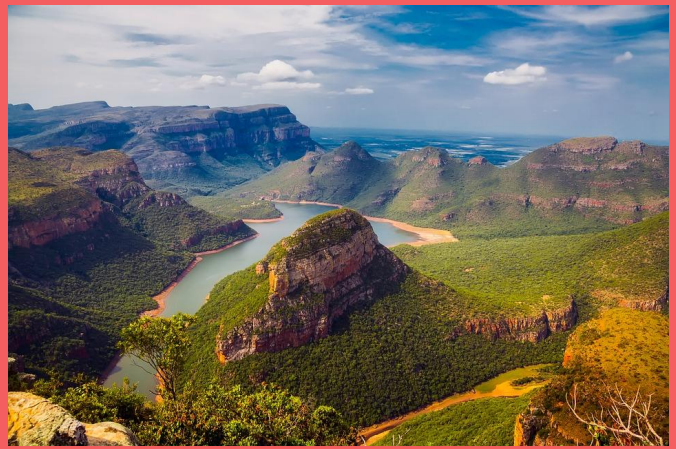
Africa Soil Property Prediction

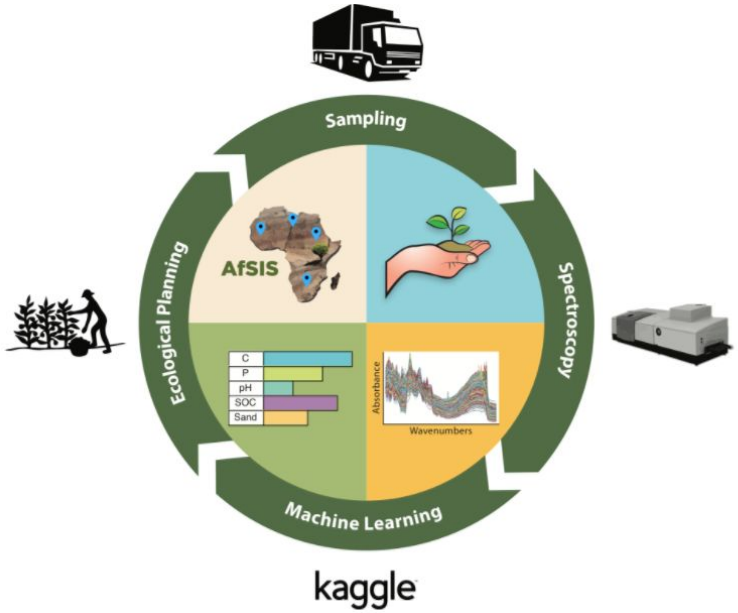
Jessica Sanders, Jennifer Casper, Carlos Castro

April 24th, 2017

W207 - Machine Learning



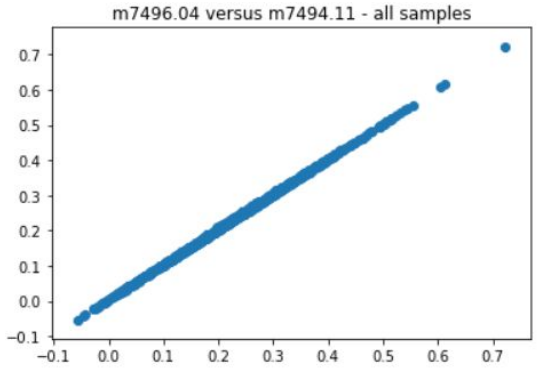
“Functional” properties can tell us how well soil will support an ecosystem



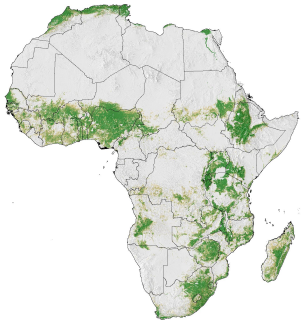
- Data: 1157 soil samples from 60 “Sentinel Landscapes” in Africa , each of which has:
  - Mid-infrared light absorption measurements f separate wavelengths.
  - Seventeen other predictor variables, such as s depth, mean annual rainfall from sample area, etc.
- Challenge: Predict 5 target soil functional properties:
  - SOC: Soil organic carbon
  - pH: pH values
  - Ca: Calcium content
  - P: Phosphorus content
  - Sand: Sand content

The data presented with some challenges

Adjacent wavelength measurements (features) are nearly co-linear across the samples.



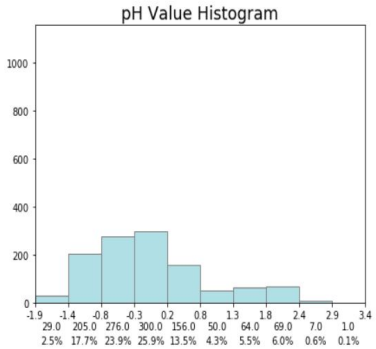
Training/test data was split along lines of Sentinel Landscapes, and thus may be inherently different in some ways.

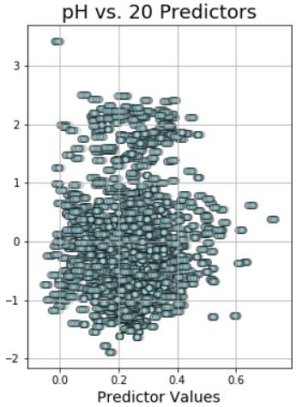


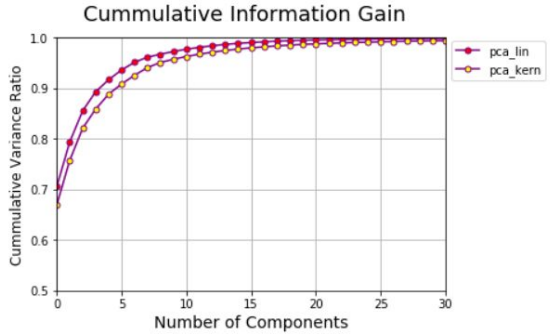
Number of features (3594) >> number of samples (1157)

Our process (Python/sklearn)

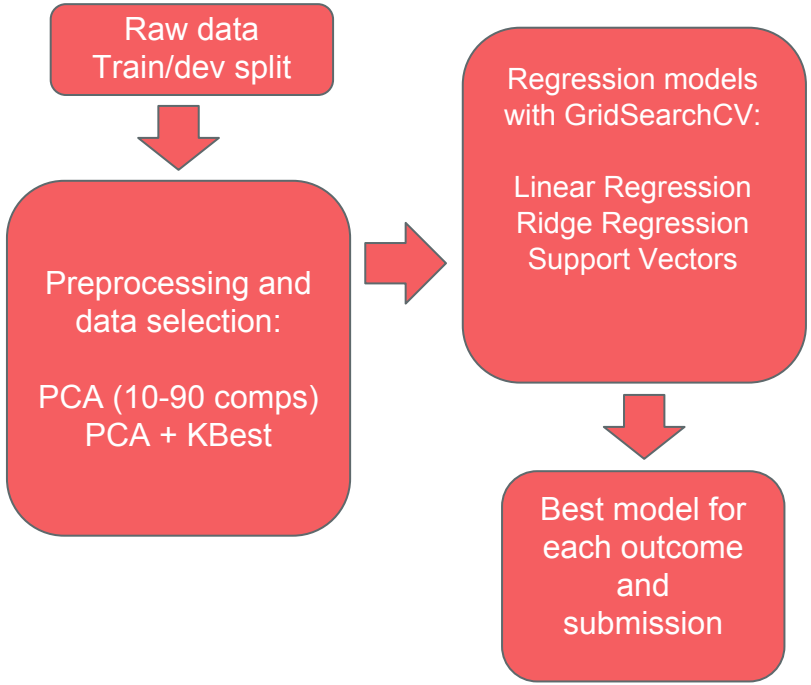
EDA and PCA experiments







Regression pipeline and model selection



Lessons Learned

- Tried many different models, including:
  - Linear Regression
  - Random Forests,
  - Neural Networks,
  - Support Vector Regression
  - K-Nearest Neighbors
  - … and linear regression tended to win out.
- Found that the scoring was pretty sensitive to changes in the train/dev split
- Encoding the categorical depth predictor variable was not helpful
  - Phorphorus content MSE varied given depth

- Future options
  - Preprocessing alternatives
  - Split the data differently
  - Explore the other data collected - weather and spatial

Results:

The scoring metric is the mean column root mean squared error:

$$\text{MCRMSE} = \frac{1}{5}\sum_{j=1}^{5}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{ij} - \hat{y}_{ij})^2},$$

Our final score: 0.53557

The competition winner scored: 0.46892