Data delivery report

Grow pact

Date: 29-11-2022

Author: Ines Lesire & Mercy Makena

Focus crop(s): Tomato and cabbage

Country: Kenya

Introduction **2**

Survey design notes **2**

Survey questions on farmer satisfaction and protection 2

Survey questions on off-taker 3

Sample characteristics **3**

Theoretical sample size calculation 3

Sample size allocation in the field 3

Data cleaning steps **4**

Introduction 4

Removing farmers from the set 5

Text cleaning 5

Determining and handling outliers 5

Addressing missing values 5

Anonymising 6

Repeated question groups 6

Case specific adjustments 6

Measurement units 6

Quantity produced, sold, lost and/or consumed 7

Focus crop price 7

Farm size 7

Calculation of productivity 7

Notes from data collection **8**

Introduction

This report elaborates on the primary data collection (PDC) on farmers working with Growpact, cultivating tomato and cabbage in Embu, Trans-Nzoia, and Uasin Ngishu Counties in Kenya. The goal of the data collection is to assess the current livelihood of these tomato and cabbage smallholder farmers as a benchmark for measuring the positive influence of IDH and Growpact activities on their livelihoods and improve on the service delivery model for the company which in turns is expected to improve the smallholder farmer income.

Further information and context will be provided on the survey design, the actual sample characteristics, the data cleaning steps, and qualitative observations in the field during data collection. The purpose of this document is to handover contextual knowledge so that analysis can be done in the most optimal and efficient way.

The data collection for Growpact took place between 14-18 November 2022.

Survey design notes

The survey consists of the core questions and some new questions:

- Questions about farmer satisfaction and protection
- Questions on off-taker of focus crops

Further, optional questions on future outlook, household roster, and the poverty probability index were added.

The survey differs from other PDC surveys because the crops of focus were tomato and cabbage which are horticulture crops and they are not grown in seasons. The questions on labour and income were set using the harvest cycles instead of seasons.

Survey questions on farmer satisfaction and protection

Please check the questions with variable names starting with "*fsp_*" for the questions on satisfaction and protection of farmers that were added upon request by the SDM team, since this is a returning topic. The section contains a total of 28 questions that address the farmers' perception on the company's - service delivery, their responsible pricing, loans taken from the company, their transparency, fair and respectful treatment, privacy of the client data, and complaint resolution. This is the second case in which these questions are integrated. The plan is to incorporate them in the question library after reviewing their use with the Intelligence team.

Survey questions on off-taker

In the section on revenues from focus crop, additional questions -on whether or not all produced focus crop was sold to the off-taker (Growpact), and what happens with the produce that is not sold to them were added.

Sample characteristics

A sample size of 234 farmers was calculated using a population of 4700 farmers that Growpact shared with Akvo through the intake form. The sample size assumed a confidence level of 95% and a margin of error of 5%. We assumed a response rate of 80%. A total of 242 farmers were interviewed from Embu, Transzoia and Uasin Gishu.

Farmers were randomly sampled by use of the snowballing technique. Snowballing was identified as the appropriate method for identifying farmers due to the long distances between the selected farmers in the villages. We stratified the population using the location. The characteristics of the farmers samples depending on whether they were part of the original sample is as below;

Survey question: was the farmer part of the original sample?		
Response	**Nr. of farmers**	**Share of farmers**
No	25	10.33%
No, he/she is an alternative for a sampled farmer that was unavailable	21	8.68%
Yes	196	80.99%

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. For this case, none of the farmers refused to participate so no removals needed to be done.

Text cleaning

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

- Set the submission date variable to date format
- 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_tomato_measurement_prod_kg and cal_cabbage_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). However, for cabbage, another measurement unit of “per head” was introduced. More on this in the section about case specific adjustments.

- A measurement of an area is generally reported by farmers in acres, kilometres squared or hectares. In this case, the farm size was measured in acres, squared meters, and plots. In the cleaned data the measurements are set to acres, which can be seen in the column heading (_acre). This is explained in more detail in section “Case specific adjustments”.
- Some redundant columns with Flow details which are unimportant for the FarmFit analyses, are removed from the data.

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’, which means that the value is missing (see next section).

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’.

Anonymising

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) are removed from the set. The farmer can still be identified by the unique number in the “identifier” column.

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 cycles, cal_focus_quant_prod_kg captures the total production during 2 cycles. For the amount produced, sold, lost and used for own consumption, we add the values of every cycle 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 cycles; farmers that reported labour for multiple cycles; and farmers that reported input use and costs for multiple cycles.

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:

- Land measurement units
- Crop 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

Land measurement units

We added a land measurement unit “Squared meters” and “plot” in the survey because they are one of the main land measurement units in this region. During data cleaning, we calculated the farm size in acre, as usual.

- 1 squared meters = 0,000247105 acres
- 1 plot = 500 squared meters = 0.100 acres

Crop measurement units

When measuring cabbage produced, sold, lost or consumed, many farmers use “per head” as a measurement unit. It is not feasible to transform this into kgs, since crop sizes vary a lot. It is thus not possible to determine an average weight for a cabbage. The variables indicating quantity produced, sold, lost and consumed are twofold for cabbage, one “per head” and one in “kgs”. Also for productivity, both productivity_per_acre_kg and productivity_per_head are calculated, to indicate what measurement unit was used.

Quantity produced, sold, lost and/or consumed

Some data entries contained unrealistic numbers on quantity produced, sold, lost and consumed. For three farmers the quantity of produced tomato was put to NA. For other farmers, we discovered that enumerators filled in a total quantity in kgs already, instead of the number that should be multiplied with the kg in a crate. These errors were solved in the data cleaning script.

Focus crop price

Extreme outliers in tomato price were deleted from the dataset. For one farmer, the outlier in cabbage price was corrected to “50”, since this outlier is related to a typo of the enumerator.

Farm size

In some cases, enumerators made a mistake with entering one zero too much. These errors were corrected in the script. One farmer reported an unrealistic farm size, and was therefore removed from the set.

Calculation of productivity

As already mentioned, the respective focus crops of tomato and cabbage are planted continuously throughout the year in rotation. Above that, most farmers cultivate only one of the two focus crops (figure 2).

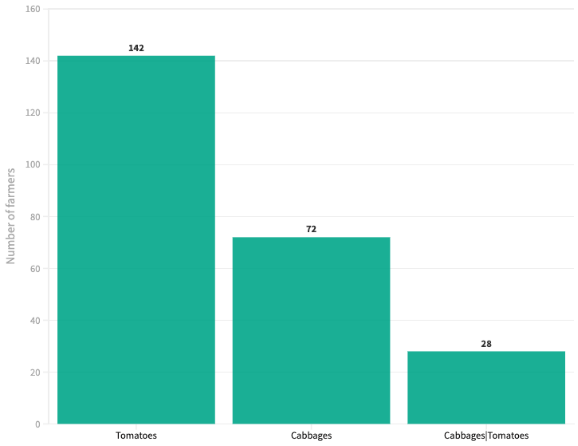

Figure 1: number of farmers cultivating each focus crop

Because of these reasons, the productivity for each focus crop is calculated separately. Also, for cabbage, a distinction is made between cabbages measured per head and cabbages measured in kgs. This leads to three productivity variables, calculated as follows:

$$cal\_tomato\_productivity\_acre \;=\; \frac{cal\_tomato\_quant\_prod\_kg}{f\_focus\_crop\_size\_acre}$$

$$cal\_cabbage\_productivity\_kg\_acre \;=\; \frac{cal\_cabbage\_quant\_prod\_kg}{f\_focus\_crop\_size\_acre}$$

$$cal\_cabbage\_productivity\_per\_head\_acre = \frac{cal\_cabbage\_quant\_prod\_per\_head}{f\_focus\_crop\_size\_acre}$$

Notes from data collection

Enumerator Selection and Training

An advert was circulated through our contacts in the region. Following receipt of many applications, qualified enumerators were selected. The selected enumerators had an understanding of the context in the concerning region and they could speak the local language. A one day training was conducted where the enumerators were taken through the use of Akvo flow for data collection and the survey.

Finding Samples farmers in the field

Finding farmers in the field was a challenge, since the farmers working with Grow pact are spread out over a large area and live very far from each other. Because the database received from the Grow pact was not complete, a method of referral was applied in the field, where farmers were asked to refer other farmers.

General feedback

- The database that was given to us by Grow pact contained only farmers who work with Grow pact regularly. Farmers only buying from Grow pact once or twice are not included in the database. We tried to include them in the sample by using the method of referral.