Pontus Olofsson and Paulo Arevalo
Boston Education in Earth Observation Data Analysis/
Department of Earth & Environment, Boston University

6. Use of buffer stratum

6.1 Introduction

A buffer stratum is particularly efficient when trying to estimate the area of a rare activity (typically deforestation) in the presence of very large strata (typically forest or non-forest). In this case, let's consider the following example from Peru. Our objective is to estimate the area of forest loss and gain. A map has been extracted from the Global Forest Change dataset (https://earthenginepartners.appspot.com/science-2013-global-forest) by University of Maryland. The map covers the Región of Madre de Dios in southeastern Peru and contains five classes: forest cover loss and gain, and area of both gain and loss, between 2001 and 2016, and stable forest, non-forest and water. We know that this map – as all maps – have errors which prohibit us from obtaining the areas of interest by counting the pixels in the map classified as forest loss and gains. Instead, we need to estimate the areas by applying an unbiased estimator to sample data of reference conditions. Even if the map has errors, it is still very useful because it helps us target the sampling to regions where forest cover change is likely to have occurred. A problem, obvious from looking at Figure 1, is that the forest map class is far larger than other classes while the change is barely noticeable. This large difference in stratum area is problematic because any sample unit in the forest stratum identified has having experienced forest cover change (i.e. omission errors) will carry a large area weight. The area proportion of such an omission error is the weight of the forest stratum divided by the number of sample units allocated to the forest stratum. From Figure 1, the forest stratum appears to occupy 95% of Madre de Dios. If assuming that 300 units of a sample of 500 units were allocated to the forest stratum, one single omission error would represent an area of 0.95/300 = 0.3% of the region. Errors with such large weights will have a detrimental impact on precision of estimates of area and accuracy. A larger sample size in the forest stratum would reduce the weight of omission errors but it would also increase cost and time. Incorporating information about where errors are likely to occur into the stratification would increase efficiency without increasing cost – that's the rationale behind the buffer stratum. We know from experience that errors of omission tend to appear in close proximity to actual forest cover change; hence, a buffer zone of a few pixels around each pixel mapped as change is likely to contain omission errors. A stratum corresponding to the buffer zone will have a substantially smaller weight, as will any omission errors in the buffer stratum.

This work is licensed under the Creative Commons Attribution-ShareAlike 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-sa/4.0/.

6.2 Stratified estimation without buffer stratum

1. In QGIS, open the map of Madre de Dios Peru ("/work/2_stratification/GFC_Madre_de_Dios.tif") that is delivered with the Virtual Machine and display it using "palette.txt". It should look like Figure 1.

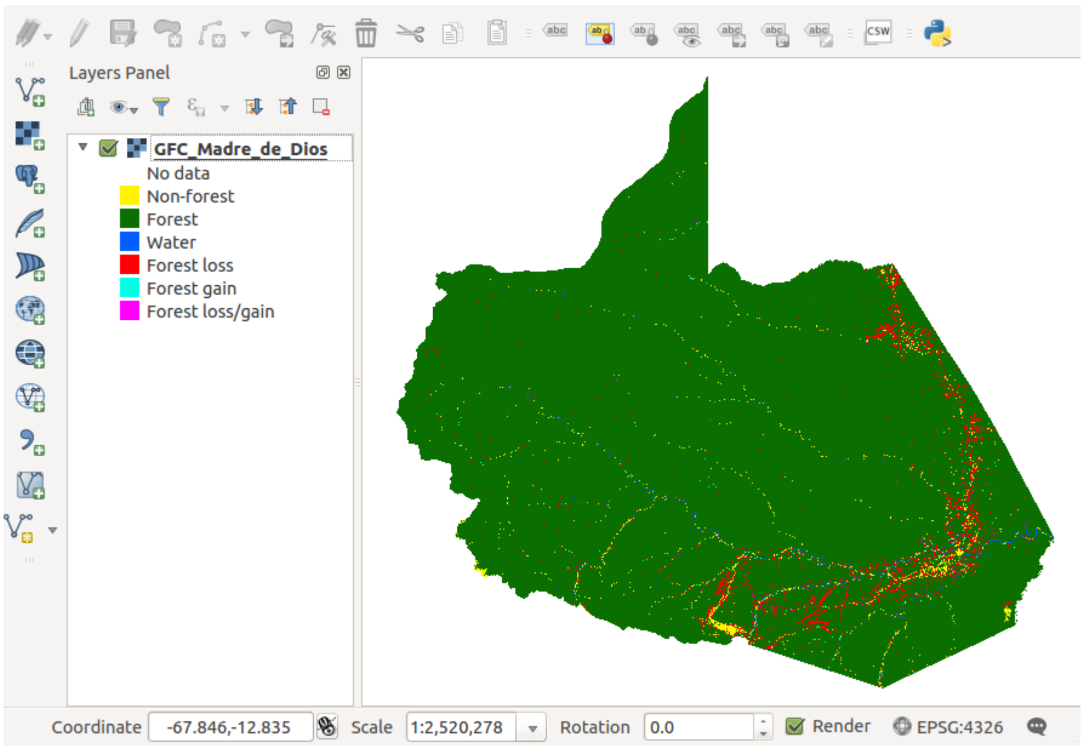

Figure 1. QGIS

2. Design a sample without the use of a buffer, First, delete any *aux.xml files in the stratification directory (the gdalinfo –hist command will return erroneous values otherwise)
3. Navigate to "/2_stratification/" and run `gdalinfo -hist GFC_Madre_de_Dios.tif` to obtain the area of each stratum
4. Copy the area in number of pixels of each stratum into a spreadsheet application (LibeOffice Calc in the VM for example, or Excel on your host computer) to start determining sample size and allocation: – the calculations are the same as in the case of using a buffer as the weight of the forest loss stratum is the same (it's only the weight of the forest stratum that is different):

Stratum	*Non-forest*	*Forest*	*Water*	*Forest loss*	*Forest gain*	*Gain/loss*	*Total*
Area [pix]	1253492	108028839	364623	2294651	110091	64500	112,116,196
Area [ha]	112,814	9,722,596	32,816	206,519	9,908	5,805	10,090,458
W_i [%]	1.12%	96.35%	0.33%	2.05%	0.10%	0.06%	100%

5. Let's design a sample to estimate the area of forest loss with a target margin of error of 33%, assuming that 90% of the forest loss is in the forest loss stratum and one error of omission in other strata. Following calculations in Section 3.2 in "3. Sampling design" gives:

pi	0.01	0.01	0.01	0.9	0.01	0.01	
Si	0.099498744	0.099498744	0.099499	0.3	0.099499	0.099499	
S(pi=4)				0.0035			
CI				0.0069			
MoE [%]				33.5%			
Wi x Si	0.001112	0.095871374	0.000324	0.00614	9.77E-05	5.72E-05	
n	-- 876 --						
n_i	50	550	50	50	50	50	800

6. In "/work/2_stratification", select the sample:
```
python ~/Desktop/scripts/bin/sample_map.py -v --size 800 --
allocation "50 550 50 50 50 50" --vector sample.shp stratified
GFC_Madre_de_Dios.tif
```
7. In a "real life situation", you would now go through each unit in the sample to observe reference conditions. The shapefile "sample_wo_buffer.shp" contains made-up reference labels that can be used instead.
8. First, let's proceed without using the buffer information by creating an error matrix using the original stratification and the sample in the point above:
```
python
~/Desktop/scripts/bin/crosstab.py -v -a ref_label
GFC_Madre_de_Dios.tif ../4_interpretation/sample_wo_buffer.shp
../5_analysis/errormatrix_wo_buffer.txt
```
9. Create stratified estimators of area with 95% confidence intervals – you should get the following estimates (use the following expression to calculate the CIs:
```
=$H$7*1.96*SQRT(($I21*B31-B31^2)/($H21-1) + ($I22*B32-
B32^2)/($H22-1) + ($I23*B33-B33^2)/($H23-1) + ($I24*B34-
B34^2)/($H24-1) + ($I25*B35-B35^2)/($H25-1) + ($I26*B36-
B36^2)/($H26-1))
```
also note that this is information is in the spreadsheet named "stratified_wo_buffer.xlsx"):

18	**ANALYSIS**								
19	Error matrix, sample counts (map in rows, reference in columns)								
20		*Non-forest*	*Forest*	*Water*	*Forest loss*	*Forest gain*	*Gain/loss*	*Total*	*Wi*
21	*Non-forest*	47	3	0	0	0	0	50	0.01118
22	*Forest*	14	530	1	4	0	1	550	0.963544
23	*Water*	0	1	49	0	0	0	50	0.003252
24	*Forest loss*	2	12	0	36	0	0	50	0.020467
25	*Forest gain*	8	18	0	0	24	0	50	0.000982
26	*Gain/loss*	10	22	0	5	4	9	50	0.000575
27	Total	81	586	50	45	28	10	800	
28									
29	Error matrix, estimated area proportions (map in rows, reference in columns)								
30		*Non-forest*	*Forest*	*Water*	*Forest loss*	*Forest gain*	*Gain/loss*	*Total*	
31	*Non-forest*	0.0105	0.0007	0.0000	0.0000	0.0000	0.0000	0.0112	
32	*Forest*	0.0245	0.9285	0.0018	0.0070	0.0000	0.0018	0.9635	
33	*Water*	0.0000	0.0001	0.0032	0.0000	0.0000	0.0000	0.0033	
34	*Forest loss*	0.0008	0.0049	0.0000	0.0147	0.0000	0.0000	0.0205	
35	*Forest gain*	0.0002	0.0004	0.0000	0.0000	0.0005	0.0000	0.0010	
36	*Gain/loss*	0.0001	0.0003	0.0000	0.0001	0.0000	0.0001	0.0006	
37	Total	0.0361	0.9348	0.0049	0.0218	0.0005	0.0019		
38	**Area [ha]**	**364,537**	**9,432,157**	**49,837**	**219,984**	**5,220**	**18,722**		
39	**95% CI [ha**	**128,821**	**154,430**	**34,672**	**73,824**	**1,454**	**34,653**		
40	**MoE**	**35%**	**2%**	**70%**	**34%**	**28%**	**185%**		
41	**Prod acc**	**29%**	**99%**	**65%**	**68%**	**91%**	**6%**		
42	**Users acc**	**94%**	**96%**	**98%**	**72%**	**48%**	**18%**		

6.2 Create the buffer stratum

1. First, start the VM; in QGIS open the map of Madre de Dios Peru ("GFC_Madre_de_Dios.tif") that is delivered with the Virtual Machine and display it using "palette.txt". It should look like Figure 1.
2. Download the script that we will use to create the buffer (save it in the bin directory on the VM Desktop, "/home/opengeo-vm/Desktop/scripts/bin/"): https://raw.githubusercontent.com/parevalo/workflow/master/multi_scene/6_stratification/create_strata.py
3. Open a terminal and navigate to the directory of the Peru map (/home/opengeo-vm/work/2_stratification/). To create buffer of three pixels around mapped forest (map class 4 in this example that has a value of 1 and 8-bit precision, type:

```
gdal_proximity.py GFC_Madre_de_Dios.tif lossbuffer.tif -values 4 -maxdist 3 -fixed-buf-val 1 -nodata 0 -ot Byte -co "COMPRESS=PACKBITS"
```

NOTE: The size of the buffer will have an impact on its effectiveness – a larger buffer is likely to catch more errors but the weight of the buffer stratum will increase. There is no magic size of the buffer but in general, you can afford a larger stratum if the forest stratum is large. In this case, the forest stratum is relatively very large and we'll use a three pixel buffer.

4. Once completed, open it in QGIS, set *Value* to "0", *Color* to 0% opacity and leave *Label* blank; and add another value, set *Value* to "1", *Color* to black and *Label* "buffer" (right click layer > *Properties* > *Style* > *Render type:* "Singleband Pseudocolor"). You should now see a black buffer around areas of forest loss as in Figure 2:

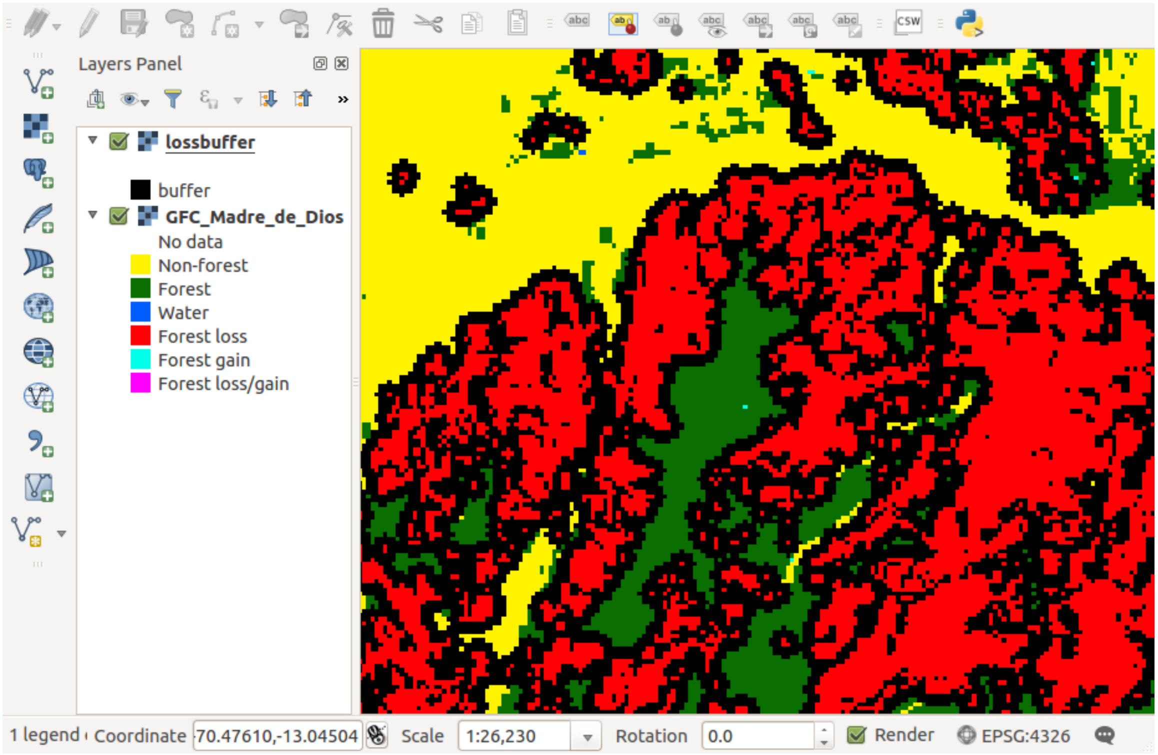

Figure 2. Buffer draped onto original stratification.

6.3 Update stratification with the buffer stratum

First, let's merge the buffer and the original stratification by giving buffered **forest** pixels a value of 7. The buffer includes all map classes but we only want to buffer pixels in the forest stratum because of its large weight. All other pixels remain unaltered.

1. We can do that using the Raster Calculator and the following expression:
```
(("lossbuffer@1" = 1) AND ("GFC_Madre_de_Dios@1" = 2))*7 +
(("lossbuffer@1" = 0) AND ("GFC_Madre_de_Dios@1" = 2))*2 +
(("lossbuffer@1" >= 0) AND ("GFC_Madre_de_Dios@1" = 0))*0 +
(("lossbuffer@1" >= 0) AND ("GFC_Madre_de_Dios@1" = 1))*1 +
(("lossbuffer@1" >= 0) AND ("GFC_Madre_de_Dios@1" = 3))*3 +
(("lossbuffer@1" >= 0) AND ("GFC_Madre_de_Dios@1" = 4))*4 +
(("lossbuffer@1" >= 0) AND ("GFC_Madre_de_Dios@1" = 5))*5 +
(("lossbuffer@1" >= 0) AND ("GFC_Madre_de_Dios@1" = 6))*6
```
2. Save as "GFC_Madre_de_Dios_32b.tif".
3. To convert to 8 bit precision, type in the Terminal:
```
gdal_translate -ot byte GFC_Madre_de_Dios_Buffered_32b.tif
GFC_Madre_de_Dios_Buffered.tif
```
4. Use the palette named "nice_colors_buffer.txt". The result should like in Figure 4.

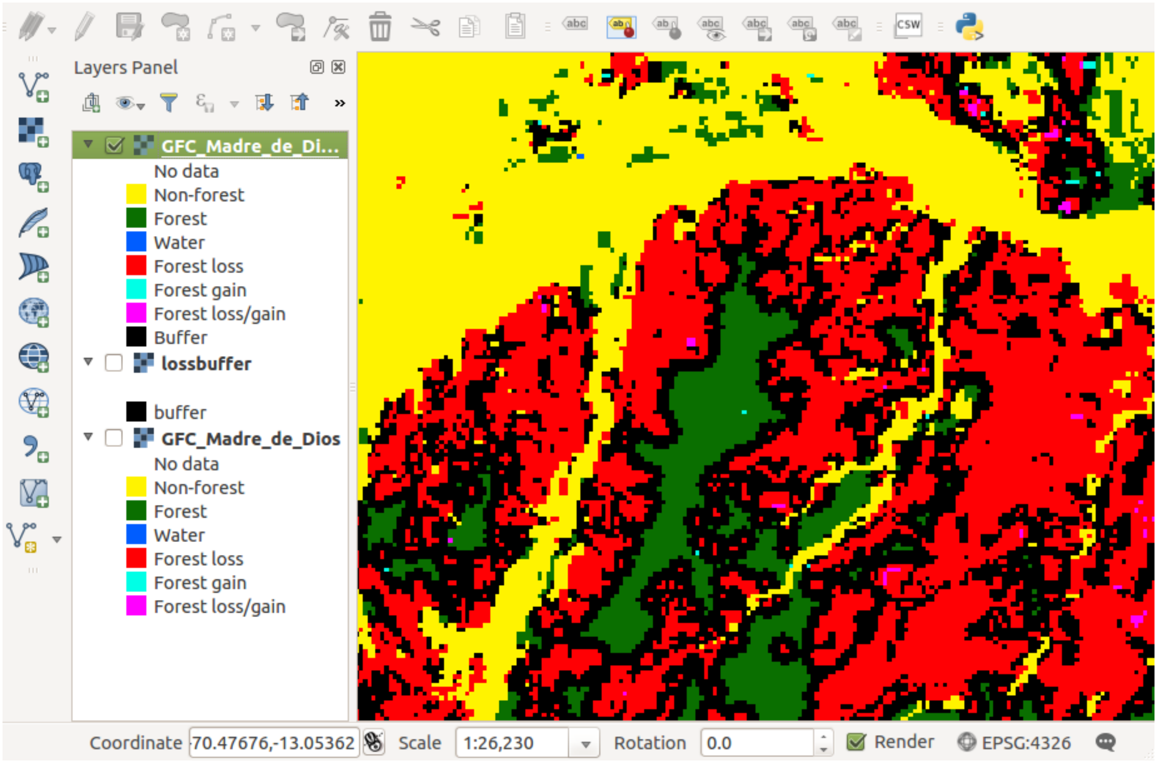

Figure 3. Buffer stratum incorporated into original stratification.

6.4 Use the buffer stratum in a pre-stratification approach

To use the stratum in a stratification of the study area, just follow the instructions in "3. Sampling Design" but add the buffer as a seventh stratum:

10. First, delete any *aux.xml files in the stratification directory (the gdal –hist command will return erroneous values otherwise)
11. Run `gdalinfo -hist GFC_Madre_de_Dios_Buffered.tif` to obtain the area of each stratum
12. Copy the area in number of pixels of each stratum into a spreadsheet application (LibeOffice Calc in the VM for example, or Excel on your host computer) to start determining sample size and allocation:

Stratum	*Non-forest*	*Forest*	*Water*	*Forest loss*	*Forest gain*	*Gain/loss*	*Buffer*	*Total*
Area [pix]	1,253,492	103,033,862	364,623	2,294,651	110,091	64,500	4,994,977	112,116,196
Area [ha]	112,814	9,273,048	32,816	206,519	9,908	5,805	449,548	10,090,458
W_i [%]	1.12%	91.90%	0.33%	2.05%	0.10%	0.06%	4.46%	100%

13. Let's design a sample to estimate the area of forest loss with a target margin of error of 33%, assuming one error of omission in each strata and five in Buffer. Following calculations in Section 3.2 in "3. Sampling design" gives:

p_i	0.01	0.01	0.01	0.9	0.05	0.01	0.01	
S_i	0.0995	0.0995	0.0995	0.3	0.218	0.0995	0.0995	
$S(p_{i=4})$				0.0034				
CI				0.0067				
MoE [%]				33%				
n	----------	----------	----------	931	----------	----------	----------	
n_i	50	550	50	50	50	50	100	900

14. Select the sample: `python ~/Desktop/scripts/bin/sample_map.py -v --size 900 --allocation "50 550 50 50 50 50 100" --vector sample.shp stratified_GFC_Madre_de_Dios_Buffered.tif`
15. In a "real life situation", you would now go through each unit in the sample to observe reference conditions. The shapefile "sample_w_buffer.shp" contains made-up reference labels that can be used instead.
16. As explained in "4. Response Design", create an error matrix: `python ~/Desktop/scripts/bin/crosstab.py -v -a ref_label GFC_Madre_de_Dios_Buffered.tif ../4_interpretation/sample_w_buffer.shp ../5_analysis/errormatrix_w_buffer.txt`
17. Follow the steps in "5. Analysis" to construct stratified area estimates with 95% confidence intervals:

17	**ANALYSIS**									
18	***Error matrix, sample counts***									
19		*Non-forest*	*Forest*	*Water*	*Forest loss*	*Forest gain*	*Gain/loss*	*Buffer*	Total	Wi
20	*Non-forest*	47	3	0	0	0	0	0	50	0.0112
21	*Forest*	13	535	1	1	0	0	0	550	0.9190
22	*Water*	0	1	49	0	0	0	0	50	0.0033
23	*Forest loss*	2	12	0	36	0	0	0	50	0.0205
24	*Forest gain*	8	18	0	0	24	0	0	50	0.0010
25	*Gain/loss*	10	22	0	5	4	9	0	50	0.0006
26	*Buffer*	0	96	0	3	0	1	0	100	0.0446
27	Total	80	687	50	45	28	10	0	900	1
28										
29	***Error matrix, estimated area proportions***									
30	*Non-forest*	0.0105	0.0007	0.0000	0.0000	0.0000	0.0000	0.0000	0.0112	
31	*Forest*	0.0217	0.8939	0.0017	0.0017	0.0000	0.0000	0.0000	0.9190	
32	*Water*	0.0000	0.0001	0.0032	0.0000	0.0000	0.0000	0.0000	0.0033	
33	*Forest loss*	0.0008	0.0049	0.0000	0.0147	0.0000	0.0000	0.0000	0.0205	
34	*Forest gain*	0.0002	0.0004	0.0000	0.0000	0.0005	0.0000	0.0000	0.0010	
35	*Gain/loss*	0.0001	0.0003	0.0000	0.0001	0.0000	0.0001	0.0000	0.0006	
36	*Buffer*	0.0000	0.0428	0.0000	0.0013	0.0000	0.0004	0.0000	0.0446	
37	Total	0.0333	0.9430	0.0049	0.0178	0.0005	0.0005	0.0000	1	
38	**Area [ha]**	**336,234**	**9,514,823**	**49,020**	**179,620**	**5,220**	**5,540**	**N/A**	10,090,458	
39	**95% CI [ha]**	**118,626**	**130,131**	**33,071**	**44,661**	**1,454**	**8,833**	**N/A**		
40	**MoE**	**35%**	**1%**	**67%**	**25%**	**28%**	**159%**	**N/A**		
41	**Prod. accur.**	**0.32**	**0.95**	**0.66**	**0.83**	**0.91**	**0.19**	**N/A**		

Note: the Producer's accuracy increased to 0.83 from 0.68 and the 95% confidence interval decreased to ±45 kha from ±74 kha – i.e. the precision of the area estimate of forest loss is almost 40% less than if using the buffer.

6.5 Use the buffer stratum in a post-stratification approach

When using the buffer stratum in pre-stratification approach, the buffer must be part of the sampling design. However, there are situations when this is not feasible. For example, because the use of buffer strata was exemplified in the literature rather recently, countries may have estimated activity data using existing maps without buffer strata and without the possibility to go back to change the sampling design. Post-stratification is necessary in such situations to increase precision in area estimates by use of buffer strata. While post-stratification is typically used with simple random or systematic sampling, in the example of Madre de Dios, we can use the buffer within the forest stratum to create two forest post-strata. Once these post-strata are formed, we proceed with the implementation of the stratified estimation.

1. Now, let's try to decrease the uncertainty in the forest loss area estimate by incorporating the buffer information. In the case of post stratification, we use the buffer information *after* the sample data have been collected. Therefore, we now need to create the buffer to create the two forest post strata. Because we created a buffered stratification above, just copy the stratification with the buffer ("Madre_de_Dios_Buffered.tif") into the post stratification directory.
2. Create an error matrix using the sample in "sample_post_strat.shp" mentioned in point 4 on the previous page:
```
python ~/Desktop/scripts/bin/crosstab.py -v -a ref_label GFC_Madre_de_Dios_Buffered.tif ../4_interpretation/sample_wo_buffer.shp ../5_analysis/errormatrix_post_buffer.txt
```
3. Open the error matrix in a spreadsheet application, copy the strata weights from the pre-stratification spreadsheet, and rename *Forest* to *ForPostStrat1* and the seventh category to *ForPostStrat2*. It should look like below:

17	**ANALYSIS**									
18	***Error matrix, sample counts***									
19		*Non-forest*	*ForPostStrat1*	*Water*	*Forest loss*	*Forest gain*	*Gain/loss*	*ForPostStrat2*	Total	Wi
20	*Non-forest*	47	3	0	0	0	0	0	50	0.0112
21	*ForPostStrat1*	13	504	1	3	0	1	0	522	0.9190
22	*Water*	0	1	49	0	0	0	0	50	0.0033
23	*Forest loss*	2	12	0	36	0	0	0	50	0.0205
24	*Forest gain*	8	18	0	0	24	0	0	50	0.0010
25	*Gain/loss*	10	22	0	5	4	9	0	50	0.0006
26	*ForPostStrat2*	1	26	0	1	0	0	0	28	0.0446
27	Total	81	586	50	45	28	10	0	800	1.0000

4. Create the error matrix of estimated area proportions and compute area estimates, confidence intervals and producer's accuracy:

29	***Error matrix, estimated area proportions***								
30		*Non-forest*	*ForPostStrat1*	*Water*	*Forest loss*	*Forest gain*	*Gain/loss*	*ForPostStrat2*	Total
31	*Non-forest*	0.0105	0.0007	0.0000	0.0000	0.0000	0.0000	0.0000	0.0112
32	*ForPostStrat1*	0.0229	0.8873	0.0018	0.0053	0.0000	0.0018	0.0000	0.9190
33	*Water*	0.0000	0.0001	0.0032	0.0000	0.0000	0.0000	0.0000	0.0033
34	*Forest loss*	0.0008	0.0049	0.0000	0.0147	0.0000	0.0000	0.0000	0.0205
35	*Forest gain*	0.0002	0.0004	0.0000	0.0000	0.0005	0.0000	0.0000	0.0010
36	*Gain/loss*	0.0001	0.0003	0.0000	0.0001	0.0000	0.0001	0.0000	0.0006
37	*ForPosStrat2*	0.0016	0.0414	0.0000	0.0016	0.0000	0.0000	0.0000	0.0446
38	Total	0.0361	0.9349	0.0049	0.0217	0.0005	0.0019	0.0000	1
39	**Area [ha]**	**364,046**	**9,433,835**	**49,924**	**218,623**	**5,220**	**18,809**	**N/A**	
40	**95% CI [ha]**	**128,738**	**153,906**	**34,842**	**72,716**	**1,454**	**34,824**	**N/A**	
41	**MoE**	**35%**	**2%**	**70%**	**33%**	**28%**	**185%**	**N/A**	
42	**Prod. accur.**	**0.29**	**0.95**	**0.64**	**0.68**	**0.91**	**0.06**	**N/A**	

5. As you can see in the error matrix, the buffer captured one of the four omissions of forest loss in the forest stratum. The gain in precision is therefore less compared to pre-stratification example, with the 95% confidence interval dropping to ±72 to ±74 kha, and the producer's accuracy staying the same after post-stratifying. Also, because we can't target the sampling in the buffer area, we might need to increase the size of the buffer when post-stratifying. But by how much is an open question? Future research in this area needs to answer such questions but it's clear from this rather simple example that using a buffer around areas of land change is a good idea.