Machine Learning Strategies for Enhancing Bathymetry Extraction from Imbalanced Lidar Point Clouds

Kim Lowell, Brian Calder
Center for Coastal and Ocean Mapping/Joint Hydrographic Center
University of New Hampshire
Durham, New Hampshire, UNITED STATES
klowell@ccom.unh.edu (Corresponding author)

***Abstract*—Density-based approaches to extract bathymetry from airborne lidar point clouds generally rely on histogram/frequency-based disambiguation rules to separate noise from signal. The present work targets the improvement of such disambiguation rules by enhancing each pulse with a machine learning-based estimate of its *p(Bathy)* – i.e., its probability of truly being bathymetry. Extreme gradient boosting (XGB) is used to assess the strength of bathymetric signal in pulse return metadata. Because lidar point clouds can be highly imbalanced between *Bathymetry* and *NotBathymetry*, three strategies for mitigating the effects of imbalanced samples were examined. Impacts of an imbalanced lidar point cloud were successfully mitigated by:**

- **Applying an "optimal" decision threshold (ODT) that equalizes accuracy for *Bathymetry* and *NotBathymetry* to *p(Bathy)* rather than using a conventional probability decision threshold (PDT) of 0.50, and**
- **Using proportional class weighting to fit the XGB model.**

However, decomposing a confusion matrix by iteratively discarding misclassified points and re-fitting an XGB model was not successful in improving the strength or detectability of the bathymetric signal in the metadata. The same was true for iteratively discarding correctly classified points.

The bathymetric signal in the metadata was found to be sufficiently strong to explore the operational incorporation of results into the disambiguation rules of density-based bathymertric extraction methods.

***Keywords*—extreme gradient boosting, imbalanced samples, bathymetric lidar, confusion matrix decomposition, probability decision threshold.**

I. Introduction

Airborne lidar ("light detection and ranging") has been an established methodology for mapping shallow water bathymetry (less than ±15m depth) since at least 2005 [1]. Well-established procedures for acquisition of airborne lidar including a recognized data standard [2] exist and are constantly being improved. Similarly, methods and workflows for separating noise from signal in lidar point clouds for terrestrial and aquatic features have also been described and implemented operationally (see, for example, [3]) and are similarly constantly evolving.

Lidar is an active sensor that produces a point cloud from returns of laser pulses. The fundamental difficulty in analyzing or processing lidar data is separating the information desired from the noise. In a terrestrial context, this may mean producing a digital elevation model (DEM) or describing vegetative cover by separating pulse returns that strike the ground from those reflected from vegetation. In bathymetric mapping the primary interest is being able to separate the pulse returns that are reflected from the ocean floor from those reflected from the ocean surface, water column, or other non-bathymetric feature.

A strategy that has been successfully operationalized for bathymetric mapping is based on the "concentration" or "density" of pulse returns. One example for terrestrial and aquatic mapping is the random consensus filter described by [4] that is based on the random sample consensus (RANSAC) approach described by [5]. A more recent example is specific to bathymetry and uses localized uncertainty. CUBE (Combined Uncertainty and Bathymetry Estimator) [6] was originally developed for processing acoustic bathymetric data by estimating the depth for each node of a grid. Since recently being modified to employ a variable-sized grid the new algorithm is known as CHRT (CUBE with Hierarchical Resolution Techniques) [7].

The density-based approaches cited operate essentially by establishing locations of interest (on a spatial grid or other spatial sampling framework), and tabulating pulse return depth frequency histograms for pulses considered to be part of the neighborhood each location. Because the histograms are generally multi-modal, disambiguation rules are applied to each points histogram to select the mode that is most likely to be the true depth for each location.

Airborne bathymetric lidar data, however, comprise more information than the depth associated with each pulse return. Additionally, information is collected about each return; herein we term these data (pulse return) "metadata." The metadata for each pulse return includes information such as scan direction (i.e., fore or aft), return number, and intensity. Furthermore, other metadata can be derived from a lidar point cloud as will be discussed and demonstrated.

The central hypothesis of this article is that lidar pulse metadata contain a sufficiently strong signal to distinguish bathymetric returns from noise. If this is true, then models can be developed that assign to each pulse return an estimated probability – *p(Bathy)* -- of a return being bathymetry. This *p(Bathy)* can then be used to improve the disambiguation rules employed in algorithms like CHRT and RANSAC thereby improving their accuracies and decreasing the amount of manual editing required.

Measuring the strength of the bathymetric signal in the pulse metadata presents potential difficulties since there is currently little understanding how some pulse metadata and associated interactions relate to bathymetry. Hence using existing knowledge to formulate a process-based or empirical model to measure the strength of the pulse metadata/bathymetry relationship is not a viable strategy for model development.

Machine learning[1] (ML) is a collective suite of analytical techniques that can be used to find hidden or unknown relationships in data. Whereas an *a priori* bathymetry/pulse metadata model structure cannot be formulated, ML has the potential to "automatically" develop models that can measure the strength of the pulse metadata-bathymetry signal.

A potential difficulty, however, is that lidar point clouds often contain relatively few returns that are truly bathymetry. That is, airborne lidar data sets can be highly imbalanced – particularly as one approaches the depth limit of the lidar sensor. And because most models are optimized relative to a global cost function, model goodness-of-fit metrics may indicate a strong relationship between bathymetry and metadata even when the model has little practical utility. For airborne bathymetric lidar, clearly it is preferable to develop models that are "optimal" for bathymetry rather than being globally optimal.

In this article, we examine the strength of the bathymetric signal in lidar metadata using ML with a view to tagging each pulse return with a precise *p(Bathy)* estimate. This is done while also considering that the bathymetric signal may be masked due to point cloud imbalance; we consider lidar point clouds to be imbalanced if either *Bathymetry* [2] or *NotBathymetry* pulse returns comprise a relatively large or small proportion of the total. We first fit ML models to estimate *p(Bathy)* based on a suite of pulse metadata variables and examine a number of global and bathymetry-focussed goodness-of-fit metrics. We then explore three options for being able to measure the strength of the bathymetric signal in the lidar point cloud and to improve the performance of the ML models.

II. Study Area, Data, and Pulse Metadata

Four 500m-by-500m tiles of lidar data located within about 12 km of the Key West (Florida, USA) airport were employed as a testbed. These data had been acquired the National Oceanographic and Atmospheric Administration (NOAA) in April 2016. Data were collected over multiple flightpaths and then subsetted to individual tiles – i.e., each tile contained data from multiple flightpaths. Data were captured using a RIEGL™ VQ-880G circular scanning lidar with a 40^0 field of view (20^0 on either side of the airplane) from a nominal altitude of 400m above mean sea level and a nominal speed of 200 km/h. NOAA post-processing classified each return as *Bathymetry* or *NotBathymetry*. Tiles represent a variety of depths and pulse return densities (Table I). Most importantly, they also include a range of sample imbalances with both "excessive" *Bathymetry* (Tile 27285n) and "excessive" *NotBathymetry* (Tile 27075n) in evidence.

TABLE I. DESCRIPTIVE INFROMATION ABOUT DATA TILES.

Tile Identifier[a] (Name)	**Depth Range (m)**	**Return density (returns/m^2)**	***Bathymetry* (%)**	**Number of Points (millions)**
27195n ("Shallow")	-1 to 1	27.6	6.5	7
27285n ("Deep")	-6 to -3	30.4	76.2	8
27080n ("Deeper")	-11 to -7	14.8	21.2	4
27075n ("Deepest")	-16 to -13	13.3	0.4	1

[a] The northern UTM coordinate of the (northern) limit of a tile divided by 100.

Three types of return metadata were employed; we term these return-based, SBET, and lidar edge (Table II).

Most return-based metadata are part of the LAS data standard [2] and are produced during data acquisition. However, three -- *azim_2_pls, pls_frm_heading*, and *inciangle* -- had to be derived by analysis of a combination of the geographic coordinates of each pulse return and the SBET flightpath data that are described in the following paragraph. Return-based metadata describe characteristics of both a pulse and its return(s), but also are hypothesized to describe environmental characteristics at the moment of data acquisition that may impact the information content of pulse returns. For example, the variables *azim_2_pls, inciangle*, and *scan_direct* may detect a wind that is sufficiently strong to impact ocean surface characteristics.

SBET ("Smoothed Best Estimate of Trajectory") data describe the flightpath of the plane. They are provided with the data acquired by the Applanix Corporation and are based on a tightly coupled extended Kalman filter. In addition to providing the location (x, y, z) and yaw, pitch, and roll of the airplane every 0.005 seconds, SBET data include standard deviations of the x, y, z, yaw, pitch, and roll values. These are associated with individual pulses based on time of acquisition. The variables *stdXYZ* and *stdYwPtRl* are measures of airplane

[1] We use "machine learning" as a blanket term that includes "deep learning" and "artificial intelligence" techniques that are sometimes considered to be distinct from "machine learning" techniques. We make no such distinction.

[2] We adopt the convention that *Bathymetry* is italicized when referring to a data class, and that no space is employed in *NotBathymetry*.

stability and may be indicative of ocean surface characteristics that impact the bathymetric information content of individual pulse returns.

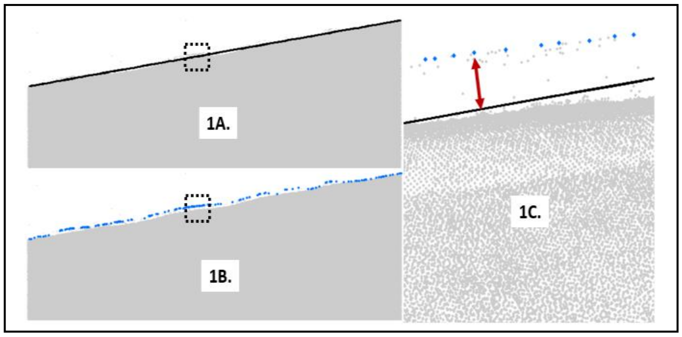
TABLE II. METADATA TYPE, NAME, AND DESCRIPTION.

Type	Name[a]	Definition (range)
Return-based	*Numreturns*	Number of returns from pulse (1 to 6)
	Return_no	Return number (1 to 6)
	Single	1 if numreturns=1; else 0
	First_of_many	1 if numreturns > 1 and single =0; else 0
	Last	1 if numreturns = return_no; else 0
	Rel_return_num	(return_no-1)/(numreturns-1); 0 for single=1 (0 to 1.0)
	Azim_2_pls	Pulse azimuth (0 to 360)
	Pls_frm_heading	Difference from airplane heading and pulse direction (0 to 90)
	Inciangle	Nominal scanning angle of 20^0 corrected for yaw pitch and roll
	Scan_direct	1 if in front of the airplane (fore); -1 if aft (-1 to 1)
SBET	*StdXYZ*	Sum of standard deviations for x, y, and z locations (0.022 to 0.030)
	StdYwPtRl	Sum of standard deviations for yaw, pitch, and roll (0.027 to 0.034)
Lidar-edge	*Absdevia*	Absolute value of orthogonal distance) in m) between point cloud edge and "corner-to-corner" flightpath (0.001 to 8.3)
	Maxabsdev	Maximum of absdevia for each flightpath on a tile (0.6 to 8.3)

[a] Here and throughout the text we adopt the convention that variable names are italicized.

Lidar-edge variables are derived from the crenularity of the edge of the point cloud. Like SBET variables, they have the potential to describe wind and surface characteristics at the time of data acquisition. These are derived by locating the "corner-to-corner" flightpath that describes the straightest possible path from one end of the lidar point cloud to the other (Fig. 1A). Edge points along the lidar point cloud are then identified (Fig. 1B), and the orthogonal distance of each from the corner-to-corner flightpath determined (Fig. 1C). Values for *absdevia* are associated with individual pulses based on time of acquisition. A single value for *maxabsdevia* is associated with all pulse returns for a given flightpath.

Fig. 1. Example showing calculation of *absdevia*. Gray area/points are lidar pulse returns. The dotted box in 1A and 1B is the area of enlargement of 1C. A. Corner-to-corner of flightpath (black). B. Lidar point cloud edge points (blue). C. Orthogonal distance from edgepoint to corner-to-corner flightpath.



III. METHODS/APPROACH

Regularized logistic regression, multi-layer perceptron neural networks, and regularized extreme gradient boosting (XGB) were explored for ML model development. NOAA's *Bathymetry*/*NotBathymetry* classification was used as the dependent variable and the metadata variables described in Table 2 and selected interactions among them were used as independent variables. Because the best-performing models were produced by XGB, for brevity, results for the other two ML methods are not shown.

Regularized XGB is a tree-based approach. Numerous separate "shallow/simplistic" decision trees are iteratively "grown" with each successive tree giving greater weight to those pulse returns having the greatest error. Once convergence is achieved – conceptually the point at which additional trees do not change the cost function value – all trees grown are combined into a single model using a weighted majority vote. That is, trees that provide the greatest improvement to the cost function are weighted most heavily. Regularization drives the impact of some variables to zero (0).

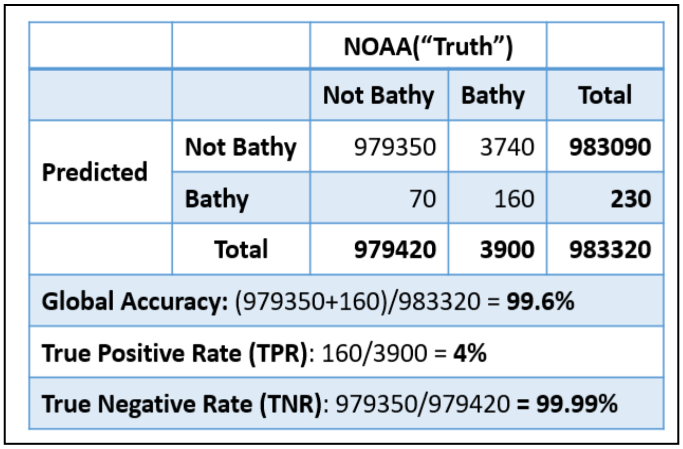
To assess global bathymetric signal strength in the metadata, the metrics employed are R^2 and global accuracy. Because a conventional R^2 based on sums of squares is inappropriate for a categorical classification model, McFadden's pseudo R^2 [8] is employed instead.

To better assess the impacts of sample imbalance and the ability to correctly classify returns for the *Bathymetry* and *NotBathymetry* classes separately, the true positive and true negative rate (TPR and TNR, respectively) expressed as a percentage are employed.

To explain and demonstrate results for all metrics, we employ Tile 27075n – the tile with the greatest sample imbalance – as an example.

Figure 2 is the confusion matrix resulting from fitting an XGB model for Tile 27075n. The R^2 for this model (not shown in Fig. 2) is 0.58 which suggests a reasonable bathymetry signal strength given that the total number of observations is approximately one million (*n*=983300). Global accuracy is high (99.6%) but this is due largely to the extreme class imbalance – i.e., only 0.4% of the pulse returns are B*athymetry*. Essentially, the XGB fitting procedure "learned" that classifying almost all (979350/983090 = 99.99%) pulse returns as *NotBathymetry* optimizes the global cost function.

Fig. 2. Confusion matrix and relevant statistics for an XGB model for Tile 2707500n.



		NOAA("Truth")		
		Not Bathy	**Bathy**	**Total**
Predicted	**Not Bathy**	979350	3740	**983090**
	Bathy	70	160	**230**
	Total	**979420**	**3900**	**983320**

Global Accuracy: (979350+160)/983320 = **99.6%**

True Positive Rate (TPR): 160/3900 = **4%**

True Negative Rate (TNR): 979350/979420 **= 99.99%**

Though analytically sensible, this produces a classification that is of little value for our purposes. The TPR for *Bathymetry* returns makes this apparent. The TPR answers the question "What percent of *Bathymetry* returns were classified as *Bathymetry*?" For this model, the TPR is only 4% again showing the problem of optimizing a global cost function for an imbalanced sample – particularly when it is the minority class that is of interest. In contrast, the TNR is virtually 100% again demonstrating that the XGB optimized the global cost function by classifying all pulses as the majority class *NotBathymetry* at the cost of incorrectly classifying almost all of the minority *Bathymetry* class. And, of course, it is the minority *Bathymetry* class that is of greatest interest in bathymetric mapping. Hence despite the high global accuracy, the large difference between the TPR and the TNR is indicative of a large sample imbalance.

Three strategies that were explored to overcome the effects of such a sample imbalance. We refer to the first as "Optimum Decision Threshold" (ODT).

A. Optimum Decision Threshold

XGB models (as well as the other two ML techniques examined) do not in reality classify each pulse return as *Bathymetry* or *NotBathymetry*. In fact, the XGB model estimates *p(Bathy)* -- the probability that each pulse return is *Bathymetry*. The confusion matrix in Figure 2 is subsequently produced by assigning a pulse return to *Bathymetry* if its *p(Bathy)* is above a certain probability decision threshold (PDT) – conventionally 0.50 for a binary classification. However, if a sample is imbalanced, the signal of the minority class may be so weak that it makes more sense to use a different PDT. Recognising that effectively the TNR was maximized at the expense of the TPR for Tile 27075n, we formulated the ODT as being the PDT at which the TNR and TPR are equal.

Fig. 3. A. Receiver operating characteristics (ROC) curve. B. Probability decision threshold where TPR (blue) is equal to the TNR (red). See text for explanation of dotted lines and arrows.

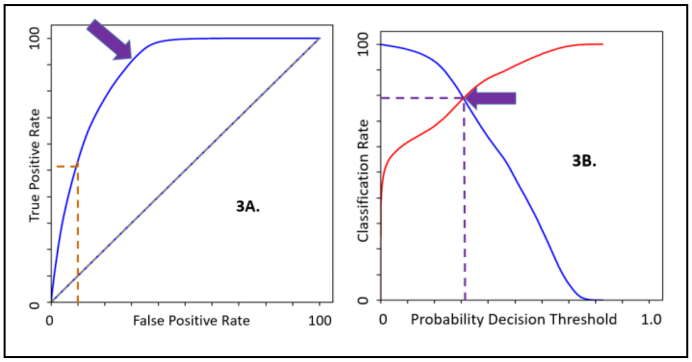
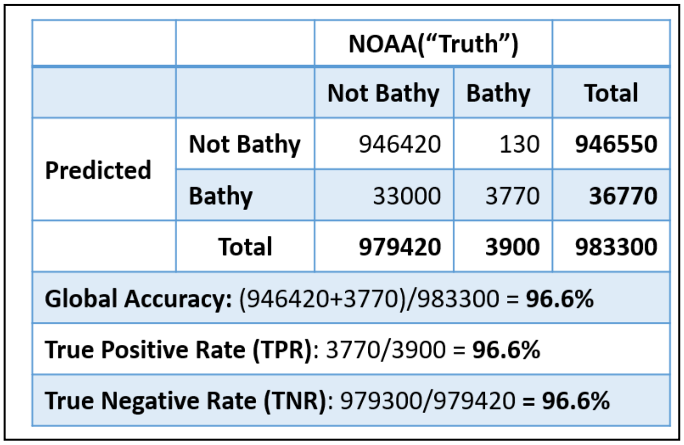


Figure 3A shows a representative receiver operating characteristics (ROC) curve – a commonly used metric for evaluating the quality of a binary classification (see, for example, [9]). A ROC curve plots the TPR against the False Positive Rate (FPR) over the full range (0 to 1.0) of PDTs. The straight diagonal line represents a random classification; a perfect classification results in a ROC curve with a vertical line located at 0.0 on the x/FPR axis and a horizontal line located at 100 on the y/TPR axis. In this example, the FPR and the TPR for the 0.50 PDT are represented by the orange dotted line. This produces a FPR of 10% (i.e., a TNR of 90%) and a TPR of 50%. The inflection point on the ROC curve indicated by the purple arrow is the point at which the TNR and TPR are equal. The PDT associated with this point – the ODT -- is found by plotting the TPR and TNR over the range of possible PDTs (0 to 1.0) (Fig. 3B). In this example, the ODT, is about 0.30 resulting in a TPR and TNR of about 80% accuracy.

For Tile 27075n the ODT was 0.02. That the ODT is small and very different from the conventional PDT of 0.50 reinforces the idea that because of the extreme *Bathymetry*/*NotBathymetry* imbalance, the bathymetric signal is weak. In fact, it is so weak that any pulse return with a *p(Bathy)* above 0.02 may in reality be *Bathymetry*. Confusion matrix and relevant statistics for the XGB model for Tile 27075n using the Optimum Decision Threshold (0.02).

Fig. 4. Confusion matrix and relevant statistics for the XGB model for Tile 2707500 using the Optimum Decision Threshold (0.02).



		NOAA("Truth")		
		Not Bathy	**Bathy**	**Total**
Predicted	**Not Bathy**	946420	130	**946550**
	Bathy	33000	3770	**36770**
	Total	**979420**	**3900**	**983300**
Global Accuracy: (946420+3770)/983300 = **96.6%**				
True Positive Rate (TPR): 3770/3900 = **96.6%**				
True Negative Rate (TNR): 979300/979420 **= 96.6%**				

Figure 4 presents the confusion matrix and metrics resulting from an ODT of 0.02 for Tile 27075n. While the ODT has decreased the global accuracy compared to the PDT of 0.50 (Fig. 2), more importantly, it has improved the TPR considerably. Trade offs for this improvement that are of little concern in bathymetric mapping is that the TNR and the global accuracy have decreased slightly. Potentially of greater concern is that to achieve a TPR of 96.6% using the ODT, 33000 pulse returns that are *NotBathymetry* have been classified as *Bathymetry*. If the XGB model were to be employed as a stand-alone classification, this would mean that the User's Accuracy [10] for *Bathymetry* would be only 10% -- i.e., only 3770 of the 36770 pulse returns identified as *Bathymetry* truly are *Bathymetry*. Thus a bathymetric chart produced from these 36770 pulse returns may be of dubious quality. However, in the alternative context of this work where results are intended to enhance an existing bathymetric extraction methodology, the use of the ODT demonstrates that despite the *Bathymetry*/*NotBathymetry* sampling imbalance, the XGB model is clearly able to detect the *Bathymetry* signal.

B. *Class Weighting/Resampling*

The second general approach explored for overcoming sample imbalance and understanding bathymetric signal strength comprises two methods: resampling and weighting. The fundamental strategy of both is to overcome the class imbalance by “emphasizing” the minority class (or class of interest in the model fitting process).

In resampling, new observations representing the minority class are created (i.e., “oversampling”) using the characteristics of the minority class observations, and/or observations from the majority class are removed (“undersampling”) to create a new more balanced data set on which a ML model is fit. ADASYN (Adaptive Synthetic Sampling) [11] and SMOTE (Synthetic Minority Oversampling Technique) [12] are among the resampling strategies that have been described; both were explored in this study.

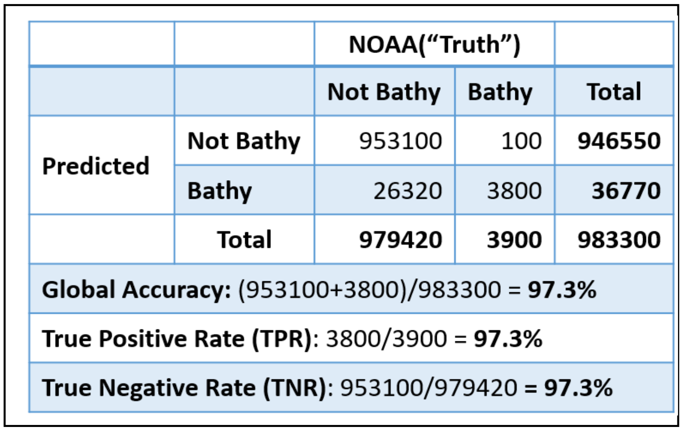
In weighting, prediction errors associated with the minority class observations are given a greater weight than those of the majority class. This complements the general XGB model-fitting approach that prioritizes observations having the greatest prediction error. We explored proportional weighting for the minority class – whether *Bathymetry* (e.g., Tile 27075) or *NotBathymetry* (Tile 27285). In proportional weighting, each observation is given a weight based on whether it is *Bathymetry* or *NotBathymetry*:

$$W_{\text{Bathy or NotBathy}} = (T/P_{\text{Bathy or NotBathy}})-1)/2 \qquad (1)$$

where W is the weight for the *Bathymetry* or *NotBathymetry* class, T is the total number of observations, and P is the number of observations of the minority or majority class. This proportional weighting has the desirable property that in a balanced binary data set, observations from both classes are given equal weight.

Proportional weighting provided better results than ADASYN and SMOTE. Figure 5 shows the confusion matrix for Tile 27075n produced by a proportionally weighted XGB model with the ODT applied. Weighting has clearly improved the XGB model with the TPR and TNR increasing (compare with Fig. 4). Most notably, the number of false positives has fallen from 33000 to 26320 – a reduction of 20%.

Fig. 5. Confusion matrix and relevant statistics for the base XGB model for Tile 2707500 using the Optimum Decision Threshold (0.66).



		NOAA(“Truth”)		
		Not Bathy	**Bathy**	**Total**
Predicted	**Not Bathy**	953100	100	**946550**
	Bathy	26320	3800	**36770**
	Total	**979420**	**3900**	**983300**
Global Accuracy: (953100+3800)/983300 = **97.3%**				
True Positive Rate (TPR): 3800/3900 = **97.3%**				
True Negative Rate (TNR): 953100/979420 **= 97.3%**				

It is also of interest that the ODT for the weighted model is 0.66. That this is closer to the conventional PDT of 0.50 demonstrates that proportional weighting is a useful strategy for improving the bathymetric signal for Tile 27075n.

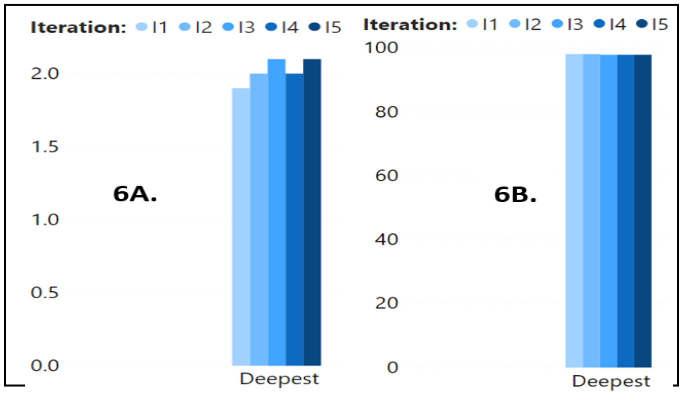
C. *Confusion Matrix Decomposition*

The final approach to strengthening and detecting the bathymetric signal in ML models is iterative resampling of the confusion matrix. In a confusion matrix, the correctly classified diagonal elements represent the observations with the strongest signal for their respective classes, and the incorrectly classified off-diagonal elements represent those with the weakest signals.

The idea behind iterative confusion matrix decomposition is that XGB models fit only on observations having the strongest or weakest signals will better “distill” the bathymetric signal for imbalanced samples, particularly since the majority class will initially have the most observations discarded. Such models can then be applied to all observations and the process repeated until a “best model” is produced. In essence, it is hoped that at each iteration a large number of highly certain or uncertain majority class observations will be removed thus increasing the proportion of minority class observations in the data set produced and resulting in XGB models fitted from the “essence” of the bathymetric signal. This was explored with a maximum of five iterations employed.

Figure 6 shows the outcome of iteratively removing false positives (FPs) and false negatives (FNs) for Tile 27075n. If this decomposition strategy is successful, the FPR (which is identical to the FNR because of the application of the ODT to the *p(Bathy)* values) (Fig. 6A) would decrease with each iteration and the TPR (identical to the TNR) (Fig. 6B) would increase. This did not occur for Tile 27075n.

Fig. 6. Accuracy rates for successive model iterations following removal of false positives (FPs) and False Negatives (FNs). A. False Positive (& False Negative) Rate. B. True Positive (& True Negative) Rate for Tile 27075n using the Optimum Decision Threshold (0.02).



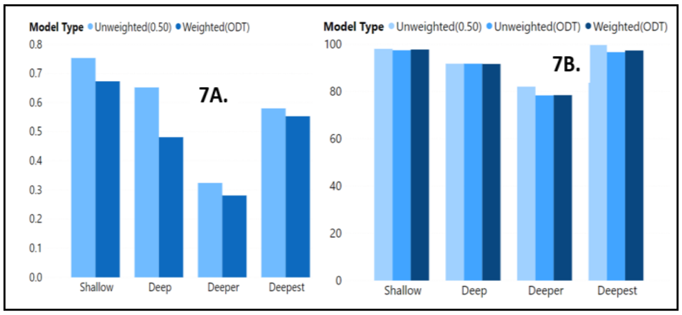
IV. Results and Discussion

Results for Tile 27075n have been presented as a means of explaining the analytical methodology. For robust evaluation of methods, however, the range of characteristics – depth, level of sampling imbalance, etc. – across all four tiles must be examined.

Recall that the overall goal of this work is to evaluate the strength of the bathymetric signal in the metadata variables considered (Table II). Furthermore, in recognition of the sampling imbalance in lidar point clouds, a further goal is to find ways of mitigating the impacts of the sampling imbalance thereby enabling better detection of the bathymetric signal to produce more precise estimates of *p(Bathy)*.

The R^2 (Fig. 7A) and the global accuracy (Fig. 7B) for the XGB models suggest that the signal strength for bathymetry in the metadata is reasonably strong[3]. Nonetheless, the danger of relying on such global statistics has been demonstrated and discussed. In particular note that the global accuracy is highest for the Shallow (27195n) and the Deepest (27075n) tiles that also are the most imbalanced. The remaining two tiles are approximately equally 80/20 imbalanced with *NotBathymetry* comprising 20% of the Deep tile and *Bathymetry* comprising 20% of the Deeper tile.

Fig. 7. (Pseudo) R^2 values (left) and global accuracy (right) for XGB models. R^2 values for the Unweighted model that employs the ODT are the same as for the Unweighted model that employs a conventional probability decision threshold of 0.50.

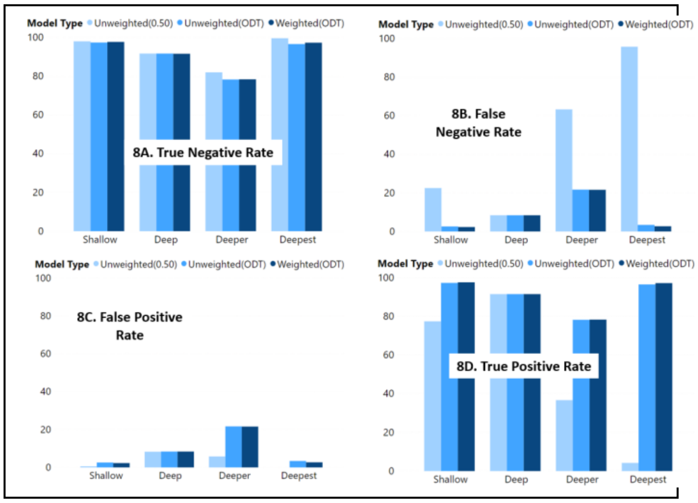


Also apparent from Figure 7 is that these global measures suggest that an unweighted XGB model that employs a conventional PDT of 0.50 performs best. This is understandable given that the cost function employed in XGB model fitting targets global optimization based on an implicit assumption that all/both classes are equally important. Hence of greater interest than global metrics is how the XGB models perform for *Bathymetry*.

Figure 8 presents class-specific information for each tile and model type. This demonstrates the ability to mitigate class imbalance of both the use of the ODT, and fitting weighted XGB models. For TNR and FPR (Figs. 8A and 8C), the unweighted models with a conventional 0.50 PDT perform better than or as well as the other models and the use of the ODT. However, the metrics of TNR and FPR are of least interest given that they address the classification accuracy of *NotBathymetry* points. Of greater interest are FNR and TPR (Figs. 8B and 8D) that reflect classification accuracy of *Bathymetry* points. FNR and TPR demonstrate that the use of the ODT considerably improves detection of Bathymetry, and that weighting may provide a further slight improvement.

Though promising, there are practical implications of these results. A positive outcome is that weighted models and the use of the ODT provided better results than unweighted models and a 0.50 PDT for *Bathymetry*-minority tiles, but performed similarly for the *Bathymetry*-majority tile "Deeper"/27285n. This means that it should be possible to operationalize results with little concern about the level or direction of *Bathymetry*/*NotBathymetry* imbalance.

Fig. 8. *NotBathymetry* ("Negatives") and *Bathymetry* ("Positives") accuracy rates for XGB models.



A potentially concerning outcome, however, is the higher FPR resulting from the use of the ODT for all tiles except the *Bathymetry*-majority "Deep" tile (27285n). Though the FPRs are indicative of classification accuracy for *NotBathymetry* pulse returns, high FPRs are associated with low User's Accuracy with both indicating the percentage of pulse returns incorrectly identified as *Bathymetry*. The accuracy of maps constructed from bathymetric classifications having a high FPR/low User's Accuracy may be low.

What is considered a "high" FPR or "low" User's Accuracy is dependent, of course, on the ultimate use of the data. The present work targets the improvement of an existing bathymetric classification method as distinct from the creation of a stand-alone system. This will be discussed further in the final section. For now, however, we suggest that the FPRs associated with the use of the ODT are sufficiently low for operationalization of this approach, particularly given the considerable improvements in the FNR and TPRs compared to the use of the unweighted model and the conventional PDT.

The final method examined to strengthen the bathymetric signal was confusion matrix decomposition; for this analysis, the ODT was applied at each iteration. Figure 9 shows the FPR and FNR (which are equal when the ODT is applied) and TPR and TNR (also equal when the ODT is applied) for the iterative removal of TPs and TNs; Figure 10 presents the same information for the iterative removal of FPs and FNs. Note that the repetitive removal of TPs and TNs (Fig. 9) removes so

[3] McFadden's pseudo R^2 value for classificatory models cannot be tested for statistical significance. A "reasonably strong" relationship is being inferred from the authors' experience with the conventional R^2 metric in consideration of the number of observations and the pseudo R^2 values.

many pulse returns s that it becomes impossible to fit a model beyond four iterations.

If confusion matrix decomposition is a successful strategy for increasing bathymetric signal, the FPR/FNR will decrease with each iteration and the TPR/TNR will increase. This did not occur for either decomposition considered. In particular, the repeated removal of correctly classified pulse returns (TNs and TPs) (Fig. 9) led to poorer accuracies for all tiles. Thus the elimination of the most certain pulse returns did not allow the bathymetric signal among the less certain pulse returns to be better manifested while also maintaining accuracy for the most certain.

Fig. 9. Accuracy rates for repetitive removal of true positives and true negatives. A. FPR & FNR. B. TPR & TNR.

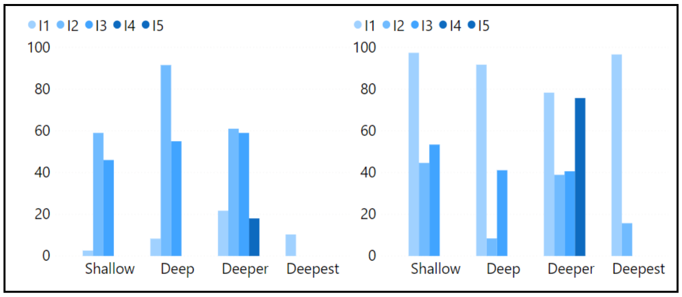
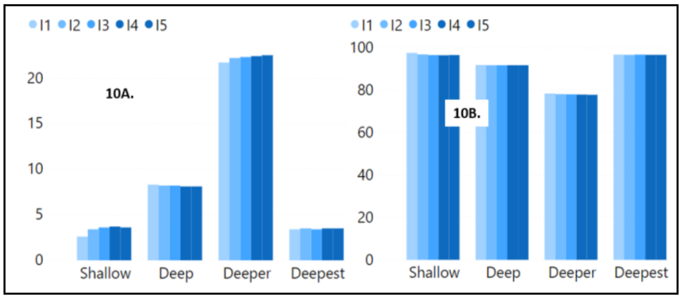


Fig. 10 Accuracy rates for repetitive removal of false positives and false negatives. A. FPR & FNR. B. TPR & TNR.



In contrast, the repeated removal of the incorrectly classified pulse returns – the FNs and FPs – has little impact on accuracy rates (Fig. 10). This is consistent with the findings associated with the removal of TPs and TNs. That is, unsurprisingly in retrospect, the bathymetric signal is most evident in the correctly classified pulse returns.

V. Conclusions and Operational Context

The scientific findings of this work are that:

- The bathymetric signal in the lidar pulse return metadata examined is sufficiently strong to warrant further attention.
- The ability to detect the bathymetric signal can be enhanced by mitigating a *Bathymetry*/*NotBathymetry* sample imbalance using an optimal decision threshold (ODT) that equalizes the TPR and the TNR.
- Proportional weighting during ML model fitting mitigates the impacts of an imbalanced sample to a lesser extent, but its effect can be combined with the use of an ODT.
- Confusion matrix decomposition was not found to be a viable strategy for mitigating the effects of an imbalanced sample.

From a practical perspective, we reemphasize that the goal of this work is to enhance the ability to detect the bathymetric signal in lidar point clouds rather than create a standalone method. The signal enhancement methods examined provide a way to better understand the strength of the bathymetric signal in pulse return metadata and also serve as a guide for how best to incorporate the results into an existing bathymetry extraction methodology.

In practice, an existing methodology such as CHRT [6] or RANSAC [5] would be used in its existing form to provide an initial *Bathymetry*/*NotBathymetry* classification of each lidar pulse return. A ML model would be fitted to this classification and used to estimate the probability of each return being bathymetry – i.e., *p(Bathy)*. This information would then be used in the disambiguation rules to produce a second classification. Assuming a relatively precise *p(Bathy)* estimate, it is anticipated that the second classification would be more accurate than the first. This process would continue until some convergence criterion was achieved. Though this criterion has not yet been defined, a number of alternative metrics are available – e.g., the number of lidar pulses changing from *Bathymetry* to *NotBathymetry* or vice versa at each iteration, or the magnitude of change in the TPR from one iteration to the next.

This work presents the initial step in a ML-based modification to existing density-based methods for processing bathymetric lidar data. That the bathymetric signal strength can be detected in the lidar pulse return metadata – and strengthened even for highly imbalanced samples – suggests that this approach is reasonably promising.

References

[1] J. Wozencraft, and D. Millar, “Airborne lidar and integrated technologies for coastal mapping and nautical charting,’ Marine Technology Society, vol. 39(3), pp. 27-35, 2005.

[2] American Society for Photogrammetry and Remote Sensing (ASPRS), “LAS Specification Version 1.3-R13,” 28 pp., 2013.

[3] D. Nagle, and C.W. Wright, “Algorithms used in the Airborn Lidar Processing System (ALPS),” United States Dept. of the Interior/ U.S. Geological Survey Open File Report 2016-1046, 45 pp., 2016.

[4] A. Nayegandhi, J. Brock, and C. Wright, “Small-footprint, waveform-resolving lidar estimation of submerged and sub-canopy topography in coastal environments,” International Journal of Remote Sensing, vol. 30(4), pp. 861-878, 2009.

[5] M. Fischler, and J. Bolles, “Random sample consensus – a paradigm for model fitting with applicatons to image analysis and automated cartograpy: Communication of the ACM,” vol. 24(6), pp. 381-395, 1981.

[6] B. Calder, and L. Mayer, “Automatic processing of high-rate, high-density multibeam echosounder data,” Geochemistry, Geophysics, Geosystems, vol. (4(6), 22pp., DOI: 10.1029/2002GC000486, 2003.

[7] B. Calder, and G. Rice, “ Computationally efficient variable resolution depth estimation,” Computers & Geosciences, vol. 106, pp. 49-49, DOI: dx.doi.org/10.1016/j.cageeo.2017.05.013.

[8] D. McFadden, “Conditional logit analysis of qualitative choice behavior,” In Frontiers in Econometrics, P. Zarembka Ed., Academic Press, pp. 105-142, 1974.

[9] T. Fawcett, “An Introduction to ROC Analysis,” Pattern Recognition Letters, vol. 27(8), pp. 861-874, DOI: 10.106/j.patrec.2005.10.010, 2006.

[10] R. Congalton, and K. Green, “Assessing the Accuracy of Remotely Sensed Data: Principles and Practices (3rd Edition),” CRC Press, 328 pp., 2019.

[11] H. He, Y., Bai, E. Garcia, and S. Li, “ADASYN: synthetic sampling approach for imbalanced learning,” IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence, pp. 1322-1328, 2008.

[12] N. Chawla, K. Bowyer, L. Hall, W.P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.