Radiant MLHub
EARTH IMAGERY FOR IMPACT

Dataset Documentation

Dataset Name: Eyes on the Ground Image Data

Description

The "Eyes on the Ground" project (lacunafund.org) is a collaboration between ACRE Africa, the International Food Policy Research Institute (IFPRI), and the Lacuna Fund, to create a large machine learning (ML) dataset for crop phenology of smallholder farmer's fields. This is a unique dataset of georeferenced and timestamped crop images, which were captured using smartphone cameras following standardized "picture-based insurance" protocol along with labels on input use, crop management, crop growth stages, crop damage, and yield estimates, collected across eight counties in Kenya. The research leading to this dataset was undertaken as part of the CGIAR research program on Policies, Institutions and Markets (PIM). The project was a collaboration between ACRE Africa, International Food Policy Research Institute, Dvara E-Registry and KALRO.

Citation

Waithaka, L., Kramer, B., Hufkens, K., Kivuva, B., & Mansabdar, S. (2022) Eyes on the Ground Image Data, Version 1.0, (Version 1.0) Radiant MLHub. [Date Accessed] https://doi.org/10.34911/rdnt.1bs2jw

License

Creator(s)

LACUNA fund (https://lacunafund.org/)

ACRE Africa (https://acreafrica.com/)

KALRO (https://www.kalro.org/)

IFPRI (https://www.ifpri.org/)

CGIAR PIM (https://pim.cgiar.org/)

BlueGreen Labs (https://bluegreenlabs.org/)

Dvara E-Registry (https://www.dvaraeregistry.com/)

Contact

Lilian Waithaka (lwaithaka@acreafrica.com)

Radiant MLHub
EARTH IMAGERY FOR IMPACT

Location and boundaries

Overall Location Method

- ☐ Ground collection only
- ☐ Ground collection with boundary drawn using imagery
- ☐ Ground collection with spatial buffer added
- ☐ Boundary drawn from imagery
- ☑ Other __village bounding boxes using GADM36 vectors to ensure contributor privacy
- ☐ Unknown

GeoLocation Device

- ☐ Industrial grade GPS (List model) _____
- ☐ Retail grade GPS
- ☑ Mobile Phone GPS
- ☐ N/A
- ☐ Unknown

Ground Boundary Method (Details explained in Appendix A)

- ☐ Live/Continuous point capture of walk-around
- ☐ Manual point capture of walk-around
- ☐ Manual point capture of polygon boundaries (not whole field)
- ☐ Manual point capture for later image annotation
- ☐ Manual point capture for spatial buffer within field
- ☐ Manual point capture while looking at but not in field, with heading recorded
- ☑ Other _village bounding boxes using GADM36 vectors to ensure contributor privacy___
- ☐ Unknown

Imagery used (Skip if no imagery used)

Sensor: _____

Date(s): _____

Imagery Annotation methods

- ☑ Boundaries drawn based on a single ground point captured
- ☐ Boundaries drawn/edited based on multiple ground points captured
- ☐ Buffer validated from ground point captured
- ☐ Boundary drawn without ground reference data
- ☐ Pixels annotated without ground reference data

- ☐ Unknown

Boundary inclusion

- ☑ Captured polygon includes the entire field/area
- ☐ Captured polygon includes only a sample of the field/area
- ☐ N/A

Classification

Classification Type

- ☐ Land cover
- ☑ Crop type
- ☑ Other ___growth phase (phenology) and disturbances (drought, pest, weeds) ____

Classes/fields used

Describe in Appendix B

Ground Referenced Classification

- ☐ Observation (Describe methods of determination in Appendix C)
- ☐ Survey/interview with land holder (Describe methods in Appendix C)
- ☑ Other (Describe methods in Appendix C)

Image Referenced Classification

Not Applicable

Data Properties

Property name	Property Description	Parameters/Allowed responses (optional)
farmer_unique_id	*Farmer ID*	
site_id	Site / field ID	
crop_name	Crop in the field	
sowing_date	Date of sowing as self reported by the farmer	YYYY-MM-DD
expected_yield	Expected yield as estimated by the farmer	Kilograms (Kg)

season	Crop growth season (SR or LR followed by the year, where SR is short rain and LR is the long rain season)	
spatial_location	Village name (as defined in GADM36)	
spatial_unit	Village source information (GADM36)	
acquired	Data of image acquisition	YYYY-MM-DDTHH:MM:SS
ML labels (manual)		
growth_stage	Phenological growth stage	S (sowing) F (flowering) V (vegetative) M (maturity) NA (not classified)
damage	Crop damage	DGT (drought) WED (weed) WND (wind) DSE (disease) FLD (flooding) PS (pest) ANM (Animals) ND (nutrient deficit) GC (Good Crop) NA (not classified)
Extent	The extent of the damage	0 - 1 (or NA)
ML labels (extended)		
drought_probability	Drought probability	0 - 1 (or NA)
drought_extent	Drought exent	0 - 1 (or NA)
growth_sowing	Growing phase (sowing)	0 - 1 (or NA)
growth_vegetative	Growing phase (vegetative)	0 - 1 (or NA)
growth_flowering	Growing phase (flowering)	0 - 1 (or NA)
growth_maturity	Growing phase (maturity)	0 - 1 (or NA)
disturbance_none	Crop disturbance (none)	0 - 1 (or NA)
disturbance_drought	Drought disturbance	0 - 1 (or NA)
disturbance_weed	Weeds	0 - 1 (or NA)
disturbance_nutrient_deficit	Nutrient deficit	0 - 1 (or NA)

Radiant MLHub
EARTH IMAGERY FOR IMPACT

Appendix A: Describe the method of geographic ground data collection

All original images were geo-located using the GPS internal to smallholder farmer cellphones. To ensure privacy exact locations are removed and only bounding boxes of the village in which the field is located are reported (along with the village name). Village names are sourced from the GADM36 dataset. For a full description of the Picture Based Insurance scheme we refer to Ceballos, Kramer and Robles (2019, https://doi.org/10.1016/j.deveng.2019.100042).

Appendix B: List all top-level classes or the classification guidance used

Images were assigned labels based on growth phases and type and extent of the damage. This includes various disturbances due to drought, weeds and pests as well as their extent (intensity).

Appendix C: Describe methods for determining classes based on direct/ground observation

All cellphone images were classified by agronomist or trained professionals to ensure consistency and quality of the manual labels. Additional automatic labels were assigned using an in house ML algorithm if manual labels were not collected. Probabilities on automatic labels are derived from various independent ML algorithms and therefore do not sum to unity, but can provide valuable information in augmentation efforts.

Appendix D: Ancillary data

Due to the lack of precise geolocation of the cellphone images (for privacy reasons) we provide point based subsets of Sentinel 2, precipitation product (ARC2 / TAMSAT) and ERA5 data covering the whole campaign (2020 - 2021). This should allow for a comparison between remotely sensed imagery at the site location - despite the absence of precise field locations. Village or regional level data can still be used, using either village identifiers or their bounding boxes.

Sentinel 2 data is provided as raw band values and quality control labels, in particular bands: B1, B2, B3, B4, B8, AOT, SCL, QA60 (with multipliers as listed on Google Earth Engine). We provide daily precipitation estimates as extracted from two remote sensing products. We use both TAMSAT (Maidment et al. 2017, Tarnavsky et al. 2014, Maidment et al. 2014), where "TAMSAT produces daily rainfall estimates for all of Africa at 4km resolution." In addition, daily precipitation values are provided through ARC2 data as provided by NOAA. Data is extracted at point locations, for units and further processing we refer to the product websites. Finally, we provide hourly ERA5 values (on a single level) for mean sea level and surface pressure, 2m air temperature, wind speed components, precipitation and 2m dew point temperature, we refer to the Copernicus Climate Data Store (Hersbach et al. 2018) for an in depth description of the products.

References:

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., Thépaut, J-N. (2018): ERA5 hourly data on single levels from 1979 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). (Accessed on < DD-MMM-YYYY >), 10.24381/cds.adbb2d47

Maidment, R. I., D. Grimes, E. Black, E. Tarnavsky, M. Young, H. Greatrex, R. P. Allan et al. (2017). A new, long-term daily satellite-based rainfall dataset for operational monitoring in Africa Nature Scientific Data 4: 170063 DOI:10.1038/sdata.2017.63.

Tarnavsky, E., D. Grimes, R. Maidment, E. Black, R. Allan, M. Stringer, R. Chadwick, F. Kayitakire (2014). Extension of the TAMSAT Satellite-based Rainfall Monitoring over Africa and from 1983 to present Journal of Applied Meteorology and Climate DOI 10.1175/JAMC-D-14-0016.1

Maidment, R., D. Grimes, R.P.Allan, E. Tarnavsky, M. Stringer, T. Hewison, R. Roebeling and E. Black (2014). The 30 year TAMSAT African Rainfall Climatology And Time series (TARCAT) data set Journal of Geophysical Research DOI: 10.1002/2014JD021927