Google Earth Engine Data Driven Sampling Scheme Breakdown

Oliver Hoffman

Colorado State University- Department of Soil and Crop Sciences

4/3/2024

**Step 1- Alteration of Dynamic Values:**

Code: A screen shot of a computer

Description automatically generated

Explanation:

The first chunk here is the customizable foundation of the code and allows for easy adjustment. The adjustable parameters are the number of clusters (‘numClusers’), the number of sampling points (‘numPoints’), the target field or area (‘geometry’), the slope data name and location (‘SlopeData’), and the temporal frame (‘dateRanges’) for Sentinel-2 image collection. By setting the number of clusters and points, the user can tune the algorithm for density and variability of sampling within a given area. The ‘geometry’ variable should be replaced with the name of your polygon added to the Earth Engine map. The ‘SlopeData’ is imported via ArcGIS Pro and ESRI should be cited for use of this data (Terrain: Slope Map). The ‘dateRanges’ specifies the periods from which to collect Sentinel-2 imagery, in the default case, for growing season over the last 5 years. Together, these dynamic variables allow for an adaptable code that can be tailored to a specific field or fields.

**Step 2- Bringing Down Requested Data:**

Code: A screenshot of a computer code

Description automatically generated

Explanation:

This step of the code processes and integrates a number of datasets for a more comprehensive soil and environmental analysis. First harmonized Sentinel-2 images are imported, followed by a number of POLARIS datasets for specific soil properties. The goal with the base datasets is to address soil, water, and topography while supplementing this with higher spatial and temporal resolution remotely sensed data.

Initially, 10-meter resolution Sentinel-2 data from the European Space Agency is loaded in for the specified date range (Sentinel-2 ESA…). Following import, the data is filtered for cloudy images with greater than 10% cloud cover to reduce clouds impacting calculated values. All the images that pass through the filters are merged into a single image using the median values for each band in each pixel that is tailored to the area specified. The Sentinel-2 Harmonized dataset is already atmospherically corrected for a number of atmospheric properties that cause disruption in how light is measured by the satellite, increasing reliability of data acquired before calculations are applied. Using the bands from Sentinel-2, a Normalized Difference Vegetation Index (NDVI = (NIR – R) / (NIR + R)) and Normalized Difference Water Index (NDWI = (G – NIR) / (G + NIR)) are calculated on the “filtered” raw image. The product of this is a variable for NDVI and NDWI, where each pixel has a value calculated that will fall between -1 and 1 (Ndwi Normalized…) (Normalized Difference Vegetation…).

The POLARIS 30-meter resolution modeled dataset from the United States Geological Survey, has detailed layers on various soil properties (Chaney, et al. 2016). POLARIS uses SSURGO combined with “soil covariates” such as relief, parent material + age, land cover, and soil properties. The “DSMART” algorithm is applied to these factors to produce the 30m pixels. Authors of this dataset mention that there is “variable performance” across the US, and specifically limitations in areas with consistent environmental characteristics. Using this dataset, bulk density, clay percentage, organic matter, hydraulic conductivity, and residual soil moisture are imported. An adjustment in organic matter values addresses the model limitations by correcting negative values to zero.

Integrating high-resolution slope data from ArcGIS Pro into Earth Engine enhances the reliability of analysis due to the size of typical agricultural fields. However, this data is owned by USGS and ESRI, limiting the availability of free data. If you do not have an ESRI license, the 10m NED data can be imported instead. The dataset in ArcGIS Pro is named “Terrain: Slope Map”, created by ESRI using a combination of a couple datasets from around the world that use LiDAR and DEM’s. For most of the USA, 1m NRCS and some 0.5m LiDAR data is used.

**Step 3- Primary Statistics Pre-Clustering for Reference:**

Code:

A screen shot of a computer code

Description automatically generated

Explanation:

The goal of this code is to display all the important data and information necessary for full understanding of the process. This code chunk calculates key statistics and generates histograms for each dataset, organized by their spatial scale. This step allows the user to understand the data distribution, central tendencies, variability, and range prior to clustering and data manipulation. The statistics and histograms are displayed in the console, therefore, making data accessible and easy to see.

**Step 4- K Mean Clustering:**

Code:

A screenshot of a computer code

Description automatically generated

Explanation:

This section of the code uses K-means clustering to segment the dataset into groups or clusters of similarity based on the stack of datasets utilizing the Weka K Means algorithm. The idea and general direction for this whole code and K-means clustering was based off the Stratifi v3.2 Earth Engine app created by Skidmore College (Bettigole, et al. 2023). To ensure data consistency and avoid overrepresentation of certain datasets, each layer is resampled to a consistent 10 meters using bilinear interpolation. After resampling, the layers are combined into a stacked multi-band image, aligning the pixels across the specified area. This consistency allows for the algorithm to effectively identify, and group based on similarity. The clustering process is driven by the ‘numClusters’ variable and segments the specified area into a user-defined number of clusters. This unsupervised classification results in a new image where pixels are assigned a cluster ID which allows for spatial analysis and visualization for each cluster or group.

**Step 5- Determine Feature Importance Using Random Forest:**

Code:

A screenshot of a computer code

Description automatically generated

Explanation:

To complement the K-means Clustering approach in step 4, a Random Forest classifier is conducted to assess feature importance, given that K-means lacks this capability. This process first labels each pixel with its corresponding cluster assignment and creates a uniform dataset that integrates cluster labels with imported datasets. This model is trained on a subsample of the dataset to increase efficiency without sacrificing representativeness. The model consists of 50 trees and is then trained to predict cluster label based on the input variable which facilitates an analysis of how each value contributes to the clustering scheme. The result is a feature importance list displayed in the console. That rank of feature importance, despite using a different model, represents an indicative insight as to what datasets may have been the driving factors in clustering.

**Step 6- Proportionally Stratified Samples:**

Code:

A screen shot of a computer code

Description automatically generated

Explanation:

This section of code calculates the spatial dynamics of the clustering along with taking the user-defined number of samples into account to assign points to the clusters that represent the spatial variability. The goal is to have each set of points be proportionally representative of the field’s variability. The process sums the area of pixels within each cluster, followed by calculating the total area of all clusters. From here the proportion of each cluster are to the total is determined which guides the distribution of the total number of sampling points. The result is that the number of points in each cluster is representative of the relative size of that cluster.

**Step 7- Producing Statistics for each Cluster on the Organic Matter Data + Area:**

Code:

A screenshot of a computer code

Description automatically generated

Explanation:

This step, like step 3, increases information provided to enhance user understanding and information. The raster clusters are converted to vector polygon clusters which then allows for analysis. First, the area of each cluster is calculated and printed to the console. Next, the resampled 10-meter organic matter dataset is overlayed and clipped to each cluster. From here, the key statistics can be calculated and displayed to the console. This allows for the user to understand organic matter dynamics in each cluster which depending on the goals of the project can be altered to be a different soil property.

**Step 8- Creating Visual Map:**

Code:

A screenshot of a computer code

Description automatically generated

Explanation:

This section is where basic map properties are set, satellite map, color palettes, transparency, etc. Also, each map layer is added and set to where only the clusters and points are visible, however, you can click on layers and add all other layers as well (this helps reduce processing time).

**Step 9- Exporting All Layers:**

Code:

A screenshot of a computer code

Description automatically generated

Explanation:

For sharing purposes, this section creates export files of each layer. The spatial data can then be moved to ArcGIS Pro or another mapping software for further analysis or visualization.

**Citations:**

Bettigole, Charles, et al. “Optimizing sampling strategies for near-surface soil carbon inventory: One size doesn’t fit all.” *Soil Systems*, vol. 7, no. 1, 17 Mar. 2023, p. 27, https://doi.org/10.3390/soilsystems7010027.

Chaney, Nathaniel W., et al. “Polaris: A 30-meter probabilistic soil series map of the contiguous United States.” *Geoderma*, vol. 274, July 2016, pp. 54–67, https://doi.org/10.1016/j.geoderma.2016.03.025.

“Ndwi Normalized Difference Water Index.” *Sentinel Hub Custom Scripts*, Sentinel-Hub by Sinergise, custom-scripts.sentinel-hub.com/custom-scripts/sentinel-2/ndwi/. Accessed 3 Apr. 2024.

“Normalized Difference Vegetation Index.” *Sentinel Hub Custom Scripts*, Sentinel-Hub by Sinergise, custom-scripts.sentinel-hub.com/custom-scripts/sentinel-2/ndvi/#:~:text=It%20normalizes%20green%20leaf%20scattering,rock%2C%20sand%2C%20or%20snow. Accessed 3 Apr. 2024.

“Sentinel-2.” *ESA*, www.esa.int/Applications/Observing\_the\_Earth/Copernicus/Sentinel-2#:~:text=The%20Copernicus%20Sentinel%2D2%20mission,of%20our%20land%20and%20vegetation. Accessed 3 Apr. 2024.

“Terrain: Slope Map.” *Arcgis.Com*, ESRI, www.arcgis.com/home/item.html?id=a1ba14d09df14f42ad6ca3c4bcebf3b4. Accessed 3 Apr. 2024.