**Sucden Data delivery report**

**7-3-2022**

**Introduction**

This document contains details of the Sucden primary data collection and data cleaning steps. Please read the PDC process guide (SamePage: Primary Data Collection) before consulting this document.

**Number of farmers removed from the set**

Initially 399 farmers were targeted for this data collection.

Of the 399, 387 farmers participated.

**Sample characteristics:**

To get an accurate picture of the SDM farmers we take a random sample of the farmer group. However, we often run into issues with the farmer list supplied by the SDM 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 Sucden can be found below:

|  |  |
| --- | --- |
| no | 28 |
| no he/she is an alternative for a sample farmer that was unavailable | 90 |
| yes | 269 |

Table 1

|  |  |
| --- | --- |
| **Cooperative name** | |
|  | 18 |
| coopagnipi bas sassandra meagui | 86 |
| ecojad | 2 |
| ecokim | 1 |
| scapcca belier toumodi | 62 |
| scoopsyat la me yakasse attobrou | 76 |
| scops coopaths | 1 |
| socoofem marahoue bouafle | 85 |
| socopad | 27 |
| socopad haut sassandra daloa | 29 |

Table 2

#### **Introduction**

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

1. Removing Farmers from the Set
2. Text cleaning
3. Determining and handling outliers
4. Looking at missing values
5. Anonymizing
6. Repeated question groups
7. Case specific adjustments

#### **Removing Farmers from the Set**

At the moment farmers are only removed with they refuse to participate to the survey. The only data we have from these farmers is the name, location and sometimes a phone number.

#### **Text Cleaning**

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

1. All columns and text values are set to lower case
2. Flow sets spaces to points; we set them to ‘\_’.
3. Dummy variables get the prefix ‘X..OPTION…’ by Flow, these are removed from the cleaned data set.
4. 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: ‘dont know’, ‘doesn’t know’, ‘I am not sure’ are all changed to: ‘I don’t know’.
5. In case the measurement of crop is supplied by farmers in bags, boxes, crates, etc. In the survey the farmer is asked about the number of kilograms the applicable measurement contains. In the cleaned data the measurements are set to kilograms, which can be seen in the column heading (\_kg).
6. A measurement of an area is generally supplied by farmers in acres, kilometers squared or hectares. In the cleaned data the measurements are set to acres, which can be seen in the column heading (\_acre).
7. 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’.

#### **Looking at 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’, 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 from the 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 procide 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.

#### **Case specific adjustments**

We learned that farmers reported the price in one of the 2 following ways: approach 1) The price per bag of cocoa or approach 2) the price per kg of cocoa. The variable cal\_focus\_price captures the price per kg, which is either calculated in case of approach 1, or the price that was reported by the farmers (approach 2).From experts we learned that the price/kg ranges from 700-1000CFA/kg depending on the time of the year. We have not adjusted data of farmers that reported a price above or below this range considering agreements made with IDH, however this is a suggested step of data cleaning.

Likewise, farmers reported the quantities produced in different ways: approach 1) The quantity in metric tons, 2) the number of bags, 3) the total quantities in kg. The variable cal\_focus\_quant\_prod\_kg, cal\_focus\_quant\_sold\_kg, cal\_focus\_quant\_lost\_kg,report the total quantities. In case farmers applied approach 1 or 2, we conducted calculations to come to the total quantity produced. The variable “quantities\_interpretation” reflects how the quantities should be interpreted. This list was consulted with all enumerators.

In the majority of cases, farmers sell all that they produce.