Area assessment and uncertainty estimation

Spatial Informatics Group SilvaCarbon

Method rationale

All **maps have errors**, and sometimes they can be significant.

Those errors make pixel-counting unreliable, leading to inaccurate estimates compared to rigorous statistical analysis.

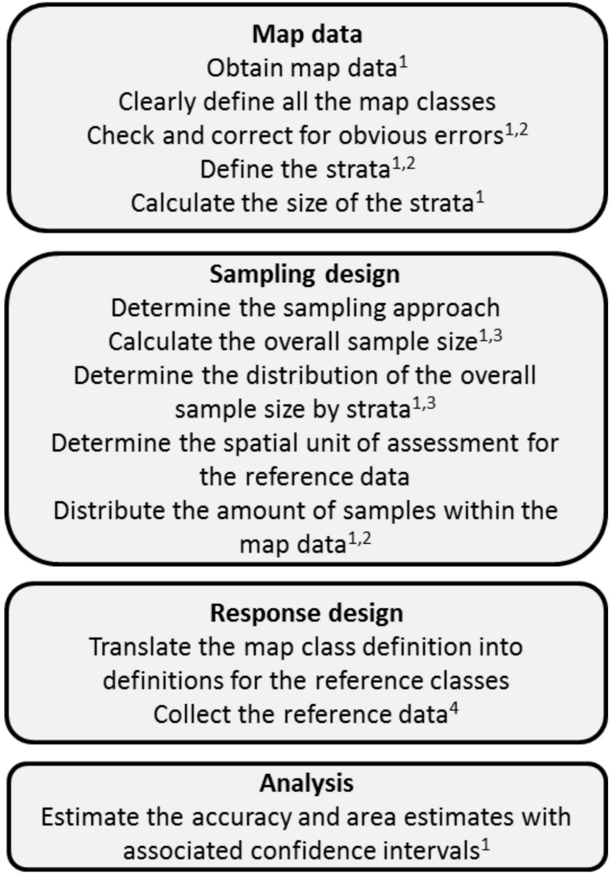
Recommended is a **sample based approach** because it:

- makes **estimates statistically defensible** and so that remote sensing can inform policy and decision-making.
- provides a statistically sound method to **estimate area and quantify uncertainty.**
- satisfies two good practice criteria set by the Intergovernmental Panel on Climate Change (IPCC) for use in Greenhouse Gas (GHG) inventories, which are:
  - **Not over- or under-estimating**, as far as possible.
  - **Reducing uncertainties** to the extent practicable.

Method overview

The **four main components** of sample based area estimation are:

- **Map Definition**: Identify **classes and strata** based on the characteristics of the target area.
- **Sampling Design**: Choose a **representative subset** of the map for collecting reference data.
- **Response Design**: Define **guidelines for reference data collection**, including type of data, used equipment, and frequency of collection.
- **Analysis**: Estimate area by **comparing remotely sensed and reference data**, adjusting the sampling and response designs if necessary.



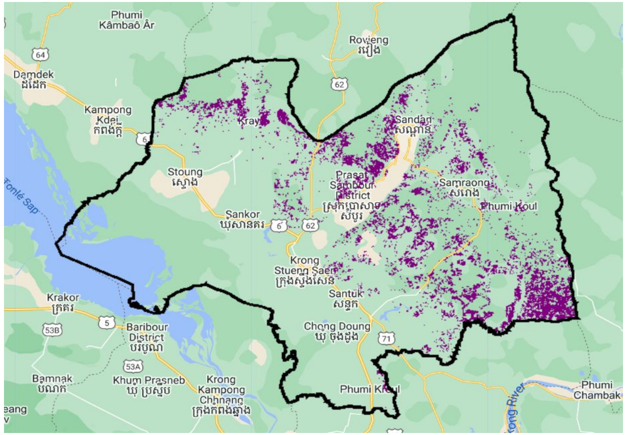
Finegold et al., 2016 (Fig 2)

Map definition

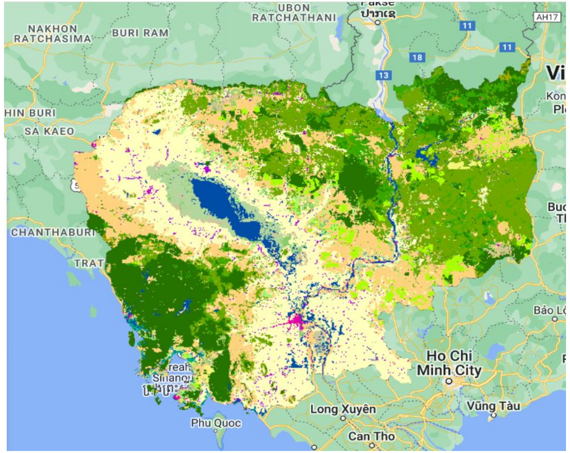
The main aspect of the "map definition" step is the definition of Strata.

Strata are "subpopulations that are non-overlapping, and together comprise the whole population" (Cochran, 1977)

In our case we use the map of **plantations** and combine it with a **national land cover** map to define strata.



Example of mapped crops



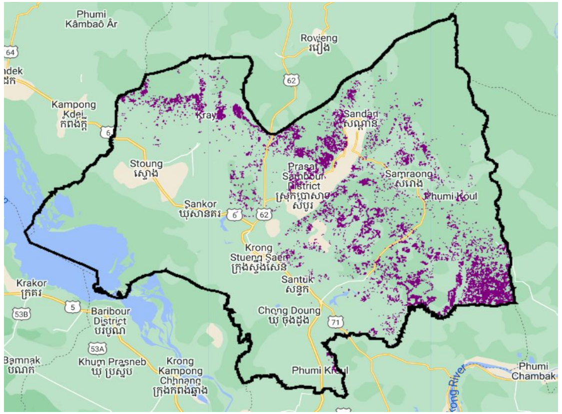
Land cover in Cambodia

Map definition

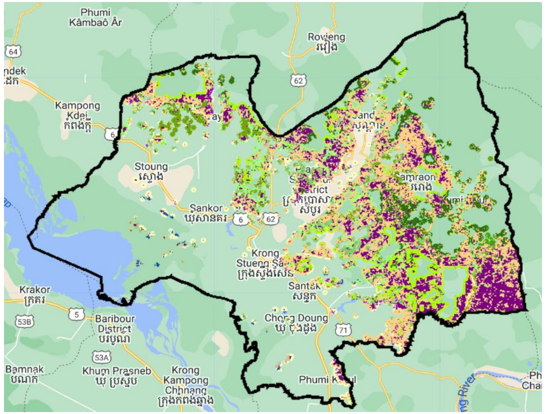
Given limited presence of plantations, there is a high **risk of omission errors**.

To reduce this error, we can focus sampling on **smaller areas with a higher target class proportion**. We do this by considering a buffer zone:

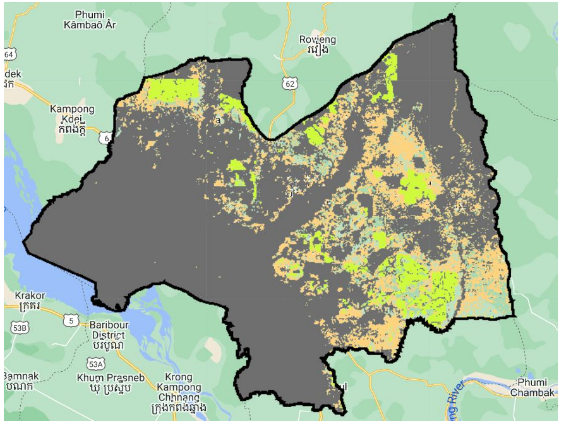
The **target land cover** (in this example cashew)



The **main land cover types in a buffer zone** (rubber cropland, other plantations)



Land cover types outside the buffer zone (rubber, cropland, “everything else”)



Sampling design

Sampling design is a crucial aspect of research that aims to select a **representative sample** from the population while **balancing cost considerations**

It should consider:

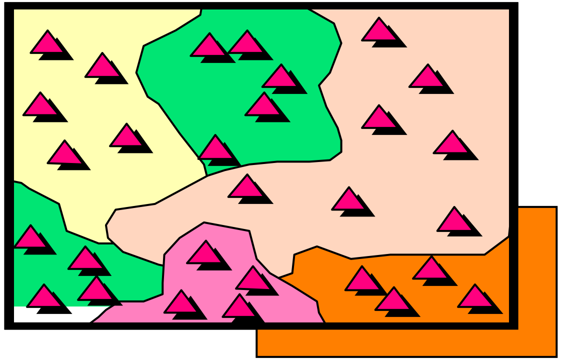
- **Randomization**: can include simple random, systematic, and stratified random methods.
- **Inclusion probability**: refers to each unit having a known, non-zero probability of being selected.
- **Sample size**: should be determined based on setting an error target.

Sampling design

Stratified random sampling is a technique that involves dividing the population into **homogeneous strata**, and **randomly selecting samples from each stratum**.

- Benefits of stratified random sampling include reducing sampling error and increasing estimate precision by **ensuring that each stratum is proportionally represented in the sample**.

- Unlike some other random sampling methods, the stratified random sampling technique **ensures that rare or important subgroups are sampled**



Sampling design

The **area for each defined stratum** is calculated to determine the number of points to be sampled per stratum.

These calculations are based on pixel counting and provide estimates that cannot be used as final area estimates for the target land cover type.

However, **these estimates can inform the sampling strategy** by indicating the relative size of each stratum and how many points should be sampled from each one.

area cashew JSON
111871.9060288113

area buffer JSON
124476.28416538882

area rubber crop JSON
132244.86788138762

area other JSON
827686.3757508714

Calculating ideal sample size

Based on **how large the area of each stratum is** via “pixel counting” method

Calculating Ideal Total Sample Size

Strata of Agreement Map		Cashew	buffer zone	rubber crop	other
variable name	*equation*				
pixel count	*only use when pixels have same resolution*				
pixel area in ha	*GEE output*	111,871	124,476	132,244	827,686
wi (Area Weight)	AreaCount / AreaSum	0.094	0.104	0.111	0.692

Spreadsheet to determine sampling size per classes

Defining ideal sample distribution

… and based on **the proportion of each strata in the land cover map**

Sample Distribution

Strata of Agreement Map	*equation*	Cashew	buffer zone	rubber crop	other
proportional % distribution	wi as a percent (*100)	9%	10%	11%	69%
proportional distribution of points	proportional % distibution * n [rounded to integers]	38	42	45	282
% distribution without "other" strata	excluding the smallest classes, % distribution or remaining pixel counts [row2value/(sum of larger row2values)]	30.35%	33.77%	35.88%	
minimum sample size	*defined*	60			
point distribution with minimum sample size	*Update this equation if calculating for more than 1DEG, forest, and nonforest* (n-[MinimumSampleSize*#SmallStrata])*%DistributionWithoutSmall	106	118	125	60

Defining ideal sample distribution

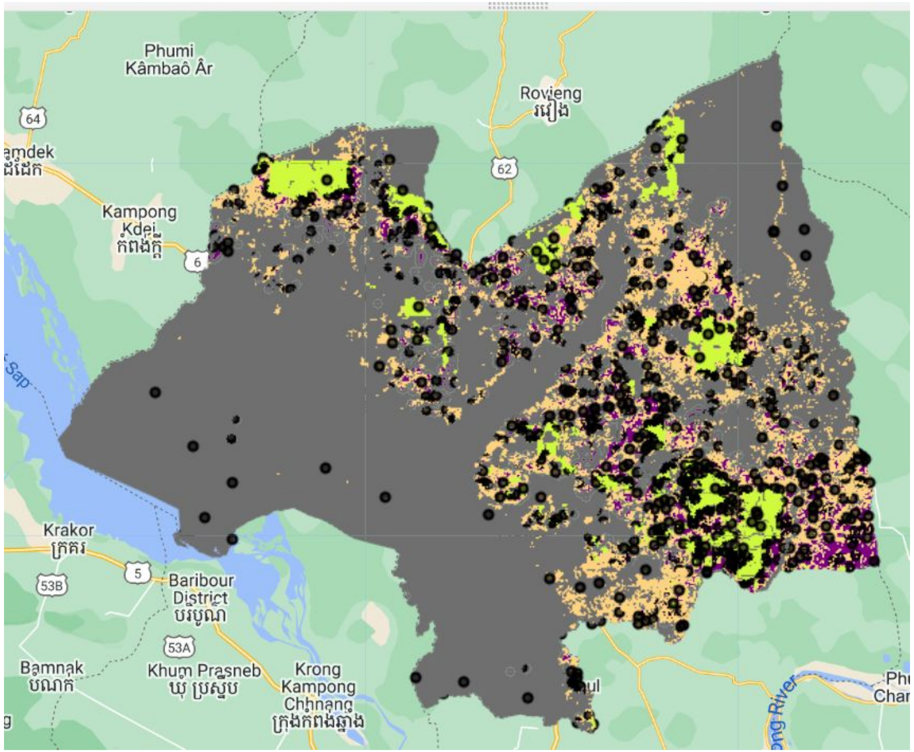
… we **define the number of sampling points** we will randomly select from each stratum

Sample Distribution					
Strata of Agreement Map	*equation*	Cashew	buffer zone	rubber crop	other
proportional % distribution	wi as a percent (*100)	9%	10%	11%	69%
proportional distribution of points	proportional % distibution * n [rounded to integers]	38	42	45	282
% distribution without "other" strata	excluding the smallest classes, % distribution or remaining pixel counts [row2value/(sum of larger row2values)]	30.35%	33.77%	35.88%	
minimum sample size	*defined*	60			
point distribution with minimum sample size	*Update this equation if calculating for more than 1DEG, forest, and nonforest* (n-[MinimumSampleSize*#SmallStrata])*%DistributionWithoutSmall	106	118	125	60
	One QA/QC option, selecting				

Creating sampling points

The resulting **sampling points** will cover land cover types across the study area.

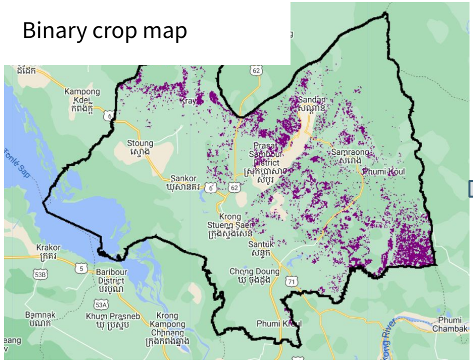
We aim to reduce omission error within the target class by **focusing on land cover in close proximity to the target class** (buffer zone) and land cover types with expected confusion.



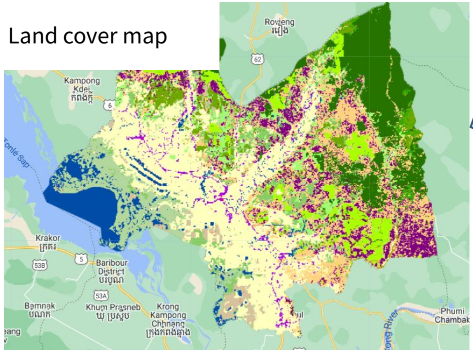
Creating sampling points

By using this method, we create a **random stratified sample** based on a **map of the target crop** and a **map of all other land cover types**

Binary crop map



Land cover map



Sample distribution

Update values for all yellow cells, the pixelcounts. Important results are in blue.
pi values were estimated from ha estimation based on pixel count within 200m buffer area of mapped cashew
S(p) was chosen as 0.01, a generally accepted target standard error, but this could be adjusted.

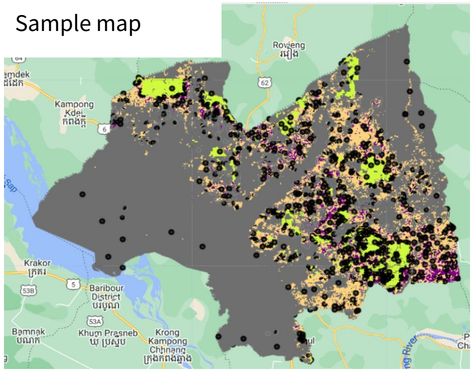
Calculating Ideal Total Sample Size

Strata of Agreement Map		Cashew	buffer zone	rubber crop	other
variable name	*equation*				
pixel count	*only use when pixels have same resolution*				
pixel area in ha	*GEE output*	111,871	124,476	132,244	827,686
wi (Area Weight)	AreaCount / AreaSum	0.094	0.104	0.111	0.692
pi (Proportion of target (cashew) in stratum; estimated from unofficial CAC data)	*defined*	0.850	0.300	0.050	0.020
Si (Standard error for stratum)	sqrt(pi*(1-pi))	0.357	0.458	0.218	0.140
S(p) (Target standard error)	*defined*	0.010			
numerator coefficients	wi*Si	0.033	0.048	0.024	0.097
numerator	sum of numerator_coefficients	0.202			
denominator	S(p)	0.010			
n (Target Sample Size)	(numerator/denominator)^2	408			

Sample Distribution

Strata of Agreement Map	*equation*	Cashew	buffer zone	rubber crop	other
proportional % distribution	wi as a percent (*100)	9%	10%	11%	69%
proportional distribution of points	proportional % distibution * n [rounded to integers]	38	42	45	282
% distribution without "other" strata	excluding the smallest classes, % distribution or remaining pixel counts [row2value/(sum of larger row2values)]	30.35%	33.77%	35.88%	
minimum sample size	*defined*	60			
	Update this equation if calculating for more than 1DEG, forest, and nonforest				
point distribution with minimum sample size	(n-[MinimumSampleSize*#SmallStrata])*%DistributionWithoutSmall	106	118	125	60
15% duplicate review points	One QA/QC option, selecting proportional 15% from strata for duplicate review	16	18	19	9
point distribution for single interpretation	total points distributed - 15% duplicate review points	90	100	106	51

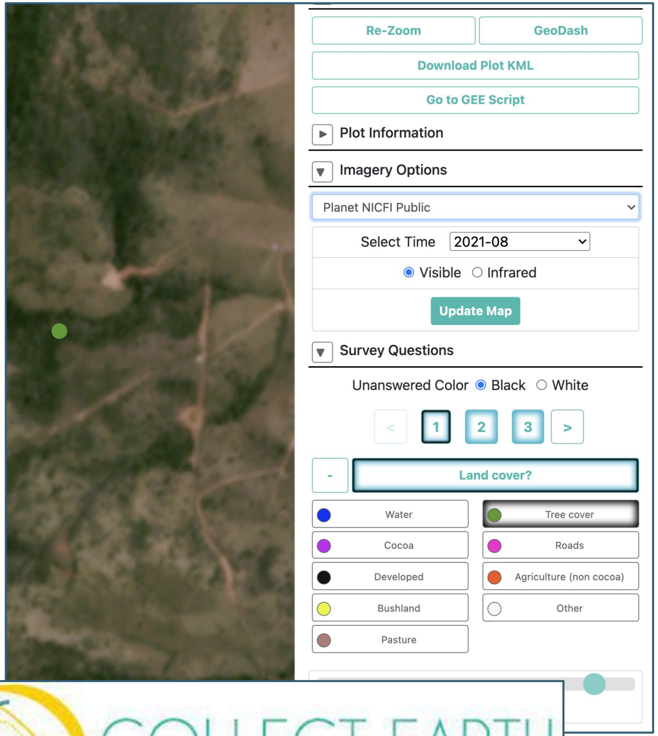
Sample map



Response Design

The main considerations for our response design are:

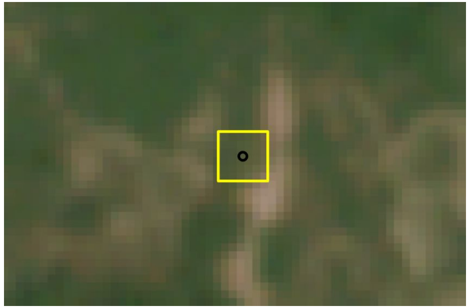
- Determining the **assessment unit** (i.e. the spatial extent) and create a frame for sampling.
- Identifying **data sources** (e.g., high-resolution imagery, Landsat data, field data).
- Developing a **labeling protocol** and classification schema.
- Defining an **agreement protocol** for interpreter consistency and label quality assurance
- Assessing **reference classification uncertainty** (e.g., geolocation, interpreter variability).

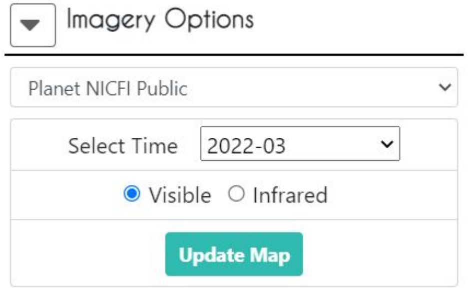


COLLECT EARTH ONLINE

Response Design

- Labelling of sampling points is performed in **Collect Earth Online** (CEO), which displays them over satellite images.
- We need to select the images taken in the year for which you want to create the crop map.
- Images from other dates can be used as additional reference.





Analysis

The two components of the analysis are:

- Creating an **error matrix** to evaluate the accuracy of the land cover map, including the proportion of correctly and falsely classified land covers.

- Estimating the **unbiased area of the target land cover** by taking the error matrix into account, which provides a more accurate estimation of the area than the initial land cover map.

		Reference Data		
		CH	NCH	ROW TOTAL
Map Data	CH	27	6	33
	NCH	4	63	67
	COLUMN TOTAL	31	69	100

Analysis

The confusion matrix **compares the mapped land cover types with the labels assigned** by interpreters at the reference points. It shows the **agreement** (green fields) and **disagreement** (yellow fields) frequency between the two.

It insight into the **quality of the mapping** method and helps to identify areas that need further refinement.

Sample Points Comparison Table	Reference Data				
	Cashew	Rubber	Other Cropland	Other Plantation	Other
Cashew	34	1	0	2	5
Rubber	2	73	0	1	0
Other Cropland	0	0	60	0	2
Other Plantation	0	0	0	100	1
Other	0	0	0	4	185
Sum	36	74	60	107	193

Spreadsheet on creating a estimating uncertainty per class

Analysis

To **estimate the area of a target land cover class** we can apply a **stratified estimator** by Cochran (1977, Eq. 5.1).

- It incorporates the **confusion matrix** and each stratum area to **estimate the true area of the target land cover class**.
- This method accounts for accuracy variability in the land cover map, improving area estimation precision.

$$\hat{\mu}_{STR} = \sum_{h=1}^{3} W_h \frac{n_{hi}}{n_h} = \sum_{h=1}^{3} p_{hi} = 0.0033$$

Stratum, h	Reference observation = Forest	Non-forest	Deforest.	Total	Weight*, W_h
h = 1 Forest	0.8781	0.0044	0.0022	0.8847	0.8847
h = 2 Non-forest	0.0137	0.1004	0.0000	0.1141	0.1141
h = 3 Deforestation	0.0001	0.0000	0.0011	0.0012	0.0012
Total	0.8918	0.1049	0.0033	1	1

*Made up example but strata weights from Arevalo et al. (2018)

Analysis

The spreadsheet below can be used to **calculate uncertainties and unbiased area estimates for the target class** by applying the stratified estimator formula and incorporating the confusion matrix values.

The spreadsheet automatically computes the confusion matrix, and the **estimated area with a confidence interval**, ensuring precise and reliable area estimate of the target land cover.

		Cashew	Rubber	Other Cropland	Other Plantation	Other	
Uncertainty Estimation	V(μ)	0.0000936	0.0000160	0.0000212	0.0000643	0.0001203	Variance Estimator
	SE(μ)	0.0096770	0.0040047	0.0046027	0.0080168	0.0109687	Standard Error
	±95% CI	0.0189668	0.0078492	0.0090214	0.0157130	0.0214987	±95% Confidence Interval
	MoE [%]	14.8%	11.9%	4.6%	76.7%	3.7%	Margin of Error
Area Estimation	Category Area (ha)	105,255.9	54,261.1	161,330.3	16,779.8	481,730.9	Area Estimate in hect
	±95% CI (ha)	15,540.6	6,431.3	7,391.7	12,874.6	17,615.1	±95% Confidence Interval in hectares

Total Study Area	819,358.0	ha