Enveritas

Defining number of samples – Detailed Approach

Carbon Footprint Baselines for Central Highlands and Southern Sumatra

April 2022

Project Scope and objectives

Main project outcome: carbon footprint assessments of robusta coffee produced in Central Highlands (Vietnam) and Southern Sumatra (Indonesia).

Objectives

- Collect inputs to the Cool Farm Tool at the coffee farm level
- Run the models to get carbon footprint estimates of coffee farmers
- Conduct data analyses to surface insights on carbon emissions (e.g., comparative analyses across defined archetypes, uncertainty assessments)

Scope

Central Highlands:
- Dak Lak
- Dak Nong
- Gia Lai
- Lam Dong

Southern Sumatra:
- Bengkulu
- Lampung
- Sumatera Selatan

Nature of collected data

- All applicable Cool Farm Tool inputs
- Other data points of interest identified by partners, with the purpose of conducting extra analyses

Sampling – Assumptions

Results	• Main result: carbon emissions estimates (kg CO2e per kg green bean or per ha where possible), disaggregated by main source (land use change, fertilizers, energy, etc.) • Secondary results: indicators used for data analyses, including but not limited to: yield, farm size, % of farmers who practice intercropping (final list to be decided at a later stage)
Desired granularity of results	• Main and secondary results representative at **province** and **origin levels** • Main and secondary results representative at district level for largest districts
Confidence level	**95%** (the goal is to minimize the margin of error at this level)
Buffer (to account for non-valid surveys)	**5%**
Level of sampling	**District** (# of required samples per district provided)
Sampling unit	**Farm**
Geographical scope	**Vietnam:** Daklak, Lam Dong, Dak Nong, Gia Lai **Indonesia:** Sumatera Selatan, Bengkulu, Lampung
Sample weights	Based on **production volumes**

Sampling strategy

1

Stratification: divide each origin into provinces in the scope of this study

2

Define a suitable minimum sample size for each stratum based on the sampling assumptions

3

Distribute provinces' samples across districts using their production volumes as weights

4

Remove districts that have 10 samples or less and reallocate samples (goal: simplify operations)

5

Round samples per district to the nearest 5

6

Do final manual adjustments to ensure that largest districts (for which representative results are required) reach a minimum sample size.

2. Define a suitable minimum sample size per province

Requirements:

- **Accuracy**: both carbon emission results and secondary indicators (e.g. yield, % intercropping, etc.) must be estimated with a low margin of error (objective: <10% MoE with a 95% confidence level, considering a 5% buffer to account for rejected surveys)

- **Operational feasibility**: sample sizes must remain reasonable to ensure that suppliers / partners will achieve the data collection targets

Approach:

- Set a maximum target that is **operationally feasible**:
  - As per discussions with suppliers and partners, a total of **650 samples per province** in both origins is seen as a reasonable target
  - Hence, **2,600 samples in Central Highlands** and **1,950 samples in Southern Sumatra** will be collected

- Check that the suggested sample sizes meet the desired level of **accuracy** for both types of variables (see next slide for details):
  - Binary variables (e.g., % farmers who use inorganic fertilizers, % of farmers who irrigate)
  - Continuous variables (e.g., CO2e emissions, yield, quantity of fertilizer used)

2. Define a suitable minimum sample size per province – accuracy assessment

For binary variables:

- Assuming an unlimited population (conservative): $n = \left(\frac{z * \sigma}{E}\right)^2 = \left(\frac{z * p * (1-p)}{E}\right)^2$
  - n is the sample size
  - z is the Z-score, equal to 1.96 (95% confidence interval assumed)
  - σ is the standard deviation of the indicator
  - E is the margin of error of the indicator
  - p is the uncertainty level (set at 50% to consider a maximum variance)
- For a sample size of n = 650 reduced by 5% to account for rejected surveys, the resulting margin of error is **4%**, below the target 10%
- This margin of error is theoretical and assumes an unbiased random sample, which is difficult to guarantee in reality. The difference between the target 10% and theoretical result of 4% will thus serve as buffer to account for this and will give room for data analyses across archetypes.

For continuous variables:

- Assuming a random sample with no bias: $n = \left(\frac{z * \sigma}{E}\right)^2$
- We are trying to determine the maximum standard deviation of the variable allowed under a 10% Margin of Error and 95% confidence level:

  $$n = \left(\frac{1.96 * \sigma}{10\% * m}\right)^2 \Rightarrow \frac{\sigma}{m} = \frac{\sqrt{n}}{19.6}$$

- For a sample size of n = 650 reduced by 5% to account for rejected surveys, the resulting relative standard deviation (σ / m) is **1.30**
- This relative standard deviation values is compared against external datasets / studies similar to the one that will be conducted:
  - Province-level carbon emission estimates in Central Highlands (USAID / JDE / IDH, 2019)*: **0.39 – 0.68**
  - Province-level yield estimates in Central Highlands (Enveritas, 2020) – one of the largest drivers of emissions: **0.47 – 0.66**
  - Province-level fertilizer usage estimates in Central Highlands (Enveritas, 2020) – one of the largest drivers of emissions: **0.47 – 0.54**
- Comparisons show that the theoretical maximum relative standard deviation of 1.30 is sufficient (roughly 2x the benchmark values). However, those comparisons are limited to Central Highlands (no usable dataset could be found for Southern Sumatra) and not perfectly relevant as the planned study does not strictly follow the same approach (e.g. the 2019 study used farmer lists for sampling instead of a purely random approach). Moreover, the suggested sampling approach is theoretical and assumes an unbiased random sample, which is difficult to guarantee in reality.
- The relatively high value of the maximum standard deviation will serve as a buffer to account for these uncertainties and will give room for data analyses across archetypes.

*: Scaling Up Sustainable Robusta Coffee Production In Vietnam: Reducing Carbon Footprints While Improving Farm Profitability (USAID / JDE / IDH, 2019)

3. Distribute samples across districts

Approach:

- Use proportional allocation to distribute samples across districts – each district will get a sample size proportional to its coffee production volume

Output:

- Suppliers/partners are given a table with the recommended number of samples per district, which are then distributed among them
- As the actual number of samples per district will most likely be different from the theoretical one, custom weights will be assigned per collected sample during the data processing stage to ensure that results remain consistent with districts' production volumes

Province	District	Sample size
Daklak	Buôn Ma Thuột	116
Daklak	Ea H'leo	316
Daklak	Ea Súp	0
Daklak	Krông Năng	281
Daklak	Krông Búk	205
Daklak	Buôn Đôn	45
Daklak	Cư M'Gar	347
Daklak	Ea Kar	45

4. Remove districts that have few samples

Rationale:

- Districts with a low number of samples would require significant operational work for a limited gain (results will be negligible when aggregated with districts that have a much larger sample size)
- Besides, districts with low production volumes (and hence few samples) would be particularly challenging to survey because suppliers' / partners' local presence is likely to be lower in those areas

Approach:

- Define a minimum number of samples under which the district will be removed from the sampling plan and the samples reassigned → minimum sample size is set at **10**
- Reallocate the dropped samples to the top 5 producing districts per origin* (equal allocation)

Results:

- The total number of reallocated samples remains low for both origins

	Central Highlands	Southern Sumatra
# districts dropped	5	12
# samples dropped	14 / 2,600	40 / 1,950
% production volume dropped	0.7%	2.7%

*: 6 districts in Central Highlands (Krong Nang added as per partners' request)

5. Round samples per district to the nearest 5

Rationale:

- The farm randomization process will happen as such: GPS points will be assigned to field agents, who will navigate to each point and collect surveys nearby
- Field agents will have the possibility to collect a maximum of 3 surveys per GPS point
- On average, it is expected that between 2 and 3 surveys will be collected per GPS point
- Hence, 2 GPS points will be created for every 5 surveys – it is therefore needed to have a number of surveys per district that is a multiple of 5

Approach:

- Sample sizes are rounded to the nearest 5 in each district

6. Manual adjustments

Rationale:

- As per suppliers' and partners' requests, some large districts will need representative results that require more samples
- Besides, rounding to the nearest 5 led to slight changes in aggregate sample sizes per province, which need to be brought back to the initial 650

Approach:

- Sample sizes of the top 5 producing districts per origin* are adjusted to reach at least 110
  - 110 represents the minimum number of samples required for a binary variable to reach 10% Margin of Error at a 95% confidence interval, with a 5% buffer (rounded up to the nearest 10).
  - Although restricted to binary variables, this model is considered to be a good balance between an expected sufficient level of accuracy at the district level and a decent sample size with no substantial impact on other districts
- Samples of other districts are adjusted to keep the province-level total at 650, while maintaining at least 10 samples per district in scope

Results:

- Large districts all have a minimum sample size of 110
- All province-level sample sizes remain at 650

		# samples
Central Highlands	Ea H'leo	115
	Krong Nang	110
	Cu M'Gar	120
	Lam Ha	165
	Di Linh	175
	Bao Lam	155
Southern Sumatra	Muara Enim	110
	Ogan Komering Ulu Selatan	175
	Empat Lawang	190
	Lampung Barat	325
	Tanggamus	200

*: 6 districts in Central Highlands (Krong Nang added as per partners' request)

Experts' comments and responses

CIRAD – Take geographical features (landscape & climate) into consideration in samples distribution, as they can have a strong impact on farming practices and resulting emissions.

- High complexity: research on landscapes & conditions (hydrological cycles, etc.) will have to be conducted for both Central Highlands and Southern Sumatra
- Lack of replicability: the goal is to define a robust and operationally simple approach to collect data that will feed into GHG emissions models (aka Cool Farm Tool). The goal is to be able to replicate this approach to other coffee origins
- Results must be representative at the province (and some districts') level: e.g., *"CO2e emissions for coffee produced in Dak Nong are …"*

➢ **Action:** Samples will be spread geographically across districts, with the aim of covering the various landscapes within each province, district. After the data collection stage, some landscape (e.g. elevation) and climatic features could be retrieved at the district / province level to conduct analyses and assess correlations, and potentially adjust sampling weights.

Experts' comments and responses (2)

CIRAD – Distribute samples by system type instead of randomly, to guarantee representativeness by system and optimize confidence level

- How to define a system/archetype? Two options:
  - *A priori system types (archetypes based on local experts + literature)*: No consensus among the 10+ suppliers & partners in the program, and lack of knowledge on systems in Southern Sumatra – "expert knowledge on type distributions" is scarce and can be subject to debate in some origins
  - *A posteriori system types (based on preliminary surveys)*: high complexity of a 2-stage survey, out of the scope of this program
- Replicability: a common approach needs to be taken for Central Highlands & Southern Sumatra (and other potential origins in future years), but systems highly differ

➢ **Action:** Random sampling across districts / provinces without considering archetypes, but with inflated sample sizes
  - Samples were reallocated to optimize MoE at province level (proportional allocation dropped, now sample sizes per province are equal)
  - Past studies in Central Highlands were used as benchmark – planned sample size was defined as largely above required minimum sample size according to the benchmark standard deviations.
  - Data analyses post data collection will be conducted to identify the main archetypes found, their distribution across the origins / landscapes, their differences in terms of carbon footprints, and recommendations for future sampling if the results for some archetypes were found not to be robust enough.

Experts' comments and responses (3)

CIRAD – Results should account for yearly variability, therefore use 3–4-year assessments instead of a single year

- This study is a 1-year program that will serve as a baseline for future similar studies
- The intent is to repeat the exercise for at least 2 more years to develop a more robust baseline covering climatic availability
- **Action:** The report will include context on this year's specific meteorological conditions (including ENSO index) to put results into perspective and assess how this year's results are representative in light of the long-term climate conditions. The report will also provide suggestion on how to conduct a follow-up multi-year program that can capitalize on the knowledge acquired during this first year.

CIRAD – The methodology should account for the whole perennial crop cycle, with two possible synchronic approaches: spatial assessment (even distribution of ages across the territories) or modular assessment (weighted average of results)

- In coffee, crop cycles can be highly diverse within the same farm (rejuvenation via stumping and/or replanting is common) – hence, it is difficult to single out farms based on their cycle stage (average tree age would be used, leading to uncertainties)
- Spatial assessment: operationally too complex to ask field agents to "survey farms with 30 years of age" for example
- **Action:** We will implement a modular assessment, using tree age categories per farm as inputs and weighing the results according to the crop cycle distribution. However, it is important to note that the nursery stage assessment is not considered in most carbon footprint models including Cool Farm Tool (often neglected in GHG models as emissions related to the first 2-5 years of crop growth are distributed over the many productive years). The carbon footprint estimate of nurseries would require a separate model and underlying questionnaire, outside of the scope of this study.

Experts' comments and responses (4)

CIRAD – Set up a structural approach to integrate existing data (e.g., already collected by suppliers / partners) into the process

- The goal is to be able to replicate the process across several origins, it is therefore important not to include any data pre-requisite for the project. However, a framework to integrate existing data – if it exists – into the process could be defined.
- **Action:** This project will not use data collected in the past, except for the purpose of validating an approach (e.g., sampling approach) or results (e.g., benchmark on GHG emissions figures), or filling the gaps left in the survey (e.g., estimate of soil organic matter content if the farmer does not know). We will however include guidelines on how existing data can inform the project and improve its precision. These guidelines will include a list of quality checks on the input data and a description of the process to integrate them into the project.

CIRAD – Characterize each district based on four pre-defined dimensions: climate/soil & altitude, farm sophistication, average age of the plantation, intensity of land use change. These dimensions will define district archetypes that will help determine the required sample sizes ahead of data collection as well as provide insights on districts' similarities that can be leveraged during the data processing stage to find substitute data and increase representativeness of results.

- Given the simplicity and replicability constraints, the archetype dimensions will need to be common to all origins, and the scoring/gradient must remain simple (max. 3 categories).
- District characterization cannot be used to inform sample sizes as those are determined using an already defined statistical approach.
- However, it can be used to help weigh samples at the data processing stage and ensure that the results are more representative, e.g., using the standard crop cycle for Robusta or average climate/soil conditions
- **Action:** We can convene a working group with experts from CIRAD and other suppliers / partners in order to characterize districts across Central Highlands and Southern Sumatra. We can then use this characterization along with external data (typical crop cycles, climate conditions, etc.) after the data collection stage to improve results' precision (to be discussed at a later stage). This characterization would be a useful way to summarize findings and identify differences across the landscapes / origins, as well as inform future sampling efforts.