File descriptions

- train.csv - the training set has 1158 rows.
- test.csv - the test set has 728 rows.
- sample_submission.csv - all zeros prediction, serving as a sample submission file in the correct format.

Data fields

SOC, pH, Ca, P, Sand are the five target variables for predictions. The data have been monotonously transformed from the original measurements and thus include negative values.

- PIDN: unique soil sample identifier
- SOC: Soil organic carbon
- pH: pH values
- Ca: Mehlich-3 extractable Calcium
- P: Mehlich-3 extractable Phosphorus
- Sand: Sand content
- m7497.96 - m599.76: There are 3,578 mid-infrared absorbance measurements. For example, the "m7497.96" column is the absorbance at wavenumber 7497.96 cm-1. We suggest you to remove spectra CO2 bands which are in the region m2379.76 to m2352.76, but you do not have to.
- Depth: Depth of the soil sample (2 categories: "Topsoil", "Subsoil")

We have also included some potential spatial predictors from remote sensing data sources. Short variable descriptions are provided below and additional descriptions can be found at AfSIS data. The data have been mean centered and scaled.

- BSA: average long-term Black Sky Albedo measurements from MODIS satellite images (BSAN = near-infrared, BSAS = shortwave, BSAV = visible)
- CTI: compound topographic index calculated from Shuttle Radar Topography Mission elevation data
- ELEV: Shuttle Radar Topography Mission elevation data
- EVI: average long-term Enhanced Vegetation Index from MODIS satellite images.
- LST: average long-term Land Surface Temperatures from MODIS satellite images (LSTD = day time temperature, LSTN = night time temperature)
- Ref: average long-term Reflectance measurements from MODIS satellite images (Ref1 = blue, Ref2 = red, Ref3 = near-infrared, Ref7 = mid-infrared)
- Reli: topographic Relief calculated from Shuttle Radar Topography mission elevation data
- TMAP & TMFI: average long-term Tropical Rainfall Monitoring Mission data (TMAP = mean annual precipitation, TMFI = modified Fournier index)

BART Example

An example model using Bayesian Additive Regression Trees can be found here.

FAQ

Why not more data?

We will not introduce additional data (e.g. georeference) at this stage of the competition. We think that would be confusing (to us), as we would really like to find out how predictive the spectral methods are/would be when they are applied in new places and/or at different points in time by data

science experts such as yourselves. Subsequent Kaggle competitions may focus on explicitly spatial and or space-time predictions.

Background on data set creation

There have been a number of questions regarding why and how the data were ordered in the training and test sets. As some of you have surmised there is certainly geographical clustering in this dataset. This is due to the spatially stratified multilevel sampling design that was used to assemble the data. The following is an abbreviated version of how this came about.

When the Africa Soil Information Service (AfSIS) project started in 2009, we were faced with the enormous logistical task of obtaining a representative sample covering ~18.1 million km2 of the non-desert portion of Africa, including Madagascar, that could be used as a baseline for monitoring soil and other ecosystem properties.

The way we chose to go about this was to select 60, 10 × 10 km sized “Sentinel Landscapes”, stratified by the major Koeppen-Geiger climate zones of Africa, excluding the true deserts and some of the African countries which we were not allowed to work in at the time, due to security reasons.

Within each of the 60 Sentinel Landscapes AfSIS field teams sampled 16, 1 km2 “Sampling Clusters” (1 km2 circular areas) with 10, 1000 m2, randomly located circular “Sampling Plots”.

Topsoil (0-20 cm) and subsoil (20-50 cm) samples were subsequently recovered by physically mixing core subsamples from 4 locations within each Sampling Plot. Hence the intent was to obtain a representative multilevel/multistage sample consisting of:

- 60 Sentinel Landscapes
- 16 Sampling Clusters per Sentinel Landscape
- 10 Sampling Plots per Sampling Cluster
- 2 composite Soil Samples (topsoil & subsoil) per Sampling Plot

Multiply those numbers and you obtain the intended number of composite soil samples (19,200) that were to be collected in the field over a 4-year period between 2009-2012.

To achieve this target, we pre-generated appropriately randomized GPS coordinates for every Sampling Plot and, AfSIS field teams then navigated to (most) of those spots on the map to collect samples (insert n/N Sampling Plots).

As might be expected with an exercise of this magnitude, the actual total number of soil samples in this dataset is somewhat smaller than intended, as some locations were either completely inaccessible by 4WD vehicle and/or on foot or that had soil depth restrictions that prevented the field teams from recovering physical samples.

All physically recovered samples went into our lab (in Nairobi) to be characterized with the MIR spectral measurements that you are currently using. The potential spatial predictors, which cover the entire African continent (and beyond), were derived from NASA remote sensing data missions.

A 10% subsample of all the soils that were measured with the MIR method, subsequently went on to be characterized with more reference measurements.

“Reference measurements” are much more expensive (potentially hundreds of U$ per sample) and time-consuming. The other 90% of samples that were not characterized with “reference” methods have been physically archived, so that we can potentially retrieve those for calibrating new analytical methods and/or validating old methods.

What is posted for this Kaggle is the complete spectral + reference dataset that we have currently, subject to the sampling procedures described above. The training and test data have been split

along Sentinel Landscape levels because we are primarily interested in predicting soil properties at new Sentinel Landscapes.