.96
Lasso
Elastic-net
Ridge
Random Forest
Method
Method
Lasso
Elastic-net
Ridge
Random Forest



10-fold CV curve for Lasso

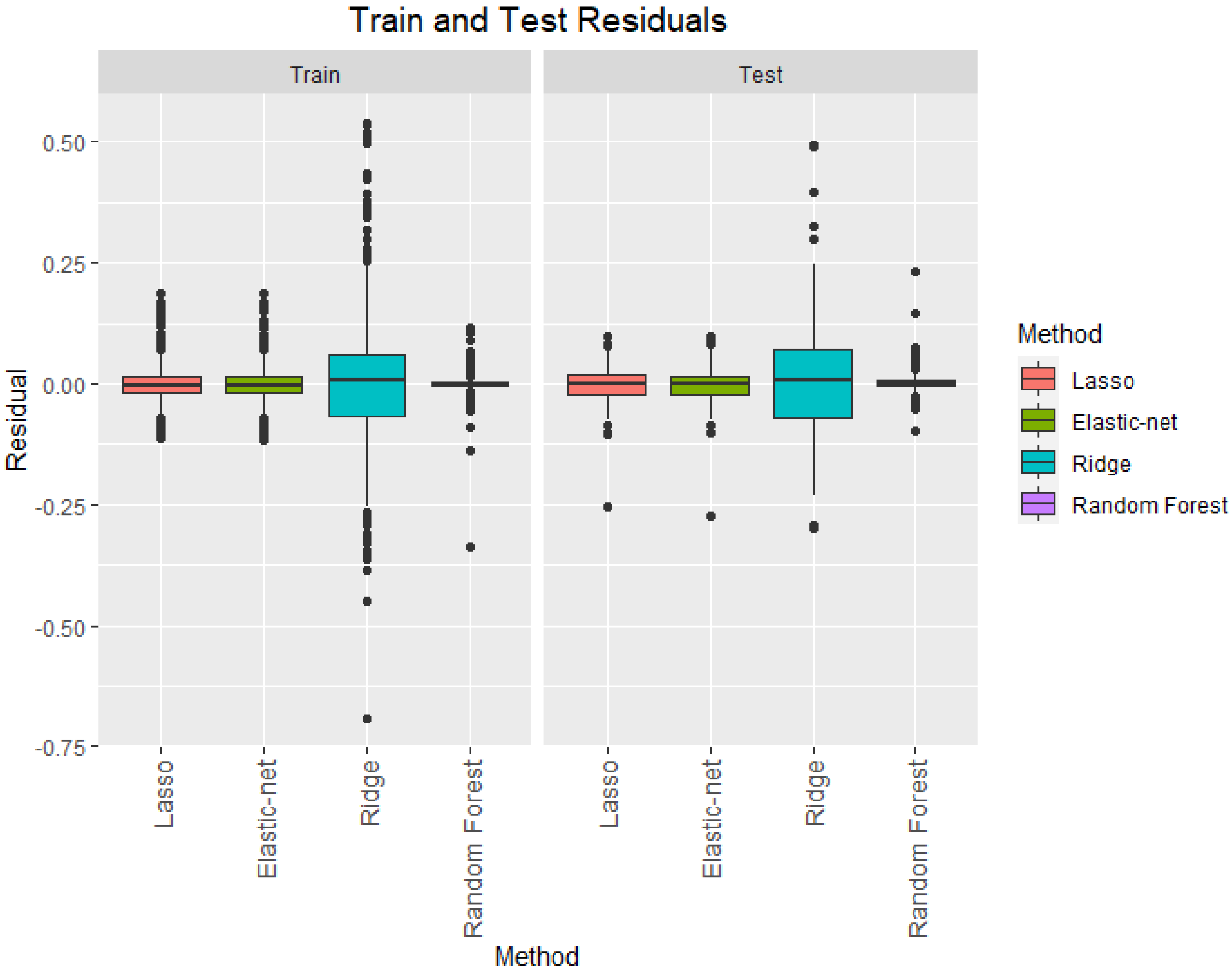


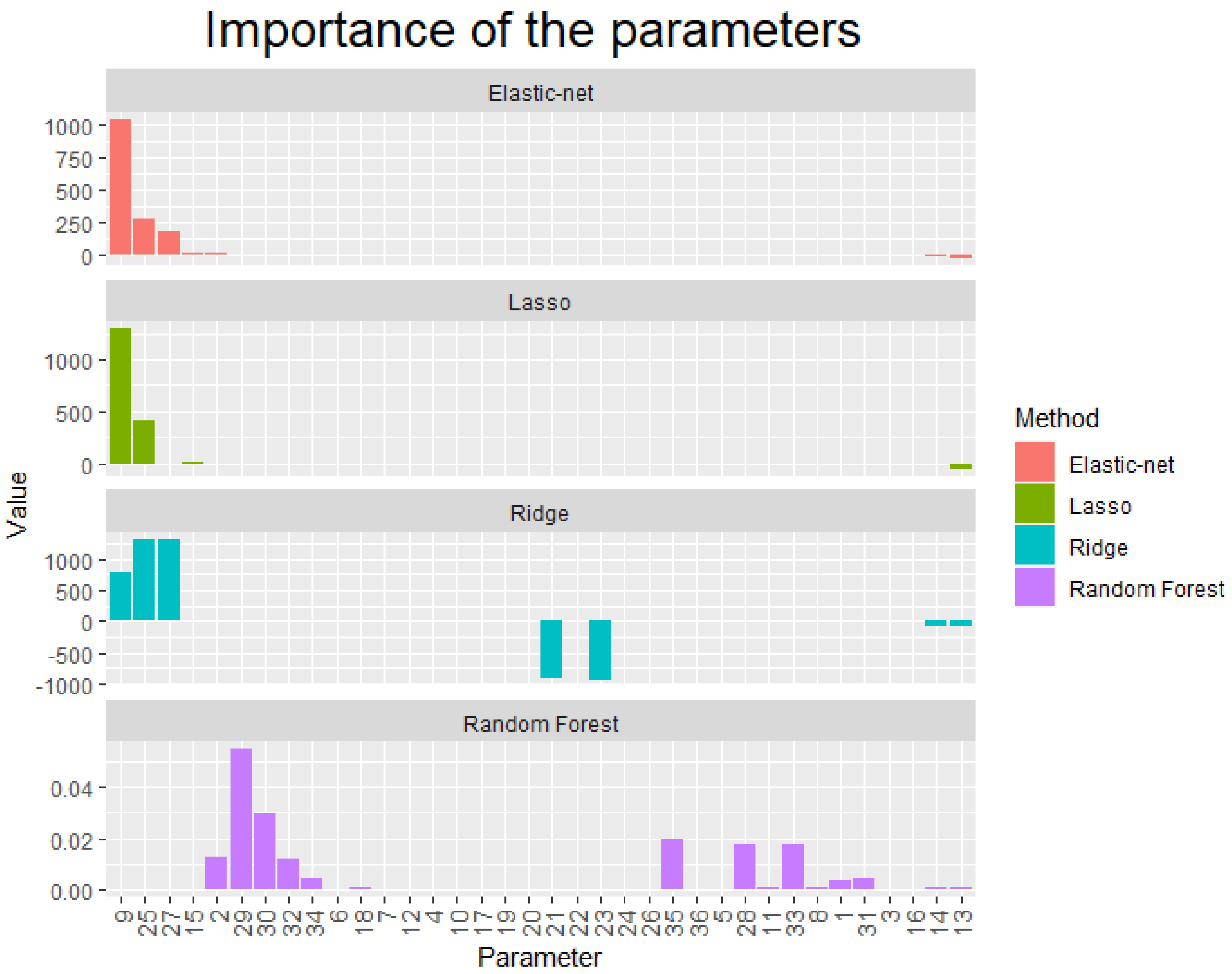
10-fold CV curve for Elastic-net



10-fold CV curve for Ridge

	Time (sec)
Lasso	0.870
El-net	1.050
Ridge	1.090


Train and Test Residuals
Train
Test
Residual
0.50
0.25
0.00
-0.25
-0.50
-0.75
Lasso
Elastic-net
Ridge
Random Forest
Method
Method
Lasso
Elastic-net
Ridge
Random Forest


Importance of the parameters
Elastic-net
1000
750
500
250
0
Lasso
1000
500
0
Ridge
1000
500
0
-500
-1000
Random Forest
0.04
0.02
0.00
Value
9
25
27
15
2
29
30
32
34
6
18
7
12
4
10
17
19
20
21
22
23
24
26
35
36
5
28
11
33
8
1
31
3
16
14
13
Parameter
Method
Elastic-net
Lasso
Ridge
Random Forest

Test R2 and time

	90% Test R2	Time (in sec)
Lasso	(0.85056 - 0.97669)	1.36
El-net	(0.8169 - 0.9473)	1.30
Ridge	(0.5393 - 0.7524)	1.48
RF	(0.99056 - 0.99206)	3.94

Concluding remarks

- In terms of time vs. efficiency Lasso gives the highest R2 and the fastest time which makes it the best model to predict water T;
- Ridge has the worst performance;
- Elastic net and lasso agree on most important features.