UAE Lab Data Analysis

There were 295 samples in the UAE Lab data which were used for building the best models for sand, silt, clay and TOC for the UAE soil. The process of finding the hyper-tuned best model was divided into the following major steps.

1. Noise Removal and Smoothing of Spectra

The plots of the original spectra highlighted the first 50 wavelengths (350 nm- 400 nm) as noise. So, the initial 50 wavelengths were removed and the final spectra comprised of wavelengths 400 nm to 2500 nm.

We applied Savitzky-Golay second order filter at window lengths 51 to get rid of the remaining noise (in the range 400 nm to 2500 nm) without losing the actual nature of spectrum. The final spectra (after smoothing) to be considered for further preprocessing was named as spec2[51].

2. Preprocessing and setting up the Spectra

For further processing we considered following three variants of the original spectra (i.e. spec2[51]).

- None:

  This category represents the smoothed spectra that we obtained from the previous step (i.e. spec2[51]).

- CR (Continuum Removed):

  This category represents the continuum removed spectra which was obtained from spec2 [51] and the obtained spectra was finally named as cr_spec.

- LOG (Logarithmic Transformation):

  This category represents the spectra obtained from applying Logarithmic transformation on the inverse of spec2[51] (i.e. log(1/spec2[51])) and the obtained spectra was finally named as log_spec.

3. Resampling at multiple bands

The main highlight of our approach in the analysis of UAE lab data was that we combined the technique of resampling with the above-mentioned spectral preprocessing as well as with FOD (first order derivative) as described below. All the spectra obtained above were resampled to the following bands (referred to as n_bands):

n_bands: [0,2,3,5,7,9,10,11,13,15,17,19,20,21,23,25,27,29,30,31,33,35,37,39,40,45,50,55,60,70,80,90,100]

We maintained a similar convention while renaming the obtained resampled spectra. For instance, spectra obtained as a result of resampling cr_spec to *n* bands is termed as sampled_cr[n], spectra obtained as a result of resampling log_spec to *n* bands is referred as

sampled_log[n] and spectra obtained as a result of resampling spec2[51] to *n* bands is referred as sampled_spec[n].

Thus, we have

- sampled_cr [50] denoting cr_spec resampled to 50 bands
- sampled_log[50] denoting log_spec resampled to 50 bands.
- sampled_spec[50] denoting spec2[51] resampled to 50 bands.

First Order Derivative

Furthermore, first order derivative was applied to sampled_spec[n] for each *n* in n_bands and the resulting spectra is termed as fod_sampled[n]. Hence, fod_sampled[50] represents first order derivative applied on the smoothed spectra resampled to 50 bands.

To summarize, we start building the model for a given attribute by considering a total of 132 preprocessed spectra resulting from the following variants.

- sampled_spec[n] : contains 33 spectra each corresponding to a different value of n.
- sampled_cr[n] : contains 33 spectra each corresponding to a different value of n.
- sampled_log[n] : contains 33 spectra each corresponding to a different value of n.
- fod_sampled[n] : contains 33 spectra each corresponding to a different value of n.

4. Obtaining best ML models (for each attribute)

Our spectra and attributes are now ready for building the best models.

In our analysis we have used the following 6 machine learning algorithms to build the models:

- Multiple Linear Regression ('mult')
- Support Vector Regression ('SVR')
- Partial Least Square Regression ('PLSR')
- Kernel Ridge Regression ('ridge')
- Cubist ('cubist')
- Gradient Boosting Regression Tree ('GBRT')

Each of these ML algorithms have parameters known as hyperparameters whose values control the learning process. Value of the hyperparameter is set before the learning algorithm begins training the model and their values cannot be changed during the training process. Hyperparameter tuning is the process of determining the right combination of hyperparameter that maximizes the model performance.

For a given attribute (say sand) we obtain best ML model (for predictions) by considering all possible combinations of ML algorithms and all the available preprocessed spectra.

For instance, for the sand attribute we have 132*6 = 792 combinations resulting from the following choices of the ML algorithms and the preprocessed spectra.

- 6 ML algorithms
- 132 preprocessed spectra.

For each of this combination we obtain the best sand model by hypertuning model parameters and evaluating the corresponding model performance (i.e. validation). Note that we used the leave one out method in hypertuning the individual model parameters as well as for evaluating the model performance.

As an illustration, assume the cubist method and the preprocessed spectra sampled_cr[35]. In our work, for the cubist method, we considered two hyperparameters n_committees (with possible values 5,10,15,20) and n_rules (with possible values 10,20,30,40,50) totalling 20 combinations for hypertuning using grid-search. To arrive at the best model (for cubist method and sampled_cr[35]) we evaluated model performances corresponding to each of these 20 combinations using the leave-one-out technique.

Note that the best model search space may vary due to different grid sizes for different ML algorithms. Following are the grid sizes corresponding to each algorithm:

- Multiple Linear Regression ('mult') : 2
- Support Vector Regression ('SVR') : 18
- Kernel Ridge Regression ('ridge') : 14
- Cubist ('cubist') : 20
- Partial Least Square Regression ('PLSR') : 10
- Gradient Boosting Regression Tree ('GBRT') : 18

Finally, (for sand) we obtain best models for each of the 792 combinations of preprocessed spectra and ML algorithms. We choose the best model among these available 792 models depending upon the R2 scores evaluated for each model using the leave one out technique. It follows that for each attribute the total search space for finding best model was 132*(2 + 18 + 14+ 20 + 10 + 18) = 10824.

In the case of UAE data, the best model for sand turned out to be the model corresponding to the method cubist and spectra fod_sampled[100] with the R2 score 0.49. Following are the optimal ML model details for each attribute considered in this project:

Attribute	**Method**	**Spectral Preprocessing**	**R2 Score**
Sand	Cubist	fod_sampled[100]	0.49
Silt	Cubist	sampled_log[80]	0.32
Clay	GBRT	sampled_cr[70]	0.63
TOC	Cubist	fod_sampled[29]	0.55

For visual comprehension of Model Accuracy see the attached detailed project report that contains Scatter Plots (R2 Score), Stem Plots (R2 vs n_bands), and Feature Importance Plots.