TETRA TECH

Enabling Crop Analytics At Scale

ENABLING CROP ANALYTICS AT SCALE (ECAAS)

Enhancing Agriculture Datasets for Remote Crop Monitoring

Contents

1

Introduction

The Need for Quality Ground Data to Improve Yield Predictions

Extreme weather events, fluctuating temperatures (and the concomitant socio-economic impacts), further exacerbated by the effects of the COVID-19 pandemic and recent food supply chain disruptions, have dramatically demonstrated the risks smallholder farmers face. In a context of climate change, adapting to the effects of more erratic weather and more intense shocks is imperative for various stakeholders — including governments, development organizations, agricultural companies, and credit providers/financial institutions. These changes are required to improve and maintain smallholder productivity and incomes.

At present, one of the most significant challenges is the lack of high-quality data that would allow stakeholders to develop adequate responses to manage risks. As a result, agricultural policies and solutions cannot be used systematically or preemptively to address climate change-related challenges as decision-makers cannot determine necessary and relevant efforts or interventions worth investing in to increase agricultural resilience, reduce losses and lift smallholder farmers out of poverty.

Further, the community-level impacts of deployed solutions, including agricultural insurance, subsidies, extension services or policy changes, are often not measured. Financial models for agricultural risks, such as extreme weather and climate change effects, are often based on satellite or topological data but rarely integrate robust and consistent field data measuring conditions on the ground including actual field boundaries, crop types, area planted, and crop health.

A key challenge in most sub-Saharan African countries is the lack of longitudinal yield data making it difficult to generate or model accurate estimates and expectations about yield trends and the factors that influence them. With ECAAS-Tetra Tech funding support, Pula is addressing this challenge by developing national yield maps for the main crops in Kenya and Zambia.

The overall objective of this project is to **enhance predictive capacity of yields and risk analysis by collecting ground data to increase the quality, depth, accuracy, and applicability for robust analysis.** Pula plans to use the agricultural data to estimate crop yields, collecting previously unavailable data, especially training datasets, for machine learning applications.

In the following sections, we have identified limitations in the current data collection approaches and how Pula aims to make data collection processes more efficient. This will involve real-time monitoring of the datasets coming in, flagging them if there are obvious discrepancies, and creating a platform that will help the team communicate back in near real-time to the field in case any corrective measures need to be carried out.

Gaps in Current Approaches for Data Collection

Currently, a wide variety of stakeholders, including Pula, use Crop Cut Experiments (CCE) for yield assessment. CCE is considered to be the most accurate method for quantifying crop yield – however, the fieldwork requirements result in high costs and time requirements. On a technical level, there are other limitations. The major challenge is that the CCE process collects coordinates on where the experiment was conducted, but does not help differentiate between areas where a crop of interest has been planted, vis-a-vis other crops. Pula uses the yield information from the CCEs to generalize production on a farm. While this indicates the expected yield from the farm, it assumes homogeneity of crop growth and health, which usually is not the case. Additionally, due to a lack of farm boundary data and accurate collection of geolocations, creating accurate and representative crop masks is usually not possible.

How the Proposed Approach Fill Key Gaps

The approach outlined here enhances current data collection methods through four key advances:

a. **Data collection app (Mavuno):** The application allows collection of hyper-accurate field data that can then be used to train forecasting models. Data points collected include: wet and dry weight of crops, hazards that may have affected crop performance; field boundaries; geolocations of farms.
b. **Field boundary data collection:** Every field data collection process we undertake now includes field boundaries. In addition, we have automated the process of polygon checks to ensure we are collecting accurate field boundaries that capture the crops of interest.
c. **Crop Type Masking and Yield Estimation:** Collection of field boundary data now makes it possible to accurately model yield at different spatial levels. In addition, we are integrating automatic data quality checks to ensure that any yield modeling done using the collected ground data, accurately reflects the conditions in the fields.
d. **Automated mapping platform (Skyfall):** Based on the lessons learnt and accurate data collection protocols developed through the ECAAS program, we are integrating cropland mapping, yield and hazard/peril models allowing the ability to map farmlands anywhere in the world within 2 hours and providing an accurate view of production cycles, expected yields and perils.

The Proposed Approach Improves Yield Predictions

In our assessment of model accuracy, we have identified improvements in yield predictions enabled due to automated data quality checks, more accurate cropland masks, and improved yield forecasting models.

1. **Automated data quality checks**
   These checks include checks to ensure that the field boundary has been properly tagged and that the crop metadata collected is from the farm that has been geo-referenced. These checks are automatic, ensuring erroneous data does not leave the source without being corrected.

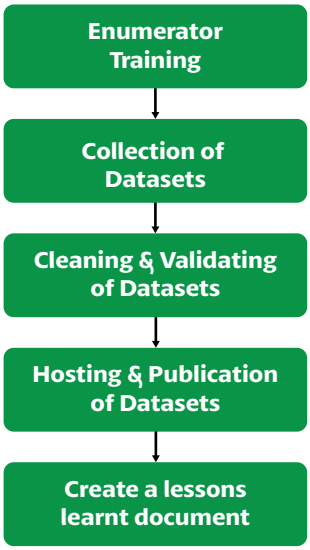
2. **Accurate cropland masks**
   With the field boundary data, crop mask models are then used to define crop masks for areas of interest. ML models are sensitive to the data they are trained on. It is the reason we have put a significant amount of effort and resources into ensuring the application used to collect the data supports accurate data collection through the automated data quality checks defined. A combination of empirical and ML models is used to generate these masks. **Our current models are now able to create masks with high accuracy (>90%) and a field variation of +/- 0.1 Ha.**

3. **Yield Forecasting**
   Based on the results from the above process, and the resulting data, we then use weather systems and environmental information to forecast yield at 40 days after planting and 1 month to harvest. Our initial tests so far closely match our field data 7/9 times.

2

Approach

Our approach to data collection generates accurate geo-tagged yield data that can be used to enhance model yield predictions. The meticulous approach was divided into 5 distinct phases including: Enumerator Training, Collection of Datasets, Cleaning and Validation of datasets, Hosting and publication of datasets, and Creating the lessons learnt document.



a) Enumerator Training

We conducted training to ensure enumerators were familiar with theoretical and practical elements associated with data collection best practices, data collection methodologies, and assembling data collection tools. Training sessions were conducted in Kenya for the long rains in 2021, for the short rains in Kenya in early 2022, and for the main season in Kenya in mid-2022. The enumerators also engaged in practical exercises in the field to get first-hand experience and to ensure that they understood the entire data collection process. A total of 118 Enumerators were trained, with 52 from Kenya and 66 from Zambia.

Figure 1.0: (left)
Training in Nakuru, Kenya for the ECAAS project in September 2021.
Figure 1.1: (right)
Training in Lusaka, Zambia for the ECAAS project in March 2022.

b) Collection of Datasets

We used the United Nations Food and Agriculture Organization's (FAO) crop calendars to create an agronomic calendar for the main crops in Kenya and Zambia including beans, maize, millet/sorghum and potatoes. **(Figures 1.2 and 1.3)**. Subsequently we selected several sample farms across FAO's Agro Ecological Zones (AEZs), and then developed a general data collection workflow for the datasets.

Kenya (Dec 16, 2021)

Harvesting of the bulk of the 2020 "long-rains" cereal crops has recently started in major uni-model rainfall growing areas of Central, Rift Valley, and Western Provinces.

Season Key

Sowing		Farmer's List	L
Growing		Recruit Enumerators	R
Harvesting		CCE- Placement	P
		CCE- Wet Weight	W
		CCE- Dry Weight	D

Primary Season Crops	JAN			FEB			MAR			APR			MAY			JUN			JUL			AUG			SEP			OCT			NOV			DEC		
Barley (Long Rains)					L								R	P	P	W	W	W	D	D	D	D	D													
Barley, Maize*, Millet, Sorghum & Beans (Short rains)		P	P	W	W	D	D																				L						R		
Beans (Long rains)							L												R	P	P	W	W	W	D	D	D	D	D							
Maize* (Long rains)						L																R	P	P	W	W	W	D	D	D	D	D				
Millet (Long rains)						L				R	P	W	W	W	D	D	D	D	D																	
Sorghum (Long rains)			L												R	P	P	W	W	W	D	D	D	D	D											
Wheat* (Long rains)												L										R	P	P	W	W	W	D	D	D	D	D				
Green grams (Short rains)		P	P	W	W	D	D																					L						R		
Green gram (Long rains)						L					R	P	W	W	D	D	D	D	D																	
Sunflower (Long rains)						L				R	P	P	W	W	W	D	D	D	D	D																
Sunflower (Short rains)	D	D	D	D																L									R	W	W	W	W	W	D	D
Soy Beans					L								R	P	P	W	W	W	D	D	D	D														

Figure 1.2:
Crop calendar for Kenya (Source: FAO Website)

Zambia (Nov 13, 2020)

Planting of 2021 cereal crops underway amid favourable weather conditions.

Season Key

Sowing		Farmer's List	L
Growing		Recruit Enumerators	R
Harvesting		CCE– Placement	P
		CCE– Wet Weight	W
		CCE– Dry Weight	D

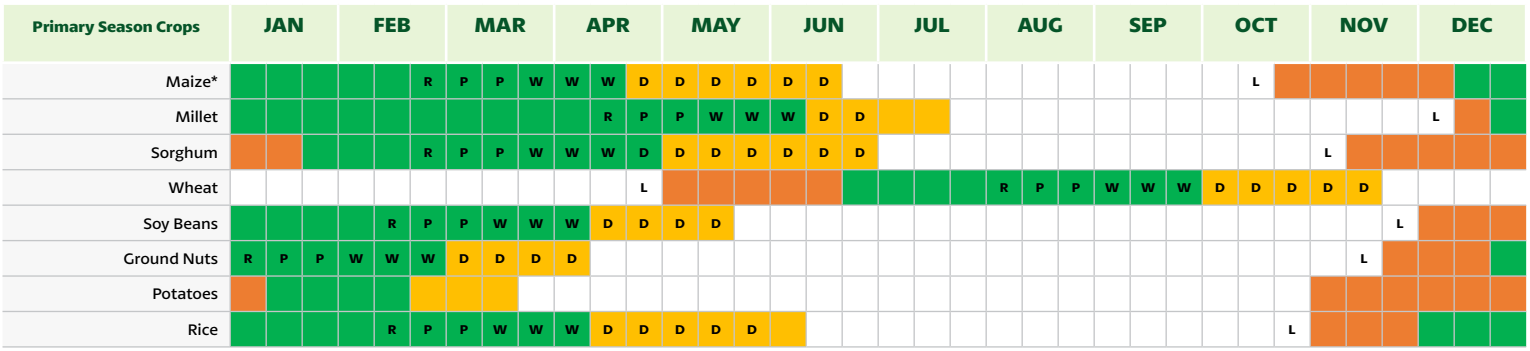


Figure 1.3:
Crop calendar for Zambia (Source: FAO Website)

We estimate yield using the crop cutting experiment (CCE) approach, which involves:

- **Box Placement.** During the Box Placement for beans, green grams, groundnuts, maize, and sorghum a box of 8 meters by 5 meters was laid by the enumerators. The boxes were placed just after the crops flowered to prevent farmers from tampering with the crops. During this visit, other attributes like crop condition, farm polygons, the boxes' center coordinates, intercropping data, and administrative boundary information were also captured.
- **Wet Harvest.** During the second visit, Wet Harvest, crops were harvested, weighed, and stored in sacks. The measurement taken in this stage is the wet weight of the freshly harvested crop which still has a lot of moisture content. Crops were labeled and left with the farmers who were directed to dry the crops to ensure they lost some moisture for the final visit.
- **Dry Harvest.** During the final visit, Dry Harvest, the harvest stored in the labeled sacks were threshed and their final dry weights were captured. The weight captured at this stage is considered the final weight of the crop and the average yield metric ton per hectare (MT/ha) is calculated depending on the size of the box placed. For Irish potatoes, only two visits were required as they are not dried after harvesting

Figure 1.4:
An enumerator placing a 5m by 5m box for Irish potatoes in Nyandarua County

Figure 1.5: (left)
Wet harvest of Maize in Uasin Gishu county

Figure 1.6: (right)
Threshing and winnowing of maize before weighing in Nandi County, Kenya

Weight measurement

The weight of the wet and dry harvested crop from the two 8m by 5m boxes are captured using a digital scale and recorded in the data collection app. We initially used Commcare, found in Google Play store. However, we found this proprietary app to be difficult to customize, and the post-processing involved in polygon capture delayed verification. Through this project we transitioned to an in-house app called Mavuno Pula, which offers greater customization. To ensure that we collected the correct weights we followed a strict protocol:

a. Making sure that the digital weighing scale is suspended freely, and the units are in kilograms.
b. When taking the measurements, the sacks holding the harvested crop should not be in contact with any surface.
c. Capturing the observed measurements twice to confirm that the correct measurements are captured into the data collection application.
d. The photos of the weighing scale and the sack holding the harvested crop are captured to ascertain that the weights were correctly captured.
e. An up-close photo of the weighing scale indicating the weight is captured as proof of the weighted values that have been captured.

Figure 1.7:
Green grams weighing scale value photo capture in Kitui, Kenya.

Lessons Learned

One of the major concerns with the box measurement sampling technique is the issue of representativeness. For instance, what happens when boxes lie on the part of the farm that performed either very well or very poorly, relative to the rest of the farm?

To mitigate potential data misrepresentation, we used satellite imagery to find the average vegetative index (VI) of a given farm. This average was then used to scan the VI map and sample out areas that exhibit the same VI. This method is, however, only ideal on relatively large farms (i.e., >2 acres) if using non-proprietary satellite imagery like Sentinel 2 with a resolution of 10m. During the vegetative stage there is also a possibility of missing a cloud-free satellite image to help in the above process.

c) Cleaning and Validation of Datasets

Despite the thorough training sessions, several human errors – principally, typographical errors (such as incorrect dates) – occurred during the data collection process. To mitigate these errors, the team set crop cycle window thresholds, as well as probable planting and harvesting date periods. For instance, the crop cycle window for common crops like maize cannot go beyond 12 months, and if this happens, such datasets are discarded. Both human errors and systematic errors were flagged via validation checks to increase the accuracy of the datasets. This exercise included removing missing or inconsistent data points and data formats, identifying and removing outliers using expert knowledge and remote sensing technology, and validating the datasets using established benchmarks. In some cases we corrected errors such as dates and polygon offsets, but in most cases erroneous data were eliminated.

Data quality assurance began with the training phase and ran through the final steps of the process. Some of the preliminary steps included:

1. Acquiring active farmer lists from working partners, which was important for planning purposes.
2. Defining targets per crop and Agroecological zone. These targets were highly dependent on the farmer lists that we had pieced together.
3. Field operations readiness, using the defined targets as a guide to organize field operations logistics like enumerator recruitment and allocation in readiness for training. During the training of the enumerators, the importance of collecting high-quality data was emphasized.
4. Audit and backchecking of the data collected, done through primary and secondary checks as summarized below.

Primary Checks

The primary data quality checks included automated scripting that verified the following parameters and flagged datasets that do not meet the set thresholds:

- **Time between successive data collection steps:** There is an allowed time allocation between collecting two data points or more for specific activities, including box placement or wet harvest surveys.
- **Weights:** Dry weights should not be higher than the wet weight.
- **Average yield outliers:** We fixed the upper limit for all the crops except ginger and cassava at 10Mt/Ha. We are looking into calculating this dynamically based on submissions. We also flag yield outliers using interquartile range (IQR) outlier detection.

L= Q1 – 1.5* IQR
H = Q3 + (1.5*IQR)

- **Where L is the lower outlier**
- **H is the higher outlier**
- **Q1 and Q3 are the average values of those quantities**
- **IQR is the interquartile range**

› **High moisture loss:** This assesses the average moisture loss between wet and dry harvest in both boxes. Historically, we have had 20–30% moisture. We thus use outlier detection methods (interquartile range (IQR) in our case).

Secondary Checks (Post processing)

Some data cleaning processes, such as polygon boundaries and planting/harvesting date, were not captured by automated scripts from the Business Intelligence dashboard.

Polygons

a. First Phase: Automatic quality control

We created a Python script to support the data validation process, specifically to determine:

1. Distance between coordinates of the two Crop Cutting Experiment (CCE) boxes. The CCE Protocol allows a maximum of 30 m between the two boxes. With an allowed GPS accuracy of 6m, we filtered out points that were more than 50 m apart.
2. Distance between box 1 and box 2 to the polygon's centroid. We considered a threshold of 50m, assuming a larger distance would indicate that the boxes belong to different farms.
3. Area of the polygon. We only kept polygons with areas more than 200 square meters. This assumes that most data users will be using Sentinel imagery for analysis and cannot work with extremely small plots.
4. Validity of the polygon. Polygons crossing each other are considered invalid and were filtered out.
5. Confirmation that the coordinates of box 1 and box 2 are inside the farm polygon.

b. Second Phase: Expert quality control

Geospatial and Data Science experts were given satellite imagery and were tasked to visually ascertain the location of the data points and make amendments on GIS software. The following activities were carried out:

- Ensuring that the data had the correct map projection(s).
- Checking whether the polygons fell within farm boundaries by visualizing data on GIS and having satellite imagery as the base map.
- Editing polygons that had inconsistencies with farm boundaries by looking at the base map.

Planting/Harvesting Dates

Most of the mistakes in the dates were a result of the enumerators picking the wrong year in the data collection app. For instance, an enumerator would erroneously pick 2023 instead of 2022 or 2020 instead of 2021. We used the crop season calendar and the difference between the planting and harvesting dates to make amendments. This issue was not found with the day and month fields.

General Data Cleaning

Removed data duplicates using the Case_ID attribute, which is the unique identifier. We also ensured all the data attributes have a uniform sentence case.

d) Hosting And Publication Of Datasets.

The datasets collected have been hosted in an accessible platform where anyone who needs to use the data will access it with ease. We developed a website that has links to both the documentation of the project as well as the link to the google drive storage. Our datasets follow the FAIR Data principles, which advocate for data that are **Findable**, which means that the data should be furnished with descriptive metadata; **Accessible**, which means that the data should be hosted on a platform that can be accessed by anyone; **Interoperable**, which means that the data should be integrated with other sources with ease; and **Reusable** which means that the datasets should have a clear usage license to enable users to use the data.

To help meet FAIR standards, our activities included:

1. Formulation of metadata for all datasets collected using the accepted metadata standards.
2. Storing the datasets in a cloud instance and developing specifications to enable easy access online.
3. Developing a common API that enables potential users to access the data.

The project aimed to create 4,000 cleaned and validated georeferenced yield datasets. We were able to create 4,063 data points from both Kenya and Zambia. The data had the following attributes

- Type of crop
- Variety of crop
- Crop season
- Planting dates
- Expected harvest dates
- Seed brands and variety
- Farm boundary
- Yield (MT/Ha)
- Crop Condition
- Intercropping

Geographies

- In Zambia, we were able to collect data in 108 out of 116 districts in the country
- In Kenya, we collected data in the Western, Eastern, Rift Valley, and Central Kenya regions.

Crop types

The crop types include beans, green grams, Irish potatoes, ground nuts, soya beans, maize, and sorghum. **Tables 1 and 2** below show the summary of the different crop types in Kenya and Zambia respectively.

	Crop_Type	Number	Percentage
	Beans	243	18%
	Green Grams	129	10%
	Irish-Potatoes	385	28%
	Maize	527	39%
	Sorghum	68	5%
	Total	**1352**	**100%**

Table 1:
Crop type datasets in Kenya

	Crop_Type	Number	Percentage
	Soybeans	261	9%
	Groundnuts	695	26%
	Irish-Potatoes	155	6%
	Maize	1552	57%
	Sorghum	48	2%
	Total	**2711**	**100%**

Table 2:
Crop type datasets in Zambia

e) Drone Data Collection

In addition to ground sampling, we had planned to collect 1,000 drone imagery data points (500 data points in Kenya and 500 data points in Zambia). We began data collection using drones in Kenya for the long rains in Nyandarua and Elgeyo Marakwet counties, where most of the long rains season crops had already matured. Drones were flown over 100 already collected data points. Our aim in using drone imagery was to try to solve some of the below issues:

- Mapping very small farms. Such farms can be 4 or 5 pixels on a 10m resolution satellite image, but most of these pixels are contaminated with reflectance over the edges of the farm boundary.
- Mapping intercropped farms, which is difficult using (especially open source) satellite imagery.
- Mapping weeds within fields

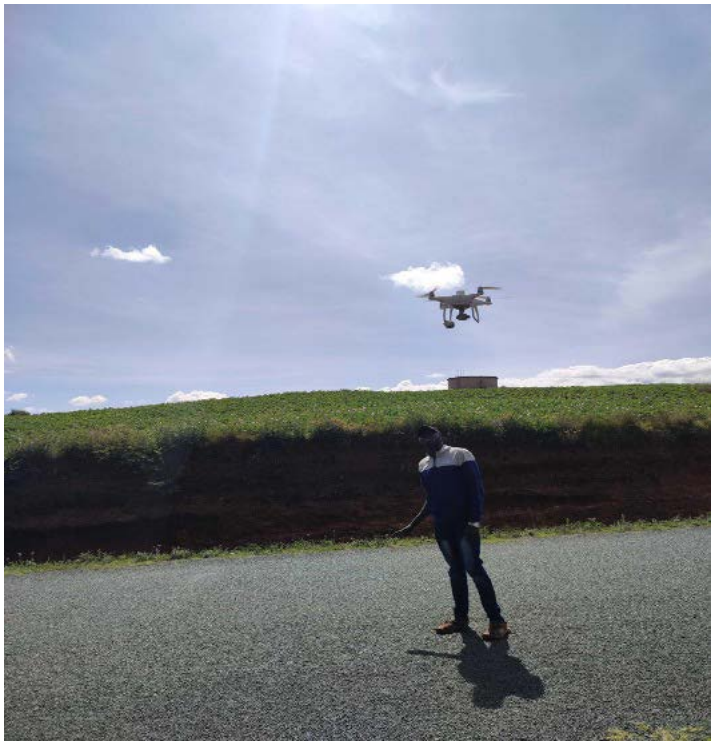


Figure 3: (left)
A drone mapping a potato farm in Elgeyo Marakwet County.
Figure 3.1: (right)
RGB imagery of a Potato farms in Elgeyo Marakwet

Figure 3.2: (left)
NDVI image for the above farm

Figure 3.3: (right)
The drone team setting up the drone to start the Drone Data collection exercise

We also collected 100 multispectral drone images for farms selected for data collection during the short rains season in Kenya.

The drone imagery campaign faced several unexpected challenges, including:

1. Heavy, unexpected rains.
2. Narrow vegetative stages resulting in very tight timelines to collect the imagery.
3. Political uncertainties brought about by the general elections in Kenya (and a court ruling that took 1 month to normalize) making travel and farm mapping perilous.
4. A drastic mid-project (300%) price increase from the drone service provider. Addrone's team had originally given a cost of $18 per acre to cover the project to completion, however market charges increased unexpectedly to $50-60 dollars.

Due to these challenges, we could not collect drone data in Zambia and most parts of Kenya during the short rains. We are currently collecting data from Kenya's long rainy season and have collected 200 data points from the coastal region; we anticipate finishing the data collection process in Mid-October.

3

Challenges

a) Drought

Kenya has faced five consecutive seasons of below-average rainfall, resulting in zero yields in some eastern parts of the country. (Please find the link to Kenya's current weather situation here). Data collection was expected to begin between January and February 2022 for the short rains, but the exercise began between late February and March 2022, owing to erratic weather patterns that led to variation in planting times across the counties.

We started the season with a farmer list of around 800 plots for the short rains data collection. This list shortened considerably since many of the crops had dried up. We were thus unable to meet our anticipated target. As a result of the drought, we only managed to clean and validate approximately 400 data points out of the targeted 1,000 data points. Figure 4 below shows the status of a farm that was affected by drought.

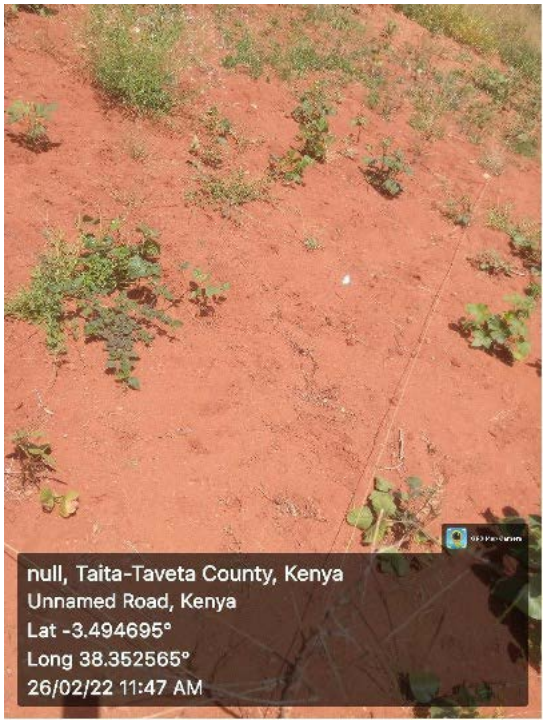


Figure 4:
A farm whose crop dried up due to drought in Taita Taveta County, Kenya

b) Small Landholding

According to FAO, smallholder farmers make up about 80% of all farmers in Kenya, and the average acreage is 0.47 ha (1.1 acres). In western Kenya, especially, small landholdings are dominant and many farmers plant on less than half an acre. This is attributed to a fairly higher population density in those regions as noted by this local daily from the 2019 Kenya Census Report.

If data from such landholdings are collected, they will be less useful since the highest spatial resolution open-source satellite data imagery is Sentinel-2, with a spatial resolution of 10m. For Sentinel-2 imagery the minimum acreage we recommend is 0.4 acres, which corresponds to more than 16 pixels. Accordingly, enumerators focused on slightly larger farms, which both delayed the data collection period and reduced the data collected from those regions.

Large-scale drone campaigns are an ideal solution for tackling the issue of small landholdings. However, the feasibility of these campaigns will be determined by several factors including the cost of carrying out such projects, the inability of drones to be used in extreme weather conditions, and government requirements and paperwork required to obtain flight permits.

c) Intercropping

Some crops, including beans, are intercropped with maize. After a certain stage of crop development, maize forms a canopy over the beans, thereby reducing visibility and accessibility for remote sensing (Figure 5). Especially for smallholder farmers, there is a financial incentive to intercrop in order to take advantage of their small farm spaces. While intercropping poses a major challenge from a data collection perspective, it also represents an interesting research opportunity to explore different approaches to map out intercropped systems. Future research should explore methods to accurately collect data on intercropped farms.

Figure 5:
A farm that has maize intercropped with beans.

d) Farm Polygons: Base map

Enumerators delineated farm boundaries by capturing the coordinates of the farm vertices. Vertices were offset, partly due to human error, and partly due to GPS accuracy (~6m). To correct this, we validated our data with Google satellite hybrid imagery. However, some of the imagery is relatively outdated (over 2 years old) and does not represent the current situation. For instance, the imagery might indicate that there are no farms whereas in reality there are several farms, or farm boundaries may have changed in the last two years. To avert the issue of using outdated base maps for the polygon cleaning, a higher spatial resolution satellite imagery (<3m) acquired during the crop life cycle would be useful and would significantly enhance the polygon data cleanup.

Figure 6:
A farm polygon with offsets (left) and corrected vertices (right)

e) Crop Vegetative Stages

During the drone data capture, we noticed that most crops' vegetative stages were narrow, and within a month from the box placement date, the crops would be past the vegetative stage. The drone team was also operating with one drone, it thus took some time to move to all the intended farms. The team had to therefore pass other farms since it was too late to capture the imagery.

Figure 7:
The drone team in a farm that it is past the vegetative stage

It was also difficult to fly the drone during the rainy season, as shown in **figure 8** below.

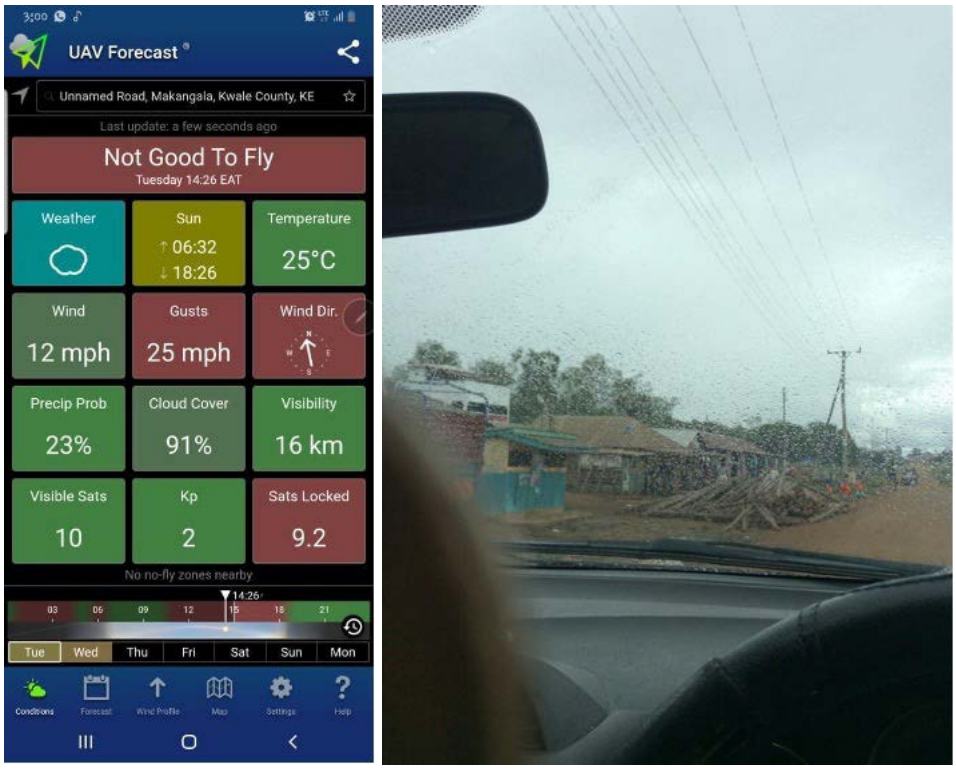


Figure 8:
Drone Flying weather forecasting application showing weather conditions

f) Drone Permits

We could not carry out data collection with drones in Zambia, as the contractor could not secure the required permits and documentation on time. This was due to bureaucratic requirements and changes in the Zambian government administration. The approval process, obtaining the required documentation and identifying a Zambia Civil Aviation observer to accompany the field team to the farms took close to two months, leading to most of the crops surpassing the vegetative stage

g) Data Capture Errors

Most of the drone imagery is taken during the box placement visit, which generally coincides with the vegetative state of the crop. After that, there is a gap of 1–2 months where the other two wet and dry visits are executed. At the end of the whole process, an accurate analysis of the datasets collected is carried out. Some of the data were removed because of quality issues. Such datasets had to be flagged even though we had already captured the drone imagery. As a result, we have examples of imagery but with limited attributes, possibly devoid of yield data. This will subsequently affect crop yield modeling if users use this data.

4

Successes

a) Farmer Sensitization

Before the project started, we focused on group discussions with farmers and county government agriculture officers. We outlined the objectives and possible outcomes of the project. We explained to the farmers and the agricultural officers that the data collected will be used for "utafiti" (Research). We further expounded that the process was intended to help us understand production in those areas. Consequently, we would be best placed to advise on the seed and type of fertilizer combinations the farmers should apply for optimal production. We also explained how the data we created would help researchers develop intelligent advisory systems, focused mainly on crop monitoring and information disseminated through commonly used media such as text messages. As a result of the awareness raising efforts, farmers were very receptive and responsive.

b) Dataset Collection and Publication

Despite the challenges observed in Kenya and Zambia, we collected 4,063 datasets across Irish potatoes, beans, maize, sorghum, groundnuts, and green grams. The datasets had several attributes which were important in crop health monitoring and the creation of yield prediction maps, among other uses. These attributes include yield data, farm polygons, and planting and harvesting dates. The datasets are published on the project website along with documentation and a repository for the datasets collected.

5

Lessons Learnt and Recommendations

Below we highlight lessons learned, together with recommendations on tackling some of these issues across different component areas of the project.

a) Survey Design

In the initial stages of survey design, some of the data attributes collected (in addition to final weight and farm polygons) were planting/harvesting dates, administrative boundaries, crop type, and season. Through the data collection process, we felt that we could use more data attributes which could be useful to help corroborate or enhance the datasets that we had already collected.

- We added a parameter to identify intercropped farms by including a Boolean response of "true" or "false". Additional detail on intercropping practices is required, however High-resolution remote sensing imagery (<3m) such as Planet's PlanetScope imagery or Airbus' Pleiades constellation enables discrimination between different types of intercropped crops. For most intercropped systems, the average spacing is about 0.5m.
- While in the field, enumerators can observe the crop conditions, which could be useful when inferring crop yields and can shed light on what to expect of the crops. This can be one of the many ways to validate the data points.

b) Crop Classification/Detection/Mapping

One of the key processes that acts as a predecessor to yield type prediction, mostly on a regional scale, is crop type mapping. Even though the project focused on yield data, future research could design a survey that facilitates crop type classification to create a robust dataset.

Before data collection, a crop signature survey should be carried out where enumerators in certain localities identify crop types and land covers and record the coordinates to form a basis for crop type classification. Prior to undertaking such research it is important to evaluate where similar projects have already collected data that can complement new research campaigns.

c) Rule-Based Crop Masking

Crop masks can also be created from the data collected using rule-based masking. This can only be achieved in smaller areas with homogenous climatic conditions with crops with a unique NDVI profile/curve. Within this context, we can create rules that could support generating specific crop masks. However, the limitations of this approach include the fact that some crops exhibit similar NDVI profile curves, especially those from the same scientific family; and climatic heterogeneity makes it difficult to zone areas, creating room for variation. To the best of our knowledge, no proven model evaluation methods can be used to determine the accuracy of these models.

Other metrics that can be used in place of NDVI are:

- EVI (Enhanced Vegetation Index): Good in areas with a dense canopy.
- NDRE (Normalized Difference Red Edge): Works well for thick crops and permanent crops.
- SAVI (Soil Adjusted Vegetative Index): Adjusts for the effect of soil brightness. Especially useful in arid and semi-arid regions.
- GCI (Green Chlorophyll Index): Does a great job at measuring the impact of seasonality.
- ARVI (Atmospheric Resistant Vegetation Index): Especially important in regions where there are atmospheric pollutants (air pollution, rain, fog, etc.)
- MTCI (MERIS Terrestrial Chlorophyll Index)

d) Data Validation of Polygons Using High-Resolution Satellite Imagery

To ensure robustness, we used automated and manual scripts to validate field boundaries. In addition, we verified each polygon using Google imagery to account for GPS measurement errors.

Unfortunately, a large amount of imagery is outdated (ranging from a few days to almost 2 years). This means there could be an error of commission or omission during the process.

The validation process is crucial – but our approach is limited by the accuracy of Google imagery. To overcome this challenge we recommend using higher resolution satellite imagery (less than 5m) for the specific crop cycle to help in the validation process. Getting such imagery would be expensive but would significantly enhance data quality and accuracy.

e) Validation Script

All validation scripts should be automated and incorporated into the data collection process. Throughout our project, the research team carried out post-processing and validation exercises, which forced the field team to spend some time moving back and forth trying to redo some of the data collection processes. When such scripts are automated, most survey anomalies will be flagged in real-time, prompting the data collection team to make amends before leaving the field.

This process would improve the data collection experience and help obtain high-quality datasets.

f) Drones Rules and Regulations

Drone regulations differ across countries. Laws governing the use of drones are still a work in progress in many African countries and hence very unpredictable. It is important to conduct due diligence before embarking on any drone project, especially where country boundaries are in play, and to keep abreast of the latest developments to act and plan accordingly.

This was our first time engaging in drone data collection, and we were unaware of the bureaucratic processes involved. Given our interest in drone data collection, we are planning to create avenues to involve ourselves in matters concerning drones, for example, establishing relationships and creating networks with civil aviation authorities in our countries of operations. We are also planning to acquire our own drone for research purposes, and this will be one way for us to keep abreast of these regulations.

g) The Value of Technology to Stakeholders

One of the critical learnings regarding drones is that stakeholders are ready to embrace technology only if we clarify the value of this technology. For this, we avoided using lots of technical jargon. These are some of the phrases that we used:

- "We are carrying out research using drones on advanced methods of crop monitoring"
- "We are using drones to research whether we can use drones in loss assessment when it comes to crop insurance"

Using drones to capture farm imagery involves the engagement of several stakeholders. The whole process of onboarding creates awareness, making them appreciate the role of technology in agriculture. This also plays a significant role in attracting youth to carve a niche in the agriculture sector, bearing in mind they are more likely to be tech-savvy and have a higher affinity to use gadgets than older generations. In the long run, this would increase research and innovation in the agriculture space.

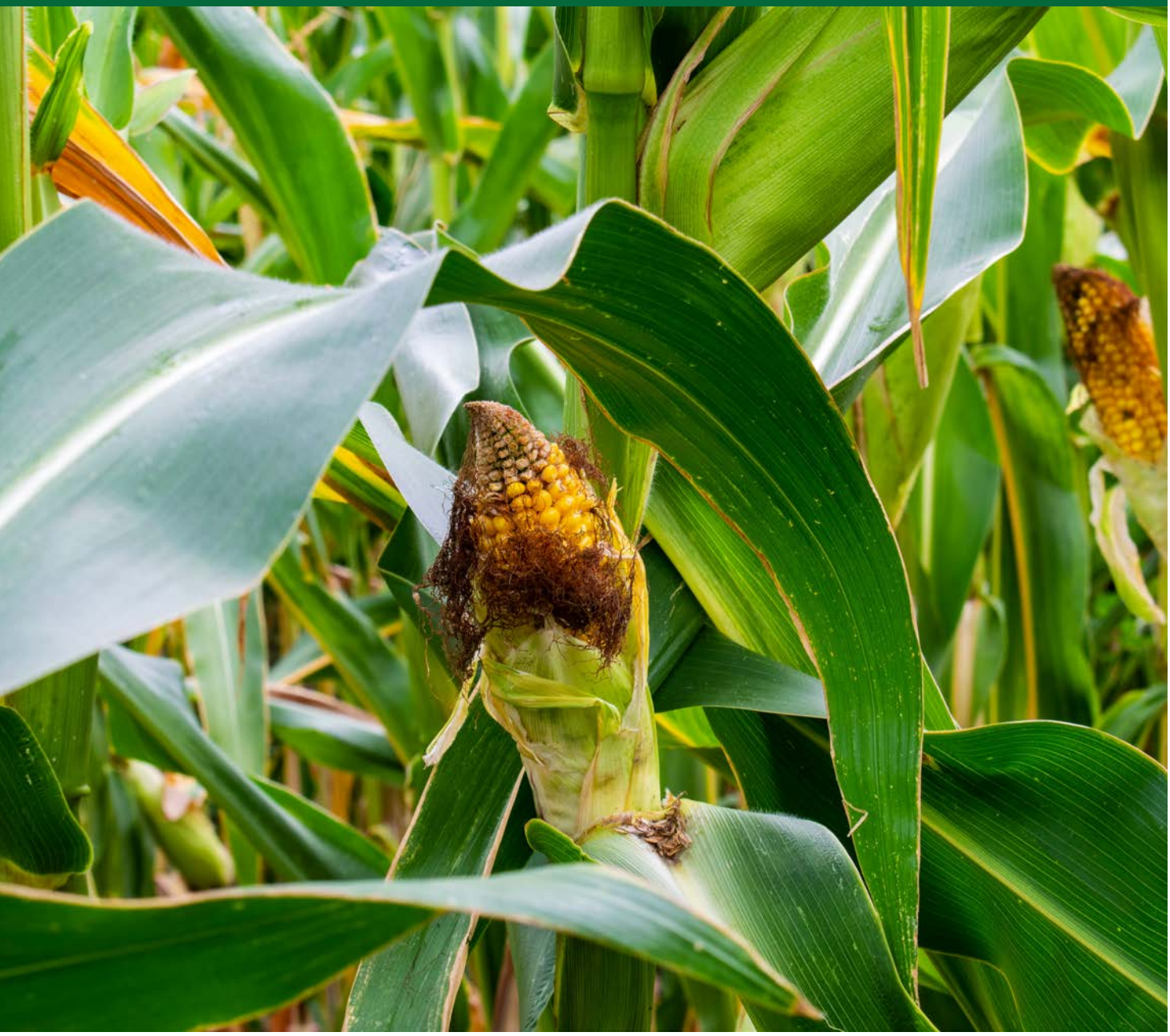
h) Weight Data Capture Automation and Checks

Accurate measurements of weight data are crucial as they inform the final yield data to be calculated. To improve data accuracy, appropriate software solutions should be used. The process of capturing these weights for the application in our data collection is not foolproof, owing to human errors when handling equipment and using the software.

To avert this, digital Bluetooth weighing scales can be programmed to capture the harvested crop weights once they are measured. Photos taken for dry and wet weights should have geo-location data in order to facilitate data verification (i.e., determining whether a data point was captured in a specific field).

6

Conclusion



Pula collected agricultural validation data to create local and national yield maps for Kenya and Zambia. Pula plans to use the yield and farm polygon data internally for research purposes, yield prediction, and crop health monitoring, and the analysis of the data has and will support the current index-based insurance products that Pula offers.

For instance, the team already created a data visualization dashboard from the data that presents actionable information and maps to agribusinesses, governments, financial institutions, and other critical stakeholders in the agriculture sector. The dashboard helps agribusinesses, governments, financial institutions, and other stakeholders make informed decisions about product development and marketing, disaster mitigation, public financing, and policies. The dashboard solution is technology-driven, analyzing remote sensing and validation data (including customized metrics as described above) and linked directly to Pula's existing business operations in the field.

Pula will also use the yield maps to inform their other digital products like iSMS and Skyfall which are used to help their clients monitor crop health and offer digital advisory services to help clients better manage their farms and subsequently increase production levels.

Enabling Crop Analytics At Scale

Enhancing Agriculture Datasets for Remote Crop Monitoring

info.ecaas@tetratech.com
cropanalytics.net

Final report
August 2022

Prepared by:
Pula Advisors GmbH