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

5. Analysis

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

5.1 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.

5.2 Estimation

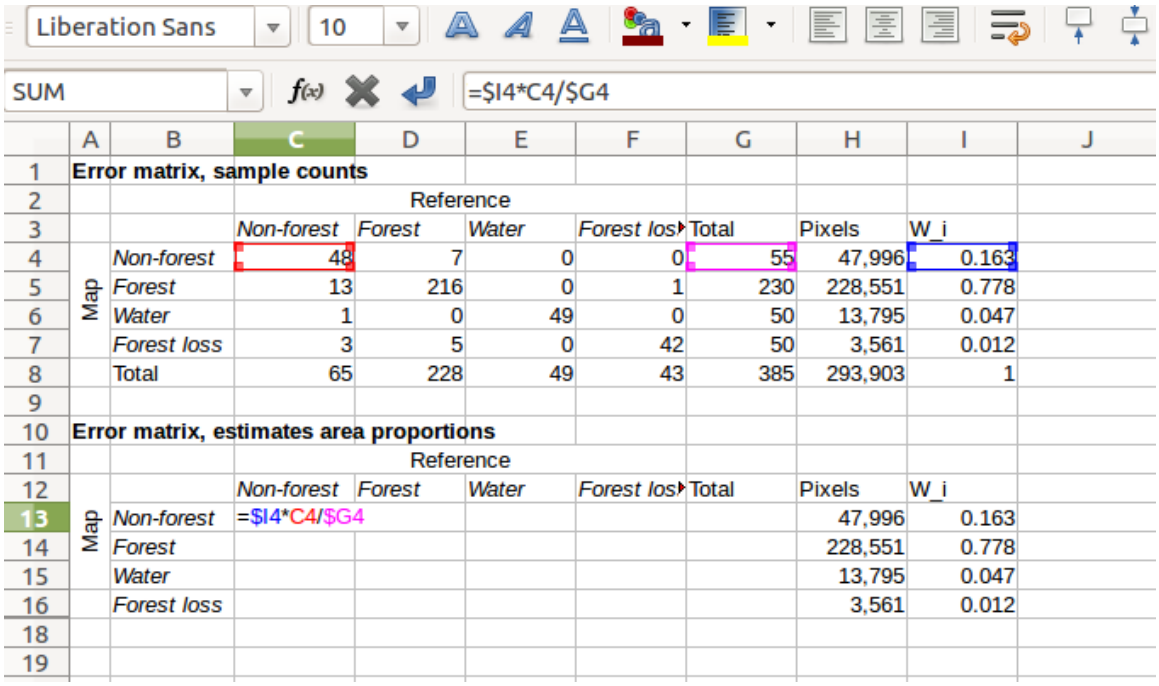
The error matrix (with the mapped areas of each map category) contains all the information needed to perform the analysis which includes stratified estimation of area and confidence intervals. Again, this can be done various way but we recommend implementation in spreadsheet program to provide the user 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 (*Office* > *LibreOffice Calc*).
2. In LibreOffice Calc > *File* > Open > browse and open the text file created in 4.3 in the previous section. The screen should like below:

G12

	A	B	C	D	E	F	G	H	I	J
1	**Error matrix, sample counts**									
2			Reference							
3			*Non-forest*	*Forest*	*Water*	*Forest los*	Total	Pixels	W_i	
4	Map	*Non-forest*	48	7	0	0	55	47,996	0.163	
5		*Forest*	13	216	0	1	230	228,551	0.778	
6		*Water*	1	0	49	0	50	13,795	0.047	
7		*Forest loss*	3	5	0	42	50	3,561	0.012	
8		Total	65	228	49	43	385	293,903	1	
9										
10										
11										
12										
13										
14										
15										

3. 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.
4. 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.
5. 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).



	A	B	C	D	E	F	G	H	I	J
1	**Error matrix, sample counts**									
2			Reference							
3			*Non-forest*	*Forest*	*Water*	*Forest los*	Total	Pixels	W_i	
4	Map	*Non-forest*	48	7	0	0	55	47,996	0.163	
5		*Forest*	13	216	0	1	230	228,551	0.778	
6		*Water*	1	0	49	0	50	13,795	0.047	
7		*Forest loss*	3	5	0	42	50	3,561	0.012	
8		Total	65	228	49	43	385	293,903	1	
9										
10	**Error matrix, estimates area proportions**									
11			Reference							
12			*Non-forest*	*Forest*	*Water*	*Forest los*	Total	Pixels	W_i	
13	Map	*Non-forest*	=$I4*C4/$G4					47,996	0.163	
14		*Forest*						228,551	0.778	
15		*Water*						13,795	0.047	
16		*Forest loss*						3,561	0.012	
18										
19										

6. 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.
7. 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.
8. 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 W_i and the totals should sum to 1:

J25

	A	B	C	D	E	F	G	H	I	J
1	**Error matrix, sample counts**									
2			Reference							
3			*Non-forest*	*Forest*	*Water*	*Forest loss*	Total	Pixels	W_i	
4	Map	*Non-forest*	48	7	0	0	55	47,996	0.163	
5		*Forest*	13	216	0	1	230	228,551	0.778	
6		*Water*	1	0	49	0	50	13,795	0.047	
7		*Forest loss*	3	5	0	42	50	3,561	0.012	
8		Total	65	228	49	43	385	293,903	1	
9										
10	**Error matrix, estimates area proportions**									
11			Reference							
12			*Non-forest*	*Forest*	*Water*	*Forest loss*	Total	Pixels	W_i	
13	Map	*Non-forest*	0.1425	0.0208	0.0000	0.0000	0.1633	47,996	0.163	
14		*Forest*	0.0440	0.7303	0.0000	0.0034	0.7776	228,551	0.778	
15		*Water*	0.0009	0.0000	0.0460	0.0000	0.0469	13,795	0.047	
16		*Forest loss*	0.0007	0.0012	0.0000	0.0102	0.0121	3,561	0.012	
18		Total	0.1881	0.7523	0.0460	0.0136	1	293,903	1	
19										
20										

9. 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). In my example, I get the following unbiased area estimates: 4,977 ha, 19,899 ha, 1,217 ha and 359 ha:

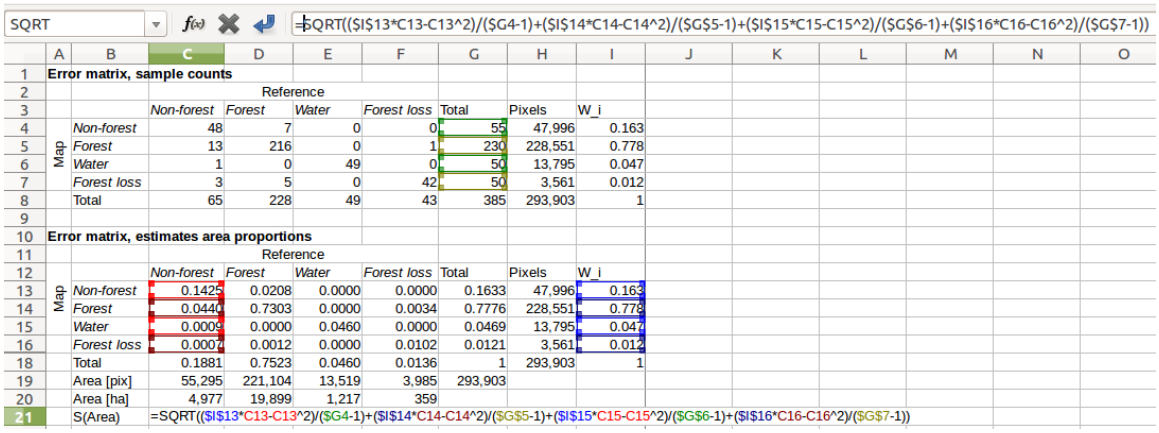
H25

	A	B	C	D	E	F	G	H	I	J
1	**Error matrix, sample counts**									
2			Reference							
3			*Non-forest*	*Forest*	*Water*	*Forest loss*	Total	Pixels	W_i	
4	Map	*Non-forest*	48	7	0	0	55	47,996	0.163	
5		*Forest*	13	216	0	1	230	228,551	0.778	
6		*Water*	1	0	49	0	50	13,795	0.047	
7		*Forest loss*	3	5	0	42	50	3,561	0.012	
8		Total	65	228	49	43	385	293,903	1	
9										
10	**Error matrix, estimates area proportions**									
11			Reference							
12			*Non-forest*	*Forest*	*Water*	*Forest loss*	Total	Pixels	W_i	
13	Map	*Non-forest*	0.1425	0.0208	0.0000	0.0000	0.1633	47,996	0.163	
14		*Forest*	0.0440	0.7303	0.0000	0.0034	0.7776	228,551	0.778	
15		*Water*	0.0009	0.0000	0.0460	0.0000	0.0469	13,795	0.047	
16		*Forest loss*	0.0007	0.0012	0.0000	0.0102	0.0121	3,561	0.012	
18		Total	0.1881	0.7523	0.0460	0.0136	1	293,903	1	
19		Area [pix]	55,295	221,104	13,519	3,985	293,903			
20		Area [ha]	4,977	19,899	1,217	359				
21										

10. 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:

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

This can be tricky to get right in a spreadsheet! Calculate the standard errors in row 21; the $S(\hat{p}_{\cdot 1})$ 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.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	**Error matrix, sample counts**														
2			Reference												
3			*Non-forest*	*Forest*	*Water*	*Forest loss*	Total	Pixels	W_i						
4		*Non-forest*	48	7	0	0	55	47,996	0.163						
5	Map	*Forest*	13	216	0	1	230	228,551	0.778						
6		*Water*	1	0	49	0	50	13,795	0.047						
7		*Forest loss*	3	5	0	42	50	3,561	0.012						
8		Total	65	228	49	43	385	293,903	1						
9															
10	**Error matrix, estimates area proportions**														
11			Reference												
12			*Non-forest*	*Forest*	*Water*	*Forest loss*	Total	Pixels	W_i						
13		*Non-forest*	0.1425	0.0208	0.0000	0.0000	0.1633	47,996	0.163						
14	Map	*Forest*	0.0440	0.7303	0.0000	0.0034	0.7776	228,551	0.778						
15		*Water*	0.0009	0.0000	0.0460	0.0000	0.0469	13,795	0.047						
16		*Forest loss*	0.0007	0.0012	0.0000	0.0102	0.0121	3,561	0.012						
18		Total	0.1881	0.7523	0.0460	0.0136	1	293,903	1						
19		Area [pix]	55,295	221,104	13,519	3,985	293,903								
20		Area [ha]	4,977	19,899	1,217	359									
21		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))												

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

K24 $f(x)$ Σ =

	A	B	C	D	E	F	G	H	I	J
1	**Error matrix, sample counts**									
2			Reference							
3			*Non-forest*	*Forest*	*Water*	*Forest loss*	Total	Pixels	W_i	
4	Map	*Non-forest*	48	7	0	0	55	47,996	0.163	
5		*Forest*	13	216	0	1	230	228,551	0.778	
6		*Water*	1	0	49	0	50	13,795	0.047	
7		*Forest loss*	3	5	0	42	50	3,561	0.012	
8		Total	65	228	49	43	385	293,903	1	
9										
10	**Error matrix, estimates area proportions**									
11			Reference							
12			*Non-forest*	*Forest*	*Water*	*Forest loss*	Total	Pixels	W_i	
13	Map	*Non-forest*	0.1425	0.0208	0.0000	0.0000	0.1633	47,996	0.163	
14		*Forest*	0.0440	0.7303	0.0000	0.0034	0.7776	228,551	0.778	
15		*Water*	0.0009	0.0000	0.0460	0.0000	0.0469	13,795	0.047	
16		*Forest loss*	0.0007	0.0012	0.0000	0.0102	0.0121	3,561	0.012	
18		Total	0.1881	0.7523	0.0460	0.0136	1	293,903	1	
19		Area [pix]	55,295	221,104	13,519	3,985	293,903			
20		Area [ha]	4,977	19,899	1,217	359				
21		S(Area)	0.0140	0.0144	0.0009	0.0034				
22		S(Area) [ha]	371	380	25	91				
23		95% CI [ha]	727	744	49	178				
24										

12. 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 $\hat{p}_{i\cdot}$ and $\hat{p}_{\cdot j}$ 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 ($\hat{P}_1$ "=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:

M8

	A	B	C	D	E	F	G	H	I	J
1	**Error matrix, sample counts**									
2			Reference							
3			*Non-forest*	*Forest*	*Water*	*Forest loss*	Total	Pixels	W_i	
4	Map	*Non-forest*	48	7	0	0	55	47,996	0.163	
5		*Forest*	13	216	0	1	230	228,551	0.778	
6		*Water*	1	0	49	0	50	13,795	0.047	
7		*Forest loss*	3	5	0	42	50	3,561	0.012	
8		Total	65	228	49	43	385	293,903	1	
9										
10	**Error matrix, estimates area proportions**									
11			Reference							
12			*Non-forest*	*Forest*	*Water*	*Forest loss*	Total	Pixels	W_i	
13	Map	*Non-forest*	0.1425	0.0208	0.0000	0.0000	0.1633	47,996	0.163	
14		*Forest*	0.0440	0.7303	0.0000	0.0034	0.7776	228,551	0.778	
15		*Water*	0.0009	0.0000	0.0460	0.0000	0.0469	13,795	0.047	
16		*Forest loss*	0.0007	0.0012	0.0000	0.0102	0.0121	3,561	0.012	
18		Total	0.1881	0.7523	0.0460	0.0136	1	293,903	1	
19		Area [pix]	55,295	221,104	13,519	3,985	293,903			
20		**Area [ha]**	**4,977**	**19,899**	**1,217**	**359**				
21		S(Area)	0.0140	0.0144	0.0009	0.0034				
22		S(Area) [ha]	371	380	25	91				
23		**95% CI [ha]**	**727**	**744**	**49**	**178**				
24		**User's**	**0.87**	**0.94**	**0.98**	**0.84**				
25		**Producer's**	**0.76**	**0.97**	**1.00**	**0.75**				
26		**Overall**	**0.92900364**							