Data delivery report

Afrokai

Date: 16-6-2022
Author: Jildemarie Brouwer
Focus crop(s): Maize & Sorghum, Potatoes
Country: Uganda

Introduction **1**

Survey design notes **2**
- Structure of both surveys 2
- Two focus crops in one survey (Maize & Sorghum) 2

Sample characteristics **3**

Data cleaning steps **3**
- Introduction 3
- Removing Farmers from the Set 4
- Text Cleaning 4
- Determining and Handling Outliers 4
- Addressing Missing Values 5
- Anonymizing 5
- Repeated question groups 5
- Case specific adjustments 5
- Other notes 6

Notes from data collection **7**
- Introduction 7
- General Feedback 7

Introduction

This document contains details of the Afrokai primary data collection and data cleaning steps. Please read the PDC process guide before consulting this document.

The data collections for Afrokai took place between 30th of May till 4th of June 2022, and consisted of two data collections. One for potato farmers in Burahya county (32 farmers) and one for Maize & Sorghum farmers in the counties Amuru, Kasese, Lira and Nwoya (365 farmers). No farmers were deleted from the datasets.

Survey design notes

Structure of both surveys

Both surveys consisted of the usual core questions. We also included questions on the farmer agent model. Please check the questions with variable names starting with "*cs_sdm_company_agent..*" for the questions on agents that were added upon request by Landmark Millers.

Note that with only using the core survey, we are not asking questions about costs for livestock. Hence, the calculation of household income as well as farm income, does not contain those costs. Next, unfortunately we made a mistake with the maize & sorghum survey by removing the question asking for income from livestock while keeping a similar question asking for the income from "other" livestock. This results is that we are not able to calculate net-income from livestock.

Two focus crops in one survey (Maize & Sorghum)

The maize & sorghum survey differs from other surveys with regards to questions about the focus crop, because we had 2 focus crops: maize and sorghum. The questions about labour and input use were unaffected. Below we have some additional information about Maize & Sorghum

Crop seasons

Farmers produced maize and/or sorghum throughout the year. They can produce maize in season 1 and season 2, maize in season 1 and sorghum in season 2 (or other way around), or sorghum in season 1 and season 2.

Certification

No certification schemes exist, hence the questions related to this topic were removed.

Sample characteristics

To get an accurate picture of the farmers we take a random sample of the farmer group. However, we often run into issues with the farmer list supplied by the company. Farmers for example might not have an address and we are dependent on a local contact person to take us to the right farmer. We often find the farmer is not the farmer that is part of the sample. As the random component of sampling is very important for the reliability of our findings we try to register whether the farmers we speak to were part of the original sample.

The amount of farmers part of the original sample in case of Landmark Millers can be found below:

Survey question: Was the farmer part of the original sample?		
Maize & Sorghum		
Response	**Nr. of farmers**	**Share of farmers**
No	113	31%
No, he/she is an alternative for a sampled farmer that was unavailable	25	7%
Yes	226	62%
Potatoes		
Response	**Nr. of farmers**	**Share of farmers**
No	20	63%
No, he/she is an alternative for a sampled farmer that was unavailable	2	6%
Yes	10	31%

Data cleaning steps

Introduction

This section contains an overview of the different steps that are taken to clean the data. These steps have been drawn up in cooperation with IDH-FarmFit analysts and will be discussed in the following order:

- Removing Farmers from the Set
- Text cleaning
- Determining and handling outliers
- Looking at missing values
- Anonymizing
- Repeated question groups
- Case specific adjustments

Removing Farmers from the Set

Farmers are only removed from the set if they refused to participate in the survey. The only data we have from these farmers is the name, location and sometimes a phone number. None of the farmers refused to participate in the survey

Text Cleaning

In order to make the FarmFit data more accessible a few general steps are taken to clean the data.

- All columns and text values are set to lowercase
- Flow sets spaces to points; we set them to '_'.
- Dummy variables get the prefix 'X..OPTION…' by Flow, these are removed from the cleaned data set.
- A few free text options that have been found often in the data are set to similar text in order to make them comparable. An example is: 'don't know', 'doesn't know', 'I am not sure' are all changed to: 'I don't know'.
- In case the measurement of crop is supplied by farmers in a measurement unit other than Kilogram (e.g bags, boxes, crates, etc.), we have identified the value of the alternative measurement units in KG. The variable cal_focus_measurement_prod_kg captures a numeric value, indicating the number of kg that is in the measurement unit that is used (similar for measurement units used to report quantities sold, lost, or own consumption)
- A measurement of an area is generally reported by farmers in acres, kilometres squared or hectares. In the cleaned data the measurements are set to acres, which can be seen in the column heading (_acre).
- Some redundant columns are removed, for example columns with Flow details unimportant for the FarmFit analyses.

Determining and Handling Outliers

To determine outliers for the numerical questions of the survey, a cut off of three standard deviations from the corresponding mean is set. All values are compared to this cut off. When the value is either higher than

three standard deviations above the mean or lower than three standard deviations below the mean, it is set to '9997'.

Addressing Missing Values

The structure of the FarmFit survey prevents having actual missing values. All multiple-choice questions have the options 'I don't know' and 'I prefer not to say' and are mandatory. The numerical questions are also mandatory. Enumerators are instructed to answer them with '9999' in case a farmer doesn't know the answer, and '9998' when the farmer doesn't want to give the answer. This way all missing values are defined. In case of numerical questions, these values are not usable in aggregations and will give incorrect descriptive values. Therefore, all values containing '9999', '9998' and '9997', including those resulting from outlier handling, are set to 'NA'.

Anonymizing

In order to anonymize the data, farmer names, phone numbers, geolocation (longitude and latitude) and location except the highest administrative level (e.g. region or district) is removed from the set.

Repeated question groups

When recording the amount of crop produced, sold, lost or used for own consumption, we use 'repeated question groups'. This means farmers can provide input per season or for the whole year. In the cleaned data we only present one row of calculated values for each farmer. So if farmers reported production for 2 seasons, cal_focus_quant_prod_kg captures the total production during 2 seasons. For the amount produced, sold, lost and used for own consumption, we add the values of every season to get an idea of what happens throughout the year.

This process is applied for farmers that reported quantities produced, sold, lost, or consumed for multiple seasons; farmers that reported labour for multiple seasons; and farmers that reported input use and costs for multiple seasons.

Case specific adjustments

During data collection, we monitored incoming data and checked for outliers and inconsistencies using our data monitoring dashboard. Variables we check are:

- Measurement units
- Quantities produced, sold and lost (and consumed) of the focus crop
- Price received for the focus crop
- Farm size, including farm size of focus crop

Outliers were detected and adjusted in collaboration with the enumerators. In the majority of cases, the outliers were caused by typos.

In addition to this and the above described standard data cleaning steps, we have made case-specific adjustments to the data to ensure that data is reported in the correct way. We only assess the need for such adjustments for the above list variables, considering we have only limited time for data cleaning:

Quantity produced, sold, lost and/or consumed

Measurement units - Maize & Sorghum

Afrokai farmers producing Maize & Sorghum use Kgs, tonnes or bags as a measurement unit for their quantities produced. A bag in the county Kasese weights 100 kg, and a bag in the other counties weights 110-120 kg (set at 115 for calculations).

We produced the values for cal_*focus*_measurement_prod_kg, cal_*focus*_measurement_sold_kg, and cal_*focus*_measurement_lost_kg accordingly.

Measurement units - Potatoes

Afrokai farmers producing potatoes use KGs as a measurement unit.

Price/kg

Sorghum and maize

-

Potatoes

Some farmers reported the price for a bag of potatoes when asking for the price/kg. For those, we calculated the price per kg (cal_focus_price) by the weight of a bag of potatoes. The weight of a bag of potatoes is 100 KG in the sub-counties Kichwamba and Karambi, and 120 KG in the sub-county Kabende.

Farm size, including farm size of focus crop

No major adjustments were made, except for manual changes during data collection after checking odd numbers with the enumerators.

Other notes

- We were for now unable to add the descriptive results of the farmer agent questions to the sheets: “Cat. desc. multi all farmers” and “Cat. desc. multi by gender” in the data delivery file. We will aim to fix this problem but we still went ahead with delivering the data.
- Since we have 2 focus crops in the maize/sorghum survey, calculated variables can have different names. Each variable that starts usually with “cal_focus_..” now has 2 versions: “cal_maize_…” or

"cal_sorghum_..". There is 1 variable that captures net-income from focus crop: cal_maize_sorghum_income

- For more information on the calculation, please check the Excel where we explain the calculation attached to this delivery report. Please note that the actual income and farm income calculation is different for this case because we only asked the core questions related to farmer income.

Notes from data collection

Introduction

Our data collection lead took notes during data collection. These notes related to qualitative findings in the field, context-specific events the SDM team needs to be aware of, and feedback.

General Feedback

- Most farmers know Afrokai but they have not benefited. In Purongo subcounty Nwoya District, they complained that at one point they were told they would get seeds but waited in vain
- The farmers from Anaka sub- county said that it was their first time to hear about Afrokai but they are willing to work with Afrokai and they want if possible in this first season or season 2. They also confessed that they were selling at a very low price to the middle man.
- The major issue our supervisor encountered in the field was lack of awareness about Afrokai, then the pricing per kilogram vs. the expenses incurred
- Farmers from Lungulu sub county, Nwoya district also said it's their first time knowing about the company. An enumerator tried briefing the farming groups he met and they were positive to the extent that some farmers had to follow him the last day for the survey.
- Also another enumerator notes that they are willing to be partners with Afrokai. This enumerator also mentioned that in Lungulu sub county, it expects more than 100 tonnes of maize which is coming out in a month's time from now. The next crop that the company can target is soya bean.
- Lastly the farmers of Lungulu ask the supervisor to send their voices in the following issues:
    - One is about visiting them for introduction and explaining to the company's major aims formally
    - Secondly, if the company could make the evaluation a bit faster, that would be great, so that in a month they will be selling to Afrokai