Data Analysis Project 2

MA8701 Advanced Methods in Statistical Inference and Learning V2021

Florian Beiser, Yaolin Ge & Helene Minge Olsen

27 april, 2021

Random Forest Model Fit

As our data set is a time series, we include the water temperature the day before as a covariate. As a result, we know that this strong predictor `water_temp` will dominate the trees produced if we choose to use bagging. To obtain decorrelated trees and improve the variance reduction, we therefore want to fit a random forest model to our data set.

We utilize the `randomForest` package in R. As the random forest allows a random selection of m covariates $m \leq p$ to be considered for the split for each node, a 5-fold cross validation was applied to find the optimum number of covariates. As the data set contains 19 variables, the general rule for regression is to set the number of randomly chosen variables to be considered as $floor(\frac{19}{3}) = 6$. However, we obtain the smallest cross-validation error for $m = 4$. This can be explained by our data set being very correlated, and it is therefore wise to set m to be small.

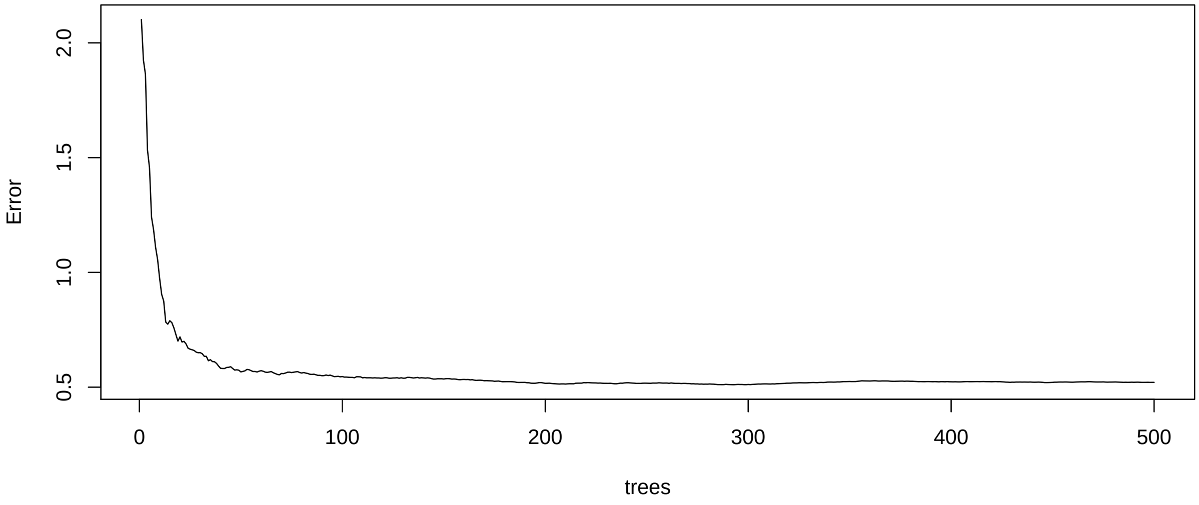

Figure 1: Plot of numbers of trees

As each tree uses a different bootstrap sample, the OOB sample is used as a validation set. From Figure 1 we observe how the OOB error estimate is reduced as a function of the nr of trees. As a result, the model fits 500 trees to average over, with a percentage of variance explained, also known as pseudo R-squared, around 91.97%

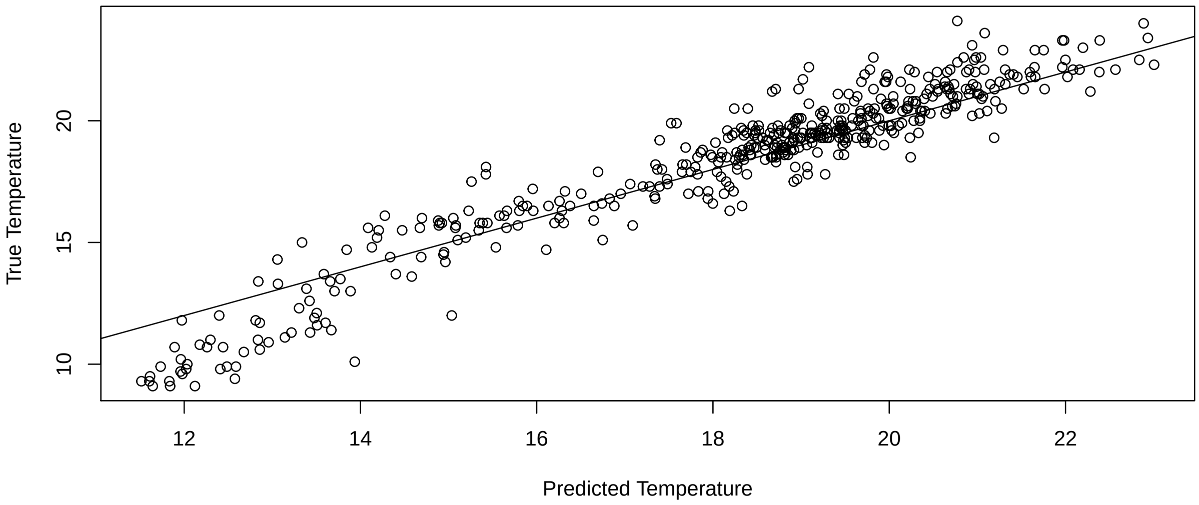

Figure 2: Predicted test values against true values.

Figure 2 depicts a good fit except for a slight overprediction for lower temperatures between 9 and 13 degrees.

Figure 3 and 4 depicts that the variable importance plot based on the OOB sample tends to spread the importances more uniformly, but `water_temp` remains a main predictor for both. We observe that there are a lot of predictors considered to be relevant after `water_temp`, which indicates a well performance from the random forest model. This is reflected in the low test MSE.

```
## Test MSE of Random Forest Model
##  1.163736
```

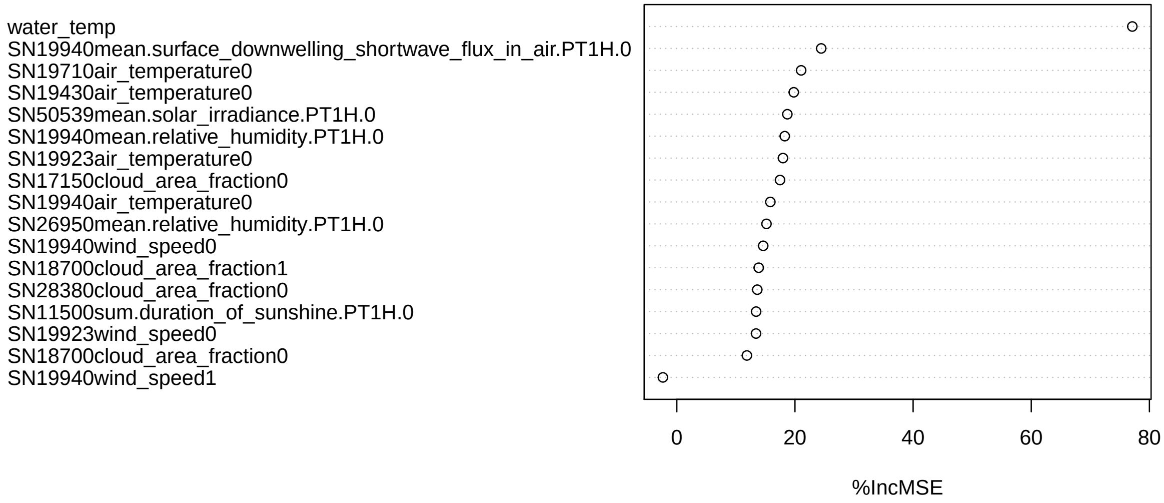

Figure 3: Variable importance plot based on increase in MSE when premuted and tested using OOB sample, higher values indicate larger impact when premuted, hence larger importance.

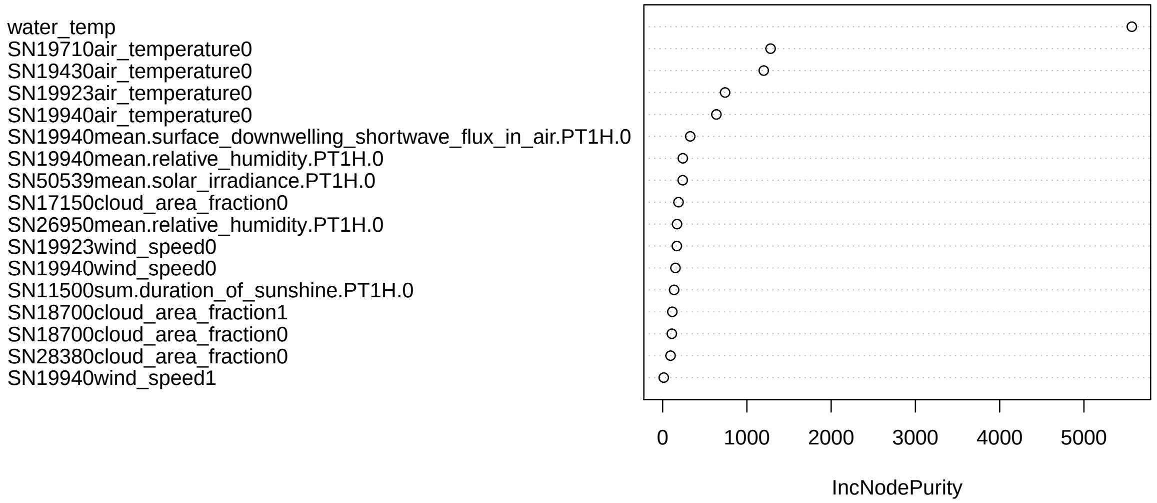

Figure 4: Variable importance plot based on increase in node purity, higher values indicate larger importance.