Culvert Classification Random Forest Model

Natural Resource Spatial Informatics Group

Andrew Cooke, Luke Rogers, Jeff Comnick

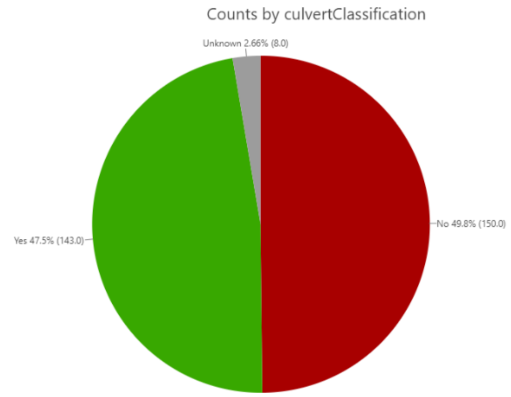
January 19, 2024

Model

A random forest classification model was developed to classify culverts as good matches (properly located on an NHD stream) and bad matches (improperly located on an NHD stream or not on an NHD stream). A balanced training dataset of 301 randomly selected culverts (3.9%) was manually classified. The random forest model was developed in ArcGIS Pro. Full model characteristics presented at the end.

Training Data and Validation Data

The results of the training data classification and model validation are below. Good matches are 'Yes' or 1, bad matches are 'No' or 0, and culverts that could not be classified are 'Unknown' or -1.



Training Data: Confusion Matrix Counts

		Unknown	No	Yes	
actual	Unknown	4	3		7
	No		73	2	75
	Yes		1	72	73
		4	77	74	155

predicted

Training Data: Confusion Matrix Percent

		Unknown	No	Yes
actual	Unknown	0.57	0.43	
	No		0.97	0.03
	Yes		0.01	0.99

predicted

Training Data: Classification Diagnostics

Category	F1-Score	Recall	Precision
1	0.96	0.92	1.00
0	0.96	1.00	0.93
-1	0.83	1.00	0.71

*Predictions for the data used to train the model compared to the observed categories for those features

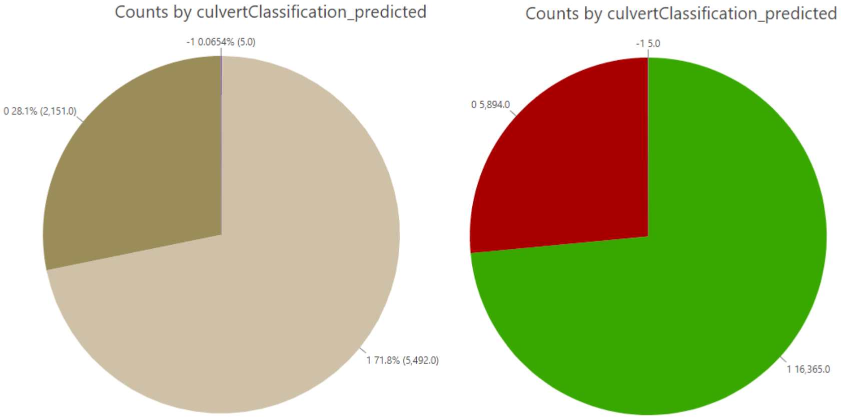
Validation Data: Classification Diagnostics

Category	F1-Score	Recall	Precision
1 (Yes)	0.91	0.83	1.00
0 (No)	0.94	1.00	0.88
-1 (Unknown)	NA	NA	NA

*Predictions for the validation data (excluded from model training) compared to the observed values for those test features

Prediction Results

Two prediction datasets were produced using the model, one for the 7,648 provided culverts, and one for 22,264 barriers of any type. Only culverts were used to train the model, so the impacts on the accuracy for other types of barriers is unknown.



Classification results: culverts (left), all barriers (right)

Validation Data: Counts Predicted

	Yes	No	Unknown	Total
Culverts Only (R2)	5,492	2,151	5	7,648
All Barriers (R3)	16,365	5,894	5	22,264

Data Attributes

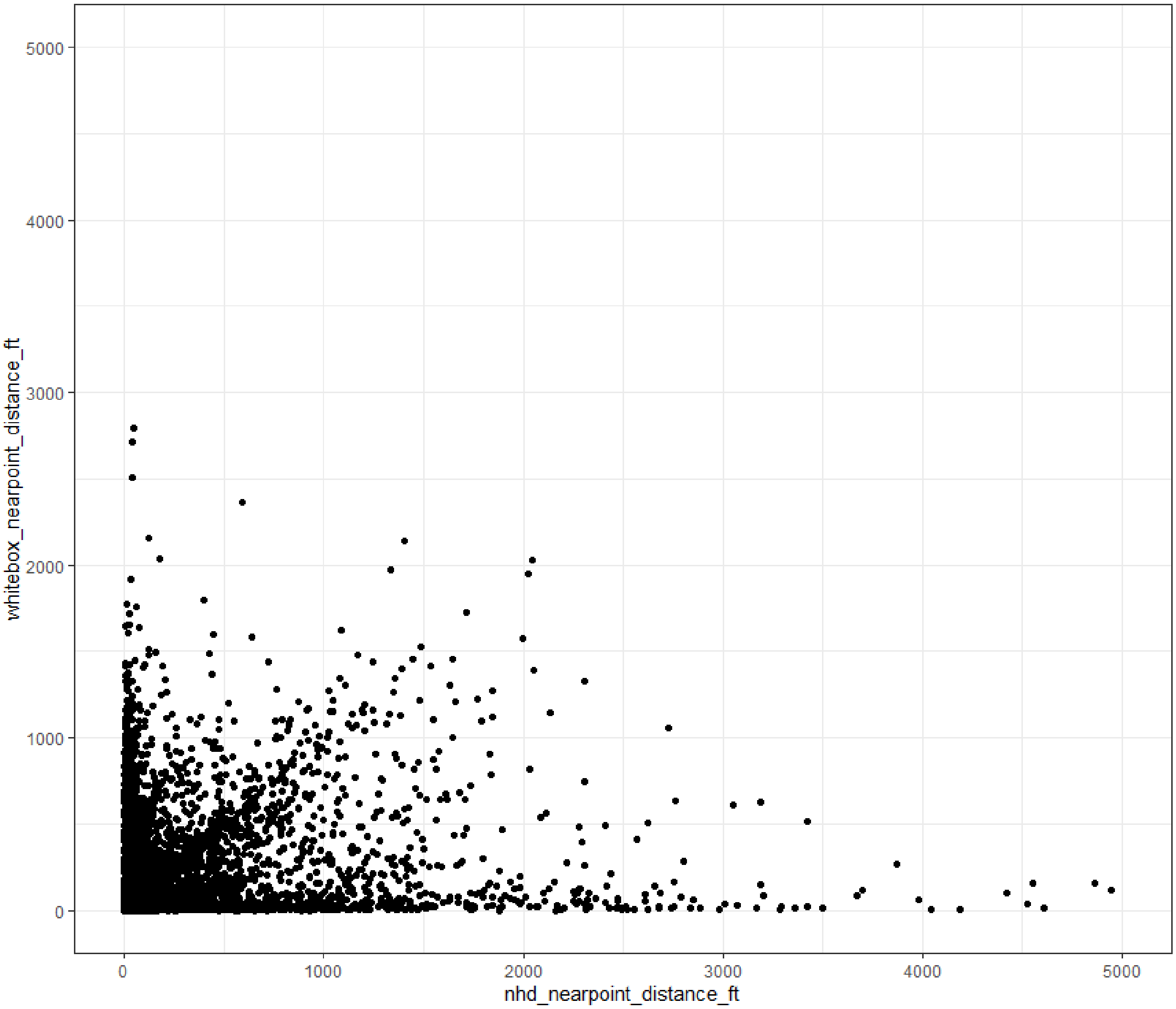
The attributes below were calculated for all barriers with the exception of culvert size and count attributes, which were only calculated for the All Culverts dataset.

Distances

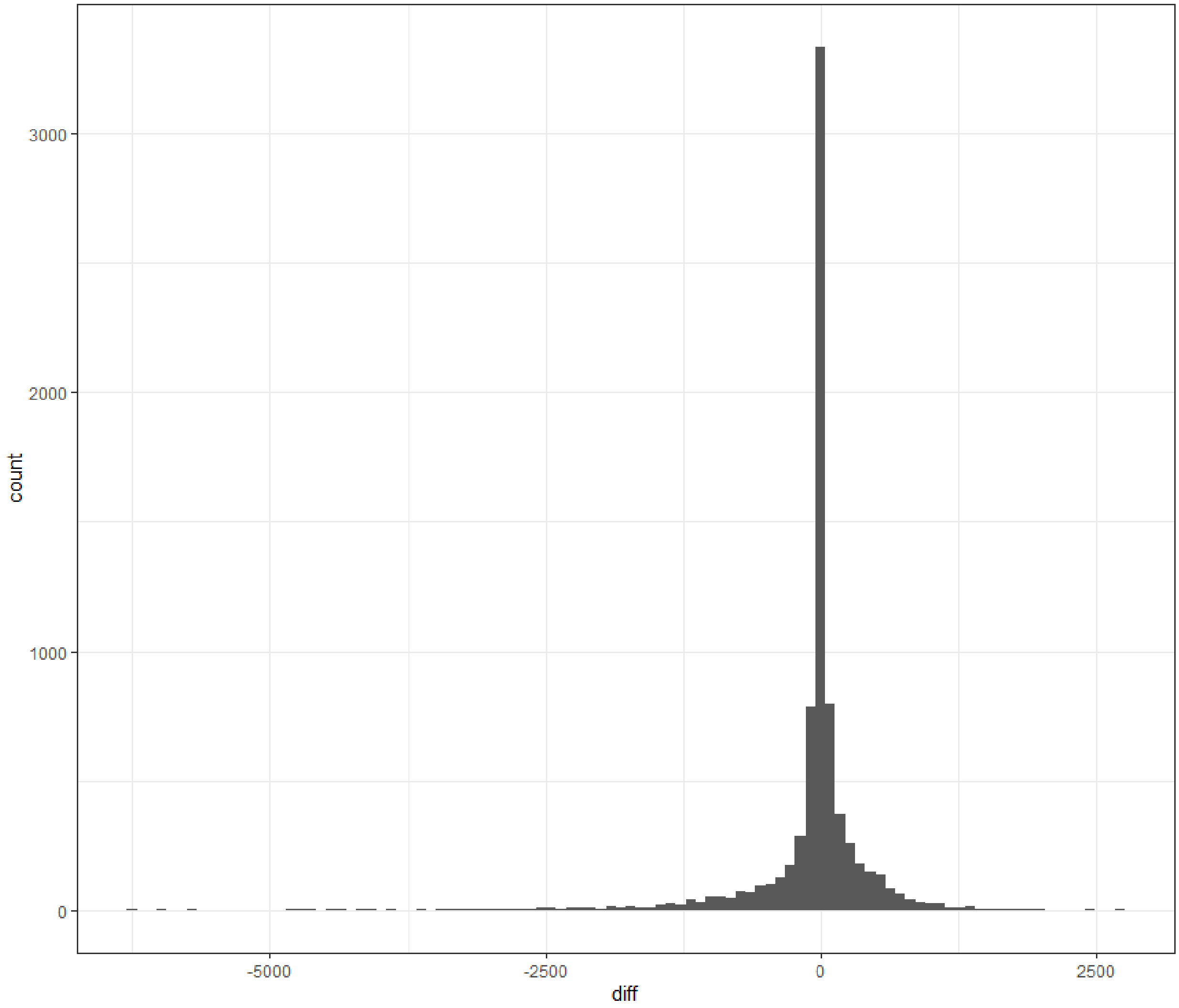
Distances from each culvert to the closest points on any NHD and LIDAR stream segment were calculated along with coordinates and azimuths were calculated.

Name	Description
nhd_nearpoint_distance_ft	Distance in feet from culvert to nearest point on any NHD stream
nhd_nearpoint_x	X coordinate in NAD83 HARN WA State Plane South US Feet (EPSG 2927) of nearest point on any NHD stream
nhd_nearpoint_y	Y coordinate in NAD83 HARN WA State Plane South US Feet (EPSG 2927) of nearest point on any NHD stream
whitebox_nearpoint_distance_ft	Distance in feet from culvert to nearest point on any LIDAR stream
whitebox_nearpoint_x	X coordinate in NAD83 HARN WA State Plane South US Feet (EPSG 2927) of nearest point on any LIDAR stream
whitebox_nearpoint_y	Y coordinate in NAD83 HARN WA State Plane South US Feet (EPSG 2927) of nearest point on any LIDAR stream
whiteboxDistanceMinusNHDDistance_ft	whitebox_nearpoint_distance_ft - nhd_nearpoint_distance_ft; positive when NHD is closer, 0 when distance are the same, negative when LIDAR is closer
distanceBetweenNearPoints_ft	Distance in feet between the nearest point on any NHD stream and the nearest point on any LIDAR stream. If this is small, the points are close together increasing confidence in culvert location
nhd_nearpoint_azimuth	Azimuth in degrees from culvert to nearest point on any NHD stream
whitebox_nearpoint_azimuth	Azimuth in degrees from culvert to nearest point on any LIDAR stream
azmiuthDifference_deg	Difference in degrees between the azimuth to nearest point on any NHD stream and the azimuth to nearest point on any LIDAR stream. If this is small, the points are in the same direction from the culvert, increasing confidence in culvert location
DistanceToNearestNHDRoadIntersection_ft	Distance in feet from culvert to nearest intersection of the NHD streams and NRSIG road dataset. Culverts should be located at the intersection of a road and a stream. The smaller this distance, the more likely a culvert is accurately located.

The chart below is nhd_nearpoint_distance_ft vs. whitebox_nearpoint_distance_ft for the culverts only. There are a large number of culverts close to NHD and far from LIDAR, and vice versa, but there is also a number of culverts near to the one to one line (similar distance to both NHD and LIDAR).



The chart below is histogram of whiteboxDistanceMinusNHDDistance_ft for culverts only. Positive values are culverts that are closer to NHD, 0 are culverts that are the same distance to both stream datasets, and negative values are culverts that are closer to LIDAR.



Stream Similarity

These are used as a confidence metric on stream location. If the NHD segment for a culvert is within the LIDAR buffer, the NHD segment is more likely to be accurately located. If the LIDAR segment for a culvert is within the NHD buffer, it is likely that it represents the same stream segment.

Two buffer sizes were developed, 100 feet and 115 meter. 100-foot buffers with a threshold (inside percentage) of 0.5 or higher are very similar to 115-meter buffers with a threshold of 1.

Name	Description
pctNHDInside100ftLidarBuffer	the percentage of the length of each NHD segment contained within a 100ft buffer on the LIDAR streams
pctNHDInside115mLidarBuffer	the percentage of the length of each NHD segment contained within a 115m buffer on the LIDAR streams
pctLidarInside100ftNHDBuffer	the percentage of the length of each LIDAR segment contained within a 100ft buffer on the NHD streams
pctLidarInside115mNHDBuffer	the percentage of the length of each LIDAR segment contained within a 115m buffer on the NHD streams

Culvert Sizes

Size attributes were parsed from Level A field survey pdfs for culverts only where possible. There may be more than one culvert for each culvert location. These are summarized for each culvert location. Small culverts on main channels or large culverts on small side channels reduce the likelihood that the culvert is on the correct stream.

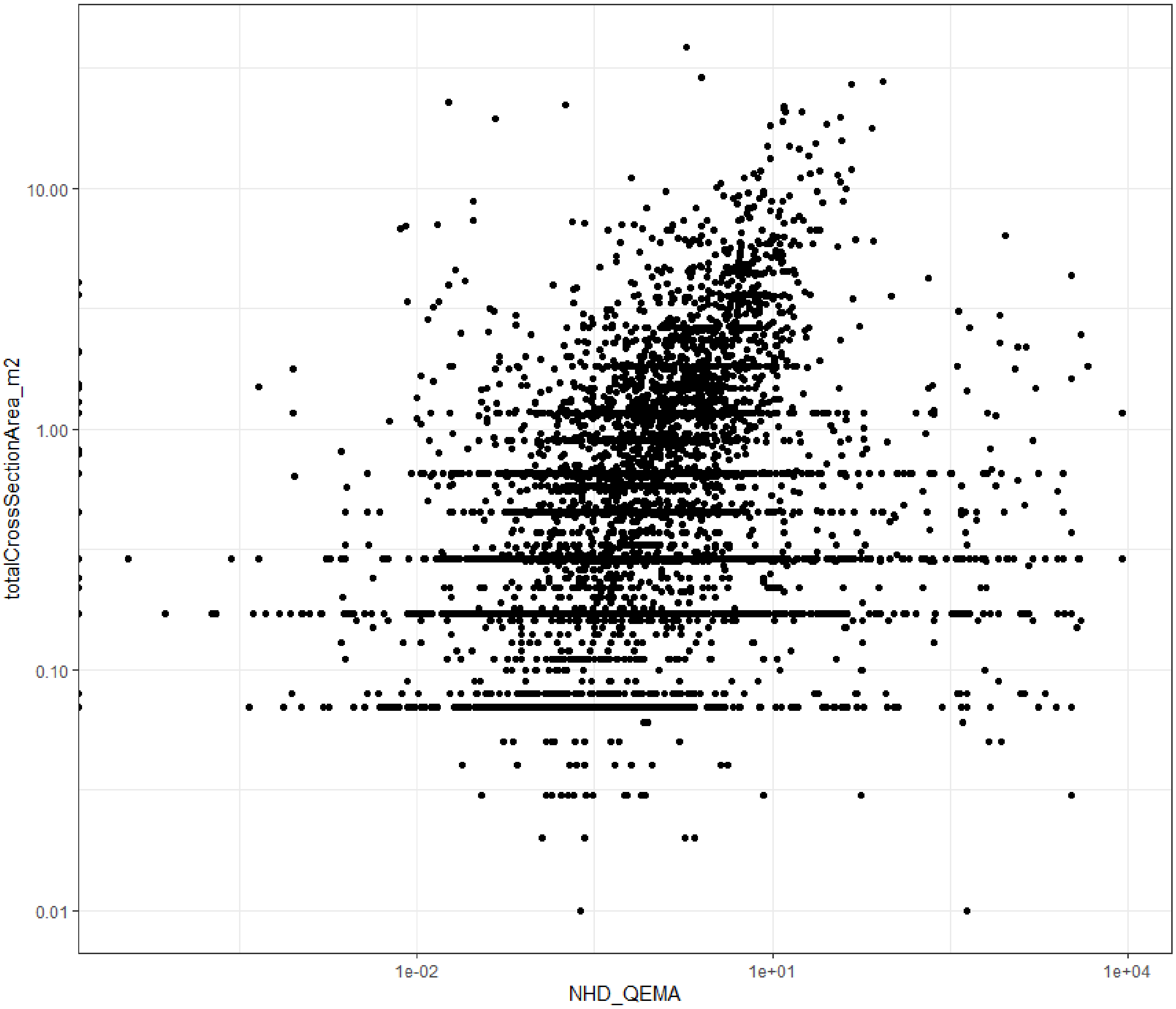
Name	Description
numberOfCulverts	number of culverts at location
totalCrossSectionArea_m2	the total cross-sectional area, in square meters, of all culverts at a location, calculated using Span, Rise, and Shape
largestCrossSectionArea_m2	the cross-sectional area, in square meters, of the largest culvert at the location. If there is only one culvert this will match totalCrossSectionArea_m2
maxSizeClass	The largest size class of all culverts at the location. Culverts are classified into large and small size classes. A large culvert has the cross-sectional area equivalent of a circular culvert with a 0.5m radius or larger. A small culvert is below that size

Stream Flow

Stream flow estimates for the nearest stream segment to each culvert were added from the QEMA attribute in the NHDPlusEROMMA table. According to the User's Guide for the National Hydrography Dataset Plus (NHDPlus) High Resolution, https://pubs.usgs.gov/of/2019/1096/ofr20191096.pdf, “The best EROM streamflow and stream-velocity estimates are the gage-adjusted values, from streamflow calculation step E (NHDPlusEROMMA.QEMA)”.

Culvert Size vs. Stream Flow

It is expected that there should be a relationship between culvert size and stream flow. The chart below is QEMA vs. totalCrossSectionArea_m2 for culverts only. The horizontal bands are specific sizes of culverts. There is a general pattern, but not a tight relationship. Points in the bottom-right corner are small culverts on high flow streams, while points in the upper-left corner are large culverts on low flow streams.



Name Matching

The name of the stream on which a culvert is located is sometimes provided in the field survey reports. There is also sometimes a name on the USGS stream segment nearest to a culvert. These names were compared with the results used as a confidence metric for the culvert location. If the culvert and nearest NHD stream have the same name, it is more likely that the culvert should be on that NHD segment. Named NHD segments are more likely to be accurately located.

Name	Description
stream_name	Name of stream on which the culvert is located according to the culvert field survey
stream_name_simple	Simplified version of stream_name used for matching
tributary_to_name	Name of channel downstream of the culvert location according to the culvert field survey
NHD_GNIS_Name	Name of NHD stream segment according to USGS
NHD_GNIS_Name_simple	Simplified version of NHD_GNIS_Name used for matching
near_name_match_code	Name matching results code (see below)
near_name_match_value	Name matching results value (see below)

Name Matching Results

Code	Value	Description
1	both name	culvert survey and USGS have the same name
2	both NULL	culvert survey and USGS have no name
-1	culvert null nhd name	culvert survey has no name, USGS has name
-2	culvert name nhd null	culvert survey has name, USGS has no name
-3	different names	culvert survey and USGS have different names

Model Characteristics

Number of Trees: 100
Leaf Size: 1
Tree Depth: 5
% of Training Available per Tree: 100
Number of Randomly Sampled Variables: 4
% of Training Data Excluded for Validation: 10

---------------------------------Top Variable of Importance---------------------------------

Variable	**Importance Percentage**
pctNHDInside115mLidarBuffer	0.01
pctLidarInside115mNHDBuffer	0.03
pctNHDInside100ftLidarBuffer	0.04
azimuthDifference_deg	0.04
whitebox_nearpoint_azimuth	0.04
nhd_nearpoint_azimuth	0.05
whitebox_nearpoint_distance_ft	0.05
NHD_QEMA	0.06
DistanceToNearestNHDRoadIntersection_ft	0.08
pctLidarInside100ftNHDBuffer	0.09
distanceBetweenNearPoints_ft	0.12
whiteboxDistanceMinusNHDDistance_ft	0.12
nhd_nearpoint_distance_ft	0.25