Overcoming Incomplete Coverage of the Domain in soil organic carbon spectral modeling - handling with the lack of representativeness of the dataset

Observations from a domain used to train a model are a sample and incomplete by definition. In statistics, a random sample refers to a collection of observations chosen from the domain without systematic bias. There will always be some bias. We will never have all of the observations. If we did, a predictive model would not be required.

A suitable level of variance and bias in the sample is required such that the sample is representative of the task or project for which the data or model will be used.

In spectroscopy we aim to collect or obtain a suitably representative random sample of observations to train and evaluate a machine learning model. Often, we split the dataset ... ramdonly

For example, we might choose to measure the size of randomly selected samples in a spectral library to teste/validate a soil organic carbono spectral model. The samples to test are randomly selected and we test whenever the variance between training and validation dataset differ.

Even when we stratify according to the histogram (quartilhes), there is no guarantee that samples will covared the domain of the variable.

Because of that we usually attemp to select samples for calibration from the histogram. However, there is no guarantee that samples in the extreme of the distribution will be selected.

We test the heterocedasticity between datasets.

And we use this dataset to evaluate how good our predictions are. In this case, everytime the test set chance, the evaluation of the model chances as well.

Or even to run several times the split and compute the mean of the evaluation.

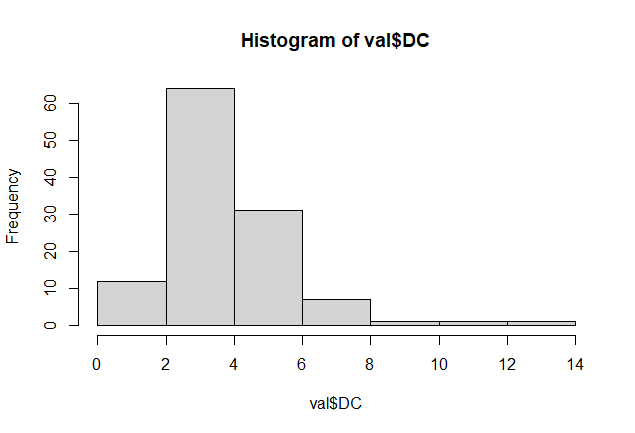
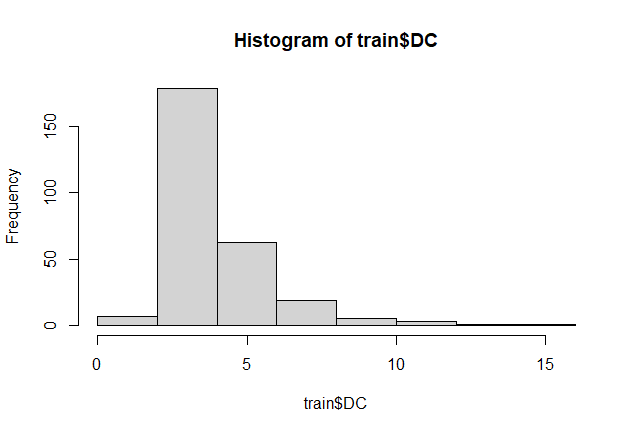
However, whenever the space cover by the test set excide the limit of the training set, there is an effect called covariate shift, that impart the predictions.

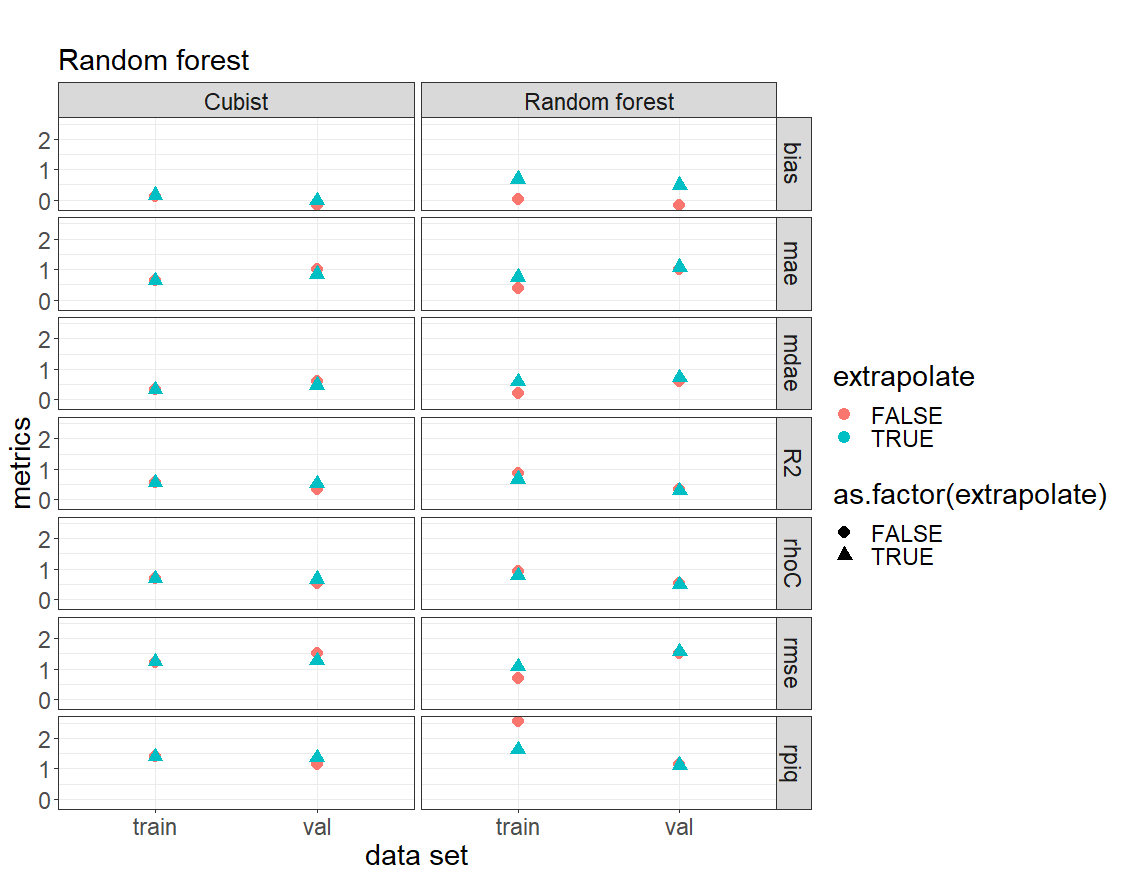
This means that there will always be some unobserved cases. There will be part of the problem domain for which we do not have coverage. No matter how well we encourage our models to generalize, we can only hope that we can cover the cases in the training dataset and the salient cases that are not.

This is why we split a dataset into train and test sets or use resampling methods like k-fold cross-validation. We do this to handle the uncertainty in the representativeness of our dataset and estimate the performance of a modeling procedure on data not used in that procedure.

This is why we split a dataset into train and test sets or use resampling methods like k-fold cross-validation. We do this to handle the uncertainty in the representativeness of our dataset and estimate the performance of a modeling procedure on data not used in that procedure.

Study case



****

**Encouraging models to extrapolate**

