Technical Documentation

UNIVERSITY OF CALIFORNIA
Santa Barbara

MEASURING AGRICULTURAL ADAPTATION TO CLIMATE CHANGE IN ZAMBIA USING SATELLITE IMAGERY AND MACHINE LEARNING

A Capstone Project submitted in partial satisfaction of the requirements for the degree of
Master of Environmental Data Science
for the
Bren School of Environmental Science & Management

by

Andrew Bartnik
Carlo Broderick
Gabrielle Smith
Hailey Veirs

Committee in charge:
Tamma Carleton
Naomi Tague
Ruth Oliver

June 2023

Technical Documentation Signature Page

MEASURING AGRICULTURAL ADAPTATION TO CLIMATE CHANGE IN ZAMBIA USING SATELLITE IMAGERY AND MACHINE LEARNING

The Bren School of Environmental Science & Management produces professionals with unrivaled training in environmental science and management who will devote their unique skills to the diagnosis, assessment, mitigation, prevention, and remedy of the environmental problems of today and the future. A guiding principle of the School is that the analysis of environmental problems requires quantitative training in more than one discipline and an awareness of the physical, biological, social, political, and economic consequences that arise from scientific or technological decisions.

The Capstone Project is required of all students in the Master of Environmental Data Science (MEDS) Program. The project is a six-month-long activity in which small groups of students contribute to data science practices, products or analyses that address a challenge or need related to a specific environmental issue. This MEDS Capstone Project Technical Documentation is authored by MEDS students and has been reviewed and approved by:

_____
Andrew Bartnik

_____
Carlo Broderick

_____
Gabrielle Smith

_____
Hailey Veirs

This Capstone Technical Documentation is authored by MEDS students and has been reviewed and approved by:

_____

Tamma Carleton

_____

Ruth Oliver

_____

Date

Table of Contents

1.0 Abstract

Food insecurity is a pressing issue in sub-Saharan Africa that is expected to be further exacerbated by climate change's effects on agriculture. Recent machine learning techniques applied to high-resolution satellite imagery have been effective in monitoring and forecasting agricultural outcomes and related environmental variables. However, there are significant barriers to widespread adoption of these techniques that impede equitable access, including substantial computational demands, data storage needs, financial costs, and specialized technical expertise. Thus, the implementation of these techniques remains largely inaccessible to low and middle income regions, which tend to be acutely vulnerable to the effects of climate change. Here, we explore an efficient machine learning approach that can be a viable alternative to traditional machine learning techniques for data and computationally constrained regions. In this project, we look at the low-income country of Zambia in a use case of the Multi-task Observations using Satellite Imagery & Kitchen Sinks (MOSAIKS) machine learning approach (Rolf et al. 2021). We use this method to extract summary imagery information, called "features" across Zambia from Sentinel-2 satellite imagery and pair it with ground-truth survey data from the Crop Forecast Survey, gathered by the Zambian Ministry of Agriculture. Using these features as predictors, we construct linear regression models that predict 28 selected agricultural variables. The resulting prediction datasets cover the time period from 2015 to 2022, which expands the geographic bounds and improves the spatial resolution of the existing agricultural survey data in Zambia. We found that much of the observed variation in many of these variables can be explained using our features as the only predictors, enabling a more comprehensive understanding of farmer adaptation to climate change.

2.0 Executive Summary

Effective monitoring of agricultural productivity plays a critical role in supporting food security, facilitating agricultural planning and policy, and mitigating the adverse effects of climate change. Although machine learning techniques applied to satellite imagery have demonstrated their effectiveness in land-use monitoring (e.g., Hansen et al., 2013), their conventional implementation is hindered by computational demands, data storage requirements, specialized expertise, and financial costs. Consequently, these techniques remain largely inaccessible to low and middle-income regions, such as sub-Saharan Africa, which face acute climate-related challenges (Inglada et al., 2017). To address this disparity, we present an implementation of an emerging technique that offers enhanced accessibility, as demonstrated through a comprehensive use case example in Zambia.

Our approach establishes a pipeline for generating high spatial resolution, temporally-consistent predictions of diverse agricultural outcomes in Zambia. We pair Random Convolutional Features (Rahimi and Recht, 2008) extracted from Sentinel-2 satellite imagery with ground-truth agricultural data. This enables us to predict 28 agricultural variables, including crop yields, harvest and production metrics, crop loss mechanisms, and tillage practices. The predictions are based on data from the Zambian Ministry of Agriculture's Crop Forecast Survey (CFS), spanning from 2008 to 2021. Our predictions fill spatial and temporal gaps in the survey's coverage, improving the spatial resolution and providing predictions for the years 2018 and 2019, where survey data was insufficient. By filling these gaps, our predictions contribute to a more comprehensive understanding of agricultural practices in Zambia, supporting efforts to combat food insecurity in the face of a changing climate.

We use an innovative machine learning approach, Multi-task Observation using SAtellite Imagery & Kitchen Sinks (MOSAIKS; Rolf et al., 2021), which includes steps for converting satellite images to "features" and producing predictions of agricultural variables. In the MOSAIKS system, we combine Sentinel-2 satellite imagery with machine learning (SIML) methods to generate monthly tabular data encoded with image pattern information, called "features", through a process called "featurization". Features capture comprehensive information from satellite images, such as color hue and saturation, patterns, and textures in a tabular format. We join these features with ground-truth data, the survey data, to allow for the training of a supervised machine learning model. We then create multiple cross-validated ridge regression models for the years 2016-202 to predict 28 agricultural variables present in the survey data. We apply these models outside of the survey area to create high temporal and spatial resolution predictions, which increases the resolution of and complements the survey data.

Previous research has demonstrated the ability of unsupervised featurization to achieve performance comparable to deep learning methods across various tasks (Rolf et al., 2021).

However, these examples have largely taken place within high income countries where ground-truthed data and imagery data are more available and of higher quality. In 2021, the CropMOSAIKS MEDS Capstone team implemented the MOSAIKS approach in Zambia, successfully predicting one variable, maize yield, at an annual temporal resolution for years 2014-2021 (Cohen, Cognac, Lewin, Molitor, 2022). This work introduced a novel temporal and resource context to applications of the technique, but focused on only one variable for prediction. Our implementation of MOSAIKS predicts 28 agricultural variables, from 2015-2022, and improves the spatial resolution of the previous iteration of the technique. The demonstrated accuracy in predicting these agricultural variables holds direct implications for researchers involved in assessing food security risks in Zambia by enabling a more comprehensive understanding of agriculture in the region.

The features, predictions, and code base generated through this application have applications beyond Zambia's agriculture and food security. The features derived from satellite imagery using MOSAIKS are task-agnostic and can be used for predicting a wide range of variables across Zambia. The reproducible code base pipeline contributes to the MOSAIKS ecosystem, as it builds on the code from the previous CropMOSAIKS capstone team, and provides public documentation of the MOSAIKS technique on dozens of agricultural variables within a resource-constrained, low-income context. The features and agricultural variable predictions created by this project will be integrated into a public-facing and freely accessible application programming interface (API). We aim to facilitate the accessibility of satellite data and methods, and empower researchers and decision-makers to embrace SIML methodologies in their standard data practices through this project.

3.0 Problem Statement

Climate change is anticipated to have numerous adverse effects on global agricultural productivity, posing a substantial threat to food security worldwide. Among the regions most vulnerable to these impacts, sub-Saharan Africa is particularly at risk (Hultgren et al., 2022). Integrating satellite imagery and machine learning (SIML) techniques has shown to allow effective monitoring and prediction of variables such as land use (Inglada et al., 2017), and forest cover (Hansen et al., 2013) and now maize crop yield (Cohen, Cognac, Lewin, Molitor, 2023). These SIML techniques can be used to monitor crucial agricultural metrics including other crop yields and various agricultural practices. These insights are essential to understand and address the profound impacts of a changing climate on existing food insecurity challenges. However, adopting satellite imagery and machine learning methods requires substantial financial and computational resources. Traditional SIML methods, such as convolutional neural networks, are often highly computationally intensive, require specific expertise and task-specific data for training, and focus on predicting only one outcome, such as crop yield (Ball et al., 2017). Furthermore, sub-Saharan Africa faces additional challenges due to limited data availability and computational capacities, a reality shared by many low and middle-income regions.

To address these barriers, a pre-processed collection of encoded satellite data can be utilized by researchers in these low and middle-income regions. Our project focuses on the use of a novel machine learning process, MOSAIKS, to expand the spatial and temporal resolution of on-the-ground agricultural survey data, and predict various agricultural metrics and practices. The Crop Forecast Survey is used by Zambian policy makers to guide agricultural investment, drive development, and support human welfare. Increasing the resolution and quality of the survey data would support Zambia’s agricultural economy and bolster the country’s food security.

4.0 Specific Objectives

This project aims to achieve the following objectives:

1. Demonstrate the efficacy and task-agnostic nature of the MOSAIKS technique in a novel application of addressing many prediction tasks within a low-income, data-limited context, at a high spatial and temporal resolution.
    a. Generate features for Zambia by encoding annual satellite imagery in Random Convolutional Features at a monthly time scale using the MOSAIKS technique.
    b. Improve the quality of raw data from the survey data through extensive cleaning, organizing, and processing.
    c. Create ridge regression models trained on the feature data paired with survey data to predict 28 agricultural variables.
2. Fill the agricultural data gap in Zambia.
    a. Enhance the spatial resolution and geographic coverage of agricultural data in Zambia by deploying the trained machine learning models to regions beyond the spatial and temporal coverage of the survey data.
    b. Enhance temporal resolution by generating predictions for all variables for missing years (2018, 2019).
3. Contribute to the broader MOSAIKS ecosystem.
    a. Extend the utility of the MOSAIKS process by providing reproducible, well-documented guidelines based on the Zambia use case.

To achieve these objectives, we applied the MOSAIKS approach to open-source Sentinel-2 satellite imagery, implementing a $1km^2$ grid cell resolution at a monthly time scale for years 2015-2022. The resulting features were combined with survey data to train a ridge regression model (refer to 7.0 Summary of Testing for detailed model validation and testing information), which we used to predict variables of interest for each year. The following variables were selected from the survey data through a combination of data quality and hypotheses about potential visibility from space:

1. Harvested Area and Production
    a. Total Area Harvested (ha)*
    b. Total Area Lost (ha)*
    c. Total Harvest (kg)*
    d. Yield (metric of overall production, kg/ha)*
    e. Fraction of Area Harvested/Fraction of Area Lost*
2. Crop yields (kg/ha)
    a. Maize*
    b. Groundnuts*
    c. Mixed Beans
    d. Popcorn
    e. Sorghum
    f. Soybeans
    g. Sweet Potatoes

    h. Log(Maize)*
    i. Log(Sweet Potatoes)
    j. Log(Groundnuts)*
    k. Log(Soybeans)
3. Crop loss mechanisms
    a. Fraction of Area Lost due to Drought
    b. Fraction of Area Lost due to Flooding
    c. Fraction of Area Lost due to Animal Destruction
    d. Fraction of Area Lost due to Pests
    e. Fraction of Area Lost due to Soil Quality
    f. Fraction of Area Lost due to Lack of Fertilizer
4. Agricultural Practice
    a. Bunding (construction of ridge-like structures along the contours of a field, used to control water flow, prevent soil erosion, and retain moisture) (ha)
    b. Monocropping* (ha)
    c. Mixture Cropping (ha)
    d. Proportion of Area Tilled Using a Conventional Plough*
    e. Proportion of Area Tilled Using Ridge*
    f. Proportion of Area Not Tilled
    g. Proportion of Area Tilled by Hand
    h. Proportion of Area Monocropped*
    i. Proportion of Area Mixed Cropped

It should be noted that not all variables demonstrate high performance in our modeling iterations. In our workflow, we extensively tested and documented all variable performance scores (refer to Appendix A for detailed variable testing scores). Our modeling iterations involve cross-validation and bootstrapping sampling methods, though the latter is still incomplete and should be considered in future work. For this reason, we focus on assessing model performance metrics for the cross-validated ridge regression models. We established a validation R^2 score threshold of 0.4 as an indicator of statistical significance. Any variables that meet or exceed this minimum threshold are marked with an asterisk (*) in the previous list. Generally, our poorest performing variables were those affected by known data quality issues. However, further testing is recommended to gain a better understanding of the feasibility of predicting specific variables using these techniques.

5.0 Summary of Solution Design

5.1 Approach and Methods

1. Data Collection and Storage
    a. Ground Truth Data: Crop Forecast Survey, Zambian Ministry of Agriculture
        i. The ground truth data used in this project comes from Zambia's Crop Forecast Survey, an annual survey designed to forecast agricultural production for the year and to record changing agricultural practices within the country. This data is made available by the Zambian government in collaboration with the Baylis lab at UCSB. Access to this data is restricted due to privacy and security concerns and will not be shared as part of this project.
        ii. The survey data was shared with our team via Box and the raw data was formatted in individual CSV/.sav files for each survey year. Supporting geofiles to spatially identify each survey enumeration area were included separately.
        iii. The survey data takes the form of individual household responses to the Crop Forecast Survey (CFS). Each response was collected and is associated with a specific geographic area known as Survey Enumeration Areas (SEAs).
    b. Satellite Imagery
        i. All satellite images used in this project were stored and accessed in the cloud on Microsoft Planetary Computer servers located in Western Europe.
        ii. All satellite imagery for this project originated from Sentinel-2.

2. Preprocessing
    a. Ground Truth Data: Survey Data (Modeling Targets)
        i. We used the programming language R and tidyverse packages to transform our individual field-level survey data into SEA/year-level data. We extensively cleaned and filtered the data, these steps were documented in our cleaning script and associated metadata. We then spatially joined our aggregated data to geo-locate each SEA.
    b. Satellite Imagery: Prepare Images to be Featurized (Modeling Features)
        i. We used Microsoft Planetary Computer's Spatio-Temporal Access Catalogue (STAC) API to access Sentinel-2 satellite imagery. We temporally filtered satellite imagery and excluded images with more than 20 percent cloud cover using the STAC API's eo:cloudcover tag. We then created a uniform grid of evenly spaced points to sample from during the featurization step.

3. Featurization
    a. Following the 2021-2022 MOSAIKS team's approach, an unsupervised machine learning featurization process was used to compute Random Convolutional Features (RCFs) for all filtered satellite images using Microsoft Planetary

Computer and Azure Machine Learning Studio. The featurization converts image data into georeferenced tables with a latitude and longitude point, month, year, and feature set in each row.

b. We joined the feature points to their associated Survey Enumeration Area (SEA) polygons.
c. We took the average values across all points for each feature within each SEA to create a single record for each SEA for each year. This new format matched the spatial and temporal resolution of the aggregated survey data, to allow for the training of a ridge regression model.

4. Modeling
    a. We joined the survey data with features using unique SEA identifiers to create a labeled data set.
    b. We split the data into training, test, and validation sets according to conventional machine learning approaches using scikit-learn's train_test_split function. We used 10% of our data for validation, 10% for testing, and 80% for training.
        i. Different iterations of our model used different sampling methods. See 7.0 Summary of Testing for further details.
    c. We then created models for each target prediction variable, using 5-fold cross-validated ridge regression.
    d. We tuned each model's ridge regression penalization coefficient, α, as necessary. We allowed RidgeCV to evaluate 75 values for α within a logspace ranging from 10^{-8} to 10^{8}, and evaluated model performance metrics on the validation set.
    e. Models with the highest performance on the validation set (here we used the R^2 metric) were then selected to create prediction maps.
        i. Prediction maps for the following target variables were created:
            1. Yield (kg/ha),
            2. Fraction of planted area harvested
            3. Fraction of planted area lost
            4. Maize yield (kg/ha)
            5. Fraction of area lost due to drought
            6. Proportion of planted area plowed
            7. Proportion of area monocropped

5. Visualizations and Presentation Materials
    a. We created feature maps demonstrating the MOSAIKS encoding process on the SEA level and over the whole country for each year (Figure 2).
    b. Once predictions were generated, we then created prediction maps demonstrating model outputs on the SEA level and over the whole country for each year (Figure 3).
    c. We generated plots to visually demonstrate model performance, using maps and bar charts to depict the spatial and temporal predicted versus actual values as well as associated scatter plots showing our lines of best fit (Figure 1).

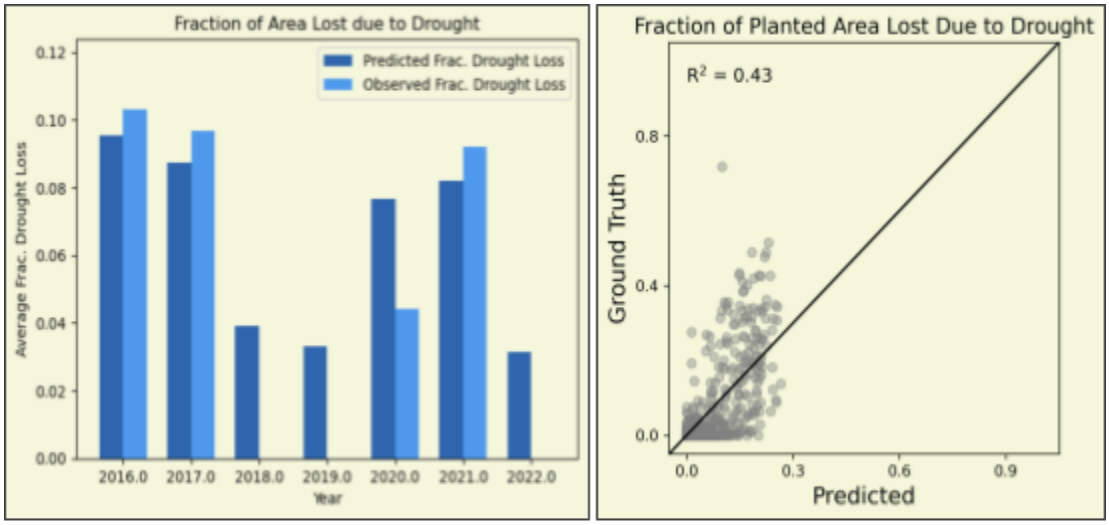

Figure 1: Predicted versus observed plot of fraction of planted area lost due to drought, filling in temporal gaps in 2018 and 2019 resulting from data quality issues (right). Scatterplot of line of best fit for predicted versus observed values for fraction of planted area lost due to drought (left).

6. Support for Future Work
    a. This capstone project built on previous MOSAIKS implementations that had either focused on a single prediction task over a range of time or multiple prediction tasks within a single time period. This project demonstrated multiple prediction tasks over a range of time. The work done in this project also created a well-documented pipeline for future users to implement MOSAIKS within their own work.

5.2 Data Management Plan

5.2.1 Zambian Agricultural Survey, Satellite Data, and Prediction Datasets

The data for this project consists of two primary components: survey data and the featurized satellite imagery. These data sets were manipulated and combined to make a joined data set that was used to train ridge regression models. These models were then used to create prediction maps and datasets to expand the temporal and spatial resolution of the CFS data.

5.2.1.1 Ground Truth Data

The ground truth data for this project comes from Zambia's Crop Forecast Survey, an annual agricultural survey meant to forecast food production that collects information about farming practices, crop species, and yields. The ground truth data were sent to us by our client, Protensia Hadunka, through Box. These data were then uploaded to Taylor and Tsosie, two remote servers located at the Bren School of Environmental Science & Management, in Santa Barbara, California. The raw ground truth data were stored in a series of folders, one for each growing season from 2007-2008 to 2021-2022. Each folder contained raw data in the form of .csv, .sav, and .dta files. Each harvest year .csv file contained the responses to survey questions collected from Zambian farmers by the Zambian government. We decided to use the crop.csv file that included 28 variables of interest that were later used in modeling. The data were spatially joined to their associated Survey Enumeration Area (SEA) then averaged over each SEA and year,

giving us 1,300 observations before imputation methods. Despite extensive cleaning efforts, these data had significant quality issues which likely impacted our model performance. Issues included inconsistently empty or missing columns between datasets, nonsensical numerical values, conversion errors, missing spatial identifiers across entire districts, and the duplication of entire datasets.

5.2.1.2 Satellite Imagery

The images used in this project originated from the European Space Agency's Sentinel-2 satellite. The satellite imagery was stored and accessed on Microsoft Planetary Computer using the STAC API and imported into the Microsoft Planetary Computer Hub and Azure Machine Learning Studio as raster files. Target locations for feature extraction were selected using a subsetted grid of roughly one kilometer square points layed over the country of Zambia. Using those target locations, the satellite image catalog on MPC was then queried for satellite images in those locations with cloud cover below 20% and within each specific time range. These satellite image rasters were then featurized using a random convolutional feature process on virtual machines accessed via MPC and Azure. This process created monthly feature files representing the features extracted from satellite images taken at specific locations within Zambia within each month. The monthly feature files were then exported from MPC and Azure and stored on Taylor and Tsosie, two remote servers at the Bren School of Environmental Science & Management. Once on Bren servers, an imputation process was used to fill the missing values created by time periods without cloudless images. The method of imputation employed backfilling to replace missing values with those from the previous month. If the preceding month's data was also absent, the missing values were filled in by calculating an average over time and space.

Here in Figure 2 we see four feature maps depicting a single feature value for a single month extracted from images across Zambia.

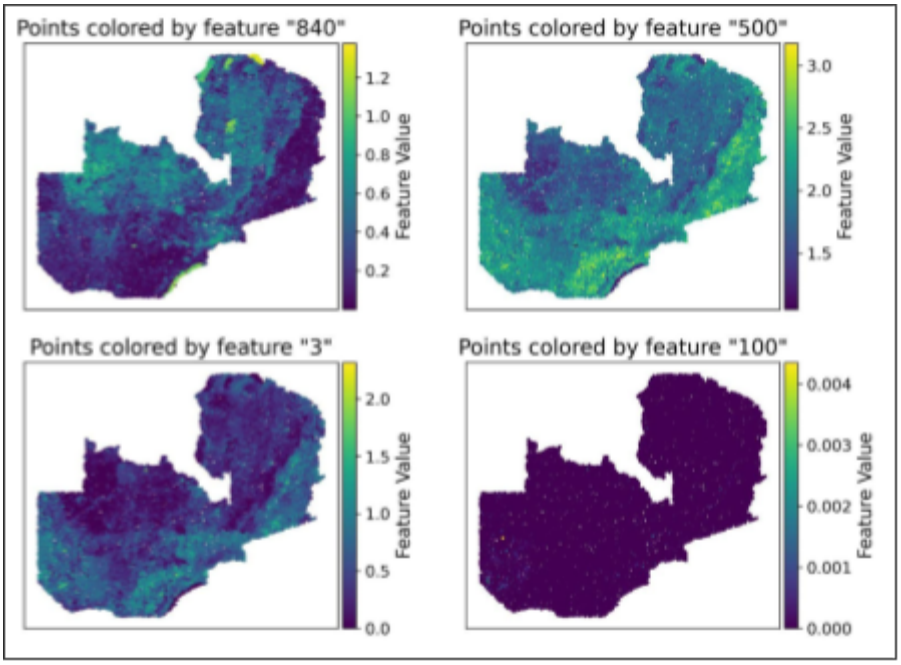

Figure 2: Four feature maps depicting featurized points across all of Zambia with colors showing feature values.

5.2.1.3 Joined Dataset

Once the ground truth data was processed and the features were aggregated yearly, we then joined the two datasets based on the year and SEA identifiers present in both datasets. Our data was then in tidy format, where each observation was an SEA/year, with the associated variables we wanted to predict and the aggregated features as columns. This format allowed us to train a ridge regression model to predict each of the 28 target survey variables.

5.2.1.3 Prediction Datasets

Features were created for points across all of Zambia. Models were trained on the features associated with agricultural survey data locations, SEAs. Once the models were trained, they were then used to predict agricultural data outside the survey areas. These predictions took the form of latitude, longitude, and prediction values for a roughly 1 km square around those points for each year that features were created.

In Figure 3, we display four prediction maps showing values for predicted target variables across the country of Zambia.

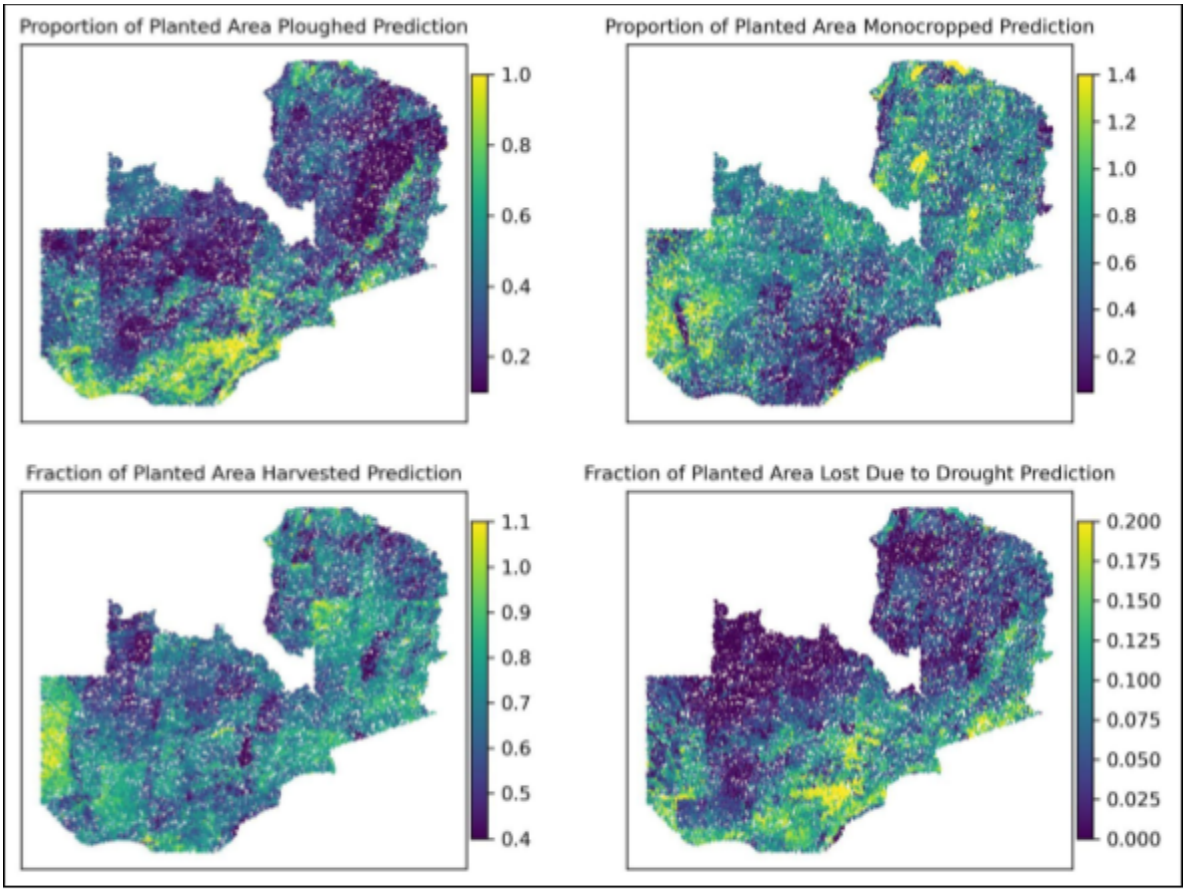

Figure 3: Four prediction maps depicting prediction points across all of Zambia with colors showing prediction values.

5.2.2 Data standards

The feature files created in this project were concatenated and stored as a single feather file that can be read in as a geodataframe with point geometries. These files are available on our public Google Drive folder. These features may be available on the MOSAIKS website for others to access and use at a later date using an API. The ground-truth survey data was cleaned and aggregated into a single dataset that was stored in a .csv format; however, this data is not publicly available due to privacy and security concerns but will be made available to our clients.

Our data and process are available on GitHub and formatted in a standard organization hierarchy that is easily accessible and searchable on GitHub.

5.2.3 Metadata

Metadata for this project is available via our GitHub organization. The organization overview contains general project purpose and structure statements. Within the organization, separate repositories are available for preprocessing, featurization, modeling, and visualizations associated with the project. Each repository contains README's that provide detailed descriptions of relevant notebooks, data sources, and file organization.

5.2.4 Data Sharing and Access

The data used in this project comes from two sources: the survey data is from the Zambian government's Crop Forecast Survey (CFS), and the imagery data is from public satellites. Access to both the raw and cleaned survey data has been restricted to the use of the capstone team. This data is sensitive in nature and may contain identifiable information, and therefore will not be made available to the public. The survey data takes the form of .csv and geofiles that allow for sub district level attribution of local agriculture practices and outcomes. The preprocessing repository contains traces of the private survey data and thus will not be accessible to the public. All other data sources for this project come from publicly available sources that can already be easily accessed and shared. All satellite data for this project can be accessed via Microsoft Planetary Computer (MPC). All repositories other than the preprocessing repository will be accessible to the public. Prediction datasets made during this project will be stored on our public facing Google Drive folder and will be accessible to the public.

5.2.5 Intellectual property and re-use

All survey data provided by the Zambian government will not be accessible to the public and its use will be restricted to the capstone team alone. The public repositories, workflows, and data derivatives will be accessible to the public via our GitHub repository and Google Drive folder.

5.2.6 Data archiving and preservation

This project used publicly available satellite data, therefore we did not store satellite image data. Additionally, due to the nature of the survey data, we did not provide the raw or cleaned survey data used in this project, the data for this will be stored on a private Google Drive folder. The images and machine learning code the team produces will be publicly hosted on GitHub. Feature and prediction data produced by the team will be available on a public Google Drive folder and may be uploaded to Zenodo or the MOSAIKS website at a later date.

5.3 Software or Tools

R and Python were used as programming languages to process data and apply the MOSAIKS machine learning algorithm. R coding was conducted using R Studio, and Python coding was conducted using JupyterLab, Visual Studio Code, and Microsoft Machine Learning Studio.

The Bren Taylor and Tsosie servers, the Microsoft Planetary Computer (MPC), and the Azure Machine Learning Studio were used for computing resources and data storage. The Taylor and Tsosie servers were made available through the project's association with the Bren School of Environmental Science & Management. MPC was accessed using the SpatioTemporal Asset Catalog (STAC) API via a free account. Azure Machine Learning Studio was utilized for supplementary computing power, in tandem with a Standard_NC12s_v3 virtual machine.

Use of the Azure platform involved an estimated cost of around $500 for cloud services. Both the MPC and Azure were used to perform the computation-intensive featurization process of the MOSAIKS machine learning approach.

The STAC API was used to access images from the Sentinel-2 satellite, which were then processed on the MPC and Azure platforms. The outcomes of the project are stored and shared through Box, Google Drive folder, and publicly accessible GitHub repositories.

6.0 Products and Deliverables

6.1 Client Products and Deliverables

1. Cleaned and consolidated survey data for 2008-2022 and documentation on the processing of this data.
    a. This deliverable is only available to our client, Tamma Carleton, the Baylis lab at UCSB, Sitian Xiong, and the Zambian Ministry of Agriculture due to data privacy and security concerns.
2. Sets of featurized satellite imagery data and featurization documentation for Zambia. These products and documentation are accessible to the public via GitHub repositories and possibly in the future via the MOSAIKS online API. There are currently 3 sets of features available:
    a. 10% sample density of entire country of Zambia (2015-2022) (Sentinel-2, RGB - Red, Green, Blue)
    b. 70% sample density at the SEA level within Zambia (2015-2022) (Sentinel-2, RGB - Red, Green, Blue)
    c. 50% sample density at the SEA level within Zambia (2015 - 2022) (Sentinel-2, RGB8 - Red, Green, Blue, Near Infra-Red)
        i. Features can be accessed via our project's public Google Drive folder.
        ii. Documentation and guidance on feature creation can be found on our Feturizaion GitHub repository.
3. Data pipeline demonstrating joining features to CFS ground truth data and the modeling process.
    a. All documentation can be found on our Modeling GitHub repository.
4. Modeled predictions and performances for each iteration of our model and associated visualizations for the entire country of Zambia for each high performing model.
    a. Comprehensive model performances are listed in Appendix A.
    b. Prediction datasets can be found on our project's Google Drive folder.

6.2 Academic and Capstone Products and Deliverables

As part of the Bren capstone requirements, we have additionally produced the following deliverables:

1. Design and Implementation Plan
2. Design and Implementation Plan Faculty Review presentation
3. Technical Documentation
4. Capstone Project Final Presentation
5. Project Repository

7.0 Summary of Testing

Tests were employed throughout our data cleaning, featurization, and modeling steps to ensure our pipeline was correctly processing and producing data. The purpose of testing was to ensure the accuracy of our intermediate results, measure uncertainty in our model, and allow for open-source, repeatable results. A summary of our testing procedures is outlined below.

7.1 Data Cleaning Testing

All survey data provided by clients were cleaned and organized in preparation for use. Tests were implemented in each cleaning notebook to ensure the aggregation and cleaning steps were working as intended. The testing for the cleaning process is listed below. For privacy purposes, data cleaning notebooks will not be made publicly available.

1. **Initial Data Integrity Checks**: We first checked for missing data and inconsistency in each of the original datasets. We removed all data for 2018-2019 since the 2019 data was a copy of the 2020 data, and the 2018 crop file was missing. Additionally, we were unable to validate that the SEAs in the Northern, Central, and Eastern districts were consistently identified across years, so we excluded all data from these districts. We filtered out a substantial amount of observations with nonsensical values (e.g. harvested area was larger than planted area, etc.)

2. **Testing after Each Transformation**: In aggregating the individual field level survey data to the SEA level, the data underwent several transformations including calculating statistics for each variable of interest, and pivoting from long to wide formats. After each of these transformations, we validated that the output was as expected. For example, after aggregating, we employed tests to ensure that the total area harvested in each SEA/year was indeed the sum of the area harvested for each individual field. After converting the data from long to wide format, we tested the data integrity by checking whether the same values were maintained. We also checked whether the total area lost due to each reason, crop yield, and area tilled in the original data matched with the reshaped data.

3. **Final Data Review:** At the end of the process, we reviewed a random sample of records for consistency and correctness. We also checked the descriptive statistics (e.g., mean, min, max, standard deviation) for each variable to ensure that they were in reasonable ranges based on our understanding of the data.

7.2 Featurization Testing

Manual exploratory testing and debugging were used throughout the random convolution featurization process. This testing incorporated code chunk output verification and the creation of intermediate visualizations, such as the mapping of generated features, grid point maps, and satellite image maps. Visualizations were used as a mechanism to confirm the code was performing as expected.

Activation maps for individual random convolutional feature sets were plotted. Activation maps, which help visualize the location of certain attributes within a defined space and time, indicate a high activation when a certain attribute was detected and were useful in validating the creation of

valid feature sets. Seen below are two feature maps depicting the spatial distribution of feature values across SEA 221 (Figure 4a) and across the country of Zambia (Figure 4b).

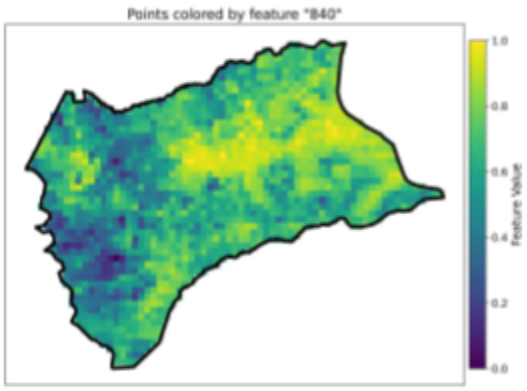

Figure 4a: Feature map on the largest Survey Enumeration Area SEA (221) with color showing feature value. Featurization done using Sentinel-2 RGB bands.

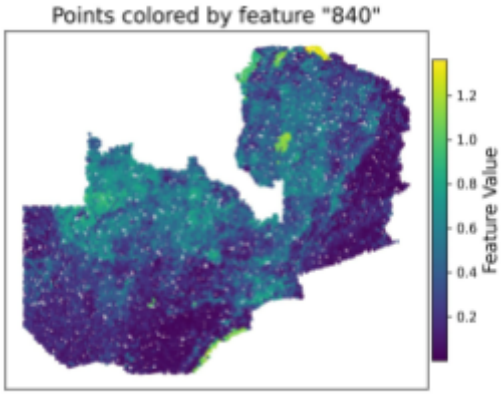

Figure 4b: Feature map depicting featurized points across all of Zambia with colors showing feature values. Featurized using Sentinel-2 RGB bands.

These activation maps were subsequently juxtaposed with visual band (RGB) satellite imagery for outlier detection and to verify the correct featurization of images. This process involved comparing the activation maps with the actual images to ensure our features accurately represented the true objects, such as trees, crops, and geographic features.

Additional visualizations were developed to evaluate output accuracy. These included plots of the uniform subset of 1 km^2 grid sample points, the 1 km^2 grid of Sentinel-2 imagery, and feature maps between years to ensure temporal stability. Debugging statements, acting as internal self-checks in the code, were used to aid in testing and troubleshooting errors. Error messages were incorporated into the code to prevent minor silent errors from escalating into major problems.

The 'assert' statement was used frequently throughout the code for similar purposes. An example of its usage was requiring the number of points. Additionally, ‘if’ statements were implemented to use only images with a certain number of pixels, ensuring the use of full, complete images. Peer code review was carried out in every coding step through group code reviews, and pull requests.

7.3 Model Testing

We used a series of tests to validate that our modeling process was working as expected. Our function provides users with the flexibility to choose between several parameters. This section details each parameter, the scenarios in which they are applicable, and the tests used to ensure each parameter was working correctly. Each testing observation is a single SEA’s feature set for a single year.

7.3.1 Sampling Methods

Our function accommodates two different sampling methods that can be used in combination alongside scikit-learn's RidgeCV function. Our model testing primarily involved printing out intermediate results and employing checks to ensure our models were working as expected. We used 5-fold cross-validation (CV) in each of our model approaches for the sampling method, building off of it as we further customized our modeling approach. We experimented with different sampling methods to address and evaluate our models' performance on unseen spatial and temporal data. We extensively used random states throughout our modeling process to ensure reproducibility and to allow us to observe the effects of the four combinations of sampling methods, the implemented tests to validate their functionality, and the specific reasons chosen for utilizing each are below:

1. 5-fold CV
    - 5-fold cross-validation was used without any other methods to first examine our baseline model performances using a standard sampling approach. Here, each SEA/year had an equal probability of being sampled. We ensured the 5-fold CV was working as intended for each variable by printing out performance metrics for each variable: the training R^2 and validation R^2 on each fold. To tune our penalty coefficient for each variable, we searched over a logspace with 75 values using RidgeCV.
2. Block Sampling
    - When block sampling is selected, all observations for a specified number of unique spatial identifiers are randomly selected to be held out from the training set. The model is then trained on the remaining unique spatial identifiers, and then evaluated on the unseen spatial identifiers. This allows users to evaluate the models' performances on unseen spatial areas. We tested this approach by printing out the unique spatial identifiers held out in the validation set, and the unique spatial identifiers kept in the training set to ensure that no unique spatial identifiers appeared in both the training and validation set to prevent data leakage.
3. Bootstrapping (Incomplete)
    - When bootstrapping is selected, the original data is resampled a specified number of times with replacement to create n_bootstrap samples. The 5-fold CV process is then applied to each sample, generating a model for each resample. The training and validation R^2 values are printed out for each resample, to ensure that the resampling process is working as expected. We were unable to complete this final step, but in future work these models should be weighted and aggregated into a final model ensemble that can be used to generate new predictions.
4. Bootstrapping + Block Sampling (Incomplete)
    - When both bootstrapping and block sampling are selected, a specified number of unique spatial identifiers are first held out of the training dataset, which is then resampled n_bootstrap times. The 5-fold CV process is then applied to each sample, again generating a model for each resample. The unique spatial identifiers held out are printed for each model fit to each bootstrap resample, the training and validation scores of which are also printed out for each fold.

Before any results are printed, our function prints out several parameters selected by the user to ensure the proper parameters are being employed. The parameters output by the function before the models run are: the target columns, the validation and test sizes, whether or not bootstrapping

and block sampling were employed, and the random state used. We measured the accuracy of our training, validation, and testing sets primarily using the R^2 metric. Below are the arguments passed to our function (in a dictionary), and an explanation of their use and testing.

1. **target_columns**: A list of outcomes to model. Users can select one or multiple target columns for the model. For each target variable, a new model is fit. The resulting model coefficients, predictions, and performance metrics are stored and returned for each target column. Each target column is printed out along with its performance metrics as the model runs.

2. **test_size**: A numerical value between 0-1. This determines the proportion of the dataset to be used as the test set. The remaining data is used as the training set. If block sampling is not used, the split is random.

3. **validation_size**: A numerical value between 0-1. This is the portion of the training data to be held out for validation purposes. The size of the validation set is printed, to ensure that the data is split as intended.

4. **bootstrap**: If set to True, the model employs the bootstrapping resampling process.

5. **n_bootstraps**: This parameter determines the number of bootstrap samples to be used if bootstrapping is enabled. Each bootstrap sample is used to fit a model and make predictions, and the final model parameters and predictions are averaged over all bootstrap samples. The performance of each model for each individual sample and each variable is printed out to ensure that our bootstrapping process is indeed randomly sampling the data.

6. **block_sample**: If set to True, the model utilizes block sampling. This allows users to, given a unique spatial identifier (in our case, the unique SEA id number), hold out a specified number of spatial areas. This provides users to see how their models perform on unseen spatial data. When this is set to true, users then specify the number of these unique observations that they want to randomly hold out. We have set our function to print out the values of the unique SEA id numbers that are being trained on, and those that are being held out, to ensure that there is no data leakage.

7. **n_seas_held_out_val**: This parameter determines the number of sea regions to hold out for the validation set when block sampling is enabled. As a test/check, this value is printed out.

8. **n_seas_held_out_test**: This parameter determines the number of sea regions to hold out for the test set when block sampling is enabled. As a test/check, this value is also printed out.

9. **random_state**: This controls the shuffling applied to the data before applying the split for reproducibility.

8.0 User Documentation

The project's expected users are the client, co-authors, and collaborators. Potential future users include researchers interested in examining food security and agriculture in Zambia, and future students, researchers, or individuals interested in using the MOSAIKS workflow.

The MOSAIKS GitHub Project Organization is the primary documentation and source of code for this project. The GitHub organization contains a homepage README directing users to four primary repositories: Preprocessing, Featurization, Modeling, Visualizations. Each repository contains a comprehensive README describing its purpose, notebooks, and user instructions for reproducibility. Detailed code explanations can be found within each notebook, presented in both markdown chunks and comments, along with guidance on adapting the code for different use cases.

The Preprocessing repository serves to join and process spatial data and ground truth data for further analysis and model training. This repository holds shapefiles used to join the Zambia CFS survey data with corresponding Survey Enumeration Areas (SEAs), as well as files to outline methodologies of raw data preprocessing. Original and cleaned Zambia CFS survey data is not made publicly accessible for privacy purposes. This repository will also not be made public and will only be available to our client, those who provided the data, and the Zambian Ministry of Agriculture.

The Featurization repository serves as an in depth guide to the MOSAIKS feature generation process. Feature data created by the MOSAIKS team will be available exclusively to our clients via a shared Google Drive. Notebooks contained within this repository support the use of either Microsoft Planetary Computer or Microsoft Azure cloud computing platforms.

The Modeling repository outlines the steps to combining tabular data and the previously generated features to input into regression models for different agricultural variables. It contains separate notebooks for preprocessing and joining the data, and for training and testing ridge regression learning models against various agricultural variables.

The Visualizations repository documents the creation of visualizations to accurately portray the generated features, predictions, and other outcomes. These visualizations served as validation for the MOSAIKS process, ground truth data quality, prediction outcomes, and as communication aids during presentations.

9.0 Archive Access

- Features, predictions, and visualizations will be stored on a shared Google Drive folder, which the client Tamma Carleton has access to. The data can be moved to a more permanent and secure storage system by the client if needed.
- Original and cleaned Crop Forecast Survey data provided by the Zambian Agricultural Ministry is private and will not be shared as per the request of the Baylis Research Group. The code and repository for cleaning this data will also not be shared due to the close proximity to the sensitive survey data.
- Project code and user documentation will be located on the GitHub Project Organization.
- User documentation is provided for all repositories within each repository (Featurization, Modeling, and Visualizations).

10.0 Acknowledgements

Much of the code and copy for this project was supported by or originates from the 2022 University of California, Santa Barbara, Bren School of Environmental Science & Management, Master of Environmental Data Science Capstone: CropMOSAIKS An Open-source Pipeline for Remote Sensing of Crop Yields: a Zambia Case Study.

From the Bren School of Environmental Science & Management, thank you to Tamma Carleton, Ruth Oliver, and Naomi Tague; the 2022 CropMOSAIKS team with Steven Cognac, Juliet Cohen, Grace Lewin, and Cullen Molitor. Thank you for your support and guidance throughout the project.

From the Baylis Research Group at the University of Illinois, thank you Kathy Baylis, and Protensia Hadunka for aiding us with the survey data.

Thank you to Sitian Xiong at Clark University, the Graduate School of Geography for additional spatial data and preprocessing help for the survey data.

Thank you to the Zambian Government for access to the Crop Forecast Survey data.

Thank you to the European Space Agency for their publicly available satellite imagery.

11.0 References

Ball, J.E., Anderson, D.T. & Chan, C.S. A comprehensive survey of deep learning in remote sensing: theories, tools and challenges for the community. J. of Appl. Remote Sens. 11, 042609 (2017). https://doi.org/10.1117/1.JRS.11.042609

Cohen, Juliet, Steven Cognac, Grace Lewin, Cullen Molitor. *AN OPEN-SOURCE PIPELINE FOR REMOTE SENSING OF CROP YIELDS: A ZAMBIA CASE STUDY.* University of California, Santa Barbara, Jun. 2022. Web. February 2023. https://bren.ucsb.edu/projects/open-source-pipeline-remote-sensing-crop-yields-zambia-case-study

Hansen, M.C. et al., High-Resolution Global Maps of 21st-Century Forest Cover Change. Science 342, 850-853 (2013). DOI: 10.1126/science.1244693

Hultgren, Andrew and Carleton, Tamma and Delgado, Michael and Gergel, Diana R. and Greenstone, Michael and Houser, Trevor and Hsiang, Solomon and Jina, Amir and Kopp, Robert E. and Malevich, Steven B. and McCusker, Kelly and Mayer, Terin and Nath, Ishan and Rising, James and Rode, Ashwin and Yuan, Jiacan, "Estimating Global Impacts to Agriculture from Climate Change Accounting for Adaptation" (September 16, 2022). Available at SSRN: https://ssrn.com/abstract=4222020 or http://dx.doi.org/10.2139/ssrn.4222020

Inglada, Jordi, et al. "Operational High Resolution Land Cover Map Production at the Country Scale Using Satellite Image Time Series." MDPI, Multidisciplinary Digital Publishing Institute, (January 22, 2017). https://www.mdpi.com/2072-4292/9/1/95.

Rolf, Esther, Jonathan Proctor, Tamma Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht, and Solomon Hsiang. "A Generalizable and Accessible Approach to Machine Learning with Global Satellite Imagery." Nature Communications 12, no. 1 (December 2021): 4392. https://doi.org/10.1038/s41467-021-24638-z.

Appendix A: Model Testing for Various Sampling Techniques

Table 1A: Model Performance Scores: Cross-Validated Ridge Regression

Variable Name	Training R2	Validation R2	Pearsons	Testing R2
Total Area Harvested (ha)	0.71	0.46	0.69	0.63
Total Area Lost (ha)	0.75	0.5	0.72	0.58
Total Harvest (kg)	0.86	0.45	0.71	0.37
Yield (kg/ha)	0.74	0.62	0.8	0.58
Fraction Area Harvested	0.64	0.46	0.71	0.2
Fraction Area Lost	0.64	0.46	0.71	0.2
Maize	0.77	0.61	0.79	0.6
Groundnuts	0.52	0.44	0.67	0.33
Mixed Beans	0.34	0.22	0.49	0.18
Popcorn	0.13	-0.04	0.08	0.01
Sorghum	0.12	-0.01	0.1	0.1
Soybeans	0.42	0.17	0.44	0.14
Sweet Potatoes	0.46	0.3	0.55	0.07
Log Maize	0.77	0.71	0.84	0.65
Log Sweet Potato	0.51	0.36	0.6	0.04
Log Groundnuts	0.58	0.42	0.66	0.38
Log Soybeans	0.4	-0.07	0.2	0
Bunding	0	-0.02	0.1	-0.03
Monocrop	0.59	0.51	0.73	0.47
Mixture	0.08	0.06	0.29	0.08
Proportion Till Plough	0.78	0.71	0.85	0.85
Proportion Till Ridge	0.77	0.54	0.74	0.61
Proportion No Till	0	-0.75	0.05	0
Proportion Hand Till	0.6	0.33	0.58	0.18
Proportion Monocrop	0.9	0.56	0.76	0.68
Proportion Mixed Crop	0.02	-0.09	-0.02	0.05
Fraction Loss Drought	0.42	0.38	0.62	0.32
Fraction Loss Flood	0	0	-0.04	0
Fraction Loss Animal	0.1	-0.12	-0.1	-0.32
Fraction Loss Pests	0	-0.03	-0.01	0
Fraction Loss Soil	0.07	0.01	0.12	-0.24
Fraction Loss Fertilizer	0.03	0.01	0.29	-0.22

Categorical Variable Name	FPR	AUC-ROC
Loss Indicator	0	0.84
Drought Indicator	0	0.79
Field Loss Indicator	0	0.46
Animal Loss Indicator	0	0.42
Pest Loss Indicator	0	0.42

Variable Group
Harvested Area and Production
Crop Yields
Crop Loss Mechanisms
Agricultural Practice

Table 1B: Model Performance Scores: Bootstrapped & Cross-Validated Ridge Regression (10 samples)

Variable Name	Training R2	Validation R2	Pearsons
Total Area Harvested (ha)	0.87	0.69	0.84
Total Area Lost (ha)	0.88	0.7	0.85
Total Harvest (kg)	0.89	0.57	0.79
Yield (kg/ha)	0.89	0.72	0.85
Fraction Area Harvested	0.87	0.6	0.8
Fraction Area Lost	0.87	0.6	0.8
Maize	0.88	0.7	0.85
Groundnuts	0.84	0.58	0.78
Mixed Beans	0.64	0.27	0.62
Popcorn	0.24	0.2	0.48
Sorghum	0.36	0.06	0.37
Soybeans	0.7	0.33	0.66
Sweet Potatoes	0.77	0.49	0.72
Log Maize	0.92	0.74	0.86
Log Sweet Potato	0.78	0.39	0.68
Log Groundnuts	0.87	0.57	0.77
Log Soybeans	0.66	0.16	0.49
Bunding	0	-147.82	0.13
Monocrop	0.79	0.5	0.76
Mixture	0.18	-0.04	0.24
Proportion Till Plough	0.94	0.78	0.89
Proportion Till Ridge	0.91	0.53	0.77
Proportion No Till	0.04	-0.13	NaN
Proportion Hand Till	0.85	0.58	0.77
Proportion Monocrop	0.94	0.78	0.89
Proportion Mixed Crop	0.39	-0.19	0.29
Fraction Loss Drought	0.66	0.47	0.71
Fraction Loss Flood	0.14	-0.01	0.23
Fraction Loss Animal	0.08	-0.09	0.18
Fraction Loss Pests	0.23	-0.44	0.2
Fraction Loss Soil	0.21	-0.08	0.36
Fraction Loss Fertilizer	0.49	-0.09	0.35
Categorical Variable Name	**FPR**	**ROC/AUC**	
Loss Indicator	0.45	0.89	
Drought Indicator	0.08	0.91	
Field Loss Indicator	NaN	NaN	
Animal Loss Indicator	NaN	NaN	
Pest Loss Indicator	NaN	NaN	

Variable Group
Harvested Area and Production
Crop Yields
Crop Loss Mechanisms
Agricultural Practice

Table 1C: Model Performance Scores: Bootstrapped and Cross-Validated Ridge Regression (100 samples)

Variable Name	Training R2	Validation R2	Pearsons
Total Area Harvested (ha)	0.87	0.63	0.81
Total Area Lost (ha)	0.88	0.64	0.82
Total Harvest (kg)	0.9	0.61	0.8
Yield (kg/ha)	0.89	0.67	0.83
Fraction Area Harvested	0.88	0.6	0.79
Fraction Area Lost	0.88	0.6	0.79
Maize	0.89	0.66	0.82
Groundnuts	0.83	0.56	0.77
Mixed Beans	0.65	0.33	0.63
Popcorn	0.26	-0.23	0.31
Sorghum	0.31	0.01	0.33
Soybeans	0.69	0.34	0.63
Sweet Potatoes	0.79	0.43	0.7
Log Maize	0.93	0.73	0.86
Log Sweet Potato	0.76	0.41	0.67
Log Groundnuts	0.84	0.52	0.74
Log Soybeans	0.64	0.21	0.54
Bunding	0.01	-21.04	0.04
Monocrop	0.79	0.5	0.74
Mixture	0.12	-9.1	0.23
Proportion Till Plough	0.91	0.75	0.87
Proportion Till Ridge	0.87	0.61	0.8
Proportion No Till	0.08	-0.88	0.13
Proportion Hand Till	0.82	0.45	0.71
Proportion Monocrop	0.95	0.75	0.88
Proportion Mixed Crop	0.28	-0.55	0.33
Fraction Loss Drought	0.74	0.42	0.7
Fraction Loss Flood	0.18	-0.13	0.24
Fraction Loss Animal	0.11	-0.15	0.19
Fraction Loss Pests	0.15	-0.17	0.16
Fraction Loss Soil	0.27	-0.05	0.28
Fraction Loss Fertilizer	0.44	0.07	0.38
Categorical Variable Name	**FPR**	**AUC-ROC**	
Loss Indicator	0.3	0.89	
Drought Indicator	0.07	0.92	
Field Loss Indicator	NaN	NaN	
Animal Loss Indicator	NaN	NaN	
Pest Loss Indicator	NaN	NaN	

Variable Group
Harvested Area and Production
Crop Yields
Crop Loss Mechanisms
Agricultural Practice