

METHODOLOGY

Open Access



# MAST (Movement Analysis Software for Telemetry data), for the semi-automated removal of false positives from radio telemetry data

K. Nebiolo<sup>1\*</sup> and T. Castro-Santos<sup>2</sup>

## Abstract

**Introduction** Radio telemetry, one of the most widely used techniques for tracking wildlife and fisheries populations, has a false-positive problem. Bias from false-positive detections can affect many important derived metrics, such as home range estimation, site occupation, survival, and migration timing. False-positive removal processes have relied upon simple filters and personal opinion. To overcome these shortcomings, we have developed MAST (Movement Analysis Software for Telemetry data) to assist with false-positive identification, removal, and data management for large-scale radio telemetry projects.

**Methods** MAST uses a naïve Bayes classifier to identify and remove false-positive detections from radio telemetry data. The semi-supervised classifier uses spurious detections from unknown tags and study tags as training data. We tested MAST on four scenarios: wide-band receiver with a single Yagi antenna, wide-band receiver that switched between two Yagi antennas, wide-band receiver with a single dipole antenna, and single-band receiver that switched between five frequencies. MAST has a built in a  $k$ -fold cross-validation and assesses model quality with sensitivity, specificity, positive and negative predictive value, false-positive rate, and precision-recall area under the curve. MAST also assesses concordance with a traditional consecutive detection filter using Cohen's  $\kappa$ .

**Results** Overall MAST performed equally well in all scenarios and was able to discriminate between known false-positive detections and valid study tag detections with low false-positive rates ( $< 0.001$ ) as determined through cross-validation, even as receivers switched between antennas and frequencies. MAST classified between 94 and 99% of study tag detections as valid.

**Conclusion** As part of a robust data management plan, MAST is able to discriminate between detections from study tags and known false positives. MAST works with multiple manufacturers and accounts for receivers that switch between antennas and frequencies. MAST provides the framework for transparent, objective, and repeatable telemetry projects for wildlife conservation surveys, and increases the efficiency of data processing.

## Introduction

Wildlife telemetry is the practice of monitoring movements of animals using systems of transmitters attached to individuals, and receivers that may be land-based or mobile. Radio telemetry is the most versatile and widely used tracking method in a variety of ecosystems. Its functionality in freshwater, including the flexibility of

\*Correspondence:

K. Nebiolo

[kevin.nebiolo@kleinschmidtgroup.com](mailto:kevin.nebiolo@kleinschmidtgroup.com)

<sup>1</sup> Kleinschmidt Associates, 35 Pratt St. Suite 201, Essex, CT 06426, USA

<sup>2</sup> U.S. Geological Survey, Leetown Science Center, S.O. Conte Anadromous Fish Research Center, One Migratory Way, Turners Falls, MA 01376, USA



© The Author(s) 2022, corrected publication 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

transmission rates and detection ranges, is the reason it is the most common tool for understanding movements of migratory fish as they approach and pass barriers like hydroelectric dams, where millions of individual fish have been tagged and tracked worldwide [2].

Unfortunately, telemetry technologies are prone to both false positives (where a random noise or other factors produce a signal that is logged as a viable code) and false negatives (where a transmission fails to be detected, even though the tag is within range of a receiver). These receiver errors can bias estimates of occupancy, movement, and survival [29]. Statistical tools developed for mark-recapture can be used to control for the bias induced by missed detections (false negatives): [9, 19, 35]. No comparable methods exist to control for false positives, however [6]. This is problematic because false positives overestimate the frequency of occurrence for an event of interest and may assign an animal to a location, habitat, or state that they do not actually occupy. McClinck et al. [25] found severe overestimation of site occupancy with as little as 1% false-positive rate. For studies that assess migratory delay, false positives may bias towards longer presences and greater delay. Mark-recapture approaches have been successfully applied to study migratory barriers [32] and the methodology inherently supports sites with imperfect detection rates at the cost of precision. However, mark recapture assumes the truth value of each detection is known with 100% accuracy. Mark recapture has no means to reduce bias from false positives; they must be identified and removed a priori. Therefore, a confirmation strategy as suggested by Chambert et al. [8] is required to assess the validity of every observation.

Methods currently in use for false-positive removal include eliminating records from transmitter codes not used in a study or occurring prior to time of release, imposing a minimum received signal strength, setting a minimum frequency of detections per unit of time at a given site, and examining the spatio-temporal distribution of detections for logical errors in site progression [37]. Beeman and Perry [6] and Simpfendorfer et al. [36] added to these steps by requiring sequential detections to be in series, e.g., if the nominal pulse rate on a tag was 3 s, records within a contiguous set of detections should be 3 s apart. Both Brownscombe et al. [7] and Heupel et al. [17] required multiple detections of an individual within a given time window, but did not require detections to be in series. When implemented in tandem, these filtering methods will remove a considerable amount of false-positive detections. However, these methods may also lead to overcorrection, which create instances of false negatives. This overcorrection is problematic, particularly for studies with low sample sizes, where insufficient data can

preclude accurate estimation. Therefore, most studies must rely heavily on manual classification, which is subjective, non-reproducible, labor intensive, and takes away critical time from analysis and interpretation of data.

The field of machine learning explores the study and construction of algorithms that can learn from a training set and can then make predictions. Here, we present a false-positive identification algorithm based on the classic Naïve Bayes (NB) classifier ([28], please see Appendix Table 9 for a complete list of abbreviations and their definitions) that provides an objective score of the likelihood that a given observation is valid, and a transparent framework and justification for the final assignment. Considerable attention has been paid to binary classification problems, with NB used to identify ‘fake news’ [14], spam [3, 20], and medical diagnoses [40]. NB approaches are widely used in ecology as well, with Fernandes et al. [13] using NB to predict fish recruitment, Feki-Sahnoun et al. [12] using NB to predict likelihood of toxic algal blooms, and Lehikoinen et al. [22] using NB to assess the effects of environmental factors on ecological indicators. NB is simple, and with ample minimally biased training data, very robust. Radio telemetry false-positive screening is yet another appropriate use for the algorithm because of the large amount of training data created by the system.

We present a module of Movement Analysis Software for Telemetry data (MAST) for the semi-automated removal of false positive detections from radio and other forms of telemetry data [31]. MAST trains a semi-supervised NB algorithm to identify false-positive detections in radio telemetry data with beacon and/or study tags. We will examine the effectiveness of the approach with a case study on the Connecticut River, USA. The large-scale study tracked 562 American Shad (*Alosa sapidissima*) and 80 Sea Lamprey (*Petromyzon marinus*) through a complex network that employed wide-and single-band receivers with Yagi and dipole antenna configurations. With over 600 active tags in the system, multiple frequencies were required, meaning single-band receivers switched between frequencies. We will demonstrate the algorithm’s ability to perform well with multiple technologies with various antenna configurations.

## Methods

The identification and removal of false-positive detections from radio telemetry with MAST starts with the quantification and description of predictor variables. Then, MAST fits (or trains) a NB model, which calculates probability that a detection is valid or false positive given a set of observations. With probabilities in hand, MAST applies a decision criterion to remove false positives from record. To assess the algorithm’s ability to discern between valid and false-positive detections, MAST

performs a k-fold cross-validation and assesses model quality with the area under the curve (AUC) statistic as well as measures of sensitivity (*sen*), specificity (*spc*), negative and positive predictive value (*npv*, *ppv*), and false-positive rate (*fpr*).

### Selecting and quantifying predictor variables

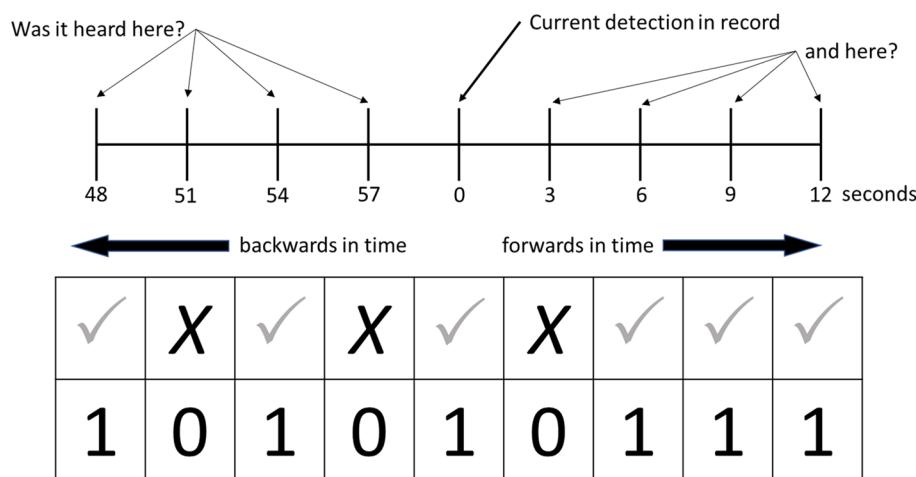
In developing the classifier, it was important to select predictor variables that maximized the ability to discriminate between valid and false-positive detections. These included: power or received signal strength (RSS), hit ratio (HR), consecutive record length (CRL), noise ratio (NR), and the difference in the time-lag between detections ( $\delta^2 L$ ). Power refers to the received signal strength of a given transmission. Depending on the receiver model used, this may be reported in arbitrary units or dB. Of the predictor variables, power is the only one that is intrinsic to a given transmission; all other predictors were derived from detections recorded within a short period of time surrounding of a given transmission.

The Proximate Detection History (PDH or detection history) refers to a series of detections of a given tag recorded during a fixed number of pulse intervals immediately preceding and following a given detection. This describes the pattern of recorded to missed detections in series from the current record. The algorithm looks forwards and backwards in time a specified number of transmission intervals. For example, say a given tag detection occurs at midnight (00:00:00), has a 3-s pulse rate and produces the pattern of heard to missed detections in (Fig. 1). The middle integer is the initial detection (00:00:00), but it was not detected 3 s prior (11:59:57) or 3 s post (00:00:03). To create the PDH, the algorithm queries the recaptures database a set number of pulse

intervals forward and backward in time from the current record. If the tag was detected in series, a '1' was added to the history, if it was not then a '0' was added. In Fig. 1, the fish was heard on the -12th, -6th, 0 (current record), 6th, 9th and 12th s, and the resulting detection history was '101010111'.

Radio telemetry receivers typically record detection times rounded to the nearest second. Sometimes, however, a given transmission might not fall within the expected second. This can arise because tags can be programmed with intervals that are not discrete integer values, or because a pulse randomizer is employed. The pulse randomizer slightly adjusts the signal burst rates by a small amount (depending on manufacturer this is typically  $\sim \pm 500$  ms), which reduces the probability of two signals colliding resulting in false negatives. Since most commercially available radio telemetry receivers record data to the nearest second, the algorithm must query the recaptures database within a 3-s moving window (the expected time, then that time plus and minus 1 s to allow for rounding). Any detection logged within this broader window is considered valid. For example, if a tag was detected at midnight (00:00:00) and has a 3-s pulse rate, MAST queries all recaptures from 11:57:56 to 11:57:58 and from 00:00:02 to 00:00:04. In the case of Fig. 1, the tag was not detected within the first interval (forwards or backwards); meaning the original detection, which occurred at 00:00:00 would not be valid from a consecutive detection in series perspective. Any detection occurring outside of the interval  $+/- \epsilon$  is not included in the PDH.

Having defined the detection history and its associated time window, we now can calculate HR, which is the number of detections within a PDH divided by the length



**Fig. 1** Creation of a detection history around the current detection (0 s)

of the detection history. For a  $\pm 4$  detection hit ratio like the one pictured in Fig. 1, the length of the history is 9, and the HR is 6/9.

The second derived predictor is the consecutive record length (CRL). This refers to the longest contiguous subset of 1's in a given detection history. In Fig. 1, the CRL is 3. Table 1 contains examples of possible detection histories and their respective HRs and CRLs. Note that these PDHs have  $\pm 4$  intervals, and that the first row in the table corresponds to the detection history pictured in Fig. 1. The middle position in a PDH is the current detection and is always a 1. Also note the second and third rows, the histories have the same HR, but different CRL (Table 1).

The next predictor is noise ratio (NR), which is simply the number of plausible study tag hits, divided by the total number of detections (i.e., including known noise detections but excluding beacon tags) within a 1-min interval around the current detection. Detections within the window were categorized into two classes, plausible and known false positive. Plausible detections are from those codes and frequencies currently active within the study area. The remaining detections are from unknown or unavailable codes, in other words, they are known false-positive detections ( $f$ ). The noise ratio (NR) is given with  $R = f/n$ , where  $f$  is the number of false-positive detections within a 1-min window around the current detection, and  $n$  is the total number of detections within a 1-min window around the current detection.

The last derived predictor calculates the second-order difference in time-lag between detections or  $\delta^2L$ . It is simply the difference of the difference in timestamps between sequential rows. When a tagged animal is within detection range, the tags will pulse and be recorded at the nominal rate set at the onset of the study. For example, if the nominal pulse rate of the tag is 3 s, one would expect to hear that tag every 3 s. For a valid detection, the time-lag between successive detections should be 3 s (first order), and the difference in time-lag between subsequent detections should be zero (second order). We expect true-positive detections to have a  $\delta^2L$  of zero, or fixed multiples of the pulse interval. The more consistent

$\delta^2 L$  is with expectations, the more belief we have in the record being true.

## Treatment of continuous variables

All continuous classifier variables were discretized into bins. Discretization of continuous features has a number of advantages. Discretization roughly approximates probability distributions and helps to overcome inaccurate shape assumptions [11, 41]. Hsu et al. [18] tested a number of discretization methods for NB classifiers, but found no performance improvements with algorithm complexity. Therefore, MAST uses a simple equal width interval discretization process.

Detection power was binned into equal width intervals of 5 dB or 5 arbitrary units depending upon manufacturer, NR was binned into 10 percentile units, and lag differences were binned into equal width intervals as wide as the tag's nominal pulse rate. HR and CRL were limited by the number of detections within the PDH, and thus limited to a set number of classes.

Discretization has one major limitation. If there were no observations for a particular bin, then the probability of it occurring is zero. This limitation negates the weight of evidence provided by the other predictor variables and is uninformative. To overcome this, we applied Laplace smoothing [21], which added a single observation to each bin and eliminated zero counts. This slight positive bias has almost no effect on training datasets with a large number of observations.

## Training methods

When MAST trains a model, it reads raw telemetry data and separates it into rows with known valid detections, and rows of known false positives. Known false-positive detections are from tags not on the study tag list, while known valid detections come from beacon tags. Detections with known validity are the training set, while detections with unknown validity (i.e., study tags) are the classification set. MAST can construct a training dataset two ways: by training on beacons (supervised) or by training on study tags themselves (semi-supervised).

**Table 1** Detection history examples with corresponding hit ratio and consecutive record length

### Training on beacons

It is common in telemetry studies for researchers to employ tags that are not on fish. These may be ‘beacons’, which are typically set to transmit at fixed intervals to provide a record of continued functioning of the receiver system; or ‘test tags’ which are typically drawn through the intended detection field in such a way as to emulate the movements of free-swimming fish. Either can be used to provide training data, but the greater verisimilitude provided by test tags makes them the better choice, provided sufficient volume of data are generated to create a suitable training dataset.

### Training on study tags

There are limitations to using beacons as training data. The beacon may not be representative of a study tag. Transmission intervals might be too long, some tags cannot be cycled on and off in realistic ways, producing unrealistically long strings of valid detections, the tags themselves are typically in fixed locations, and it is possible for there to be false positives among the beacons (after all, it is the removal of data that look valid that drives this effort). When training on study tags, MAST constructs a training dataset that assumes all study tag records are true. This poses a dilemma as we anticipate there being false positives mis-labeled as valid. Therefore, it is advised to re-classify the data by training on the previous iteration’s valid detections and known false positives from the initial training. There exists a tradeoff: with the beacons we had a limited number of tags (usually just one per receiver) so the likelihood of a false positive with that exact code is small. Because of that we were able to make the simplifying assumption that all beacon data were valid. We cannot make the same assumption with study tags because the purpose of this effort is to remove false positives.

The solution to this is to use an iterative approach, where on the first iteration, we train on beacon tags (or study tags themselves) and classify study tags. On subsequent iterations, we train on the previous iterations’ valid detections and known false-positive detections from the first iteration. This alters the density functions of the predictor variables, with fewer known, but higher quality valid detections. A new iteration uses these new frequencies to re-classify the remaining study tag detections. This process continues until convergence when no new observations are classified as false positive.

### False-positive classification

Supervised learning algorithms use observed data with known classifications (training data) to classify unknown data. Bayes theorem takes training data and quantifies

the probability that a record is either true or false positive given what we have observed about it [5]. This probability, known as the posterior, is given with

$$P(C_i|F_1, \dots, F_n) \propto P(C_i) \prod_{j=1}^n P(F_j|C_i), \quad (1)$$

where  $P(C_i|F_1, \dots, F_n)$  is the posterior probability of a valid (or false positive) detection given the values of each observed predictor ( $F_1, \dots, F_n$ );  $P(C_i)$  is the prior probability of the  $i$ th detection class occurring ( $C \in \{\text{Valid}, \text{FalsePositive}\}$ ), and  $P(F_j|C_i)$  is the likelihood (conditional probability) of the  $j$ th observed predictor ( $F_j$ ) value given the  $i$ th detection class ( $C_i$ ). Naïve Bayes assumes that all predictor variables are independent, and hence, the likelihood of the observed predictor values given a detection class is a product. The posterior probability expresses our belief in the record being true or false positive given what we have observed.

The prior  $P(C_i)$  is the marginal probability of the  $i$ th detection classification occurring in the training dataset, where ( $C \in \{\text{Valid}, \text{FalsePositive}\}$ ). MAST calculates the prior probability a record is valid  $P(T)$  with a simple frequency analysis;  $P(T) = n_T/n$  where  $n_T$  is the number of valid records in the training dataset divided by the total number ( $n$ ) of records in the training dataset. Since the priors are marginal, the prior probability that a record is false positive is given with  $1 - P(T)$ .

The likelihood  $P(F_j|C_i)$  is the conditional probability of the  $j$ th observed predictor value  $F_j$  given the  $i$ th detection class ( $C_i$ ). MAST calculates the likelihood using a frequency table:  $P(F_j|C_i) = n_F/n_C$ , where  $n_F$  is the number of records within detection class  $C_i$  that match the observed predictor value  $F_j$ , and  $n_C$  is the number of records with the detection class  $C_i$ .

To classify, MAST applies the *maximum a posteriori (MAP)* hypothesis, and chooses the detection class that is most true. The algorithm’s decision rule becomes

$$\operatorname{argmax}_{C_i} \left\{ P(C_i) \prod_{j=1}^n P(F_j|C_i) \right\}. \quad (2)$$

Under this hypothesis, the detection class with the larger posterior probability is chosen. Under the *MAP* hypothesis, any detection with a valid to false-positive ratio ( $\frac{P(T)}{1-P(T)}$ ) less than 1.0 as false positive.

Tables 2 and 3 follows the classification of two records from initial observation and description to the calculation of prior, likelihood, and posterior probabilities, and then the application of the *MAP* criterion. Table 2 contains two records from two different study tags. The first detection was recorded on May 5th, and the second on July 4th. The first detection had a moderately full PDH

**Table 2** Observation of two study tag records at Receiver T21. Abbreviations and variable names and units are as described in Appendix Table 9

Time stamp	$\delta^2 L$	PDH	HR	CRL	RSS	NR
2015-05-18 13:55:27	-10	00010111100	0.45	4	-101	0
2015-07-04 16:12:03	33,339	00000100000	0.09	1	-107	0.8

**Table 3** The prior, likelihood and posterior calculations for detections identified in Table 2. Abbreviations are defined in Appendix Table 9

Model component	Term	Valid		Term	False positive	
		2015-05-18 13:55	2015-07-04 16:12		2015-05-18 13:55	2015-07-04 16:12
Prior	$P(T)$	0.996	0.996	$P(F)$	0.004	0.004
Likelihood	$P(HR T)$	0.05	0.259	$P(HR F)$	0.051	0.083
	$P(CRL T)$	0.052	0.466	$P(CRL F)$	0.008	0.34
	$P(RSS T)$	0.339	0.193	$P(RSS F)$	0.02	0.036
	$P(NR T)$	0.427	<0.001	$P(NR F)$	0.045	0.018
	$P(\delta^2 L T)$	0.011	2.841	$P(\delta^2 L F)$	0.002	0.001
Posterior	$P(T HR, CRL, RSS, NR, \delta^2 L)$	4.13E-06	7.49E-12	$P(F HR, CRL, RSS, NR, \delta^2 L)$	3.02E-12	5.23E-11

with an HR of 0.45 and CRL of 4. The second record had a very sparse PDH, with an HR of 0.09. NR was also high for the second detection (Table 2).

Table 3 contains the prior, likelihood and posterior of the two detections identified in Table 3. Note, the posterior is simply the product of all rows above it. The prior probability that a detection is false positive  $P(F)$  at this receiver is only 0.004 (Table 3), meaning there is overwhelming evidence that a detection will be valid. The next 5 rows identify the likelihood of each observation occurring given the detection classification. The MAP hypothesis chooses the detection class with the larger posterior, therefore, the detection occurring on May 18 was valid, while the detection occurring on July 4th was false positive.

### Cross-validation

MAST assesses the ability of the algorithm to discern between classified valid study tags and known false-positive detections with a  $k$ -fold cross-validation technique [39]. For studies that train-on-study tags, the training dataset includes those records classified as valid from the final iteration and known false-positive detections. The procedure partitions the training dataset into  $k$  equal sized subsamples. In each iteration, a single subsample (or fold) was retained as the test dataset (to be classified) and the remaining  $k - 1$  subsamples are retained as the training dataset. The cross-validation process is then repeated  $k$  times over each fold, with each of the  $k$  subsamples used exactly once as validation data. This

**Table 4**  $2 \times 2$  contingency table summarizing results of  $k$ -fold cross-validation. Variables are defined in Appendix Table 9

	True	False positive
Classified valid	$t_p$	$f_p$
Classified false	$f_n$	$t_n$

procedure allows a 1:1 comparison of known classifications against the algorithm's classifications. A classification can have 1 of 4 states, true positive ( $t_p$ ), true negative ( $t_n$ ), false positive ( $f_p$ ), and false negative ( $f_n$ ). Results of the  $k$ -fold cross-validation are summarized into a  $2 \times 2$  contingency table (Table 4).

Accuracy metrics derived from the  $2 \times 2$  cross-validation contingency table include sensitivity and specificity. Sensitivity, or the true-positive rate, is given with:  $sen = t_p / (t_p + f_n)$ , and measures the probability that all valid detections were correctly classified as valid. Specificity, or true negative rate, is given with:  $spc = t_n / (f_p + t_n)$ , and quantifies probability that all false-positive detections were correctly classified.

Precision metrics include the positive and negative predictive value or  $ppv$  and  $npv$ . The positive predictive value ( $ppv$ ) is the proportion of detections classified as valid that were valid;  $ppv = t_p / (t_p + f_p)$ . The negative predictive value ( $npv$ ), which measures the proportion of detections classified as false that were false, is given with:  $npv = t_n / (f_n + t_n)$ . Again, our objective is to maximize both measures. The higher the  $ppv$  the lower the number

of potential false-positive detections in the dataset. Likewise, a high  $npv$  means a lower number of false negatives.

Since we are identifying and removing false-positive detections, and false-positive detections are rare, the most important algorithm metric is the false-positive rate ( $fpr$ ), which calculates the proportion of detections classified as valid that are in fact false positive with:  $fpr = f_p / (f_p + t_n)$ , or  $1 - spc$ . Our objective is to minimize  $fpr$ , the lower the rate, the fewer known false-positive detections were classified as valid.

MAST also produces the precision-recall curve (PRC) and calculates the area under the curve (AUC) statistic. Precision quantifies the number of correct false-positive predictions while recall quantifies the number of correct false-positive predictions made out of all false-positive predictions that could have been made. The AUC statistic summarizes area under the PRC curve for a range of threshold values. The PRC is a better performer on imbalanced datasets [34], which are typical of radio telemetry studies.

The results of the  $k$ -fold cross-validation can inform on the selection of predictor variables. Ling et al. [24] found AUC to be statistically consistent and more discriminating than accuracy alone. Rosset [33] used AUC as an evaluation criterion for scoring classification models where models with higher AUC are preferred. With MAST, it is possible to construct suite of classifiers that use different combinations of predictor variables; the model that maximizes measures of AUC,  $sen$ ,  $spc$ ,  $ppv$ , and  $npv$  while minimizing  $fpr$  is the best.

### Case study

MAST was implemented on a large-scale telemetry project on the Connecticut River in 2015 that tracked 560 American Shad and 80 Sea Lamprey with 30 continuous radio telemetry monitoring stations. The subset of receivers highlighted in this paper (Fig. 2) created 4 scenarios, which included multiple receiver manufacturers Sigma Eight Orion and Lotek SRx800 receivers, dipole and Yagi antenna configurations, receivers that scanned multiple frequencies, and receivers that switched between antennas.

The receivers in scenario 1 were Orion units manufactured by Sigma Eight and consisted of detection zones T13, T15, T18, T21 and T22 (Fig. 2). These units spanned the full width and depth of the river and from a noise perspective were similar. Scenario 2 consisted of the detection zones T12E and T12W (Fig. 2), which was a single Orion receiver switching between two Yagi antennas. Scenario 3 consisted of detection zones T09, T07 and T30. The Orion receivers had a single dipole antenna and were typically deployed in areas where specimens were known to congregate. Scenario 4 included detection

zones T03, T06, and T24. These were Lotek SRx800 receivers with a single Yagi switching between 5 frequencies. MAST accounts for number of frequencies (or antennas) and the scan time devoted to each while building the PDH and deriving CRL and HR statistics.

After training and classifying receivers within each scenario, we performed a  $k$ -fold cross-validation that assessed the ability of MAST to correctly identify and remove known false-positive detections from record with measures  $sen$ ,  $spc$ ,  $ppv$ ,  $npv$ ,  $fpr$ , and precision-recall AUC (PRC-AUC). Aside from assessing the quality of the model, these metrics also assist in model selection, as we will demonstrate.

As a last measure, we compared the saturated model in MAST with the filtering method proposed by Beeman and Perry [6], which stated for a detection to be classified as valid, it must be within a consecutive series of detections. We assessed concordance between methods with Cohen's Kappa ( $\kappa$ ) [26], which takes into account the possibility of agreement occurring by chance. A value of  $\kappa = 1$  suggests perfect agreement between MAST and the consecutive detection requirement.

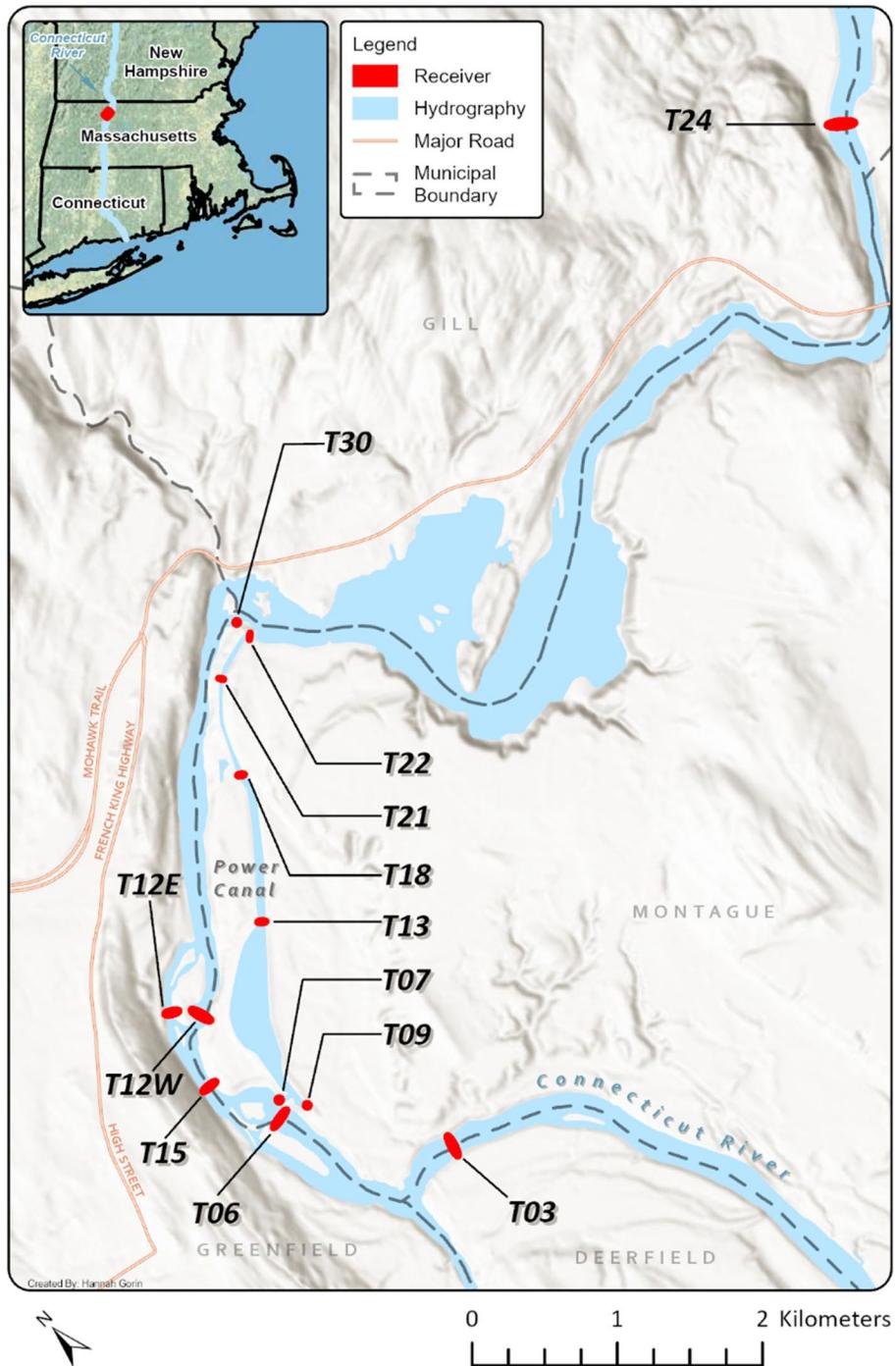
## Results

Results are organized by case study scenario. For each scenario, a panel figure compares the distributions of each predictor variable across detection classes. Then, we determine concordance between MAST and the consecutive detection filter with Cohen's  $\kappa$ . Then, we present results of the  $k$ -fold cross-validation procedure with a comparison of models using  $sen$ ,  $spc$ ,  $ppv$ ,  $npv$ ,  $fpr$ , and precision-recall AUC (PRC-AUC).

### Scenario 1: Sigma Eight Orion receivers with single Yagi antenna

Scenario 1 included five Sigma Eight Orion receivers placed at optimal sites with a single Yagi antenna capable of sampling the full width and depth of the water body. MAST classified 94.4% of detections received by this group of receivers as valid. Cohen's  $\kappa$  was 0.24, suggesting a rather low agreement between MAST and the consecutive detection requirement. Figure 3 is a panel of histograms (A–F) showing the distributions of each predictor by detection class. Detections classified as valid had much higher  $HRs$  (Fig. 3A) and longer  $CRLs$  (Fig. 3B), while false-positive detections had higher  $\delta^2L$  (Fig. 3E). There does not appear to be much difference between valid and false-positive detections with respect to RSS (Fig. 3C) or NR (Fig. 3D).

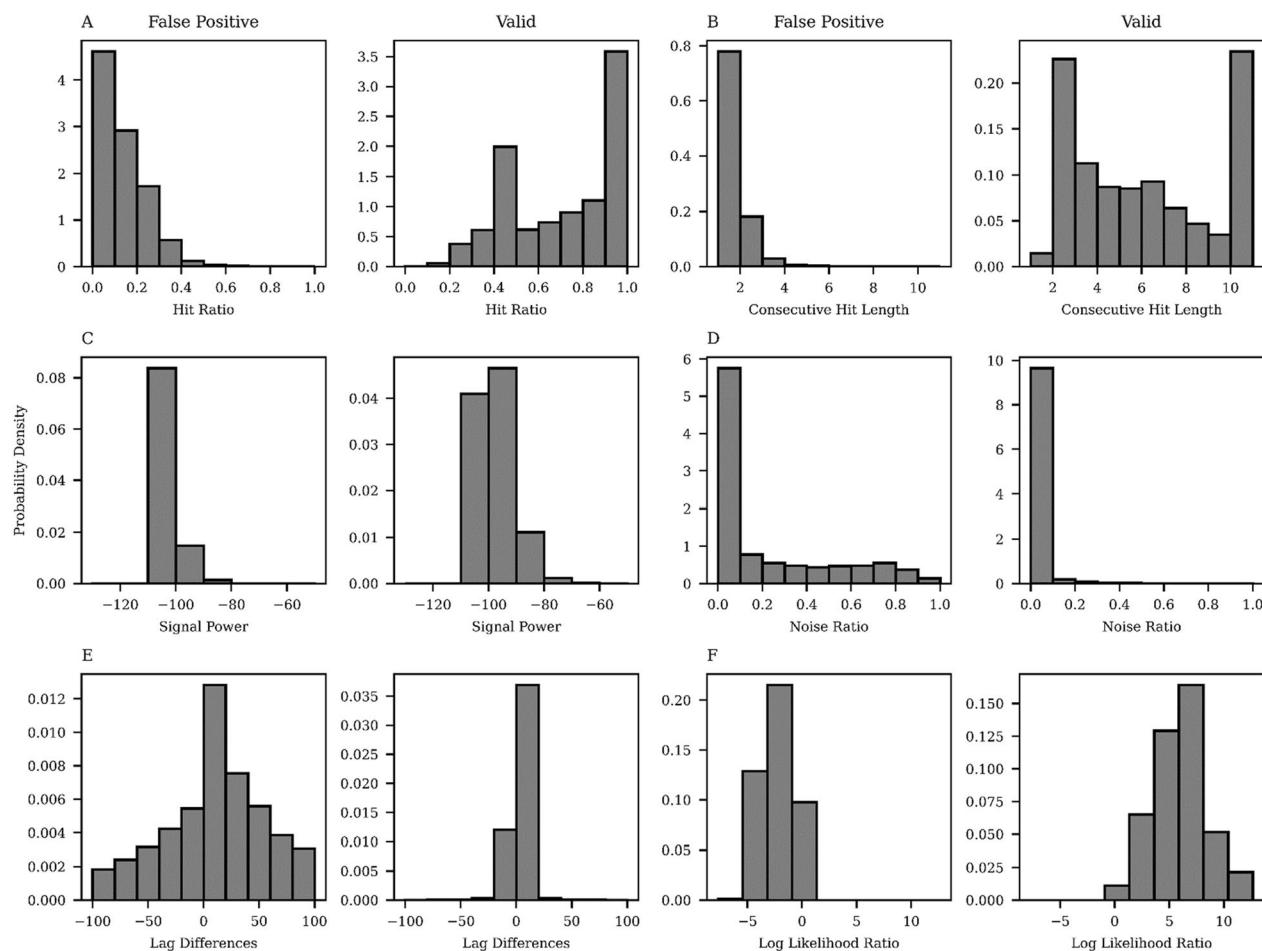
The cross-validation procedure produced favorable metrics for the saturated model with a PRC-AUC of 0.850 and  $fpr$  of 0.001 (Table 5). HR and CRL were highly correlated ( $R^2=0.9048$ ). When we removed CRL as a



**Fig. 2** Site location map with receivers from the 2015 study

predictor, the reduced model produced only slightly better *fpr* (<0.001 vs 0.001) (Table 5), which suggests there was no penalty when including correlated variables. A third model was created that removed NR and *P*, which appeared to be the least predictive variables (Fig. 3). The reduced model (*HR* \* *L*) appears to perform the best by

not producing any false-positive detections, thus minimizing *fpr*. A model was constructed that included *RSS* as the only predictor. In Fig. 3, it appears as though valid detections have higher power, meaning the likelihood of a high-powered detection given that it was true is greater than the likelihood of a high-powered detection given



**Fig. 3** MAST predictor variable probability mass functions for detections at Sigma Eight Orion receivers (T13, T15, T18, T21, T22; Fig. 2).

**Table 5** Cross-validation results for Scenario 1

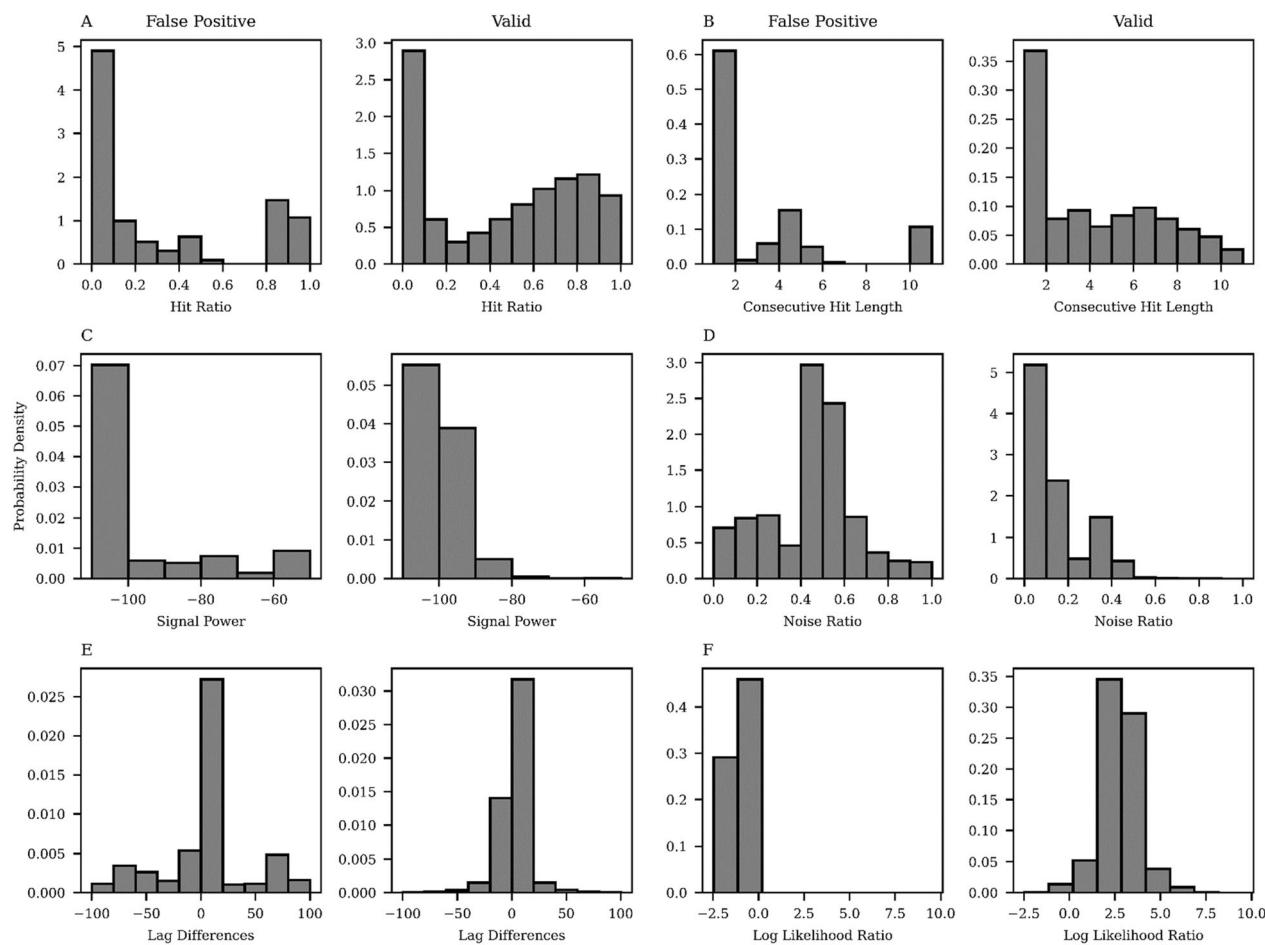
Model	$t_p$	$t_n$	$f_p$	$f_n$	$se$	$sp$	$ppv$	$npv$	$fpr$	PRC-AUC
CRL * HR * NR * RSS * $\delta^2 L$	2,062,067	101,292	63	17	1.0	0.999	1.0	0.999	0.001	0.850
HR * NR * RSS * $\delta^2 L$	2,062,079	101,337	18	5	1.0	0.999	1.0	1.0	<0.001	0.850
HR * $\delta^2 L$	2,062,084	101,351	4	0	1.0	1.0	1.0	1.0	0.0	0.850
RSS	2,062,084	67	101,288	0	1.0	0.001	0.953	1.0	0.999	0.999

that it was false positive. This resulted in most of the known false-positive detections classified as valid, which increased the  $fpr$  to 0.999 (Table 5).

#### Scenario 2: single Sigma Eight Orion receiver switching between antennas

Scenario 2 includes information from a single receiver that switched between two antennas to cover two different channels in the river. This meant there was a portion of time a river channel was not observed. MAST

accounts for the time spent observing the second channel when constructing the PDH. Therefore, detections classified as valid still have high  $HR$  (Fig. 4A) and  $CRL$  (Fig. 4B). Detections classified as false positive tended to have larger  $\delta^2 L$  (Fig. 4E) and higher NR (Fig. 4D). Overall, MAST identified <1% of records as false positive. Concordance between methods was very low ( $\kappa=0.0010$ ) as evident by 134,107 records classified as valid by MAST but false positive by the consecutive detection requirement.



**Fig. 4** MAST predictor variable probability mass functions for detections at a single Sigma Eight Orion receiver that switched between antennas (T12E, T12W)

**Table 6** Cross-validation results for Scenario 2

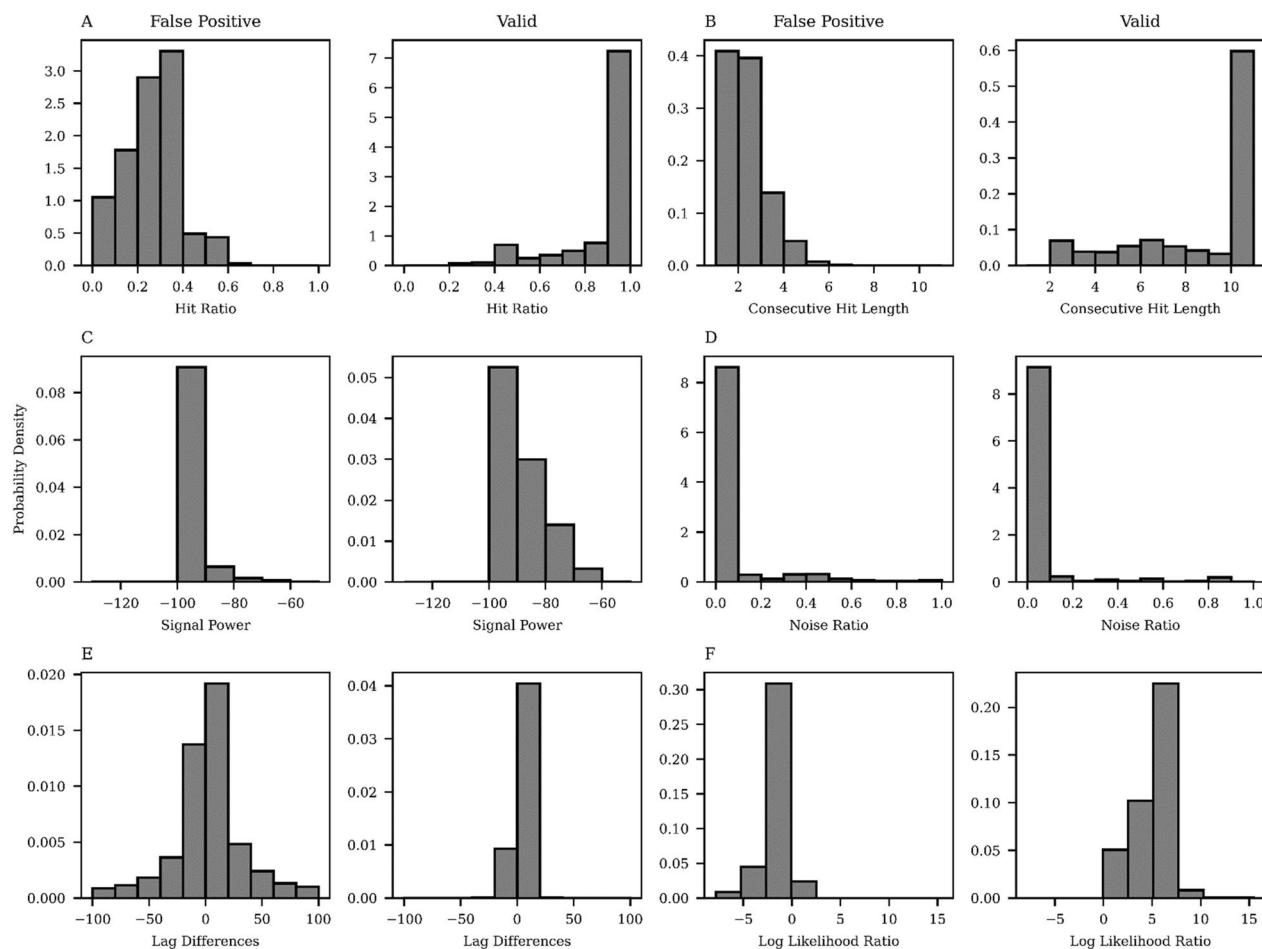
Model	$t_p$	$t_n$	$f_p$	$f_n$	$se$	$sp$	$ppv$	$npv$	$fpr$	PRC-AUC
CRL * HR * NR * RSS * $\delta^2 L$	319,526	1589	1	3	1.0	0.999	1.0	0.998	0.001	0.974
HR * NR * RSS * $\delta^2 L$	319,529	1589	1	0	1.0	0.999	1.0	1.0	0.001	0.974
HR * NR * RSS	319,529	1572	18	0	1.0	0.988	0.999	1.0	0.011	0.974
RSS	319,529	1260	330	0	1.0	0.793	0.999	1.0	0.208	0.979

The cross-validation procedure produced excellent results again with the  $fpr$  for all but the RSS-only model at or far below < 1% (Table 6).  $HR$  and  $CRL$  were highly correlated ( $R^2=0.9137$ ); however, there was no difference in  $fpr$  between the saturated and reduced model (Table 6). Removing  $\delta^2 L$  from the model increased the  $fpr$  to 0.11, this demonstrates that  $\delta^2 L$  still provides some discriminatory power. Figure 4C shows little difference in  $RSS$  between detections classified as valid and those classified as false positive. A model with just

$RSS$  still had high sensitivity (1.0), but  $fpr$  increased to 21%.

#### Scenario 3: Sigma Eight Orion receivers with single dipole antenna

Scenario 3 includes information from Orion receivers T07, T09, and T30, which had a single dipole antenna each. These antennas were placed adjacent to areas of congregation, such as fish passage infrastructure, and classified 97% of the study tag detections as valid. Unlike



**Fig. 5** MAST predictor variable probability mass functions for detections at a single Sigma Eight Orion receiver that switched between antennas (T07, T09, T30)

other receivers, there were detections classified as false positive that had moderately high HR values (Fig. 5A). These dipole antennas sampled almost no noise, as is evident with little difference in NR between false positive and valid detections (Fig. 5D). We found little agreement ( $\kappa = 0.11$ ) between MAST and the consecutive detection requirement; MAST classified 22,041 records as false positive while the consecutive detection filter did not.

The saturated model had an *fpr* of less than 0.01 (Table 7). HR and CRL were highly correlated ( $R^2=0.9440$ ) and removing CRL from the model

lowered *fpr* even further. Removal of NR from the model improved *fpr* even more with no false positives recorded (*fpr* = 0.0). We tested a single predictor (RSS) model (Table 7) and found an *fpr* of 23%.

#### Scenario 4: Lotek SRX 800 with single Yagi antenna

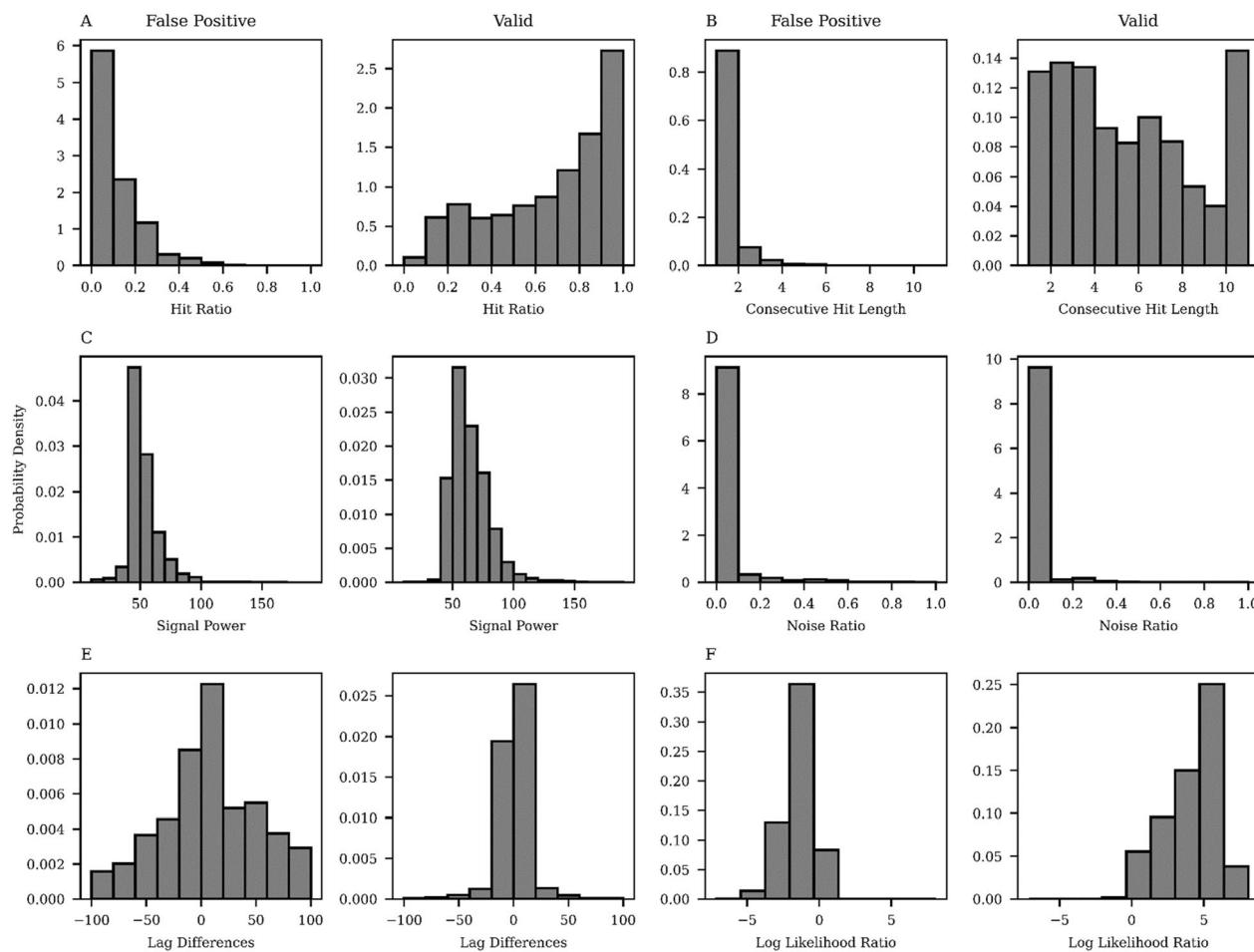
Scenario 4 includes information from three Lotek SRx800 (T03, T06, and T24) receivers scanning five frequencies. MAST parses Lotek SRx800 receiver raw data, extracts the number of channels and their scan time from the header information, then incorporates these data into

**Table 7** Cross-validation results for scenario 3

Model	<i>t<sub>p</sub></i>	<i>t<sub>n</sub></i>	<i>f<sub>p</sub></i>	<i>f<sub>n</sub></i>	<i>se</i>	<i>sp</i>	<i>ppv</i>	<i>npv</i>	<i>fpr</i>	PRC-AUC
CRL * HR * NR * RSS * $\delta^2 L$	2,644,723	12,597	33	25	1.0	0.997	1.0	0.998	0.003	0.975
HR * NR * RSS * $\delta^2 L$	2,644,744	12,622	8	4	1.0	0.999	1.0	0.999	0.001	0.975
HR * RSS * $\delta^2 L$	2,644,748	12,630	0	0	1.0	1.0	1.0	1.0	0.0	0.975
RSS	2,644,747	9740	2890	1	1.0	0.771	0.999	0.999	0.229	0.981

the logic behind the PDH creation. Like the Orion receivers, valid Lotek SRx800 detections generally had high HR (Fig. 6A) and CRL (Fig. 6B). Detections classified as false positive generally had lower RSS (Fig. 6C). MAST classified 95% of study tag detections as valid. When compared with the consecutive detection filter, Cohen's  $\kappa$  was 0.20 suggesting low concordance between methods as is evident with 191,282 records classified as false positive by the consecutive detection requirement but valid by MAST.

The saturated model performed very well with a low  $fpr$  of 0.001 (Table 8); however, HR was highly correlated with CRL ( $R^2=0.8810$ ). A reduced model demonstrated no change in  $fpr$  (Table 8), which suggests there is no penalty for including correlated predictors. In Fig. 6, there does not appear to be much noise present within these three receivers (Fig. 6D), nor is there much difference in RSS (Fig. 6C). We constructed single predictor models for NR and RSS to demonstrate the effect of variables with low discriminatory power. The  $fpr$  for a model



**Fig. 6** MAST predictor variable probability mass functions for detections at a single Sigma Eight Orion receiver that switched between antennas (T03, T06, T24)

**Table 8** Cross-validation results for scenario 4

Model	$t_p$	$t_n$	$f_p$	$f_n$	$se$	$sp$	$ppv$	$npv$	$fpr$	PRC-AUC
CRL * HR * NR * RSS * $\delta^2 L$	765,976	3858	3	3	1.0	0.999	1.0	0.999	0.001	0.973
HR * NR * RSS * $\delta^2 L$	765,977	3858	3	2	1.0	0.999	1.0	0.999	0.001	0.973
NR	765,889	868	2993	90	0.999	0.225	0.996	0.906	0.7752	0.993
RSS	765,979	0	3861	0	1.0	0.0	0.995	0.0	1.0	1.0

with  $NR$  only was 0.76, while the  $fpr$  for a model with only  $P$  was 1.0 (Table 8).

## Discussion and conclusion

In this paper, we have shown that MAST is effective at removing false-positive detections from radio telemetry data. MAST provides a framework for transparent, objective, and repeatable telemetry projects for wildlife conservation surveys, and increases the efficiency of data processing. We have demonstrated this effectiveness with a range of scenarios that included: multiple manufacturers, multiple antenna configurations, wide-band receivers switching between antennas, single-band receivers scanning multiple frequencies, and highly correlated predictor variables. The cross-validation procedure assessed the ability of the algorithm to correctly identify and remove false-positive detections and provided a means to rank and compare models. MAST is available open-source on the Python Package Index (<https://pypi.org/project/pymast/>) and is copyrighted under the MIT License. While MAST was developed with radio telemetry in mind, applications such as Simpfendorfer et al. [36] suggest that the PDH approach can be extended to acoustic telemetry.

The predictor variables used for MAST describe the characteristics of valid and false-positive detections and were able to discriminate between detection classes with high degrees of accuracy (Tables 5–8). For unbalanced NB applications such as this, predictor variables that provide the greatest degree of difference between valid and false-positive detections will perform the best [15]. Almost all known false-positive detections had low HR, while valid detections had high HR, even in cases where a receiver was switching between frequencies or antennas (Figs. 4, 6). However, valid detections had only slightly higher power (Figs. 3–6), which suggests that this predictor may not be strong enough on its own; as is evident in Tables 5–8 where a reduced model using only RSS as a predictor was compared with a more saturated model. Higher power was associated with valid detections  $P(\text{highpower}|\text{valid}) > P(\text{highpower}|\text{falsepositive})$ ; however, there were known false-positive detections with high power. Therefore, if power is the only predictor, then all high-powered detections are classified as valid, which resulted in a higher  $fpr$  for those models (Tables 5–8). Likewise, higher  $NR$  was associated with false-positive detections (Figs. 3, 4), inclusion of  $NR$  as the only predictor in the model may bias the model towards classifying most detections as false positive. The Lotek receivers with Yagi antennas produced lower noise (Fig. 6) than Sigma Eight Orion receivers with Yagi antennas (Figs. 3, 4), although this was probably related to site specific conditions and not manufacturer. Results of the  $k$ -fold

cross-validation procedure suggest the predictor variables designed for MAST provide excellent discriminatory power regardless of manufacturer, or antenna configuration.

The naïve assumption that all predictor variables were independent is remarkably robust, even when multicollinearity is present. This assumption almost never holds for natural data sets [23], including MAST where HR and CRL were highly correlated. However, MAST still performed well. Stephens et al. [38] found the naïve assumption valid even in the presence of strong attribute dependence due to cancellations between errors in the likelihoods of different classes. They do caution that this assumption will breakdown when the set of predictor variables is large and suggest combining correlated features then judging the relative performance of the algorithm with and without the combined variables. While we did not combine variables, we found that removing CRL as a predictor improved the  $fpr$  (Tables 5 and 7) or had no effect (Tables 6 and 8). The  $fpr$  of the saturated model was generally low and within acceptable ranges.

Discretization of continuous predictor variables did not appear to affect the performance of the NB algorithm as evidenced by high  $sen$  and low  $fpr$ . We found discretization simplified the calculation of likelihood and was able to describe complex multi-modal PMFs. However, Dougherty et al. [11] have shown that the equal width interval method chosen for MAST is vulnerable to outliers that may drastically skew ranges. While an examination of PMFs produced for this project did not demonstrate an effect of outliers, the authors will amend MAST in future iterations to include other discretization methods should the need arise. We do caution against too many bins. One could use the PDH itself as a predictor, however, the number of bins produced and low number of observations in each would not produce informative likelihoods and could lead to single precision floating point decimal errors.

MAST is able to account for receivers that switch between antennas or frequencies. However, care must be taken when setting up study parameters such as the scan time of the receiver and the pulse rate of the tag. Poor setup will result in missed detections and sparse PDHs, which would reduce the power of the HR and CRL predictors. For this study, the scan time was set to an interval slightly longer than the pulse rate of the tags. This increased the likelihood the tag was detected when the receiver cycled back to the original frequency and reduced the likelihood of missed detections. While we were careful, our setup still resulted in missing detections. However, it only slightly affected HR and CRL as evident by little to no difference in HR for valid detections across scenarios.

By relaxing the *MAP* criterion, a researcher can control their level of confidence in a given score considered to be true. For example, a ratio close to unity is weak evidence of a false positive or negative detection. This application used a strict interpretation of the *MAP* hypothesis, but rather than implementing *MAP*, we can require the rejection ratio to be smaller than a certain threshold before we prompt classification as false positive. Conversely, we can require overwhelming evidence to accept that a record is true and mandate that the ratio is greater than 1.0 by a certain threshold. However, in doing this, one would expect to create instances of false negatives, where marginal detections are classified as false positive. The algorithm is meant to be adaptable to study goals and site conditions, where one may want a stricter classifier due to electronic or radio interference, or where the consequences of mis-assignment dictate a conservative approach. Researchers must recognize and balance the two types of errors in setting the classification threshold.

Likewise, the researcher may want to negate the weight of evidence provided by the prior probability. There may be instances where there is an overwhelming number of detections in one class versus another. So much so, that obvious misclassifications occur. In these cases, the researcher can implement an uninformative prior, where there is an equal split between detections classes. With an uninformative prior, our problem reduces to a maximum likelihood approach.

This study lumped multiple receivers of like-types together when creating training data for classification and cross-validation purposes. By lumping receivers together, we reduced positive bias that was introduced by Laplace smoothing, increased statistical power, and generalized training data making it more applicable to other locations within the same study. However, care must be taken when lumping. Single-band receivers should not be lumped with wide-band receivers. Likewise, receivers with Yagi antennas should not be grouped with receivers that have dipole antennas. Even if two receivers have the same type of antenna, they should not be grouped together if their noise profiles are entirely different. Noisy sites happen, and they are fundamentally different from low-noise sites or receiver-antenna configurations that produce less noise.

The algorithm only classifies a single receiver at a time; however, it may be possible to use recapture histories from multiple receivers to inform on movement or overlap between receivers. For example, it is possible to build receiver-to-receiver logic that identifies improbable site progressions. Once a telemetry network has been expressed as a graph with telemetry receivers for nodes and logical migratory pathways described with directed edges, we can calculate two important matrices:

the adjacency and distance matrix. With an adjacency matrix, we can identify illogical movements. For example, a fish cannot swim upstream through the powerhouse of a dam.

Choice of a tag's burst rate is critical. With short burst rates, we are able to assess fine scale movements and increase the temporal accuracy of assessments of migratory delay. However, short burst rates come at the cost of reduced predictive power. We can increase the discriminatory power with longer burst rates, as detections occurring between defined pulses are not even considered for inclusion in the PDH. However, this problem will become moot when all receiver manufacturers increase time-step precision to sub-second intervals. Until then, the likelihood of two false-positive detections occurring in series is low, as is evident in Figs. 3B, 4B, and 6B, which suggests that this problem may not bias posteriors too much.

The intention of this effort was to develop a method that assists the researcher in culling the volume of false-positive data that telemetry projects produce. The data must still go through quality control procedures. When coupled with a robust data management system, MAST increases the efficiency of data processing and provides the researcher with a quantitative measure backed up with evidence. A researcher is then able to diagnose why a detection was classified a certain way leading to objective and repeatable studies.

Telemetry methods have undergone substantial development and evolution in recent decades, with radio and acoustic telemetry methods offering complementary capabilities [2, 10, 27]. With improved technologies, costs have declined and dimensions of radio (including passive integrated transponders) and acoustic tags have gotten smaller, broadening the scope and scale of species and studies that these tools can support. All telemetry approaches have the potential to yield false positives, and the ability to reliably score individual detections is needed to support this development. This will allow for new applications, such as the use of received power from receivers with overlapping detection ranges among multiple antennas to position tags in space [4, 16]. Retaining valid detections improves per-transmission detection efficiency, which will improve the power and reliability of mark-recapture studies [32] and other investigations of habitat use, movement, and survival [1, 30]. These data are essential for improving understanding of habitat connectivity and movement corridors as well as identifying and mitigating risks to populations. By providing an objective measure of data quality this approach should help managers working in a variety of environments and habitat types to better manage species of interest.

## Appendix 1

See Table 9

**Table 9** A list of terms and their meaning

Variable	Meaning
NB	Naïve Bayes Classifier
RSS	Received signal strength or detection power, model parameter
HR	Hit ratio, model parameter
CRL	Consecutive record length, model parameter
NR	Noise ratio, model parameter
$\delta^2 L$	Difference in time-lab between detections, model parameter
PDH	Proximate detection history
$P(C_i F_1, \dots, F_n)$	Posterior probability of a detection belonging to detection class $C_i$ given the observed predictor variables $F_1, \dots, F_n$ .
$P(C_i)$	Prior probability of the $i$ th detection class occurring where ( $C \in \{\text{Valid}, \text{FalsePositive}\}$ )
$P(F_j C_i)$	Likelihood (conditional probability) of the $j$ th observed predictor ( $F_j$ ) value given the $i$ th detection class ( $C_i$ )
$t_p$	True positive
$f_p$	False positive
$f_n$	False negative
$t_n$	True negative
sen	Sensitivity, model metric, $sen = t_p/(t_p + f_n)$
spc	Specificity, model metric, $spc = t_n/(f_p + t_n)$
npv	Negative predictive value, model metric, $npv = t_n/(f_n + t_n)$
ppv	Positive predictive value, model metric, $ppv = t_p/(t_p + f_p)$
fpr	False positive rate, model metric, $fpr = f_p/(f_p + t_n)$
PRC	Precision-recall curve
AUC	Area under the curve statistic, integral of the PRC
$\kappa$	Cohen's Kappa measure of concordance

### Acknowledgements

Steven Leach and the FirstLight Relicensing Team for use of data and facilities, Hannah Gorin for the map, and to Thomas Meyer for review. Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

### Authors' contributions

The lead author, KPN, was the primary Python developer. TC-S developed the iterative reclassification routine and train-on-study tag methodology. Each author made significant contributions to the paper. Both the authors have read and approved the final manuscript.

### Funding

The authors did not receive funding to produce this manuscript.

### Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request. A thorough ReadMe describing the software, installation, and use guiding researchers through a radio-telemetry project, including false positive removal routines, can be found at the project's GitHub: <https://github.com/knebiolo/MAST>.

### Declarations

#### Ethics approval and consent to participate

The study, methods, and handling procedures were approved as part of FERC federal-relicensing resource-studies for the Northfield Mountain Pumped

Storage Project (P-2485) and Turners Falls Hydroelectric Project (P-1889) on projects 3.3.2: Evaluation of Upstream and Downstream Passage of Adult American Shad, and 3.3.15: Assessment of Adult Sea Lamprey Spawning within the Turners Falls Project and Northfield Mountain Project Area.

#### Consent for publication

Authors and their respective organizations consent to publication.

#### Competing interests

The authors have no competing interests.

Received: 28 February 2021 Accepted: 4 January 2022

Published: 13 January 2022

### References

1. Aarts G, MacKenzie M, McConnell B, Fedak M, Matthiopoulos J. Estimating space-use and habitat preference from wildlife telemetry data. Ecography. 2008;31:140–60. <https://doi.org/10.1111/j.2007.0906-7590.05236.x>.
2. Adams NS, Beeman JW, Eiler JH. Telemetry techniques: a user guide for fisheries research. Bethesda: American Fisheries Society; 2012.
3. Almeida TA, Yamakami A. Content-based spam filtering. In: The 2010 International Joint Conference on Neural Networks IJCNN, IEEE. 2010. pp. 1–7.

4. Anderson-Sprecher R, Ledolter J. State-space analysis of wildlife telemetry data. *J Am Stat Assoc.* 1991;86:596–602.
5. Barber D. Bayesian reasoning and machine learning. Cambridge: Cambridge University Press; 2014.
6. Beeman JW, Perry RW. Bias from false-positive detections and strategies for their removal in studies using telemetry. New York: American Fisheries Society; 2012.
7. Brownscombe JW, Lédée EJ, Raby GD, Struthers DP, Gutowsky LF, Nguyen VM, et al. Conducting and interpreting fish telemetry studies: considerations for researchers and resource managers. *Rev Fish Biol Fisheries.* 2019;29:369–400.
8. Chambert T, Miller DA, Nichols JD. Modeling false positive detections in species occurrence data under different study designs. *Ecology.* 2015;96:332–9.
9. Cormack RM. Estimates of survival from the sighting of marked animals. *Biometrika.* 1964;51:429–38.
10. Craighead FC, Craighead JJ. Data on grizzly bear denning activities and behavior obtained by using wildlife telemetry. *Bears Biol Manag.* 1972. <https://doi.org/10.2307/3872573>.
11. Dougherty J, Kohavi R, Sahami M. Supervised and unsupervised discretization of continuous features. *Mach Learn Proc.* 1995. <https://doi.org/10.1016/B978-1-55860-377-6.50032-3>.
12. Feki-Sahnoun W, Njah H, Hamza A, Barraj N, Mahfoudi M, Rebai A, Hassen MB. Using general linear model, Bayesian Networks and Naïve Bayes classifier for prediction of *Karenia selliformis* occurrences and blooms. *Eco Inform.* 2018;43:12–23.
13. Fernandes JA, Irigoien X, Goikoetxea N, Lozano JA, Inza I, Pérez A, Bode A. Fish recruitment prediction, using robust supervised classification methods. *Ecol Model.* 2010;221:338–52.
14. Granik M, Mesyura V. Fake news detection using naïve Bayes classifier. In: 2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON). IEEE. 2017. p. 900–3.
15. Mladenic, Dunja and Marko Grobelnik. "Feature selection for unbalanced class distribution and Naïve Bayes." In Proceedings of the 16th International Conference on Machine Learning (ICML). Burlington, MA: Morgan Kaufmann Publishers, 1999. p. 258–267.
16. Harbicht AB, Castro-Santos T, Ardren WR, Gorsky D, Fraser DJ. Novel, continuous monitoring of fine-scale movement using fixed-position radiotelemetry arrays and random forest location fingerprinting. *Methods Ecol Evol.* 2017;8:850–9.
17. Heupel MR, Semmens JM, Hobday AJ. Automated acoustic tracking of aquatic animals: scales, design and deployment of listening station arrays. *Mar Freshw Res.* 2006;57:1–13. <https://doi.org/10.1071/MF05091>.
18. Hsu, Chun-Nan, Hung-Ju Huang and Tzu-Tsung Wong. "Why Discretization Works for Naïve Bayesian Classifiers." Proceedings of the Seventeenth International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann Publishers, 2000. p. 399–406.
19. Jolly GM. Explicit estimates from capture-recapture data with both death and immigration-stochastic model. *Biometrika.* 1965;52:225–47.
20. Kosmopoulos, Aris, Georgios Palioras and Ion Androutsopoulos. "Adaptive spam filtering using only naïve bayes text classifiers." Proceedings of the Fifth Conference on Email and Anti-Spam (CEAS). Mountain View, CA: Microsoft Research Silicon Valley, 2008. p. 1–2.
21. Laplace, Pierre Simon. Théorie analytique des probabilités. Paris: Courcier, 1820.
22. Lehtinen A, Olsson J, Bergström L, Bergström U, Bryhn A, Fredriksson R, Uusitalo L. Evaluating complex relationships between ecological indicators and environmental factors in the Baltic Sea: a machine learning approach. *Ecol Ind.* 2019;101:117–25.
23. Lewis DD. Naïve (Bayes) at forty: the independence assumption in information retrieval. *Eur Conf Mach Learn.* 1998. <https://doi.org/10.1007/BFb0026666>.
24. Ling, Charles X., Jin Huang and Harry Zhang. "AUC: A Better Measure than Accuracy in Comparing Learning Algorithms." Advances in Artificial Intelligence. Ed. Yang Xiang and Brahim Chaib-draa. Berlin: Springer Berlin Heidelberg, 2003. 329–341.
25. McClintock BT, Bailey LL, Pollock KH, Simons TR. Unmodeled observation error induces bias when inferring patterns and dynamics of species occurrence via aural detections. *Ecology.* 2010;91:2446–54.
26. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med.* Vol. 22, p. 276–82. 2012. <https://pubmed.ncbi.nlm.nih.gov/23092060>. Accessed 10 Aug 2021
27. McMichael GA, Eppard MB, Carlson TJ, Carter JA, Ebberts BD, Brown RS, Deng ZD. The juvenile salmon acoustic telemetry system: a new tool. *Fisheries.* 2010;35:9–22.
28. Minsky M. Steps toward artificial intelligence. *Proc IRE.* 1961;49:8–30.
29. Montgomery RA, Roloff GJ, Hoef JM. Implications of ignoring telemetry error on inference in wildlife resource use models. *J Wildl Manag.* 2011;75:702–8.
30. Montgomery RA, Roloff GJ, Hoef JM, Millspaugh JJ. Can we accurately characterize wildlife resource use when telemetry data are imprecise? *J Wildl Manag.* 2010;74:1917–25.
31. Nebiolo KP, Castro-Santos TR. (2021). mast. github: <https://github.com/knebiolo/MAST>. 17 July 2021.
32. Perry RW, Castro-Santos T, Holbrook CM, Sandford BP. Using mark-recapture models to estimate survival from telemetry data. Bethesda: American Fisheries Society; 2012.
33. Rosset S. "Model Selection via the AUC." Proceedings of the Twenty-First International Conference on Machine Learning. New York, NY, USA: Association for Computing Machinery, 2004. 89. <https://doi.org/10.1145/1015330.1015400>.
34. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE.* Vol. 10, p. e0118432. 2015. <https://pubmed.ncbi.nlm.nih.gov/25738806>. Accessed 6 Aug 2021
35. Seber GA. A note on the multiple-recapture census. *Biometrika.* 1965;52:249–59.
36. Simpfendorfer CA, Huveneers C, Steckenreuter A, Tattersall K, Hoenner X, Harcourt R, Heupel MR. Ghosts in the data: false detections in VEMCO pulse position modulation acoustic telemetry monitoring equipment. *Anim Biotelemetry.* 2015;3:1–10.
37. Skalski JR, Lady J, Townsend R, Giorgi AE, Stevenson JR, Peven CM, McDonald RD. Estimating in-river survival of migrating salmonid smolts using radiotelemetry. *Can J Fish Aquat Sci.* 2001;58:1987–97.
38. Stephens CR, Huerta HF, Linares AR. When is the Naïve Bayes approximation not so naïve? *Mach Learn.* 2018;107:397–441.
39. Stone M. Cross-validatory choice and assessment of statistical predictions. *J Roy Stat Soc Ser B (Methodol).* 1974;36:111–33.
40. Yang M-C, Huang C-S, Chen J-H, Chang R-F. Whole breast lesion detection using naïve bayes classifier for portable ultrasound. *Ultrasound Med Biol.* 2012;38:1870–80.
41. Yang Y, Webb GI. A comparative study of discretization methods for naïve-bayes classifiers. *Proceedings of PKAW.* Vol. 2002. 2002.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.