Pontus Olofsson *(olofsson@bu.edu)*
Earth & Environment, Boston University

BOSTON UNIVERSITY

3. Sampling Design

Training module: Use of global tree cover and change datasets in REDD+ Measuring, Reporting and Verifying. Boston, 2015.

3.1 Introduction

In this Module we will use a map of land cover and land cover change for stratifying a random sample with the aim of estimating the area of forest change. We will also use the map for estimating accuracy of the map classes. Any map can be used but the instructions will refer to the map that was extracted from the global dataset described in the second Module. The sampling design is the protocol for selecting the subset of spatial units (e.g., pixels) that will form the basis of the analysis of area and accuracy. Is it recommended that the sampling design is a probability sampling design, which incorporates randomization in the selection protocol and is defined in terms of inclusion probabilities such that the inclusion probability is known and greater than zero for each unit in the sample. A variety of probability sampling designs are applicable, with the most commonly used designs being simple random, stratified random, systematic and clustered. When choosing a design, three main decisions are whether to use clusters, whether to use strata, and whether to use a systematic or simple random protocol. The primary motivation for cluster sampling is to reduce the cost of data collection – for example, if the map is large and high resolution data need to be collected for each unit in the sample, a clustered design will allow for collection only for the primary sampling units and not for the entire population (cluster designs as defined in this text include 2-stage designs where the first and second sampling stages include selection of primary and secondary sampling units, respectively). However, the use of clusters is recommended only if cost savings or practical advantages are substantial as it results in a more complex analysis and because the potential correlation among units within a cluster (i.e., intracluster correlation) often reduces precision relative to a simple random sample of equal size. The use of strata is usually motivated by the fact that activity data is small proportion of the total map and if not stratifying the sample, a very large sample might be required to implement the analysis.

That map that was created in Module 2 contains seven classes (*no data*, *water*, *non-forest*, *forest cover*, *forest cover loss, forest cover gain* and *forest cover gain/loss*) but the theory and methodology is generic and could be applied to any thematic map regardless of how the map was made and regardless of the nature and number of map categories. As the aim is to estimate the area of forest change, it is recommended to use the map classes as strata. This will ensure that a sufficient sample size for estimation can be allocated to the change classes. The stable forest class is defined as the percentage canopy closure for all vegetation taller than 5 m in height, and a threshold is required to create a forest and non-

forest stratum. The estimation approach we will employ is called **stratified estimation** (Cochran , 1977) and has proven useful for estimating area of discrete map categories and the uncertainty of area estimates (Olofsson et al., 2013; Stehman, 2013). In short, we will use the map as a source of **stratification** for the sample. This is important as the area of forest loss is typically small relative the total map area and a good stratification allows for more precise estimation. Without stratification it might be hard to obtain enough units in small categories to allow for inference of area and accuracy. Note that stratified estimation can be used with simple or systematic random samples too.

Once we designed the sample and a stratified random sample is drawn, it needs to be interpreted using a suitable source of reference data. This step is referred to as the response design and is described in Module 4. With each unit having a map label and a reference label we can construct an **error matrix**, which contains all the information needed to perform the analysis (Module 5).

3.2 Determine sample size and allocation

1. Display your map in QGIS by clicking *Layers* > *Add Raster Layer*.
2. Color it if you haven't already: right-click the map in the layer pane and click *Properties* > *Style*; set *Render Type* to *Singleband pseudocolor*; click the green plus-sign 7 times and set values to 1-7, and give each category an appropriate name and color.
3. Determine the areas of each map category: open a terminal, navigate to you directory and type: `gdalinfo -hist stratification_massachusetts_utm.tif`
4. This gives the number of pixels of each map class; in the New Brunswick example, gdalinfo gives the following areas in pixels (third row percent, calculated from pixels):

	Non-forest	*Forest*	*Water*	*Forest loss*	*Forest gain*	*For. loss/gain*
Area	47,996	228,551	13,795	3,561	293	87
W_i	16.3%	77.6%	4.69%	1.21%	0.10%	0.03%

5. To determine the sample size for a stratified random sample, we will use Eq. 5.25 in Cochran (1977): $n \approx \left(\frac{\sum W_i S_i}{S(\hat{P})}\right)^2$ where W_i is the stratum weight and S_i is the standard error for stratum *i*; the latter is estimated as $\sqrt{p_i(1-p_i)}$ where p_i is the proportion of forest loss in stratum *i*. $S(\hat{P})$ is the target standard error of the forest loss estimate. If assuming one error of omission of forest loss in *non-forest*, *forest*, and *forest loss/gain* per 100 units and a user's accuracy of 0.8 and a target standard error of the forest loss estimate of 0.5% (i.e. a confidence interval of 1%); we get following information for determining the sample size:

	Non-forest	*Forest*	*Water*	*Forest loss*	*Forest gain*	*For. loss/gain*
p_i	0.01	0.01	0	0.8	0	0.01
S_i	0.099	0.099	0.000	0.400	0.000	0.099
$S(\hat{P})$				0.005		

This in turn gives: $n \approx \left(\frac{\sum W_i S_i}{S(\hat{P})}\right)^2 = \left(\frac{0.098}{0.005}\right)^2 = 387$ (note that this is just an example and users need to specify their own target errors and expected accuracy and omission errors).

6. The second step is to determine how to allocate these units to strata. Good practices stipulate that 50, 75 or 100 units are allocated to the smaller classes depending on the total sample size and that the rest is proportionally allocated to the larger strata. In this all strata are small relative forest and the sample is allocated to strata as (*Forest gain* and *Forest gain/loss* are so small fractions of the map that we will not attempt to estimate them; if they were larger they would be included):

	Non-forest	*Forest*	*Water*	*Forest loss*	*Forest gain*	*For. loss/gain*
n_i	55	230	50	50	0	0

3.3 Select sample

1. QGIS does not have built-in tools for drawing samples (this hold true also for most propretiary software) so we need to make use of Python script: copy the script sample_map.py from "~/Documents/scripts/bin" (or download it from https://raw.githubusercontent.com/ceholden/accuracy_sampler/master/script/sample_map.py) to the "3_sampling_design" directory.
2. In the "3_sampling_design" directory, type: `python sample_map.py -v --size 385 --allocation "55 230 50 50 0 0" --vector sample.shp stratified stratification_massachusetts_utm.tif` to select a stratified random sample with a sample size of 385 pixels allocated as in the table above.
3. This will create a shapefile "sample.shp" that contains the sample. **Note:** if the script halts with the message "MemoryError", the memory allocation when starting the Virtual Machine needs to be increased (in *Oracle VirtualBox Manager*: *Settings* > *System* > increase *Base Memory* before launching the VM).