Smallholder maize area and yield mapping at national scales with Google Earth Engine

Keywords: Sentinel-1, Sentinel-2, Data fusion, Cropland classification, Yield mapping

https://doi.org/10.1016/j.rse.2019.04.016
Remote Sensing of Environment
April 2019

Ziyi Guo

Background

Smallholder Farms
—dominating agriculture in many food-insecure regions, especially in Sub-Saharan Africa and South Asia
often lack detailed, accurate data on crop area and yields.

Traditional Ground Surveys
expensive and limited in scope
makes it difficult to assess productivity & guide policy for these critical systems.

LSMS-ISA:
Uses household interviews on inputs and outputs, relying on farmer recall and thus susceptible to bias.
One Acre Fund (1AF) Crop Cuts:
Physically measures small 5 × 8 m plots, harvests, and weighs maize, providing accurate yields but at high labor cost.
TAMASA:
Uses similar crop-cut protocols in Tanzania to measure maize yields on selected fields.
AFSIS:
Primarily a soil survey, but also records land-cover information (including crop presence) during field visits.
IPA Field Polygon Mapping:
Enumerators walk field boundaries with GPS to map precise shapes, though this is expensive at scale.
National Agricultural Surveys:
Government enumerators gather area and yield data via interviews, crop cuts, or observations, but often face limited budgets and under-sampling.

Background

Study Area: Kenya and Tanzania

Kenya and Tanzania rank among Africa's top maize producers, with large areas devoted to smallholder maize—often intercropped and using low inputs—across diverse climates.

Study Time Frame: Primary Maize Season

Each country has multiple rainy seasons, but this study focuses on the primary maize season (February–July in Tanzania, March–September in Kenya).

Data Collection Process

Table 1
Bands and spectral indices from Sentinel-1 and Sentinel-2 data.

Band or index	Central wavelength/index formula	Satellite
VV	Vertically polarized backscatter	Sentinel-1
VH	Horizontally polarized backscatter	Sentinel-1
VV_RLSPCK	VV with refined Lee speckle filter	Sentinel-1
VH_RLSPCK	VH with refined Lee speckle filter	Sentinel-1
RATIO	VH/VV	Sentinel-1
DIFF	$VV - VH$	Sentinel-1
RATIO_RLSPCK	VH_{RLSPCK}/VV_{RLSPCK}	Sentinel-1
DIFF_RLSPCK	$VV_{RLSPCK} - VH_{RLSPCK}$	Sentinel-1
AEROS	443 nm	Sentinel-2
BLUE	490 nm	Sentinel-2
GREEN	560 nm	Sentinel-2
RED	665 nm	Sentinel-2
RDED1	705 nm	Sentinel-2
RDED2	740 nm	Sentinel-2
RDED3	783 nm	Sentinel-2
NIR	842 nm	Sentinel-2
RDED4	865 nm	Sentinel-2
VAPOR	940 nm	Sentinel-2
CIRRU	1375 nm	Sentinel-2
SWIR1	1610 nm	Sentinel-2
SWIR2	2190 nm	Sentinel-2
NDVI	$(NIR - RED)/(NIR + RED)$	Sentinel-2
RDNDVI1	$(NIR - RDED1)/(NIR + RDED1)$	Sentinel-2
RDNDVI2	$(NIR - RDED2)/(NIR + RDED2)$	Sentinel-2
GCVI	$(NIR/GREEN) - 1$	Sentinel-2
RDGCVI1	$(NIR/RDED1) - 1$	Sentinel-2
RDGCVI2	$(NIR/RDED2) - 1$	Sentinel-2
MTCI	$(NIR - RDED1)/(RDED1 - RED)$	Sentinel-2
MTCI2	$(RDED2 - RDED1)/(RDED1 - RED)$	Sentinel-2
REIP	$700 + 40 * \left(\frac{RED + RDED3}{2} - RDED1\right)/(RDED3 - RDED1)$	Sentinel-2
NBR1	$(NIR - SWIR1)/(NIR + SWIR1)$	Sentinel-2
NBR2	$(NIR - SWIR2)/(NIR + SWIR2)$	Sentinel-2
NDTI	$(SWIR1 - SWIR2)/(SWIR1 + SWIR2)$	Sentinel-2
CRC	$(SWIR1 - GREEN)/(SWIR1 + GREEN)$	Sentinel-2
STI	$SWIR1/SWIR2$	Sentinel-2

Data Acquisition

Sentinel-1: Collect dual-polarized SAR data (VV & VH).
Sentinel-2: Gather multi-spectral optical imagery (13 bands).

Preprocessing

Sentinel-1:

Apply noise corrections, use a refined Lee filter to reduce speckle noise.
Derive backscatter values and compute the VH/VV ratio.

Sentinel-2:

Implement cloud and shadow masking using a custom decision tree.
Create seasonal median composites from the cleaned images.

Feature Extraction

From Sentinel-1: Extract backscatter measures and derived ratios.
From Sentinel-2: Compute vegetation indices (like NDVI and GCVI) and capture seasonal phenological changes.

Data Integration

Combine features from both datasets into a unified multi-sensor dataset.

Modelling Process

Collecting Ground Truth Data

Data Labeling: true value for crop/non-crop data.

Training Datasets

(Using the ground truth data & Sentirel1, 2)
Two separate classification tasks are defined:
- Cropland & non-crop areas.
- Within cropland areas, maize & other crops.

Training the Model

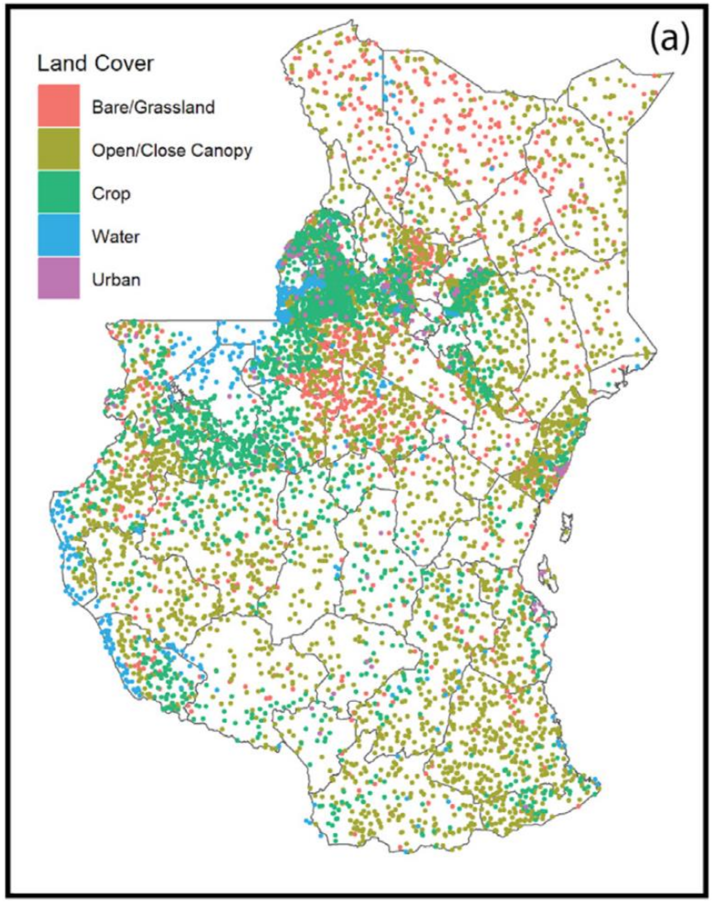
RF Model

Random Forest models are deployed.
- Classifies each pixel into cropland / non-crop
- Cropland are classified into maize / other crops.

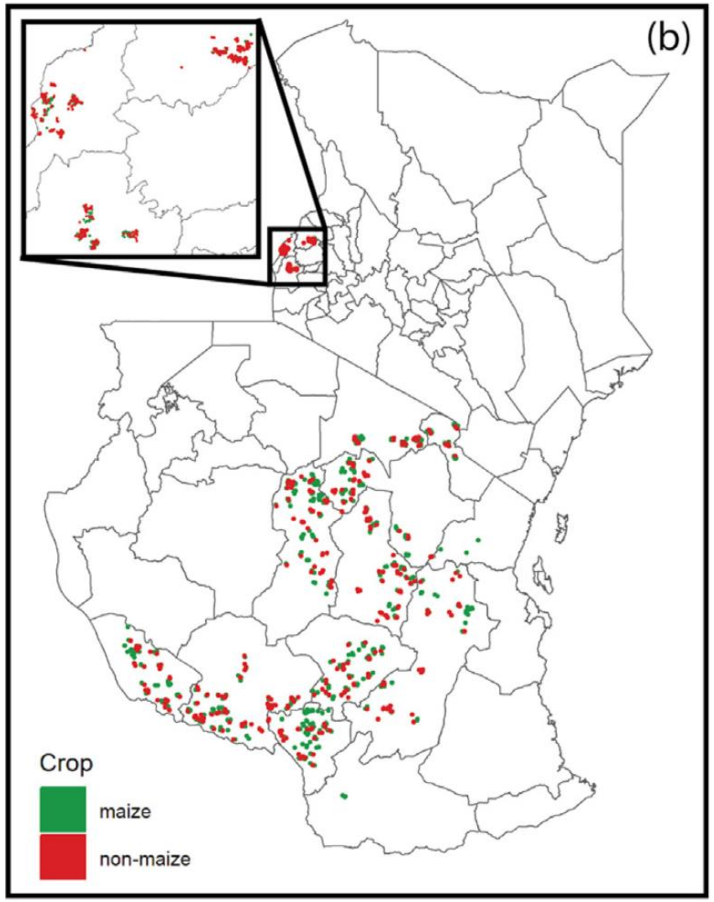
Validation

The cropland (crop vs. non-crop) model achieved about ***85% accuracy out-of-sample in both Kenya and Tanzania***.
The maize-specific model had a lower accuracy, with around ***67% accuracy in Kenya and 79% in Tanzania***

Country	Provider	Year	Geometry	Points	Crop type task
Kenya	1AF	2016	Points	3944	
Kenya	1AF	2017	Points	4202	
Kenya	IPA	2017	Polygons	16,964	Yes
Tanzania	AFSIS	2015	Points	2075	Yes
Tanzania	AFSIS	2016	Points	7999	Yes
Tanzania	AFSIS	2017	Points	35	Yes
Tanzania	TAMASA	2015	Points	136	Yes
Tanzania	TAMASA	2016	Polygons	2383	Yes
Tanzania	TAMASA	2017	Polygons	3219	Yes



1. Crop/Non-Crop



2. Maize/Non-Maize

Takeaway 1:
Integration of Multiple Data Sources and Data Fusion

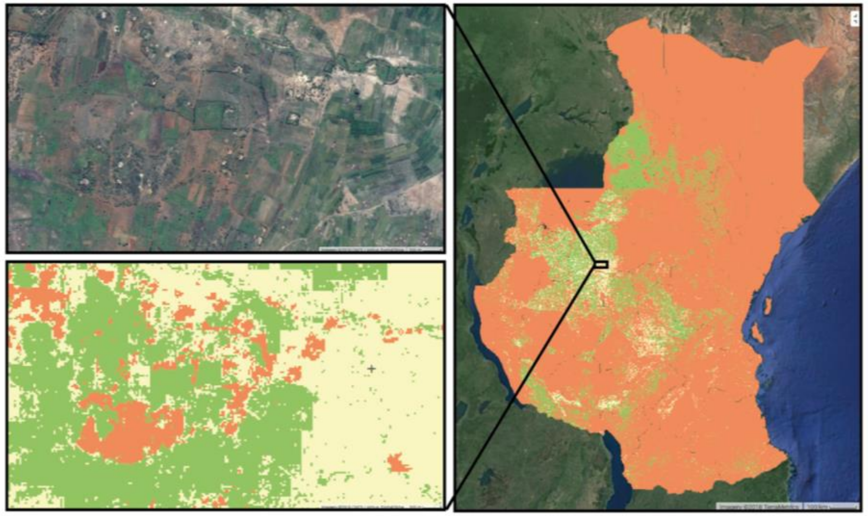
- **Complementary Sensor Strengths:**
  - Sentinel-1 Radar can capture ***field structure and moisture*** even under cloudy conditions, which is crucial for regions with persistent cloud cover.
  - Sentinel-2 Optical provides ***detailed spectral information*** (via indices like NDVI and GCVI) that accurately tracks crop health and seasonal growth patterns.

    By combining these, the researchers could ensure that if one sensor's data was compromised (clouds in Sentinel-2), the other (Sentinel-1) could still provide useful information.

- **Data Fusion Process:**
  - **Spatial Alignment**
    - Common Grid: Both Sentinel-1 and Sentinel-2 data are resampled or clipped to the same 10 m grid.
    - CRS: Ensures that each pixel in both datasets represents the same ground location.
  - **Temporal Matching**
    - Seasonal Windows: Imagery is grouped into key phases of the maize growing season
    - Median Composites: Within each window, Sentinel-1 and Sentinel-2 images are combined (e.g., median value per pixel) to reduce noise and cloud contamination.
  - **Merging into a Unified Dataset**

    By "stacking" these features together can create a vector like [VV-S1, VH-S1, NDVI-S2, GCVI-S2, ...] for that pixel.
    This unified dataset allows the machine learning model to ***consider all these different types of information simultaneously***

Takeaway 2:
Model Choice – Random Forest



Why choosing Random Forest?

- RF models have been successfully used in various satellite-based mapping studies (e.g., crop type mapping).
- They're less prone to overfitting because they average results across many decision trees.
- An implementation is available in Google Earth Engine (GEE), allowing the team to make large-scale, pixel-level predictions across Kenya and Tanzania.

Three K-Fold Validation Methods

- **Random Split** (Crop/Non-Crop Classifier)
  - The dataset is divided into K folds by randomly assigning data points to each fold, regardless of their geographical location.
  - contains a diverse mix of samples, often yielding a stable average performance estimate.
- **District-Based Split** (Crop/Non-Crop Classifier)
  - Each fold corresponds to data from different administrative districts, so training and validation data are drawn from non-overlapping areas.
  - tests the spatial generalization of the model
- **Field-Level Split** (within Crop, Maize/Non-Maize Classifier)
  - Data points from the same field are kept together in one fold.
  - Ensures the model isn't simply memorizing the exact spectral patterns from a field it already "saw" in training.

Takeaway 3:
Google Earth Engine Application

Access to Vast Satellite Archives

- GEE maintains ready-to-use repositories of Sentinel-1 and Sentinel-2 data, eliminating the need for users to download or store these massive datasets locally.

Built-In Machine Learning Capabilities

- GEE provides implementations of Random Forest and other algorithms that can be directly trained and applied on satellite imagery. This allows pixel-level classification (e.g., cropland vs. non-crop, maize vs. non-maize) without exporting data to external software.

User Interface API for Label Collection

- The paper specifically mentions a custom interface built within GEE that enabled researchers to visually inspect high-resolution basemaps and label points as “crop” or “non-crop.” These labeled datasets were then used to train the cropland classification model.

On-the-Fly Data Fusion and Preprocessing

- GEE made it straightforward to filter out contaminated (cloudy or hazy) pixels, apply a custom decision tree for cloud masking, and perform noise correction for radar data.
- These steps, crucial for generating clean, consistent inputs, are done at scale and integrated seamlessly with the classification workflow.

Rapid Prototyping and Iteration

- The authors note that being able to adjust parameters (e.g., cloud masking thresholds, date ranges, feature selection) in real time and see immediate results was essential for refining their methods quickly.

Questions:

1. How can researchers ensure the accuracy and representativeness of ground truth labels, especially when field surveys are sparse or prone to measurement errors?

2. What innovative validation methods (like drone imagery, sensor networks, or farmer self-reporting apps) could be employed in data-sparse regions to verify satellite-derived estimates?

3. What modifications (like additional ground data, different feature engineering) would be required to extend this approach to other crops, regions, or countries with different farming practices or climatic conditions?

4. How could socioeconomic data—like farm size, household income, or market access—be incorporated into satellite-driven models to better understand yield constraints and adoption of best practices?