Pontus Olofsson, Christopher E. Holden, Eric L. Bullock
Earth & Environment, Boston University

BOSTON UNIVERSITY

S4. Methods: Estimation

S4.1 Sample design

S4.1.1 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_newbrunswick.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	5,944,827	60,666,366	1,849,855	7,389,701	4,237,172	506,588
W_i	7.4%	75.3%	2.3%	9.2%	5.3%	0.6%

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: $n \approx \left(\frac{\sum W_i S_i}{S(\hat{P})}\right)^2 = \left(\frac{0.119}{0.005}\right)^2 = 572$ (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 less than 10% of the map except forest and the sample is allocated to strata as (75 units are allocated to forest loss stratum as it is relatively large):

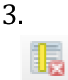
	Non-forest	*Forest*	*Water*	*Forest loss*	*Forest gain*	*For. loss/gain*
n_i	50	300	50	75	50	50

S4.1.2 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 "sample_map.py" and "docopt.py" from *Desktop* to your working directory (or download from https://github.com/ceholden/accuracy_sampler/tree/master/script and https://github.com/docopt/docopt); make sure both files are stored in the same folder.
2. If not using the Virtual Machine but a Windows operating system and *OS4Geo*, click the Windows Start button > *QGIS* > *OSGeo4W Shell*; in the terminal, navigate to the working directory. In the Virtual Machine, open a terminal. Type `python sample_map.py -h` and read about the different options.
3. To select a stratified random sample, type: `python sample_map.py -v --size 575 --allocation "50 300 50 75 50 50" --vector sample.shp stratified stratification_newbrunswick_utm.tif`
4. 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).

S4.2 Response Design

S4.2.1 Interpreting sample

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.
2. Display the shapefile containing the sample, i.e. the file you created in Section 3.
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 labels that correspond to the grid codes of the map: for example, if

the forest loss class has the grid code "4" in the map, then provide each sample unit exhibiting forest loss with label "4". **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. NOTE: 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`

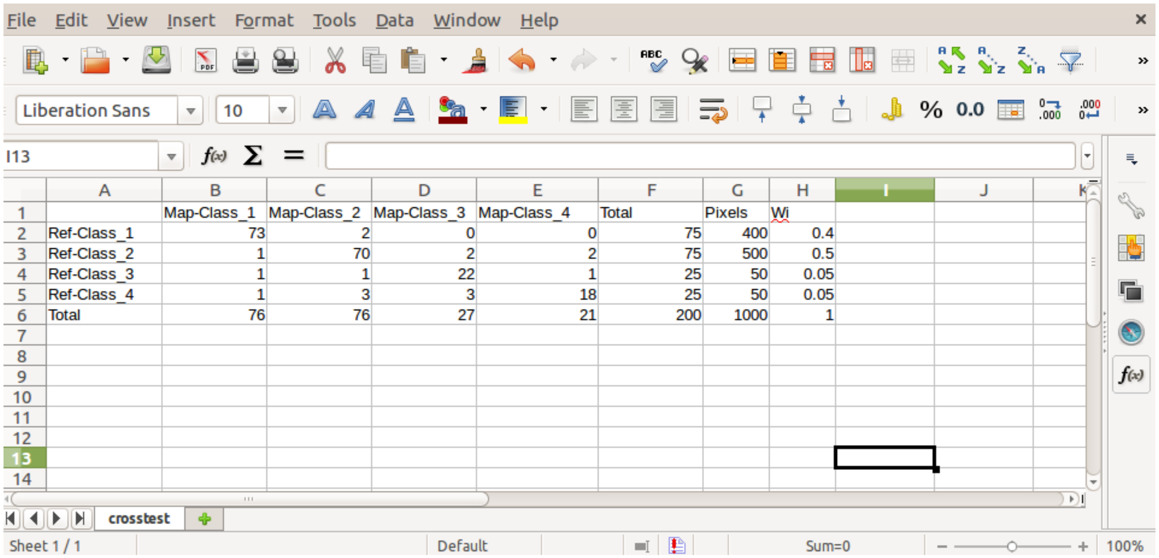
S4.2.2 Construct the error matrix

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; if not present, download the script from https://raw.githubusercontent.com/ceholden/accuracy_sampler/master/script/crosstab.py and place it in the directory where the sample shapefile is located.
2. Open a MATE terminal and navigate to the directory where the sample shapefile and "crosstab.py" are located.
3. Type `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".

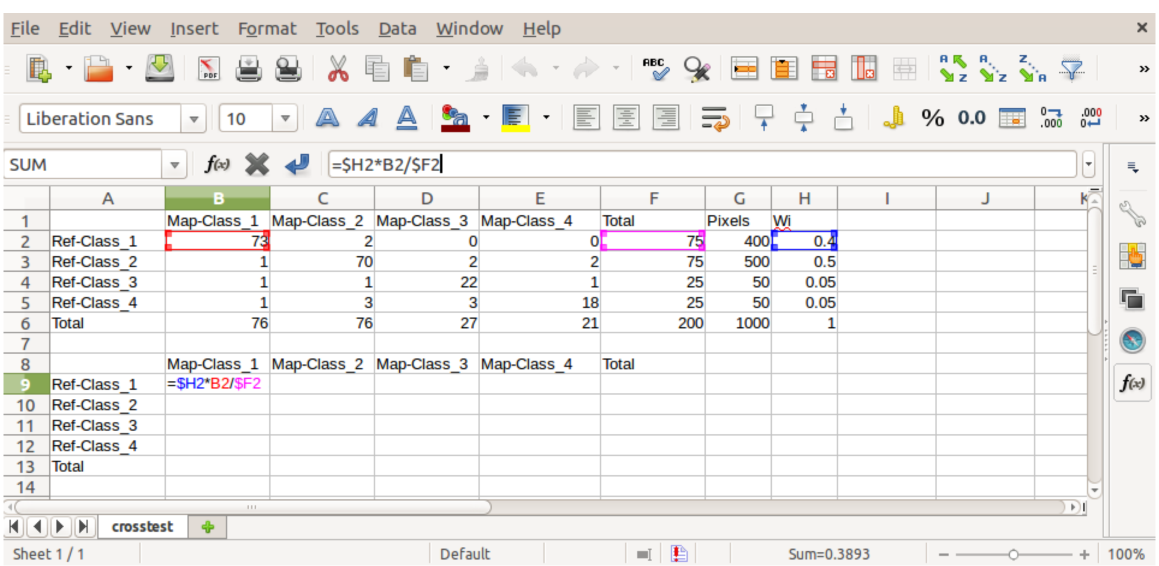
S4.3 Analysis

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.

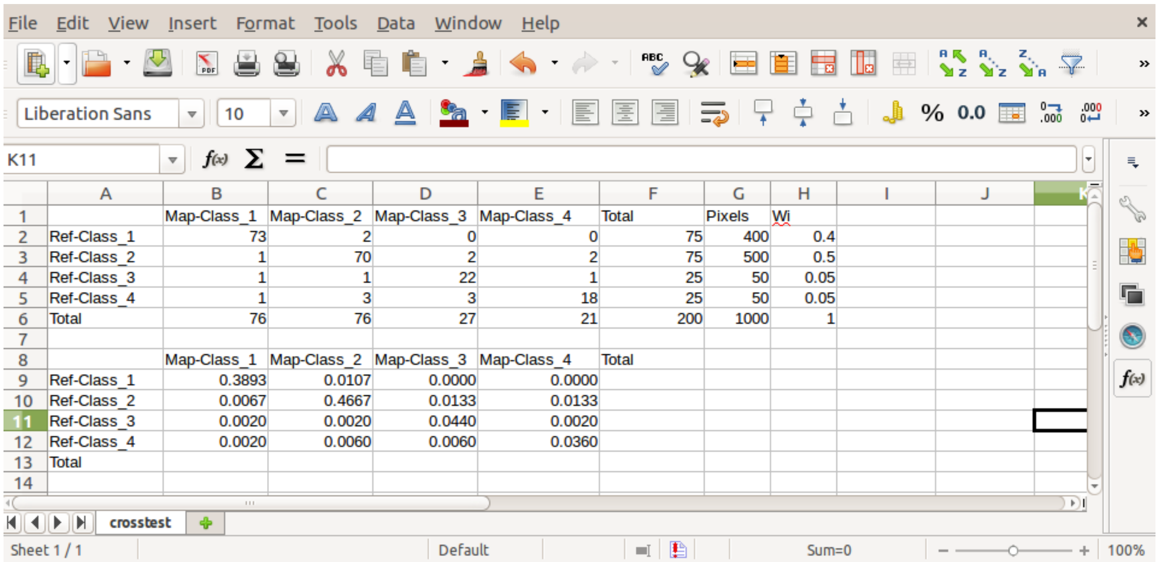
1. The first step of the analysis open the error matrix in a spreadsheet software: open "LibreOffice Calc" from the Desktop menu in the VM (*Office* > *LibreOffice Calc*).
2. In LibreOffice Calc > *File* > Open > browse and open the text file created in subsection S4.2.2 above. The screen should like below:



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 (i.e. copy cells B2:G2 and paste into cells B8:G8, and cells A2:A6 into A9:A13 in the example above).
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 “=$G2*B2/$F2” without the quotation marks; see screenshot below).

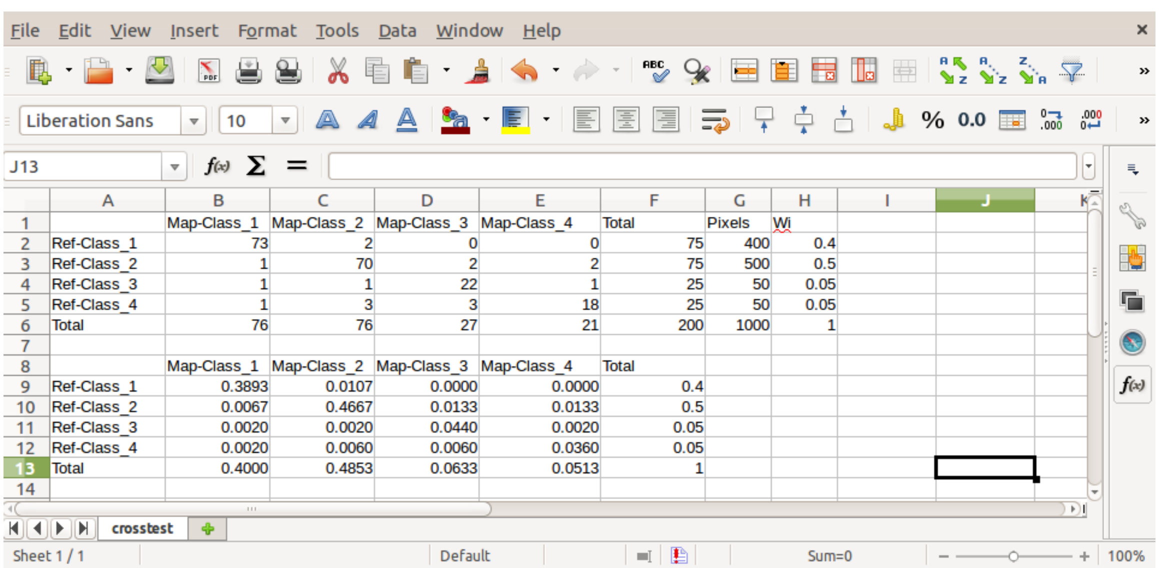


6. Then just populate the rest of the first row of the matrix by highlighting the first cell and then “grabbing” the little black square at the bottom right of the cell (mouse pointer turns into a plus sign) and drag to the end of the row.
7. Then highlight the first row of the matric and and drag down to populate the entire matric; highlight all cells > right click > *Format cells...* > set format to *Number* with 4 decimals. It should look like the screen shot below:



	A	B	C	D	E	F	G	H
1		Map-Class_1	Map-Class_2	Map-Class_3	Map-Class_4	Total	Pixels	Wi
2	Ref-Class_1	73	2	0	0	75	400	0.4
3	Ref-Class_2	1	70	2	2	75	500	0.5
4	Ref-Class_3	1	1	22	1	25	50	0.05
5	Ref-Class_4	1	3	3	18	25	50	0.05
6	Total	76	76	27	21	200	1000	1
7								
8		Map-Class_1	Map-Class_2	Map-Class_3	Map-Class_4	Total		
9	Ref-Class_1	0.3893	0.0107	0.0000	0.0000			
10	Ref-Class_2	0.0067	0.4667	0.0133	0.0133			
11	Ref-Class_3	0.0020	0.0020	0.0440	0.0020			
12	Ref-Class_4	0.0020	0.0060	0.0060	0.0360			
13	Total							
14								

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:

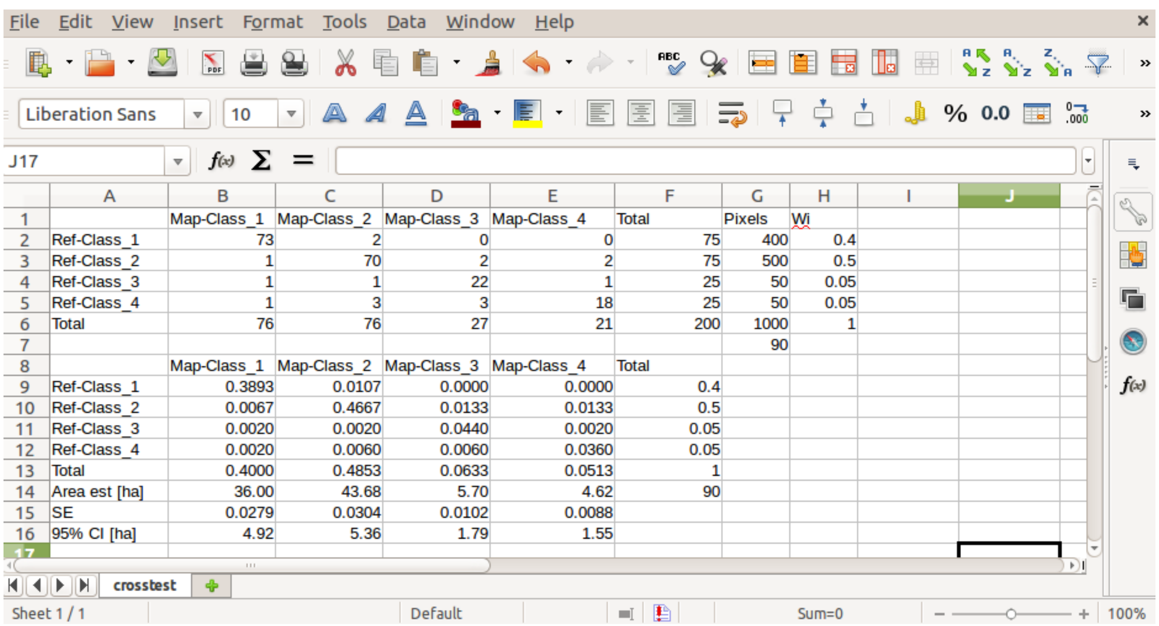


	A	B	C	D	E	F	G	H
1		Map-Class_1	Map-Class_2	Map-Class_3	Map-Class_4	Total	Pixels	Wi
2	Ref-Class_1	73	2	0	0	75	400	0.4
3	Ref-Class_2	1	70	2	2	75	500	0.5
4	Ref-Class_3	1	1	22	1	25	50	0.05
5	Ref-Class_4	1	3	3	18	25	50	0.05
6	Total	76	76	27	21	200	1000	1
7								
8		Map-Class_1	Map-Class_2	Map-Class_3	Map-Class_4	Total		
9	Ref-Class_1	0.3893	0.0107	0.0000	0.0000	0.4		
10	Ref-Class_2	0.0067	0.4667	0.0133	0.0133	0.5		
11	Ref-Class_3	0.0020	0.0020	0.0440	0.0020	0.05		
12	Ref-Class_4	0.0020	0.0060	0.0060	0.0360	0.05		
13	Total	0.4000	0.4853	0.0633	0.0513	1		
14								

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 "=B13*G2*30^2/100^2". Do this calculation on row 14 for all classes. In my example, I get the following unbiased area estimates: 36 ha, 44 ha, 6 ha and 5 ha.
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:

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

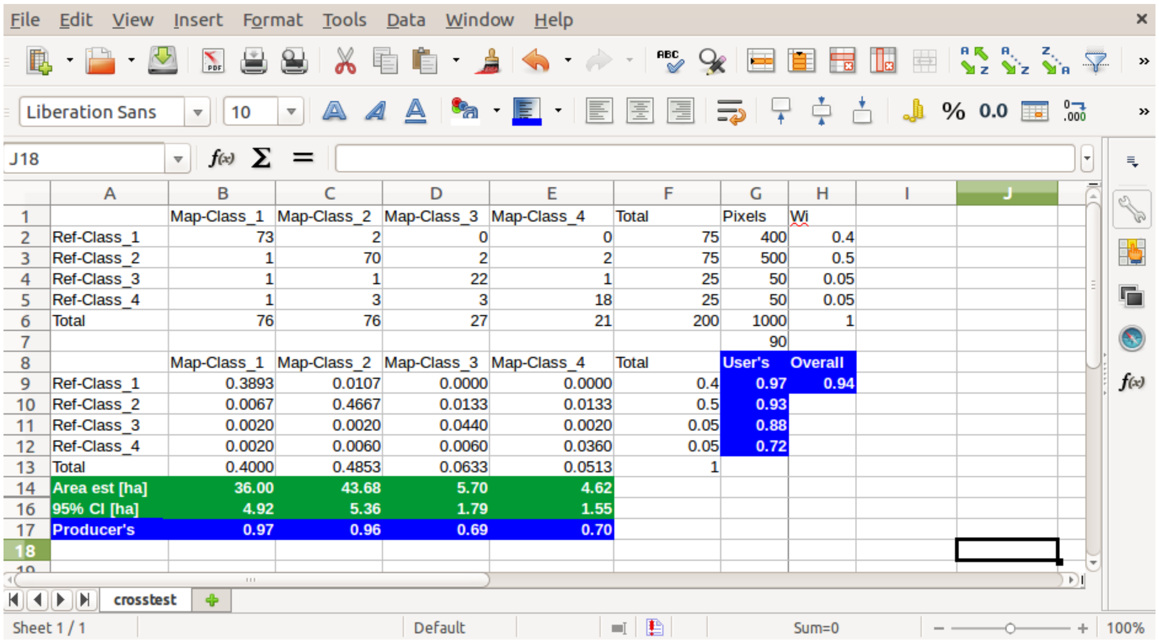
This can be tricky to get right in a spreadsheet! Calculate the standard errors in row 15; the $S(\hat{p}_{\cdot 1})$ which is the standard error for map class 1 (first column total) is calculated as "=SQRT(($H$2*B9 + B9^2)/$G2 + ($H$3*B10 + B10^2)/$G$3 +($H$4*B11 + B11^2)/$G$4+ ($H$5*B12 + B12^2)/$G$5)"; then just can drag the expression to complete the row.
11. Confidence intervals are given by multiplying the standard errors by 1.96. Again, to express the confidence intervals in areal units, multiply by the total map area ("=1.96*B15*$G$6*30^2/100^2"). The spreadsheet should look like below:



	A	B	C	D	E	F	G	H
1		Map-Class_1	Map-Class_2	Map-Class_3	Map-Class_4	Total	Pixels	Wi
2	Ref-Class_1	73	2	0	0	75	400	0.4
3	Ref-Class_2	1	70	2	2	75	500	0.5
4	Ref-Class_3	1	1	22	1	25	50	0.05
5	Ref-Class_4	1	3	3	18	25	50	0.05
6	Total	76	76	27	21	200	1000	1
7							90	
8		Map-Class_1	Map-Class_2	Map-Class_3	Map-Class_4	Total		
9	Ref-Class_1	0.3893	0.0107	0.0000	0.0000	0.4		
10	Ref-Class_2	0.0067	0.4667	0.0133	0.0133	0.5		
11	Ref-Class_3	0.0020	0.0020	0.0440	0.0020	0.05		
12	Ref-Class_4	0.0020	0.0060	0.0060	0.0360	0.05		
13	Total	0.4000	0.4853	0.0633	0.0513	1		
14	Area est [ha]	36.00	43.68	5.70	4.62	90		
15	SE	0.0279	0.0304	0.0102	0.0088			
16	95% CI [ha]	4.92	5.36	1.79	1.55			

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 $\widehat{U}_i = \hat{p}_{ii} \div \hat{p}_{i\cdot}$ and iii) **producer's accuracy** for map category *j* given by $\widehat{P}_i = \hat{p}_{jj} \div \hat{p}_{\cdot j}$ 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 column G ($\widehat{U}_1$"=B9/B13"), producer's in row 17 ($\widehat{P}_1$ "=B9/B13") and overall in H9

("=sum(B9,C10,D11,E12)"). This gives the final spreadsheet with areas in green cells and accuracies in blue cells:



	A	B	C	D	E	F	G	H
1		Map-Class_1	Map-Class_2	Map-Class_3	Map-Class_4	Total	Pixels	Wi
2	Ref-Class_1	73	2	0	0	75	400	0.4
3	Ref-Class_2	1	70	2	2	75	500	0.5
4	Ref-Class_3	1	1	22	1	25	50	0.05
5	Ref-Class_4	1	3	3	18	25	50	0.05
6	Total	76	76	27	21	200	1000	1
7							90	
8		Map-Class_1	Map-Class_2	Map-Class_3	Map-Class_4	Total	User's	Overall
9	Ref-Class_1	0.3893	0.0107	0.0000	0.0000	0.4	0.97	0.94
10	Ref-Class_2	0.0067	0.4667	0.0133	0.0133	0.5	0.93	
11	Ref-Class_3	0.0020	0.0020	0.0440	0.0020	0.05	0.88	
12	Ref-Class_4	0.0020	0.0060	0.0060	0.0360	0.05	0.72	
13	Total	0.4000	0.4853	0.0633	0.0513	1		
14	Area est [ha]	36.00	43.68	5.70	4.62			
16	95% CI [ha]	4.92	5.36	1.79	1.55			
17	Producer's	0.97	0.96	0.69	0.70			