High Resolution LULC Classification Accuracy Assessment Methodology

Introduction

This document defines the Chesapeake Bay Program Partnership's (CBPP) High-Resolution Land Cover Classification (HRLCC) accuracy assessment methodology. For more information on the project regarding its creation and application, visit CBPP Land Use Work Group's website, under the Projects & Resources tab.

Accurate and precise land cover data is critical in planning for a myriad of conservation and restoration goals, including pollution contribution modeling and riparian forest buffer coverage. The increased accessibility of high-resolution imagery has led to the need for updated protocols in the assessment of land cover classification accuracy. In response, the Chesapeake Conservancy has employed an object-based approach and assessed the agreement between multiple reviewers to ensure an unbiased evaluation. Confusion between classes also has been allowed in order to take into account the possibility that one accuracy assessment sample may contain the transition from one land cover class to another. The protocol has been designed such that these conditions do not negatively impact the accuracy of the dataset.

This accuracy assessment methodology has been developed because approaches for evaluating coarser-resolution data, such as land cover derived from Landsat or MODIS satellites, often are not appropriate for high-resolution datasets. Such as in the case with the HRLCC, one-meter data is too precise in scale to support a pixel-based assessment. Whereas the more commonly available 30-meter resolution data allows the analyst to zoom into the area captured by a pixel and, with higher resolution imagery, observe the land cover class within that 30-meter square, at one-meter resolution it is usually too difficult to identify a single pixel's classification without taking the area surrounding it into context. Consequently, object-based accuracy assessments have been proven to be a more effective accuracy assessment methodology for high-resolution land cover classifications. Please refer to *Assessing the Accuracy of Remotely Sensed Data* by Russell G. Congalton and Kass Green for more information.

Definitions

The following terminology is used in this document:

Sample(s)

The base unit of the accuracy assessment workflow. Each sample is a three-meter radius circle, created by buffering a set of points generated with the Equalized Stratified Random technique. A set comprised of a minimum of 50 samples was used to represent the accuracy of each of the individual land cover classes.

Confusion

Method whereby an object containing a progressive transition rather than a hard line between two classified features can be considered correct if assessed to be either class. This can also be applied to objects with distinct boundaries, but that are smaller than the Minimum Mapping Unit.

Minimum Mapping Unit (MMU)
The resolution at which the Chesapeake Conservancy was able to capture, deliberately classify, and hand-correct features. If the features were too small to capture and classify in the rule-based classification, they also were regarded as too small to hand-edit in the classification.

Confusion matrix
Also referred to in accuracy assessments as an error matrix, a table that records information about the correctness of each sample, as well as user's and producer's accuracies for land cover classes.

Supplementary data
External vector datasets, for example planimetric data, that provide two-dimensional geographic representation of features.

Accuracy Assessment Evaluation Data

Accuracy was assessed for CBPP's final statewide land cover data in Maryland, West Virginia, Pennsylvania, Delaware, Washington D.C., and New York. Virginia was evaluated by a separate contractor.

Component imagery of the classifications was used to inspect accuracy of the sets of samples. This included one-meter resolution leaf-on imagery from the United States Department of Agriculture's National Agricultural Inventory Program (NAIP) and one-meter or higher-resolution orthophotos collected on a state-by-state basis. States varied in their most recent orthophoto collection from 2005 to 2015. Imagery source links and years are listed in Table 1.

Table 1: Accuracy assessment imagery sources.

State	NAIP	Orthophotos
Delaware	2013	2012
Maryland	2011 (Harford County) & 2013	2013-2014
New York	2013	2009-2015
Pennsylvania	2013	2005-2008 Cycle 1 Cycle 2
Washington D.C.	2013	2013
West Virginia	2014	2011-2015

Since it was not available across the watershed, supplementary data such as LiDAR and planimetrics was not included in the evaluation. Other imagery sources such as that available on Google Earth or as ESRI Basemaps were not used in the accuracy assessments, as these are not true to the imagery collection years listed in Table 1.

Methods

Sample creation

The Chesapeake Conservancy's accuracy assessment methodology began with the generation of approximately 10,000 equalized stratified random points per state. Each point was then buffered by a radius of three meters to create an area in which accuracy was assessed, henceforth referred to as the accuracy assessment "sample." After the set of points was buffered, it was removed from the assessment workflow. This approach allowed for consideration of context in the landscape, as well as for consideration of consistency between the component images.

Sample Selection

Samples that overlapped with other samples or contained more than one classified land cover type (as determined with the HRLCC dataset) were removed from consideration; this maintained sample uniqueness, and guaranteed that the reviewers were not dividing their decisions over two classes. Of the remaining samples, 75 were randomly selected to represent each land cover class. Based on test runs, this number of samples (i.e. 75) consistently yielded a set of at least 50 after completion of the accuracy assessment protocol, which included opportunities to discard samples. One exception to that was the Barren class, which had an initial set of 150 samples in each state to ensure that the class was properly represented. Barren samples were discarded more often than samples of other classes because the transitional nature of barren land resulted in frequent discrepancies between leaf-on and leaf-off imagery. An additional 200 qualifying samples (i.e. those that did not overlap other samples and contained only a single land cover class) were distributed among each state's classes according to the relative percentages of each land cover class in that state. For instance, a state with 40% forest coverage would be assigned an additional 80 samples to the Forest class. A minimum of 50 samples needed to remain in the set of samples after the review was completed to be input into the confusion matrix.

ArcGIS Pro Format

Using ArcGIS Pro, reviewers observed each sample on top of leaf-on and leaf-off imagery concurrently (see Figure 1 and Figure 2 below) and recorded what they believed to be the land cover class within the sample, based on their interpretation of the imagery. In the case of Figures 1 and 2, an analyst trained on the methods and land cover type definitions of the HRLCC would record Low Vegetation as their assessment of the sample.

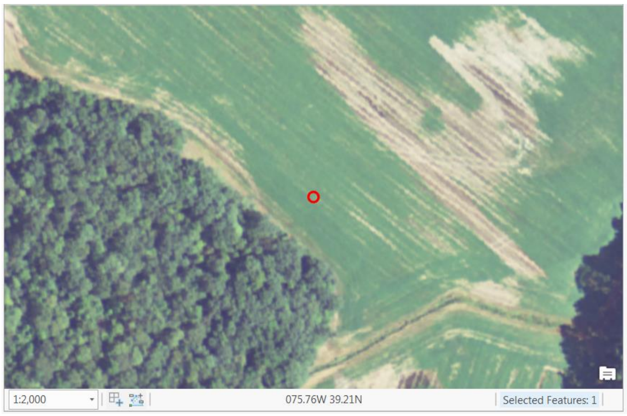


Figure 1. Leaf-on imagery under accuracy assessment sample.

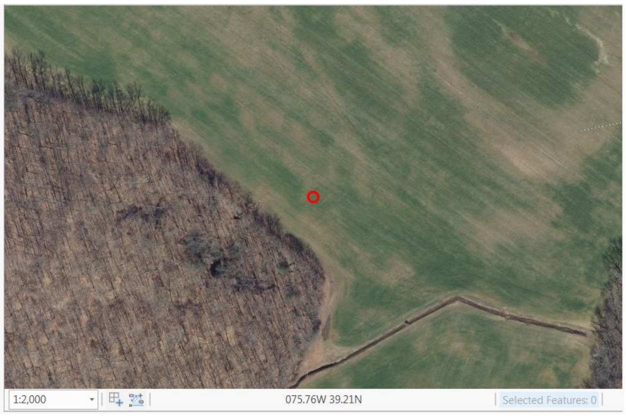


Figure 2. Leaf-off imagery under accuracy assessment sample.

Each sample was reviewed for a primary land cover class and a secondary land cover class by at least two reviewers, and conditionally a third (conditions discussed in "Accuracy Assessment – Reviewer #3" below). Primary choices signified what a reviewer believed to be the most likely land cover class value within the sample, while secondary choices signified what the reviewer believed to be the following most likely land cover class value within the sample. For example, a sample that fell on wet-looking low vegetation within close proximity to water would likely be entered as Wetland as a primary class, and Low Vegetation as a secondary class.

All primary and secondary choices were recorded directly into the attribute table of each sample file, as a numbered classification value:

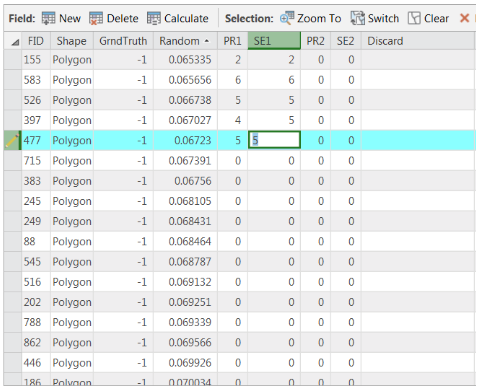
1	=	Water
2	=	Wetlands
3	=	Tree Canopy
4	=	Shrubland
5	=	Low Vegetation
6	=	Barren
7	=	Structures
8	=	Impervious Surfaces
9	=	Impervious Roads
95	=	Tree Canopy over Structures/Impervious Surfaces/Impervious Roads (TCOI)

Accuracy Assessment – Reviewer #1 & Reviewer #2

All samples were assessed by a minimum of two reviewers trained in remote sensing and aerial imagery classification. In order to reduce bias in the assessments, reviewers were unaware of the HRLCC class belonging to the accuracy assessment samples.

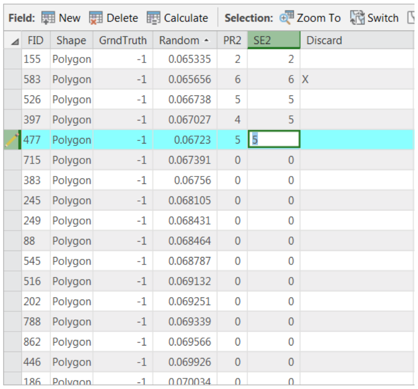
Reviewer #1 randomized the set of samples and marked their primary and secondary choices for each sample in the “PR1” (i.e. Primary Choice, Reviewer #1) and “SE1” (i.e. Secondary Choice, Reviewer #1) fields, respectively; they kept hidden the “Classified” field that contained the actual Land Cover Dataset class value. Next, while keeping Reviewer #1's primary and secondary choices hidden, as well as the “Classified” field, Reviewer #2 went through the same process with the samples in the same order. Reviewer #2’s primary and secondary choices for each sample were marked in the “PR2” (i.e. Primary Choice, Reviewer #2) and “SE2” (i.e. Secondary Choice, Reviewer #1) fields.

If a reviewer was confident that the land cover within the sample area was homogeneous and clearly identifiable, they would record the same land cover type as both their primary and secondary choice, as demonstrated in Figure 3 and Figure 4 below. If the land cover within the sample area appeared to be more than one type, the reviewer was allowed to record a secondary choice that differed from their primary choice.



FID	Shape	GrndTruth	Random	PR1	SE1	PR2	SE2	Discard
155	Polygon	-1	0.065335	2	2	0	0	
583	Polygon	-1	0.065656	6	6	0	0	
526	Polygon	-1	0.066738	5	5	0	0	
397	Polygon	-1	0.067027	4	5	0	0	
477	Polygon	-1	0.06723	5	5	0	0	
715	Polygon	-1	0.067391	0	0	0	0	
383	Polygon	-1	0.06756	0	0	0	0	
245	Polygon	-1	0.068105	0	0	0	0	
249	Polygon	-1	0.068431	0	0	0	0	
88	Polygon	-1	0.068464	0	0	0	0	
545	Polygon	-1	0.068787	0	0	0	0	
516	Polygon	-1	0.069132	0	0	0	0	
202	Polygon	-1	0.069251	0	0	0	0	
788	Polygon	-1	0.069339	0	0	0	0	
862	Polygon	-1	0.069566	0	0	0	0	
446	Polygon	-1	0.069926	0	0	0	0	
186	Polygon	-1	0.070034	0	0	0	0	

Figure 3. Attribute table for Reviewer #1; “Classified” field has been hidden.



FID	Shape	GrndTruth	Random	PR2	SE2	Discard
155	Polygon	-1	0.065335	2	2	
583	Polygon	-1	0.065656	6	6	X
526	Polygon	-1	0.066738	5	5	
397	Polygon	-1	0.067027	4	5	
477	Polygon	-1	0.06723	5	5	
715	Polygon	-1	0.067391	0	0	
383	Polygon	-1	0.06756	0	0	
245	Polygon	-1	0.068105	0	0	
249	Polygon	-1	0.068431	0	0	
88	Polygon	-1	0.068464	0	0	
545	Polygon	-1	0.068787	0	0	
516	Polygon	-1	0.069132	0	0	
202	Polygon	-1	0.069251	0	0	
788	Polygon	-1	0.069339	0	0	
862	Polygon	-1	0.069566	0	0	
446	Polygon	-1	0.069926	0	0	
186	Polygon	-1	0.070034	0	0	

Figure 4. Attribute table for Reviewer #2; “Classified” field and primary and secondary choices for Reviewer #1 have been hidden.

In the special case that the land cover type within a sample's buffered area was not consistent between the leaf-on and leaf-off imagery, either reviewer was able to discard the sample without making note of it. The imagery was occasionally inconsistent for a number of reasons, the two most common being temporal and physical shifting. A temporal change was anything that reflected the passing of time. Leaf-off and leaf-on imagery were not collected at the same time of year, which allowed for landscape alterations in between these periods. One common example of this was housing development: in one set of imagery, a grassy field was present, and in the other, the beginnings of a new suburb were captured. This on-the-ground change resulted in two different land cover classes being reflected in the source imagery. The second most common reason for differences in imagery was shifting due to changes in the angle at which the imagery was collected. This manifested itself as a sample not falling in the same place in both images, while the imagery's land cover types remained unchanged.

In other circumstances, reviewers had to indicate intent to discard a sample. These were instances where the leaf-on and leaf-off imagery were congruent in and around the sample area, but the land cover type within the sample area was indistinguishable (see Figure 5 below). In this situation, if either of the first two reviewers thought that the sample was indistinguishable, they recorded their primary and secondary choices to the best of their ability, plus marked an "X" in the Discard field in the sample attribute table.

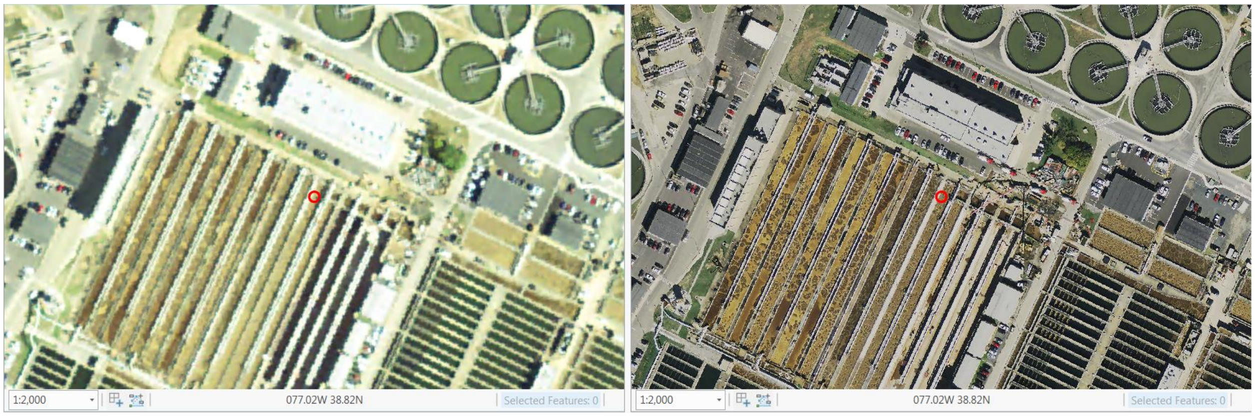


Figure 5. Sample deemed indistinguishable by reviewers.

Initial Accuracy Evaluation

After Reviewer #1 and Reviewer #2 assessed every sample in the set, results were compared to the HRLCC class value and reviewed by an analyst to determine if there was agreement in the sample values. Three designations of a sample were possible after Reviewer #1 and Reviewer #2 completed their assessments: "Correct," "Incorrect," and "Disagree." If the sample was "Correct" under both reviewers, it would be designated as "Correct" and no further evaluation was needed. If the sample was "Incorrect" under both reviewers, it would be designated as "Incorrect" and no further evaluation was needed. If the sample was "Correct" under one of the reviewers and "Incorrect" under the other reviewer, the sample would be designated as "Disagree" which indicated that there was disagreement between the reviewers. The exception to these rules was if there was no agreement between the two reviewers, the sample was marked as "Disagree" regardless of the correctness of either of the individual reviewers.

"Correctness" of a sample was evaluated for each reviewer individually, and the results were evaluated as outlined above. Two scenarios produced a "Correct" sample. 1) A sample was "Correct" if the reviewer marked their primary choice as the actual Land Cover Dataset class value. 2) Alternatively, a sample also was correct if the reviewer marked their primary choice as a class value confusable with the actual HRLCC class type while the secondary choice was the actual HRLCC class value. Otherwise, the sample was "Incorrect." Confusable classes were determined based on classification methods:

- Impervious Surfaces over water (i.e. small docks and piers) were confusable with Water because the MMU of these small impervious features often resulted in them being missed in the land cover classification. These sample confusions were separated by hand from the other samples that had primary and secondary choice combinations of Water and Impervious Surfaces to ensure the features under the samples were indeed docks and open water.
- Wetlands were confusable with Low Vegetation due to the feature extraction methods used for these classes. Wetlands were classified from land cover that had been designated Low Vegetation using a combination of LiDAR derivatives, visual interpretation of imagery, and federal, state, and local ancillary data. Use of ancillary information resulted in conflicts between visual interpretation of imagery during the accuracy assessment and designations from supporting datasets. Because of these mixed methods, confusion was permitted with the parent land cover class.
- For the same reasons, Shrubland was confusable with Tree Canopy. Shrubland was separated from other tall vegetation based on visual interpretation of imagery and LiDAR derivatives. Again, use of ancillary information resulted in conflicts between visual interpretation of imagery during the accuracy assessment and designations from supporting datasets.
- Shrubland also was confusable with Low Vegetation under the special circumstance that the first two reviewers both chose Low Vegetation as their primary choice and Shrubland as their secondary choice. This followed from the feature extraction methods, which, in most counties with poor LiDAR, required reliance only on visual interpretation of imagery. In appearance, tall Low Vegetation looked similar to low Shrubland. Additionally, what was visible in aerial imagery was not represented consistently in LiDAR collected in a different year, as demonstrated in Figure 6. Therefore, interpretations made solely based on visual analysis necessitated this confusion.

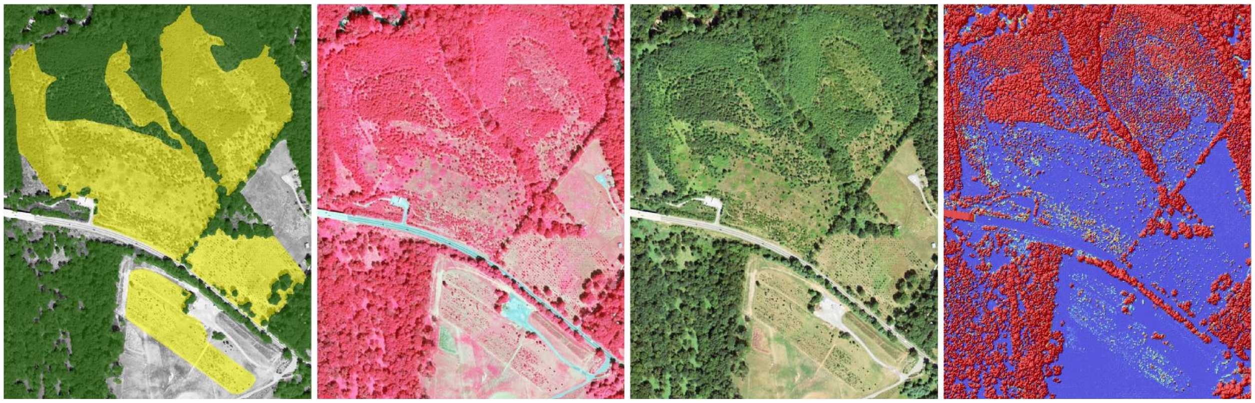


Figure 6. Example of Tree Canopy and Shrubland transition boundary; far-right panel shows LiDAR capture.

- Barren was confusable with Impervious only where the features were interpreted as highly compacted dirt (e.g. dirt roads) or in mines/extractive sites. In these two cases, Barren and Impervious Surfaces were spectrally similar, and were therefore confused in the land cover classification methodology. This mandated allowable confusion in the accuracy assessment.
- Impervious Surfaces were confusable with Impervious Roads. This land cover type was separated from features classified as Impervious Surfaces with supplementary federal, state, and local vector data. Where the road edge was not clear, or the reviewer had some evidence that a feature was a road but not enough to exclude Impervious Surfaces as a possibility, confusion between these classes was permitted. An example of this is highlighted in Figure 7.

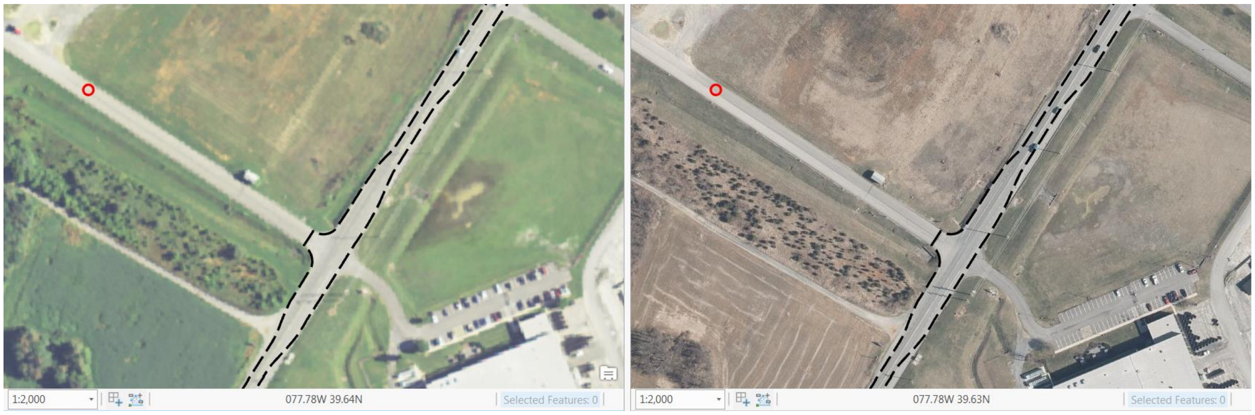
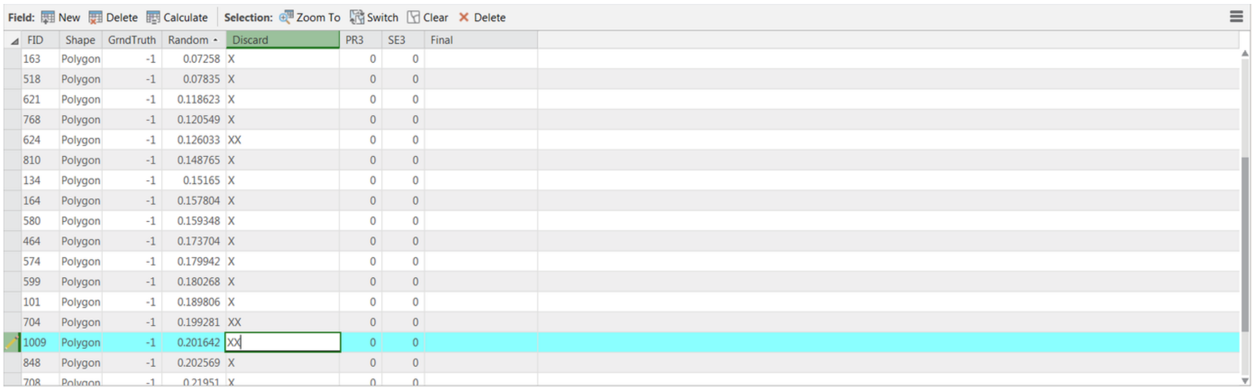


Figure 7. Road planimetric data designated by dashed outline does not cover sample that could be visually interpreted as either an Impervious Road or Impervious Surface.

- Impervious Surfaces were confusable with Structures. This was due to spectral similarity between the two classes, especially in states without LiDAR to clarify structure height.
- Tree Canopy over Structures/Impervious Surfaces/Impervious Roads (TCOI) was confusable with Tree Canopy or the respective features underneath it, again because of feature extraction methods. Each state's TCOI classes were made by intersecting Tree Canopy, extracted from NAIP imagery, with a combination of local planimetric data and the impervious features that were identified in leaf-off orthoimagery. Spatial inaccuracies of planimetric data that intersected with tree canopy (particularly road datasets), difficulties of remotely sensing the actual edge of tree canopy (due to shadows, LiDAR point density, and the porosity of natural canopy which allowed light to reflect off of underlying surfaces), as well as spatial and temporal variability in aerial imagery led to the confusion of these classes.

Accuracy Assessment – Reviewer #3

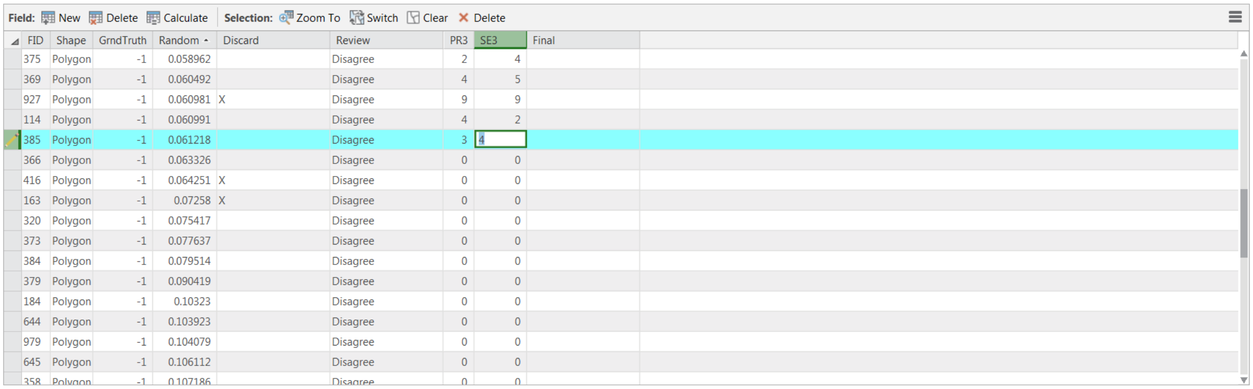
As mentioned above, some samples required further review. A third reviewer analyzed 1) the samples that were marked with an “X” in the Discard field, which indicated that at least one of the previous reviewers had felt the samples were indistinguishable, and 2) samples with results that were in disagreement. The first step for Reviewer #3 was to query out only the samples that had been marked with an “X” in the Discard field by either Reviewer #1 or Reviewer #2. The primary and secondary choices of these reviewers were kept hidden, as well as the HRLCC class value and the initial accuracy evaluation. Then, Reviewer #3 randomized this subset of samples and went through them to assess whether they believed the sample was indistinguishable. If Reviewer #3 believed the sample was indistinguishable, they would add another “X” to the Discard field so that the cell read “XX;” this indicated their agreement that the sample was indistinguishable and thus was not used in the accuracy assessment process (see Figure 8 below). If Reviewer #3 believed the sample was not indistinguishable, they left the single “X” in the Discard field and the sample remained part of the accuracy assessment process.


Field: New Delete Calculate
Selection: Zoom To Switch Clear Delete

FID	Shape	GrndTruth	Random	Discard	PR3	SE3	Final
163	Polygon	-1	0.07258	X	0	0	
518	Polygon	-1	0.07835	X	0	0	
621	Polygon	-1	0.118623	X	0	0	
768	Polygon	-1	0.120549	X	0	0	
624	Polygon	-1	0.126033	XX	0	0	
810	Polygon	-1	0.148765	X	0	0	
134	Polygon	-1	0.15165	X	0	0	
164	Polygon	-1	0.157804	X	0	0	
580	Polygon	-1	0.159348	X	0	0	
464	Polygon	-1	0.173704	X	0	0	
574	Polygon	-1	0.179942	X	0	0	
599	Polygon	-1	0.180268	X	0	0	
101	Polygon	-1	0.189806	X	0	0	
704	Polygon	-1	0.199281	XX	0	0	
1009	Polygon	-1	0.201642	XX	0	0	
848	Polygon	-1	0.202569	X	0	0	
708	Polygon	-1	0.21951	X	0	0	

Figure 8. Attribute table for reviewing samples deemed indistinguishable by either Reviewer #1 and/or Reviewer #2; if sample was deemed indistinguishable, it was marked as “XX.”

Then, Reviewer #3 evaluated all of the samples marked as "Disagree” in the Review field. Reviewer #3 followed the same workflow as Reviewers #1 and #2: the reviewer marked their primary and secondary choices for each sample in the “PR3” (i.e. Primary Choice, Reviewer #3) and “SE3” (Secondary Choice, Reviewer #3) fields, while all other information about the samples was kept hidden. Fields visible to Reviewer #3 matched those shown in Figure 9. This was so that the reviewer would remain blind and unbiased to the first two reviewers’ choices.

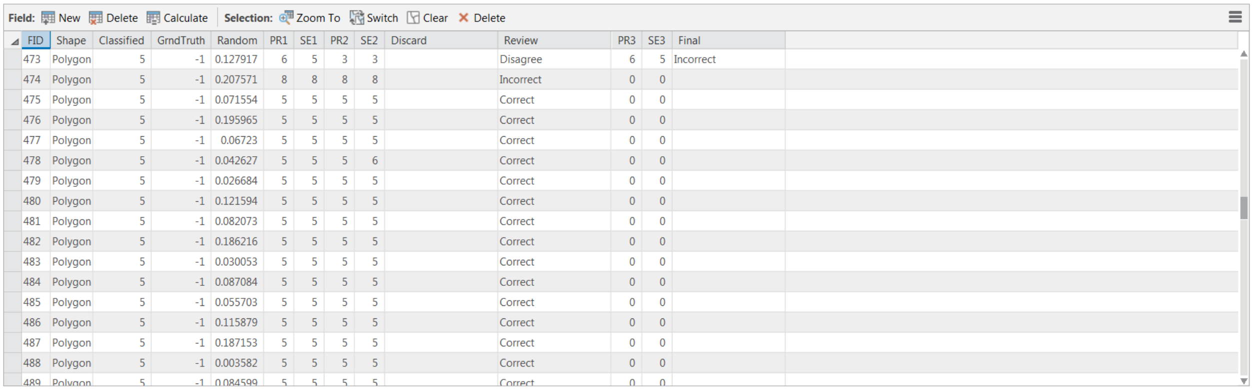

Field: New Delete Calculate
Selection: Zoom To Switch Clear Delete

FID	Shape	GrndTruth	Random	Discard	Review	PR3	SE3	Final
375	Polygon	-1	0.058962		Disagree	2	4	
369	Polygon	-1	0.060492		Disagree	4	5	
927	Polygon	-1	0.060981	X	Disagree	9	9	
114	Polygon	-1	0.060991		Disagree	4	2	
385	Polygon	-1	0.061218		Disagree	3	4	
366	Polygon	-1	0.063326		Disagree	0	0	
416	Polygon	-1	0.064251	X	Disagree	0	0	
163	Polygon	-1	0.07258	X	Disagree	0	0	
320	Polygon	-1	0.075417		Disagree	0	0	
373	Polygon	-1	0.077637		Disagree	0	0	
384	Polygon	-1	0.079514		Disagree	0	0	
379	Polygon	-1	0.090419		Disagree	0	0	
184	Polygon	-1	0.10323		Disagree	0	0	
644	Polygon	-1	0.103923		Disagree	0	0	
979	Polygon	-1	0.104079		Disagree	0	0	
645	Polygon	-1	0.106112		Disagree	0	0	
358	Polygon	-1	0.107186		Disagree	0	0	

Figure 9. Attribute table for Reviewer #3; only samples marked as “Disagree” in the Review field were assessed by Reviewer #3.

Accuracy Calculations

After Reviewer #3 assessed all samples that were marked as “Disagree,” an analyst compared the results to the HRLCC class value to determine the correctness of the samples, and marked the decisions as “Correct” or “Incorrect” in the Final field, as in Figure 10.

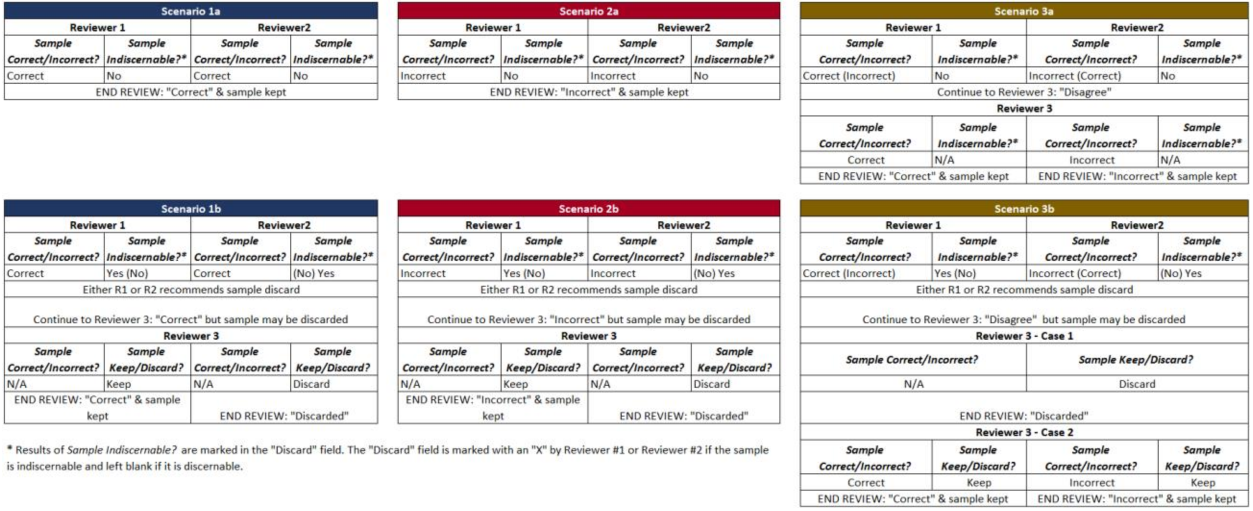


FID	Shape	Classified	GrndTruth	Random	PR1	SE1	PR2	SE2	Discard	Review	PR3	SE3	Final
473	Polygon	5	-1	0.127917	6	5	3	3		Disagree	6	5	Incorrect
474	Polygon	5	-1	0.207571	8	8	8	8		Incorrect	0	0	
475	Polygon	5	-1	0.071554	5	5	5	5		Correct	0	0	
476	Polygon	5	-1	0.195965	5	5	5	5		Correct	0	0	
477	Polygon	5	-1	0.06723	5	5	5	5		Correct	0	0	
478	Polygon	5	-1	0.042627	5	5	5	6		Correct	0	0	
479	Polygon	5	-1	0.026684	5	5	5	5		Correct	0	0	
480	Polygon	5	-1	0.121594	5	5	5	5		Correct	0	0	
481	Polygon	5	-1	0.082073	5	5	5	5		Correct	0	0	
482	Polygon	5	-1	0.186216	5	5	5	5		Correct	0	0	
483	Polygon	5	-1	0.030053	5	5	5	5		Correct	0	0	
484	Polygon	5	-1	0.087084	5	5	5	5		Correct	0	0	
485	Polygon	5	-1	0.055703	5	5	5	5		Correct	0	0	
486	Polygon	5	-1	0.115879	5	5	5	5		Correct	0	0	
487	Polygon	5	-1	0.187153	5	5	5	5		Correct	0	0	
488	Polygon	5	-1	0.003582	5	5	5	5		Correct	0	0	
489	Polygon	5	-1	0.084599	5	5	5	5		Correct	0	0	

Figure 10. Completed attribute table for every reviewer and every sample; confusion matrix is calculated from separate point file.

Finally, the correctness of a sample was determined in the same ways as in the initial accuracy evaluation, summarized in Table 2.

Table 2: Accuracy assessment sample review workflow.



Scenario 1a			
Reviewer 1		**Reviewer2**	
Sample Correct/Incorrect?	***Sample Indiscernable?****	***Sample Correct/Incorrect?***	***Sample Indiscernable?****
Correct	No	Correct	No
END REVIEW: "Correct" & sample kept			

Scenario 2a			
Reviewer 1		**Reviewer2**	
Sample Correct/Incorrect?	***Sample Indiscernable?****	***Sample Correct/Incorrect?***	***Sample Indiscernable?****
Incorrect	No	Incorrect	No
END REVIEW: "Incorrect" & sample kept			

Scenario 3a			
Reviewer 1		**Reviewer2**	
Sample Correct/Incorrect?	***Sample Indiscernable?****	***Sample Correct/Incorrect?***	***Sample Indiscernable?****
Correct (Incorrect)	No	Incorrect (Correct)	No
Continue to Reviewer 3: "Disagree"			
Reviewer 3			
Sample Correct/Incorrect?	***Sample Indiscernable?****	***Sample Correct/Incorrect?***	***Sample Indiscernable?****
Correct	N/A	Incorrect	N/A
END REVIEW: "Correct" & sample kept		END REVIEW: "Incorrect" & sample kept	

Scenario 1b			
Reviewer 1		**Reviewer2**	
Sample Correct/Incorrect?	***Sample Indiscernable?****	***Sample Correct/Incorrect?***	***Sample Indiscernable?****
Correct	Yes (No)	Correct	(No) Yes
Either R1 or R2 recommends sample discard			
Continue to Reviewer 3: "Correct" but sample may be discarded			
Reviewer 3			
Sample Correct/Incorrect?	***Sample Keep/Discard?***	***Sample Correct/Incorrect?***	***Sample Keep/Discard?***
N/A	Keep	N/A	Discard
END REVIEW: "Correct" & sample kept		END REVIEW: "Discarded"	

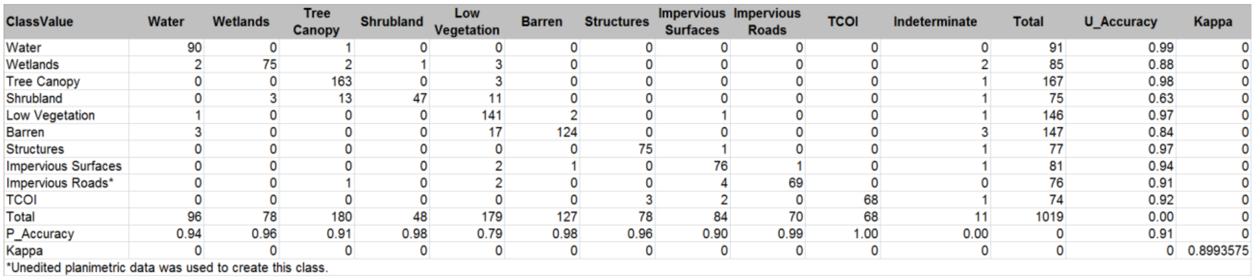
Scenario 2b			
Reviewer 1		**Reviewer2**	
Sample Correct/Incorrect?	***Sample Indiscernable?****	***Sample Correct/Incorrect?***	***Sample Indiscernable?****
Incorrect	Yes (No)	Incorrect	(No) Yes
Either R1 or R2 recommends sample discard			
Continue to Reviewer 3: "Incorrect" but sample may be discarded			
Reviewer 3			
Sample Correct/Incorrect?	***Sample Keep/Discard?***	***Sample Correct/Incorrect?***	***Sample Keep/Discard?***
N/A	Keep	N/A	Discard
END REVIEW: "Incorrect" & sample kept		END REVIEW: "Discarded"	

Scenario 3b			
Reviewer 1		**Reviewer2**	
Sample Correct/Incorrect?	***Sample Indiscernable?****	***Sample Correct/Incorrect?***	***Sample Indiscernable?****
Correct (Incorrect)	Yes (No)	Incorrect (Correct)	(No) Yes
Either R1 or R2 recommends sample discard			
Continue to Reviewer 3: "Disagree" but sample may be discarded			
Reviewer 3 - Case 1			
Sample Correct/Incorrect?		***Sample Keep/Discard?***	
N/A		Discard	
END REVIEW: "Discarded"			
Reviewer 3 - Case 2			
Sample Correct/Incorrect?	***Sample Keep/Discard?***	***Sample Correct/Incorrect?***	***Sample Keep/Discard?***
Correct	Keep	Incorrect	Keep
END REVIEW: "Correct" & sample kept		END REVIEW: "Incorrect" & sample kept	

* Results of *Sample Indiscernable?* are marked in the "Discard" field. The "Discard" field is marked with an "X" by Reviewer #1 or Reviewer #2 if the sample is indiscernable and left blank if it is discernable.

The exception to this was if none of the class values among all three reviewers in the primary choices and secondary choices were the same, then the sample was discarded regardless of the correctness under any of the individual reviewers. This scenario indicated that the sample's land cover type was unknown and, as a result, was not a fair representation of the classification.

Once the entire review was completed, the samples marked with "XX" in the Discard field were removed. Next, a confusion matrix (e.g. Figure 11) was created based on the correctness of each remaining sample.

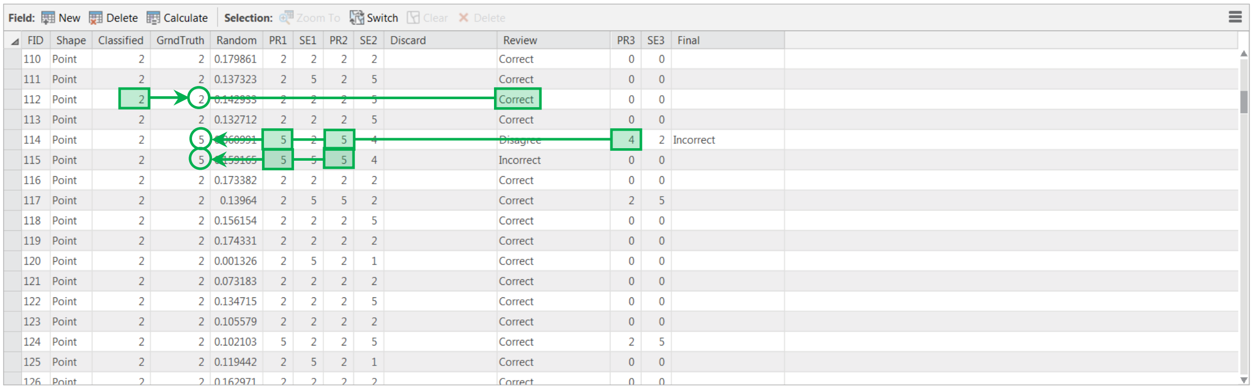


ClassValue	Water	Wetlands	Tree Canopy	Shrubland	Low Vegetation	Barren	Structures	Impervious Surfaces	Impervious Roads	TCOI	Indeterminate	Total	U_Accuracy	Kappa
Water	90	0	1	0	0	0	0	0	0	0	0	91	0.99	0
Wetlands	2	75	2	1	3	0	0	0	0	0	2	85	0.88	0
Tree Canopy	0	0	163	0	3	0	0	0	0	0	1	167	0.98	0
Shrubland	0	3	13	47	11	0	0	0	0	0	1	75	0.63	0
Low Vegetation	1	0	0	0	141	2	0	1	0	0	1	146	0.97	0
Barren	3	0	0	0	17	124	0	0	0	0	3	147	0.84	0
Structures	0	0	0	0	0	0	75	1	0	0	1	77	0.97	0
Impervious Surfaces	0	0	0	0	2	1	0	76	1	0	1	81	0.94	0
Impervious Roads*	0	0	1	0	2	0	0	4	69	0	0	76	0.91	0
TCOI	0	0	0	0	0	0	3	2	0	68	1	74	0.92	0
Total	96	78	180	48	179	127	78	84	70	68	11	1019	0.00	0
P_Accuracy	0.94	0.96	0.91	0.98	0.79	0.98	0.96	0.90	0.99	1.00	0.00	0	0.91	0
Kappa	0	0	0	0	0	0	0	0	0	0	0	0	0	0.8993575

*Unedited planimetric data was used to create this class.

Figure 11. Final confusion matrix showing user's and producer's accuracies for each class value.

In order to do this, the set of samples was converted from the three-meter radius circles back into a new point shapefile that fit the specifications required by ESRI's Compute Confusion Matrix tool. A class value was assigned to each sample in the set so that it could be represented in the confusion matrix. Class value assignments were determined from the "correctness" of each sample. If the sample was labelled "Correct," in the confusion matrix it represented the actual Land Cover Dataset class value. If the sample was labelled "Incorrect" then it represented the majority class value of the primary choices in the confusion matrix (see Figure 12 below). Alternatively, if it was "Incorrect" but there was not a majority class value of the primary choices, it was given an indeterminate value of 97 in the confusion matrix. Labelling it with this value indicated that it was incorrectly classified yet did not have a unified class value among the primary choices to input into the confusion matrix.



FID	Shape	Classified	GrndTruth	Random	PR1	SE1	PR2	SE2	Discard	Review	PR3	SE3	Final
110	Point	2	2	0.179861	2	2	2	2		Correct	0	0	
111	Point	2	2	0.137323	2	5	2	5		Correct	0	0	
112	Point	2	2	0.142933	2	2	2	5		Correct	0	0	
113	Point	2	2	0.132712	2	2	2	5		Correct	0	0	
114	Point	2	5	0.060991	5	2	5	4		Disagree	4	2	Incorrect
115	Point	2	5	0.159165	5	5	5	4		Incorrect	0	0	
116	Point	2	2	0.173382	2	2	2	2		Correct	0	0	
117	Point	2	2	0.13964	2	5	5	2		Correct	2	5	
118	Point	2	2	0.156154	2	2	2	5		Correct	0	0	
119	Point	2	2	0.174331	2	2	2	2		Correct	0	0	
120	Point	2	2	0.001326	2	5	2	1		Correct	0	0	
121	Point	2	2	0.073183	2	2	2	2		Correct	0	0	
122	Point	2	2	0.134715	2	2	2	5		Correct	0	0	
123	Point	2	2	0.105579	2	2	2	2		Correct	0	0	
124	Point	2	2	0.102103	5	2	2	5		Correct	2	5	
125	Point	2	2	0.119442	2	5	2	1		Correct	0	0	
126	Point	2	2	0.162971	2	2	2	2		Correct	0	0	

Figure 12. Assigned class values for each sample to be input into confusion matrix; values are assigned in the "GrndTruth" field.

Results

The entire extent of the HRLCC was mapped with an average of approximately 90% overall accuracy. These results were a measure of the spatial accuracy of the Land Cover Dataset; they reflected the degree to which physical features in the landscape were mapped in the correct locations. Statistics did not take into account whether or not the total area of features was represented accurately in the dataset.

Under those specifications, Delaware achieved a statewide accuracy of 93%, Maryland 91%, New York 88%, West Virginia 86%, Pennsylvania 82%, and Washington D.C. 91%.

Average class accuracies across all states were as follows:

Class	Accuracy	Class	Accuracy
Water	98%	Barren	89%
Wetlands*	80%	Structures	97%
Tree Canopy	98%	Impervious Surfaces	90%
Shrubland†	61%	Impervious Roads	92%
Low Vegetation	95%	TCOI	57%

*excluded WV and NY; this class was not delineated in the Land Cover Dataset for these states
†excluded WV; this class was not delineated in the Land Cover Dataset for this state

Discussion

Accuracy is an important concern when working with any spatial dataset, and objectivity in assessing the accuracy of spatial datasets is a necessity. Since the transition from one land cover type to another is not always clear, unique challenges arise in assessing the accuracy of high-resolution land cover classifications. The Chesapeake Conservancy has developed this accuracy assessment protocol to address these challenges by evaluating multi-pixel samples rather than independent, individual pixels. It utilizes agreement among multiple reviewers to define correctness and compares these definitions against classifications in the Land Cover Dataset. Due to visual and functional similarities between certain land cover types, the approach presented here allows for confusion between classes under special circumstances.