Regression analysis on oceanographic dataset

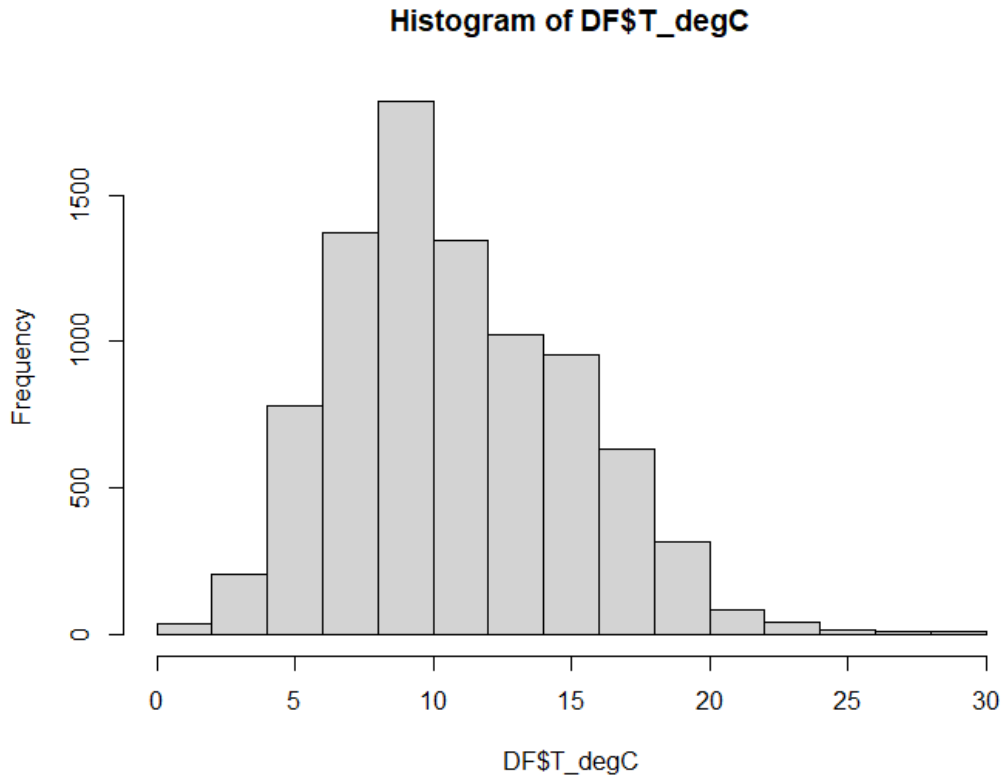
Group 17: Diana Gorshechnikova and Tatiana Parashina

STA 9890

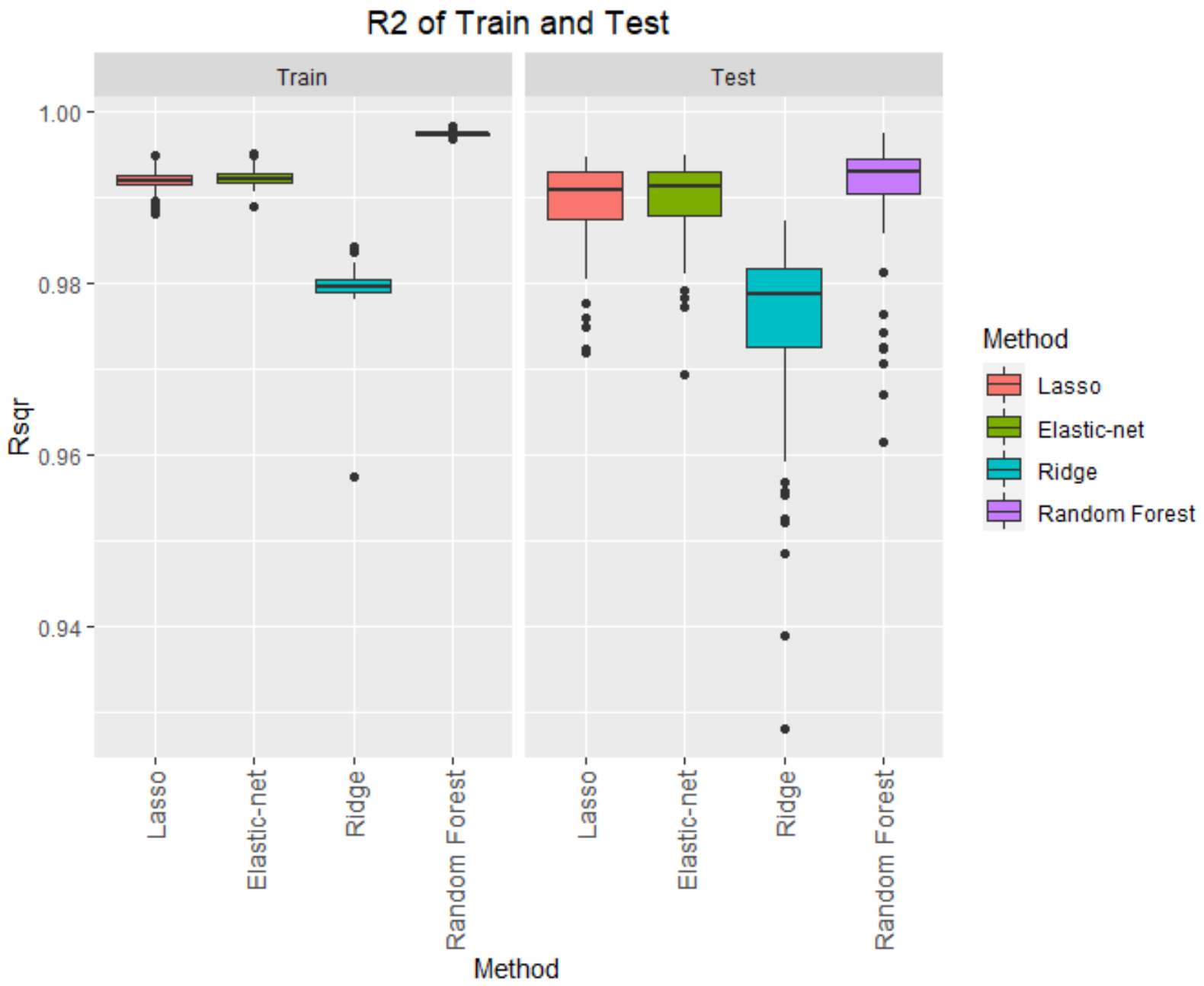
The California Cooperative Oceanic Fisheries Investigations (CalCOFI) dataset

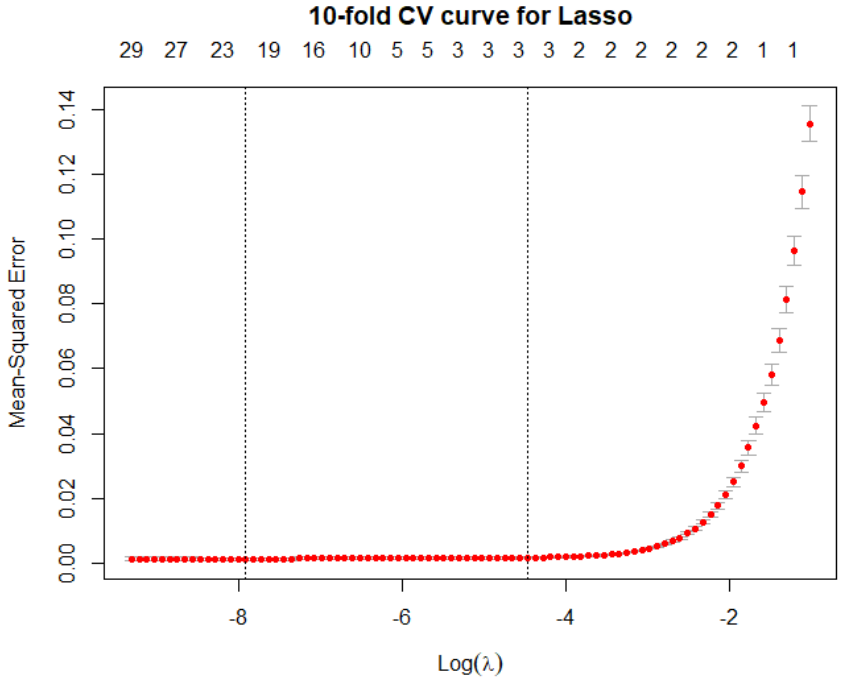
- Original dataset has 74 columns and 864863 observations;
- Response variable is water temperature;
- Increasing ocean temperatures severely affect marine species and ecosystems;
- Rising temperatures can contribute to coral bleaching and the loss of breeding grounds for marine fishes and mammals;
-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

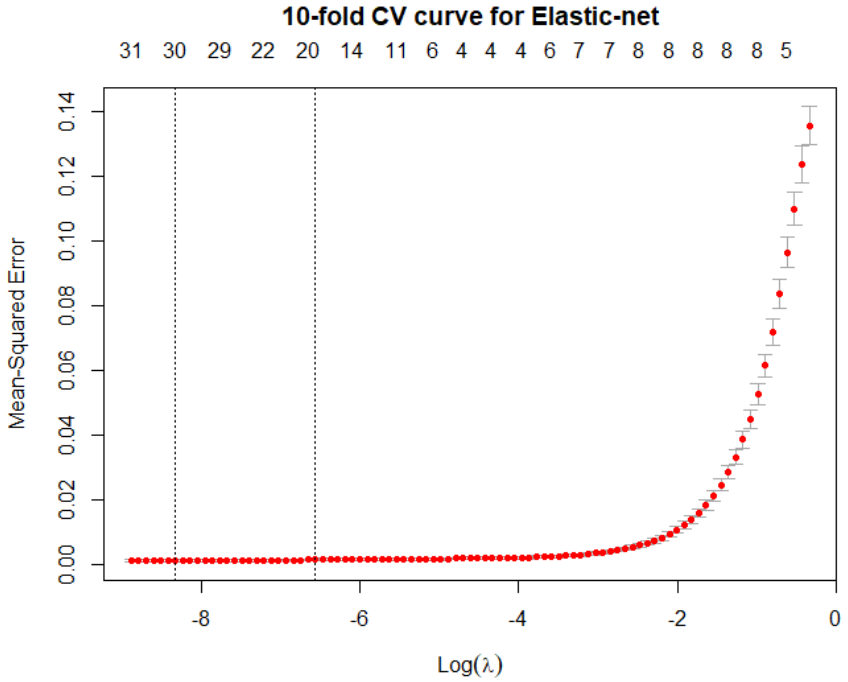
Minimum t	Maximum t	Mean t
1.44 C	31.14 C	10.8 C

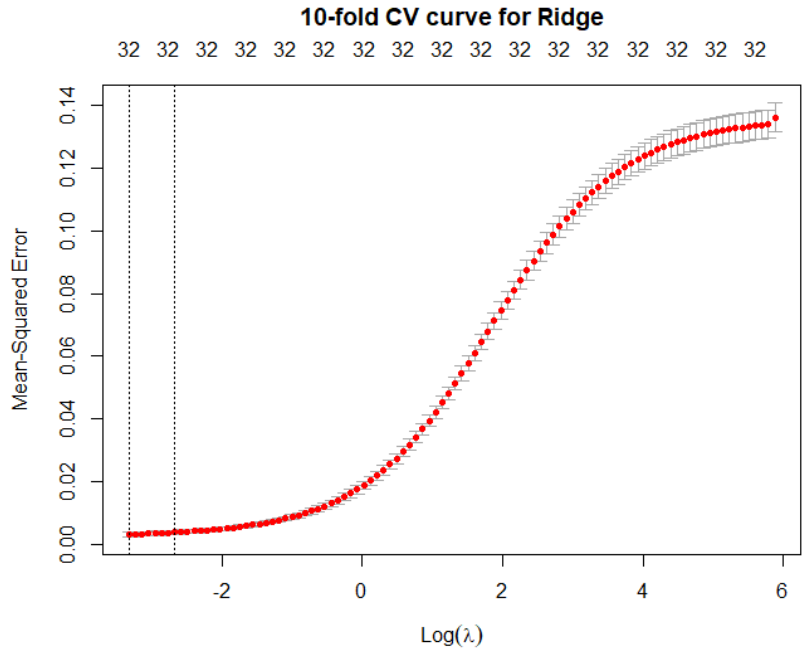


- The predictors are: salinity, oxygen, phosphate, silicate, nitrate and nitrite, chlorophyll, transmissometer, PAR, C14 primary productivity, phytoplankton biodiversity, zooplankton biomass, zooplankton biodiversity, etc.
- N = 1300
- P = 40

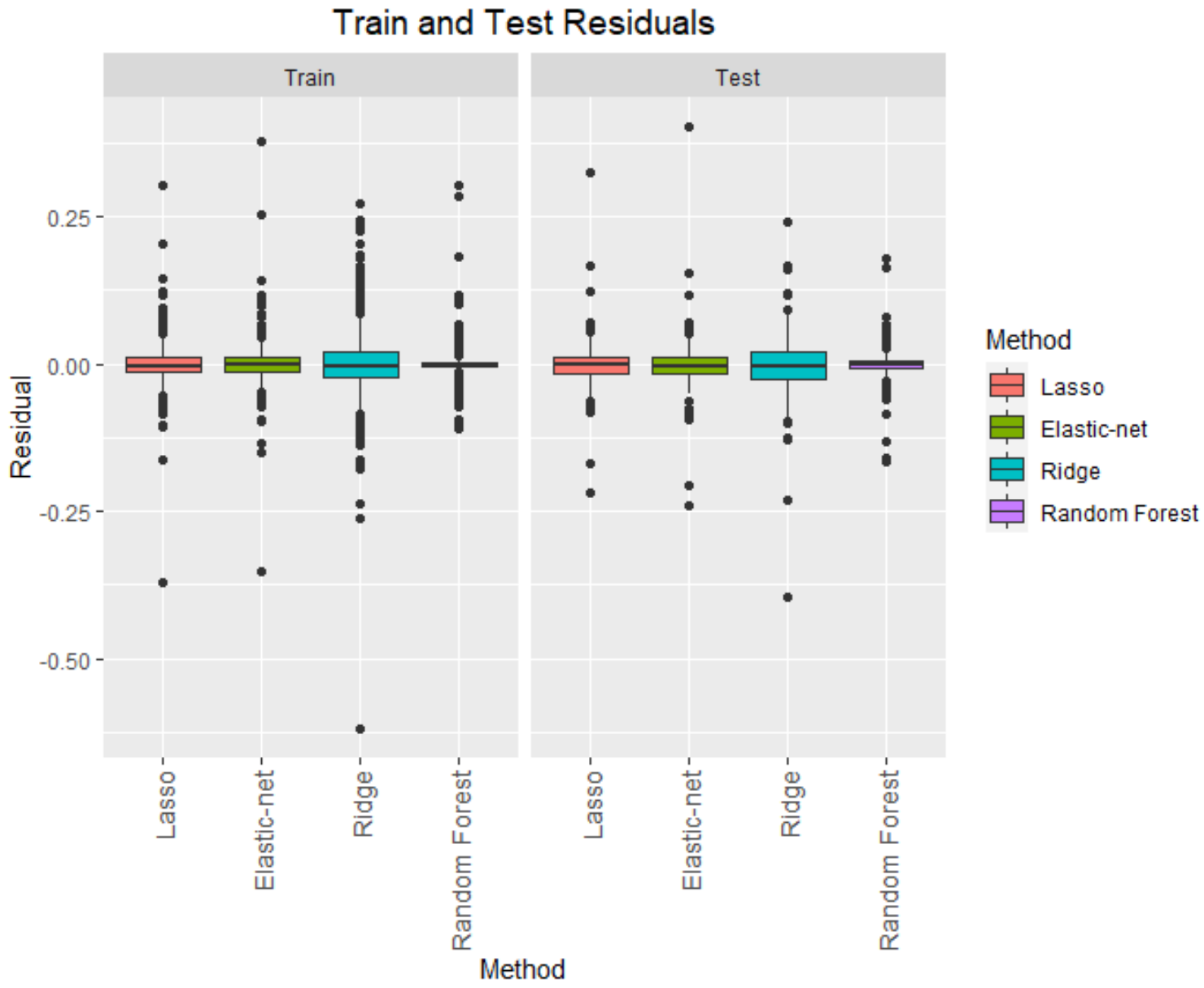

R2 of Train and Test
Train
Test
1.00
0.98
0.96
0.94
Rsqr
Lasso
Elastic-net
Ridge
Random Forest
Method
Method
Lasso
Elastic-net
Ridge
Random Forest



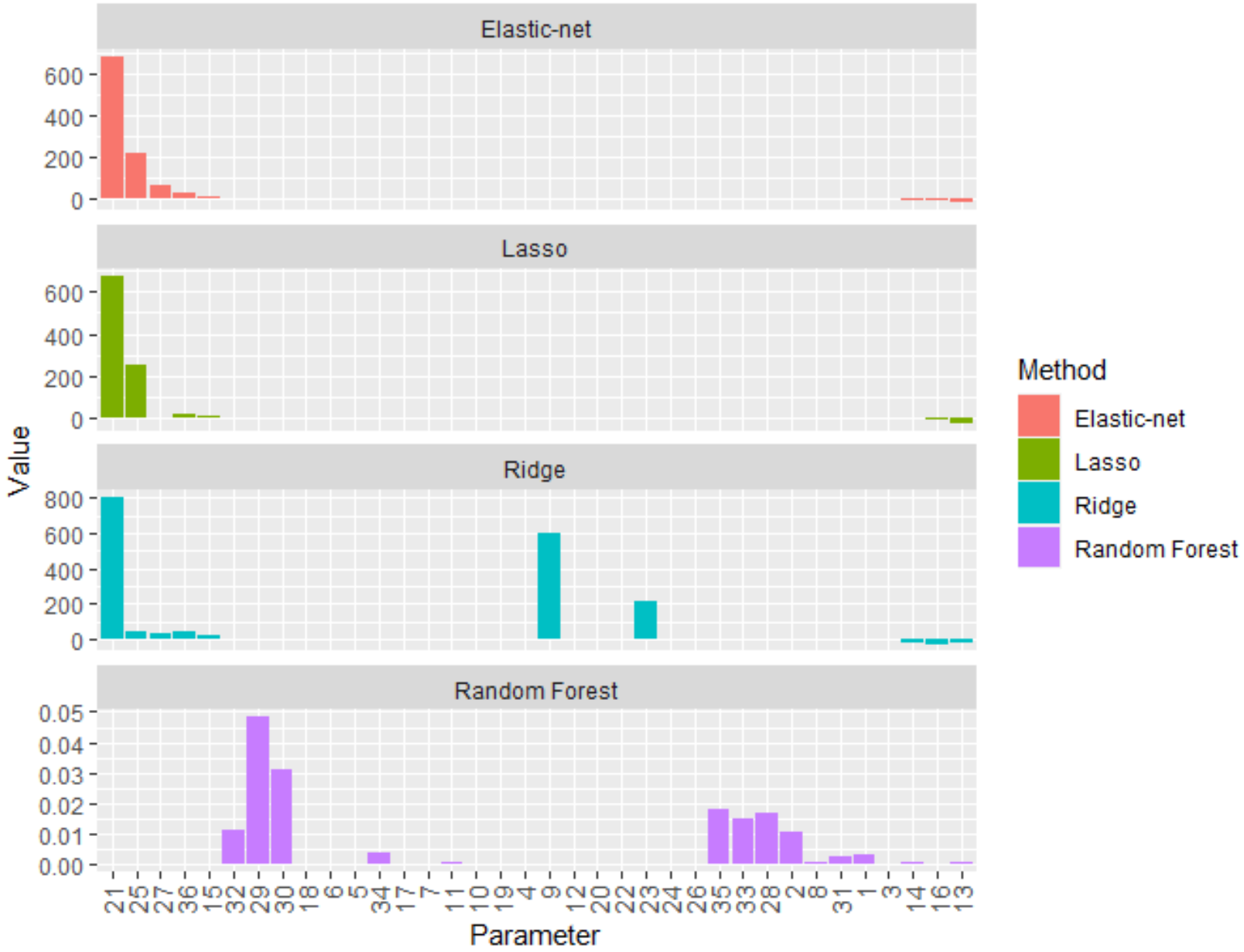




	Time (sec)
Lasso	1.14
El-net	0.88
Ridge	0.95


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

Importance of the parameters



Test R2 and time

	90% Test R2	Time (in sec)
Lasso	(0.9882 - 0.9899)	0.64
El-net	(0.9891 - 0.9906)	0.14
Ridge	(0.9738 - 0.9771)	0.14
RF	(0.9900 - 0.9922)	3.53

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.