A region-wide, multi-year set of crop field boundary labels for Africa

Technical Report on Label Development and Processing

Wussah, A.[1], Asipunu, M.[1], Gathigi, M.[2], Kovačič, P.[2], Muhando, J.[2], Yeboah, V.[1], Addai, F.[1], Akakpo, E.S.[1], Allotey, M.[1], Amkoya, P.[2], Amponsem, E.[1], Dadon, K.D.[1], Harrison X.G.[1], Heltzel, E.[3], Juma, C.[2], Mdawida, R.[2], Miroyo, A.[2], Mucha, J.[2], Mugami, J.[2], Mwawaza, F.[2], Nyarko, D.[1], Oduor, P.[2], Ohemeng, K.[1], Segbefia, S.I.D.[1], Tumbula, T.[2], Wambua, F.[2], Yeboah, F.[1], and Estes, L.D.[3]

[1] Farmerline, Kumasi, Ghana
[2] Spatial Collective, Nairobi, Kenya
[3] Clark University, Worcester, MA, USA

Table of contents

1 Background

This project was undertaken to provide a comprehensive set of annotated, high resolution satellite imagery that indicate the boundaries and extent of crop fields throughout the continent of Africa, across multiple years (2017-2023), in order to enable the training and assessment of models designed to map agricultural fields in remote sensing imagery. Accurate cropland mapping, in particular field boundary delineation, is important for applications ranging from food security assessment to environmental impact assessment (Fritz et al. 2015, e.g. Estes et al. 2022, Rufin et al. 2023).

The imagery selected for this project were Planet imagery provided through Norway's International Climate and Forest Initiative (NICFI n.d.), which are mosaics provided at <5 m resolution at monthly to 6-monthly time steps from 2016 until present. These data, developed as they are from daily PlanetScope imagery, provide a capability that is not currently provided by other observing platforms, in that the imagery has sufficient spatial resolution for a human observer to identify a large proportion of the smaller agricultural fields that predominate throughout the continent, while providing sufficient temporal resolution to provide annual to sub-annual over wide areas (Estes et al. 2022). These combined characteristics–high spatial and temporal resolution and continental coverage–enable the collection of a sample that accounts for the immense geographic diversity in agricultural systems, as well as the temporal shifts that occur between years. Accounting for both factors is important in training and assessing the machine learning models used to map agricultural fields.

2 Methods

2.1 Sample selection

We collected a sample of imagery from likely cropland areas (as determined by an existing cropland layer; Potapov et al. (2022)) across continental Africa south of the Sahara where annual rainfall is above 150 mm, down to 30° latitude in South Africa, the southern extent of availability for NICFI basemaps. To create the label set, we draw a random sample of >37,000 cells from a ~500 m (0.005°) pre-existing sampling grid (see Estes et al. 2022), stratified by 9 different agro-ecoregions (ranging from Arid to Humid, as defined by the FAO), and selected from cells having a minimum level of cropland cover, as determined by the input cropland layer, to ensure to ensure that most labels had a mixture of cropland and non-cropland in them, and that the number of purely negative (non-cropland) labels were minimized.

2.2 Image processing

The selected samples cells were randomly assigned to one of the 7 years in the pre-determined study timeframe (2017-2023), with a month assigned that corresponded the least cloudy month for the given location. Cloudiness was determined by calculating the monthly frequency of bad quality MODIS pixels for the year 2022, with imagery accessed through the Google Earth Engine platform (Gorelick et al. 2017). We then used a set of pythonroutines to query the Planet API for each location and its corresponding date in the sample. The Planet NICFI quads intersecting a larger 0.0592° tile that each sample cell was nested in was downloaded, cropped, reprojected to geographic coordinates, and resampled to a 0.000025° resolution (approximately 3 m). The resampling enabled slightly better visual identification of field boundaries.

We then further processed the imagery by normalizing each image band within its 1st and 99th percentile ranges, which further improved image contrast and the ability to distinguish fields. The normalized images were converted to Cloud-Optimized Geotiffs, and uploaded to SentinelHub using the Bring Your Own Cog (BYOC) service, from where they were accessed

using the Web Map Service (WMS) protocol. We gratefully acknowledge credits provided by the European Space Agency and SentinelHub for use of the service.

2.3 Label collection

Labels were collected using two methods. The primary method was through a custom cloud-hosted labelling platform designed specifically for field boundary delineation on Planet imagery (see Estes et al. 2016, Estes et al. 2022). This platform was not immediately accessible, as it had to be redesigned to work with SentinelHub's WMS services (it previously used Element84's rasterfoundry service). While this was being established, a team of expert labellers who were experienced spatial analysts using QGIS to collect samples. The basic labelling task entails mapping the fields visible in the processed Planet imagery within the target box, which represents the boundaries of the initial 0.005° grid sample Figure 1. The objective was to delineate boundaries of recently active fields that appeared to be under annual cultivation (i.e. planted to field crops), with fields interpeted to be inactive, abandoned, or fallow left unlabelled.

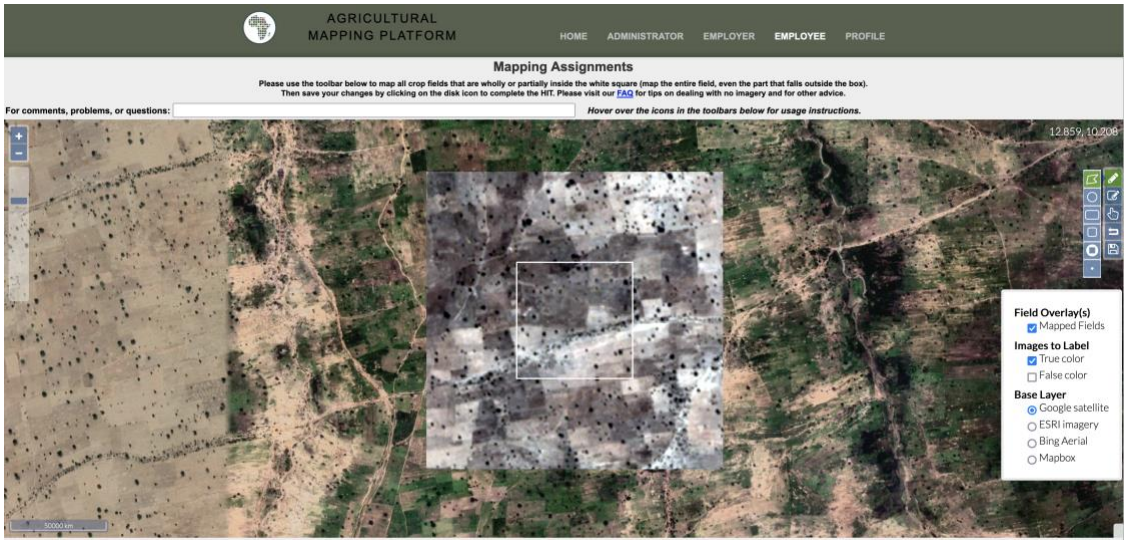

Figure 1: A view of labeller's interface, showing the target box to label, and a true color rendering of the Planet imagery to label at that location. Labellers were provided with a larger extent of the imagery, in order to provide additional context. Imagery was also presented in false color rendering, and a variety of virtual globe basemaps provided higher resolution views to aid interpretation.

Sites were randomly separated into the following categories:

- Class 1: Expert-labeled sites that were used as quality control (Q) in the labelling platform (n=2000);
- Class 2: Individual samples sites mapped one time by one individual labeller (n=31,000);
- Class 4: Individual samples sites mapped one time each by three different labeller (n=1,000).

Initially, a Class 3 sample, which represented locations with up to 6 years of imagery to be labelled, were selected, but these were determined after beginning to be impractical to label, given the substantial variability in image interpretations. Class 3 locations where therefore re-allocated to Class 2 to boost that sample size.

Class 1 labels were collected by a team of 5 individual experts using QGIS. A subset of their samples were reviewed and graded for quality by a separate set of geospatial analysts at Clark University, and 800 were selected to load into the labelling platform prior to the collection of Class 2 and 4 sites. Class 1 sites were used as quality control (Q) assignments, slipped surreptiously into each labeller's workflow, in order to measure differences between the two sets of collected labels.

2.4 Quality control

The labelling task was inherently difficult, as the nature of the target-smallholder-dominated cropland–and the resolution of the imagery–high resolution, but still coarse enough so that it can be difficult to discern boundaries–made it such that there are often multiple reasonable interpretations as to what constitutes a field. We therefore applied several different forms of quality control to assess label quality during the course of the project.

2.4.1 Training

The labelling teams were hired following an initial selection process in which their ability to map a small number of sites provided in a QGIS project was assessed. After hiring and before commencing ordinary mapping assignments, labellers were provided with an inital round of training by the managements teams, and were required to pass a 10-site qualification test provided on the labelling platform, in which each person's maps were assessed against previously labelled boundaries using the system's quality control routines.

2.4.2 On platform quality control

During labelling, each team member's work was assessed occassionally against Class 1 labels that were automatically assigned by the platform's assignment scheduler. The scheduler randomly assigned different labelling tasks to individual labellers according to a frequency parameter assigned to each label class. Quality control tasks (Q assignments) were initially served to labellers at a rate of 1 in 10 assignments, dropping to 2 in 100 assignments towards the end of the labelling period. Q assignments were assessed against Class 1 labels using a multi-dimensional metric Figure 2. For this project, 4 of 5 metrics were applied in the overall Score calculated after each Q assignment was completed, which was calculated as follows:

$$\text{Score} = 0.55\text{Area} + 0.225\text{N} + 0.1\text{Edge} + 0.125\text{Categorical}$$

Here Area measures how well the areas delineated by the labeller digitized polygons agreed with the Class 1 polygons, as the sum of area of agreement for the field (positive agreement) and non-field (negative agreement) areas divided by the total area of the labelling target (which sums positive and negative agreement and positive and negative disagreement). N represents the agreement in the number of fields digitized by the labeller and the expert (Class 1), while Edge is a measure of how close the boundaries of the labeller's polygons were to those of the Class 1 label, and Categorical is the agreement between the class value assigned by the labeller and the expert. The 5th category in the algorithm, the agreement in area of fields extending outside the target, was not used in this assessment.

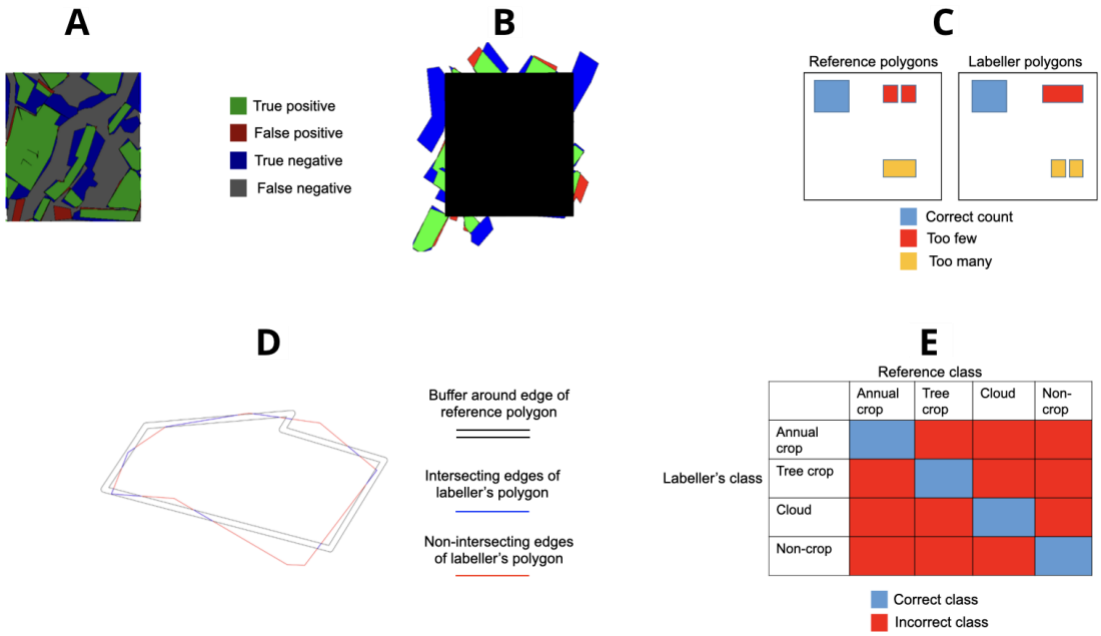

Figure 2: The 5 dimensions used to assess label accuracy relative to Category 1 reference polygons. A = Correctly labelled areas inside the target cell; B=Correctly labelled portions of fields that extend outside the cell (not used in this project); C=Fragmentation accuracy, or the agreement between the number of fields mapped by the reference (Class 1) label and the labeller; D = Edge accuracy; C = Categorical accuracy.

2.4.3 Label reviews

In addition to the on-board quality metrics assessed, a parallel review process was undertaken by the project supervisory team. Two supervisors undertook independent reviews of randomly selected assignments using a standalone Jupyter Lab environment that provided a rendering of the imagery and digitized labels for each selected site. Each site was given a score of 0-4, with 0-3 indicating increasing levels of accuracy in assignments where fields were present in the imagery, and 4 indicating sites where no fields were present that were correctly left undigitized by the labelling team.

3 Results

3.1 Total assignments

The main labelling tasks (Classes 2 and 4) were undertaken between November, 2023 and completed in mid-March, 2024. A total of 42403 assignments were completed, broken down into the classes detailed in Table 1. Class 1 assignments were divided into 4 sub-categories, with Class 1a representing the subset of Class 1 samples that were assessed as being of passing quality by an initial independent review, and served as the reference labels for the Q type assignments on the platform. Class 1b were the remaining Class 1 assignments completed by the expert teams, which together sum to >2000 as each expert was assigned an overlapping set of 50 samples, which was used to assess agreement among experts. Class 1c were the total number of Q assignments completed by main labelling teams, i.e. those scored against the Class 1a labels. Finally Class 1d assignments corresponded to the Class 1b sites that were re-mapped by 1-3 members of the primary labelling teams, to provide extra assignments to compare to the initially mapped Class 1 labels that did not get loaded into the platform.

Table 1: Number of assignments mapped per Class.

Class	n
1a	797
1b	1376
1c	2598
1d	2433
2	32167
4	3032

3.2 Overall Quality

The overall quality of the labels was assessed using several metrics, which are presented in detail in the detailed analysis of assignments, which provides overall and weekly average quality score metrics for each labeller, as assessed using the platform quality control algorithm and the independent label reviews Table 2.

Table 2: The average score against the various quality metrics assessed during the project. Qscore is the weighted mean of N, Edge, and Area, which were calculated using the platform's quality control algorithm, while Rscore is overall proportion of assignments reviewed by expert as passing.

Qscore	N	Edge	Area	Rscore
0.61	0.33	0.05	0.75	0.7

An additional assessment of agreement between the two supervisors was conducted in which the two experts ratings of assignments were compared for an overlapping set of 190 assignments out of 4348 reviewed assignments. The two experts agreed 46.8% of the time along the 5 category rating system, and 74% of the time on a simplified binary version (ratings 0-1 grouped as failing, and 2-4 as passing). The overall mean rating difference was 0.12 (expert 2 - expert 1 rating), indicating no appreciable rating bias between experts. A total of 2999 expert reviewed assignments were rated as passing, of which 1787 were given the highest ratings (3-4).

3.3 Label Outputs

The resulting labels were extracted from the platform's database and post-processed to produce a label catalog, containing the details of each assignment, as detailed in Table 3.

Table 3: The names and descriptions of variables provided in the label catalog.

Variable	Description
name	Unique site ID, prefixed with two character ISO country code
Class	Label class (1a-1d, 2, 4)
assignment_id	Identifier for each unique mapping assignment (1 mapping by 1 labeller)
Labeller	Anonymous identifiers for each labeller
completion_time	Date and time the labelling assignment was completed
label_time	Total time spent on the assignment
status	A system assigned value, including "Rejected" (failed Q assignment), "Untrusted"

	(assignment completed during time when labeller had low rolling average Q score); “Approved” assignment passing the Q threshold, or non-Q assignment passed when labeller’s average Q score was above the quality threshold
Score	Weighted mean Quality score, comprised of N, Edge, Area, and Categorical metrics derived from the Class 1a labels.
N	Agreement between number of fields mapped by a labeller and the corresponding Class 1a labels
Edge	Nearness of labeller’s field boundaries to those in a corresponding Class 1a label set
Area	Agreement between a labeller’s mapped area of fields and non-fields and those of the corresponding Class 1a labels
FieldSkill, NoFieldSkill	A Bayesian metric of a labeller’s skill in mapping field and non-field areas, respectively (see Estes et al. 2022)
Categorical	Agreement in the label assigned to each polygon delineated by the labeller and those in the Class 1a labels.
rscore	A 1-4 ranking assigned to a given assignment by a supervising expert during independent review. Note: decimal values appear for cases where the same assignment was assessed more than once by experts
rscore2	A simplified binary version of rscore, where 0 indicates failing and 1 indicates passing.
Qscore	Each labeller’s overall average Score, assigned as a general confidence measure applied to all Class 2 and 4 assignments undertaken by the labeller
QN	Each labeller’s overall average N score, to provide a general measure of each labeller’s tendency to over or under-segment fields (lower score typically mean under-segmentation).
Rscore	Each labeller’s overall average rscore2, assigned as a second generalized measure of confidence. Experts (except one) were assigned the same measure, based on initial reviews conducted by a separate team at Clark
x, y	The centroid of each site, in decimal degrees
farea	The average area (in ha) of fields digitized in each assignment
nflds	The average number of fields digitized in each assignment
tile	The unique identifier of the 0.05° image tile containing the labelling site
image_date	The collection date of the image being labelled. The month and day represent the central date of images from August, 2020 and earlier, which were drawn from 6-month composites, while later images were collected in the month indicated, with the 15th being the central date for the month
chip	The chip identifier (a concatenation of name and image-date)
label	(In the demonstration catalog only). The identifier of rasterized label chip, concatenating the name, image-date, and assignment_id.

The images labelled for each site were processed into chips of 224 X 224 pixels, cropped to the borders of the labelling target. A full label catalog was developed from the assignments with the highest Rscore for each site, with the assignment process and label rasterization process demonstrated in a companion Jupyter notebook that can be adapted to recreate labels using different criteria. The mapped field polygons (n=825,395) accompanying each assignment are provided in a separate geoparquet file.

The spatial distribution of sites, broken down by the main assignment types, and associated variables (average field area, number of times mapped) are shown in Figure 3.

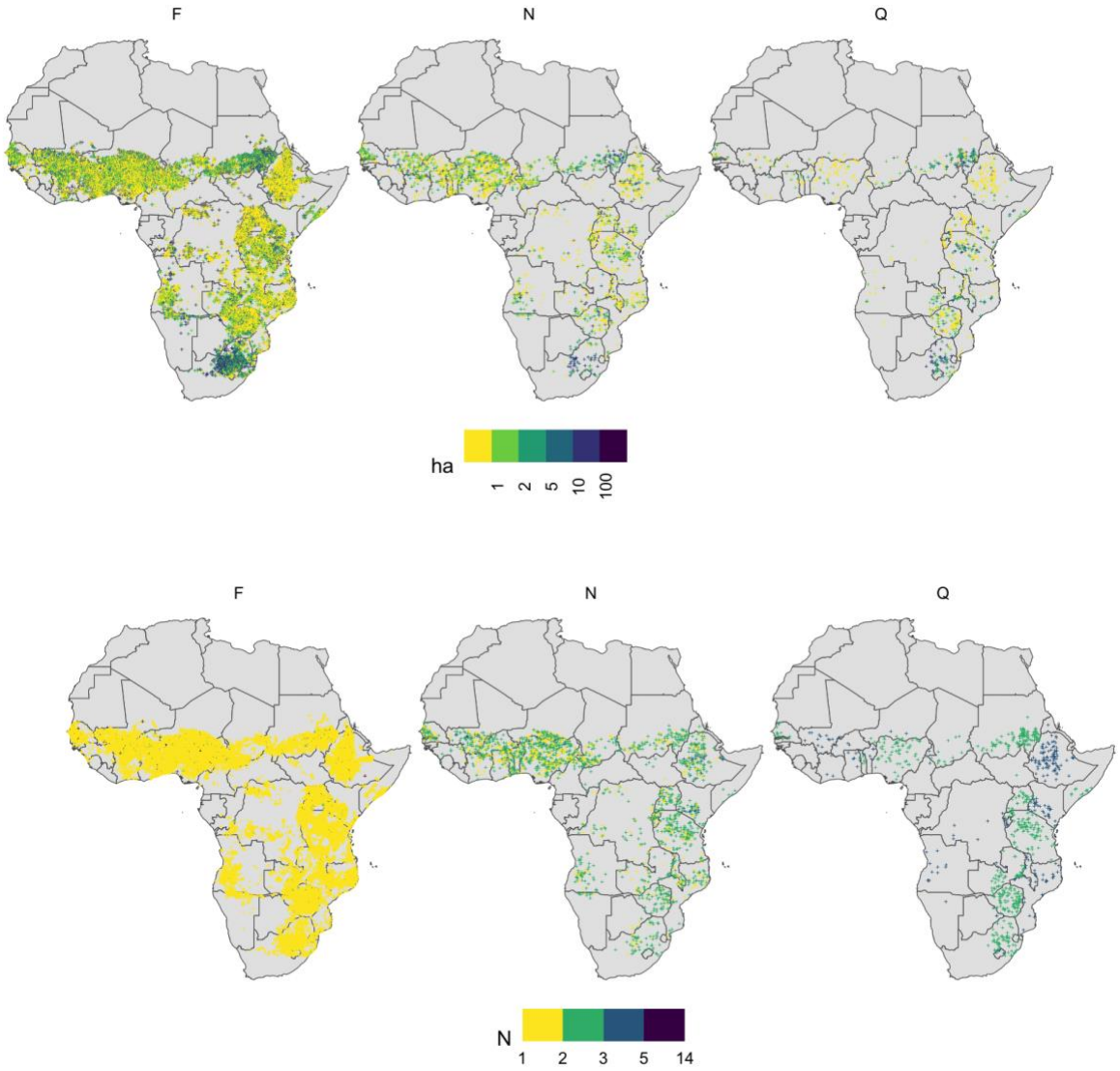

Figure 3: The distribution of mapped sites across the continent, showing the average area (in ha) of fields digitized for each site (top), broken down according to assignment type (F = Class 2, Q = Class 1c, N = Class 4), and the number of times each site was labelled (bottom).

4 Applications and Usage Guidelines

The provided data can be used in a variety of different ways to train and assesing machine learning models. As one example, the demonstration label catalog was created to provide input for boundary aware semantic segmentation models (see Khallaghi et al. 2023), in which the model learns to distinguish between the boundaries and interiors of fields.

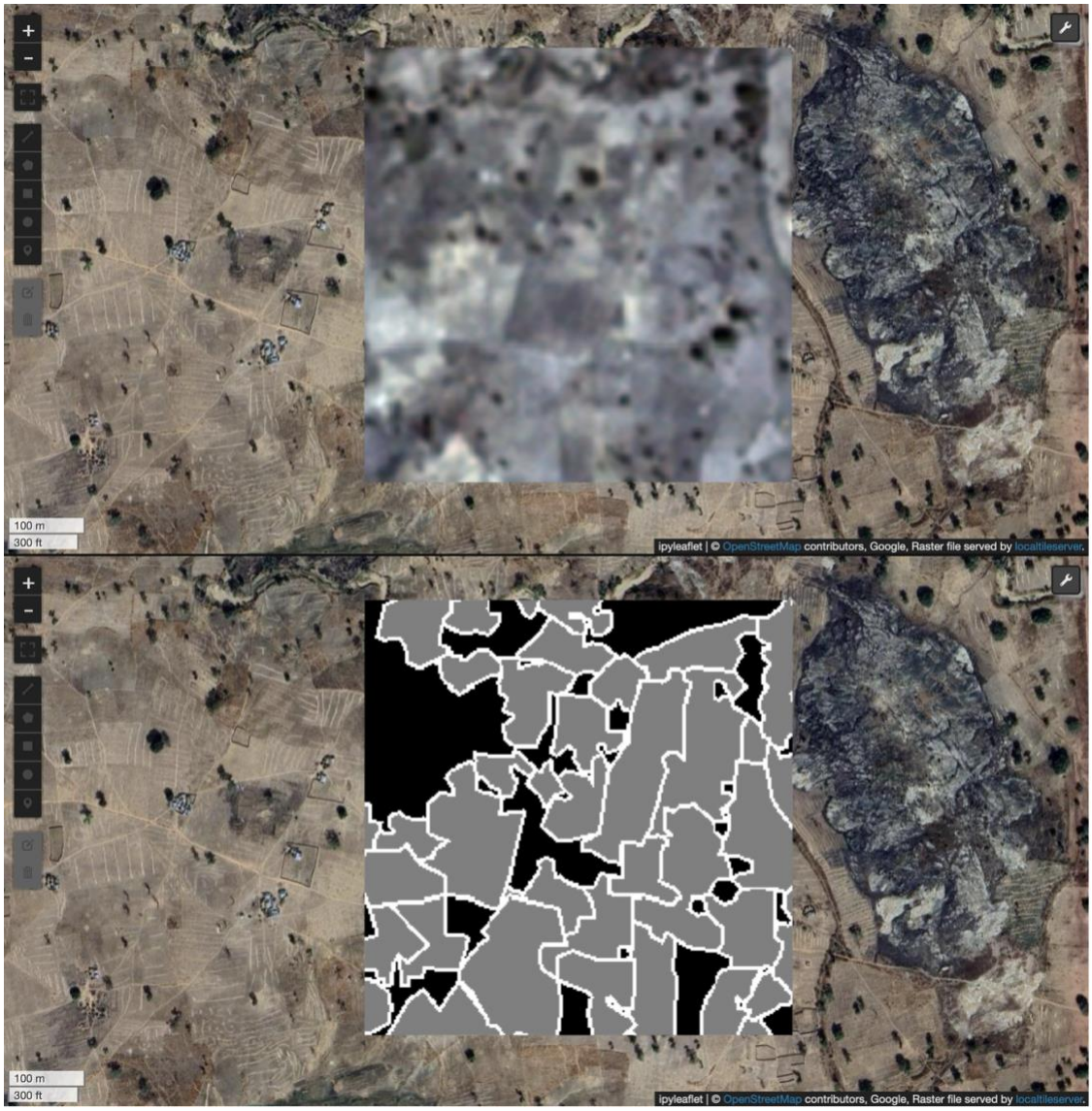

Figure 4: An example of an image chip and the corresponding label, rasterized into 3 classes, representing non-field, field interior, and field edge.

The full catalog may also be used to test the impact of label quality on the overall performance of models, which is a recommended practice for developing machine learning models (Elmes et al. 2020). In addition to the measures of label quality that are associated with each assignment, additional measures associated with label quality can be derived from those Class 1 and Class 4 labels that had 3 or more assignments were completed for each site. The labels can be combined into single maps representing the labeling frequency for each class. A simple approach is to rasterize each assignment, assigning a value of 1 for any pixel included in a field, and 0 for any non-field pixel, then sum the rasters and divide by the total number of assignments to calculate the frequency that each pixel was labelled as a field, which can serve as a simple measure of confidence. A more sophisticated approach is to weight each assignment by one of its corresponding quality metrics, or to use a Bayesian-based approach that assesses label risk (Estes et al. 2022).

The labels could be further converted to simple binary crop/non-crop labels for a general purposes semantic segmentation model, or used to extract a point based crop/non-crop sample to train models that do not require spatial context.

The label quality measures may also be used to define specific label subsets. For example, lower quality labels could be used to initially train a model, while the highest quality labels could be divided into subsets used for model fine-tuning and final validation (Burke et al. 2021).

4.1 Appropriate Usage

It is important that users of these labels realize their limitations in terms of how close they are to the “truth”. Image-based labelling of satellite imagery is inherently an error-prone endeavour, particularly over complex, smallholder-dominated agricultural landscapes. In many cases, expert labellers may come to different, but reasonable, interpretations as to which qualifies as a field and what doesn’t, making the practice of labeling is as much an art as a science. The overall quality metrics and measures of between-labeller agreement indicate that 70-75% appears to be the upper bound of confidence when it comes to labelling the total extent of fields. The values are much lower for measures of individual fields, such as the total number and their individual edges.

Users should therefore apply these labels with the understanding that the overall performance of their models should be assessed with these upper limits of knowable accuracy (Elmes et al. 2020). More confident measures of accuracy may be obtained from labels developed on higher resolution imagery (Wang et al. 2022), such as WorldVew-3, but the ability to compile a continent-wide, multi-year dataset of such imagery that would enable the creation of such a catalog is practically out of reach.

4.2 Data and Code Availability

The resulting label data and associated catalogs are available for download from Zenodo (10.5281/zenodo.11060871) and in this repository. Pending approval of the results, the data will be published on Source Cooperative. The code used to process the imagery and post-process and analyze the labels is available at the lacunalabels repository. The source code for the labelling platform is on the labeller repository.

5 Acknowledgements

This project was supported by the Lacuna Fund.

6 References

Burke, M., A. Driscoll, D. B. Lobell, and S. Ermon. 2021. Using satellite imagery to understand and promote sustainable development. Science 371.

Elmes, A., H. Alemohammad, R. Avery, K. Caylor, J. R. Eastman, L. Fishgold, M. A. Friedl, M. Jain, D. Kohli, J. C. Laso Bayas, D. Lunga, J. L. McCarty, R. G. Pontius, A. B. Reinmann, J. Rogan, L. Song, H. Stoynova, S. Ye, Z.-F. Yi, and L. Estes. 2020. Accounting for training data error in machine learning applied to Earth Observations. Remote Sensing 12:1034.

Estes, L. D., D. McRitchie, J. Choi, S. Debats, T. Evans, W. Guthe, D. Luo, G. Ragazzo, R. Zempleni, and K. K. Caylor. 2016. A platform for crowdsourcing the creation of representative, accurate landcover maps. Environmental Modelling & Software 80:41–53.

Estes, L. D., S. Ye, L. Song, B. Luo, J. R. Eastman, Z. Meng, Q. Zhang, D. McRitchie, S. R. Debats, J. Muhando, A. H. Amukoa, B. W. Kaloo, J. Makuru, B. K. Mbatia, I. M. Muasa, J. Mucha, A. M. Mugami, J.

M. Mugami, F. W. Muinde, F. M. Mwawaza, J. Ochieng, C. J. Oduol, P. Oduor, T. Wanjiku, J. G. Wanyoike, R. B. Avery, and K. K. Caylor. 2022. High resolution, annual maps of field boundaries for smallholder-dominated croplands at national scales. Frontiers in Artificial Intelligence 4:744863.

Fritz, S., L. See, I. McCallum, L. You, A. Bun, E. Moltchanova, M. Duerauer, F. Albrecht, C. Schill, C. Perger, P. Havlik, A. Mosnier, P. Thornton, U. Wood-Sichra, M. Herrero, I. Becker-Reshef, C. Justice, M. Hansen, P. Gong, S. Abdel Aziz, A. Cipriani, R. Cumani, G. Cecchi, G. Conchedda, S. Ferreira, A. Gomez, M. Haffani, F. Kayitakire, J. Malanding, R. Mueller, T. Newby, A. Nonguierma, A. Olusegun, S. Ortner, D. R. Rajak, J. Rocha, D. Schepaschenko, M. Schepaschenko, A. Terekhov, A. Tiangwa, C. Vancutsem, E. Vintrou, W. Wenbin, M. van der Velde, A. Dunwoody, F. Kraxner, and M. Obersteiner. 2015. Mapping global cropland and field size. Global Change Biology 21:1980–1992.

Gorelick, N., M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore. 2017. Google earth engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment 202:18–27.

Khallaghi, S., J. R. Eastman, and L. D. Estes. 2023. A review of technical factors to consider when designing neural networks for semantic segmentation of earth observation imagery.

NICFI. 2023. Norway's international climate and forest initiative. https://www.planet.com/nicfi/

Potapov, P., S. Turubanova, M. C. Hansen, A. Tyukavina, V. Zalles, A. Khan, X.-P. Song, A. Pickens, Q. Shen, and J. Cortez. 2022. Global maps of cropland extent and change show accelerated cropland expansion in the twenty-first century. Nature Food 3:19–28.

Rufin, P., S. Wang, S. N. Lisboa, J. Hemmerling, M. G. Tulbure, and P. Meyfroidt. 2023. Taking it further: Leveraging pseudo labels for field delineation across label-scarce smallholder regions. http://arxiv.org/abs/2312.08384

Wang, S., F. Waldner, and D. B. Lobell. 2022. Unlocking large-scale crop field delineation in smallholder farming systems with transfer learning and weak supervision. Remote Sensing 14:5738.