Project 2: Early Prediction of Corn Yields Across U.S. from Satellite Images (Google Earth Engines API)

1. Introduction

Growing up in a family whose business is primarily distribution of agricultural produce, it is always a challenge deciding when we will sell the product, and for how much as these ultimately depend on how much of the produce will be harvested at the end of the season. If there is a way to predict how much will be obtained at the end of the season, we would be able to make decision much easier. Previous studies were able to show that satellite images can be used to predict the area where each type of crop is planted [1]. This leaves the question of knowing the yields in those planted areas. To this end, this project aims to use data from several satellite images to predict the yields of a crop. We chose corn as an example crop in this study. The implication for this project is much more than just my family business of course, big businesses can use this model to optimize their price and inventory, government can prepare for food shortage, even farmers can be informed of appropriate selling price if they know the regional yields.

This project aims to tackle this data using a data-driven approach, particularly we hope to:

- Identify correlations between satellites images and crop yields.
- Build a regression models to predict yields from these images using data from year 2010-2015 as a training and yields in 2016 as a test set.
- Determine how early can we accurately predict the yields.

Solutions to all problems start with gathering data and seeing the big picture through big data analytics lens, here I queried images by 4 satellites for each time point from Google Earth Engine including 1) MODIS Terra Surface Reflectance, 2) MODIS Surface Temperature, 3) USDA-FAS Surface and Subsurface Moisture, and 4) USDA-NASS for masking (total of 146 GB). The ground truth annual yields were collected in a county-level from USDA QuickStats.

Topics that will be covered using these datasets include

1. Exploration of value distribution in each satellites in the area that is corn fields
2. Exploration of the correlation between these values to the corn yields
3. Feature engineering and image processing
4. Selection of deep regression models

Data source:
1. https://explorer.earthengine.google.com/#detail/MODIS%2F006%2FMOD09A1
2. https://explorer.earthengine.google.com/#detail/MODIS%2F006%2FMYD11A2
3. https://explorer.earthengine.google.com/#detail/NASA_USDA%2FHSL%2Fsoil_ moisture
4. https://explorer.earthengine.google.com/#detail/USDA%2FNASS%2FCDL
5. https://www.nass.usda.gov/Quick_Stats/Lite/index.php