Pontus Olofsson, Christopher E. Holden, Eric L. Bullock
Boston Education in Earth Observation Data Analysis/
Department of Earth & Environment, Boston University

Use of global tree cover and change datasets in REDD+ Measuring, Reporting and Verifying.

Introduction

The global dataset contains several layers that need to be merged to create a map that can serve as a stratification for the sample. The layers of interest are tree cover, forest loss, forest gain, forest loss and gain, and the data mask that shows unmapped areas and water bodies. The tree cover layer is continuous and each pixel represents the percentage of tree canopy closure with trees being defined as vegetation taller than 5 m in height, and a threshold is required to create discrete forest and non-forest strata. This is a non-trivial task and the reader is referred to MGD Module 2 for further guidance. The global data is delivered in granules of 10 by 10 degrees and several granules might need to be downloaded to cover the area of interest. If the study area is a country it is likely that the user wants to clip the map according to the country outline. In this case, a vector file showing the outline of the country is required. **Note:** For this example, a small subset is provided. In the instructions below, menus, buttons and other components of the graphical user interface are italicized while filenames and directories are put in quotation marks.

Part 1. Install Software

1. Install Virtual Machine

1. Download a binary of *Oracle VirtualBox* compatible for your operating system from https://www.virtualbox.org/wiki/Downloads; follow the instructions to install.
2. Connect the USB stick containing the virtual machine file.
3. Open *VirtualBox* and click *File* > *Import Appliance* and browse to the USB containing the ".ova" file. Click *Continue* and then *Import.*
4. Before launching the VM, it is necessary to pay attention to the memory allocation: In *VirtualBox*, highlight the VM in the left-hand pane and click *Settings* (yellow cogwheel) > *System*> *Motherboard* tab > increase *Base Memory* on the scale line; the amount of allocated memory will depend on the host computer but it is recommended to use slightly more than half of the memory of the host computer (upper part of the green part of the *Base Memory* scale line).

5. In order to copy and paste between your host computer and the virtual machine go to *Settings* > *General* > *Advanced* and change both ‘Shared Clipboard’ and ‘Drag’n’Drop’ to ‘Bidirectional’.
6. Another useful feature is *Shared folders* which are directories on the host computer or mounted devices that can be accessed from within the VM (although not technically not the same thing, in this and other modules the terms *folder* and *directory* means the same thing). You can add shared directories from *Settings* > *Shared Folders*.
7. In *VirtualBox*, highlight the VM in the left-hand pane and click *Start* (green arrow). This will launch the VM; the username and password are both “opengeo-vm” (without quotation marks).
8. To view in fullscreen mode click *View* > *Switch to Fullscreen* in the VM top menu.
9. The interface of the VM can be changed to style of Windows and OS X and various Linux distributions by the *Menu* in the upper left corner > *Preferences* > *MATE Tweak* > *Interface* (the Windows-style interface is referred to as Redmond and the OS X interface as Cupertino according to the locations of the headquarters of Microsoft and Apple).
10. Some of the more useful applications are the *Caja* file browser (similar to the file browsers in Windows and OS X), the *MATE Terminal* which provides text-based access to the operating system, *QGIS* which is the main graphical user interface used in this training material, the *LibreOffice* suite which is similar to *Microsoft Office*, and *Atom* or *Pluma* both ASCII text editors. You can add these software to the quick launch panel by right clicking the name in the menu > *Add to panel.*
11. Make yourself acquainted to the VM and these software.

Part 2. Download and prepare data

2. Extract data

1. Create a directory on the desktop named ‘Hansen_Global”.
    a. Right-click on the desktop -> “Create Folder”.
2. **Skip this step.** To work on a specific area of interest, in the Firefox Web Browser in the virtual machine, go to http://earthenginepartners.appspot.com/science-2013-global-forest/download_v1.2.html and identify a granule that you would like to download.
3. **For this exercise:** From the BEEODA Github repository, download a subset of the Global Forest Loss dataset to the “Hansen_Global” folder.
    a. The easiest way to do this is using ‘wget’ from the command line.
    b. Open the MATE terminal.
    c. To navigate to a directory you can use the ‘cd’ command (**c**hange **d**irectory).
    d. Type the following command to navigate to the “Hansen_Global” directory:

```
cd Desktop/Hansen_Global/
```

e. Copy and paste the following command to download the script:

```
wget -q
"https://github.com/beeoda/tutorials/raw/master/6_Global_Tree_Cover/Hansen_Peru_Su
bset.zip"
```

f. Unzip the file either in the folder browser or in the terminal with:

```
Unzip Hansen_Subset_Peru.zip
```

3. Create stratification

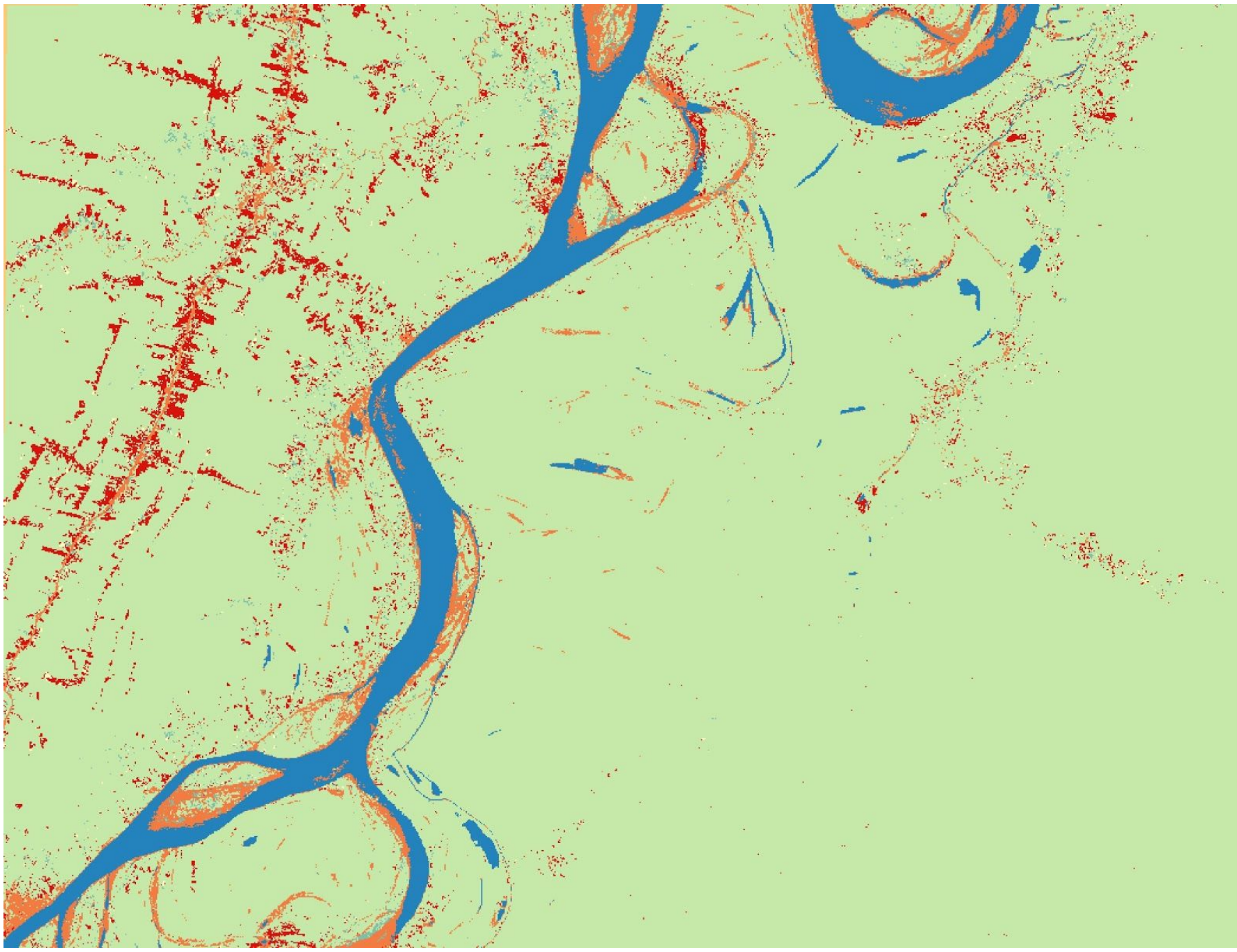
1. Within the virtual machine, start the QGIS Desktop
2. Clicking Layer > Add Raster Layer and navigate to the download directory and select the files “treecover2000*”, “*datamask* *gain* *loss*”; rename these to “treecover”, “loss”, “gain” and “mask” by right clicking the layers in the Layers panel > Rename.
3. Open the Raster calculator from the Raster menu to create the final stratification. Make an expression that makes map where the map classes have the following values:

   No data 0
   Non-forest 1
   Forest 2
   Water 3
   Forest loss 4
   Forest gain 5
   Forest loss/gain 6

   The expression should be the following (you can copy and paste into the Raster calculator expression) and specify an output file name in the stratification directory “/home/opengeo-vm/Desktop/Hansen_Global” (call it “stratification_0N08W” for example). (Note that in this example forest is defined as 10% tree cover – please specify a threshold for the tree cover.)

```
("mask@1" = 0)*0 + (("mask@1" = 1) AND (("treecover@1" > 10) = 0)
AND ("gain@1" = 0) AND ("loss@1" = 0))*1 + (("mask@1" = 1) AND
(("treecover@1" > 10) = 1) AND ("gain@1" = 0) AND ("loss@1" = 0))*2
+ ("mask@1" = 2)*3 + (("mask@1" = 1) AND ("gain@1" = 0) AND
("loss@1" = 1))*4 + (("mask@1" = 1) AND ("gain@1" = 1) AND ("loss@1"
= 0))*5 + (("mask@1" = 1) AND ("gain@1" = 1) AND ("loss@1" = 1))*6
```

4. Right-click the newly created stratification in the Layers panel and click Properties > Style. In Render type select Singleband psuedocolor, click the plus sign 7 times and give each class a name according to the list of classes in point 7 and an appropriate color. Click OK. It should look like the screenshot below.



4. Project

1. The data is delivered in angle-correct projection and we want to reproject it to an area correct projection: click Raster > Warp; set input file “stratification_0N08W” and output to “stratification_0N08W_utm”; set Source SRS to “EPSG:4326” (WGS 84); set Target SRS to “ESPG: 26918” (UTM Zone 18 N); click Edit at the bottom right (pencil) and add -tr 30 30 (to specify pixel size) such that the expression reads:

```
gdalwarp -overwrite -s_srs EPSG:4326 -t_srs EPSG:26918 -tr 30 30 -of GTiff
/home/opengeo-vm/Desktop/Hansen_Global/stratification_0N08W.tif
/home/opengeo-vm/Desktop/Hansen_Global/stratification_0N08W_utm.tif
```

2. You have now created a stratification that you will use in the next section, Sampling Design.

Part 3. Sampling Design

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, nonforest, 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- 2 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 the following section. 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.

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 your directory and type: `gdalinfo -hist stratification_0N08W_utm.tif.` Note: Since the image is automatically 32-bit instead of 8, there will be many bins with 0 values. Ignore those, and focus on the ones with values.
4. This gives the number of pixels of each map class; in the Peru example, gdalinfo gives the following areas in pixels (third row percent, calculated from pixels):

Label	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Class	No Data	Non-Forest	Forest	Water	Forest Loss	Forest Gain	Forest Loss/Gain
Pixels	30098	67188	2129531	124067	67306	13987	11364
Area (ha)	90.29	201.56	6388.59	372.2	201.92	41.96	34.09

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 Wi is the stratum weight and Si is the standard error for stratum i; the latter is estimated as $\sqrt{p_i(1-p_i)}$ where pi is the proportion of forest loss in stratum i. In the following example: sqrt(.8*(1-.8)) = .4. $S(\hat{P})$ is the target standard error of the forest loss estimate. If assuming one error of omission of forest loss in

non-forest and forest 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:

Label	0	1	2	3	4	5	6	Total
Class	No Data	Non-Fores t	Forest	Water	Forest Loss	Forest Gain	Forest Loss/Gain	
Area (ha)	90.29	201.56	6388.59	372.2	201.92	41.96	34.09	7296.52
Wi(%)	1.2	2.7	87	5.1	2.8	0.6	0.5	1
Pi (%)	0	.01	.01	0	.80	0	0	
Si	0	.0995	.0995	0	0.4	0	0	
S(P^)(%)					1			

6. This in turn gives: $n \approx \left(\frac{\sum W_i S_i}{S(\hat{P})}\right)^2$ = (.1 / .005)^2 = 400
7. 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):

Label	1	2	3	4	5	6	Total
Class	Non-For est	Forest	Water	Forest Loss	Forest Gain	Forest Loss/Gain	
Allocation	50	225	75	50	0	0	400

3. Select Sample

1. QGIS does not have built-in tools for drawing samples (this hold true also for most proprietary software) so we need to make use of Python script: copy the script

sample_map.py from “~/Desktop/scripts/bin” (or download it from https://raw.githubusercontent.com/ceholden/accuracy_sampler/master/script/sample_map.py) to the “Hansen_Global” directory.

2. In the “Hansen_Global” directory, to select a stratified random sample with a sample size of 400 pixels allocated as in the table above, type (note: the first 0 is for the no data class):

```
python sample_map.py -v --size 400 --allocation "0 50 225 75 50 0 0" --vector sample.shp
stratified stratification_0N08W_utm.tif
```

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).

Part 4. Response Design

1. Introduction

At this stage a sample has been selected which needs to be interpreted using a suitable source of reference data. This step is referred to as the response design and includes providing reference labels for each unit in the sample. 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 (Section 5). The use of TimeSync (http://timesync.forestry.oregonstate.edu/) is recommended, especially if the stratification covers a large area as Landsat data can be retrieved without having to download it.

2. Interpret sample. Note for this exercise, to save time, skip to section 2b.

1. Display the reference data in QGIS, i.e., display the data you will use to interpret the sample you just created. This is likely a combination of different data sources, such as Landsat, RapidEye and Google Earth, acquired around the same times as the data used to create the map (in this case 2000 and 2012), and preferably also in-between. Do not display the map!
2. Display the shapefile containing the sample. You may need to define the projection if you do not see the sample points. Right click on the sample and go to ‘Properties’ and choose the projection (in this case it is EPSG:32618) under ‘Coordinate reference system’.
3. Right-click shapefile in Layer pane; Open Attribute Table; then and then ; delete the STRATUM column.

4. Click the New column button to add a column; name it “reference”; leave options as default except Width which should be set to 3.
5. Now provide a label for each of the units in the sample by manually examining the reference data. Add label
   “1” for stable non-forest
   “2” for forest
   “3” for water,
   “4” for forest loss,
   “5” for forest gain (note that this stratum does not exist in this example),
   “6” for forest loss/gain (note that this stratum does not exist in this example).
   Since your final area estimates are based on the interpretation of this sample it is important that the labels are correct – if you can’t provide a correct label then delete the unit rather than guessing.
   You can click to jump to the highlighted unit. Make sure you save the shapefile regularly.
6. If you want to open the sample in Google Earth TM, right click the shapefile with the sample > Save As... > in the Save As dialog, set Format to Keyhole Markup Language [KML], specify an output file and set NameField to ID; leave other options as default > click OK . You can also use the GDAL program “ogr2ogr” (www.gdal.org/ogr2ogr.html) to create the KML file: either paste the following into the terminal: ogr2ogr -f "KML" test_ge.kml test.shp -dsco NameField=ID
7. In addition to (or instead of) the workflow described above, the use of the TimeSync software is highly recommended (http://timesync.forestry.oregonstate.edu/). TimeSync is by the time of writing not fully operational but is expected to be soon. The main benefit of TimeSync is the retrieval of annual Landsat data for each unit in the sample without having to download the data.

2B. Download pre-validated results.

1. Due to time constraints, we will skip the validation step and move on with a pre-validated sample.
2. Go to the “Hansen_Global” folder in the terminal and download the validated result:

```
wget -q
"https://github.com/beeoda/tutorials/raw/master/6_Global_Tree_Cover/Validated_Sampl
e.zip"
```

3. Unzip the shapefile:

```
Unzip Validated_Sample.zip
```

3. Construct the error matrix

1. With each unit having a map label and a reference label we can construct an error matrix. This can be done in various ways but we recommend using a home-made script that executes in the terminal.
    a. Open the terminal from the desktop.
    b. Copy the crosstab.py script from the ‘scripts/bin’ directory to the ‘Hansen_Global’ directory with the ‘cp’ command:

```
cp scripts/bin/crosstab.py Hansen_Global/
```

2. Open a MATE terminal and navigate to the directory where the sample shapefile and “crosstab.py” are located.
3. Type `python crosstab.py -v -a [column] [map].tif [shapefile].shp errormatrix.txt` where “[column]” is the column in the shapefile that contains the reference labels, “[map].tif” is the map that is being assessed (the stratification created in Section 3 in this case) and “[shapefile].shp” is the sample shapefile. This will create textfile that contains the error matrix called “errormatrix.txt”. An example command is:

```
Python crosstab.py -v -a label stratification_0N08W_utm.tif Validated_Sample/sample.shp
errormatrix.txt
```

Part 5. Analysis

1. Introduction

This step is referred to as the response design and includes providing reference labels for each unit in the sample. With each unit having a map label and a reference label we can construct an error matrix. The error matrix (with the mapped areas of each map category) contains all the information needed to perform the analysis which includes estimation of area and confidence intervals.

2. Estimation

The error matrix (with the mapped areas of each map category) contains all the information needed to perform the analysis which includes stratified estimation of area and confidence intervals. Again, this can be done various way but we recommend implementation in spreadsheet program to provide the user with an understanding of the estimation procedure.

1. The first step of the analysis open the error matrix in a spreadsheet software: open “LibreOffice Calc” from the Desktop menu in the VM (Menu > Applications > Office > LibreOffice Calc).

2. In LibreOffice Calc > File > Open > browse and open the text file created in the previous section. The screen should like below:

	Ref-Class_0.0	Ref-Class_1.0	Ref-Class_2.0	Ref-Class_3.0	Ref-Class_4.0	Ref-Class_5.0	Ref-Class_6.0
Map-Class_0.0	0	0	0	0	0	0	0
Map-Class_1.0	0	45	5	0	0	0	0
Map-Class_2.0	0	8	210	3	4	0	0
Map-Class_3.0	0	0	0	75	0	0	0
Map-Class_4.0	0	0	14	0	36	0	0
Map-Class_5.0	0	0	0	0	0	0	0
Map-Class_6.0	0	0	0	0	0	0	0

3. Add a column on the far right called ‘Total’. Sum the values of each of the reference classes across the row. Since we were not trying to estimate classes 0, 5, or 6, these can then be deleted. However, while these area proportions were small, they were not 0. Therefore, we must post-classify these classes as forest by adding their area proportion to the forest class.
4. In this case, the sample is stratified and the number of sample units per stratum is disproportionate relative to the area of the stratum; it is therefore necessary to estimate the area proportions $(\hat{p}_{ij})$ for each cell in the error matrix rather than sample counts before proceeding with the analysis. The area proportions are estimated as $\hat{p}_{ij} = W_i \times n_{ij} \div n_i$ where W_i are the stratum weights (the area proportion of stratum i), n_{ij} is the sample count in cell i,j, and n_i is the total number of sample counts in map category i.
5. Add a column on the right of ‘Ref Class_4.0’ called ‘Pixels’ and another “Wi’. Fill these values in with their corresponding values in the previous section.
6. In “LibreOffice Calc” copy the column and row headers and paste below the matrix, and the “Pixels” and “W_i” columns to below the sample counts error matrix. Optionally add a ‘Map’ label to the rows and ‘Reference’ label to the columns. The top matrix will be your ‘Sample count’ matrix and the bottom ‘Area proportions’. The spreadsheet should look like:

Error Matrix, Sample Counts

		Reference						
		Non-Forest	Forest	Water	Forest Loss	Total	Pixels	Wi
Map	Non-Forest	45	5	0	0	50	67188	0.027
	Forest	8	210	3	4	225	2129531	0.894
	Water	0	0	75	0	75	124067	0.051
	Forest Loss	0	14	0	36	50	67306	0.028

Error Matrix, Area proportions

		Reference						
		Non-Forest	Forest	Water	Forest Loss	Total	Pixels	Wi
Map	Non-Forest						67188	0.027
	Forest						2129531	0.894
	Water						124067	0.051
	Forest-Loss						67306	0.028

7. In the first cell in the area proportions matrix, calculate $\hat{p}_{11} = W_1 \times n_{11} \div n_1$ (the spreadsheet expression should be "=I4*C4/G4" without the quotation marks; see screenshot below).

Error Matrix, Sample Counts

		Reference						
		Non-Forest	Forest	Water	Forest Loss	Total	Pixels	Wi
Map	Non-Forest	45	5	0	0	50	67188	0.027
	Forest	8	210	3	4	225	2129531	0.894
	Water	0	0	75	0	75	124067	0.051
	Forest Loss	0	14	0	36	50	67306	0.028

Error Matrix, Area proportions

		Reference						
		Non-Forest	Forest	Water	Forest Loss	Total	Pixels	Wi
Map	Non-Forest	=I4*(C4/G4)					67188	0.027
	Forest						2129531	0.894
	Water						124067	0.051
	Forest-Loss						67306	0.028

8. Then just populate the rest of the first row of the matrix by highlighting the first cell and then "grabbing" the little black square at the bottom right of the cell (mouse pointer turns into a plus sign) and drag to the end of the row.
9. Then highlight the first row of the matric and and drag down to populate the entire matric; highlight all cells > right click > Format cells... > set format to Number with 4 decimals.
10. The error matrix you just created contains all of the information required for stratified estimation area! And estimators are now easily obtained as the column totals of the estimated area proportions. Calculate the row and columns totals by highlighting the row or the cell and clicking the sum sign (∑) above the B column. To check if you got it right: the row totals should equal Wi and the totals should sum to 1:

Error Matrix, Sample Counts								
			Reference					
		Non-Forest	Forest	Water	Forest Loss	Total	Pixels	Wi
Map	Non-Forest	45	5	0	0	50	67188	0.027
	Forest	8	210	3	4	225	2129531	0.894
	Water	0	0	75	0	75	124067	0.051
	Forest Loss	0	14	0	36	50	67306	0.028
Error Matrix, Area proportions								
			Reference					
		Non-Forest	Forest	Water	Forest Loss	Total	Pixels	Wi
Map	Non-Forest	0.0243	0.0027	0.0000	0.0000	0.0270	67188	0.027
	Forest	0.0318	0.8344	0.0119	0.0159	0.8940	2129531	0.894
	Water	0.0000	0.0000	0.0510	0.0000	0.0510	124067	0.051
	Forest-Loss	0.0000	0.0078	0.0000	0.0202	0.0280	67306	0.028
	Total	0.0560866667	0.84494	0.06292	0.0360533333	1		

11. You have just calculated unbiased estimates of area! I.e. the column totals. To express these in hectares rather proportions multiply the column totals by the stratum size and the pixel size in hectares (30^2/100^2). For example, an unbiased area estimate of map class 1 in hectares is calculated as "=C18*H18*30^2/100^2". Do this calculation on row 14 for all classes. (It's a good idea to first calculate the area in pixels and calculate the sum to make sure it matches the total map area).

A	B	C	D	E	F	G	H	I
Error Matrix, Sample Counts								
			Reference					
		Non-Forest	Forest	Water	Forest Loss	Total	Pixels	Wi
Map	Non-Forest	45	5	0	0	50	67188	0.027
	Forest	8	210	3	4	225	2129531	0.894
	Water	0	0	75	0	75	124067	0.051
	Forest Loss	0	14	0	36	50	67306	0.028
Error Matrix, Area proportions								
			Reference					
		Non-Forest	Forest	Water	Forest Loss	Total	Pixels	Wi
Map	Non-Forest	0.0243	0.0027	0.0000	0.0000	0.0270	67188	0.027
	Forest	0.0318	0.8344	0.0119	0.0159	0.8940	2129531	0.894
	Water	0.0000	0.0000	0.0510	0.0000	0.0510	124067	0.051
	Forest-Loss	0.0000	0.0078	0.0000	0.0202	0.0280	67306	0.028
	Total	0.1	0.8	0.1	0.0	1	2388092	1
	Area [pix]	133940.1	2017794.5	150258.7	86098.7			
	Area [hex]	12054.6	181601.5	13523.3	7748.9			

12. The next step is to calculate the standard errors of the area estimates, which are given by the following equation for a stratified random sample:

$$S(\hat{p}_{\cdot j}) = \sqrt{\sum_i \frac{W_i \hat{p}_{ij} - \hat{p}_{ij}^2}{n_{i\cdot} - 1}}$$

.

13. This can be tricky to get right in a spreadsheet! Calculate the standard errors in row 21; the S($\hat{p}$1) which is the standard error for map class 1 (first column total) is calculated as =SQRT((I13*C13-C13^2)/($G4-1)+($I$14*C14-C14^2)/($G$5-1)+($I$15*C15-C15^2)/($G$6-1)+($I$16*C16-C16^2)/($G$7-1))”; then just can drag the expression to complete the row.

SQRT | =SQRT((I13*C13-C13^2)/($G4-1)+($I$14*C14-C14^2)/($G$5-1)+($I$15*C15- C15^2)/($G$6-1)+($I$16*C16-C16^2)/($G$7-1))

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**Error Matrix, Sample Counts**												
2				**Reference**									
3			Non-Forest	Forest	Water	Forest Loss	Total	Pixels	Wi				
4		Non-Forest	45	5	0	0	50	67188	0.027				
5	**Map**	Forest	8	210	3	4	225	2129531	0.894				
6		Water	0	0	75	0	75	124067	0.051				
7		Forest Loss	0	14	0	36	50	67306	0.028				
8													
9													
10	**Error Matrix, Area proportions**												
11				**Reference**									
12			Non-Forest	Forest	Water	Forest Loss	Total	Pixels	Wi				
13		Non-Forest	0.0243	0.0027	0.0000	0.0000	0.0270	67188	0.027				
14	**Map**	Forest	0.0318	0.8344	0.0119	0.0159	0.8940	2129531	0.894				
15		Water	0.0000	0.0000	0.0510	0.0000	0.0510	124067	0.051				
16		Forest-Loss	0.0000	0.0078	0.0000	0.0202	0.0280	67306	0.028				
17		Total	0.1	0.8	0.1	0.0	1	2388092	1				
18		Area [pix]	133940.1	2017794.5	150258.7	86098.7							
19		Area [hex]	12054.6	181601.5	13523.3	7748.9							
20		S(Area)	=SQRT((I13*C13-C13^2)/($G4-1)+($I$14*C14-C14^2)/($G$5-1)+($I$15*C15- C15^2)/($G$6-1)+($I$16*C16-C16^2)/($G$7-1))										
21													

14. Now, calculate the standard errors in the units of hectares by multiplying S(area) by the total number of pixels times 30^/100^2; 95% confidence intervals are given by multiplying the standard errors by 1.96. The spreadsheet should look like below:

Error Matrix, Sample Counts								
			Reference					
		Non-Forest	Forest	Water	Forest Loss	Total	Pixels	Wi
	Non-Forest	45	5	0	0	50	67188	0.027
Map	Forest	8	210	3	4	225	2129531	0.894
	Water	0	0	75	0	75	124067	0.051
	Forest Loss	0	14	0	36	50	67306	0.028
Error Matrix, Area proportions								
			Reference					
		Non-Forest	Forest	Water	Forest Loss	Total	Pixels	Wi
	Non-Forest	0.0243	0.0027	0.0000	0.0000	0.0270	67188	0.027
Map	Forest	0.0318	0.8344	0.0119	0.0159	0.8940	2129531	0.894
	Water	0.0000	0.0000	0.0510	0.0000	0.0510	124067	0.051
	Forest-Loss	0.0000	0.0078	0.0000	0.0202	0.0280	67306	0.028
	Total	0.1	0.8	0.1	0.0	1	2388092	1
	Area [pix]	133940.1	2017794.5	150258.7	86098.7			
	Area [hec]	12054.6	181601.5	13523.3	7748.9			
	S(Area)	0.0	0.0	0.0	0.0			
	S(Area) [hec]	2390.4	3235.2	1472.5	1739.8			
	95% CI [hec]	4685.1	6341.0	2886.1	3410.1			

15. Finally, we can estimate the accuracy of the map. Three different accuracy measures are of interest: i) overall accuracy which is simply the sum of the diagonals in the error matrix of estimated area proportions; ii) user’s accuracy which for a map category i is

given by $\hat{U}_i = \hat{p}_{ii} \div \hat{p}_{i\cdot}$ and iii) producer’s accuracy for map category j given by $\hat{P}_i = \hat{p}_{jj} \div \hat{p}_{\cdot j}$ where $p\hat{\cdot i}$ and $p\hat{\cdot j}$ are the row and columns totals respectively. In my example, I calculated user’s accuracy in row 24 ($\hat{U}_1$=C13/G13), producer’s in row 25 ($p\hat{1}$=C13/C18”) and overall in row 26 (“=sum(C13,D14,E15,F16)”). This gives the final spreadsheet with areas in green cells and accuracies in blue cells (Note: Make sure the ‘Total’ row has is formated to have at least 3 decimals):

Error Matrix, Sample Counts								
			Reference					
		Non-Forest	Forest	Water	Forest Loss	Total	Pixels	Wi
Map	Non-Forest	45	5	0	0	50	67188	0.027
	Forest	8	210	3	4	225	2129531	0.894
	Water	0	0	75	0	75	124067	0.051
	Forest Loss	0	14	0	36	50	67306	0.028

Error Matrix, Area proportions								
			Reference					
		Non-Forest	Forest	Water	Forest Loss	Total	Pixels	Wi
Map	Non-Forest	0.0243	0.0027	0.0000	0.0000	0.0270	67188	0.027
	Forest	0.0318	0.8344	0.0119	0.0159	0.8940	2129531	0.894
	Water	0.0000	0.0000	0.0510	0.0000	0.0510	124067	0.051
	Forest-Loss	0.0000	0.0078	0.0000	0.0202	0.0280	67306	0.028
	Total	0.056	0.845	0.063	0.036	1	2388092	1
	Area [pix]	133940.1	2017794.5	150258.7	86098.7			
	Area [hec]	12054.6	181601.5	13523.3	7748.9			
	S(Area)	0.0	0.0	0.0	0.0			
	S(Area) [hec]	2390.4	3235.2	1472.5	1739.8			
	95% CI [hec]	4685.1	6341.0	2886.1	3410.1			
	Users	0.90	0.93	1.00	0.72			
	Producers	0.43	0.99	0.81	0.56			
	Overall	0.92986						