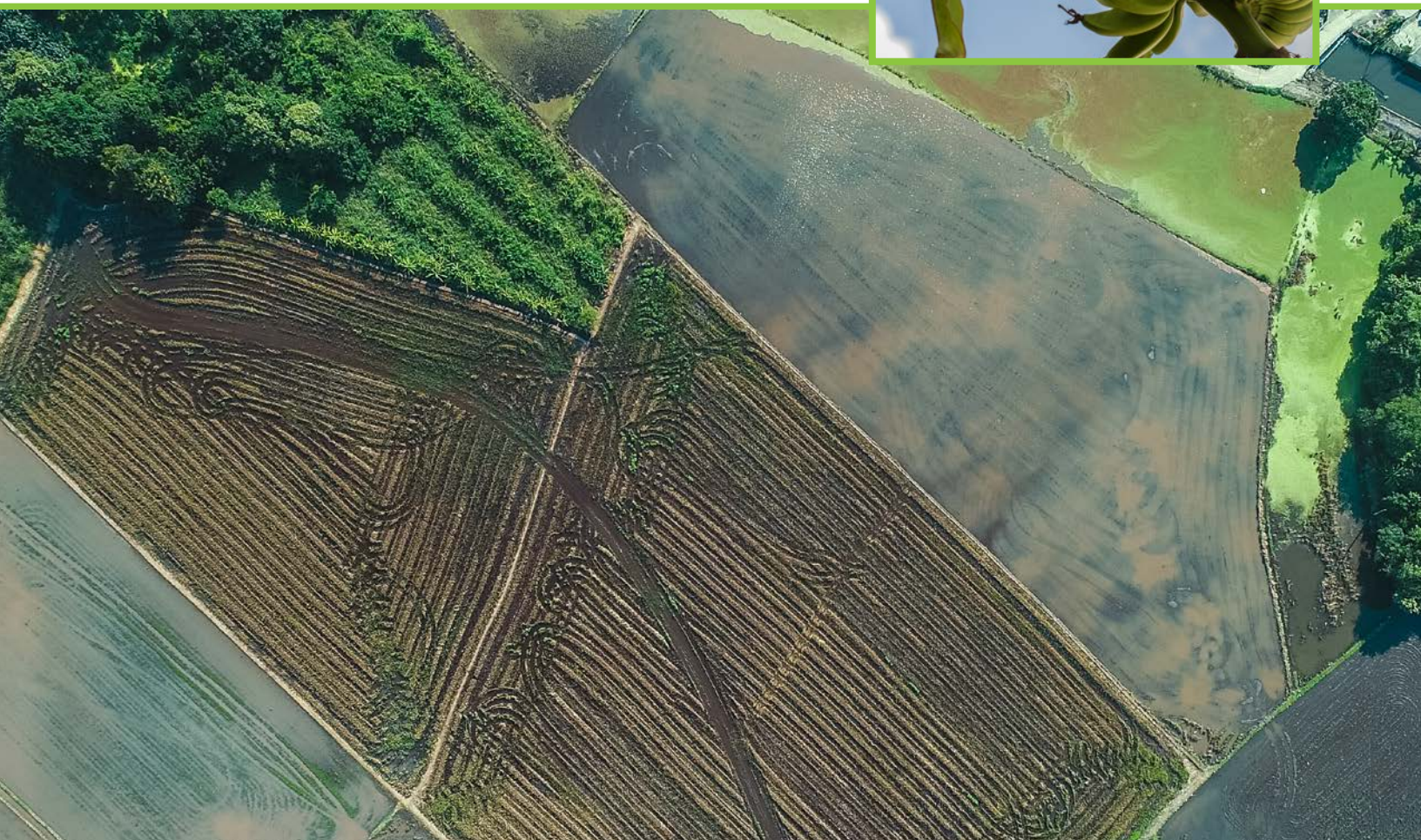
TETRA TECH

Enabling Crop Analytics At Scale

ENABLING CROP ANALYTICS AT SCALE (ECAAS)

# Streamlining Ground Truth Data Collection:

## Rwanda Case Study



# Contents

## Appendices



## Figures



## Tables

# Key Takeaways

RTI investigated three research questions to advance crop analytics in smallholder settings:

**How well do existing crop classification models perform in subsequent growing seasons?** We found that the best-performing existing crop classification model that we had developed based on 2019 Season A did not perform well when used with 2022 Sentinel-2 imagery and compared against either field observation or drone imagery ground truth data (**Table KT1**). We attribute the overall poor performance to two factors. First, the model from 2019 Season A was trained on a relatively small number of label points (< 1,000 across all land cover categories) which may have led to overfitting to characteristics specific to the training data areas. Second, the model only used the spectral bands captured by the Sentinel-1 and Sentinel-2 sensors. Other factors such as slope, soil type, fertilization, irrigation, and rainfall were not included, though such environmental/biophysical data should ideally be included in future analyses.

**How do field observations compare with drone imagery when generating ground truth data?** Overall, RTI determined that using field observations provided better results than hand-labeled drone imagery in terms of model performance, time to collect/generate, and cost. We also support using a field data collection app that captures polygons for generating data to train satellite models. However, the number of field observation points collected will be tied directly to the budget, whereas drone imagery can be revisited to obtain additional data. Training and a field test are strongly recommended to ensure the best quality data is obtained and meets project requirements.

**Can machine learning generate labels from drone imagery with adequate accuracy to train a satellite-based crop type model?** The use of machine learning to generate data for calibrating a satellite-based model from drone imagery presents an interesting possibility. We rapidly generated a large number of labels by classifying the entirety of the drone images using a pre-trained computer vision model that produced better satellite model performance for certain crops than either ground-based or drone-based labels in less time and at lower cost. This has the potential to minimize the amount of human labor, as the computer vision model produced good results for key land covers without having to be retrained. One issue to be addressed in future work is that it produces many mixed class labels. Additional work to classify specific crop mixes in heavily intercropped areas is important for assessing agriculture in smallholder settings.

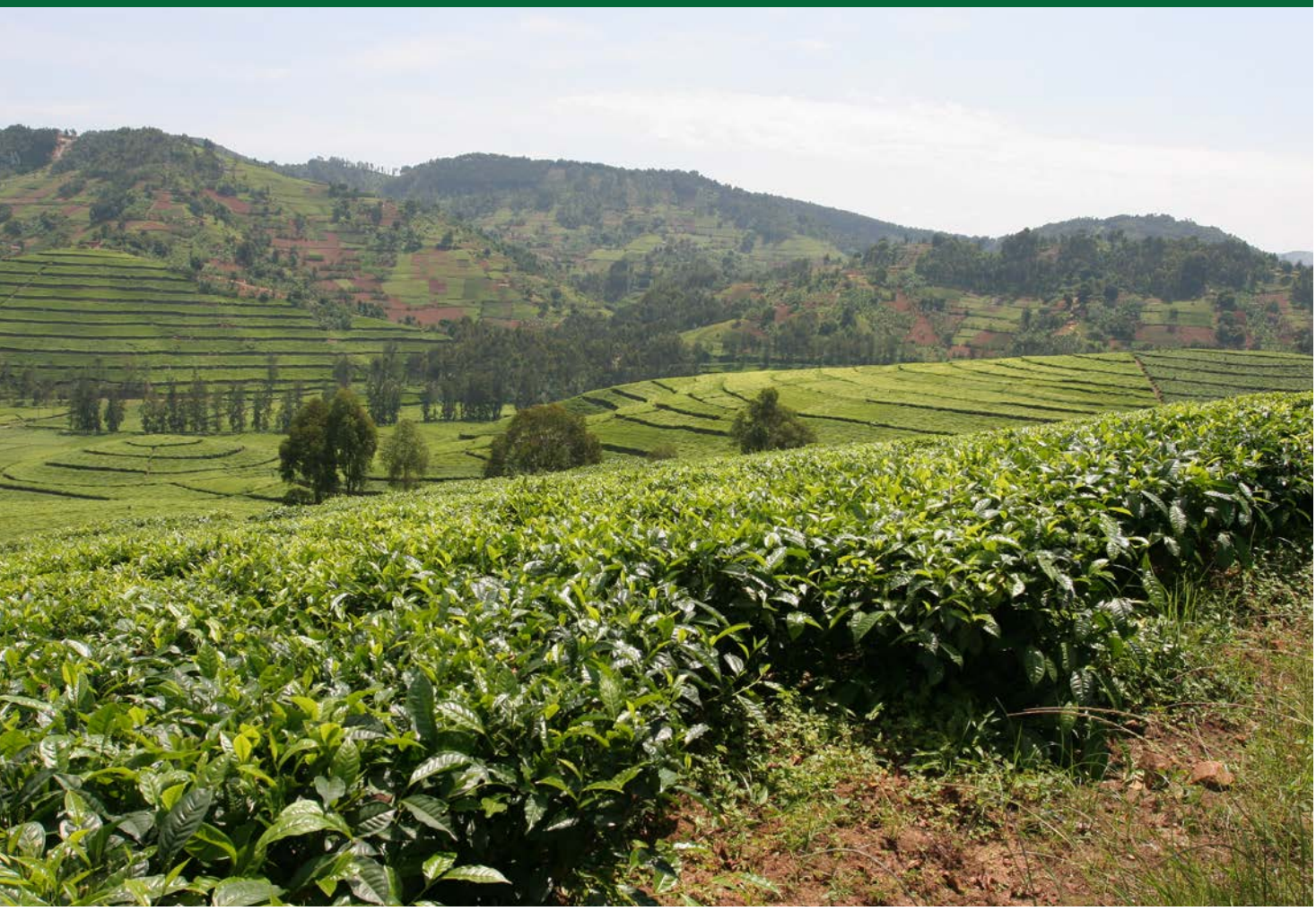
**Table KT1:**

Comparison of Model Accuracy Results

| Model | Classes | Sentinel Imagery Year | Trained Using | Validated Using | Overall Accuracy |
|---|---|---|---|---|---|
| 2019 Landcover Type | 8 | 2019 | Drone Imagery Labels (2019) | Drone Imagery Labels (2019) | 85.6% |
| 2019 Landcover Type | 8 | 2022 | Drone Imagery Labels (2019) | Field Observation Labels (2022) | 18.9% |
| 2019 Landcover Type | 8 | 2022 | Drone Imagery Labels (2019) | Drone Imagery Labels (2022) | 24.9% |
| 2022 Crop/NonCrop | 2 | 2022 | Field Observations Labels (2022) | Field Observations Labels (2022) | 84.5% |
| 2022 Crop/NonCrop | 2 | 2022 | Drone Imagery Labels (2022) | Drone Imagery Labels (2022) | 75.3% |
| 2022 Crop/NonCrop | 2 | 2022 | Computer Vision Labels (2022) | Field Observations Labels (2022) | 68.4% |
| 2022 Crop/NonCrop | 2 | 2022 | Computer Vision Labels (2022) | Drone Imagery Labels (2022) | 73.1% |
| 2022 Crop Type | 3 | 2022 | Field Observations Labels (2022) | Field Observations Labels (2022) | 68.1% |
| 2022 Crop Type | 3 | 2022 | Drone Imagery Labels (2022) | Drone Imagery Labels (2022) | 58.4% |
| 2022 Crop Type | 3 | 2022 | Computer Vision Labels (2022) | Field Observations Labels (2022) | 70.3% |
| 2022 Crop Type | 3 | 2022 | Computer Vision Labels (2022) | Drone Imagery Labels (2022) | 73.4% |

## Executive Summary

Crop analytics leveraging artificial intelligence (AI)- and machine learning (ML)-trained models based on remotely sensed data hold great promise for providing actionable data to public and private stakeholders worldwide. However, the current lack of reliable ground truth or labeled data that can be paired with sensor measurements for calibration purposes is a major barrier to widespread adoption of satellite-based crop analytics to calculate these measures (Burke et al., 2021).



Traditionally, field data have been collected for a specific analysis at a specific location, resulting in non-public, non-reusable training data. RTI had previously conducted agricultural research in Rwanda in 2018–2020 as part of an internally funded study that utilized unmanned aerial vehicle (UAV or drone) imagery to train satellite-based models to classify crops at the national level for Rwanda (Hegarty-Craver et al., 2020). We focused on growing Season A, which runs September–February and is one of the two major growing seasons in Rwanda.[1] RTI posed three research questions; the answers to which could potentially streamline ground truth data collection, reduce the burden to crop modelers, and add data to the growing public repositories of training data:

1. How well do existing crop classification models perform in subsequent growing seasons?

2. How do field observations compare with drone imagery when generating ground truth data?

3. Can machine learning generate labels from drone imagery with adequate accuracy to train a satellite-based crop type model?

**To answer Research Question #1**, we applied the best performing machine learning model developed based on data from 2019 Season A to Sentinel-1 and Sentinel-2 data from 2022 Season A to generate a predictive surface of crop types. The model predicted the most likely land cover type from eight categories: banana, beans, cassava, other crops, trees, natural vegetation, maize, and bare ground. We used both our 2022 drone-based and field observation-based label points to evaluate how well the 2022 predictive surface agreed with them. The original model produced an overall accuracy of 85.6%. However, when the 2022 drone-based and field observation-based labels were overlaid and compared against the 2022 predicted values, there was very little agreement, indicating limitations in transferring the existing model across years. The overall accuracy for the drone-based labels was 24.9%, while the overall accuracy for the field observation-base labels was 18.9%.

**To answer Research Question #2**, we compared ground truth data collection via field observations with ground truth data collection from drone imagery in terms of speed, cost, and model accuracy. Overall, our assessment determined that using field observations provided better results than hand-labeled drone imagery in terms of model performance, time to collect/generate, and cost.

[1] The primary growing seasons in Rwanda are Season A (September–February) and Season B (March–June). In low-lying areas where there is sufficient water during the dry season, there is also a Season C from July to August that tends to be more focused on vegetable production.

For field observation data collection, our subcontractor Vanguard Economics was able to complete one site per day, collecting approximately 1,000 land cover examples per site. This was possible because of the relatively small area (80 hectares), large field team (10 persons), and small number of attributes collected. The drone imagery collection was also completed at a rate of approximately one site per day. In terms of time to acquire data there was not a significant difference, although field observation teams can work through weather that might delay drone operations.

Drone imagery acquisition was approximately $5,000 less than the cost of deploying a field observation team. Drone imagery is even less expensive on a per unit area basis since the field collection costs were for a specific number of land cover observations per site (~1,000), whereas the drone imagery can yield up to 7,000–8,000 labeled Sentinel-2 grid cells per site if the entire drone image were classified.

The field observation data required processing to map it into categories by Sentinel-2 grid cells. Since there could be multiple land cover types per observation point, and multiple land cover observations per Sentinel-2 cell, algorithms were developed to create a dominant category per point, and a dominant category per grid cell to account for cases with multiple points falling within a given grid cell boundary.

One concern we had regarding field data collection was the accuracy of the native GPS receivers in the tablets while in the field. In our initial discussions with Vanguard Economics, we stated that we were hoping to limit GPS errors to no more than 2–3 m. They indicated that they thought this was possible and that the application they used (SurveyCTO) provided excellent accuracy. In practice the field teams were able to get accuracies that ranged between 1.8 m and 5.0 m with a mean error value of 4.4 m. This increased the likelihood that points used to define the land cover in a given Sentinel-2 cell could have in reality been in the adjacent cell. We were able to compensate for this by using a buffer when developing the land cover prediction models.

The biggest challenge with the drone imagery was training analysts to recognize the six crop types of interest, in the imagery. We had a wealth of crop examples from our previous work in Rwanda, as well as the photos taken by the ground observation team, which we used to train the analysts independently from the drone labeling process. Maize and bananas were easily recognizable, but cassava, beans, sweet potatoes, and Irish potatoes required more training to reliably identify. The time and effort required to train the analysts and do the actual drone imagery labeling was significantly higher (152 hours) than the labor to process the field data (36 hours).

Neither ground truth generation method produced as strong of an overall predictive model as we had anticipated. The overall crop-type model accuracy was 68.1% when trained using field-based ground-truth data, slightly outperforming the drone-based crop-type model which had an overall accuracy of 58.4%. We attribute this to the ability of the ground observation team to more reliably identify the crops while standing next to them, as opposed to interpreting a drone image. This is especially true for beans,

sweet potatoes, and Irish potatoes which are sometimes difficult to distinguish from each other, or from fallow and natural vegetation.

**To answer Research Question #3** we utilized machine learning methods (specifically a Deep Neural Network (DNN)) to automatically generate labels from drone imagery, and then assessed model performance when using those labels to train a predictive land cover classification satellite model. RTI used a computer vision model generated during the Rwanda Grand Challenge project (Chew et al., 2020) to classify the entire drone image into six land cover types (bananas, maize, legumes, forest, structure, and other). The computer vision model was generated using the visual spectrum drone imagery captured over the same six areas, by the same drone operator, during the same growing season (Season A) as we used for our 2019 Season A data work (Chew et al.,2020). The 2022 drone imagery was broken into 5m x 5m image chits, each one was classified, and then aggregated into a single classification for each 10m x 10m Sentinel-2 cell that overlapped the drone imagery. This produced 7,000–8,000 label points per study site. This method produced an overall accuracy of 66.3%, when compared against field observation labels and 65.6% when compared against drone imagery labels, suggesting that transferring a DNN model across seasons works better than transferring a Random Forest (RF) model. These label points were then used to train/evaluate a satellite model. The three-crop class model performed better (73.4%) than using either the human-labeled drone-based or ground observation-based labels, suggesting that using a machine learning model to generate a large volume of labels benefits the modeling process. We envision a workflow whereby a pre-trained computer vision model could be used to automatically generate labels from drone imagery for satellite models. This would eliminate the need for field observations and draw upon existing examples of crop/land cover types held in public repositories.

Overall, we identified limitations in transferring our earlier ML crop classification model across years. RTI also determined that using field observations provided better results than hand-labeled drone imagery in terms of model performance, time to collect/generate, and cost in our application. We recommend using a field data collection app that captures polygons for generating data to train satellite models. This will however produce a finite number of observations dependent on the budget. Training and a field test are strongly recommended to ensure the best quality data are obtained and meets project requirements. The use of machine learning to generate data for calibrating a satellite-based model from drone imagery presents an interesting possibility. We were able to generate a very large number of labels within hours by classifying the entirety of the drone images using a pre-trained computer vision model that produced better satellite model performance for certain crops than either ground-based or drone-based labels generated by humans in less time and at lower cost. This has the potential to minimize the amount of human labor, as the computer vision model produced good results for key land covers without having to be retrained. One issue to be addressed in future work is that this approach produces a large number of mixed class labels. Additional work to classify specific crop mixes in heavily intercropped areas is important for assessing agriculture in smallholder settings.

contain summarized statistics by different analytical units such as blocks, administrative areas, or regions (using area and fraction weights based on the extent of each LULC–zone present).

Sampling data must also account for changing weather patterns. Agroecological zone–(AEZ[4]) and CPSZ–based zoning account for climate but are relatively static and do not capture season–specific performance differences due to weather–specific anomalies frequently occur within and across zones. The Long Term Normal (LTN), is another typical weather measurement that presents seasonal differences in rainfall at regional levels (at ± 7km$^2$ grids). The weather anomalies occur mostly following patterns of larger weather systems such as El Niño Southern Oscillation (ENSO). Differences in landforms and terrain only marginally influence such patterns (**Figure 1**).

While crop performance can be affected by the severity of large–scale weather anomalies, the local aspects of terrain, soil, and land management are far more significant (**Figure 2**). Therefore, performance indicators must include impacts of terrain, soil, and land management. One such indicator is NDVI, a widely accepted land cover greenness metric representing the performance of cropping systems and other land cover classes. For this work, we focus on anomalies in the response of systems rather than anomalies in inputs; the latter is, however, highly relevant regarding the production of timely performance predictions (**see Annex 3**).

NDVI anomalies can be overlaid on static CPSZ–maps (representing current land use) to create a Dynamic Sample Frame (a season–specific dynamic area frame, DAF). That DAF represents an ideal solution to scale up site–specific yield data from various sources such as Crop Cutting Estimate surveys (CCE surveys) to crop production estimates by area or region. Typically, a DAF does not guide sample schemes but instead creates a layer that presents a season–specific stratification on crop performance.

# Introduction

To provide timely data in support of national and development goals and to mitigate food insecurity, countries need accurate, crop-specific measures of areas under cultivation and of crop productivity. Crop analytics derived from artificial intelligence (AI)- and machine learning (ML)-trained models based on remotely sensed data hold great promise for providing timely and actionable data in support of these goals.

Although satellite revisit intervals are shrinking and spatial resolution is increasing for key environmental variables (Fu et al., 2020), sensor readings are often misaligned with varied planting and subsequent growth stages over large areas, and except for active cloud-penetrating wavelengths, cloud cover remains a significant barrier to creating large collections of paired ground truth (labeled) and satellite pixel information, especially in key growing periods (which coincide with rainy seasons). Other barriers include the current publicly accessible satellite pixel size which are incompatible with small plot sizes, intercropping practices in some areas, and potentially abrupt agroclimatic differences observed over short distances. With unmanned aerial vehicle (UAV or drone)-based images, we can overcome many of these constraints but are currently restricted to smaller coverage areas.

Researchers who perform crop analytics at scale are aware that satellite model prediction accuracies of crop type, crop extents, and crop yield vary directly with the training data they are supplied. As noted by Burke et al. (2021), ground truth data are often collected in the form of household surveys that frequently lack geospatial accuracy, and an overall lack of reliable ground truth data stands in contrast to the ever-increasing quantity and quality of satellite data.

RTI developed three research questions to determine if there was an opportunity to streamline ground truth data collection:

1. How well do existing crop prediction models perform in a subsequent growing season?

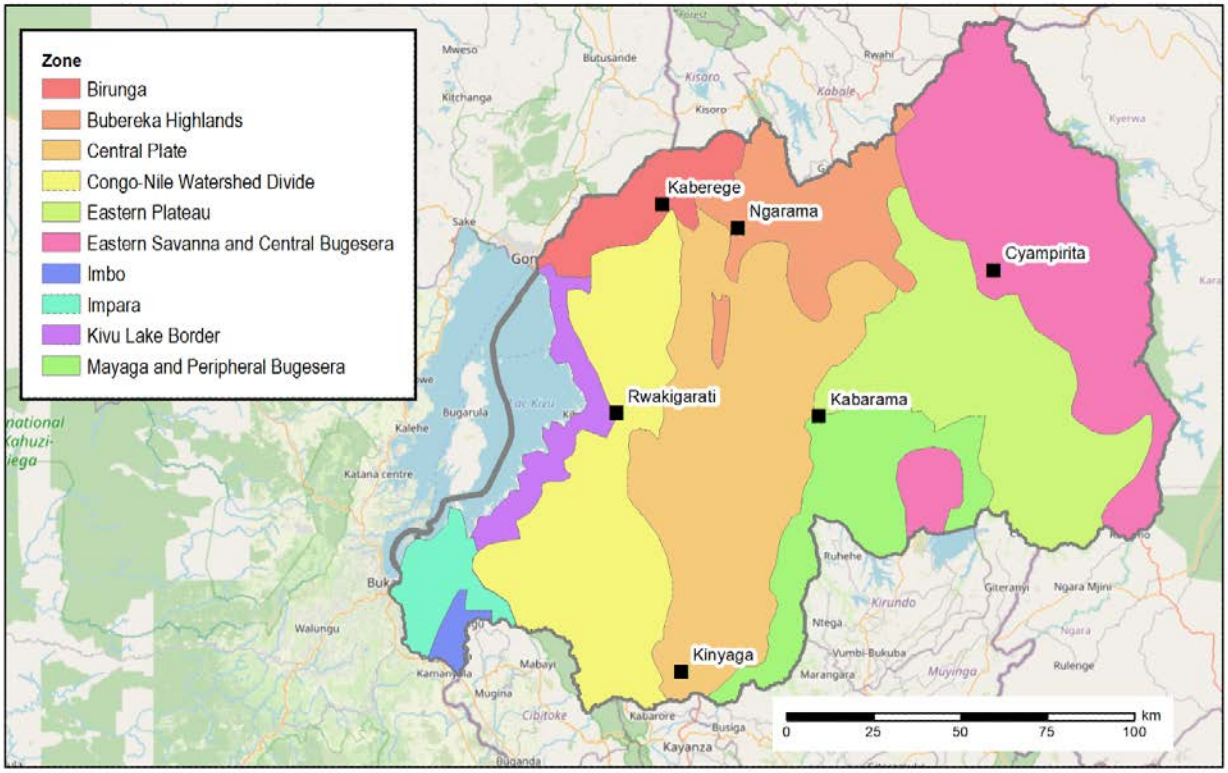
2. How do field observations compare with drone imagery when generating ground truth data?

3. Can machine learning generate labels from drone imagery with adequate accuracy to train a satellite-based crop type model?

Research Question #1 focused on finding out how well existing models performed in subsequent growing seasons, since an ideal solution to the dearth of training data is to re-use models developed on a set of training data collected in a previous season or year. In theory, a given crop type should have a similar spectral signature from one year to the next. RTI developed crop type models from its work in Rwanda in 2018–2020. This presented an opportunity to test model performance using ground truth data from field observations and drone imagery collected in 2022 at the same six study sites and in the same crop season (Season A, which is from September through February in Rwanda). These six sites were chosen to represent six different agroecological zones found within Rwanda as shown in **Figure 1.** Each study site was about 80 ha in area.

This also provided an opportunity to answer Research Question #2 which aimed to compare the time, cost to collect and process ground truth data, and accuracy achieved using two distinct methods: field observations and drone imagery interpretation. These data were collected at the same six study sites between December 2021 and February 2022. We hypothesized that creating training data using drones would be faster, less expensive, and provide comparable accuracy to traditional ground observation data collection.

**Figure 1:**
Study Sites Within Rwanda




Research Question #3 can be answered by using the drone imagery collected during January/February 2022 and classifying it into six specific land covers using a pre-trained computer vision model created during RTI's 2018–2020 Grand Challenge project to create thousands of label points. These label points can then be used to train a satellite model, which can be evaluated using both the field observations and drone imagery labels as ground truth.

## Research Question 1:

# How Well do Existing Crop Classification Models Perform in Subsequent Growing Seasons?

One way to reduce the need for additional collection of ground truth data is to re-use models developed for different areas or time frames if sufficient accuracy can be achieved. Ideally, crop models developed based on a given study area and growing season could be applied to that area in different seasons and potentially to other, similar areas, while maintaining high accuracy. However, there may be challenges in maintaining high accuracy when applying models to growing seasons and/or regions other than those on which they were trained (Orynbaikyzy et al., 2022).

The output from a satellite-based crop type model is a predictive surface that produces the most likely crop/land cover class for each satellite cell. RTI created several crop-type models during our work from 2018–2020 using the visible bands from 10 m resolution Sentinel-2 imagery as well as several Sentinel-1 SAR bands resampled from 20 m down to 10 m. Those models used the following crop/land cover categories:

1. Banana
2. Beans
3. Cassava
4. Maize
5. Other crops and vegetation
6. Natural vegetation
7. Trees
8. Bare ground/structures/water

The previous models were trained using 904 drone-based observations and evaluated using 222 drone-based observations. The best performing model produced an overall accuracy of 85.6% (**Table 1**).

This model was used to create a predictive surface using the Sentinel-2 imagery composite from December 2021 to February 2022. This surface was evaluated using the two sets of "ground truth" data—those derived from field observations and those derived from drone imagery—by overlaying the label data with the predictive surface, and then comparing the predicted crop/land cover category with the labeled one. The crop land cover categories were not quite the same since we identified sweet potatoes and Irish potatoes separately. We mapped our 10 land cover categories into the eight categories used for labeling 2019 Season A as shown in **Annex Table C2**. When the 2022 drone-based and field observation-based labels were overlaid and compared against the 2022 predicted values, there was very little category agreement.

**Table 1:**

Confusion Matrix and Model Performance of Best 2019 Season A Model

| | Banana Labels | Beans Labels | Cassava Labels | Other Vegetation Labels | Trees Labels | Natural Vegetation Labels | Maize Labels | Bare Ground Labels | Prediction Accuracy |
|---|---|---|---|---|---|---|---|---|---|
| Banana (Pred) | 12 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 75% |
| Beans (Pred) | 0 | 15 | 0 | 0 | 0 | 1 | 2 | 1 | 78.9% |
| Cassava (Pred) | 0 | 1 | 11 | 0 | 1 | 3 | 1 | 1 | 61.1% |
| Other Vegetation (Pred) | 0 | 1 | 0 | 2 | 0 | 0 | 0 | 0 | 66.7% |
| Trees (Pred) | 0 | 0 | 0 | 0 | 57 | 1 | 1 | 0 | 96.6% |
| Natural (Pred) | 0 | 2 | 2 | 0 | 1 | 20 | 0 | 3 | 71.4% |
| Maize (Pred) | 0 | 0 | 0 | 0 | 1 | 0 | 54 | 2 | 94.7% |
| Bare Ground (Pred) | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 19 | 86.4% |
| Label Accuracy | 92.3% | 75% | 84.6% | 100% | 93.4% | 80% | 87.1% | 73.1% | 85.6% |

## Evaluation Using Field Data Ground Truth

The 2,726 field observation label points were compared with the predictive surface to produce the confusion matrix shown in **Table 2**.

### Table 2:

Confusion Matrix and Model Performance Using 2022 Sentinel Imagery and Field Observation Labels as Ground Truth

| | Banana Labels | Beans Labels | Cassava Labels | Other Vegetation Labels | Trees Labels | Natural Vegetation Labels | Maize Labels | Bare Ground Labels | Prediction Accuracy |
|---|---|---|---|---|---|---|---|---|---|
| Banana (Pred) | 36 | 19 | 14 | 86 | 32 | 12 | 118 | 15 | 10.85 |
| Beans (Pred) | 11 | 11 | 8 | 26 | 16 | 5 | 43 | 10 | 8.5% |
| Cassava (Pred) | 14 | 29 | 8 | 61 | 26 | 15 | 103 | 5 | 3.1% |
| Other Vegetation (Pred) | 0 | 1 | 1 | 1 | 3 | 0 | 2 | 0 | 12.5% |
| Trees (Pred) | 4 | 7 | 11 | 39 | 142 | 11 | 46 | 8 | 53.0% |
| Natural (Pred) | 13 | 13 | 6 | 25 | 49 | 12 | 40 | 5 | 7.4% |
| Maize (Pred) | 7 | 10 | 3 | 17 | 36 | 6 | 109 | 4 | 56.8% |
| Bare Ground (Pred) | 26 | 39 | 40 | 113 | 111 | 25 | 122 | 34 | 6.7% |
| Label Accuracy | 32.4% | 8.5% | 8.8% | 0.3% | 34.2% | 14.0% | 18.7% | 42.0% | 18.9% |

Trees and maize performed the best for the ground observation ground truth points, ranging from 53.0% (trees) to 56.8% (maize). The performance for other crops, including banana, was poor, with predictions near random.

## Evaluation Using Drone Imagery Ground Truth

The 3,038 drone-generated label points were compared with the predictive surface to produce the confusion matrix shown in **Table 3**.

### Table 3:

Confusion Matrix and Model Performance Using 2022 Sentinel Imagery and Drone Labels as Ground Truth

| | Banana Labels | Beans Labels | Cassava Labels | Other Vegetation Labels | Trees Labels | Natural Vegetation Labels | Maize Labels | Bare Ground Labels | Prediction Accuracy |
|---|---|---|---|---|---|---|---|---|---|
| Banana (Pred) | 61 | 30 | 13 | 40 | 45 | 28 | 91 | 28 | 18.2% |
| Beans (Pred) | 16 | 15 | 3 | 21 | 4 | 24 | 24 | 7 | 13.2% |
| Cassava (Pred) | 19 | 15 | 12 | 36 | 13 | 29 | 49 | 17 | 6.3% |
| Other Vegetation (Pred) | 0 | 0 | 3 | 0 | 0 | 6 | 0 | 2 | 0% |
| Trees (Pred) | 5 | 12 | 23 | 16 | 119 | 30 | 48 | 16 | 44.2% |
| Natural (Pred) | 16 | 16 | 10 | 29 | 33 | 68 | 18 | 17 | 32.9% |
| Maize (Pred) | 29 | 13 | 6 | 16 | 37 | 30 | 92 | 4 | 40.5% |
| Bare Ground (Pred) | 26 | 29 | 32 | 76 | 61 | 82 | 81 | 89 | 18.7% |
| Label Accuracy | 18.2% | 13.2% | 6.3% | 0% | 44.2% | 32.9% | 40.5% | 18.7% | 24.9% |

For drone imagery ground truth, the overall match rate was better than ground observation ground truth. Again, maize and trees performed the best, ranging from 40.5% (maize) to 44.2% (trees). The performance for other crops and land covers was poor, ranging from 13.2% (beans) to 32.9% (natural vegetation).

## Conclusions

Unfortunately the best performing model we developed in 2019 did not perform as expected with the 2022 datasets. The overall accuracy for the drone-based labels was 24.9%, while the overall accuracy for the field observation-based labels was 18.9%. While some of this inability to correctly predict the land cover for a given Sentinel-2 cell can be attributed to changing staff and labeling inconsistencies (i.e., we did not interpret the same crops the same way between 2019 Season A and 2022 Season A), it does not explain how poorly bananas and maize performed even though they may look different spectrally to the RF model. What is more likely is that the previous model used a relatively small number of training and evaluation labels, and the model may have been overfit to locations used for calibration and less accurate when transferred to locations outside those used for model training. Indeed, inspection of land cover predictions against drone imagery in areas where training labels were not selected revealed many questionable categorizations. In the previous application, drone-based labels were generally applied in areas where the land cover type was clear for use in training the model, but with less focus on use of cells where the land cover was mixed or unclear for training. These would have trained the model to identify other non-mixed land cover examples with strong accuracy but may have performed less well when applied to intercropped or other mixed land covers. The 2022 training labels were selected in a mostly random fashion, which may have presented less certain land cover examples, and therefore were less likely to match the predictive surface. Lastly, it is possible that since the previous model used only Sentinel spectral bands, and not other measures such as slope, soil type, fertilization, irrigation status, and rainfall, that in the bands used to create the predictive model, the same crops appear different spectrally from season to season and year to year. In that case, it would be important to build a larger calibration data set comprised of data from multiple years to improve performance over time (Wen et al., 2022).

- Our research indicates that a simple spectral band model created using training data from one year may not perform well when used to predict land covers in a subsequent year.

## Research Question 2:

# How do Field Observations Compare with Drone Imagery when Generating Ground Truth Data?

For this evaluation we compared the time to collect, the cost to collect, the time to process, and the model performance of two types of ground truth data: 1) traditional field observations, and 2) crop type labels derived from high resolution drone imagery. A summary of the comparisons is presented in **Table 4**.

**Table 4:**
Comparison of Field Observation and Drone Imagery Label Data

| | Field Observation | Drone Imagery |
|---|---|---|
| Time to Collect | 6 days | 5 days |
| Cost to Collect | $23,037 | $17,984 |
| Hours to Process Data After Receiving from Contractors | 36 | 152 |
| Number of Observations | 5,850 | 2,726 (but possible to create up to 45,523) |
| Model Accuracy Overall | 68.1% | 58.4% |

## Time to Collect

The time to collect (as measured in calendar days) was comparable between the field data collection and drone imagery collection. Both methods could process a 80 ha site in a single day. The field data collection was done using an 11-person team. Presumably if a larger area was required, more people could be trained and utilized. However, the more field staff employed, the greater the challenge to maintain data consistency. From the drone perspective, if a larger area were required the drone would need to fly for a longer period. Charis has reported that 80 ha is approximately the amount of area they can fly with a single drone in a single day, therefore a larger area might require more than one day. Weather (rain, wind) is a consideration for drones as well. Adverse conditions could delay drone imagery acquisition, whereas field staff are able to work regardless of the weather. A more detailed description of the field observation, and drone data collection processes is presented in **Annexes A and B** respectively.

## Cost to Collect

The cost of field data acquisition was $23,037, which included labor, internal training, travel, and taxes and fees. Field staff collected almost 6,000 observations, which consisted of land cover type(s), a GPS location, and a photo. GPS accuracy was a concern since it averaged approximately 4.4 m of positional error, whereas drone imagery contained less than 10 cm horizontal error. The field data cost compares with $17,984 for the drone imagery acquisition, which included travel, labor, processing, and the provision of imagery, digital terrain models, and video. It is worth noting that RTI paid about $2,000 per flight in 2019, but that was for three flights per study site for a total of 18 flights. Charis

indicates that the cost per flight is reduced by 30–40% when the same area is revisited. But even so the cost of acquiring drone imagery in Rwanda has increased. However, despite the increased cost, the data acquisition with drones was approximately $5,000 less than the field data collection.

## Time to Process

Once the field data were received, it was necessary to perform quality control (QC) and process the data so they could be used in a GIS and run scripts against them. The scripts to aggregate points into standard categories and then into consensus Sentinel-2 cell labels had to be written, run, and the result checked. Those steps amounted to approximately 4.5 person days of labor.

The drone data did not take long to download and incorporate into a labeling application within ArcGIS Portal, a web-based hosting platform. The biggest effort was the time required to train the analysts on crop identification, and then the actual labeling, editing, and QC. This amounted to approximately 19 person days of labor. Despite the creation of a streamlined labeling application, interpreting and classifying land cover is a labor-intensive process. A more detailed description of the data processing procedures is presented in **Annex C**.

## Cost to Process

The cost to process depends on salaries or daily rates for personnel involved, as well as how the hours are divided among the personnel. For our analysis, we found that the cost of processing the drone data was approximately three times the cost of processing the field observation data. This makes sense because the field observation data were largely already interpreted, whereas the drone imagery required interpretation at each randomly selected Sentinel-2 cell location. More details about the cost of data processing of both the field observations and drone data are presented in **Annex C**.

## Model Performance

Models trained and evaluated using field observation data consistently performed better than models trained with drone data, which logically follows as one would expect it to be easier to identify land cover types visually while in the field observing them, than from imagery. However, we expected the performance of the model trained with drone-derived labels to be closer to the one trained with field data that it was. Maize and banana accuracies for the drone label-trained model were mediocre, and not as good as the predictions created by the field observations. We attribute this to inaccuracies contained in the drone label dataset, implying that human labeling of the drone imagery introduced more labeling error than in the case of field data. Crops with smaller leaves are harder to correctly and consistently identify on drone imagery than during field observations. Some crops, such as beans, also look similar to other crops as well as natural vegetation. Examples of land cover types as seen by field staff on the ground, and on drone imagery are presented in **Annex Figures G1-G8**. For crops like banana

and maize, where leaf size, color, and texture are characteristic, there may have been a tendency to classify a Sentinel-2 grid cell as that single crop, whereas field observations would have been more likely to include a mixture of crops since they are more readily apparent on the ground when they are under the larger primary crop. Information about the satellite imagery selected, and the modeling environment used is found in **Annexes D and E** respectively.

## Conclusions

Although the time to obtain the drone and field observation data for this scope of work was similar, the costs were approximately 28% higher for the field data collection as compared to the drone imagery, making drone imagery a less expensive option. However, processing time greatly favors field data, since most of the processing has already been done by the time the data are delivered. Labeling satellite imagery from drone data requires training if the staff performing the task are not already familiar with how crops look at various growth stages. And even trained staff are required to assess each randomly selected location, which can take a significant amount of time. While we feel that using randomly selected points for labeling will improve overall applicability and spatial transferability of models developed over time, it presents more labeling challenges where there is a mix of multiple land cover types.

In terms of satellite model performance, models trained using field observations had higher accuracies than those trained on drone images (68.1% overall versus 58.4% for the drone imagery model). This was a surprising result, as we expected the drone labels would be similar if not more accurate than the field observations, especially given the built in GPS error of the field observations. However, our best fitting models compensated for GPS inaccuracies by using a 100 m buffer when modeling that predicted land cover class for any given Sentinel-2 cell which likely benefitted the ground observation trained model more than the drone imagery trained model which did not possess these inaccuracies.

- Field observation data were preferred to drone-based observations for this scope of work. Although their initial acquisition cost was more expensive, they required far less processing as compared to drone imagery. In addition, models trained using field observation-based labels performed better than those using drone-based labels.

## Research Question 3:

# Can Machine Learning Generate Labels from Drone Imagery with Adequate Accuracy to Train a Satellite-Based Crop Type Model?

RTI created a computer vision model during its Rwanda Grand Challenge project (Chew et al., 2020) to classify an entire drone image into six land cover types (bananas, maize, legumes, forest, structure, and other). The model was created using visual spectrum drone imagery captured over the same six study sites in Rwanda during growing season A in 2019. This model can classify individual images within a drone area in a few hours. Our third research question focused on two topics: 1) the automated generation of a training dataset, and 2) the performance of a model trained using these training data.

## Automated Label Generation

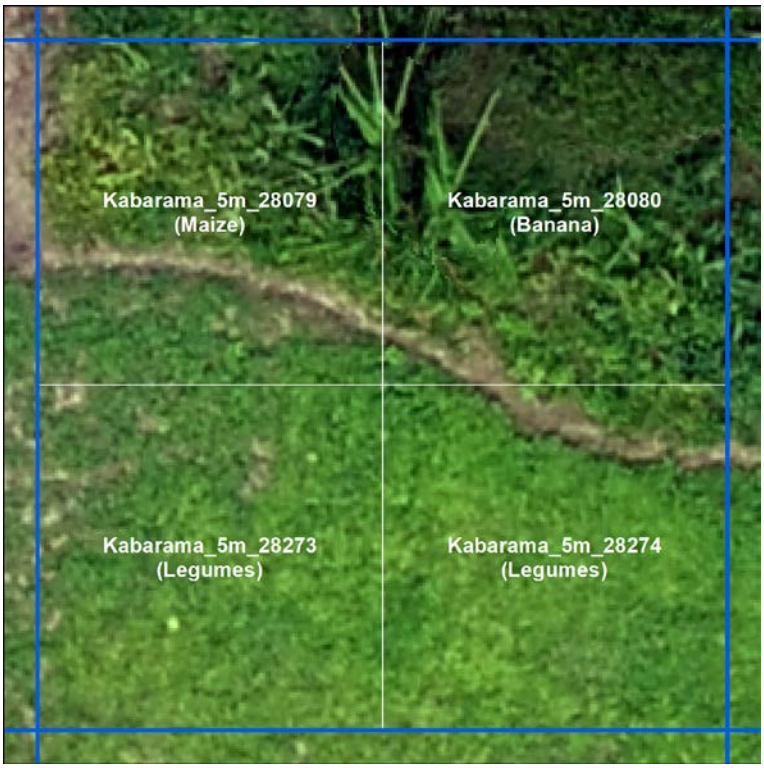
To begin the label generation process, the 2022 drone imagery was broken into 5m x 5m image files (chits). The pre-trained computer vision model classified each image as being in one of six land cover categories: banana, cassava, forest, legumes, and structures. The number of chits per category is presented in **Table 5**.

**Table 5:**
Count of Chits by Land Cover Classified by Computer Vision Model

| Site | Banana | Maize | Legume | Forest | Structure | Other | Total |
|---|---|---|---|---|---|---|---|
| Cyamparita | 7,129 | 18,001 | 35 | 3,473 | 4,566 | 1,376 | 34,580 |
| Kabarama | 5,210 | 23,312 | 192 | 1,989 | 684 | 1,136 | 32,523 |
| Kaberege | 1,253 | 8,345 | 715 | 13,414 | 1,369 | 6,400 | 31,496 |
| Kinyaga | 1,369 | 11,800 | 1,417 | 12,321 | 146 | 3,809 | 30,862 |
| Ngarama | 3,794 | 18,203 | 856 | 5,955 | 755 | 2,191 | 31,754 |
| Rwakigarati | 7,846 | 8,801 | 1,038 | 8,875 | 720 | 2,160 | 29,440 |
| **Total** | 26,601 (14.0%) | 88,462 (46.4%) | 4,253 (2.2%) | 46,027 (24.1%) | 8,240 (4.3%) | 17,072 (9.0%) | 190,655 |

Once the computer vision model classified each of the 5m x 5m image chits, an algorithm was created to reconcile the land cover types for each 10m x 10m Sentinel-2 cell. The general logic of this algorithm was to consider the land covers in each of the 5 m chits that makes up a 10 m Sentinel-2 cell and assign them a consensus category. An example of four image chits that make up a single Sentinel-2 cell is presented in **Figure 2**.

**Figure 2:**
Drone Imagery Division into Chits




Sentinel-2 cells with either 3 or 4 of the chits in agreement were assigned the land cover category of those chits. Sentinel-2 cells with two or less chits in agreement were assigned "mixed." In **Figure 2**, two cells are labelled as "Legumes", one as "Maize", and one as "Banana" resulting in a "Mixed" Sentinel label. This process produced 7,000–8,000 label points per study site. The number of Sentinel-2 labels by land cover class is presented in **Table 6**.

**Table 6:**
Count of Labels by Land Cover Classified by Computer Vision Model

| Site | Banana | Maize | Legume | Forest | Structure | Other | Mixed | Total |
|---|---|---|---|---|---|---|---|---|
| Cyamparita | 999 | 3,760 | 0 | 341 | 64 | 72 | 2,367 | 7,603 |
| Kabarama | 934 | 5,555 | 1 | 157 | 2 | 27 | 1,070 | 7,746 |
| Kaberege | 64 | 1,221 | 11 | 2,436 | 49 | 656 | 3,063 | 7,500 |
| Kinyaga | 183 | 2,152 | 37 | 2,227 | 8 | 295 | 2,816 | 7,718 |
| Ngarama | 345 | 3,729 | 18 | 857 | 18 | 103 | 2,527 | 7,597 |
| Rwakigarati | 1,417 | 1,203 | 23 | 1,416 | 134 | 62 | 3,105 | 7,360 |
| **Total** | 3,942 (8.6%) | 17,620 (38.7%) | 90 (0.2%) | 7,434 (16.3%) | 275 (0.6%) | 1,215 (2.3%) | 14,948 (32.8%) | 45,524 |

The aggregation process generated a sizeable number of banana, maize, and forest category labels but also created many (14,948) mixed pixel labels. This process reduced all categories from their prevalence in the chits dataset but had the greatest impact on legumes, which went from 2.2% of the chits dataset, to only 0.2% in the labels dataset. This indicates that landcovers that cover larger areas and are more homogeneous are more likely to retain their category when aggregated from the 5 m chit level to the 10 m Sentinel-2 cell level, while smaller plots are more likely to be reclassified as "Mixed".

Before using the labels to train a satellite model, we evaluated them for accuracy by overlaying the ground observations and drone imagery labels, which served as ground truth. The results of the evaluation using field observations labels is presented in **Annex C Table C4**, and the results of the evaluation using drone imagery labels is presented in **Annex C Table C5**.

Although the overall agreement between predicted categories and ground truth categories was only about 67%, both evaluations indicated that banana, maize, and forest predictions agreed well with the ground truth labels, ranging from 81.0% to 92.5%. These results were sufficiently encouraging that we used the labels to train a Sentinel satellite model.

## Model Performance

Several models were creating using the computer vision labels. Models were run using all Sentinel-1 and Sentinel-2 bands, and a 100 m buffer. Two types of models were created: crop/noncrop and crop type. A comparison of model accuracies and Kappa coefficients is presented in **Table 7**.

**Table 7:**
Comparison of Model Performance Using Computer Vision Labels

| Site | Banana | Maize | Legume |
|---|---|---|---|
| Crop/Noncrop (computer vision evaluated against field observation labels) | 68.4% | 0.377 | Noncrop included natural vegetation, forest, bare earth, other, mixed. Crop included banana, maize, cassava, beans, sweet potatoes, Irish potatoes, other crops. |
| Crop/Noncrop (computer vision evaluated against drone imagery labels) | 73.1% | 0.465 | |
| Crop Type (computer vision evaluated against field observation labels) | 70.3% | 0.365 | Crop types were cassava, banana, maize, beans, sweet potatoes, Irish potatoes, and other crops. Classifier used both full crop and intercropped categories. |
| Crop Type (computer vision evaluated against drone imagery labels) | 73.4% | 0.496 | |

The computer vision label trained crop/non-crop model did not perform particularly well, with only 68.4% accuracy when validated against the field observation labels, and 73.1% when validated against the drone imagery labels. The performance improved for the crop type model with 70.3% agreement with field observation labels, and 73.4% agreement with the drone imagery labels. The better performance when validating against the drone imagery labels follows logically, since both sets of labels were derived from the same set of drone images.

Where the computer vision labels seemed to make the biggest difference was with the crop type classifier as validated by drone imagery labels. This classifier used all the crop types, including the intercropped version of each crop. Only three crops were included in this evaluation since bananas, maize, and legumes(beans) were the only three crops in common between the computer vision labels and the drone imagery labels. The confusion matrix of this best performing model are shown in **Table 8**.

**Table 8:**

Confusion Matrix of Model Trained with Computer Vision Labels and Evaluated Using Drone Imagery Labels

| | Banana Labels | Maize Labels | Beans Labels | Prediction Accuracy |
|---|---|---|---|---|
| Banana (Pred) | 251 | 23 | 17 | 86.3% |
| Maize (Pred) | 67 | 609 | 204 | 69.2% |
| Beans (Pred) | 0 | 0 | 0 | 0.0 |
| Label Accuracy | 78.9% | 96.4% | 0.0% | 73.4% |

The success of this model is driven by the large number of predicted and verified banana examples. Maize fared well but predicted 204 maize cells where the label indicated that the crop present was beans. This is likely due to the small number of bean examples in the training dataset as compared with the other land cover types. The model did not predict any cells as being beans. The accuracy of this model is not directly comparable to the other model performance accuracy values however. Both the field observation and drone imagery trained models has more crop types (six plus "other") so worse accuracy would be expected when evaluating those models. So, although the computer vision trained model performed the best, it was on fewer (three) categories.

## Conclusions

Several conclusions can be drawn from this analysis. The first is that using a pre-trained computer vision can produce a large number of chit labels with relatively little effort. Although our computer vision model was trained on data from 2019 Season A, it held up reasonably well, and could be further improved by adding new land cover examples. The second is that it is possible to generate Sentinel satellite labels from computer vision labels with similar accuracy to that achieved with human labeling. The process only produced an accuracy of ~67%, but we feel this number could be improved with better training data, and better ground truth evaluation data. The process produced 15 times more labels than either the field observation labels or drone imagery labels and provides the satellite model with additional data points allowing it to recognize a greater variability of Sentinel reflectance values. The third is that the model performance is decent and these values compare favorably with the models trained with either field observations or drone imagery labels. This indicates that satellite-based models can be trained using labels generated by a machine learning process. Given that this entire workflow can be added to over time, rather than recreated each season, and requires the least amount of human interaction, we feel this has the greatest potential to produce a paradigm shift in ground truth data generation that could accelerate data availability at a lower cost.

- It is possible to automatically create labels from drone imagery that serve as training data for a satellite model. The process filters out smaller plots and creates a large number of mixed labels. The model performance for the landcovers that are well represented is good.

# Overall Lessons Learned

## Things That Went Well

1. In general, field staff were able to more reliably identify crop types through direct observation, than drone imagery interpreters. If study protocols are followed, field observations should provide an even more reliable dataset.
2. The study site size was acceptable in terms of field logistics, project timelines, and budget. If the study area were larger, we expect drone imagery acquisition would have performed relatively better due to the consistency in data collection afforded by drone-based assessments.
3. High-resolution drone imagery was collected relatively quickly across our areas of interest within our specifications, which we anticipated would allow for high accuracy labeling (though in practice we found that it was difficult to discern land cover type for some of the randomly selected locations).
4. The computer vision model developed using data from 2019 Season A transferred to 2022 Season A relatively well for the major land cover types included (maize, bananas, forest) and was used to label the entire area flown by drones in less than a day.
5. There appears to be strong potential for continued development of computer vision models that incorporate additional data to improve accuracy (particularly for mixed Sentinel-2 cells) to be applied for very rapid labeling of drone imagery. The large number of labels that can be readily generated seems to offer good potential for improving performance of future machine learning models as accuracy of the underlying computer vision model is enhanced

## Things to Improve Upon

6. Drone imagery resolution was high-resolution, but it would be worth trying the drone labeling with even higher resolution imagery (0.5-1 cm) in order to enable better classification of images.
7. The drone imagery labeling would have benefited from more training for the labelers, review by a trained agronomist, or it might be possible to contract this work out entirely to a trained agronomist (or group of agronomists). While we relied on numerous samples labeled by a Rwandan agronomist to train labelers under this project, it may nonetheless be more difficult for non-agronomist analysts to tell the difference between certain land covers.
8. Ground truth data might benefit from using polygonal labels instead of point labels for field data collection, as it might reduce GPS error. In addition, an external GPS receiver would help bring GPS error down to ~1m, if budgets allowed.
9. Conducting an in-country training for field data collection staff would improve the quality of ground truth data. This would ensure that staff understand what specific data are required and improve consistency between staff. This would increase the project budget, however.
10. Gathering data at a training site, and then evaluating those data would help the project leaders understand the nature of the data they are getting, as well as understand any limitations. Adjustments could be made to minimize those limitations before the field work was done.
11. Refining the way image chits are aggregated into Sentinel-2 cells might result in better Sentinel-2 cell level labels. Instead of a 3+ category agreement cutoff, the probabilities of the category of each image chit could be used in the calculation.
12. It may be worth reassessing the land cover categories utilized to ensure use of categories that balance interest in more land cover category disaggregation with selection of classes that are sufficiently spectrally distinct. In addition to focusing on remote sensing data, incorporation of spatially disaggregated environmental/biophysical data (e.g., soil type, climate, phenological information) and data on historical cropping patterns by region may aid in improving model accuracy.

# References


Burke, M., Driscoll, A., Lobell, D. B., & Ermon, S. (2021). Using satellite imagery to understand and promote sustainable development. *Science*, 371(6535), eabe8628. https://doi.org/10.1126/science.abe8628

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(1), 7. https://doi.org/10.3390/drones4010007

Fu, W., Ma, J., Chen, P., & Chen, F. (2020). Remote sensing satellites for digital earth. In H. Guo, M. F. Goodchild, & A. Annoni (Eds.), *Manual of Digital Earth* (pp. 55-123). Springer Singapore. https://doi.org/10.1007/978-981-32-9915-3_3

Hegarty-Craver, M., Polly, J., O'Neil, M., Ujeneza, N., Rineer, J., Beach, R., Lapidus, D., & Temple, D. (2020, 06/20). Remote crop mapping at scale: Using satellite imagery and UAV-acquired data as ground truth. *Remote Sensing*, 12, 1984. https://doi.org/10.3390/rs12121984

Orynbaikyzy, A., Gessner, U., Conrad, C. (2022) Spatial Transferability of Random Forest Models for Crop Type Classification Using Sentinel-1 and Sentinel-2. Remote Sens. 2022, 14,1493. https://doi.org/10.3390/ rs14061493

Ren, S., Malof, J., Fetter, T.R., Beach, R., Rineer, J., and Bradbury, K. (2022). Utilizing geospatial data for assessing energy security: Mapping small solar home systems using unmanned aerial vehicles and deep learning. *ISPRS International Journal of Geo-Information* 11(4):222. https://doi.org/10.3390/ijgi11040222

Wen, Y., Li, X., Mu, H., Zhong, L., Chen, H., Zeng, Y., Miao, S., Su, W., Gong, P., Li, B., and Huang, J. (2022). Mapping corn dynamics using limited but representative samples with adaptive strategies. *ISPRS Journal of Photogrammetry and Remote Sensing* 190: 252-266. https://doi.org/10.1016/j.isprsjprs.2022.06.012

Enabling Crop Analytics At Scale

## Supporting Documents

### Annexes A-G



# Streamlining Ground Truth Data Collection: Rwanda Case Study


**info.ecaas@tetratech.com**
**cropanalytics.net**


**Final report**
August 2022

**Prepared for:**


Drew Wheadon
**Tetra Tech:** International Development Services
159 Bank Street, Suite 300, Burlington, VT 05401


**Prepared by:**


Jamie Cajka, Robert Beach, Gray Martin, Nick Kruskamp
**RTI International** 3040 E. Cornwallis Road, Research Triangle Park, NC 27709



RTI Project Number 0218248.000

# A

## Annex: Field Data Collection

RTI contracted with Vanguard Economics, a small research and consulting firm specializing in agricultural development, and located in Kigali, Rwanda. Vanguard has experience collecting field data using an Open Data Kit (ODK)-based data collection platform called SurveyCTO. We determined that this was the preferred software for this effort since it could collect point data, crop and non-crop data, and as Vanguard was familiar with this platform, it would minimize training. Prior to visiting the sites, Vanguard reached out to the local farmers and farmers' associations to get permission to perform the data collection on their plots.

## Field Protocols

RTI divided each of the six study areas into 40 equal-sized rectangles. Within each rectangle, we generated a randomly located point (**Figure A1**) and calculated its latitude and longitude. These coordinates were provided to Vanguard in the form of a spreadsheet. The field protocol was for a staff member to navigate to the specified coordinate within each rectangle. They were then to collect 20 examples of crop types or land covers by walking in a random direction. The examples were to be at least 10 m from each other. At each of the 20 locations, field staff were to identify which crops/land covers were present within a 5 m radius and their proportion. They recorded the GPS location as well as took a photo. Example photos can be seen in **Annex G**, **Figures G1-G8**.

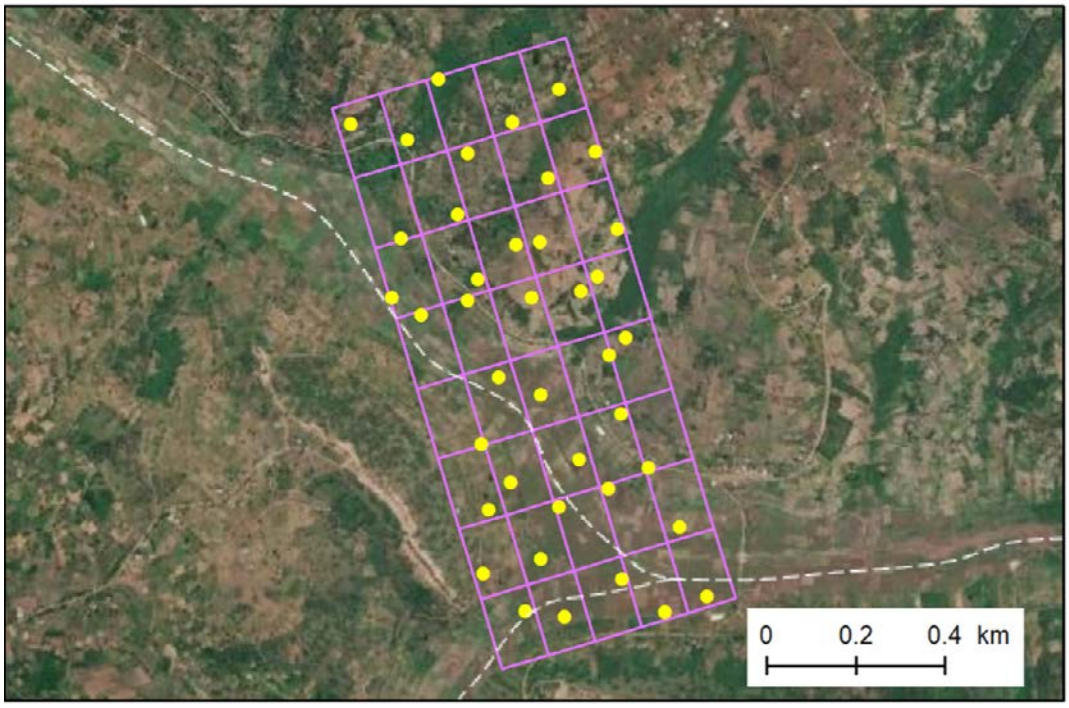



**Figure A1:**
Random Field Observation Locations in Kabarama

## Data Received

Data were collected by an 11-person team (10 data collectors and 1 supervisor). Vanguard was able to complete the required number of samples for a given study area of approximately 80 ha in a single day. In total, Vanguard provided crop locations for over 5,000 locations within the six study areas. The data were provided in the form of Stata files, which were converted to Excel spreadsheets. The data dictionary for the data is presented in **Table A1**.

**Table A1:**

Data Dictionary of SurveyCTO form

| Field Name | Description |
|---|---|
| deviceid | Id number of tablet or phone |
| subscriberid | Id number of subscriber if applicable |
| simid | SIM card number if applicable |
| devicephonenum | Phone number of device if applicable |
| username | User name of current project |
| researcher | Field staff name |
| duration | Number of seconds to record observation |
| time_check_sec_0 | Date/Time of beginning of observation |
| Site | Cyampirita/Kabarama/Kaberege/Kinyaga/Ngarama/Rwakigarati |
| Fid | |
| cid | |
| Crop_yes | Yes/No if crop present at location |
| Crop_name | Crops present at location |
| Maize | Binary flag for presence of maize |
| Banana | Binary flag for presence of banana |
| Cassava | Binary flag for presence of cassava |
| Sweet_potato | Binary flag for presence of sweet potato |
| Beans | Binary flag for presence of beans |
| Irish_potato | Binary flag for presence of Irish potato |
| Other_crop | Binary flag for presence of other crop |
| Crop_name_other | Name of other crop if applicable |
| Maize_size | Estimated percentage of maize in 5 m radius |
| Banana_size | Estimated percentage of banana in 5 m radius |
| Cassava_size | Estimated percentage of cassava in 5 m radius |
| Spotato_size | Estimated percentage of sweet potato in 5 m radius |
| Beans_size | Estimated percentage of beans in 5 m radius |
| Ipotato_size | Estimated percentage of Irish potato in 5 m radius |
| Size_oth | Estimated percentage of other crop in 5 m radius |
| Cover_yes | Yes/No if other type of land cover present |
| Cover | Name of other type of land cover (forest, natural vegetation, bare earth) |
| Cover_oth | Name of other type of land cover not in forest, natural vegetation, bare earth |
| Geo_locationlatitude | Latitude in decimal degrees |
| Geo_locationlongitude | Longitude in decimal degrees |
| Geo_locationaltitude | Altitude in meters |
| Geo_locationaccuracy | Accuracy to 1 standard deviation in meters |
| Picture | Path to image taken at location |
| Instancename | Not used |
| Formdef_version | Version of data collection form |
| Key | Unique key for record |
| Submissiondate | Date/time of record submission |
| Starttime | Date/time of start of record |
| Endtime | Date/time of end of record |
| Date_interview | Date of record |

**Table A2:**
Field Observation Data Collection Parameters

| Site | Data Collected | Number of Points |
|---|---|---|
| Kabarama | December 18, 2021 | 801 |
| Kinyaga | December 27, 2021 | 1,008 |
| Rwakigarati | December 28, 2021 | 1,001 |
| Kaberege | January 15, 2022 (originally scheduled for December 29, 2021) | 1,009 |
| Ngarama | December 30, 2021 | 1,029 |
| Cyampirita | December 31, 2021 | 1,002 |
| **Total** | | **5,850** |

## Lessons Learned

Ideally, RTI would have traveled to Rwanda to conduct field data collection training and perform a field test. Due to COVID-19 restrictions however, travel was not feasible, and we used the first study site as our field test. Although RTI created what we believed to be a simple, easy-to-implement sampling and data collection protocol, the Vanguard team did not always follow it. In general, the field team seemed to be able get to the approximately correct starting point, but in a debrief they indicated that they had trouble doing so using the GPS built into their data collection app. In addition, we measured the average distance between adjacent sample points, and found that it was 5.2m, almost half of the 10m we had specified in the protocols. We also noted that many examples of the same crop/land cover types were captured in the same general area. This was helpful for crops, but not helpful for examples of buildings, rocks, and water features. In the future, a trip to the study site(s) to provide training and a field test would be valuable to further improve data quality.

A key attribute of the field data collected was the GPS accuracy. This measure was provided as part of each sample point's metadata. The measure defines the radius of a circle in meters that would include 1 standard deviation of the data assuming a normal data distribution. Taken another way, it means that there is a 68% chance that the actual location is within a circle of the GPS error reported. The GPS error for all sites ranged between 1.8 m and 5.0 m with a mean error value of 4.4 m. Although this is higher than we would have liked, it is understandable given the use of the native GPS receiver in a consumer-grade tablet, and the study locations in rural Rwanda, which do not necessarily have the benefit of cellular or Wi-Fi networks to augment the satellites available for GPS positional calculations. We realized that this was going to reduce the utility of our crop data after getting the data from the first site (Kabarama) and brought this to Vanguard's attention. However, the GPS error remained unchanged for the remaining five sites. Vanguard attributed the error to the lack of these cellular and Wi-Fi networks, as well as tree and cloud cover obscuring the sky. We asked about using an external GPS

device, but Vanguard reports that high accuracy (< 10 cm) units costs approximately US$4,000 so they do not use them. Less expensive (< US$200) external GPS receivers promise faster satellite fixes and longer battery life, but only modest accuracy gains (~ +/- 2.5m).

As an alternative, a few companies make lower cost/higher accuracy GPS devices that work with tablets and smart phones. Juniper Systems makes a product called Geode that achieves 40-60 cm accuracy and costs approximately US $1,500. Bad Elf makes a receiver called the GNSS Surveyor that delivers ~1m accuracy and costs approximately US $650. Given the need to use field data collection apps running on consumer-grade mobile wireless signals and to reduce GPS error as much as possible, we recommend testing the Bad Elf GNSS Surveyor Bluetooth GPS receiver. The drawback is that each field team member collecting data would need to have their own GPS receiver so any project budget would need to account for this.

Although we explored the use of the Field Mapper tool, Vanguard preferred to use SurveyCTO, their normal data collection platform, which therefore had a low training burden. Another reason for selecting SurveyCTO was that it allows the collection of discreet points, whereas Field Mapper requires that data are collected as polygons. However, given the GPS error inherent to any individual point, it might be advantageous to collect field boundaries, which would define the area using hundreds of points and thus may smooth out errors. The advantage of collecting point data is that it is faster than identifying a field that has a consistent mixture of crops and walking around the perimeter. Collecting polygonal data would take more training, and more time in the field, resulting in longer data collection times and higher data costs. It may also introduce bias into selecting fields that are more uniform in their crop composition and less intercropped, thus reducing the range of training data.

# B

## Annex: Drone Data Collection

One way to potentially reduce data collection time and cost is to use drones. Drones can collect very high-resolution (~3–4 cm) imagery over small, targeted areas in a short period. The imagery can then be viewed and labeled by a trained analyst using a simple web application. These data can also be used to train a computer vision model (e.g., using deep neural networks), which can then be used to classify the entirety of the drone imagery. This could provide thousands of labeled examples with which to calibrate a satellite-based model. In addition to visual spectrum imagery, drones can be fitted with a multi-spectral sensor that can record infrared bands that are both useful in discriminating crop types and providing a measure of crop health through indices such as the Normalized Difference Vegetation Index (NDVI).

RTI contracted with Charis UAS, a drone service provider located in Kigali, Rwanda. Charis has the reputation of being one of the top drone operators in Africa. The purpose of obtaining drone imagery was to provide a way to generate ground truth data for the crop types and land uses of interest in the project. RTI used Charis to fly the same site locations as were flown during its work there in 2018–2020.

## Protocols

Since the main purpose of drone imagery is to identify crop types using a web-based viewer, the imagery must be of sufficiently high resolution, but must also be mosaicked correctly, free of blurriness, and properly color balanced. For these drone flights, RTI specified an orthophoto mosaic using the red, green, and blue (RGB) bands to form a true color image. Additional bands can be captured using an infrared (IR) sensor, instead of an RGB sensor, but this would require a second set of flights, effectively doubling the cost and imagery collection time. For applications that involve applying machine learning to classify crops directly from the drone imagery, it may be helpful to acquire these additional near infrared bans for improving accuracy. Resolution was to be 3-4 cm, and GPS accuracy was to be better than 10 cm. The vendor was responsible for obtaining the necessary permissions and permits.

## Data Received

Charis delivered a single color orthophoto mosaic for each of the six sites. The dates and specifications are presented in **Table B1**.

**Table B1:**
Field Observation Data Collection Parameters

| Study Site | Acquisition Date | Area (ha) | Resolution (cm) | Notes |
|---|---|---|---|---|
| Cyampirita | January 24, 2022 | 77.7 | 3.97 | |
| Kabarama | January 24, 2022 | 79.7 | 3.87 | |
| Kaberege | February 17, 2022 | 77.0 | 2.99 | Wrong site was flown initially. Operator returned 3 weeks later to re-fly. |
| Kinyaga | January 28, 2022 | 79.5 | 3.38 | |
| Ngarama | January 25, 2022 | 78.4 | 3.94 | |
| Rwakigarati | January 27, 2022 | 77.3 | 3.54 + 3.62 | Site comprised of two separate polygons flown at slightly different resolutions. |

## Lessons Learned

From a crop identification standpoint, it is important to capture the crops as close to maturity as possible, so that reliable visual interpretation of crop types is facilitated. The optimal date in the growing season will depend on the range of crops that are being identified. Getting imagery too early will make similar crops indistinguishable since they will not have leafed out sufficiently. Waiting too long will result in some crops being harvested. Bracketing the growth stages is a good strategy, although our budget and timeline did not allow for multiple drone flights in this case. In our experience, two or three flights over the same areas helps identify crops that might not be identifiable (or present) on a single image. Choosing an optimal window and starting the drone planning/acquisition process as early as possible is important. The time it takes between deciding to fly and getting a drone in the air is minimally one month due to flight planning, payments, and permit acquisition. In our case, the Kaberege site was flown using an old boundary, which necessitated Charis returning to the correct area approximately three weeks later. This resulted in some of the crops being already harvested, and therefore slightly reduced the number of random label points. We were still able to achieve over 3,000 label points and more than 100 in each land cover category.

# C

## Annex: Ground Truth Dataset Creation

Creation of ground truth data is a necessary, albeit costly and time-consuming, step in the creation of satellite-based predictive crop models. RTI created ground truth data two ways: through field observations and through drone imagery interpretation. The two datasets were created independently from each other. To create a common reference to evaluate the ground truth datasets, Sentinel-2 imagery was converted from raster format to polygons and clipped at the study area boundary. Each Sentinel-2 cell was given a unique ID. A process was then performed to classify the entire Sentinel-2 cell, based on either the field observations or visual interpretation. The possible categories are listed in **Table C1**.

# Field Data

## Processing

The field-derived point data contained a flag indicating whether any of the six main crop types were present at the sample location. In addition, field staff were asked to estimate the proportion of each crop type in a 5 m radius surrounding the sample location. Based on this, each individual point was classified in the categories in **Table C1**. If any given crop occupied 80% or more of the surrounding area, we classified the point as being entirely that crop. Each of the 5,850 field observations was then spatially joined to the Sentinel-2 polygons giving them the unique id. A frequency of each land cover by grid cell ID was generated. In total, 2,726 grid cells were labeled. Some cells contained as many as 18 points, but most (58%) contained a single point. Ideally, we desired a single field observation per Sentinel-2 cell and attempted to avoid multiple observation points by creating a protocol whereby staff were to move at least 10 meters between observations. As previously mentioned, in-country training would have helped, as staff did not always follow this protocol.

Once the field observations had been given Sentinel-2 cell ids, we ran a script that automatically categorized the Sentinel-2 cells based on the points they contained. Where there was a single field observation or multiple observations of the same value, the Sentinel-2 cell was given that crop label. Where there were intercropped values, the cell was given an intercropped value. A total of 702 (26%) of the 2,726 cell labels were given an intercropped classification (**Table C1**). Where there were two or more conflicting field observations, the script wrote out the Sentinel-2 ID to a file for manual review. Approximately 10% of the Sentinel-2 cells needed manual review. This required looking at the number of observations in each land cover category as well as the proportion of each crop present at each observation. When additional information was required, the photo taken at the location was consulted. In most cases, the conflicting land cover categories produced an "intercropped" classification.

**Table C1:**
Land Cover Counts for Field Observations

| Land Cover | Homogenous Cells | Mixed Cells | Total | Percent Total |
|---|---|---|---|---|
| Banana | 111 | 118 | 229 | 8.4% |
| Beans | 129 | 116 | 245 | 9.0% |
| Cassava | 91 | 85 | 176 | 6.5% |
| Irish Potatoes | 149 | 21 | 170 | 6.2% |
| Maize | 583 | 202 | 785 | 28.8% |
| Sweet Potatoes | 137 | 43 | 180 | 6.6% |
| Other Crop | 82 | 102 | 184 | 6.7% |
| Fallow | 160 | 0 | 160 | 5.9% |
| Harvested | 0 | 0 | 0 | 0.0% |
| Natural Vegetation | 86 | 0 | 86 | 3.2% |
| Forest | 415 | 15 | 430 | 15.8% |
| Bare Ground | 1 | 0 | 0 | 0.0% |
| Buildings/Structures | 80 | 0 | 80 | 2.9% |
| **Total** | 2,024 | 702 | 2,726 | 100.0% |

We also set up a crosswalk between the categories mapping in 2022, and those mapped in 2019. This was used to created labels that would later be used to validate the predictive surface created by the 2019 model and 2022 Sentinel imagery (**Table C2**).

**Table C2:**
Landcover Category Mapping Between 2022 and 2019

| 2022 Landcover Class | 2019 Landcover Class |
|---|---|
| Banana | Banana |
| Beans | Beans |
| Cassava | Cassava |
| Forest | Trees |
| Irish Potatoes | Other Vegetative |
| Maize | Maize |
| Natural Vegetation | Natural Vegetation |
| Other | Other Vegetative |
| Structure | Bare Ground/Non-vegetative |
| Sweet Potatoes | Other Vegetation |

## Lessons Learned

The field data have several limitations. The first is that while the field staff are experienced with identifying crop types, their estimation of the mix of crops growing within a 5 m radius of each sample point was sometimes unreliable. We noticed that the columns containing the proportion of each crop sometimes did not add up to 100%. We also noticed that sometimes the field staff appeared to have focused on the land cover feature that caught their attention, rather than trying to accurately characterize the land cover at the sample location. Additional training would have helped both these issues.

The GPS error was enough that it may have placed some land cover type locations in the wrong Sentinel-2 cell. This is not a problem when identifying larger fields of a given crop, or large areas of forest or natural vegetation where the adjacent Sentinel-2 cell has the same composition. But in a smallholder setting, where crops can change in a short distance, the possibility of attributing a given land cover to the wrong Sentinel-2 cell is increased. In the future, it would be beneficial to test the use of capturing crop and land covers using field polygons. This may reduce the GPS error since there are more points captured using a polygon, but it may also bias field staff to selecting and delineating a field that is more homogenous, thus reducing the prevalence of intercropped fields in the training dataset.

## Drone Data

Interpretation of high-resolution drone imagery to identify crop types requires that the analyst has a knowledge of what certain crop types look like at each crop stage. For this work, RTI relied upon our previous work in Rwanda, which was conducted at the same six sites, using comparable drone imagery, taken during the same growing season, and encompassing the same crop types. We drew upon a catalog of hundreds of existing crop examples we compiled in 2018–2019 that were created with training provided by an agronomist in Rwanda. This was reinforced by scanning through the photos taken by Vanguard during their field data collection (examples: **Annex G Figures G1-G8**), and associating a given crop with its location on the drone imagery.

One key advantage of collecting drone imagery is that there is comprehensive imagery across the entire region where data were collected. These data can potentially be used for additional applications by reviewing the imagery to collect more information whereas there is no way to extract information that was not originally collected from the field data. For instance, the imagery used for crop classification could also be used for assessing crop condition, proximity to roads or processing infrastructure, and other research questions. Drone imagery collected during 2019 Season A was later used to assess the ability of machine learning models to accurately identify small solar home systems within the six regions of Rwanda in which drones were flown (Ren et al., 2022).

## Processing

RTI randomly selected a primary sample of 500 Sentinel-2 grid cells at each site to be labeled using the drone imagery. We selected an additional random sample of 250 cells in case we did not get a minimum of 100 examples of each of the main six crop types. We also created a web-based crop labeler in ArcGIS Portal comprised of the land cover labels (purple dots), the randomly selected Sentinel-2 cells (purple squares), and the drone orthophotos as shown in **Figure C1**.

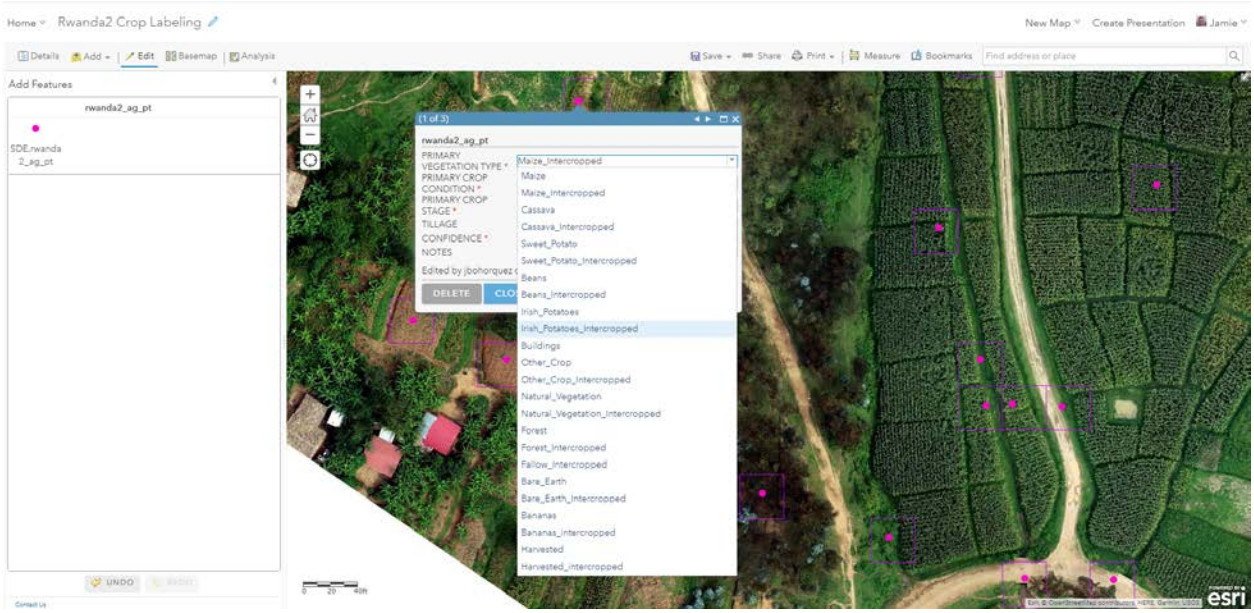



**Figure C1:**
Drone Labeler Interface

Two analysts scanned through each selected cell and created a label based on the choices in **Table 3**. Analysts were asked to select the land cover type that they felt was the most likely to characterize what they saw within the boundaries of the Sentinel-2 grid cell. Each land cover category also had a corresponding intercropped version. This produced a dataset of cell labels that contained approximately 33% intercropped cells. The analysts were also allowed to give a degree of confidence which might help when evaluating the training data. After the initial 3,000 grid cells were tallied, we found that cassava, sweet potatoes, and Irish potatoes had less than 100 examples by themselves. Therefore, we scanned through the additional sample of 250 random grid cells and added additional examples of these three crops. The total number of labels by land cover types can be seen in **Table C3**.

**Table C3:**
Land Cover Counts for Drone Imagery

| Land Cover Type | Homogenous Cells | Mixed Cells | Total | Percent Total |
|---|---|---|---|---|
| Bananas | 173 | 145 | 318 | 10.5% |
| Beans | 130 | 91 | 221 | 7.3% |
| Cassava | 102 | 93 | 195 | 6.4% |
| Irish Potatoes | 102 | 17 | 119 | 3.9% |
| Maize | 403 | 229 | 632 | 20.9% |
| Sweet Potatoes | 102 | 40 | 142 | 4.7% |
| Other Crop | 30 | 23 | 53 | 1.8% |
| Fallow | 64 | 0 | 64 | 2.1% |
| Harvested | 138 | 160 | 298 | 9.9% |
| Natural Vegetation | 297 | 85 | 382 | 12.6% |
| Forest | 312 | 55 | 367 | 12.1% |
| Bare Ground | 67 | 56 | 123 | 4.1% |
| Buildings/Structures | 112 | 0 | 112 | 3.7% |
| **Total** | 2,032 | 994 | 3,026 | 100.0% |

## Lessons Learned

Training of analysts, unless they are already trained agronomists, is essential. Analysts will have difficulty telling some land covers apart at the beginning of their training.

Not all crops are discernible on drone imagery. Some crops look similar (beans, sweet potatoes, Irish potatoes), it is only possible to tell them apart with training and experience. Additionally, some crops such as beans, are easily mistaken for fallow or natural vegetation. Although we focused on six main crop types, it would have been much more difficult to include all the crops found in the six study sites. Higher resolution drone imagery would help, but there is a tradeoff between drone imagery acquisition cost and resolution. The difference between 3 cm and 4 cm drone imagery was not significant in its land cover identification utility. Other factors such as image clarity and color balance are just as important as image resolution. Future work should consider testing the utility of drone imagery in the 0.5 – 1 cm range. This will have budget and data storage implications, but it may be possible to reduce the drone area, thereby keeping the budget and file sizes intact. Reducing the drone area will, however, potentially reduce the variety of examples of land covers encountered.

The labeling and review of drone-based labels takes time. With the field-based observations, the only additional time required was that to aggregate the observations into Sentinel-2 cells. But with drone imagery, there is time spent labeling and reviewing the classification. Drone imagery does have one major advantage, which is that it allows an analyst to come back to the data, to either clarify classification or to label additional satellite grid cells.

The classification of drone imagery is more likely to result in an intercropped classification since the analyst can see the entire cell.

## Computer Vision Data

The results of the evaluation of the computer vision labels using the field observation labels as ground truth are presented in **Table C4**.

### Table C4:

Evaluation of Computer Vision Labels Using Field Observation Labels

| | Banana Labels | Forest Labels | Legumes Labels | Maize Labels | Other Labels | Prediction Accuracy |
|---|---|---|---|---|---|---|
| Banana (Pred) | 63 | 2 | 0 | 11 | 0 | 82.9% |
| Forest (Pred) | 1 | 249 | 0 | 34 | 7 | 85.6% |
| Legumes (Pred) | 7 | 28 | 1 | 36 | 4 | 1.3% |
| Maize (Pred) | 18 | 14 | 0 | 425 | 4 | 92.2% |
| Other (Pred) | 18 | 75 | 4 | 128 | 31 | 12.1% |
| Label Accuracy | 58.9% | 67.7% | 20.0% | 67.0% | 67.4% | 66.3% |

Overall, accuracy is only fair, but this is due to the large number of "Other" classifications that were actually maize and forest. The prediction accuracies for banana, forest, and maize are very good.

The results of the evaluation of the computer vision labels using the field observation labels as ground truth are presented in **Table C5**.

### Table C5:

Evaluation of Computer Vision Labels Using Drone Imagery Labels

| | Banana Labels | Forest Labels | Legumes Labels | Maize Labels | Other Labels | Prediction Accuracy |
|---|---|---|---|---|---|---|
| Banana (Pred) | 147 | 2 | 0 | 10 | 0 | 92.5% |
| Forest (Pred) | 7 | 200 | 0 | 40 | 0 | 81.0% |
| Legumes (Pred) | 1 | 0 | 0 | 43 | 18 | 0.0% |
| Maize (Pred) | 4 | 40 | 0 | 365 | 1 | 89.0% |
| Other (Pred) | 6 | 0 | 2 | 228 | 53 | 18.3% |
| Label Accuracy | 89.1% | 82.6% | 0.0% | 53.2% | 73.6% | 65.6% |

Again, the overall accuracy is fair, but similar to the field observation labels, this is due to the large number of "Other" classifications that were actually maize. The prediction accuracies for banana, forest, and maize are very good.

## Lessons Learned

The computer vision model does a good job of correctly categorizing individual chits, but the aggregation algorithm creates many "mixed" Sentinel cell labels. In doing so, it removes land covers that are small and not well represented. The aggregation process can potentially be improved by including other metrics such as the probability score that a chit belongs to a given category. Additionally, we may also be able to define other more refined categories such as "maize intercropped" rather than just "mixed".

# D

## Annex: Selection of Satellite Imagery

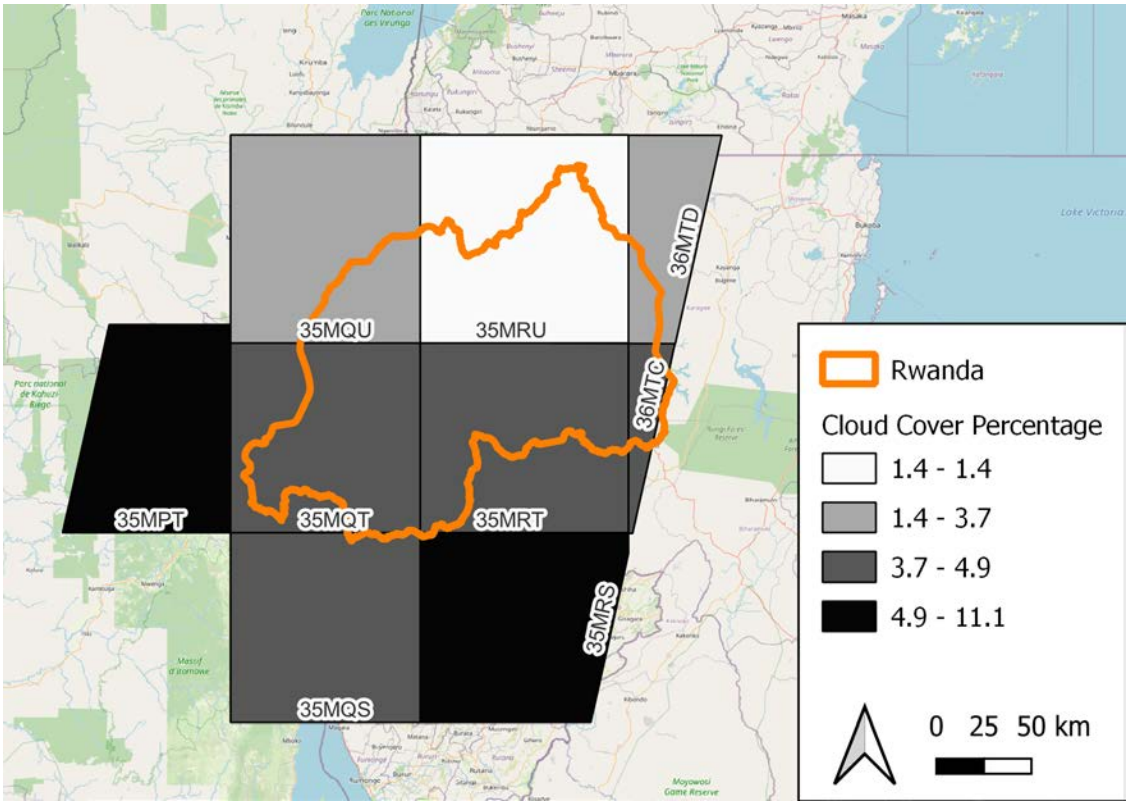
## Date Range

The remote sensing data were created from a mosaic of Sentinel-1 and Sentinel-2 tiles. The Sentinel-2 Level 2A images provide bottom of atmosphere reflectance data that have been geometrically and atmospherically corrected. In addition, the processing level provides pixel-level classification of the scene. The Sentinel-1 and Sentinel-2 data are both collected by two satellites with identical sensors. The Sentinel-1 sensors collect three cloud penetrating Synthetic Aperture radar (SAR) bands at 20 m resolution. The Sentinel-2 sensors collect 12 broad electromagnetic spectral or thermal bands at various resolutions (10, 20, and 60 m) depending on the band. There are a total of nine tiles that cover the country of Rwanda. These tiles were selected over the region to encompass the time period during field collection, from December 1, 2021, through February 28, 2022.

## Cloud Cover

For each Sentinel-2 tile, the scene with the lowest cloud cover percentage was selected from this period. This resulted in eight images from January 12 and one image from January 7 yielding the lowest cloud cover percentage. The images were retrieved from the European Space Agency Copernicus API and downloaded in the standard SAFE file format. Image bands at 20 m and 60 m were downscaled using nearest neighbor resampling to 10 m to match the finest resolution available. These bands were then stacked together for each scene. Each scene was masked using the Level 2A classification to remove clouds, water, shadows, and other pixels that were not valid data. The two clear classifications—values 4 (vegetation) and 5 (not vegetated, i.e., clear ground, urban)—were used to ensure only clear pixels existed in the final mosaic. The scenes were then mosaiced together to provide nearly full coverage of the country. Two images, encompassing the easternmost portion of the country (tiles 36 MTC and 36 MTD), were excluded because they do not cover any current study area and caused pixel misalignment due to being provided in a different projection. The final mosaic with cloud cover percentage is shown in **Figure D1**.

**Figure D1:**
Coverage of Sentinel Imagery



# E

# Annex: Modeling Environment

RTI performed environment setup and data management using a Python workflow. For data access and preprocessing, we used the GeoPandas package to import the labeled field and drone data and to determine buffer zones around each data point. Additionally, we used the Rasterio package to access the Sentinel-2 and Sentinel-1 imagery stack.

For each labeled data point, we extracted the set of pixel values from the imagery stack that fell within a circular buffer, taking the average value as the value for the data point. We repeated this process for buffer radii of 10 meters, 30 meters, and 100 meters, yielding three separate datasets. The intention was to smooth the data, removing spectral outliers. In our exploratory analysis, we found the best results with data buffered to 100 meters; we used this data to train the models as described below.

For modeling, we developed an R workflow using the caret package to set up, train, and cross-validate random forest classifiers. These classifiers used the following inputs:

1. 12 average pixel values (one per spectral band) extracted from the Sentinel-2 imagery stack
2. 3 average pixel values (one per spectral band) extracted from the Sentinel-1 imagery stack
3. Label data in the form of latitude, longitude, and consensus category

We allocated 70% of labeled data points to training and 30% to model evaluation. To prevent overfitting, we performed 3 sets of 10-fold cross-validation for each model and took the best result. Each random forest classifier used 1000 estimators with a minimum terminal node size of 1.

# F

# Annex: Modeling Results

## Using Field Observations as Training Data

The first model created classified satellite pixels as crop or noncrop. Natural vegetation and forest were considered noncrop for this analysis. This produced the confusion matrix shown in **Table F1**.

**Table F1:**
Crop/Noncrop Model Accuracy Using Field Observation Labels

| | Noncrop Labels | Crop Labels | Prediction Accuracy |
|---|---|---|---|
| Noncrop (Pred) | 209 | 54 | 79.5% |
| Crop (Pred) | 73 | 481 | 86.8% |
| Label Accuracy | 74.1% | 89.9% | 84.5% |

The model had an overall accuracy of 84.5%, performing better with respect to predicting crop (86.5%) than noncrop cells (79.5%). This type of model is valuable since information on cropland area is important for government agencies and other stakeholders interested in assessing overall planting and high-level data on agricultural area planted by region.

For specific crops within the area identified as cropland, the field observations produced an overall accuracy of 68.1 %. The success varied widely between crop types however, as shown in **Table F2**. Irish potato was correctly predicted 79.6% of the time, beans were correctly predicted 51.4%, bananas were correctly predicted 74.3%, maize was correctly predicted 75.4%, cassava was correctly predicted 42.2%, and sweet potatoes were correctly predicted 57.1%.

Overall, the ground observation trained model performed fairly well given the relatively large number of crop categories (7), and the relatively small number of training points in some of the less common crop categories. For example, bananas, beans, cassava, and sweet potatoes had approximately 100 labels each, with a 70:30 split leaving less than 100 to train the model and 30 or so to evaluate it. In general, we found that the model performed better in categories with the largest number of training labels.

**Table F2:**
Confusion Matrix for Model Trained Using Field Observation Labels

| | Cassava Labels | Bananas Labels | Maize Labels | Other Crop Labels | Sweet Potato Labels | Beans Labels | Irish Potato Labels | Prediction Accuracy |
|---|---|---|---|---|---|---|---|---|
| Cassava (Pred) | 19 | 5 | 9 | 0 | 3 | 9 | 0 | 42.2% |
| Banana (Pred) | 5 | 52 | 7 | 0 | 1 | 5 | 0 | 74.3% |
| Maize (Pred) | 13 | 6 | 187 | 0 | 22 | 12 | 48 | 75.4% |
| Other Crop (Pred) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0% |
| Sweet Potato (Pred) | 5 | 1 | 8 | 0 | 24 | 4 | 0 | 57.1% |
| Beans (Pred) | 10 | 1 | 15 | 0 | 4 | 38 | 0 | 51.4% |
| Irish Potato (Pred) | 0 | 0 | 6 | 0 | 0 | 5 | 43 | 79.6% |
| Label Accuracy | 36.5% | 76.5% | 79.6% | 0% | 44.4% | 52.1% | 84.3% | 68.1% |

## Using Drone Imagery Labels as Training Data

We also trained a crop/noncrop model using drone imagery labels. Again, natural vegetation and forest were considered noncrop. The validation produced a confusion matrix as shown in **Table F3**.

**Table F3:**
Crop/Noncrop Model Accuracy Using Drone Imagery Labels

| | Noncrop Labels | Crop Labels | Prediction Accuracy |
|---|---|---|---|
| Noncrop (Pred) | 271 | 92 | 74.7% |
| Crop (Pred) | 132 | 412 | 75.7% |
| Label Accuracy | 67.2% | 81.8% | 75.3% |

The crop/noncrop model had an overall accuracy of 75.3%, which is 6.2% worse than the model trained using field observations. The model performed slightly better when predicting crop (75.7%) as compared to noncrop cells (74.7%).

For specific crops, the drone observations produced an overall accuracy of 58.4%, which is 9.7% worse than the model trained using field observations. The success varied widely between crop types, however, as is shown in **Table F4**. Irish potato was correctly predicted 57.1% of the time, beans were correctly predicted 43.6%, bananas were correctly predicted 60.2%, maize was correctly predicted 65.9%, cassava was correctly predicted 50.0%, and sweet potatoes were correctly predicted 44.4%.

**Table F4:**

Confusion Matrix for Model Trained Using Drone Imagery Labels

| | Cassava Labels | Banana Labels | Maize Labels | Other Crop Labels | Sweet Potato Labels | Beans Labels | Irish Potato Labels | Prediction Accuracy |
|---|---|---|---|---|---|---|---|---|
| Cassava (Pred) | 23 | 5 | 12 | 0 | 2 | 4 | 0 | 50.0% |
| Banana (Pred) | 12 | 62 | 12 | 5 | 5 | 5 | 2 | 60.2% |
| Maize (Pred) | 16 | 19 | 145 | 5 | 15 | 16 | 4 | 65.9% |
| Other Crop (Pred) | 1 | 0 | 1 | 2 | 0 | 0 | 1 | 40.0% |
| Sweet Potato (Pred) | 1 | 3 | 3 | 1 | 16 | 11 | 1 | 44.4% |
| Beans (Pred) | 5 | 6 | 10 | 1 | 2 | 24 | 7 | 43.6% |
| Irish Potato (Pred) | 0 | 0 | 6 | 1 | 2 | 6 | 20 | 57.1% |
| Label Accuracy | 39.7% | 65.3% | 76.7% | 13.3% | 38.1% | 36.4% | 57.1% | 58.4% |

## Using Machine Learning-Generated Labels as Training Data

A total of 45,524 Sentinel-2 labels were generated from the drone imagery using an existing computer vision model. This model was trained to discern six crop/land cover types: banana, maize, legumes, forest, structure, and other. Before using them to train a satellite model, we evaluated them for accuracy by overlaying the ground observations and drone imagery labels, which served as ground truth. The results of the evaluation using ground observations are presented in **Table F5**.

**Table F5:**

Evaluation of Computer Vision Labels Using Field Observation Labels

| | Banana Labels | Forest Labels | Legumes Labels | Maize Labels | Other Labels | Prediction Accuracy |
|---|---|---|---|---|---|---|
| Banana (Pred) | 63 | 2 | 0 | 11 | 0 | 82.9% |
| Forest (Pred) | 1 | 249 | 0 | 34 | 7 | 85.6% |
| Legumes (Pred) | 7 | 28 | 1 | 36 | 4 | 1.3% |
| Maize (Pred) | 18 | 14 | 0 | 425 | 4 | 92.2% |
| Other (Pred) | 18 | 75 | 4 | 128 | 31 | 12.1% |
| Label Accuracy | 58.9% | 67.7% | 20.0% | 67.0% | 67.4% | 66.3% |

The accuracy of the aggregation for the three main land covers (banana, forest, and maize) is reasonably good, even though we are transferring the computer vision model across growing seasons. As previously mentioned, legumes were mostly remapped to "mixed" so are not well represented. Matching to the "Other" category presents a challenge, since for the computer vision model it was used as a catch-all, whereas for our 2022 modeling it was used to include other crops. Therefore, the categories are not directly comparable. The lack of legume labels, and the broad range of land cover types in the "Other" category brought the overall accuracy of the model down significantly.

The accuracy for the Computer Vison labels evaluated using drone imagery labels is presented in **Table F6**.

**Table F6:**

Evaluation of Computer Vision Labels Using Drone Imagery Labels

| | Banana Labels | Forest Labels | Legumes Labels | Maize Labels | Other Labels | Prediction Accuracy |
|---|---|---|---|---|---|---|
| Banana (Pred) | 147 | 2 | 0 | 10 | 0 | 92.5% |
| Forest (Pred) | 7 | 200 | 0 | 40 | 0 | 81.0% |
| Legumes (Pred) | 1 | 0 | 0 | 43 | 18 | 0.0% |
| Maize (Pred) | 4 | 40 | 0 | 365 | 1 | 89.0% |
| Other (Pred) | 6 | 0 | 2 | 228 | 53 | 18.3% |
| Label Accuracy | 89.1% | 82.6% | 0.0% | 53.2% | 73.6% | 65.6% |

Again, the computer vision aggregation process evaluation indicated that it went well, with an overall accuracy of 65.6%. Banana, maize, and forest matched very well, but "Other" brought the overall accuracy of the aggregation down significantly as it turned out to be maize most of the time. However, these results were sufficiently encouraging that we used the labels to train a Sentinel satellite model.

## Creation of Satellite Model Using Computer Vision Labels

Several models were created using the computer vision labels. Models were run using all Sentinel-1 and Sentinel-2 bands, and a 100 m buffer. Two types of models were created: crop/noncrop and crop type. A comparison of model accuracies and Kappa coefficients is presented in **Table F7**.

Of all the crop/noncrop classifiers, the best performance was still achieved using the field observation data (84.5%). The computer vision label trained model did not perform as well, with only 68.4% accuracy when validated against the field observation labels, and 73.1% when validated against the drone imagery labels. The better performance when validating against the drone imagery labels makes sense, since both sets of labels were derived from the same set of drone images.

**Table F7:**
Evaluation of Computer Vision Labels Using Drone Imagery Labels

| Classifier | Accuracy | Kappa Coefficient | Classifier Description |
|---|---|---|---|
| Crop/Noncrop (field) | 84.5% | 0.650 | Noncrop included natural vegetation, forest, bare earth, other, mixed. Crop included banana, maize, cassava, beans, sweet potatoes, Irish potatoes, other crops. |
| Crop/Noncrop (drone) | 75.3% | 0.494 | |
| Crop/Noncrop (computer vision evaluated against field observation labels) | 68.4% | 0.377 | |
| Crop/Noncrop (computer vision evaluated against drone imagery labels) | 73.1% | 0.465 | |
| Crop Type (field) | 68.1% | 0.565 | Crop types were cassava, banana, maize, beans, sweet potatoes, Irish potatoes, and other crops. Classifier used both full crop and intercropped categories. |
| Crop Type (drone) | 58.4% | 0.451 | |
| Crop Type (computer vision evaluated against field observation labels) | 70.3% | 0.365 | |
| Crop Type (computer vision evaluated against drone imagery labels) | 73.4% | 0.496 | |

Where the computer vision labels did make a difference was with the crop type classifier. This classifier used all the crop types, included the intercropped version of each crop. As reported, the drone imagery trained model produced an overall accuracy of 58.4%, and the field observation trained model produced an overall accuracy of 68.1%. When evaluating the computer vision predictions against field observation labels, the accuracy was 70.3%. The crop type computer vision trained model produced a higher accuracy of 73.4%. Only three crops were included in this evaluation since bananas, maize, and legumes(beans) were the only three crops in common between the computer vision labels and the drone imagery labels. The validation of this best performing model produced the confusion matrix shown in **Table F8**.

**Table F8:**
Confusion Matrix of Model Trained with Computer Vision Labels and Evaluated Using Drone Imagery Labels

| | Banana Labels | Maize Labels | Beans Labels | Prediction Accuracy |
|---|---|---|---|---|
| Banana (Pred) | 251 | 23 | 17 | 86.3% |
| Maize (Pred) | 67 | 609 | 204 | 69.2% |
| Beans (Pred) | 0 | 0 | 0 | 0.0 |
| Label Accuracy | 78.9% | 96.4% | 0.0% | 73.4% |

The success of this model is driven by the large number of predicted and verified banana examples. Maize fared well but predicted 204 maize cells where the label indicated that the crop present was beans. This is likely due to the small number of bean examples in the training dataset. The model did not predict any cells as being beans. The accuracy of this model is not directly comparable to the other model performance accuracy values however. Both the field observation and drone imagery trained models has more crop types (six plus "other") so worse accuracy would be expected when evaluating those models. So, although the computer vision trained model performed the best, it was on fewer (three) categories.

Several conclusions can be drawn from this analysis. The first is that using a pre-trained computer vision can produce a large number of labels with relatively little effort. Although our computer vision model was a few years old, it held up well, and could even be improved by adding new land cover examples. The second is that it is possible to generate Sentinel satellite labels from computer vision labels. The process only produced an accuracy of ~67%, but we feel this number could be improved with better training data, and better ground truth evaluation data. The process produced more than 15 times more labels than either the field observation or drone imagery labeling process and providing the model with additional data points so it was able to recognize greater variability of Sentinel-2 reflectance values. Given that this is a process that can be added to over time, rather than recreated each season, and requires the least amount of human interaction, we feel this has the greatest potential to produce a paradigm shift in ground truth data generation that could accelerate data availability at a lower cost.

# G

# Annex: Examples of Land Cover Types

The following figures provide examples of the six main crop types we identified: maize, banana, beans, cassava, sweet potatoes, and Irish potatoes. Also included is an example of fallow and natural vegetation, which can be mistaken for crops. The yellow lines indicate the extent of Sentinel-2 grid cells, while the yellow dots represent field observations locations as recorded by the built-in GPS receiver in the tablet.

**Figure G1:**
Example Imagery of Maize

**Ground Image**

**Drone Image**

**Figure G2:**
Example Imagery of Banana

**Ground Image**

**Drone Image**

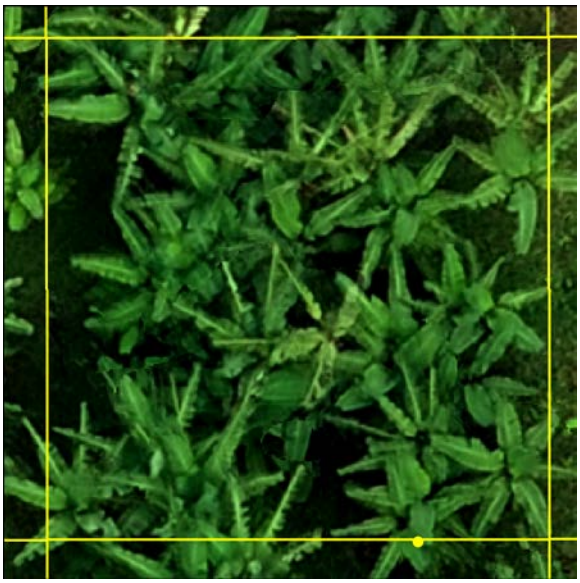


**Figure G3:**
Example Imagery of Climbing Beans

**Ground Image**

**Drone Image**

**Figure G4:**
Example Imagery of Fallow

**Ground Image**

**Drone Image**

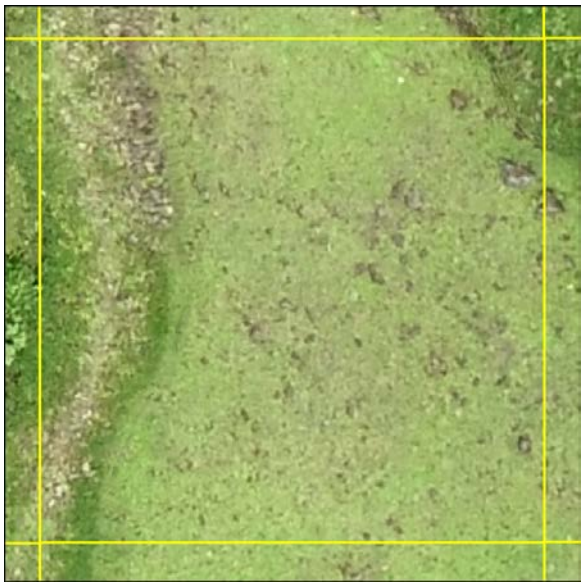


**Figure G5:**
Example Imagery of Sweet Potatoes

**Ground Image**

**Drone Image**

**Figure G6:**
Example Imagery of Natural Vegetation

**Ground Image**

**Drone Image**

**Figure G7:**
Example Imagery of Cassava

**Ground Image**

**Drone Image**

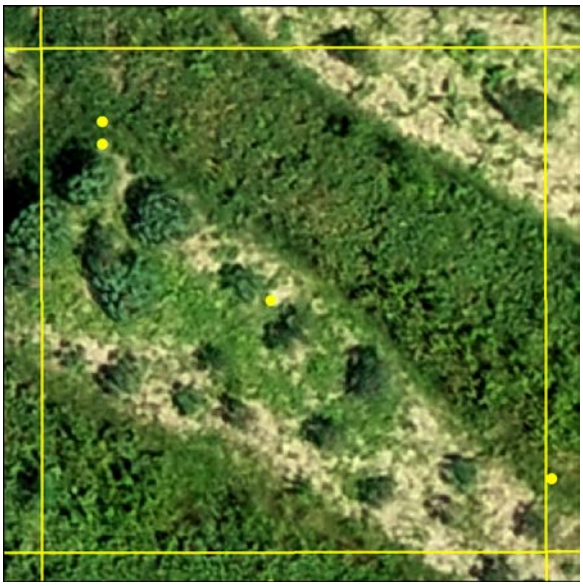
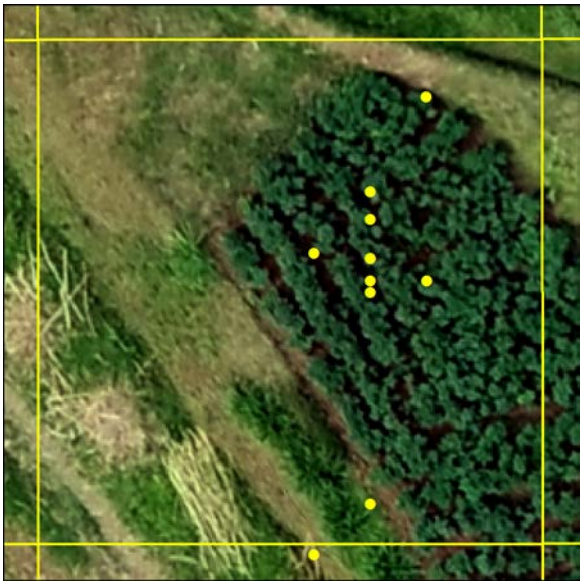


**Figure G8:**
Example of Irish Potatoes

**Ground Image**

**Drone Image**




# Streamlining Ground Truth Data Collection

## Rwanda Case Study

**info.ecaas@tetratech.com**
**cropanalytics.net**

### Final report

August 2022

### Prepared for:

Drew Wheadon
**Tetra Tech:** International Development Services
159 Bank Street, Suite 300, Burlington, VT 05401

### Prepared by:

Jamie Cajka, Robert Beach, Gray Martin, Nick Kruskamp
**RTI International** 3040 E. Cornwallis Road, Research Triangle Park, NC 27709

RTI Project Number 0218248.000