TETRA TECH

Enabling Crop Analytics At Scale

ENABLING CROP ANALYTICS AT SCALE (ECAAS)

Creating Open Agricultural Maps and Ground Truth Data to Better Deliver Farm Extension Services

PREPARED BY

CLARK UNIVERSITY

IN COLLABORATION WITH

Contents

1 Introduction

This document presents a scale-up plan for the 'Enabling Crop Analytics at Scale' project. The current project is designed to support smallholders who lack access to the inputs and information that can help boost their productivity and resilience to major sources of volatility in some regions of Ghana. The project uses cropland boundary data generated by an advanced cropland mapping system to enable the extension team to collect crop type observations on the ground, which are then used to develop crop type maps from farmers. The process of making the maps entails combining:

1. An agricultural extension team equipped with an award-winning, high-tech farmer engagement platform; Mergdata, with
2. An advanced cropland mapping system.
   The process of generating this essential ground truth and map data will help Farmerline to improve the quality and reach of our extension services as this project will fold into our existing Farmer services (B2F) and Business to Business (B2B) models.

The first project focuses on four districts covering two seasons: Ejura Sekyedumase, Sekyere West, Nkoranza, and Tain (enclosed by the red outline in Figure 1.1). During the first season, field collection efforts focused on Ejura Sekyedumase and Sekyere West districts (orange in Figure 1.1). For the second season, data collection was expanded to the Tain and Nkoranza districts (yellow in Figure 1.1). The crop types considered for both collections were maize, rice, and a general grouping of other crop types, including vegetables such as pepper and okra. The scaling-up plan describes improvements that we propose to implement, based on lessons learned during the first season, and a plan for extending the project scope to cover a broader region that includes 31 districts in northern Ghana, including the Northern, Upper East, and Upper West regions (blue outline in Figure 1.1).

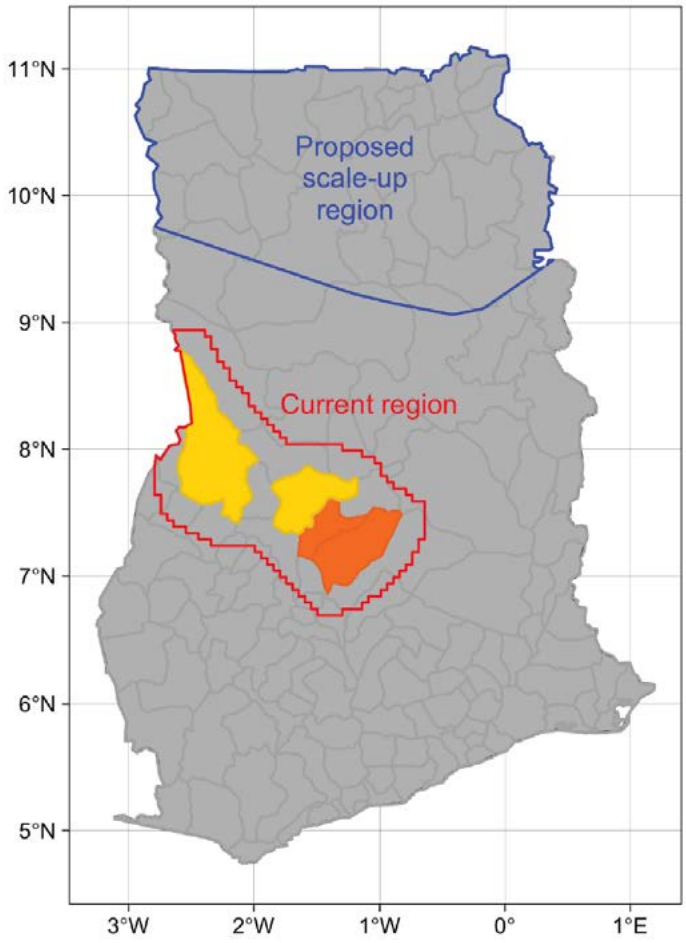

Figure 1.1:
The mapping region of the current project (red outline), showing the four districts where collections are being focused. The orange cluster was the focus of season 1, and the yellow was the focus of season 2. The mapping region extends slightly beyond the boundaries of the districts. The proposed scale-up region is outlined in blue.

2 Field Team Recruitment and Training

We will recruit field teams from Farmerline's network of agents, spearheaded by Farmerline's Agent Development Manager, for the geographical expansion of the project. Following that, the recruited agents will participate in a series of training for data collection particularly focusing on the protocols for field mapping. The new recruits will connect with the existing agents in the Ejura Sekyedumase, Sekyere West, Nkoranza, and Tain districts who assisted in the delineating of the field boundaries in the previous data collection seasons.

The trainings will provide an overview and refresher of Mergdata mobile (Android) application which has both online and offline capabilities. Agents will practice with Mergdata and will practice interviewing respondents and guiding them through questions so they better understand the overall application. In addition, we will provide the field team a step-by-step training on how to use the app for field mapping using both theoretical and practical approaches. A user manual with high-quality screenshots will accompany this training so that field teams can subsequently refer to the manual for information and guidance. During training, we will brief the field agents on the project objectives to help them answer questions from farmer participants.

The field teams will consist of two agents each, who will visit farming communities for the mapping and data collection. The pairing is mainly to provide some partnership and security during the mapping of farms. Each field team will be assigned a group of farmers to visit, and each team will arrange on farm visits with the farmer for mapping and data collection. The farmers will direct the field team on their farm boundaries to ensure effective mapping and data collection.

The results from the first crop type mapping exercise highlighted several improvements that we will make regarding the training and management of field teams, and how field data are collected:

3. **Calibration of GPS devices used with the Mergdata platform:**
   As part of new and refresher trainings, the field agents will undertake an exercise that demonstrates best practices for calibrating GPS devices used in delineating field boundaries. This includes a Mergdata protocol that allows capture only when GPS error is less than 5m and ensures that polygons are closed and completed as soon as the perimeter is completed. These steps will minimize the amount of post-collection data processing, which hampers the ability to automate model training and validation.

4. **Improving capture of photos associated with field plots:**
   We will further train the agents on the best process of image capture, including correct camera angle (in landscape view and looking across the field to include crops and background in all the frames) and recommended points of capture relative to the field (from each of the four sides of the field). Making these practices consistent will enhance the understanding of field context. In addition, we will improve the Mergdata application so that the field-captured reference images include EXIF data with GPS coordinates embedded in the photo. These data can help verify image location to help in understanding the field/scene context.

5. **Hiring new teams to expand geographic coverage:**
   Transportation was a significant impediment for field teams during the first season, which limited the geographic spread of sampling efforts and the diversity of crop types captured. To overcome these limitations, field agent selection will largely depend upon where their home bases are, with the home bases selected to maximize the geographic spread of field teams. This will help minimize the cost of transportation, spread, and increase sampling efforts.

6. **Overcoming lack of cell network and electricity:**
   Field collection protocols will ensure that agents find a strong internet connection and sync data on Friday of every week, to ensure that data collected outside connectivity areas are uploaded in a timely manner. We will provide agents with extra battery packs to overcome the inability to charge equipment in areas without electricity.

7. **Improved liaison with farmer groups, to increase the number of crop types and spread of regions sampled:**
   The project team will liaise with farmer group heads and opinion leaders to obtain details regarding the dates and locations of group meeting days so that the team can be present and have interactions with farmers of the individuals' groups, and not just the leaders.

3 Mergdata Database Improvements

After assessment of the application, we will implement a round of improvements to the Mergedata application. As an improvement, the Mergdata android app will present already mapped farms so that field agents can prompt the farmer to ensure the right boundaries are collected when there are overlaps. A feature will also be implemented to the mapped field that will include geometries collected by GPS in smartphones that are synced to the Mergdata web platform to ensure that the polygons are topologically valid.

To further improve the post-acquisition control, we will add an initial automated cleaning step in which raw field geometries are passed through a cleaning script that removes overlapping fields boundaries and self-intersections. We will run this process prior to the manual procedure we have already developed to check the integrity of the mapped area by visualizing the plot borders against satellite imagery.

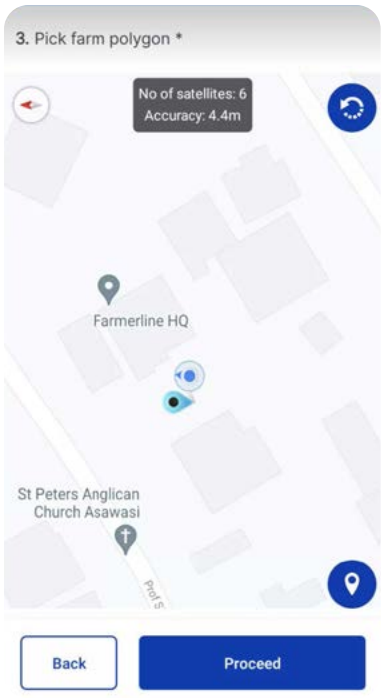

Figure 3.1:
Current Mapping Interface

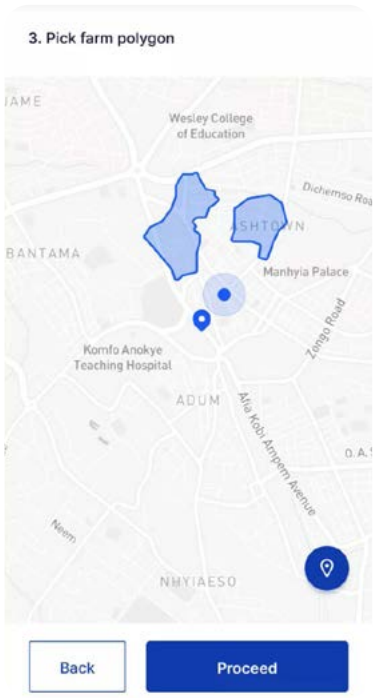

Figure. 3.2:
Interface with geometries

4 Using Drones to Improve Sample Characteristics

4.1 Overview

The first season revealed several operational constraints (see D1.7) that undermined the geographic coverage and balance of the sample, which negatively impacted the ability to develop and assess the reliability of crop type maps. These constraints included:

- The need to enroll farmers and secure their consent for data collection, which often included working through specific farmer organizations, thus limiting the geographic extent of sampling and biasing the collection towards specific crops;
- Transportation challenges, including poor roads and other movement barriers (e.g. flooded river crossings) further limited the geographic spread, as field teams were generally constrained to operate within a one-hour travel (by motorcycle) radius around their home bases;
- Lack of electricity in many areas further reduced the number of samples that could be collected during operations, thereby further constraining the sampling geography.

The initial sample was thus geographically clustered and skewed towards maize fields. The sample imbalance made it difficult to train a robust and unbiased model to distinguish between maize, rice, and other crop types. Furthermore, the constraints meant that field data were not collected following a probability-based design, which typically requires randomization (Stehman and Foody, 2019). The map reference sample we extracted from the data was therefore unlikely to be fully representative of the distribution of crops in our mapping region, which undermined our ability to objectively assess map accuracy.

To address this shortcoming, particularly the need to obtain a representative map reference sample and to facilitate the sustained and efficient collection of crop type groundtruth data over large areas, we propose to expand our sampling scheme to include the use of unoccupied aerial vehicles (UAS or drones) to collect <5 cm resolution, near-ground imagery. We will use this imagery to label (both manually and through separately developed classification models) the boundaries of individual fields and the primary crop types growing within them, as well as additional features that may affect the performance of mapping models (e.g. the frequency of tree cover in fields; variability in soil cover). The advantage of using UAS is that they can cover large areas for far less cost than a field agent, enabling the collection of data over many more fields and crop types. Their capacity to collect accurate crop type ground-truth was demonstrated by recent work in Rwanda (Chew et al, 2020; Hegarty-Craver et al, 2020). Here we propose to use UAS to address a critical but hard-to-satisfy requirement of ground-truth, which is to provide a sample that not only improves the size, spatial coverage, and balance of the training dataset but can be more readily designed to provide a probability sample

(Stehman and Foody, 2019). This latter point is critical for developing an independent map reference sample that can be used to rigorously evaluate map accuracy.

4.2 Design

We propose to use a design developed for a related project in Ghana that will focus on collecting oil palm reference data in the Ashanti and Eastern Regions of Ghana. The design has three objectives: 1) To expand the geographic coverage of crop type samples beyond that which is possible to cover on the ground; 2) To conform to the requirements of a probability sample (Stehman and Foody, 2019); 3) to overlap with field sampling efforts, so that overhead imagery can be validated against known ground observations. The design has four stages: Stratification, Clustering, Randomized route selection, and Sampling (Figure 4.1). The design presented here will use a UAS capable of beyond-line-of-sight operations. It is suitable for multi-copters with 25-30 minute flight times (which Farmerline already owns), provided extra charged batteries are available to enable 100-125 ha of mapping within a half-day period.

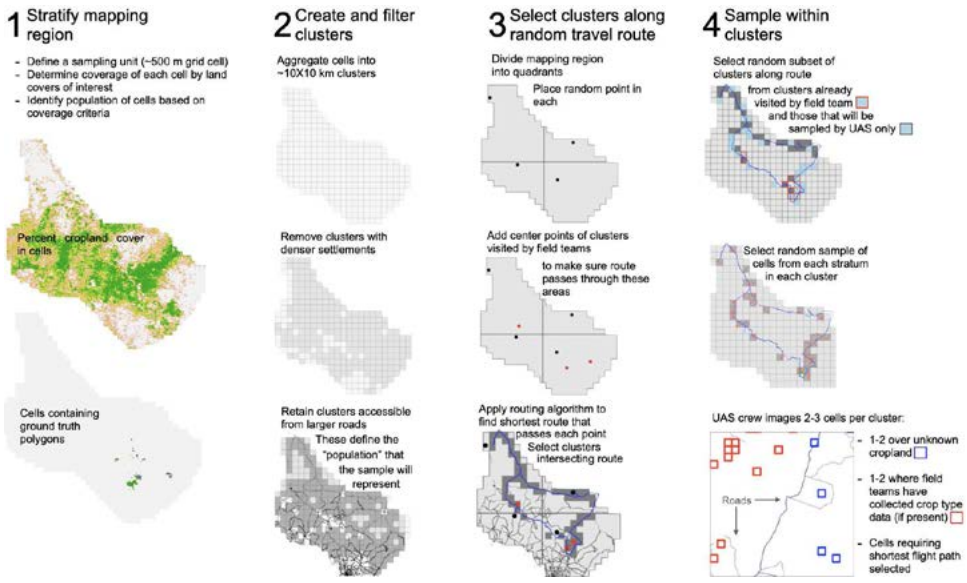

Figure 4.1:
An overview of the proposed UAS-based sample design

In the first stage, stratification, a sample unit is defined (a ~500 m grid, with the cell representing the unit to be completely imaged by a UAS) and the coverage of each cell by the targeted land cover type (strata) is calculated in a GIS. Here we have two strata of interest: 1) cropland, and 2) cells that contain (or will contain) ground-truth polygons collected by the field teams.

The first stratum is selected to ensure that cells containing cropland are sampled, and the second to ensure that UAS and in-field sampling are co-located with a subset of cells.

In the next stage, cells are grouped into clusters of ~10X10 km (0.1 degrees), which provide a unit into which sampling efforts can be concentrated, to reduce costs (Stehman and Foody, 2019). The clusters are twice filtered. Clusters that are more densely populated (determined using Facebook's high-resolution settlement layer) are removed, to avoid flying over settlements, as are clusters that are not accessible from larger roads (based on the OpenStreetMap roads layers for Ghana). This second filtering is necessary because the UAS sampling team will launch from roads, and the UAS should not fly more than 3-4 kilometers beyond the line of sight (a distance that might have to be shortened depending on operational rules and constraints). This filtering limits the domain of applicability for the resulting sample area covered by the retained clusters, but for the current study region that constitutes 72% of the total mapping geography, which is a much larger area than a purely field-based sample would represent.

In the third stage, the study region is divided into four quadrants, and a random point is placed in each quadrant, along with points representing the center of clusters in which the field teams collected (or will collect) data. An open source routing algorithm (implemented in R's stplanr package) is used to select the shortest driving route to pass each point, and the clusters intersecting that route are selected.

In the fourth and final stage, subsets of the clusters visited and not visited by the field teams are randomly selected. This selection reduces the overall sampling effort for the UAS team and ensures some overlap between field based and UAS sampling efforts. Within each cluster, we will draw a random sample of several cells from each stratum that is present. If the sample size is 5 cells per stratum, and the ground-truth stratum is present and can be further stratified by crop type, and there are 3 crop types, as many as 20 cells could be selected within a single cluster (5 cells per crop type, and 5 from the generic cropland stratum where cropland cover is at least 10%). In clusters without the ground-truth stratum, the sample would be a maximum of 5 cells. The UAS team will then collect complete imagery over 3-4 cells from this sample (at least one from each stratum, and 2-3 if only generic cropland is present), selecting the cells that minimize the total flight time to reduce collection time and accident risk. Upon completing the sample, the UAS team will proceed to the next cluster. We expect that 1-2 clusters can be collected per day, requiring a campaign of 3-5 weeks for a single UAS team (2-3 people).

The collected imagery will then be processed into orthomosaics and used to identify crop types within and between cells (the UAS will capture imagery while flying transiting between cells), using both visual interpretation and machine learning classifiers (see section 5).

This design will first be tested in the current four study districts, adapted as needed, and then extended to the northern expansion regions in subsequent seasons. A more complete description of the design is in the Appendix.

5 Crop Type Modelling

We will improve and expand crop type mapping by updating the approach used for training and testing the Random Forests-based mapping used in the first season, and by testing more advanced mapping models.

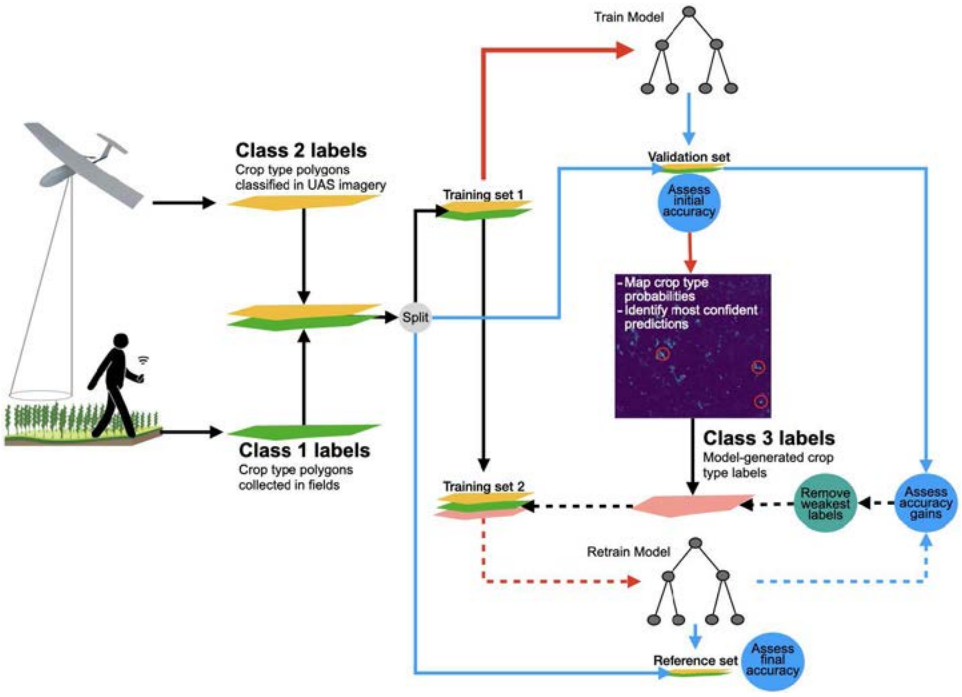

Figure 5.1:
An overview of the 3 classes of labels to be used in modelling crop types over larger regions and subsequent seasons, and how the labels will be generated. Beyond field-collected crop type polygons data (Class 1 labels), UAS-captured labels (Class 2) will also be developed (see section 5.1), as well as model-generated labels (Class 3).

5.1 Improved training/testing of Random Forests

The updated sampling approach that extends field sampling with drone imagery will be applied to improve the size and representativeness of the data used to train, refine, and validate the Random Forests-based modelling approach that has already been developed (D1.7). This process will involve the use of three classes of labels, which are grouped based on the level of confidence that can be assigned to them.

Class 1 labels: These are the labels collected by 1) Farmerline field agents on the ground, using protocols developed during season 1 and refined based on lessons learned during the first season. These labels are the most accurate in terms of classification and boundary delineation but are likely to be less representative, both in terms of geographical distribution and crop type. Our goal is that the collection of these labels become a routine part of farm visitation by field agents, rather than the focus of dedicated, standalone efforts.

Class 2 labels: Class 2 labels include those captured by the drone in grid cells not visited by the field teams. We will generate these labels based on human image interpretation and manual digitization of crop types visible in the imagery (e.g. Hegarty-Craver et al., 2020), and we will also test the efficacy of whether a model can be trained to predict crop types within the drone imagery (e.g. Chew et al., 2020), and thereby reduce the amount of manual labelling effort. Based on prior work (Chew et al., 2020; Hegarty-Craver et al., 2020), we expect that Class 2 labels will be suitable for maize, rice, and tree crops, but less accurate for vegetable crops and legumes, which will continue to be grouped under "Other crops ".

Class 3 labels: We will develop a third class of synthetic labels based on predictions from an initial version of Random Forests. After initial training of the model with a training sample consisting of Class 1 and Class 2 labels, we will assess model performance against the reserved validation sample, and then apply the trained model to map class probabilities across the expanded mapping region. For each crop, we will identify contiguous clusters of pixels where the predicted probabilities exceed 0.7 or the upper 90% percentile probability value, and convert these into new training labels, providing a third class of labels. We will then test whether adding batches (by decreasing order of confidence) of these Class 3 labels to the initial training pool improves model performance. We will retain the subset of Class 3 labels that lead to the largest gains. Class 3 labels will be suitable only for training purposes.

5.2 More advanced models

In addition to the Random Forests model, we will also test DL-based models based on convolutional and recurrent neural networks, focusing on architectures capable of utilizing the spectral-temporal domain and learning with little or no spatial context, as the identity of the land cover next to field boundary polygons is often unknown. Appropriate models include networks based on stacking modules of long short-term memory (LSTM), gated recurrent units (GRU), or temporal-attention mechanisms (e.g. Interdonato et al, 2019; Paoletti et al, 2020; Pelletier et al., 2019; Xu et al., 2020; Rußwurm & Körner, 2017, 2018, 2020). We will focus initial efforts on two of these: LSTMs and Transformer (based on temporal attention). The advantage of such models lies in their potentially greater ability to consider temporal relationships in assigning a label to each pixel, and the promise of greater transferability between regions and years than the Random Forests. We plan to develop these models using a combination of existing open-source crop types labels available for this region on Radiant MLHub, including a set available for Northern Ghana (Rustowicz et al, 2019).

We will compare these results against our existing Random Forests model. Our primary assessment will compare the performance (in terms of Overall, User's and Producer's Accuracy, as well as the F1 score) of the deep learning models versus Random Forests against the Type 2 labels. We will also perform several transferability experiments, in which we compare the ability of models:

- Trained with data from just Radiant MLHub to predict crop types within our current mapping region;
- Trained with data from our current regions to predict crop types in the Radiant MLHub region (using a reserved portion of that dataset for reference);
- Trained with pooled data from our current region and from Radiant MLHub to predict crop types in our region (on Type 2 data) and for the Radiant MLHub region (on the reserved validation sample).

6 Expansion Regions

The steps proposed in sections 3–5 will be implemented and tested in the coming season within the mapping zone around the four districts that are the focus of this current project. We will assess the improvements resulting from these changes and will develop additional customer offerings based on updated maps. If the viability of the business case is demonstrated, or additional funding is secured, our planned next step is to 1) Continue these ground-truth collection and crop type mapping approaches for our current four districts, and 2) To expand them to Northern Ghana, focusing and expanding from three clusters of already enrolled farmers (Figure 6.1), whose numbers total nearly 31,000 individuals.

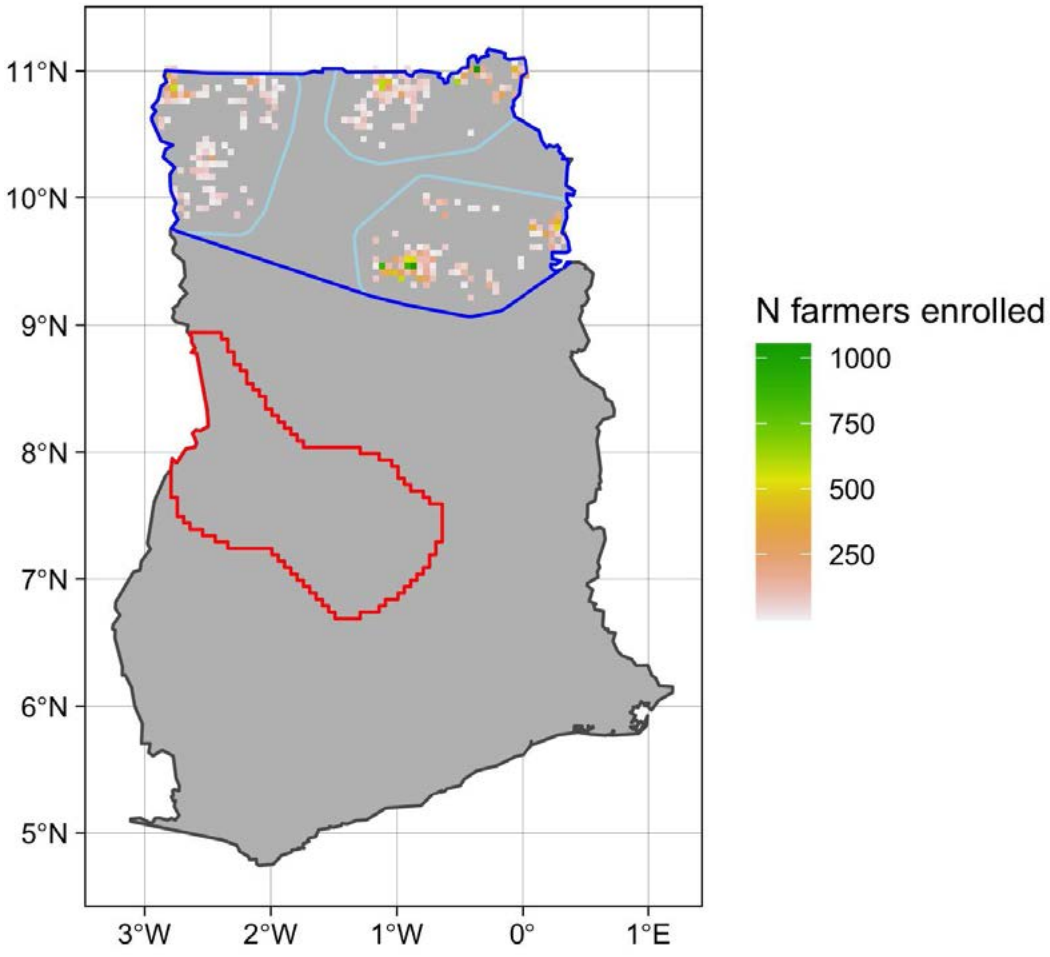

Figure 6.1:
The region (blue outline) intended for expansion of ground-truth data collection/crop type mapping, shown in relation to the region that is the focus of current activities (red outline). This region encloses three groups of already enrolled farmers. The location and number of farmers enrolled per ~5X5 km is indicated by the color bar.

Expanding to these regions will increase the geographic scope of efforts, allowing us to offer enhanced services over a greater portion of our current operational area in Ghana. This region also represents a different agro-ecological zone (Guinea Savanna), which has a single growing season (the current project area falls primarily within the Transitional zones, which has a major and minor growing season). This provides an opportunity to expand our mapping focus to include groundnuts and soyabean, which are cultivated in this region, and to collect observations of maize and rice grown under a different climate regime, which will allow us to develop more generalizable models. We will use the UAS-based sampling approach to cover the areas between farmer enrollments to enable the collection of training and reference data in these areas. Subsequent geographic expansions will track the spread of Farmerline enrollments.

7 Use of the crop maps

7.1 Integration in Farmerline Business Model

Farmerline is quickly advancing the agricultural sector in West Africa, increasing yields and income for small-holder farmers (avg 37% and 29% respectively). We leverage mobile and voice technology, data, and light in-person trainings to provide a digital whole value-chain solution. This approach is suited for smallholder farmers living at or under the poverty line. Farmerline also seeks to efficiently integrate and use these crop maps to ensure a successful and sustainable path in creating and maintaining products, services, and partnership with stakeholders, including local government, agri-businesses, development partners, academia, and global food and beverage companies, among others.

7.1.1 Overlay into maps for Farmer support

Farmerline is working with Clark University to integrate the cropland boundary and crop type maps through API, to strengthen the support that is provided to the farmers such as input needs assessment and extension services. Clark University will add all maps to its MapBox-based web maps and enable them to be embedded in the Mergdata platform. This will allow farmers to rapidly view crop type and field boundary maps at specific locations and will help field agents to plan when to conduct specific training for farmers based on their crops and stages of growth. With the crop maps, field agents will be able to understand input demand (for example seeds & fertilizer) based on accurate farm size and crop type, compared to the current approach of using national average estimates, and make better informed recommendations. Extension teams can then further verify maps when collecting additional points on farm visits. These points can be used as additional training data that helps to continuously improve the maps.

7.1.2 Crop Marketing Estimation and Sourcing

Obtaining data on commodities estimation on smallholders outside of our farmer network is currently a challenge. Using the Farmerline service, we can estimate the commodity market opportunity and engage partners before the start of the trading season, and effectively increasing the efficiency of our commodity sourcing by an estimated 50%. Data from these crop maps lowers costs of commodity search while increasing commodity supply by engaging agents to source products across regions. Farmerline provides and supports the supply chain by delivering affordable, high-quality inputs on time to farmers, and increases their market access. This could increase our revenue and reduce the cost of customer acquisition significantly.

7.1.3 Repository for Agribusiness

The Agribusinesses are entrepreneurs who support smallholder farmers with their farm input as well as training needs and are usually repaid with farm commodities after harvest. These local, often family-run agro-vet dealers sell inputs solely (seeds, fertilizer, agro-chemicals) to local, last-mile communities. They don't have working capital, nor access to working capital. They speak the local language of the farmers they serve and are therefore trusted and known.

These agribusinesses run analog for inventory and customer tracking. Farmerline digitizes their processes, allowing them to communicate and easily reach their customers and professionalize their business. Farmerline provides free marketing to the Farmerline farmers who become customers by purchasing Farmerline inputs. With the integration of the cropland boundary data, the agribusinesses will be able to have more comprehensive data on the characteristics of farms in their area which will influence their input acquisition for the farmers as well aggregation preparations for the farm produce they receive as repayment for the inputs.

7.2 Code and Data Availability

As part of the scale-up plan, we will ensure that cropland boundary data developed under this project will be captured and shared as vector polygons enclosing the target crop types. All datasets are collected with farmers' informed consent, documented according to the SpatioTemporal Asset Catalog (STAC) label extension specification, and will be publicly released on Radiant Earth's MLHub data registry. These will be fully open to the public, under a CC BY-SA 4.0 license. To avoid any risk of personal identification, the anonymized name field will be removed from these publicly released data. To aid users, we will provide a tutorial accessible through MLHub showing how to load the data, convert it to raster labels and combine it with satellite imagery (e.g. Sentinel-2). Image processing and model code will be made publicly available on GitHub (through Clark University's agroimpacts organization).

8 Sustainability plan

Maintaining and growing the dataset beyond this initial funding is critical and Farmerline's Business to Farmer (B2F) and Business to Business (B2B) models will ensure this is achieved. The project will exit into our B2F strategy which is partnering with governments to scale our Mergdata platform through existing and new program interventions to improve the quality, speed, and efficiency of farmer extension service delivery, creating a fair and transparent market based on data-driven insights for the agricultural sector. The B2F model is scaling efficient delivery of farmer/farm extension services directly to smallholders (target of 140,000 and 2,000,000 farmers in 2021 and 2025 respectively) across Ghana and Ivory Coast through data-driven insights with training, inputs, and commodity trade financing. The B2B model, which currently supports partners across 16 African countries, would offer a sustainable pipeline to introduce value-added services/products derived from our field boundary and crop type maps, which could currently serve over half a million farmers and 8 million by 2025.

We see a clear opportunity to maintain and scale this project intervention into countries across Africa. Our business expansion through key partnerships and ongoing working relationships with stakeholders including governments will enable the existing dataset to be sustainably maintained and grow. For example in Benin, Farmerline is supporting the International Fertilizer Development Center (IFDC), Fédération des Unions de Producteurs du Bénin (FUPRO-BENIN) - the largest farmer association, among other stakeholders to enhance learning and training, and improve access to market information for over 50,000 farmers. Secondly, Farmerline's B2B model, which has licensed its Mergdata technology platform that is currently used across 16 African countries to support about half a million farmers, could make it possible for agribusinesses and development partners to contribute to sustainably growing the dataset in exchange for beneficial use of the technology.

Farmerline's models are projected to be financially sustainable without being the sole company to be offering these services. However, our partnership with governments and multinational organizations will significantly increase our reach. This expanded reach is also projected to increase revenue which would continue to support, maintain, and scale the technology.

Appendix

The following provides fuller details on the proposed UAS-based sampling design, with an example provided for the four current study districts covered in this project.

A.1 Stage 1: Stratification

This entails mapping the coverage by several different land cover types in our study region (the four districts) so that we can ensure that samples are placed within each cover type. These are based on two layers:

1. Our PlanetScope-derived cropland data layer, which informs us about the distribution and density of crop fields in the region.
2. The already collected crop types from season 1, which identifies areas where we already have ground samples. Here these serve as a stand-in for sampling that might occur or would be planned to occur during the upcoming field season so that a portion of the UAS sampling effort occurs over sites where the field teams have identified and mapped crop types.

To define these strata (Figure A.1) the fractional coverage of each layer was calculated within a 0.005 degree (~500 m) sampling grid, which represents the unit that should be imaged by a UAS, as well as the sampling grid used in our cropland mapping platform (Estes et al., 2021).

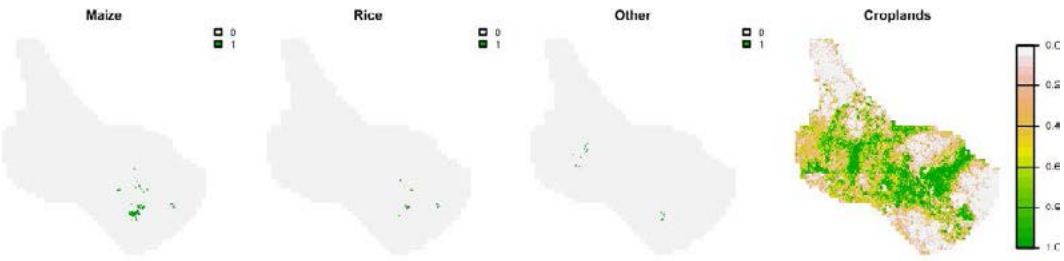

Figure A.1:
Crop type and cropland layers at 0.005-degree resolution, providing the basis for stratifying UAS-based sampling in the study region. Crop types are here shown as binary rasters (rather than proportional coverage), with any cell in the study having a value of 1, regardless of coverage, for ease of visualization.

A.2 Stage 2: Clustering

Cells are grouped into larger clusters of 0.1 degrees (~10 km) and further filtered by excluding those with more than 10% of their area under settlements, as defined by a map derived from the Facebook population grid for Ghana. Those remaining are further reduced by selecting those having 10% or more of their areas within 5 km of a road (using the OpenStreetMaps [OSM] roads layer), as the UAS crew will have to launch from a road, and the aircraft has a limited flight range (3–4 kilometers ferry distance from road). To ensure roads are more likely to be traversable, we confined roads to those designated as the trunk, primary, secondary, tertiary, or motorway in the OSM dataset.

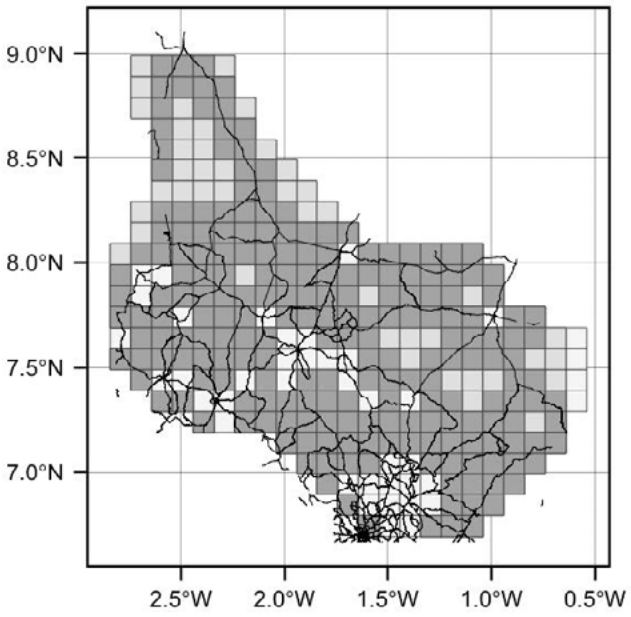

Figure A.2:
Clusters (darker grey) that have at least 50% of their areas within 5 km of the main road, and which have less than 10% of their areas occupied by settlements.

The retained clusters, therefore, contain the "population" of cells that the sample will represent, as only these cells have a non-zero chance of being included in the sample (Stehman and Foody, 2019). Map accuracy measures calculated using the sample will apply only to this area, and not to the excluded clusters, which in this example comprises 28% of the total mapping geography. Insight into how well the sample-derived accuracy measures could be extrapolated to the excluded area may be obtained by comparing the characteristics of cropland in the included and excluded clusters, in terms of density and size classes, which can be calculated from our Ghana-wide field boundary maps (Estes et al, 2021).

A.3 Stage 3:
Selecting clusters along the randomized driving route

The next step is to select a sample of the eligible clusters determined in the previous section. The UAS team will visit these clusters and collect imagery within each of them. To select a sample of clusters, we have several requirements: 1) each of the remaining clusters should have a chance of being selected; 2) some of the clusters should include those containing fields collected by (or to be collected by) the field teams; 3) the amount of travel between clusters should be minimized; 4) geographic spread should be maxi-mized.

To meet these requirements, we use a road routing algorithm to select the roads that are driven, and therefore which clusters are visited. We provide the algorithm with points it must travel through, allowing us to increase geographic coverage and to ensure that certain areas are visited, which in this case are field team sampling areas. To do this, we divide the region into four quadrants, place a random point in each quadrant, and add the center points for several clusters that contain field boundaries collected during the last season. The routing algorithm is then run between these points. The selected route is shown in Figure A.3.

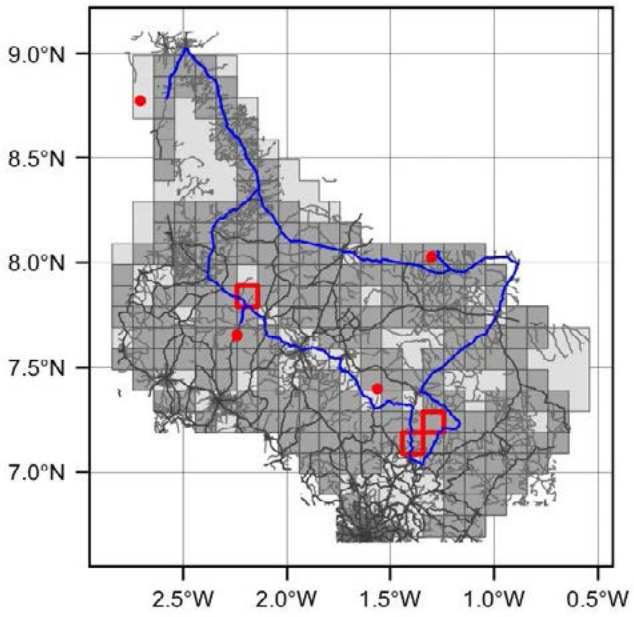

Figure A.3:
The selected driving route to follow when collecting UAS samples. The route was selected by dividing the region into four quadrants, then randomly placing a point in each (red points) and selecting the shortest round-trip route that could be travelled to reach all four points. The route was also required to visit three clusters in which ground samples had already been placed (red polygons). All roads, including minor roads and tracks, are shown for reference.

Once the route is chosen, the clusters falling on the route are selected (Figure A.4).

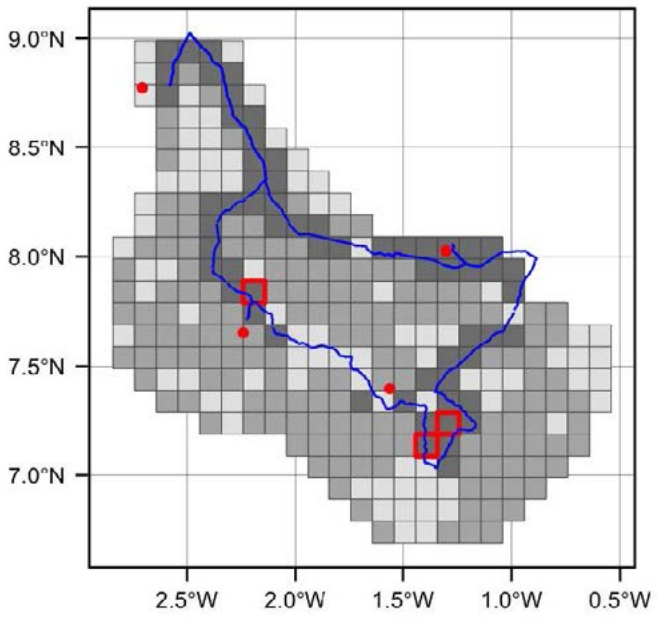

Figure A.4:
Intersecting clusters (dark grey) along the selected route. The eligible clusters are shown in light grey.

A.4 Stage 4: Sampling within clusters

We further reduce the road-intersecting clusters by randomly selecting a subset of the clusters containing only generic cropland, and a subset of those containing field-collected polygons, to minimize the overall sampling effort while ensuring overlap between field and aerial sampling efforts (Figure A.5).

Figure A.5: Randomly selected clusters along the driving route (blue fill). Clusters containing existing ground samples from the past season are outlined in orange and red, with the latter being those used in the routing process.

Having finalized the cluster selection, we next select the sample units within them. These are the grid cells (0.005 degrees) that the UAS will fly over and image. In this example, we have four classes to consider, ground-collected maize, rice, and other crop types, and generic cropland (i.e. cells where the cropland layer shows at least 10% coverage by fields) and can identify cells that have coverages in any one of or up to all four classes. Our sample cells will thus the following typology:

- **Type 1:** Cells containing previously collected maize field boundaries;
- **Type 2:** Cells containing previously collected rice or other crop field boundaries;
- **Type 3:** Cells that contain unknown/generic cropland

Cells of each type can have the other types in them, but we ensure that we select up to 5 of each type, depending on availability. We combine rice and other crops into a single class here given their low numbers in the first season's dataset (see D1.7). Each cell can thus have up to 15 selected cells (this could be expanded to 20 if we separated other crops into a fourth type). Figure A.6 shows the distribution of selected cells within the cluster.

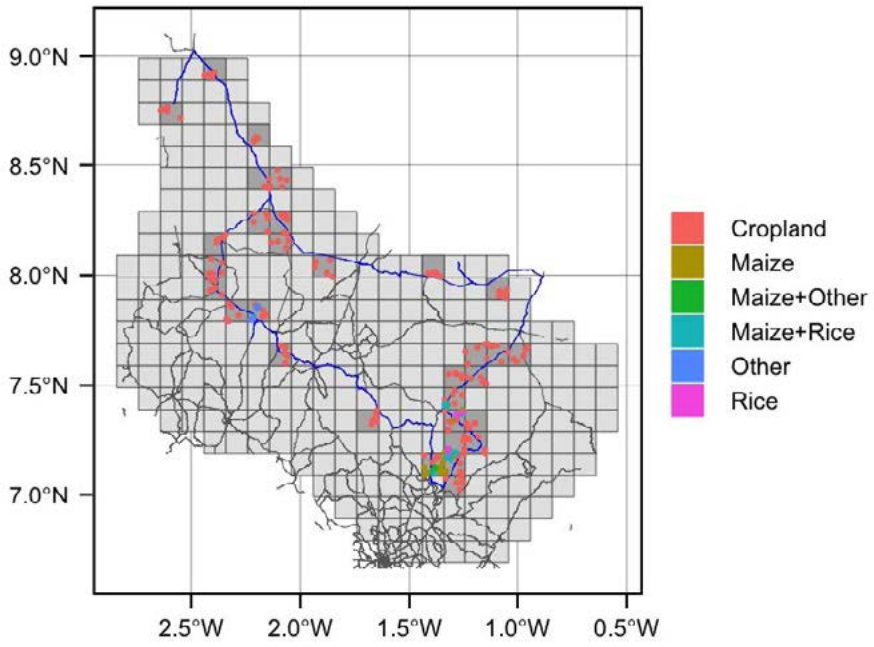

Figure A.6:
Selected sample cells within each cluster. Cells are color-coded according to the combination of crop types or cropland represented within them, with those labelled as 'cropland' being those with no known crop type, while those with combinations of maize, rice, or other crop types include fields whose crop types were mapped on the ground.

The larger number of cells selected for each cluster gives the UAS team the ability to select just 2-4 images to sample per cluster, flying first those that contain an already sampled location (Type 1 and 2 cells, i.e. those containing maize/rice/other), and where those are not available, cells containing cropland of unknown type. The team should choose the cells closest to the launch point to minimize beyond-line-of-sight flight time, altering the selection needed to deal with flight logistics (e.g. uneven terrain in a selected cell; need to return to swap out batteries).

Figure A.7 provides four examples of how samples are laid out within clusters. Cluster 909 only has Type 3 cells, so the UAS team should choose 2–3 of these–cell 26 is over a road and could be captured quickly, along with its closest neighbor 24. A third cell, either 15 or 43, could be also be captured, possibly during the same flight, or by relaunching from where the cell intersects the road. Similarly, in Cluster 1089, cells 116 and 78 are the closest pair, and cell 233 is a nearby road–intersecting cell that could also be readily imaged. Clusters 1414 and 1432 contain all three cell types. In Cluster 1414, capturing 40 (Type 1), 61 (Type 2), and 82 (Type 2) could be imaged on a single flight, while road–intersecting 22 or 202 could be imaged to provide a Type 3 sample. In Cluster 1432, the most efficient triplet to capture would be 35 (Type 3), 49 (Type 1), and 104 (Type 2).

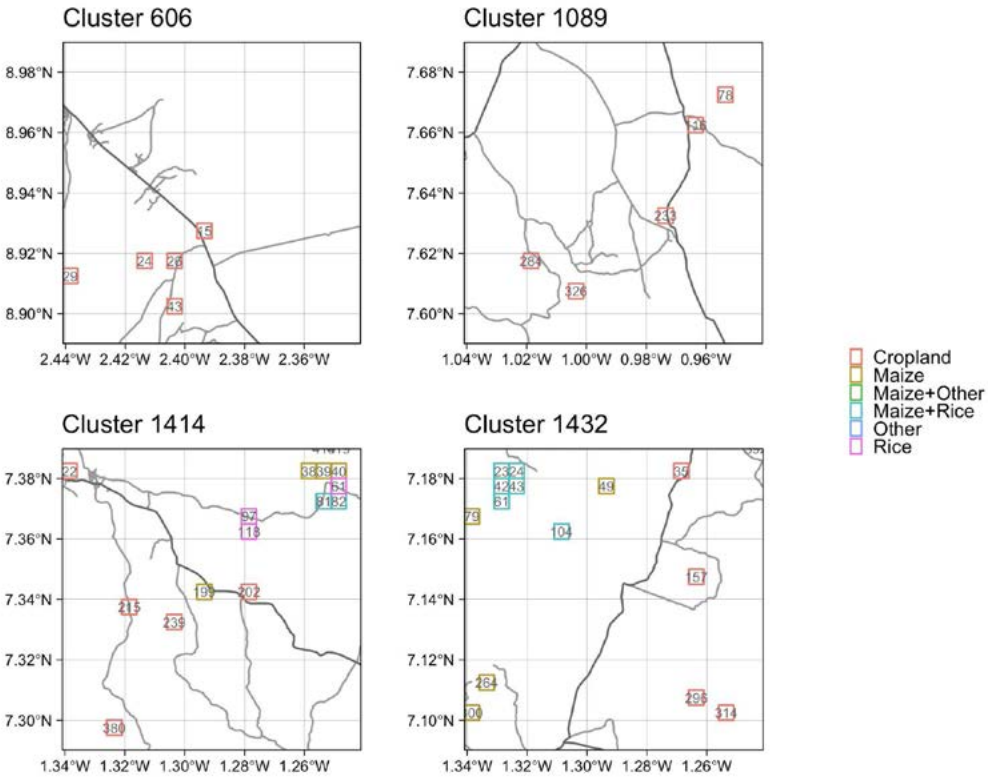

Figure A.7:
Close up of randomly selected cells, color–coded by cover type, within four selected clusters. The cover type legend provides the text description of the types contained within each cell. Cells IDs are provided inside each cell. Main roads are shown in black, while smaller roads are shown in a lighter grey.

A.5 Summary

This design presents an approach that could be paired with the ongoing work of collecting crop types on the ground by the Farmerline teams. Proposed for September–October, 2021 (if project extension is granted), when crops are well grown during the second growing and visible in the UAS imagery, this should be combined with the planned work by the field teams so that a subset of clusters are visited by both UAS and field teams. Ideally, the field teams' work should slightly precede the UAS team, so that newly collected field data can be integrated and the sample design updated, to ensure that known crops are being imaged. The UAS team will also image a larger number of clusters not visited by the field team, thereby increasing geographic coverage. As noted previously, this design will not provide a reference sample that is applicable to the entire mapping region. However, it provides a statistically representative sample for 72% of the region, which is substantially larger and less biased than would be possible with a purely ground-based approach.

We will develop a cost estimate of this proposed approach following the development of an estimate for the aforementioned oil palm collection work, which is expected to be completed by the end of May. We expect the costs to fit within the current scope of the ECAAS budget, without any impact on existing deliverables. Operational details (e.g. flight team and who they are contracted by) can be determined following the estimate of a budget and approval of this proposed sampling design.

References

Chew, R., Rineer, J., Beach, R., O'Neil, M., Ujeneza, N., Lapidus, D., Miano, T., Hegarty-Craver, M., Polly, J. & Temple, D.S. (2020) Deep Neural Networks and Transfer Learning for Food Crop Identification in UAV Images. Drones, 4, 7.

Estes, L.D., Ye, S., Song, L., Luo, B., Eastman, J.R., Meng, Z., Zhang, Q., McRitchie, D., Debats, S.R., Muhando, J., Amukoa, A.H., Kaloo, B.W., Makuru, J., Mbatia, B.K., Muasa, I.M., Mucha, J., Mugami, A.M., Mugami, J.M., Muinde, F.W., Mwawaza, F.M., Ochieng, J., Oduol, C.J., Oduor, P., Wanjiku, T., Wanyoike, J.G., Avery, R. & Caylor, K. (2021) High resolution, annual maps of the characteristics of smallholder-dominated croplands at national scales.

Hegarty-Craver, M., Polly, J., O'Neil, M., Ujeneza, N., Rineer, J., Beach, R.H., Lapidus, D. & Temple, D.S. (2020) Remote Crop Mapping at Scale: Using Satellite Imagery and UAV-Acquired Data as Ground Truth. Remote Sensing, 12, 1984

Interdonato, R., Ienco, D., Gaetano, R., & Ose, K. (2019). DuPLO: A DUal view Point deep Learning architecture for time series classificatiOn. ISPRS Journal of Photogrammetry and Remote Sensing, 149, 91-104.

Paoletti, M. E., Haut, J. M., Plaza, J., & Plaza, A. (2020). Scalable recurrent neural network for hyperspectral image classification. The Journal of Supercomputing, 76(11), 8866-8882.

Pelletier, C., Webb, G. I., & Petitjean, F. (2019). Temporal convolutional neural network for the classification of satellite image time series. Remote Sensing, 11(5), 523.

Stehman, S.V. & Foody, G.M. (2019) Key issues in rigorous accuracy assessment of land cover products. Remote Sensing of Environment, 231, 111199.

Xu, J., Zhu, Y., Zhong, R., Lin, Z., Xu, J., Jiang, H., ... & Lin, T. (2020). DeepCropMapping: A multi-temporal deep learning approach with improved spatial generalizability for dynamic corn and soybean mapping. Remote Sensing of Environment, 247, 111946.

Unlocking the Potential of Satellite-based Data and Analytics for Smallholder Farmers

The Enabling Satellite-based Crop Analytics at Scale (ECAAS) Initiative is a multi-phase project that aims to catalyze the development, availability, and uptake of agricultural remote-sensing data and subsequent applications in smallholder farming systems. The initiative is funded by The Bill & Melinda Gates Foundation and implemented by Tetra Tech.

info.ecaas@tetratech.com
cropanalytics.net