Creating NFI Phase 2 sample location with SFC

2024-02-29

# Context

This technical report covers the establishment of field sample locations for the Improved “National Forest Inventory Plan for Indonesia Forest”. The main document proposes a two-phases systematic stratified design with target sampling sizes for seven strata of dryland forest and one strata of mangrove forest.

Three annexes to the main report describe: 1. Optimal sampling grids for the NFI phase 1 visual interpretation. 1. Results of the NFI Phase 1 visual interpretation. 1. Distribution of the NFI Phase 2 field sample location (this report).

A separate annex is dedicated to Mangrove, Mangrove sampling is not addressed in this report.

# Method

## General approach

Given a final dataset of NFI phase 1 visually interpreted samples, the general approach to selecting the NFI phase 2 samples was as follows:

1. Determine the distance of each sample along a space filling curve (SFC), following Hilbert’s distance and computed with Python dask-geopandas [library](https://dask-geopandas.readthedocs.io/en/stable/docs/reference/api/dask_geopandas.GeoSeries.hilbert_distance.html)).
2. Select samples along the curve based on a sampling proportion calculated with the number of plots in the pool (i.e. the NFI phase 1 plots that are accessible forest), the number of target plots to sample (i.e. the number of target plots for the Phase 2) and adjustment factors to take into consideration uncertainty around the identification of forest / non-forest in the Phase 1.
3. A random starting point is assigned then the selection is systematic following the space filling curve.
4. The adjustment factors are tailored manually to provide a number of forest plots slightly exceeding the target.
5. All natural forest samples are kept, 70 samples are randomly selected among the accessible Trees Outside Forest samples.
6. Separately, additional samples in primary forest are provided for an intensification study on forest types that cover an area too small to be effectively represented in a national level sampling design. These samples can be measured according to the NFI field manual and integrated in the NFI database management system. However they should be filtered out from the national level statistics calculation chains and separate calculation chains should be designed specifically for these samples.

The sampling proportion is calculated as follows:

in R:

sampling\_proportion = ceiling(n\_plot \* (1 + U) \* sfc\_adj / n\_pool)

With a manually adjusted variable to reach the target number of samples in accessible forest and the uncertainty around forest/non-forest identification in the Phase 1.

## Input Data

## Rows: 9894 Columns: 66  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (40): email, analysis\_duration, imagery\_title, pl\_island, smpl\_island, ...  
## dbl (21): plotid, sampleid, lon, lat, pl\_fra\_hexid, smpl\_fra\_hexid, perc\_fo...  
## lgl (3): flagged, imagery\_attributions, gain\_driver\_17\_23  
## dttm (1): collection\_time  
## date (1): date  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.  
## Rows: 8 Columns: 5  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (2): pl\_island, perc  
## dbl (3): agree, disagree, total  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

## Reading layer `sf\_ddggrid' from data source   
## `/Users/gaelsola/Github-repos/Indonesia-NFI-prepa-Phase1-dashboard/data/sf\_ddggrid.gpkg'   
## using driver `GPKG'  
## Simple feature collection with 9896 features and 1 field  
## Geometry type: POLYGON  
## Dimension: XY  
## Bounding box: xmin: 95.21834 ymin: -10.95941 xmax: 141.0978 ymax: 5.880254  
## Geodetic CRS: WGS 84

The following datasets where used for the analysis:

* CEO final dataset containing 9894 unique values from the visual interpretation phase. This data is stored as a CSV file, but it contains the plot locations (lat/long) and can easily be transformed to a geospatial dataset.

Sample of the NFI phase 1 CEO data with user anonymized.

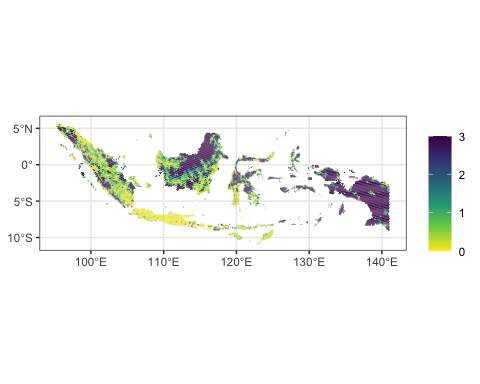
| plotid | user\_code | collection\_time | lon | lat | lu\_cat |
| --- | --- | --- | --- | --- | --- |
| 1 | user43 | 2023-10-24 02:31:00 | 124.27248 | -9.710096 | Other Land |
| 2 | user13 | 2023-10-23 07:14:00 | 116.03786 | 2.441534 | Forest |
| 3 | user16 | 2023-10-23 06:21:00 | 100.38283 | 2.278349 | Other Land |
| 4 | user02 | 2023-10-24 02:40:00 | 119.49038 | -3.071154 | Forest |
| 5 | user33 | 2023-10-23 05:51:00 | 136.32690 | -4.254708 | Forest |
| 6 | user08 | 2023-10-24 02:46:00 | 107.65248 | -6.730734 | Other Land |
| 7 | user08 | 2023-10-24 02:51:00 | 113.20639 | -7.039453 | Other Land |
| 8 | user44 | 2023-10-23 08:56:00 | 97.69162 | 1.073404 | Other Land |
| 9 | user14 | 2023-10-23 06:08:00 | 113.51342 | 0.951133 | Forest |
| 10 | user28 | 2023-10-23 07:11:00 | 126.42722 | -3.317357 | Forest |

* The results of the QAQC dataset with the percentage of disagreement between 2 observers on land use category (CSV file).

Evaluation of agreement between observers on land use category in the NFI Phase 1 QAQC dataset.

| pl\_island | agree | disagree | total | perc |
| --- | --- | --- | --- | --- |
| BaliNusa | 94 | 38 | 132 | 29 % |
| Jawa | 169 | 23 | 192 | 12 % |
| Kalimantan | 231 | 44 | 275 | 16 % |
| Maluku | 108 | 25 | 133 | 19 % |
| Papua | 223 | 32 | 255 | 13 % |
| Sulawesi | 287 | 31 | 318 | 10 % |
| Sumatera | 230 | 44 | 274 | 16 % |
| Total | 1342 | 237 | 1579 | 15 % |

* The Discrete Global Grid hexagons matching the CEO NFI Phase 1 plot locations. For mapping purposes, the spatial information is easier to display on surface accurate polygons rather than points, because point data tends not to scale well with image resolution.



Plot accessibility based on Weiss and al, 2020 motorized travel time to healthcare in hours. The 3 hours color code represents 3 hours and more.

## Data analysis workflow

The analysis workflow consisted in (1) setting rules for plot accessibility to determine the existing pool of accessible forest plots, then (2) running the plot selection based on manually taylored sampling proportion to reach the target number of plots. The plot selection process started with assigning a random starting point at the begining of the space filling curve then selecting the following plots based on the sampling proportion. For example, with a sampling proportion of 0.1, one plot is selected every 10 plots.

### Options to determine plot accessibility

At the end of the NFI Phase 1 visual interpretation, the preliminary results showed that in Papua, and probably also in Kalimantan, the accessibility by river was too loosely defined. It resulted in many plots being assessed as accessible by river because a river was found within 5 km to plot centers but in reality these rivers might not be navigable or the travel time to plot center form the nearest harbor would be too long. Three options were therefore tested to address this issue:

1. Access by road, river or sea based on the NFI Phase 1 visual interpretation, i.e. a road, river or sea is within 5 km of the plot center.
2. Restricted version where access by road only for Kalimantan and Papua. This more restrictive definition is due to many plot having a river nearby but the journey to reach the plots would be too long, or the river is not navigable.
3. Using the [Global Motorized Travel time to Healthcare](https://malariaatlas.org/) friction spatial data to define accessibility. This data is based on road infrastructure and healthcare facilities databases from Google map, Open Street Map and a number of other datasets (See [Weiss and al., 2020](https://doi.org/10.1038/s41591-020-1059-1)) to evaluate how far in minutes each pixel is from a healthcare facility. Note that this data show fewer accessible plots that

* Create a table of space filling curve input parameters for each strata

## Results