Assessing the Value of Remotely Sensed Environmental Indicator Datasets on NDVI for Food Security in Kenya

Andrew Yen

Capstone Project for JHU MS in Geographic Information Systems - 430.800.81

Abstract

Kenya has been experiencing severe drought conditions in the last few years, threatening food security within the country. Food security for Kenyans living in the arid or semi-arid lands (ASALs) in the central and eastern portions of the country is particularly challenged by drought as grazing land and almost all agriculture is rainfed (Ducheyne et al, 2014; ReliefWeb, 2019; Shisanya et al, 2011).

The Normalized Difference Vegetation Index (NDVI) is a measure of vegetation vigor and common predictor for agricultural yield (Fernandez & Soria-Ruiz, 2017; Huang et al, 2014; Kant & Mishra, 2017; Lewis et al, 1998; Lopresti et al, 2014; Maselli & Rembold, 2001; Wu et al, 2015). Early warning of yield and drought conditions (possible in part by NDVI data) enable policy makers to mitigate negative impacts by directing resources to locations that need support (http://www.fao.org/emergencies/emergency-types/drought/en/). While NDVI can serve as an indication of plant health, many other remotely sensible factors contribute to successful yields, and can be tied to NDVI values (Al-Shehhi et al, 2011; Boke-Olen, 2018; Indeje et al, 2006; Shisanya et al, 2011; Wang et al, 2000).

NDVI values however, respond to other environmental variables differently depending on regional climates. NDVI values in ASALs have been found to be more sensitive to other variables such as rainfall and soil moisture (Indeje et al, 2006; Maselli & Rembold, 2001).

Statement of the Problem

This study aims to determine the strength of the relationship between NDVI and a number of other environmental variables: soil moisture, evapotranspiration, and rainfall.

The purpose of this analysis will be to explore regression between these values and how regression might be different based on local climates (highly productive agricultural zones, ASALs, and arid).

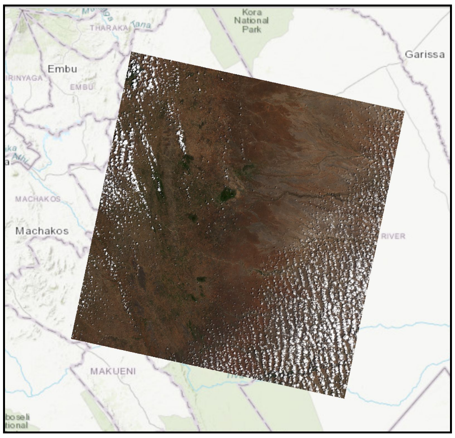
The Random Forest (RF) regression algorithm (seen in other studies relating NDVI to environmental variables: Park et al, 2019; Wang et al, 2016) will be used to measure these factors’ relationships to NDVI, the relative strength of these factors as predictors, as well as to compare differences between predicted and actual NDVI across the three study areas.

Study Area

This study considers three areas:

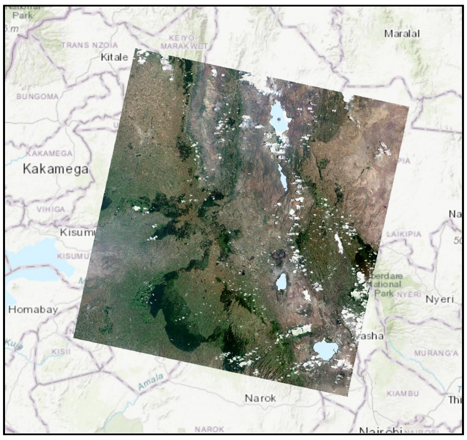
Landsat 8 - Path 167 Row 59 - Arid, Semi-Arid Lands Study Area
November 6th, 2018

An area in the ASAL region was selected because of ASALs particular sensitivity of NDVI to rainfall, soil moisture and evapotranspiration (Indeje et al, 2006; Maselli & Rembold, 2001), as well as rainfall's importance for rainfed crops and livestock health (Indeje et al, 2006; Shisanya et al, 2011; Speca, 2013).



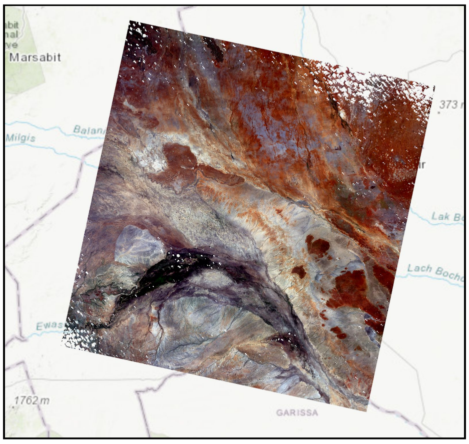
Landsat 8 - Path 169 Row 60 - Highly Productive Study Area
October 3rd, 2018

This area covers the southwestern 'Rift Valley', bordered by Lake Victoria on the west, Nairobi to the southeast and Mount Kenya National Park to the east. Kenya's Rift Valley Highlands are characterized by highly productive cropland, reliable rains, and relative food security (Speca, 2013; Indeje et al, 2006). These areas account for more than 70% of the country's population (Speca, 2013). One study that considered the relationship between NDVI and rainfall found that this region demonstrated low correlation coefficients between the two factors, as rainfall beyond what is needed for plant growth does not result in higher NDVI values (Indeje et al, 2006).



Landsat 8 - Path 167 Row 61 - Arid Study Area
December 24, 2018

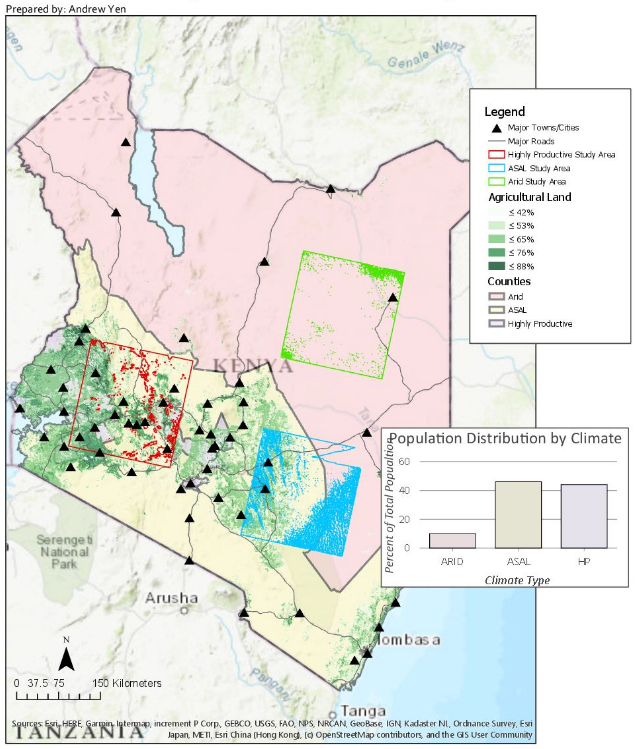
This area covers the arid region in Wajir and Isiolo counties, Kenya. Arid areas have almost no agricultural land, but may have rivers that support vegetation. This study area was considered as a basis for comparison against the ASAL and HP study areas, and is Kenya's third major climate type.



These study areas were selected based on the climate type distribution outlined in the map below (Red extent polygon = Highly Productive, Green = Arid, Blue = ASAL) :

Study Area Overview Map

Prepared by: Andrew Yen



Sources: Esri, HERE, Garmin, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), (c) OpenStreetMap contributors, and the GIS User Community

Kenya can be roughly divided into three unique climate types for the purposes of this study: Arid, Arid-Semi-Arid Lands (ASALs), and Highly Productive (HP). HP Areas are dominated by agricultural land, have higher population density, larger cities and enjoy relative food security. ASALs (the regions of particular interest for this study) rely on agriculture for subsistence farming and livestock feed. These areas have been food insecure in recent years due to severe drought and are home to the largest proportion of Kenya's population.

Study Area Overview Map

Climate-type data was retrieved from the Government of Kenya’s Ministry of Devolution and ASALs (2018). For a larger-detail map, please see the attached PDF file.

All images are taken from days during the “short rains” season (October-November-December) for two reasons:

1. This period is considered critical for food security in the ASAL, dictating grazing migration patterns and providing water for mainly rainfed crops in the region.
2. Because the OND season might be the best portion of the year for demonstrating the impact of precipitation on NDVI between the climate regions. NDVI values in drier regions (like in ASALs) are reported to have strong reactions to changes in rainfall and soil moisture.

Data

This study considers three Landsat-8 images (as described above) acquired from the USGS EarthExplorer data portal (https://earthexplorer.usgs.gov/). NDVI images are calculated from the NIR and Red bands of each image. The normalized difference between the Near-Infrared and Red reflection leverages the spectral characteristics of plant cellular biology, where healthy plants strongly reflect NIR light and weaker/stressed plants (or plants that die-off seasonally) more strongly reflect red light.

Other environmental datasets to be related with NDVI include:

- Precipitation data: From UCSB's CHIRPS program via the Climate Engine (https://clim-engine.appspot.com/). This data comes in 1/20 degree resolution and represents the total rainfall across Kenya from October 1st, 2018 to January 1st, 2019, covering the entirety of the short-rains season.

- Soil Moisture: was acquired from the National Snow and Ice Data Center data portal which serves SMAP passive radiometer data. The "Soil_Moisture_Retrieval_Data_1km_soil_moisture_1km" subset was extracted from the Level 2 Soil Moisture HDF product, L2_SM_SP. (Information on data source: https://nsidc.org/sites/nsidc.org/files/technical-references/SMAP%20L2_SM_SP%20PSD_20180531.pdf, page 50)

- Evapotranspiration Anomaly: was measured using the USGS MODIS Simplified Surface Energy Balance model Dekadal dataset, which derives ET from thermal MODIS images. The image is at 1/96 degree spatial resolution and was averaged over the 2018 short-rains season (in line with precipitation data described above). https://clim-engine.appspot.com/

Data for study area reference maps will include:

- IIASA-IFPRI Cropland area maps: available from https://application.geo-wiki.org/Application/index.php
- Kenya Population Distribution: from the Kenya Open Data Initiative (KODI) available on the ArcGIS Online Open Data Hub, https://hub.arcgis.com/datasets/68cb965bb2a847c0a27af771cf46064f_0?selectedAttribute=UrbanP
- Climate type distribution from the Government of Kenya Ministry of Devolution and ASALs, http://www.devolutionasals.go.ke/county-information/#79b4edb80adb3e6af
- Major Cities and Major Roads layers were also available through the KODI portal.

Methods

The raster and vector data processing for this project was done using ArcGIS Pro 2.2.0.

Preprocessing Steps:

- Calculate NDVI rasters from Landsat-8 Images
    - remove large bodies of water from images to mitigate the influence of water on the NDVI distribution across the image.
    - NDVI values were calculated by taking the normalized difference of two Landsat 8's NIR image (Band 5) and Red image (Band 4):

$$NDVI = \frac{(NIR - Red)}{(NIR + Red)}$$

- Average, and then majority filter SMAP rasters when data from different dates needed to be used. Only data from days before the Landsat imagery was captured was used. Changes in any of the environmental variables used to calculate the model that occurred after the imagery was captured would have caused inaccuracies when modelling NDVI.
- Create two randomly-distributed point feature classes for each study area, one for storing training data (from the acquired raster data), and the other for holding predictions.
- Values from NDVI, soil moisture, evapotranspiration and rainfall rasters were extracted to the training point feature classes in each study area.

Classification Steps:

- This analysis used the "Forest-based Classification and Regression" geoprocessing tool in ArcGIS Pro (tool documentation: https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/forestbasedclassificationregression.htm).
- Train classifier to test the performance of different variable combinations for the model and record the results. The training data can be compared with a validation dataset to determine the effect size of the model when targeting data outside of the training set.
- Test training with different parameterization schemes to find the best possible fit with the training data. Iterations that featured larger numbers of samples and larger ensemble of trees only had a very marginal improvement in model fit at the cost of additional processing time.
- After recording the statistical results of training and validation, use the same random seed generated for training to predict NDVI values to a raster.

Study Area Contextualization & GIS methods:

- Landsat coverage areas were cross-referenced with the Agricultural Areas feature class to make sure they qualified as representative areas for each of the climate types. This was accomplished by selecting agricultural area polygons within the study area polygons generated in the cloud removal step, summing the area of all agricultural area polygons, and then finding the percentage of agricultural area within the study area by taking the normalized difference between the agricultural area and the study area multiplied by 100. The proportion of agricultural area in each study area was consistent with climatological definitions.

Study Area	Agricultural Area
Arid	0.04%
ASAL	30.0%
Highly Productive	51.6%

- To determine population distribution across climate regions, county-level population data was summarized over the climate type areas (also at the county-level); the total population of counties with the same climate type were expressed as a bar chart. Major cities points were overlaid on the map to reinforce population summary analysis; the majority of cities are co-located with ASALs and agriculture, and there are only a small number of cities in arid areas. The population distribution analysis is a useful way of showing the spatial variability in the drought experience as illustrated by remote sensing indicators.

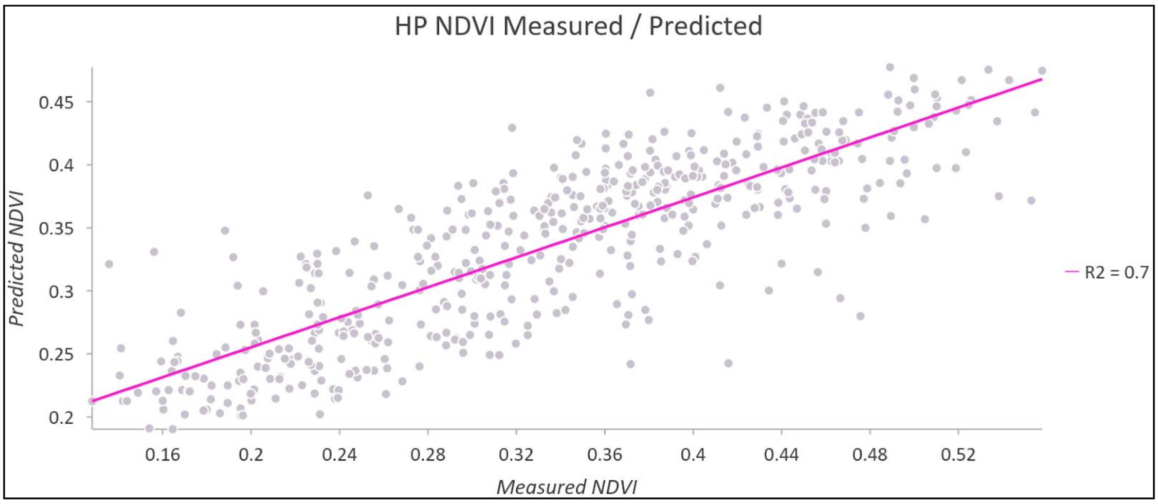
Results

Expected Results

Based on my review of the relevant literature, I expected NDVI to be closely related to all of the selected environmental variables. And I anticipated weaker correlations between the chosen variables and NDVI in the highly productive study area compared to the ASAL study area. It follows then that the predicted NDVI surface for the ASAL study area will more closely resemble the actual NDVI surface, and the percent difference in NDVI values across the images will be lower than for the Rift Valley study area.

Actual Results

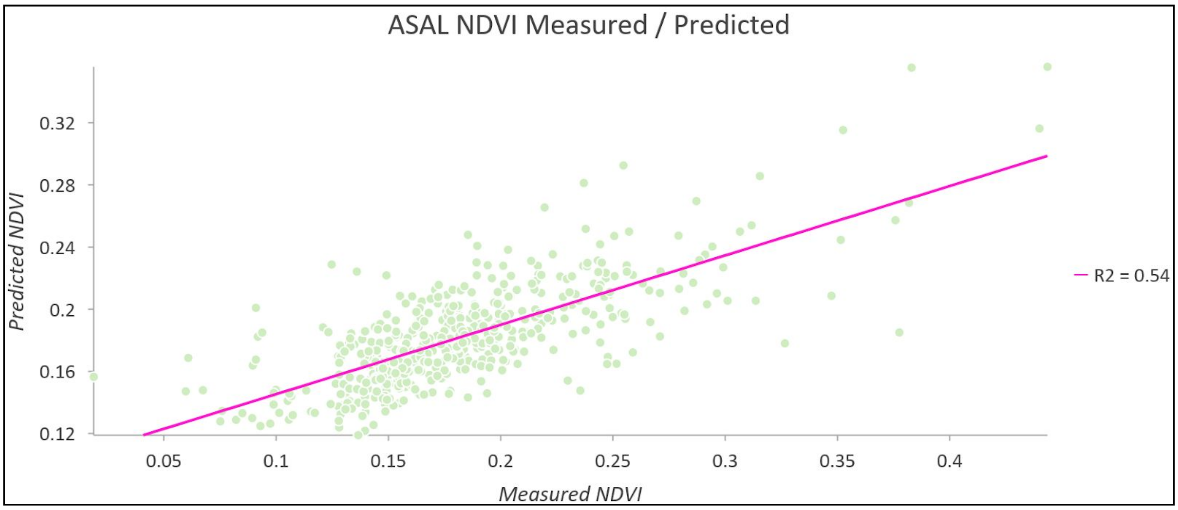
Contrary to the expected results, the comparison of RF regression predictions against measured NDVI found that predictions were the most accurate for the HP study area ($R^2 = 0.70$), and were progressively weaker in the ASAL ($R^2 = 0.54$) and arid ($R^2 = 0.32$) study areas. These values indicate the relative strength of the explanatory variables as predictors for NDVI.



Scatterplot comparing Measured and Predicted NDVI values.

HP regression equation:

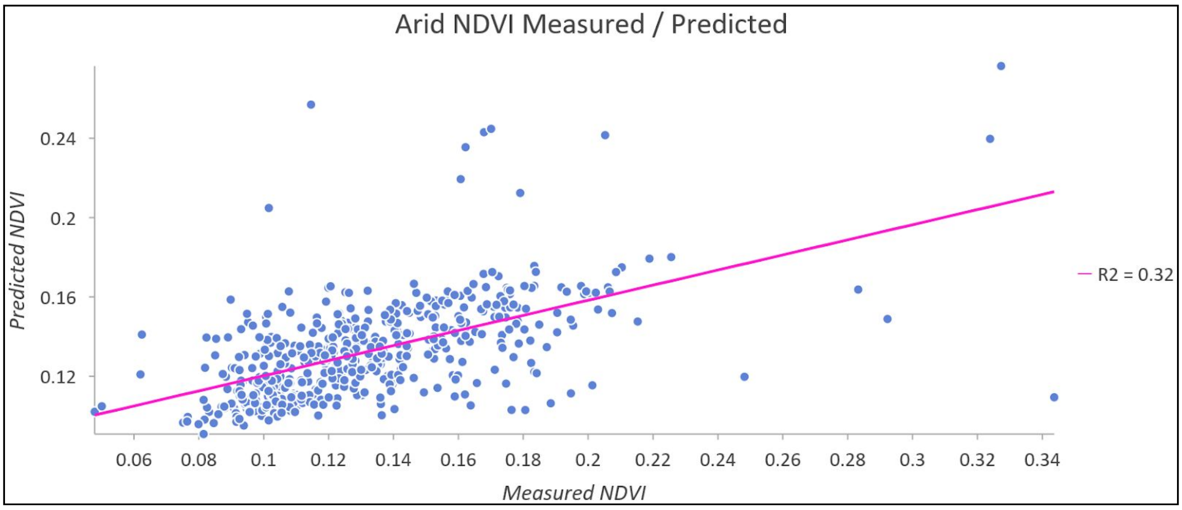
$y = 0.13652 + 0.59410x$



Scatterplot comparing Measured and Predicted NDVI values in ASAL study area

ASAL regression equation:

$y = 0.10073 + 0.44628x$

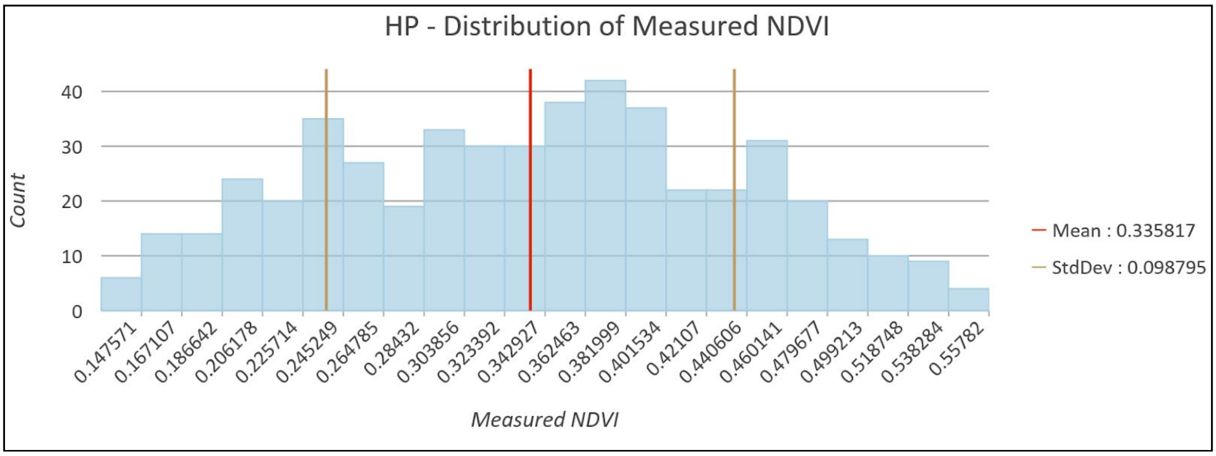


Scatterplot comparing Measured and Predicted NDVI values in Arid study area

Arid regression equation:

$y = 0.08215 + 0.38090x$

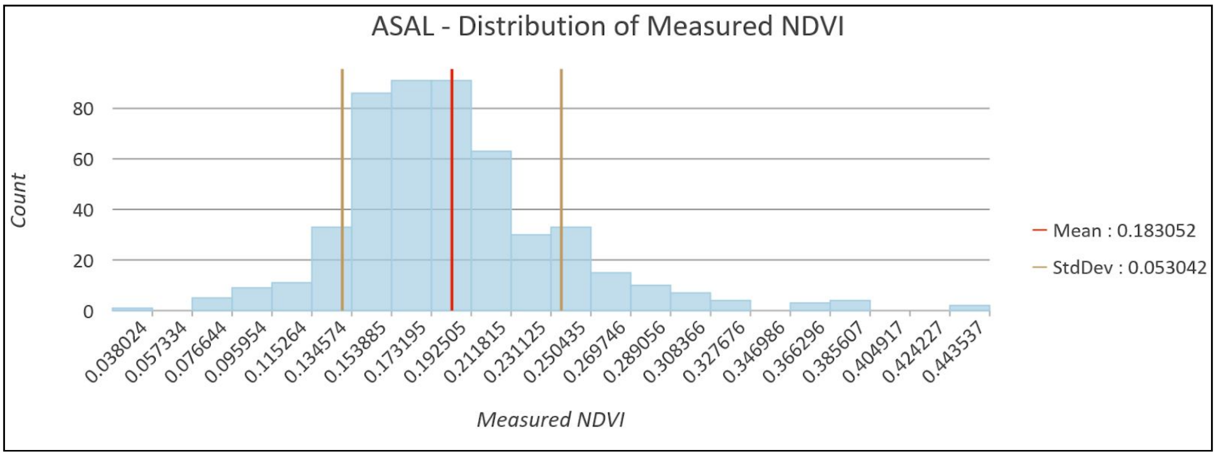
This trend in correlation strength across study areas may have been due to how strongly clustered NDVI values were in the sample data. ASAL and arid study areas showed much stronger kurtosis values compared to HP. Less variation of NDVI might have made it harder for the RF regression model to differentiate NDVI in predictions. The arid comparison shows some notable instances where NDVI was severely over and under predicted. This might have been due to the change in spatial resolution from the Landsat NDVI to NDVI prediction raster. Because bodies of water (the most significant contributors to higher NDVI values) were sparse, pixels in the prediction raster near water bodies may have been conflated with much drier pixels with lower NDVI values.



Histogram showing the distribution of NDVI values in the HP study area

Kurtosis: *2.15732* [Wide distribution]

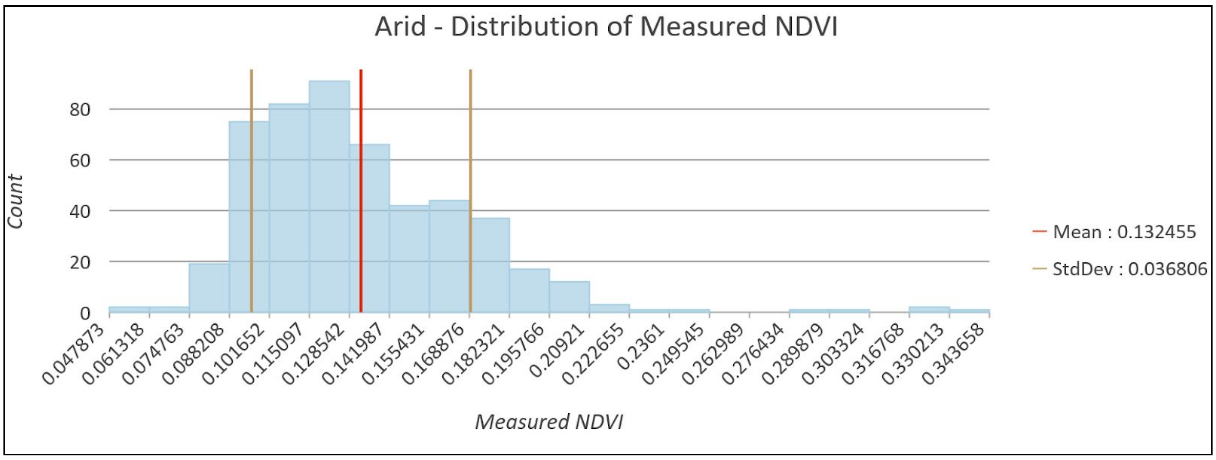
Skewness: - *0.00618* [Nearly Symmetrical]



Histogram showing the distribution of NDVI values in the ASAL study area

Kurtosis: *6.34767* [Tight distribution]

Skewness: *1.17583* [Skewed Left]



Histogram showing the distribution of NDVI values in the Arid study area

Kurtosis: *8.12913* [Tight distribution]

Skewness: *1.51610* [Skewed Left]

Gini Coefficient / Importance Scores

The RF regression also outputs importance scores for each explanatory variable to help determine their relative usefulness for the model. This score is based on the probability of a correct classification when using the respective explanatory variable to determine a branch in the tree. Relatively high importance scores indicate that an explanatory variable had a higher probability of predicting the correct value when fitting the model to the training data.

Large differences in importance scores may be an indication that certain explanatory variables are negatively impacting the model. In this study importance scores were relatively close to one another, with SMAP trailing slightly for all study areas (see table below). The exceptionally wide difference between SMAP and CHIRPS precipitation importance values for the HP study area (25% vs. 41%) may be a sign that soil moisture is a less decisive value when explaining variation in NDVI. Adding additional explanatory variables may further separate the importance of these two factors.

Relative Importance of Variables (Table):

	HP	**ASAL**	**ARID**
CHIRPS	41%	39%	38%
MODIS_ET	33%	37%	37%
SMAP	25%	24%	25%

Note: importance values are expressed here in terms of share of importance rather than Gini Coefficients, which change based on input training data, number of explanatory variables and other factors.

Model Performance - Training Data

The models all had comparable R^2 values (~ 0.880) when fitting to the training dataset — a good signal for the capacity of the model to explain the variation of data supplied to the model when training. Validation R^2 values, which measure the capacity of the model to explain variation on 10% of the reserved data, were all fairly high, ranging from 0.381 to 0.589. This conforms to expected correlations concerning training vs. validation data in general; training R^2 should be the highest as it is the only data that is “seen” by the model when it is being built. See the table below.

	HP	**ASAL**	**ARID**
Training r-squared	0.899	0.878	0.875
Training Standard Error	0.012	0.011	0.012
Validation r-squared	0.478	0.381	0.589
Validation Standard Error	0.069	0.054	0.05

Discussion

Advantages of the Random Forest Regression model

Each decision tree in the random forest is a high variance model that will result in widely different regression determinations when supplied with different subsets of the data. These inaccuracies are mitigated by “bagging” - **‘B’**ootstrapping and **‘AG’**gregating the results of many decision trees each considering random bootstrapped datasets.

In categorical classification models, aggregation is a “majority rules” approach where the most common categorical value across all decision trees is the final determination for the sample to predict.

For continuous values, the mean of all values at terminal nodes of the decision trees is the predicted value.

Potential Sources of Error

There are several sources of error that could potentially have caused NDVI predictions to be less accurate:

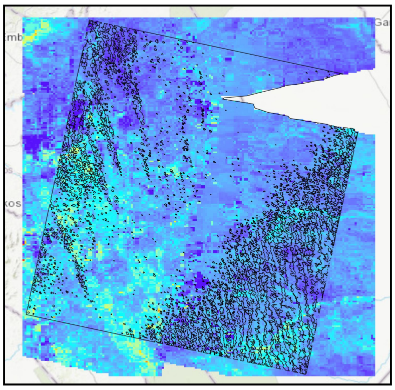
- *Cloud Cover*

Cloud cover, especially in imagery taken during the rainy season, makes it impossible to take full advantage of all of the pixel values in the scene. Data removed from the image because it was covered by clouds or cloud shadows might have helped better explain NDVI. The ASAL study in particular was impacted by cloud cover (~11%) and had a substantial number of pixels removed before RF regression.

- *SMAP mixed dates*

Soil moisture data needed to be aggregated for the ASAL and arid study areas due to limitations in geographic coverage. Only data captured before the Landsat imagery acquisition date was used, as the measured NDVI values could not possibly be caused by conditions in the future. No more than two different dates were used to generalize a soil moisture surface for each study area.

The available SMAP data did not permit complete coverage of the ASAL Landsat data (see image below).



SMAP images could not cover the entire study area; the northwest corner of the study area needed to be excluded from sample data collection.

- *Landsat mixed dates*

The Landsat data needed to be taken from different parts of the season based on geographic coverage availability as well as cloud cover. Dates ranged from October 3rd (HP), November 6th (ASAL) and December 24th, 2018 (arid). While the different dates wouldn't affect the regression model as the data were all contained in each respective study area, it does mean that comparisons between regression model performance across study areas may be impacted.

- *SMAP not aggregated and summarized*

Unlike the rainfall and evapotranspiration data, there was no option to acquire SMAP data aggregated over the season. Discrete dates needed to be aggregated to approximate soil moisture for the area. This specificity may be one explanation for why soil moisture was the explanatory variable variable with the lowest importance score across all of the RF models; aggregating values over longer periods smooth out the impacts of isolated events (e.g. short-lived heavy rains in an otherwise dry season).

- *Misspecification - Missing explanatory variables*

More variables (like PET, temperature, agricultural land, etc.) might have explained more of the variation in NDVI resulted in more accurate regression models. The inclusion of more explanatory variables would also help make better determinations on variable importance. Greater numbers of explanatory variables will result in a wider distribution of importance scores and enable easier differentiation between high/low importance score variables.

- *Decreased spatial resolution for output prediction images*

I did not have adequate processing power to produce NDVI predictions that matched the spatial resolution of Landsat Images (30 m) and was required to predict to cells that were the minimum of the explanatory raster inputs (~ 1079 m). Matching the resolution of the Landsat images would have been helpful for determining the accuracy predictions for scales as small as individual farm plots. The output prediction surface is far too coarse to make localized accuracy determinations for anything smaller than the county level.

Conclusions

In this study, I demonstrated the predictive power of three environmental variables (rainfall, evapotranspiration, and soil moisture) for NDVI. The relationship between environmental variables and NDVI is important because NDVI is often used to project crop yield in countries that struggle with food security. Based on the results of the RF regression models, the environmental variables used in this study were better able to predict NDVI in high production areas in Kenya, and were less useful predictors in the food insecure ASAL and arid areas. These findings are contrary to the expected results.

In future work, I would attempt to see what other environmental variables might be leveraged to predict NDVI, and conduct the same RF regression across several years. Multi-temporal data would likely help determine how well NDVI could be predicted by other remotely sensed environmental data.

References

Al-Shehhi M.R., Saffarini, R., Farhat, A., Al-Meqbali, N.K., Ghedira, H. (2011). Evaluating the Effect of Soil Moisture, Surface Temperature, and Humidity Variations on MODIS-derived NDVI Values. *2011 IEEE International Geoscience and Remote Sensing Symposium*, Vancouver, BC, 2011, pp. 3160-3163. doi: 10.1109/IGARSS.2011.6049889

Boke-Olen, N., Ardö, J., Eklundh, L., Holst, T., Lehsten, V. (2018). Remotely sensed soil moisture to estimate savannah NDVI. *PLoS ONE, 13(7): e0200328.* https://doi.org/10.1371/journal.pone.0200328

Ducheyne, E., Tack, W., Hendrickx, G. (2014). Remote sensing: a key tool for monitoring food resources in a changing world. Retrieved from: http://www.kaowarsom.be/documents/Conferences/DUCHEYNE.pdf

FAO (2019). Drought: FAO in Emergencies. Retrieved from: http://www.fao.org/emergencies/emergency-types/drought/en/

Fernandez, Y. & Soria-Ruiz, J. (2017). Maize crop yield estimation with remote sensing and empirical models. doi: 10.1109/IGARSS.2017.8127638

Huang, J., Dai, Q., Wang, H., Han, D. (2014). Empirical Regression Model Using NDVI, Meteorological Factors For Estimation of Wheat Yield in Yunnan, China. *CUNY Academic Works.* Retrieved from: http://academicworks.cuny.edu/cc_conf_hic/5

Huang, J., Wang, H., Dai, Q., Han, D. (2014). Analysis of NDVI Data for Crop Identification and Yield Estimation. *IEEE Journal of Selected Topics In Applied Earth Observations And Remote Sensing, 7(11).* Retrieved from: https://ieeexplore.ieee.org/document/6871305

Indeje, M. & Ward, M. (2006). Predictability of the Normalized Difference Vegetation Index in Kenya and Potential Applications as an Indicator of Rift Valley Fever Outbreaks in the Greater Horn of Africa. *International Research Institute for Climate and Society.* https://doi.org/10.1175/JCLI3708.1

Kant, C. & Mishra, M. (2017). Crop Yield Estimation Based on Landsat-NDVI a Case Study of Sitapur District, Uttar Pradesh, India. *International Journal of Applied Remote Sensing and GIS 4(1&2).* Retrieved from: http://gssjournals.org/Research_Paper/Volume_4_Issue_1_2_June_Dec-2017_IJARSG/3_Manisha_Mishra-1.pdf

Lewis, J., Rowland, J., Nadeau, A. (1998). Estimating maize production in Kenya using NDVI: Some statistical considerations, *International Journal of Remote Sensing, 19(13), 2609-2617.* https://doi.org/10.1080/014311698214677

Lopresti, M., Di Bella, C., Degioanni, A. (2015). Relationship between MODIS-NDVI data and wheat yield: A case study in Northern Buenos Aires province, Argentina. *Information Processing in Agriculture 2(2), 73-84.* https://doi.org/10.1016/j.inpa.2015.06.001

Maselli, F. & Rembold, F. (2001). Analysis of GAC NDVI Data for Cropland Identification and Yield Forecasting in Mediterranean African Countries. *Photogrammetric Engineering and Remote Sensing, 67(5), 593-602.* Retrieved from: https://www.researchgate.net/publication/279965255_Analysis_of_GAC_NDVI_Data_for_Cropland_Identification_and_Yield_Forecasting_in_Mediterranean_African_Countries

Ministry of Devolution and ASALs (Government of Kenya). (2018). ASALs Categorization. Retreived from: http://www.devolutionasals.go.ke/county-information/#79b4edb80adb3e6af

Park, H., Kim, K., Lee, D. (2019). Prediction of Severe Drought Area Based on Random Forest: Using Satellite Image and Topography Data. *Water 11(705).* doi:10.3390/w11040705

ReliefWeb (2019). Kenya: Drought - 2014-2019. Retrieved from: https://reliefweb.int/disaster/dr-2014-000131-ken

Shisanya, C. & Recha, C. (2011). Rainfall Variability and its Impact on Normalized Difference Vegetation Index in Arid and Semi-Arid Lands of Kenya. *International Journal of Geosciences 2, 36-47.* doi: 10.4236/ijg.2011.21004

Speca, A. (2013). Kenya Food Security Brief. *USAID FEWS Net.* Retrieved from: http://fews.net/sites/default/files/documents/reports/Kenya_Food%20Security_In_Brief_2013_final_0.pdf

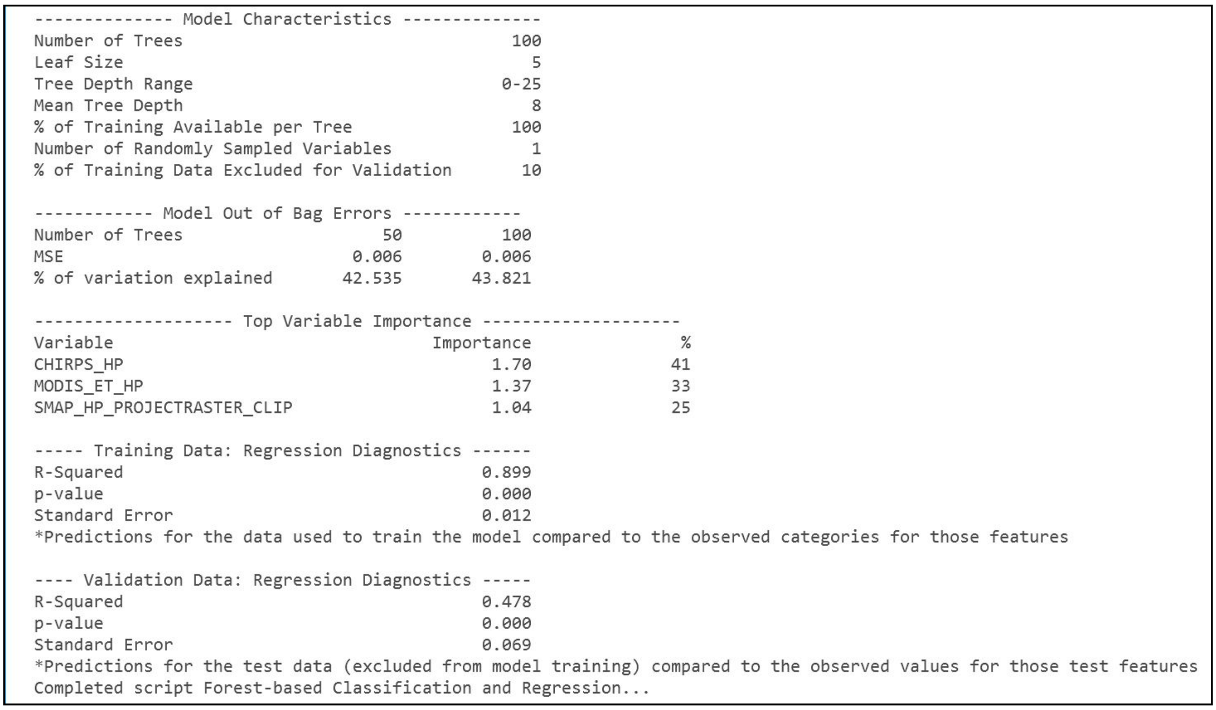
Wang, J., Price, K., Rich, P. (2001). Spatial patterns of NDVI in response to precipitation and temperature in the central Great Plains. *International Journal Remote Sensing 22(18), 3827-3844.* doi: 10.1080/01431160010007033

Wang L., Zhou, X., Zhu, X., Dong, Z., Guo, W. (2016). Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. *The Crop Journal 4(3), 212-219.* https://doi.org/10.1016/j.cj.2016.01.008

Wu, B., Gommes, R., Zhang, M., Zeng, H., Yan, N., Zou, W., Zheng, Y., Zhang, N., Chang, S., Xing, Q., Heijden, A. (2015). Global Crop Monitoring: A Satellite-Based Hierarchical Approach. *Remote Sensing 7(4), 3907-3933.* https://doi.org/10.3390/rs70403907

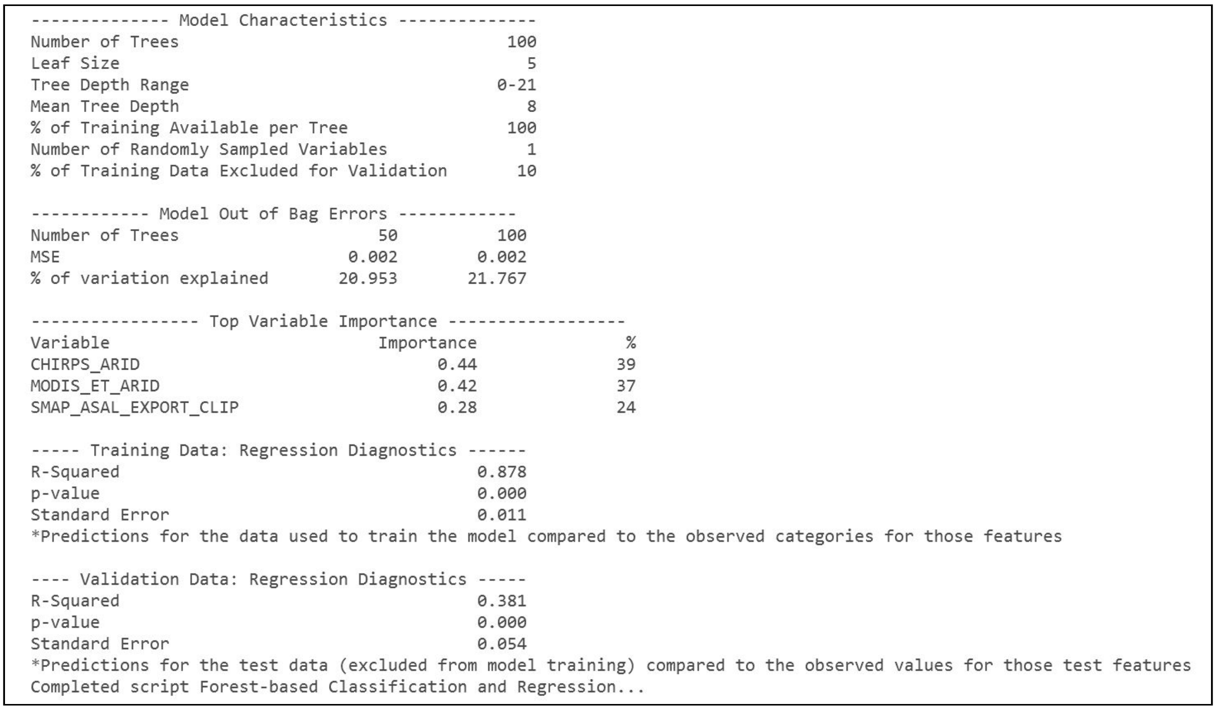
Appendix

HP Regression Report:



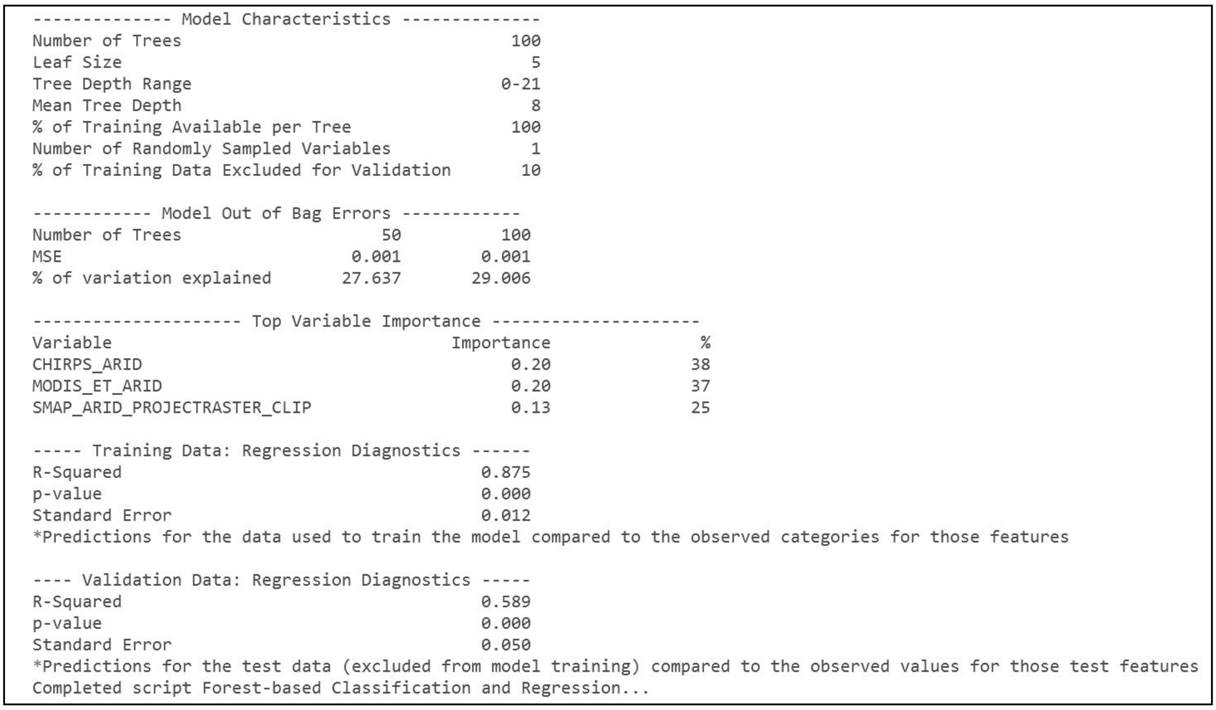
```
-------------- Model Characteristics --------------
Number of Trees                                 100
Leaf Size                                         5
Tree Depth Range                               0-25
Mean Tree Depth                                   8
% of Training Available per Tree                100
Number of Randomly Sampled Variables              1
% of Training Data Excluded for Validation       10

------------ Model Out of Bag Errors ------------
Number of Trees                    50          100
MSE                             0.006        0.006
% of variation explained       42.535       43.821

-------------------- Top Variable Importance --------------------
Variable                                Importance               %
CHIRPS_HP                                     1.70              41
MODIS_ET_HP                                   1.37              33
SMAP_HP_PROJECTRASTER_CLIP                    1.04              25

----- Training Data: Regression Diagnostics ------
R-Squared                                    0.899
p-value                                      0.000
Standard Error                               0.012
*Predictions for the data used to train the model compared to the observed categories for those features

---- Validation Data: Regression Diagnostics -----
R-Squared                                    0.478
p-value                                      0.000
Standard Error                               0.069
*Predictions for the test data (excluded from model training) compared to the observed values for those test features
Completed script Forest-based Classification and Regression...
```

ASAL Regression Report:



```
-------------- Model Characteristics --------------
Number of Trees                                 100
Leaf Size                                         5
Tree Depth Range                               0-21
Mean Tree Depth                                   8
% of Training Available per Tree                100
Number of Randomly Sampled Variables              1
% of Training Data Excluded for Validation       10

------------ Model Out of Bag Errors ------------
Number of Trees                    50          100
MSE                             0.002        0.002
% of variation explained       20.953       21.767

----------------- Top Variable Importance ------------------
Variable                           Importance               %
CHIRPS_ARID                              0.44              39
MODIS_ET_ARID                            0.42              37
SMAP_ASAL_EXPORT_CLIP                    0.28              24

----- Training Data: Regression Diagnostics ------
R-Squared                                    0.878
p-value                                      0.000
Standard Error                               0.011
*Predictions for the data used to train the model compared to the observed categories for those features

---- Validation Data: Regression Diagnostics -----
R-Squared                                    0.381
p-value                                      0.000
Standard Error                               0.054
*Predictions for the test data (excluded from model training) compared to the observed values for those test features
Completed script Forest-based Classification and Regression...
```

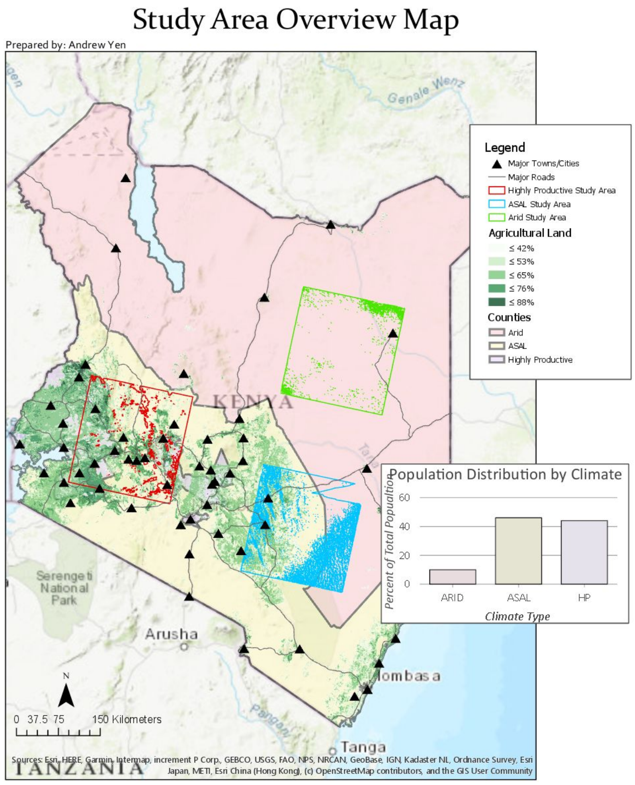
Arid Regression Report:



```
-------------- Model Characteristics --------------
Number of Trees                                   100
Leaf Size                                           5
Tree Depth Range                                 0-21
Mean Tree Depth                                     8
% of Training Available per Tree                  100
Number of Randomly Sampled Variables                1
% of Training Data Excluded for Validation         10

------------ Model Out of Bag Errors ------------
Number of Trees                      50          100
MSE                               0.001        0.001
% of variation explained         27.637       29.006

--------------------- Top Variable Importance ---------------------
Variable                                    Importance                %
CHIRPS_ARID                                       0.20               38
MODIS_ET_ARID                                     0.20               37
SMAP_ARID_PROJECTRASTER_CLIP                      0.13               25

----- Training Data: Regression Diagnostics ------
R-Squared                                      0.875
p-value                                        0.000
Standard Error                                 0.012
*Predictions for the data used to train the model compared to the observed categories for those features

---- Validation Data: Regression Diagnostics -----
R-Squared                                      0.589
p-value                                        0.000
Standard Error                                 0.050
*Predictions for the test data (excluded from model training) compared to the observed values for those test features
Completed script Forest-based Classification and Regression...
```

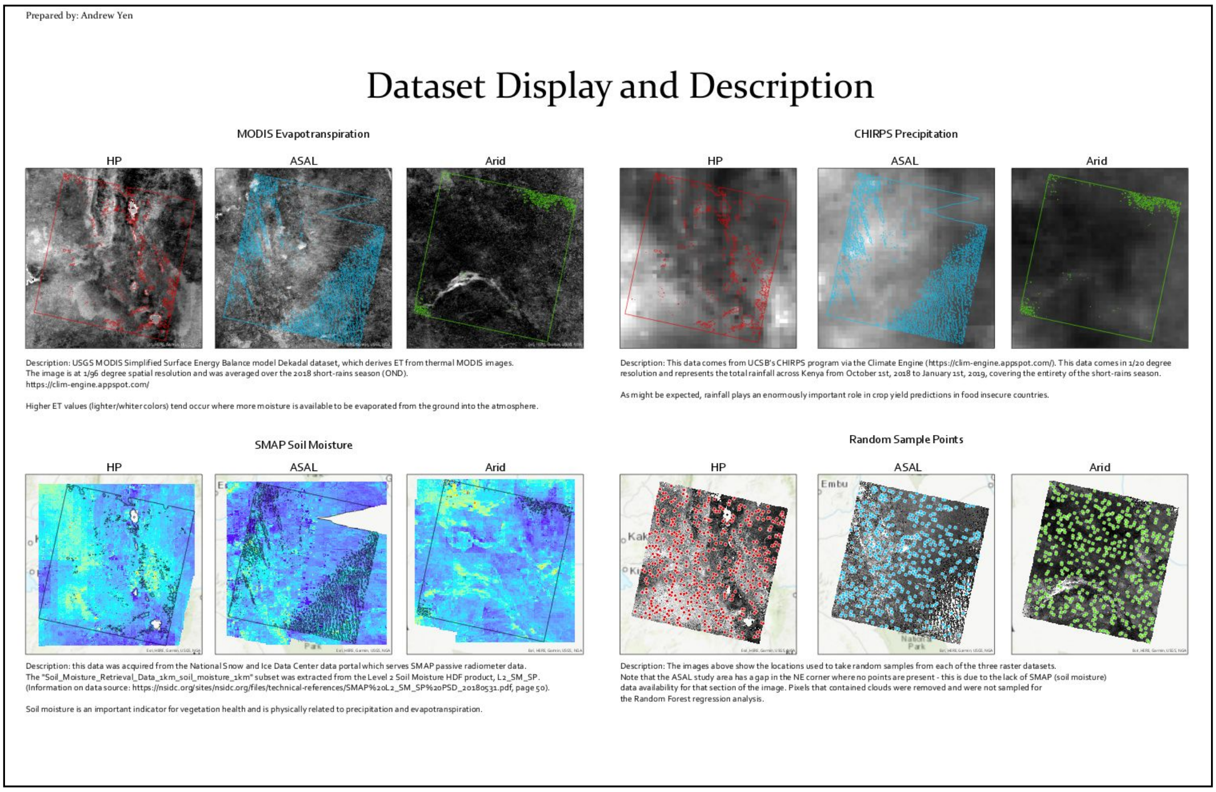
For high-resolution viewing, refer to the pdf versions

Study Area Overview Map:


Study Area Overview Map
Prepared by: Andrew Yen
Genale Wenz
KENYA
Serengeti National Park
Arusha
Mombasa
Pangani
Tanga
TANZANIA
N
0 37.5 75 150 Kilometers
Legend
Major Towns/Cities
Major Roads
Highly Productive Study Area
ASAL Study Area
Arid Study Area
Agricultural Land
≤ 42%
≤ 53%
≤ 65%
≤ 76%
≤ 88%
Counties
Arid
ASAL
Highly Productive
Population Distribution by Climate
Percent of Total Population
60
40
20
0
ARID
ASAL
HP
Climate Type
Sources: Esri, HERE, Garmin, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), (c) OpenStreetMap contributors, and the GIS User Community

Kenya can be roughly divided into three unique climate types for the purposes of this study: Arid, Arid-Semi-Arid Lands (ASALs), and Highly Productive (HP). HP Areas are dominated by agricultural land, have higher population density, larger cities and enjoy relative food security. ASALs (the regions of particular interest for this study) rely on agriculture for subsistence farming and livestock feed. These areas have been food insecure in recent years due to severe drought and are home to the largest proportion of Kenya's population.

Dataset Display Map:



Prepared by: Andrew Yen

Dataset Display and Description

Description: USGS MODIS Simplified Surface Energy Balance model Dekadal dataset, which derives ET from thermal MODIS images.
The image is at 1/96 degree spatial resolution and was averaged over the 2018 short-rains season (OND).
https://clim-engine.appspot.com/

Higher ET values (lighter/whiter colors) tend occur where more moisture is available to be evaporated from the ground into the atmosphere.

Description: This data comes from UCSB's CHIRPS program via the Climate Engine (https://clim-engine.appspot.com/). This data comes in 1/20 degree resolution and represents the total rainfall across Kenya from October 1st, 2018 to January 1st, 2019, covering the entirety of the short-rains season.

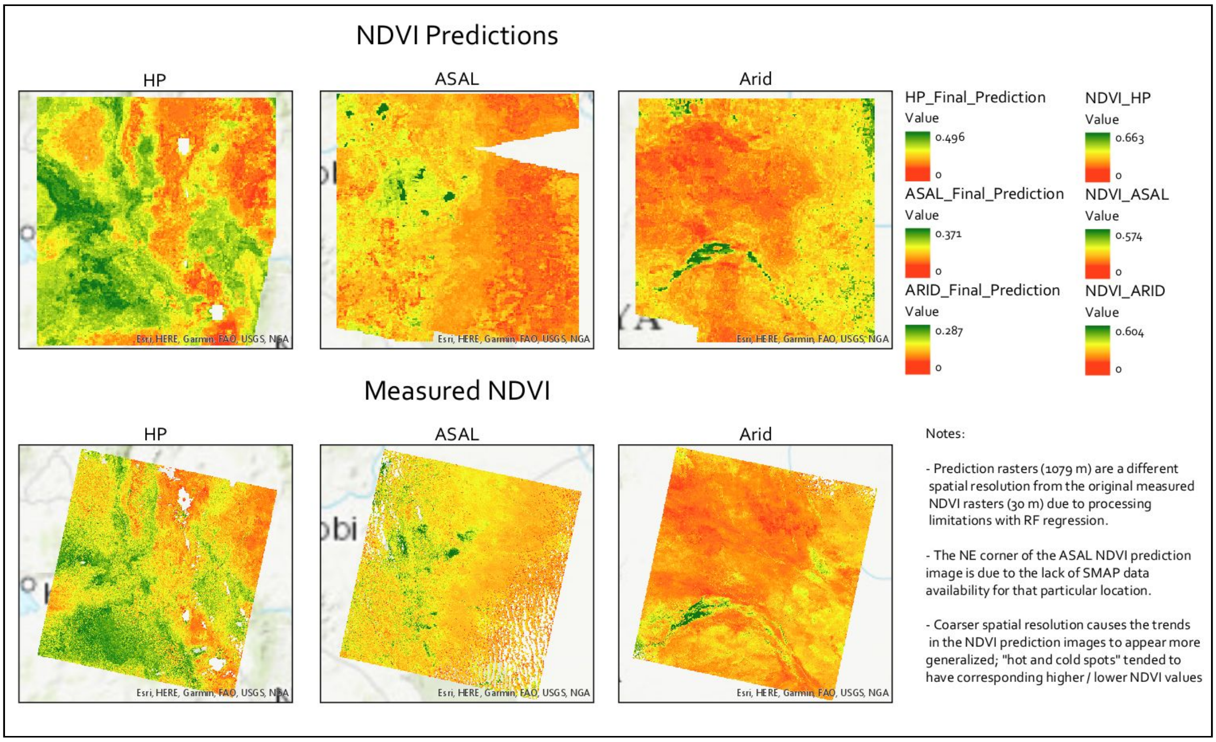
As might be expected, rainfall plays an enormously important role in crop yield predictions in food insecure countries.

Description: this data was acquired from the National Snow and Ice Data Center data portal which serves SMAP passive radiometer data.
The "Soil_Moisture_Retrieval_Data_1km_soil_moisture_1km" subset was extracted from the Level 2 Soil Moisture HDF product, L2_SM_SP.
(Information on data source: https://nsidc.org/sites/nsidc.org/files/technical-references/SMAP%20L2_SM_SP%20PSD_20180531.pdf, page 50).

Soil moisture is an important indicator for vegetation health and is physically related to precipitation and evapotranspiration.

Description: The images above show the locations used to take random samples from each of the three raster datasets.
Note that the ASAL study area has a gap in the NE corner where no points are present - this is due to the lack of SMAP (soil moisture) data availability for that section of the image. Pixels that contained clouds were removed and were not sampled for the Random Forest regression analysis.

NDVI Prediction vs. Measured NDVI Comparison Map:



Notes:

- Prediction rasters (1079 m) are a different spatial resolution from the original measured NDVI rasters (30 m) due to processing limitations with RF regression.

- The NE corner of the ASAL NDVI prediction image is due to the lack of SMAP data availability for that particular location.

- Coarser spatial resolution causes the trends in the NDVI prediction images to appear more generalized; "hot and cold spots" tended to have corresponding higher / lower NDVI values