Food and Agriculture Organization of the United Nations

FAO Statistics Working Paper Series

Issue 24-40

SAMPLING AND ESTIMATION GUIDE FOR SDG INDICATOR 2.4.1 UNDER MULTIFRAME DESIGNS

Second edition

FAO Statistics Working Paper Series / 24-40

SAMPLING AND ESTIMATION GUIDE FOR SDG INDICATOR 2.4.1 UNDER MULTIFRAME DESIGNS

Second edition

Cristiano Ferraz
Universidade Federal de Pernambuco (UFPE)

Food and Agriculture Organization of the United Nations
Rome, 2024

Required citation:
Ferraz, C. 2024. *Sampling and estimation guide for SDG Indicator 2.4.1 under multiframe designs*. FAO Statistics Working Paper Series, No. 24-40. Second Edition. Rome, FAO.
https://doi.org/10.4060/cc9550en

ISBN 978-92-5-138577-7

Abstract

This guide describes the estimation process for Sustainable Development Goal (SDG) Indicator 2.4.1 on agricultural sustainability, considering its subindicators, across different levels of disaggregation within a country. It complements the 2021 *Sampling guidance for SDG Indicator 2.4.1* published by the Food and Agriculture Organization of the United Nations (FAO). While the sampling guidance brings useful information concerning a general sampling and estimation perspective, this guide considers the use of dual-frame sampling designs in national agricultural surveys, which combine area frames with list frames. Although the focus is on dual-frame designs, the results described in this guide also apply to situations relevant to countries using agricultural surveys based on single area or list frames. Methodological documents and reports of practical experiences that reflect the realities of various countries, including Colombia, Costa Rica, Mexico and Peru, were used as the basis for developing the topics in this guide, as well as for providing examples of the fundamental concepts in this initial version.

Contents

Acknowledgements

The author would like to thank Regional Statistician Michael Rahija (FAO) for the invitation to develop this guide; Alethea Gabriela Candia (FAO), Gloria Vargas (FAO), and Xinia Andrade Ruiz (INEC) for proofreading and making suggestions to improve this guide. Any errors or inaccuracies are the sole responsibility of the author. The author also thanks the General Secretariat of the Latin American Faculty of Social Sciences (FLACSO) for the work carried out to enable the development of this guide under the Letter of Agreement signed with FAO in the "Support to national statistical systems for strengthening SDG indicators 2.4.1 and 12.3.1.a". Finally, thanks go to Tal Pinto (FAO) for editing the document.

1 Introduction

Created to monitor the progress made towards fulfilling Target 2.4 – ensure sustainable food production systems and implement resilient agricultural practices that increase productivity and production, that help maintain ecosystems, that strengthen capacity for adaptation to climate change, extreme weather, drought, flooding and other disasters, and that progressively improve land and soil quality – SDG Indicator 2.4.1 is the amount of agricultural land where productive and sustainable agriculture is practiced. To encompass economic, environmental and social dimensions in the definition of what constitutes productive and sustainable agriculture, Indicator 2.4.1 is defined as a summary function, typically the minimum value, of a series of 11 subindicators. Each subindicator corresponds to a proportion of productive and sustainable agricultural land according to a specific criterion. Therefore, Indicator 2.4.1 can be interpreted as the largest amount of sustainable national agricultural land, according to all 11 criteria. Although it is possible to analyse sustainability by changing the summary function and adapting its interpretation, this guide uses only the summary function of the minimum value as an example, as it represents a broad interpretation of sustainability, as emphasized by FAO.

To achieve proper estimates of all the parameters related to SDG Indicator 2.4.1, regardless of the level of disaggregation, it is first necessary to consider the specific set of conceptual definitions for each of them. For example, subindicator #1 reports the ratio of national agricultural land with a high agricultural production value per hectare. This can only be understood fully by having a definition for "agricultural production value per hectare", and for the adjective "high" in the expression "high production value". Various FAO publications[1] define and discuss these concepts in depth. For this reason, this guide only presents summarized versions of said definitions.

Finally, to ensure proper estimations of the parameters, these must be formalized as functions of measurable quantities based on a typical measurement process in agricultural surveys. Understanding the functional aspect of each parameter of interest, as well as understanding the sampling design used in a national survey, ensures statistically consistent estimators.

This guide is an initial effort to help obtain consistent estimators for each of the subindicators, and consequently, for SDG Indicator 2.4.1.

[1] See FAO (2023a, 2023b) for example.

2 SDG Indicator 2.4.1 as a population parameter

The descriptions of SDG Indicator 2.4.1 and its subindicators as population parameters emphasize the functional aspect of each of them. The amount of productive and sustainable agricultural land is defined as a ratio between the total areas of interest. The FAO methodological notes recommend reporting on SDG Indicator 2.4.1 as:

$$R_{241} = min(R_{\#1}, \dots, R_{\#11}), \qquad (1)$$

where each of the subindicators $R_{\#i}: i \in \{1, \dots, 11\}$ corresponds to the expression:

$$R_{\#i} = \frac{TSAS\#i}{TSA}, \qquad (2)$$

in which:

- $TSAS\#i$ represents the total agricultural land area within the country where agriculture is considered productive and sustainable according to the criterion $C_{\#i}$ used by the subindicator $R_{\#i}$.
- TSA represents the total agricultural land area.

So, R_{241} represents the amount of agricultural land where agriculture is considered productive and sustainable in the country, from a multidimensional perspective, in which all eleven criteria $C_{\#i}: i \in \{1, \dots, 11\}$ of productivity and sustainability, related to the subindicators, are met.

$R_{\#i}$ is considered a complex parameter because it consists of a ratio between totals. However, it is possible to generate a statistically consistent estimate of the subindicator by estimating each total separately. Therefore, it is recommended to use the estimator $\hat{R}_{\#i}$.

$$\hat{R}_{\#i} = \frac{\widehat{TSAS\#\imath}}{\widehat{TSA}}. \qquad (3)$$

The form of the estimators $\widehat{TSAS\#\imath}$ and $\widehat{TSA}$ depends on the sample design, which will be discussed later. Following the same principle, the recommended estimator for the indicator R_{241}, is:

$$\hat{R}_{241} = \acute{m}in\left(\hat{R}_{\#1}, \dots, \hat{R}_{\#11}\right). \qquad (4)$$

Although each subindicator has its own sustainability criterion $C_{\#i}$, the way in which each of them is reported follows a simple classification structure: “Green” (desirable), “Yellow” (acceptable) and “Red” (unsustainable). This opens the possibility of having a general estimator applicable to all subindicators. To advance in a common quantitative characterization, it is necessary to consider the measurement process of agricultural surveys.

Typically, agricultural surveys may use different types of sample units (such as area segments, producers and production units), depending on the type of frame used, either area or list. Sometimes, national surveys use both types of frames in a dual-frame design. However, in any case, the data collected to measure the sustainability criteria $C_{\#i}$ is obtained through a structured questionnaire interview, where the reference unit is the agricultural establishment. Therefore, the responses obtained are based on the agricultural land area of the agricultural establishment (agro), meaning the land area used for crop

cultivation or livestock rearing. This is the scope of the definition of Indicator 2.4.1, which allows establishing as the universe or target population the set of all agricultural establishments in a country. In this context, it is possible to define each subindicator as follows, using the following notation:

- $U = \{1, \dots, N\}$ represents the universe of N national agricultural establishments.
- y_k represents the agricultural land area of agricultural establishment $k \in U$.
- $V_{(i)k}$ represents an indicator variable that the agricultural establishment $k \in U$ is classified in the "green" sustainability category according to criterion $C_{\#i}$.

$$V_{(i)k} = \begin{cases} 1, \text{if the agricultural establishment } k \in U \text{ is "Green" by the criterion } C_{\#i}\ ; \\ 0, \text{if the agricultural establishment } k \in U \text{ is not "Green"}. \end{cases}$$

- $A_{(i)k}$ represents an indicator variable that the agricultural establishment $k \in U$ is classified in the "yellow" sustainability category according to criterion $C_{\#i}$:

$$A_{(i)k} = \begin{cases} 1, \text{if the agricultural establishment } k \in U \text{ is "yellow" by the criterion } C_{\#i}; \\ 0, \text{if the agricultural establishment } k \in U \text{ is not "yellow"}. \end{cases}$$

So, population parameters can be described as related to each subindicator based on the desired category. For example, based on expression (2), as

$$TSAS\#i = \sum_{k \in U} y_k V_{(i)k}, \qquad (5)$$

and

$$TSA = \sum_{k \in U} y_k, \qquad (6)$$

the expression

$$R_{\#i} = \frac{\sum_{k \in U} y_k V_{(i)k}}{\sum_{k \in U} y_k} \qquad (7)$$

represents the parametric form of subindicator $R_{\#i}$ in terms of the "Green" sustainability category. In other words, in this case $R_{\#i}$ represents the proportion of agricultural land with the **desirable level of sustainability** for criterion $C_{\#i}$. In parallel, the population parameters

$$R_{\#i}1 = \frac{\sum_{k \in U} y_k A_{(i)k}}{\sum_{k \in U} y_k}, \qquad (8)$$

$$R_{\#i}2 = \frac{\sum_{k \in U} y_k (V_{(i)k} + A_{(i)k})}{\sum_{k \in U} y_k}, \qquad (9)$$

and

$$R_{\#i}3 = \frac{\sum_{k \in U} y_k (1 - V_{(i)k} A_{(i)k})}{\sum_{k \in U} y_k} \qquad (10)$$

represent the proportion of agricultural land for the same subindicator using different sustainability categories, where:

- $R_{\#i}1$ is the proportion of agricultural land with an **acceptable level of sustainability** (yellow) for criterion $C_{\#i}$.
- $R_{\#i}2$ s the proportion of agricultural land with an **acceptable or desirable level of sustainability** (yellow or green*)* for criterion $C_{\#i}$.
- $R_{\#i}3$ is the proportion of agricultural land with an **unsustainable level** (red) for criterion $\mathrm{C}_{\#\mathrm{i}}$.

Without losing the more general aspect of the problem, in this guide, the estimation examples consider the subindicators defined based on the "Green" sustainability category, as in (7), as well as the others, $R_{\#i}1$, $R_{\#i}2$ and $R_{\#i}3$, defined in a general way in (8), (9) and (10), respectively.

3 Sustainability criteria in each subindicator

Understanding the sustainability criteria for each subindicator is essential for the proper estimation of Indicator 2.4.1. Peru and Costa Rica have experience in using the concepts related to each criterion and in how the information was collected through questionnaires, considering the classification of each subindicator by sustainability category. This chapter highlights the more general aspects of the criteria, grouped by dimension. Table 1 presents a summary of the topics considered for each subindicator.

Table 1. Subindicators by dimension

Subindicators	Criteria	Dimension
1	Farm output value per hectare	Economic
2	Net farm income	Economic
3	Risk mitigation mechanisms	Economic
4	Prevalence of soil degradation	Environmental
5	Variation in water availability	Environmental
6	Management of fertilizers	Environmental
7	Management of pesticides	Environmental
8	Use of agrobiodiversity-supportive practices	Environmental
9	Wage rate in agriculture	Social
10	Food insecurity experience scale (FIES)	Social
11	Secure tenure rights to land	Social

Source: Author's own elaboration.

3.1 Subindicators of the economic dimension

The three economic subindicators are $R_{\#1}$, $R_{\#2}$ and $R_{\#3}$. The sustainability criteria are $C_{\#1}$, $C_{\#2}$ and $C_{\#3}$, respectively:

- $\boldsymbol{C_{\#1}}$: Compares the value of VPA_k, the **agricultural production per hectare** of the agricultural establishment $k \in U$, with a percentile 90 ($P90$) of this same amount, resulting in:
  - Green, if $VPA_k \geq \frac{2}{3}P90$.
  - Yellow, if $\frac{1}{3}P90 \leq VPA_k < \frac{2}{3}P90$.
  - Red, if $VPA_k < \frac{1}{3}P90$.
- $\boldsymbol{C_{\#2}}$: Compares the values of IAN_k, the **net agricultural income** of the agricultural establishment $k \in U$, for the last three years, resulting in:
  - Green, if $IAN_k \geq 0$ for the last three years.

- Yellow, if $VPA_k > 0$ for at least one of the last three years.
- Red, if $VPA_k < 0$ for the last three years.

- $\boldsymbol{C_{\#3}}$: Identifies if an agricultural establishment $k \in U$ uses **risk mitigation mechanisms**, that is, has access to credit, insurance and diversifies the agricultural establishment, resulting in:
  - Green, if the agricultural establishment has had access to or used at least two of the mechanisms.
  - Yellow, if the agricultural establishment has had access to or used only one of the mechanisms.
  - Red, the agricultural establishment has had no access to any of the three mechanisms.

3.2 Subindicators of the environmental dimension

There are five environmental subindicators: $R_{\#4}$, $R_{\#5}$, $R_{\#6}$, $R_{\#7}$ and $R_{\#8}$. The sustainability criteria are $C_{\#4}$, $C_{\#5}$, $C_{\#6}$, $C_{\#7}$ and $C_{\#8}$, respectively:

- $\boldsymbol{C_{\#4}}$: Compares the **level of soil degradation** from erosion, reduced fertility, salinization, and waterlogging. It is measured as PAD_k, the amount of agricultural land area $k \in U$ affected by degradation, resulting in:
  - Green, if $PAD_k \leq 10\%$.
  - Yellow, if $10\% < PAD_k \leq 50\%$.
  - Red, if $PDA_k > 50\%$.

- $\boldsymbol{C_{\#5}}$: Identifies if there is **variation in water availability** over the years, resulting in:
  - Green, if the availability of water is stable over the years or does not use water to irrigate crops in more than 10 percent of the agricultural area of the farm.
  - Yellow, when it uses water to irrigate crops in at least 10 percent of the agricultural area of the farm, it is not known if water availability stays the same or drops, but an organization responsible for the effective distribution of water among users is in operation.
  - Red, in all other cases.

- $\boldsymbol{C_{\#6}}$: Identifies if an agricultural establishment $k \in U$ takes measures (out of a total of eight) to mitigate **the risk of contamination by fertilizers**.
  - Green, if the agricultural establishment $k \in U$ implements at least four of the measures considered.
  - Yellow, if the agricultural establishment $k \in U$ implements two or three of the measures considered.
  - Red, in all other cases.

- $\boldsymbol{C_{\#7}}$: Identifies if an agriculture establishment $k \in U$ takes measures to mitigate the **risk of pesticide contamination**. In addition to implementing mitigation measures considered in a list, the type of pesticide used is also considered, which results in:
    - Green, when the agricultural establishment $k \in U$ only uses moderately hazardous or low-hazardous pesticides (World Health Organization [WHO] classes II and III), complies with three considered health-related measures and at least four of the considered environment-related measures.
    - Yellow, when the agriculture establishment $k \in U$ only uses moderately hazardous or low-hazardous pesticides (WHO classes II and III) and takes at least two measures from each group to mitigate risks to the environment and to health.
    - Red, when the agricultural establishment $k \in U$ uses extremely dangerous, very dangerous (WHO classes Ia and Ib) or illegal pesticides, or uses moderately hazardous or slightly hazardous pesticides without taking specific measures to mitigate the environmental and health risks associated with the use of such pesticides (less than two measures from any of the lists considered).
- $\boldsymbol{C_{\#8}}$: Identifies if an agricultural establishment $k \in U$ follows **practices that support agricultural biodiversity** (from a total of six), resulting in:
    - Green, when the agricultural establishment $k \in U$ implements at least three of the six practices.
    - Yellow, when the agricultural establishment $k \in U$ implements one or two of the practices.
    - Red, when the agricultural establishment $k \in U$ does not implement any of the practices.

3.3 Subindicators of the social dimension

There are three social subindicators: $R_{\#9}$, $R_{\#10}$ and $R_{\#11}$. The sustainability criteria are $C_{\#9}$, $C_{\#10}$ and $C_{\#11}$, respectively:

- $\boldsymbol{C_{\#9}}$: Related to **agricultural wages,** it identifies whether the wage paid, on average, to unskilled agricultural workers by an agricultural establishment $k \in U$ is equal, higher, or lower than the country's minimum wage, resulting in:
    - Green, when the agricultural establishment $k \in U$ pays on average higher wages than the minimum wage.
    - Yellow, when the agricultural establishment $k \in U$ pays on average wages equal to the minimum wage.
    - Red, when the agricultural establishment $k \in U$ pays on average lower wages than the minimum wage.

- $\boldsymbol{C_{\#10}}$: It is defined based on the **Food Insecurity Experience Scale** (FIES) and identifies whether the degree of food insecurity is mild, moderate, or severe, from the probability of each of them, resulting in:
    - Green, when the probability of food insecurity is both less than 0.5 for moderate or severe cases, as well as for mild ones.
    - Yellow, when the probability of food insecurity is greater than 0.5 for moderate or severe cases, and the probability that it is severe is less than 0.5.
    - Red, when the probability that it is severe is greater than 0.5.
- $\boldsymbol{C_{\#11}}$: It is defined in terms of **security of land tenure rights**. It measures if the owner of the agricultural establishment $k \in U$ has an official land tenure document, resulting in:
    - Green, when the owner of the agricultural establishment has an official document with the name of the owner of the agricultural establishment or the right to sell or inherit any of the plots of the establishment.
    - Yellow, when the owner of the agricultural establishment has an official document, even if the name of the owner or the agricultural establishment does not appear in said document.
    - Red, in all other cases.

4 Sampling in agricultural surveys

Although in theory a national survey to address only the Sustainable Development Goals could be designed, each country has its own data collection needs. Hence, it is more efficient to integrate the sustainability indicator estimation goals with the specific objectives of existing national agricultural sample surveys or to develop an integrated agricultural survey system that addresses all these objectives at the same time. The AGRIS programme (FAO, 2018) is an example of an integrated system that can be adopted as a reference for the development of national systems considering the reality of each country.

Developing a document considering each of the sample designs of agricultural surveys in use by countries would not be reasonable. However, it is possible to identify the most used sample design elements and use them in an intuitive description so that, on the one hand, the adequacy of the sampling method to the estimation problem of Indicator 2.4.1. is perceived, and on the other, there is a benchmark for countries that need to develop or renew a method for their probability sample surveys.

Regardless of the type of frame in use, three elements are commonly used for estimating efficiency gains:

- stratification;
- multistage sampling; and
- probability proportional to size measure (PPT) sampling.

4.1 Stratification factors

Usually, a target population is stratified according to at least one of two groups of factors. The first group is comprised by the factors that define subnational domains of interest for estimation; and the second by factors that add important auxiliary information to generate precision gain. So, a stratum h represents one of the levels of a stratification factor, or a combination of factor levels of the two categories.

If each stratum is a combination of factor levels, we consider a country with a recent agricultural census, in which it is possible to identify an area frame based on terrains with known physical boundaries. We envision that the target population of all agricultural establishments in the country is completely covered by such an area frame. In addition, in some cases a univocal relationship between parcels and each agricultural producer is identified, although each producer may have more than one land, as in Mexico, in its 2022 agricultural census.

Developing a national agricultural survey for the country could benefit from a stratification structure that uses geopolitical factors to ensure subnational estimates, as well as factors related to land use intensity, measured using satellite imagery, to add precision to the estimates. Singling out the geopolitical factor of interest, such as country states in Mexico, results in 32 levels (states). If the land use intensity factor has, for example, three levels (high, moderate and low intensity), the stratification process would use $32 \times 3 = 96$ strata.

4.2 Cluster sampling

The use of clusters in sample designs can be a cheaper data collection process. By definition, a cluster is a set of elements of a target population. In agricultural surveys, where the target population is all farms, a cluster can take different forms, depending on the type of sample frame. In area frame sampling, for

example, it is common for a conglomerate to take the form of a larger area where it is possible to identify a set of agricultural establishments. Sometimes, the large number of farms in the cluster can justifies dividing the cluster into smaller sub-areas, each containing a smaller set of farms. On the other hand, in list sampling frames, clusters can have other definitions. A frame listing all producer associations grouped by village, for example, is a frame in which each village represents a set of producers, with a known relationship to the concept of farms in a country.

Costa Rica's National Agricultural Survey (ENA) subdivides its area frame into several smaller areas, called area segments, in which a set of farms (agricultural establishments) can be identified. Each area segment is a cluster that is selected by probability sampling. A screening of all area segments contained in the selected clusters is performed, and all identified farms are investigated. It would still be possible to add a selection stage to the sampling process, in which farms would be selected from within each cluster, thus becoming a two-stage (or multistage) sampling.

4.3 Weighted probability sampling

When sampling frames have auxiliary information available, it is possible to use weighted probability sampling using values proportional to a measure of importance, commonly called a size measure. In this way, the probabilities are calculated to include this data, improving precision, as long as the auxiliary information has some degree of correlation with the variable of interest.

5 Estimating Indicator 2.4.1 for a national territory

FAO recommends estimating Indicator 2.4.1 for the entire national territory, generating, when possible, disaggregated estimates for subnational divisions, as it provides useful information for sustainability efforts in accordance with the 2030 Agenda.

In this chapter, the problem of generating indicator estimates for a national territory is addressed by considering the sample design elements introduced above. In addition to stratification and varying probability sampling, two design situations are considered: one with two-stage sampling and the other with one-stage sampling.

For notation purposes, we imagine that H strata are identified, and that the elements of analysis (the agricultural establishments) coincide with the last level sampling unit of each stratum, or are included in them, when the last level is a conglomerate. The following notation is introduced:

- $U = \{1, \dots, N\}$ represents the universe of N national agricultural establishments, identified in a population structure in which:
    - $U = \bigcup_{h=1}^{H} U_h$, that is, the population is stratified into H strata, with

$U_h = \{1, \dots, M_h\}$ represents the set of M_h clusters from stratum h, such that $M = \sum_{h=1}^{H} M_h$ is the total number of clusters in the population.

    - $U_{hi} = \{1, \dots, N_{hi}\}$ represents the set of N_{hi} agricultural establishments of the cluster i in the stratum h, such that $N_h = \sum_{i=1}^{M_h} N_{hi}$ is the total number of agricultural establishments in the stratum h.
    - $N = \sum_{h=1}^{H} N_h$.
- y_{hij} stands for the agricultural area of the agricultural establishment j of the cluster i, in the stratum h.

For example, in Costa Rica's ENA the unit of analysis is the farm. The survey uses a dual sampling frame, with an area frame and a list of large producers to cover the entire population. On the one hand, a census is applied to the list frame. On the other hand, the area frame of the ENA is divided into large areas for the purpose of stratification by intensity of agricultural land use (five strata in total). In each stratum, a sample of clusters (i.e. segments) from smaller areas is selected and all farms with plots that fall within these clusters are interviewed. In this case, y_{hij} represents the agricultural area of the farm (agricultural establishment) j of the cluster (minor areas) i, in the stratum h.

For convenience, in this guide agricultural establishments will sometimes be referred to simply as "hij", instead of the agricultural establishment j of cluster i, in the stratum h.

The notation of the variables previously defined as "$V_{(i)k}$" and "$A_{(i)k}$" varies slightly, changing the index i by c, written now as:

- $V_{(c)hij}$, which represents an indicator variable that the agricultural establishment j of cluster i in the stratum h is classified in the "Green" category of sustainability, according to the criteria $c \in \{C_{\#1}, \dots, C_{\#11}\}$:

- $V_{(c)hij} =$
  $\begin{cases} 1, \text{if the agricultural establishment "}hij\text{" is "Green" according to criterion } c\ ; \\ 0, \text{if the agricultural establishment "}hij\text{" is not "Green";} \end{cases}$

and

- $A_{(c)hij}$ represents an indicator variable that the agricultural establishment j of the cluster i in the stratum h is classified in the "Yellow" category of sustainability, according to the criteria $c \in \{C_{\#1}, \dots, C_{\#11}\}$:

  - $A_{(c)hij} =$
    $\begin{cases} 1, \text{if the agricultural establishment "}hij\text{" is "Yellow" according to criterion } c; \\ 0, \text{ if the agricultural establishment "}hij\text{" is not "Yellow".} \end{cases}$

So, the population parameters (7), (8), (9) and (10), relative to each subindicator, can be written as a ratio between:

$$TSAS\#c = \sum_{h=1}^{H} \sum_{i \in U_h} \sum_{j \in U_{hi}} y_{hij} V_{(c)hij}, \qquad (11)$$

and

$$TSA = \sum_{h=1}^{H} \sum_{i \in U_h} \sum_{j \in U_{hi}} y_{hij}, \qquad (12)$$

as follows:

$$R_{\#c} = \frac{\sum_{h=1}^{H} \sum_{i \in U_h} \sum_{j \in U_{hi}} y_{hij} V_{(c)hij}}{\sum_{h=1}^{H} \sum_{i \in U_h} \sum_{j \in U_{hi}} y_{hij}}, \qquad (13)$$

$$R_{\#c}1 = \frac{\sum_{h=1}^{H} \sum_{i \in U_h} \sum_{j \in U_{hi}} y_{hij} A_{(c)hij}}{\sum_{h=1}^{H} \sum_{i \in U_h} \sum_{j \in U_{hi}} y_{hij}}, \qquad (14)$$

$$R_{\#c}2 = \frac{\sum_{h=1}^{H} \sum_{i \in U_h} \sum_{j \in U_{hi}} y_{hij} \left(V_{(c)hij} + A_{(c)hij}\right)}{\sum_{h=1}^{H} \sum_{i \in U_h} \sum_{j \in U_{hi}} y_{hij}}, \qquad (15)$$

$$R_{\#c}3 = \frac{\sum_{h=1}^{H} \sum_{i \in U_h} \sum_{j \in U_{hi}} y_{hij} (1 - V_{(c)hij} A_{(c)hij})}{\sum_{h=1}^{H} \sum_{i \in U_h} \sum_{j \in Uh_i} y_{hij}}. \qquad (16)$$

Two-stage estimation

To generate consistent national estimates from the point of view of the sample design for parameters (13), (14), (15) and (16), it is necessary to consider the randomization process used in the design. With a sample selection in two stages in which, in each stratum, a sample S_h of clusters is taken, and in each cluster a sample S_{hi} of establishments is taken, such that π_{hij} is the probability of selection of the establishment j of cluster i in the stratum h, it is possible to write $\pi_{hij} = \pi_{hi}\pi_{j|hi}$, where π_{hi} is the selection probability of the cluster i in the stratum h and $\pi_{j|hi}$ is the probability of selection of the production unit j given that the segment i to which it belongs, within the stratum h, was selected for the

sample. As a result, in a two-stage estimation, the estimators of the indicators of interest assume the following formulas:

$$\hat{R}_{\#c} = \frac{\sum_{h=1}^{H} \sum_{i \in S_h} \sum_{j \in S_{hi}} \frac{y_{hij} V_{(c)hij}}{\pi_{hi} \pi_{j|hi}}}{\sum_{h=1}^{H} \sum_{i \in S_h} \sum_{j \in S_{hi}} \frac{y_{hij}}{\pi_{hi} \pi_{j|hi}}}, \qquad (17)$$

$$\hat{R}_{\#c}1 = \frac{\sum_{h=1}^{H} \sum_{i \in S_h} \sum_{j \in S_{hi}} \frac{y_{hij} A_{(c)hij}}{\pi_{hi} \pi_{j|hi}}}{\sum_{h=1}^{H} \sum_{i \in S_h} \sum_{j \in S_{hi}} \frac{y_{hij}}{\pi_{hi} \pi_{j|hi}}}, \qquad (18)$$

$$\hat{R}_{\#c}2 = \frac{\sum_{h=1}^{H} \sum_{i \in S_h} \sum_{j \in S_{hi}} \frac{y_{hij} (V_{(c)hij} + A_{(c)hij})}{\pi_{hi} \pi_{j|hi}}}{\sum_{h=1}^{H} \sum_{i \in S_h} \sum_{j \in S_{hi}} \frac{y_{hij}}{\pi_{hi} \pi_{j|hi}}}, \qquad (19)$$

and

$$\hat{R}_{\#c}3 = \frac{\sum_{h=1}^{H} \sum_{i \in S_h} \sum_{j \in S_{hi}} \frac{y_{hij} (1 - V_{(c)hij} A_{(c)hij})}{\pi_{hi} \pi_{j|hi}}}{\sum_{h=1}^{H} \sum_{i \in S_h} \sum_{j \in S_{hi}} \frac{y_{hij}}{\pi_{hi} \pi_{j|hi}}}. \qquad (20)$$

One-stage estimation

In a sample selection in one stage in which, in each stratum, a sample S_h of clusters is taken, and in each cluster a screening of all agricultural establishments is made, such that π_{hi} is the selection probability of cluster i in the stratum h, estimators (17), (18), (19) and (20) need adjusting, resulting respectively in:

$$\hat{R}_{\#c} = \frac{\sum_{h=1}^{H} \sum_{i \in S_h} \sum_{j \in U_{hi}} \frac{y_{hij} V_{(c)hij}}{\pi_{hi}}}{\sum_{h=1}^{H} \sum_{i \in S_h} \sum_{j \in U_{hi}} \frac{y_{hij}}{\pi_{hi}}}, \qquad (21)$$

$$\hat{R}_{\#c}1 = \frac{\sum_{h=1}^{H} \sum_{i \in S_h} \sum_{j \in U_{hi}} \frac{y_{hij} A_{(c)hij}}{\pi_{hi}}}{\sum_{h=1}^{H} \sum_{i \in S_h} \sum_{j \in U_{hi}} \frac{y_{hij}}{\pi_{hi}}}, \qquad (22)$$

$$\hat{R}_{\#c}2 = \frac{\sum_{h=1}^{H} \sum_{i \in S_h} \sum_{j \in U_{hi}} \frac{y_{hij} (V_{(c)hij} + A_{(c)hij})}{\pi_{hi}}}{\sum_{h=1}^{H} \sum_{i \in S_h} \sum_{j \in U_{hi}} \frac{y_{hij}}{\pi_{hi}}}, \qquad (23)$$

and

$$\hat{R}_{\#c}3 = \frac{\sum_{h=1}^{H} \sum_{i \in S_h} \sum_{j \in U_{hi}} \frac{y_{hij} (1 - V_{(c)hij} A_{(c)hij})}{\pi_{hi}}}{\sum_{h=1}^{H} \sum_{i \in S_h} \sum_{j \in U_{hi}} \frac{y_{hij}}{\pi_{hi}}}. \qquad (24)$$

The estimators introduced so far are in general enough to be used in practice. However, it is possible to make adaptations according to the type of frame used. In this guide, adaptations are introduced for the most common agricultural survey designs using only one sampling frame, either an area frame or a list frame, and a dual frame system, with the simultaneous use area and list frame samplings.

5.1 Using an area frame

Area frames use the national territory as a reference to cover the target population of a survey. In this guide, area frames that use area segments as sampling units, and production units as observation and analysis units are considered as the standard.

The concept of area segments used in this guide is broad, i.e. an area segment is a parcel of land that has a regular or irregular shape. Square segments are examples of segments of regular shape, while segments formed from physical boundaries are examples of irregular segments. The area frame used for Colombia's ENA, prepared by the country's Department of National and Administrative Statistics (DANE) is a frame in which primary sampling units form a partition of the country's national territory through irregular area segments delimited by physical, natural, or cultural accidents, identified by satellite images.

The concept of production units is also general enough to represent various cases. For example, in Costa Rica's national agricultural survey the farm is the observation and analysis unit. This is the ideal situation to collect the amount of information necessary to generate estimates of the subindicators of interest for the definition of the sustainability Indicator 2.4.1.

Typically, area frame sampling designs applied to an agricultural survey are stratified by the agricultural land use intensity criterion and use area segments as primary sampling units in each stratum. Two widely used area segment sampling strategies for selecting production units are the open segment selection strategy and the weighted segment selection strategy.

5.1.1 Using the open segment strategy

Taking a sample of production units using the open segment selection strategy is equivalent to having a questionnaire applied for each production unit (agricultural establishment) that has a "headquarters" located within the boundaries of the segment. The answers to the questions are related to the entire production unit, even though its delineated area extrapolates the boundaries of the segment. The definition of "headquarters" needs to be a clear criterion that allows an association of each production unit to only one area segment, so that the probability of selection of a production unit is the same as the area segment in which it is located.

Using the open segment strategy, the estimators of Indicator 2.4.1 for the entire national territory coincide with those defined above for a one-stage sampling. So, in sample designs with equal probability of segment selection (i.e. simple or systematic random sampling) within each stratum, the probability of selection is $\pi_{hi} = \pi_h = m_h/M_h$, where m_h is the number of segments selected for the stratum sample h, and M_h is the total number of area segments in the stratum h of the area frame. Under these conditions, the estimators (21), (22), (23) and (24) take the following form, respectively:

$$\hat{R}_{\#c} = \frac{\sum_{h=1}^{H} \sum_{i \in S_h} \sum_{j \in U_{hi}} \frac{M_h y_{hij} V_{(c)hij}}{m_h}}{\sum_{h=1}^{H} \sum_{i \in S_h} \sum_{j \in U_{hi}} \frac{M_h y_{hij}}{m_h}}, \qquad (25)$$

$$\hat{R}_{\#c}1 = \frac{\sum_{h=1}^{H}\sum_{i\in S_h}\sum_{j\in U_{hi}}\frac{M_h y_{hij}A_{(c)hij}}{m_h}}{\sum_{h=1}^{H}\sum_{i\in S_h}\sum_{j\in U_{hi}}\frac{M_h y_{hij}}{m_h}}, \qquad (26)$$

$$\hat{R}_{\#c}2 = \frac{\sum_{h=1}^{H}\sum_{i\in S_h}\sum_{j\in U_{hi}}\frac{M_h y_{hij}\left(V_{(c)hij}+A_{(c)hij}\right)}{m_h}}{\sum_{h=1}^{H}\sum_{i\in S_h}\sum_{j\in U_{hi}}\frac{M_h y_{hij}}{m_h}}, \qquad (27)$$

and

$$\hat{R}_{\#c}3 = \frac{\sum_{h=1}^{H}\sum_{i\in S_h}\sum_{j\in U_{hi}}\frac{M_h y_{hij}(1-V_{(c)hij}A_{(c)hij})}{m_h}}{\sum_{h=1}^{H}\sum_{i\in S_h}\sum_{j\in U_{hi}}\frac{M_h y_{hij}}{m_h}}. \qquad (28)$$

5.1.2 Using the weighted segment strategy (in one stage)

Taking a sample of production units using the weighted segment strategy is equivalent to having a questionnaire applied to each production unit that has a total or partial intersection of its parcels of agricultural areas with the selected segment. The area formed by the intersection of one or more parcels of agricultural areas with the segment, under the same production unit, is here called “tract”.

We consider T_{hij} as the area of the tract of the production unit j in the segment i within the stratum h, and A_{hij} as the total area of the production unit j , chosen through segment i, within the stratum h. Using the weighted segment strategy, the variables of interest y_{hij} receive a weighting proportional to the area of the tract of the selected production unit, so that the observed weighted variable for estimation purposes is x_{hij} such that:

$$x_{hij} = \frac{T_{hij}}{A_{hij}} y_{hij}.$$

The estimation formulas for indicator 2.4.1 with this sampling strategy coincide with the same estimation formulas in one stage, changing y_{hij} by x_{hij}. Considering sample designs with equal probabilities for selection of segments within each stratum, derive $\pi_{hi} = \pi_h = m_h/M_h$, where m_h is the number of segments selected for the stratum sample h, and M_h is the total number of area segments in the stratum h of the area frame. Under these conditions, the estimators (21), (22), (23) and (24) have the following form, respectively:

$$\hat{R}_{\#c} = \frac{\sum_{h=1}^{H}\sum_{i\in S_h}\sum_{j\in U_{hi}}\frac{M_h x_{hij}V_{(c)hij}}{m_h}}{\sum_{h=1}^{H}\sum_{i\in S_h}\sum_{j\in U_{hi}}\frac{M_h x_{hij}}{m_h}}, \qquad (29)$$

$$\hat{R}_{\#c}1 = \frac{\sum_{h=1}^{H}\sum_{i\in S_h}\sum_{j\in U_{hi}}\frac{M_h x_{hij}A_{(c)hij}}{m_h}}{\sum_{h=1}^{H}\sum_{i\in S_h}\sum_{j\in U_{hi}}\frac{M_h x_{hij}}{m_h}}, \qquad (30)$$

$$\hat{R}_{\#c}2 = \frac{\sum_{h=1}^{H}\sum_{i\in S_h}\sum_{j\in U_{hi}}\frac{M_h x_{hij}\left(V_{(c)hij}+A_{(c)hij}\right)}{m_h}}{\sum_{h=1}^{H}\sum_{i\in S_h}\sum_{j\in U_{hi}}\frac{M_h x_{hij}}{m_h}}, \qquad (31)$$

and

$$\hat{R}_{\#c}3 = \frac{\sum_{h=1}^{H}\sum_{i\in S_h}\sum_{j\in U_{hi}}\frac{M_h x_{hij}(1-V_{(c)hij}A_{(c)hij})}{m_h}}{\sum_{h=1}^{H}\sum_{i\in S_h}\sum_{j\in U_{hi}}\frac{M_h x_{hij}}{m_h}}. \qquad (32)$$

One disadvantage of this strategy is the need to know the value of T_{hij} to calculate x_{hij}.

5.1.3 Using the two-stage weighted segment strategy

Sampling production units by this strategy is equivalent to selecting a subsample of production units within each segment through sampling with replacement and selection probability proportional to the size of the tract, simulated by the use of points.

To arrive at the formula of the estimators with this sampling strategy it is necessary to consider the same auxiliary variable x_{hij} previously introduced,

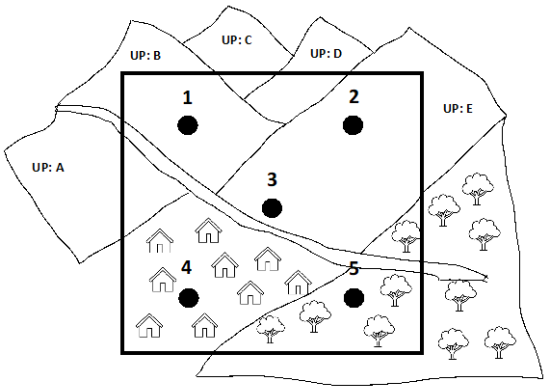
$$x_{hij} = \frac{T_{hij}}{A_{hij}} y_{hij}, \qquad (33)$$

as well as using the general expressions (17), (18), (19) and (20). To illustrate this reasoning, we consider just the expression (17), changing y_{hij} by x_{hij}:

$$\hat{R}_{\#c} = \frac{\sum_{h=1}^{H}\sum_{i\in S_h}\sum_{j\in S_{hi}}\frac{x_{hij}V_{(c)hij}}{\pi_{hi}\pi_{j|hi}}}{\sum_{h=1}^{H}\sum_{i\in S_h}\sum_{j\in S_{hi}}\frac{x_{hij}}{\pi_{hi}\pi_{j|hi}}}. \qquad (34)$$

The next step is to adapt the formula (34) to accommodate the point subsampling process. We consider, as an example, an area segment with P points. Figure 1 shows a square area segment with $P = 5$ points. It is possible to identify, in Figure 1, five production units (UPs): A, B, C, D and E. It is also possible to identify the five points: 1, 2, 3, 4 and 5. No point remains in the tracts of production units A, C and D, so these UPs are not part of the subsample of the segment illustrated in Figure 1. Point 1 was located at a tract of the UP B, so the UP B is part of the subsample and a questionnaire is applied to this production unit. Points 2 and 3 will be in the same UP E and therefore, a questionnaire is also applied to this UP. Point 4 is in an urban area and point 5 stays in a forest area.

Figure 1. Square segment with five points



Source: Author's own elaboration.

We consider that a sample of m_h segments of area is taken out of the stratum h so that each of the M_h segments have equal probability of selection, and also that a subsample of production units is selected by points, as illustrated in Figure 1. That way,

$$\pi_{hi} = \frac{m_h}{M_h}, \qquad (35)$$

and

$$\pi_{j|hi} = \frac{T_{hij}}{T_{hi}} P, \qquad (36)$$

where T_{hi} represents the area surface of the segment i in the stratum h, P is the total points used in each segment, and T_{hij} is the area of the tract of the production unit j located in the segment i, within the stratum h. Using the values of x_{hij}, π_{hi} and $\pi_{j|hi}$ from expressions (33), (35) and (36), it can be seen that:

$$\frac{x_{hij}}{\pi_{hi}\pi_{j|hi}} = \frac{M_h}{m_h}\frac{T_{hi}}{T_{hij}P}\frac{T_{hij}}{A_{hij}}y_{hij} = \frac{M_h}{m_h P}\frac{T_{hi}}{A_{hij}}y_{hij}.$$

so

$$I_{hijk} = \begin{cases} 1, \text{if the point k was located in the "tract" of the UP } hij; and \\ 0, \text{otherwise,} \end{cases}$$

it is possible to rewrite the formula (34) as follows:

$$\hat{R}_{\#c} = \frac{\sum_{h=1}^{H}\sum_{i\in S_h}\sum_{j\in U_{hi}}\sum_{k=1}^{P}\frac{M_h T_{hi}}{m_h\ P\ A_{hij}}y_{hij}V_{(c)hij}I_{hijk}}{\sum_{h=1}^{H}\sum_{i\in S_h}\sum_{j\in U_{hi}}\sum_{k=1}^{P}\frac{M_h T_{hi}}{m_h\ P\ A_{hij}}y_{hij}I_{hijk}}. \qquad (37)$$

In the expression (37), U_{hi} represents the set of all agricultural establishments that have a tract with the segment i within the stratum h.

Using the same reasoning, the following estimators can be derived:

$$\hat{R}_{\#c}1 = \frac{\sum_{h=1}^{H}\sum_{i\in S_h}\sum_{j\in U_{hi}}\sum_{k=1}^{P}\frac{M_h T_{hi}}{m_h\ P\ A_{hij}} y_{hij} A_{(c)hij} I_{hijk}}{\sum_{h=1}^{H}\sum_{i\in S_h}\sum_{j\in U_{hi}}\sum_{k=1}^{P}\frac{M_h T_{hi}}{m_h\ P\ A_{hij}} y_{hij} I_{hijk}}, \qquad (38)$$

$$\hat{R}_{\#c}2 = \frac{\sum_{h=1}^{H}\sum_{i\in S_h}\sum_{j\in U_{hi}}\sum_{k=1}^{P}\frac{M_h T_{hi}}{m_h\ P\ A_{hij}} y_{hij} (V_{(c)hij} + A_{(c)hij})\ I_{hijk}}{\sum_{h=1}^{H}\sum_{i\in S_h}\sum_{j\in U_{hi}}\sum_{k=1}^{P}\frac{M_h T_{hi}}{m_h\ P\ A_{hij}} y_{hij} I_{hijk}}, \qquad (39)$$

and

$$\hat{R}_{\#c}3 = \frac{\sum_{h=1}^{H}\sum_{i\in S_h}\sum_{j\in U_{hi}}\sum_{k=1}^{P}\frac{M_h T_{hi}}{m_h\ P\ A_{hij}} y_{hij} (1 - V_{(c)hij}\ A_{(c)hij}) I_{hijk}}{\sum_{h=1}^{H}\sum_{i\in S_h}\sum_{j\in U_{hi}}\sum_{k=1}^{P}\frac{M_h T_{hi}}{m_h\ P\ A_{hij}} y_{hij} I_{hijk}}. \qquad (40)$$

Using Figure 1 as an example, the values of the variables of interest y_{hij} would be computed as follows:

- For the point $k = 1$, $I_{hij1} = 1$ just for the UP B, then the value of y_{hij} is the response recorded by the UP B.
- For the point $k = 2$, $I_{hij2} = 1$ just for the UP E, then the value of y_{hij} is the response recorded by the UP E.
- For the point $k = 3$, $I_{hij3} = 1$ just for the UP E, then the value of y_{hij} is the response recorded by the UP E, that is, there is a repetition of the same observation recorded for the point $k = 2$.
- For the point $k = 4$, $I_{hij4} = 0$ because no UP is selected, then the value of the recorded response is null.
- For the point $k = 5$, $I_{hij5} = 0$ because, no UP is selected, then the value of the recorded response is also null.

5.2 Using a list frame

Using only one list frame as a reference for the survey simplifies the sampling context since, typically, list frames have production units (farms) as sampling units. In these conditions, it is natural to assume the use of different stratification and probability elements in the sample design. For notation purposes, we imagine that H strata are identified and that each unit of production (agricultural establishment) is selected at a stage with probability proportional to a variable of convenient size. The following notation is introduced:

- $U = \{1, \dots, N\}$ represents the universe of N national agricultural establishments, identified in a population structure in which:
    - $U = \bigcup_{h=1}^{H} U_h$, that is, the population is stratified into H strata, with

$U_h = \{1, \dots, N_h\}$ representing the set of N_h agricultural establishments of the stratum h, such that $N = \sum_{h=1}^{H} N_h$.

- y_{hi} represents the agricultural area of the agricultural establishment i of the stratum h.

- S_h represents a probabilistic sample, of size n_h, selected within the stratum h, such that π_{hi} represents the probability of selection of the agricultural establishment i in the stratum h.

The notation of the variables previously defined as "$V_{(c)hij}$" and "$A_{(c)hij}$", needs to be adjusted slightly, suppressing the index j, to be written as:

- $V_{(c)hi}$ represents an indicator variable that the agricultural establishment i of the stratum h is classified in the "Green" category of sustainability, according to the criteria $c \in \{C_{\#1}, \dots, C_{\#11}\}$:

$$V_{(c)hi} = \begin{cases} 1, \text{if the agricultural establishment "}hi\text{" is "Green" according to criterion } c \\ 0, \text{if the agricultural establishment "}hi\text{" is not "Green"} \end{cases}$$

and

- $A_{(c)hi}$ represents an indicator variable that the agricultural establishment i in the stratum h is classified in the "Yellow" category of sustainability, according to the criteria $c \in \{C_{\#1}, \dots, C_{\#11}\}$:

$$A_{(c)hi} = \begin{cases} 1, \text{if the agricultural establishment "hi" is "Yellow" according to criterion c} \\ 0, \text{if the agricultural establishment "hi" is not "Yellow"} \end{cases}$$

Under these conditions, the interest estimators remain as follows:

$$\hat{R}_{\#c} = \frac{\sum_{h=1}^{H} \sum_{i \in S_h} \frac{y_{hi} V_{(c)hi}}{\pi_{hi}}}{\sum_{h=1}^{H} \sum_{i \in S_h} \frac{y_{hi}}{\pi_{hi}}}, \qquad (41)$$

$$\hat{R}_{\#c}1 = \frac{\sum_{h=1}^{H} \sum_{i \in S_h} \frac{y_{hi} A_{(c)hi}}{\pi_{hi}}}{\sum_{h=1}^{H} \sum_{i \in S_h} \frac{y_{hi}}{\pi_{hi}}}, \qquad (42)$$

$$\hat{R}_{\#c}2 = \frac{\sum_{h=1}^{H} \sum_{i \in S_h} \frac{y_{hi}\left(V_{(c)hi} + A_{(c)hi}\right)}{\pi_{hi}}}{\sum_{h=1}^{H} \sum_{i \in S_h} \frac{y_{hi}}{\pi_{hi}}}, \qquad (43)$$

and

$$\hat{R}_{\#c}3 = \frac{\sum_{h=1}^{H} \sum_{i \in S_h} \frac{y_{hi}(1 - V_{(c)hi} A_{(c)hi})}{\pi_{hi}}}{\sum_{h=1}^{H} \sum_{i \in S_h} \frac{y_{hi}}{\pi_{hi}}}. \qquad (44)$$

5.3 Using a dual frame

The sample designs of dual frames for agricultural surveys commonly use an area frame and a list frame simultaneously. Hence, the notation introduced above for each type of frame is retained. A probabilistic sample is drawn independently from each frame. In this scenario, three types of list frame usage include:

i. The frame identifies all N agricultural establishments of the target population and a sample of n agricultural establishments ($n < N$) is taken from this frame using probability sampling.

ii. The frame identifies the largest agricultural establishments; all of them are part of the sample, that is, a census is applied to the list frame.

iii. The frame identifies the N_G largest agricultural establishments ($N_G < N$), and a sample of n establishments ($n < N_G$) is taken from it using probability sampling.

Scenarios i. and iii. have the same statistical treatment in terms of the estimator formulas. Scenario ii. does not present difficulties, because it implies the use of the formulas for the area frame, introduced in 5.1, added from the observation of the list frame.

Regarding the use of the area frame, three scenarios were considered in this guide: estimation using the open segment strategy (5.1), estimation using the one-stage weighted segment strategy (5.2) and estimation using the two-stage weighted segment strategy (5.3).

To exemplify the derivation of the estimator formulas in a dual-frame design, the case where the area frame uses the two-stage weighted segment strategy will be used (5.3), and the list frame identifies the largest producers, as described in iii. The other scenarios can be derived in a similar way.

In this guide, two estimators for dual frames are presented: simple multiplicity and screening.

5.3.1 Using simple multiplicity

Mecatti (2007) introduced an estimator of multiple frames based on a simple multiplicity factor. In this guide, this factor is represented by f_{hij}^{A}, such that f_{hij}^{A} represents the number of frames to which an agricultural establishment hij, identified through the area frame, belongs to. In this case,

$$f_{hij}^{A} = \begin{cases} 1, \text{if the agricultural establishment } hij \text{ belongs to the area frame only} \\ 2, \text{if the agricultural establishment } hij \text{ belongs to the list and the area frame} \end{cases}$$

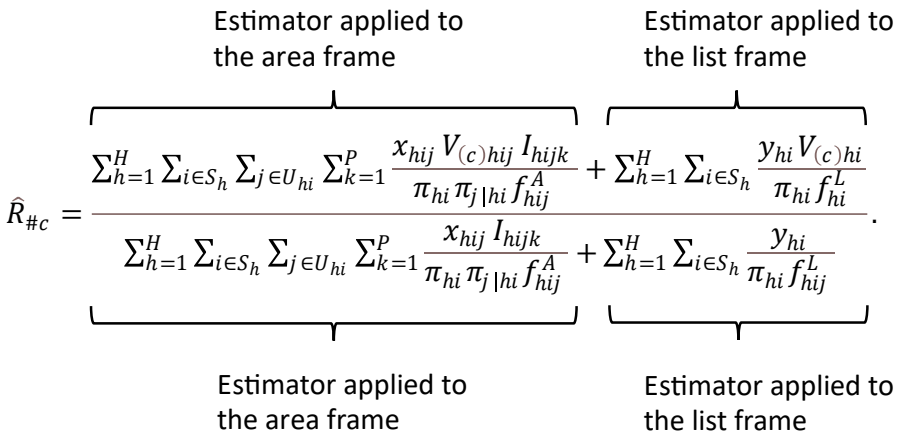
When, in this guide, the agricultural establishment is identified through the list frame, the multiplicity factor is represented by f_{hi}^{L}, so that f_{hi}^{L} represents the number of frames to which an agricultural establishment hi, identified through the area frame, belongs to. In agricultural surveys, the list frame is embedded in the area frame, so the multiplicity factor assumes the value $f_{hi}^{L} = 2$.

We consider the use of subsampling of agricultural establishments, within each segment of the area frame, using P points. We also consider the use of a simple multiplicity factor, associated with the agricultural establishments within each stratum. So, it is possible to write $\hat{R}_{\#c}$ for a dual frame design, taking advantage of formulas (34) and (41), with appropriate adaptations. The formula would be presented as follows:

$$\hat{R}_{\#c} = \frac{\sum_{h=1}^{H}\sum_{i\in S_h}\sum_{j\in U_{hi}}\sum_{k=1}^{P}\frac{x_{hij}V_{(c)hij}I_{hijk}}{\pi_{hi}\pi_{j|hi}f_{hij}^{A}} + \sum_{h=1}^{H}\sum_{i\in S_h}\frac{y_{hi}V_{(c)hi}}{\pi_{hi}f_{hi}^{L}}}{\sum_{h=1}^{H}\sum_{i\in S_h}\sum_{j\in U_{hi}}\sum_{k=1}^{P}\frac{x_{hij}I_{hijk}}{\pi_{hi}\pi_{j|hi}f_{hij}^{A}} + \sum_{h=1}^{H}\sum_{i\in S_h}\frac{y_{hi}}{\pi_{hi}f_{hi}^{L}}}. \qquad (45)$$

The expression (45) contains an abuse of notation because it deals at the same time with two different contexts, one belonging to area frame and other to list frame, as it can be seen in Figure 2.

Figure 2. Dual frame estimator by frame type



(45)

Source: Author's own elaboration.

So, for example, in the area frame estimator part, a total of H strata are used, while in the list frame part, although the same notation is used (H strata), that number is probably different. To arrive at a more precise expression, the following notation is introduced:

Notation used in the area frame

- $U_A = \{1, \dots, M\}$ represents the universe of M area segments that form the area frame, such that:
  - $U_A = \bigcup_{h=1}^{H} U_h^A$, i.e., the area frame is stratified into H strata, with:
    - $U_h^A = \{1, \dots, M_h\}$, representing the set of M_h area segments located in the stratum h, so $M = \sum_{h=1}^{H} M_h$ is the total number of clusters in the population.
    - U_{hi}^A represents the set of agricultural establishments that has tracts within segment i, of the stratum h.
  - A sample of P agricultural establishments is selected from U_{hi}^A, with replacement, and a probability of selection $\pi_{j|hi}^A$.
- x_{hij}^A represents the variable of weighted agricultural area of the agricultural establishment j which has tracts within area segment i, in the stratum h.
- S_h^A represents a probabilistic sample of m_h segments, selected within the stratum h of the area frame, such that π_{hi}^A represents the probability of selection of the agricultural establishment i in the stratum h of the area frame.

Notation used in the list frame

- U_L represents the universe of N_G largest agricultural establishments identified in the list frame, such that:
  - $U_L = \bigcup_{h=1}^{H'} U_h^L$, i.e., the list frame is stratified into H' strata, with U_h^L representing the set of N_{Gh} largest agricultural establishments in the stratum h, such that $N_G = \sum_{h=1}^{H} N_{Gh}$;

- y_{hi}^{L} represents the agricultural area of the agricultural establishment i of the stratum h of the list frame.
- S_h^L represents a probabilistic sample, of size n_{Gh}, selected within the stratum h of the list frame, such that π_{hi}^{L} represents the probability of selection of the agricultural establishment i in the stratum h of the list frame.

The formula (45) can be rewritten more formally as:

$$\hat{R}_{\#c} = \frac{\sum_{h=1}^{H} \sum_{i \in S_h^A} \sum_{j \in U_{hi}^A} \sum_{k=1}^{P} \frac{x_{hij}^A V_{(c)hij} I_{hijk}}{\pi_{hi}^A \pi_{j|hi}^A f_{hij}^A} + \sum_{h=1}^{H'} \sum_{i \in S_h^L} \frac{y_{hi}^L V_{(c)hi}}{\pi_{hi}^L f_{hi}^L}}{\sum_{h=1}^{H} \sum_{i \in S_h^A} \sum_{j \in U_{hi}^A} \sum_{k=1}^{P} \frac{x_{hij}^A I_{hijk}}{\pi_{hi}^A \pi_{j|hi}^A f_{hij}^A} + \sum_{h=1}^{H'} \sum_{i \in S_h^L} \frac{y_{hi}^L}{\pi_{hi}^L f_{hi}^L}}. \qquad (46)$$

Remembering the expression (33), it is still possible to write:

$$x_{hij}^A = \frac{T_{hij}}{A_{hij}} y_{hij}^A,$$

Where:

y_{hij}^A represents the agricultural area of the agricultural establishment j that has "tract" within the segment i of the stratum h, in the area frame; and T_{hij} and A_{hij} have the same definition previously introduced, meaning:

- T_{hij} represents the area of the "tract" of the agricultural establishment hij; and
- A_{hij} represents the total area of the agricultural establishment hij.

Now, the expression (46) can finally be rewritten as:

$$\hat{R}_{\#c} = \frac{\sum_{h=1}^{H} \sum_{i \in S_h^A} \sum_{j \in U_{hi}^A} \sum_{k=1}^{P} \frac{T_{hij}}{A_{hij}} \frac{y_{hij}^A V_{(c)hij} I_{hijk}}{\pi_{hi}^A \pi_{j|hi}^A f_{hij}^A} + \sum_{h=1}^{H'} \sum_{i \in S_h^L} \frac{y_{hi}^L V_{(c)hi}}{\pi_{hi}^L f_{hi}^L}}{\sum_{h=1}^{H} \sum_{i \in S_h^A} \sum_{j \in U_{hi}^A} \sum_{k=1}^{P} \frac{T_{hij}}{A_{hij}} \frac{y_{hij}^A I_{hijk}}{\pi_{hi}^A \pi_{j|hi}^A f_{hij}^A} + \sum_{h=1}^{H'} \sum_{i \in S_h^L} \frac{y_{hi}^L}{\pi_{hi}^L f_{hi}^L}}. \qquad (47)$$

Considering the case where a simple random sampling is used for selection of area segments in the area frame within a stratum h, and a sampling with different probabilities is used for selection of farms in the list frame, and in addition, using the value of $f_{hi}^L = 2$, then the expression (47) is written as

$$\hat{R}_{\#c} = \frac{\sum_{h=1}^{H} \sum_{i \in S_h^A} \sum_{j \in U_{hi}^A} \sum_{k=1}^{P} \frac{M_h \, T_{hi} \, y_{hij}^A V_{(c)hij}}{P \, m_h \, A_{hij} \, f_{hij}^A} I_{hijk} + \sum_{h=1}^{H'} \sum_{i \in S_h^L} \frac{y_{hi}^L V_{(c)hi}}{2 \, \pi_{hi}^L}}{\sum_{h=1}^{H} \sum_{i \in S_h^A} \sum_{j \in U_{hi}^A} \sum_{k=1}^{P} \frac{M_h \, T_{hi} \, y_{hij}^A}{P \, m_h \, A_{hij} \, f_{hij}^A} I_{hijk} + \sum_{h=1}^{H'} \sum_{i \in S_h^L} \frac{y_{hi}^L}{2 \, \pi_{hi}^L}}. \qquad (48)$$

The other estimators are:

$$\hat{R}_{\#c}1 = \frac{\sum_{h=1}^{H}\sum_{i\in S_h^A}\sum_{j\in U_{hi}^A}\sum_{k=1}^{P}\frac{M_h\, T_{hi}\, y_{hij}^A A_{(c)hij}}{P\, m_h\, A_{hij}\, f_{hij}^A} I_{hijk} + \sum_{h=1}^{H'}\sum_{i\in S_h^L}\frac{y_{hi}^L A_{(c)hi}}{2\,\pi_{hi}^L}}{\sum_{h=1}^{H}\sum_{i\in S_h^A}\sum_{j\in U_{hi}^A}\sum_{k=1}^{P}\frac{M_h\, T_{hi}\, y_{hij}^A}{P\, m_h\, A_{hij}\, f_{hij}^A} I_{hijk} + \sum_{h=1}^{H'}\sum_{i\in S_h^L}\frac{y_{hi}^L}{2\,\pi_{hi}^L}}. \quad (49)$$

$$\hat{R}_{\#c}2 = \frac{\sum_{h=1}^{H}\sum_{i\in S_h^A}\sum_{j\in U_{hi}^A}\sum_{k=1}^{P}\frac{M_h\, T_{hi}\, y_{hij}^A (V_{(c)hij} + A_{(c)hij})}{P\, m_h\, A_{hij}\, f_{hij}^A} I_{hijk} + \sum_{h=1}^{H'}\sum_{i\in S_h^L}\frac{y_{hi}^L (V_{(c)hi} + A_{(c)hi})}{2\,\pi_{hi}^L}}{\sum_{h=1}^{H}\sum_{i\in S_h^A}\sum_{j\in U_{hi}^A}\sum_{k=1}^{P}\frac{M_h\, T_{hi}\, y_{hij}^A}{P\, m_h\, A_{hij}\, f_{hij}^A} I_{hijk} + \sum_{h=1}^{H'}\sum_{i\in S_h^L}\frac{y_{hi}^L}{2\,\pi_{hi}^L}}. \quad (50)$$

$$\hat{R}_{\#c}3 = \frac{\sum_{h=1}^{H}\sum_{i\in S_h^A}\sum_{j\in U_{hi}^A}\sum_{k=1}^{P}\frac{M_h\, T_{hi}\, y_{hij}^A (1 - V_{(c)hij} A_{(c)hij})}{P\, m_h\, A_{hij}\, f_{hij}^A} I_{hijk} + \sum_{h=1}^{H'}\sum_{i\in S_h^L}\frac{y_{hi}^L (1 - V_{(c)hi} A_{(c)hij})}{2\,\pi_{hi}^L}}{\sum_{h=1}^{H}\sum_{i\in S_h^A}\sum_{j\in U_{hi}^A}\sum_{k=1}^{P}\frac{M_h\, T_{hi}\, y_{hij}^A}{P\, m_h\, A_{hij}\, f_{hij}^A} I_{hijk} + \sum_{h=1}^{H'}\sum_{i\in S_h^L}\frac{y_{hi}^L}{2\,\pi_{hi}^L}}. \quad (51)$$

5.3.2 Using the screening estimator

The screening estimator is a composition of information from the two frames, area and list frames, in which information referring to the largest producers is computed only through the sample of the list frame, and information referring to the other producers is computed only through the area frame.

To take advantage of reasoning in the development of formulas (48), (49), (50) and (51), we consider the following notation:

- $L_{hij} = \begin{cases} 1, \text{if the establishment } j \text{ with tracts within the segment } i \text{ in the stratum } h \\ \text{belongs to the list frame; and} \\ 0, \text{otherwise.} \end{cases}$

The formulas referring to the screening estimator are:

$$\hat{R}_{\#c} = \frac{\sum_{h=1}^{H}\sum_{i\in S_h^A}\sum_{j\in U_{hi}^A}\sum_{k=1}^{P}\frac{M_h\, T_{hi}\, y_{hij}^A V_{(c)hij}}{P\, m_h\, A_{hij}\, f_{hij}^A}(1 - L_{hij})\, I_{hijk} + \sum_{h=1}^{H'}\sum_{i\in S_h^L}\frac{y_{hi}^L V_{(c)hi}}{\pi_{hi}^L}}{\sum_{h=1}^{H}\sum_{i\in S_h^A}\sum_{j\in U_{hi}^A}\sum_{k=1}^{P}\frac{M_h\, T_{hi}\, y_{hij}^A}{P\, m_h\, A_{hij}\, f_{hij}^A}(1 - L_{hij})\, I_{hijk} + \sum_{h=1}^{H'}\sum_{i\in S_h^L}\frac{y_{hi}^L}{\pi_{hi}^L}}. \quad (52)$$

$$\hat{R}_{\#c}1 = \frac{\sum_{h=1}^{H}\sum_{i\in S_h^A}\sum_{j\in U_{hi}^A}\sum_{k=1}^{P}\frac{M_h\, T_{hi}\, y_{hij}^A A_{(c)hij}}{P\, m_h\, A_{hij}\, f_{hij}^A}(1 - L_{hij}) I_{hijk} + \sum_{h=1}^{H'}\sum_{i\in S_h^L}\frac{y_{hi}^L A_{(c)hi}}{\pi_{hi}^L}}{\sum_{h=1}^{H}\sum_{i\in S_h^A}\sum_{j\in U_{hi}^A}\sum_{k=1}^{P}\frac{M_h\, T_{hi}\, y_{hij}^A}{P\, m_h\, A_{hij}\, f_{hij}^A}(1 - L_{hij}) I_{hijk} + \sum_{h=1}^{H'}\sum_{i\in S_h^L}\frac{y_{hi}^L}{\pi_{hi}^L}}. \quad (53)$$

$$\hat{R}_{\#c}2 = \frac{\sum_{h=1}^{H}\sum_{i\in S_h^A}\sum_{j\in U_{hi}^A}\sum_{k=1}^{P}\frac{M_h\, T_{hi}\, y_{hij}^A (V_{(c)hij} + A_{(c)hij})}{P\, m_h\, A_{hij}\, f_{hij}^A}(1 - L_{hij}) I_{hijk} + \sum_{h=1}^{H'}\sum_{i\in S_h^L}\frac{y_{hi}^L (V_{(c)hi} + A_{(c)hi})}{2\,\pi_{hi}^L}}{\sum_{h=1}^{H}\sum_{i\in S_h^A}\sum_{j\in U_{hi}^A}\sum_{k=1}^{P}\frac{M_h\, T_{hi}\, y_{hij}^A}{P\, m_h\, A_{hij}\, f_{hij}^A}(1 - L_{hij}) I_{hijk} + \sum_{h=1}^{H'}\sum_{i\in S_h^L}\frac{y_{hi}^L}{2\,\pi_{hi}^L}}. \quad (54)$$

$$\hat{R}_{\#c}3 = \frac{\sum_{h=1}^{H}\sum_{i\in S_h^A}\sum_{j\in U_{hi}^A}\sum_{k=1}^{P}\frac{M_h\, T_{hi}\, y_{hij}^A(1-V_{(c)hij}A_{(c)hij})}{P\, m_h\, A_{hij}\, f_{hij}^A}(1-L_{hij})I_{hijk} + \sum_{h=1}^{H'}\sum_{i\in S_h^L}\frac{y_{hi}^L(1-V_{(c)hi}A_{(c)hij})}{2\,\pi_{hi}^L}}{\sum_{h=1}^{H}\sum_{i\in S_h^A}\sum_{j\in U_{hi}^A}\sum_{k=1}^{P}\frac{M_h\, T_{hi}\, y_{hij}^A}{P\, m_h\, A_{hij}\, f_{hij}^A}(1-L_{hij})I_{hijk} + \sum_{h=1}^{H'}\sum_{i\in S_h^L}\frac{y_{hi}^L}{2\,\pi_{hi}^L}}. \quad (55)$$

6 Estimation for subnational domains

The process of estimating Indicator 2.4.1 for an entire national territory includes the collection of a wide variety of data, since it depends on an evaluation of each of the 11 subindicators. That has, of course, a cost that can be translated in terms of sample size. When subnational divisions of interest coincide with the domains of interest that are expected in the agricultural survey, the sample sizes are expected to be large enough to generate estimates of Indicator 2.4.1 for each of them. This is the case, for example, of surveys that use a stratum definition factor relative to the subnational domain of interest. However, planning a survey with stratification factors coinciding with more advanced levels of disaggregation may result in prohibitive financial costs. Therefore, in many situations, estimates may be generated for domains that do not match previously planned stratification factors and, therefore, the quality of inference depends on the sample size observed within each domain.

To define Indicator 2.4.1 as a parameter in a subnational domain of interest, the following notation is entered:

- $U_d = \{1, \dots, N_d\}$ represents the universe of N_d agricultural establishments belonging to the subnational domain d, identified in a population structure in which:
    - $U_d = \bigcup_{h=1}^{H}(U_h \cap U_{dh})$, i.e. the factor that defines the domain d is crossed with the factor defining the stratum, with:
        - U_h representing the set of N_h agricultural establishments of the stratum h.
        - U_{dh} represents the set of N_{dh} agricultural establishments of the stratum h belonging to the subnational domain d.
- y_{hi} represents the agricultural area of the agricultural establishment i in the stratum h.

Indicator 2.4.1 has the following form in the subnational domain of interest d:

$$R_{241_d} = min\left(R_{\#1_d}, \dots, R_{\#11_d}\right), \qquad (56)$$

where each of the subindicators $R_{\#c_d}: c \in \{1, \dots, 11\}$ corresponds to an expression similar to:

$$R_{\#c_d} = \frac{TSAS\#c_d}{TSA_d}, \qquad (57)$$

Where:

$$TSAS\#c_d = \sum_{h=1}^{H} \sum_{i \in U_{dh}} y_{hi} V_{(c)hi}, \qquad (58)$$

and

$$TSA_d = \sum_{h=1}^{H} \sum_{i \in U_{dh}} y_{hi}. \qquad (59)$$

- $TSAS\#c_d$ represents the total agricultural area of the subnational domain d where agriculture is practiced as productive and sustainable in accordance with the criterion $C_{\#i}$ used by the subindicator $R_{\#c_d}$.
- TSA_d represents the entire agricultural area of subnational domain d.

Estimator (56) represents the parametric form of the subindicator $R_{\#c_d}$ concerning the subnational domain d, based on the "Green" category of sustainability. In this case, $R_{\#c_d}$ represents the amount of agricultural area of the subnational domain d with **desirable levels of sustainability** according to criterion $C_{\#c}$. In parallel, the population parameters

$$R_{\#c_d}1 = \frac{\sum_{h=1}^{H}\sum_{i\in U_{dh}} y_{hi}A_{(c)hi}}{\sum_{h=1}^{H}\sum_{i\in U_{dh}} y_{hi}}, \qquad (60)$$

$$R_{\#c_d}2 = \frac{\sum_{h=1}^{H}\sum_{i\in U_{dh}} y_{hi}(V_{(c)hi} + A_{(c)hi})}{\sum_{h=1}^{H}\sum_{i\in U_{dh}} y_{hi}}, \qquad (61)$$

and

$$R_{\#c_d}3 = \frac{\sum_{h=1}^{H}\sum_{i\in U_{dh}} y_{hi}(1 - V_{(c)hi}A_{(c)hi})}{\sum_{h=1}^{H}\sum_{i\in U_{dh}} y_{hi}} \qquad (62)$$

represent the amount of agricultural area of the same subindicator for the same subnational domain d of interest, using different categories of sustainability:

- $R_{\#c_d}1$ is the proportion of subnational agricultural area with **acceptable levels of sustainability** (Yellow) according to criterion $C_{\#c}$.
- $R_{\#c_d}2$ is the proportion of subnational agricultural area with **acceptable or desirable levels of sustainability** (Yellow or Green) according to criterion $C_{\#c}$.
- $R_{\#c_d}3$ is the amount of **unsustainable** subnational agricultural area (Red) according to criterion $C_{\#c}$.

The development of subnational estimators in all situations in this guide follows steps similar to those described for the estimation of the indicator for the entire national territory.

7 Conclusion

This guide has shown some ways to estimate SDG Indicator 2.4.1 on agricultural sustainability, considering various statistical aspects of sample design.

Given this is a first version, several elements can and should be improved in future editions of this guide. For instance, we expect to include more references to documentation that can help with concepts related to the topic. We also hope to have a more complete document that includes efforts to demonstrate the use of questionnaires for data collection, delving into the experiences of Colombia, Costa Rica and Peru. In addition, two other important topics for discussion are the problem of estimation in subnational domains considered small, and the problem of estimating a quality measure of the indicator estimate.

Regarding the issue of the quality measure for the indicator, it is important to engage in a discussion about the estimation of variance, which is not a trivial problem. Furthermore, considering alternatives for quality measures that may have a less restricted interpretation for countries is crucial. The use of the range between subindicators, for example, could be a more interesting measure. Other possibilities could also be explored.

References

Candia, A. 2022. *Análisis de datos e implementación del cálculo del indicador 2.4.1 en Perú.* Documento interno Informe de consultoría proyecto para el cálculo de los indicadores de ODS. Sucre.

Candia, A. 2023. *Apoyo a los sistemas estadísticos nacionales para el fortalecimiento de los indicadores ODS 2.4.1 en Costa Rica.* Documento interno Informe de consultoría proyecto para el cálculo de los indicadores de ODS. Sucre.

FAO. 2018. *AGRIS Handbook on the Agricultural Integrated Survey.* Rome. https://www.fao.org/3/ca6412en/ca6412en.pdf

FAO. 2021. *Sampling guidance for SDG Indicator 2.4.1.* Rome. https://www.fao.org/3/ca7439EN/ca7439EN.pdf

FAO. 2023a. *SDG indicator metadata 2.4.1.* Rome. https://unstats.un.org/sdgs/metadata/files/Metadata-02-04-01.pdf

FAO. 2023b. *Proportion of agricultural area under productive and sustainable agriculture.* Methodological note, Revision 11. Rome. https://www.fao.org/3/ca7154en/ca7154en.pdf

Mecatti, F. 2007. A single frame multiplicity estimator for multiple frame surveys*. Survey Methodology,* 33(2): 151–158.

Nova, D. 2021. *Ejercicio piloto para la medición del indicador 2.4.1 a partir de la Encuesta Nacional Agropecuaria (ENA) en Colombia.* Documento interno Informe de consultoría proyecto para el cálculo de los indicadores de ODS. Bogotá.

Contact:

Statistics Division – Economic and Social Development
FAO-statistics@fao.org
www.fao.org/food-agriculture-statistics/resources/publications/working-papers/en/

Food and Agriculture Organization of the United Nations
Rome, Italy

ISBN 978-92-5-138577-7

CC9550EN/1/02.24