



HAL
open science

Detecting and reducing heterogeneity of error in acoustic classification

Oliver C Metcalf, Jos Barlow, Yves Bas, Erika Berenguer, Christian Devenish, Filipe França, Stuart Marsden, Charlotte Smith, Alexander C Lees

► **To cite this version:**

Oliver C Metcalf, Jos Barlow, Yves Bas, Erika Berenguer, Christian Devenish, et al.. Detecting and reducing heterogeneity of error in acoustic classification. *Methods in Ecology and Evolution*, 2022, 13 (11), pp.2559-2571. 10.1111/2041-210X.13967 . hal-03948391

HAL Id: hal-03948391

<https://hal.science/hal-03948391v1>









Submitted on 20 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

RESEARCH ARTICLE

Detecting and reducing heterogeneity of error in acoustic classification

Oliver C. Metcalf¹  | Jos Barlow²  | Yves Bas^{3,4} | Erika Berenguer^{3,5}  |
Christian Devenish^{1,6}  | Filipe França⁷  | Stuart Marsden¹  | Charlotte Smith³  |
Alexander C. Lees^{1,8} 

¹Division of Biology and Conservation Ecology, Department of Natural Sciences, Manchester Metropolitan University, Manchester, UK; ²Lancaster Environment Centre, Lancaster University, Lancaster, UK; ³CESCO, MNHN, CNRS, Sorbonne Univ, Paris, France; ⁴CEFE, Univ Montpellier, CNRS, EPHE, IRD, Montpellier, France; ⁵Environmental Change Institute, University of Oxford, Oxford, UK; ⁶Nature Metrics, Surrey Research Park, Guildford, UK; ⁷School of Biological Sciences, University of Bristol, Bristol, UK and ⁸Cornell Lab of Ornithology, Cornell University, Ithaca, New York, USA

Correspondence

Oliver C. Metcalf

Email: o.metcalf@mmu.ac.uk**Funding information**

BNP Paribas Foundation's Climate and Biodiversity Initiative (Project Bioclimate); PELD-RAS, Grant/Award Number: CNPq/CAPES/PELD 441659/2016-0; AFIRE, Grant/Award Number: NE/P004512/1; ECOFOR, Grant/Award Number: NE/K016431/1

Handling Editor: Camille Desjonquères

[Correction added on 26 September 2022, after first online publication: Affiliation 4 has been changed from 'CEFE, Univ Montpellier, CNRS, EPHE, IRD, Univ Paul Valéry Montpellier 3, Montpellier, France' to 'CEFE, Univ Montpellier, CNRS, EPHE, IRD, Montpellier, France'.]

Abstract

1. Passive acoustic monitoring can be an effective method for monitoring species, allowing the assembly of large audio datasets, removing logistical constraints in data collection and reducing anthropogenic monitoring disturbances. However, the analysis of large acoustic datasets is challenging and fully automated machine learning processes are rarely developed or implemented in ecological field studies. One of the greatest uncertainties hindering the development of these methods is spatial generalisability—can an algorithm trained on data from one place be used elsewhere?
2. We demonstrate that heterogeneity of error across space is a problem that could go undetected using common classification accuracy metrics. Second, we develop a method to assess the extent of heterogeneity of error in a random forest classification model for six Amazonian bird species. Finally, we propose two complementary ways to reduce heterogeneity of error, by (i) accounting for it in the thresholding process and (ii) using a secondary classifier that uses contextual data.
3. We found that using a thresholding approach that accounted for heterogeneity of precision error reduced the coefficient of variation of the precision score from a mean of 0.61 ± 0.17 (*SD*) to 0.41 ± 0.25 in comparison to the initial classification with threshold selection based on *F*-score. The use of a secondary, contextual classification with thresholding selection accounting for heterogeneity of precision reduced it further still, to 0.16 ± 0.13 , and was significantly lower than the initial classification in all but one species. Mean average precision scores increased, from 0.66 ± 0.4 for the initial classification, to 0.95 ± 0.19 , a significant improvement for all species.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of British Ecological Society.

4. We recommend assessing—and if necessary correcting for—heterogeneity of precision error when using automated classification on acoustic data to quantify species presence as a function of an environmental, spatial or temporal predictor variable.

KEYWORDS

automated signal recognition, autonomous recording unit, bioacoustics, ecoacoustics, machine-learning

1 | INTRODUCTION

Passive acoustic monitoring (PAM) is an increasingly common ecological survey tool. PAM has many advantages over traditional survey methods, facilitating sampling across larger spatiotemporal scales, and in places where access is logistically challenging, thus increasing cost-efficiency (Darras et al., 2018; Gibb et al., 2019). Given this capacity to record audio data autonomously, PAM can accrue very large datasets, often too large to analyse manually. Automated classification presents a solution to this challenge. A variety of approaches have been trialled including template matching, machine learning techniques such as clustering and random forests along with deep-learning algorithms such as convolutional neural networks (Priyadarshani et al., 2018). However, outside of chiropterology, few studies have used fully automated classification to answer applied ecological questions in terrestrial landscapes, and especially for the challenge of multi-species classification across large audio datasets from tropical forests.

One of the issues facing the applied use of automated classification methods is how readily algorithms can be generalised—how well they can be applied to new data across time and space (Priyadarshani et al., 2018; Stowell, Petrusková, et al., 2019). The accuracy and generalisability of supervised machine learning techniques—those using labels generated by humans as training data—are heavily dependent on the nature of the labelled training data used (Knight et al., 2019, 2020; Towsey et al., 2012). Achieving high classification performance generalised across a range of conditions requires training data that is representative of the variation in the data on which the classifier is to be used, while also achieving a balance between classes (Towsey et al., 2012; Zhong et al., 2020).

Classification performance is impacted by variation in the underlying audio data which may be intrinsic or extrinsic to the target species. Intrinsic sources include variation in an individual animal's vocalisations—between individuals, populations, geographically and temporally. For classifiers to perform well considering these forms of intrinsic variation, a representative sample of vocal variation for each species as training data is needed. Although potentially challenging for many species, obtaining such representative data is facilitated when ecological knowledge can be used to anticipate when and where training data can be obtained.

Sources of potential extrinsic variation include; other sounds that overlap with target signals in time and/or frequency such as vocalisations from other species, the prevalence of which depends

on variation in co-occurrence and abundance of species with similar vocalisations (Tobias et al., 2014), plus other sources of biophony, geophony or anthropophony; and environmental factors that can impact sound propagation, such as weather conditions and the density of surrounding vegetation (Yip et al., 2017). There are also potential sources of audio variation that fit between the two categories—for example when a source extrinsic to the target animal causes an intrinsic change to the vocalisation of the target species. Responses to predators or competitors, such as duetting or lekking birds can fit in this category (Mennill & Vehrencamp, 2008), as do broadband sources of noise such as cicadas or vehicle engines that may cause a change in the frequency at which calls are made.

In comparison to intrinsic variation, these sources may be more challenging to represent well in a training dataset, as they are likely to be both more variable and less predictable. In particular, the use of online libraries such as Macaulay Library (<https://www.macaulaylibrary.org>), xeno-canto (<https://www.xeno-canto.org>) or AmphibiaWeb (<https://amphibiaweb.org>) could cause training datasets to be less representative, as recordings made with directional microphones of single species have high signal-to-noise ratios (Priyadarshani et al., 2018; Towsey et al., 2012), and are unrepresentative of external sources of acoustic variation.

Errors associated with variation in noise could be resolved by applying noise reduction techniques (e.g. Priyadarshani et al., 2020). While this approach is undoubtedly effective in some or even many, cases, it is a difficult approach when dealing with a multi-species classifier. Here, one target species call is 'signal' when considering its own classification, but can also be considered as 'noise', and impact classification accuracy, for all the other species included in the classifier. Therefore, even removing all of the sound that could be considered 'noise' in all circumstances still leaves a considerable amount of variable sonotypes in the training data. Furthermore, noise reduction is not a universally agreed approach to improve classification accuracy, for instance the OPEN SOUNDSCAPE package (Lapp et al., 2022) offers many approaches to augment and increase noise in training data—for instance overlaying training samples on top of each other.

Intrinsic and extrinsic variation in audio data make obtaining truly representative datasets extremely difficult if the classifier is intended to operate across large spatial extents, long periods of time or across heterogeneous habitats (Zhong et al., 2020). In these cases, providing representative data labels at local scales would require huge increases in labelling effort (LeBien et al., 2020). The inevitable

shortfalls in obtaining representative datasets mean that classification accuracies obtained on test datasets may not translate to field conditions (Stowell, Wood, et al., 2019)—a common problem in supervised learning fields termed covariate shift (Shimodaira, 2000). Many classification algorithms may exhibit biases that will lead to heterogeneous error structure when exposed to these variations. Heterogeneity of error could be especially problematic if the covariate responsible for shifting classification accuracy is the same as that being studied for ecological purposes. This may occur, for example, when error varies; spatially in a space-for-time swap experimental design, temporally in a phenological study or at ecotones where replacement species with similar vocalisations may overlap when examining habitat preference.

Automated classification models are typically assessed by deriving a range of accuracy metrics from a manually labelled test dataset that has been randomly subsampled from the training data, or independently subsampled from the dataset on which the algorithm is to be applied (Knight et al., 2017; Priyadarshani et al., 2018). Following Knight et al. (2017), precision, recall, *F*-score and area under the curve (AUC) have been widely adopted to determine classification algorithm performance (Table 1). However, these methods fail to explicitly test the generalisability of the algorithm across the gradient of a shifting covariate. Consequently, using only these metrics to assess classification performance risks masking high variability in false-positive error and subsequent confounding of results if error covaries with a variable of ecological interest.

Although both heterogeneity in false-positive and false-negative errors can be detrimental, we focus here on false-positive errors, and consequently the precision metric, in keeping with other studies highlighting precision as important for acoustic surveys (e.g. Juodakis et al., 2021). Unlike false negatives, false positives cannot be mitigated by summarising presence across longer time periods or spatial extents (Metcalf et al., 2019), violate the assumptions of many standard methods for modelling detection probability and can lead to poor model inference (Royle & Link, 2006). While research has been conducted into the overall reduction in false-positive error in ecoacoustic datasets (e.g. Balantic & Donovan, 2020; Clare et al., 2021; Knight et al., 2020), and in reducing the variation in error when analysing the ecological variables through the use of occupancy models (e.g. Chambert et al., 2018; Rempel et al., 2019), there has been limited research into methods to reduce variability of false-positive error at the classification stage.

We present a case study using automated classification of an Amazonian PAM dataset to highlight the challenges in detecting heterogeneity of precision error. First, by incorporating a measure of heterogeneity into the metric used to set a threshold for classification confidence score, so that thresholding goes beyond a two-way trade-off between precision and recall, and instead becomes a three-way trade-off between precision, recall and heterogeneity of (precision) error. Second, by incorporating a secondary classification that accounts for the context in which the classification is made, through the use of neighbouring classification scores for the target and other species, environmental and temporal data.

TABLE 1 A glossary of key terms relating to classification accuracy

Metric	Definition	Formula
True Positive (TP)	One of four potential outcomes of classification. True positives are a correct positive prediction (i.e. a species is actually present, and predicted to be present)	
False Positive (FP)	An incorrect positive prediction (i.e. a species is predicted to be present, but is actually absent)	
True Negative (TN)	A correct negative prediction (i.e. a species is predicted to be absent and is actually absent)	
False Negative (FN)	An incorrect negative prediction (i.e. a species is predicted to be absent, but is actually present)	
Precision	The percentage of correct positive predictions in all positive predictions	$\frac{TP}{(TP + FP)}$
Recall	The percentage of all possible positive events that are correctly predicted	$\frac{TP}{(TP + FN)}$
Precision-Recall Area-Under-the-Curve (PR-AUC)	Precision-Recall curves summarise the trade-off between precision and recall at different classification confidence score thresholds. The area-under-the-curve is therefore a good metric for comparing classification accuracy independent of threshold selection	
<i>F</i> -Score	The harmonic mean of precision and recall	$\beta = 1$ $\frac{(\beta^2 + 1) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}$
<i>F</i> ^{0.5} -Score	The harmonic mean of precision and recall, weighting precision as twice as important as recall	$\beta = 0.5$ $\frac{(\beta^2 + 1) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}$

2 | MATERIALS AND METHODS

2.1 | Data collection

We used PAM to collect data from 29 survey points from a 10,000 km² study area in eastern Amazonia encompassing parts of the municipalities of Santarém, Belterra and Mojuí dos Campos (≈3°2'45.6"S, 54°56'49.2"W) in the Brazilian state of Pará. Survey points located in upland *terra firme* rainforest had a minimum separation of 2 km to minimise spatial dependence. The experimental design was intended to investigate differences in socioecological responses to forest disturbance (see Metcalf et al., 2021 for full details), and the survey points are thus distributed across an anthropogenic disturbance gradient. Consequently, we address spatial heterogeneity of error in this study across the 29 points.

We collected two sets of acoustic data. The first was used exclusively for the purpose of training the classification algorithm (hereafter 'Acoustic Dataset 1') between 20 November and 30 December 2017 from two of our survey points, each with three recorders 150 m apart. The second set of acoustic data (hereafter 'Study Dataset') was collected from a single recording location at each of the 29 survey points between 12 June and 16 August 2018. Recordings for the Study Dataset were made over 1–2 recording periods at each point, with recording period varying in length between 3 and 22 days. This gave as optimal a coverage of nocturnal species as logistical limitations would allow (nocturnal species vocalisation rate may be impacted by the lunar cycle). A minimum of 13 days were surveyed at each location. Both datasets were collected using Frontier Labs Bioacoustic Recorders (Frontier Labs, 2015). Recordings were continuous across the diel cycle, and were filtered afterwards to only include astronomical night, measured using the `SUNCALC` package in R (Thieurmél & Elmarhraoui, 2019) using the location of our field station (coordinates given above). Full details of recording periods, equipment and protocols for each location are given in SOM Appendix 1. This work was conducted under SISBIO permit number 53271–13.

2.2 | Automated classification and verification datasets

Tadarida is an open-source toolbox for detecting, labelling and classifying sounds (Bas et al., 2017) and shown to be effective at classifying various European species of insects and mammals (Barré et al., 2019; Newson et al., 2017). We used Tadarida to build a classifier in R (R Core Team, 2020) for seven nocturnal bird species—four owl species; Southern Tawny-bellied Screech-owl *Megascops usta*, Crested Owl *Lophostrix cristata*, Spectacled Owl *Pulsatrix perspicillata*, Amazonian Pygmy-owl *Glaucidium hardyi* and three nightjar species; Ocellated Poorwill *Nyctiphrynus ocellatus*, Silky-tailed Nightjar *Antristomus sericeicaudatus* and Common Pauraque *Nyctidromus albicollis*. Tadarida first identifies sound events using a hysteresis function; the sound event starts when a high amplitude threshold is passed and ends when the signal-to-noise ratio drops below a second lower threshold. The program extracts 269 acoustic features (e.g. minimum and maximum frequency, peak frequency, duration) from each detected sound event and facilitates feature labelling for use as training data in a random forest classifier (see Bas et al., 2017 for full details). As multiple detected sound events may be identified from a single animal vocalisation, Tadarida uses simple rules to group events and makes classification predictions. Consequently, Tadarida works best over short-duration sound files, so we split all the recordings into 15 s files for all further processes. We limited all detections to those with the point of highest amplitude between 0.2 and 4.2 kHz which includes most terrestrial nocturnal vertebrates in the region.

To create training data for the classification algorithm, we undertook manual labelling of sound events detected by Tadarida (this labelled dataset hereafter referred to as 'Training Dataset 1') (Figure 1). Tadarida classifies every detected sound event, potentially

comprising tens of millions of sound events of which only a fraction are made by target species. Consequently, we chose to label additional classes beyond those of the target species so that common non-target sounds would be classified into those groups. We were unconcerned about classification accuracy for these non-target classes. During the labelling process, in addition to vocalisations of the seven target species, we identified 293 potential non-target classes by grouping similar sounds together, which included a range of biophony, geophony and rarely anthropophony. These sound types were simplified to a final set of sonotypes, either by merger or removal to give a final set of 59 sound types, including the seven classes for target species, as the classes the Tadarida algorithm classified detected sound events into. We identified each sonotype to species level where appropriate and possible. Where identification was not apparent, online resources such as the Macaulay Library, xeno-canto and AmphibiaWeb were consulted, and some call types were shared with relevant regional experts. If identification was still not possible, the sound type was left unidentified.

To obtain training data, we systematically searched for discrete sound types in our recording datasets. First, we labelled data from a subset of Acoustic Dataset 1. This subset consisted of 3 hr of recording per night—1 hr up to 30 min before sunrise, 1 hr commencing 30 min after sunset and 00:00–01:00, every 3rd night from each of the three recording units deployed, totalling 96 hr of data. While each sound file was searched systematically, training data were added based on the labeller's discretion so that not all calls in an extended vocalisation bout were necessarily included, especially for common sound types. As this data only came from three survey points, we additionally labelled data from eight other survey points in the Study Dataset, to increase spatial coverage and representation of forest disturbance, which can impact species abundance and composition. As the systematic search method adopted for Acoustic Dataset 1 was extremely time-consuming, we adopted a more targeted approach to labelling the additional data from the Study Dataset, choosing data from periods of time and places which we knew were most likely to contain vocalisations of species that were known to be present in the region (Lees et al., 2013), but were thus far under-represented in Training Set 1.

Finally, we supplemented labels generated from our own audio data using recordings from online archives (Macaulay Library, xeno-canto), also split into 15 s .wav files. We augmented all recordings by adding noise at six amplitude levels by combining each labelled file with three files identified in our own recordings as containing only heavy rainfall, and manipulating each by increasing the amplitude, giving six 'rain' files. Each rain file was then combined with each labelled file, increasing six-fold the number of labels in the training dataset. For full details of the Tadarida labelling and data augmentation process, see SOM Appendix 2.

2.3 | Assessing classifier performance

To assess classifier performance, we followed Knight et al. (2017) in using precision, recall, F-score and the area under the precision/

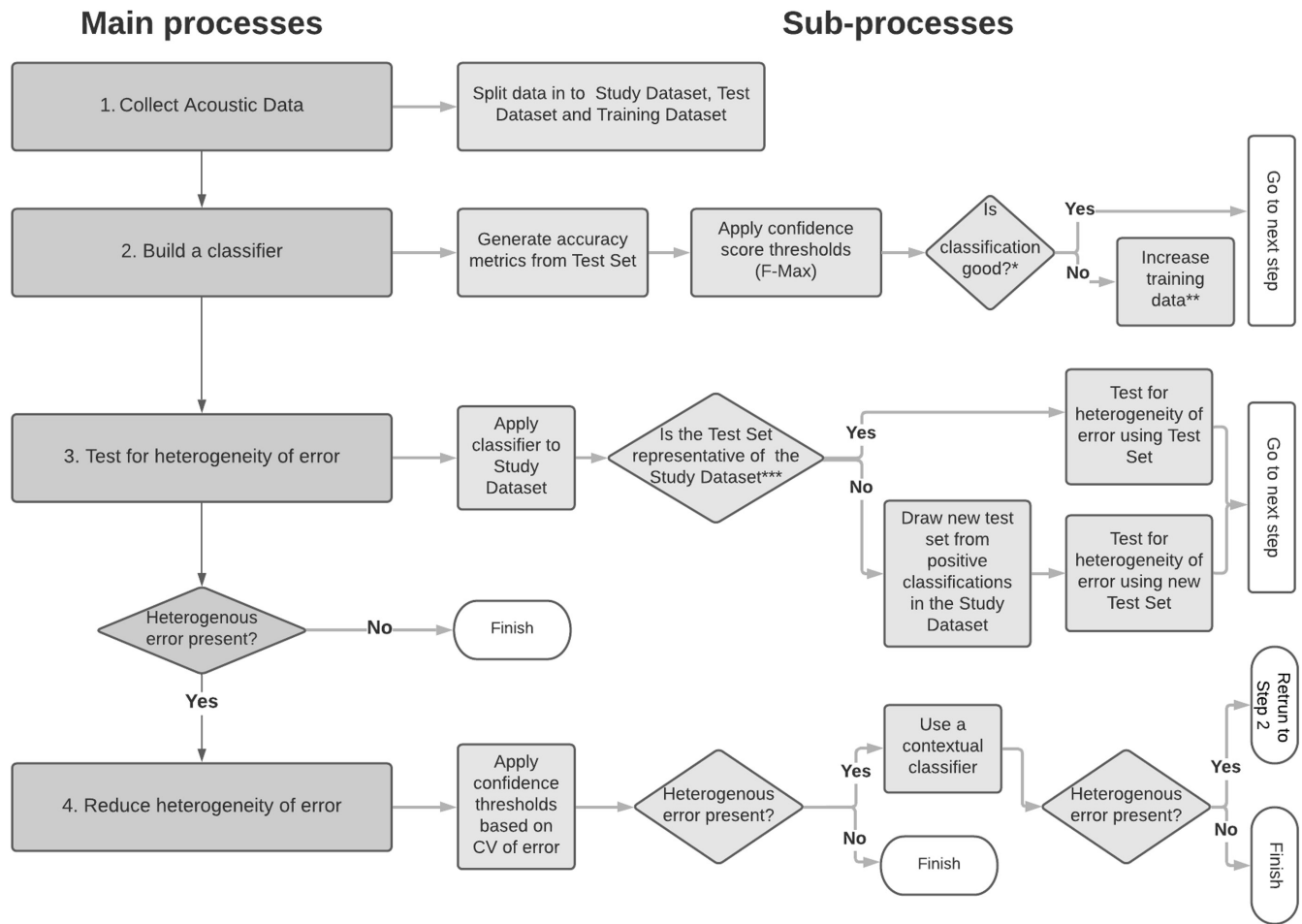


FIGURE 1 A flowchart showing the proposed processes required to assess and reduce heterogeneity of error in automated eoaacoustic classification.

recall curve (hereafter 'PR-AUC'). We calculated these scores from two test subsets of the Study Dataset by comparing Tadarida predictions against manual assessment of the same audio files. We detect the presence or absence of a species at 15s resolution. Tadarida gives a confidence score for each class included in the classification algorithm ($n = 59$) for each group of detected sound events (Tadarida made classification predictions on $n = 28,030,710$ grouped sound events in the Study Dataset), and we considered the species with the highest score in each grouped sound event to be the species predicted as present by Tadarida. We then summarised the predictions to the 15-s file level, taking the maximum score for each predicted species per file so that a species only had one score per file. All predictions for non-target species were discounted. Manual assessment of target species present within 15 s files was conducted using Raven Pro (Center for Conservation Bioacoustics, 2019).

When creating a test dataset for manual labelling, it is vital to use a sample large enough to be representative of the original dataset to accurately assess classification performance (Knight et al., 2017). However, test sets of 10% or even 1% may not be practical for very large datasets, as the subsample would take too long for manual assessment, as is the case here ($n = 1,081,780$ fifteen second files in

the Study Dataset). Instead, we made a subjective decision on test set sample size based on the trade-off between manual labelling effort and representativeness. Consequently, the first test set (Test Set 1) consisted of 2900 15 s files stratified such that 100 files were taken from each survey point, just under 0.3% of the total number of unique 15 s files in the study dataset.

Test Set 1 was randomly subsampled from the study dataset prior to classification following Knight et al. (2017), ensuring independence from the training dataset. A 15 s file was considered to be a true positive when Tadarida predicted a species presence within that file and the species was also detected in the file during manual assessment. A 15 s file was considered a false positive when Tadarida predicted a species presence within that file, but the species could not be detected in the file during manual assessment, and so on, respectively, for true negative and false negatives. A confusion matrix of true positives, true negatives, false positives and false negatives was created, from which we calculated precision, recall, F-Score and PR-AUC, following Knight et al. (2017). F-Scores can be weighted between Precision and Recall. We weighted precision as being twice as important as recall. This is because false positives are likely to have more severe consequences in spatial analysis of species occurrence (Balantic

& Donovan, 2020; Royle & Link, 2006), and as we used short files likely to be summarised over larger time-scales to avoid temporal autocorrelation in a later analysis, thus mitigating the impact of false negatives. Calculation of mean accuracy metrics was done by first calculating the scores for each target species across all sites, as opposed to at each survey site, and taking the unweighted mean from the seven values generated.

Knight et al. (2017) highlight the importance of score thresholds in assessing classifier performance, so we calculated $F^{0.5}$ -score at each possible threshold between 0 and 100 with intervals of 0.001. We used the maximum $F^{0.5}$ -score to determine the optimal threshold for each target species, and recalculated precision, recall and $F^{0.5}$ -score with that threshold applied.

2.4 | Detecting heterogeneous error

To look for heterogeneity in precision across our survey points, we calculated precision values for each target species at each survey point and computed the sample coefficient of variation (CV) using the `ENVSTATS` R package (Millard, 2013). However, we were concerned that the random subsampling approach used to select Test Set 1 may result in a non-representative subsample. In particular, selecting too few true-positive data points (e.g. instances in which rarer species were present) at intervals across the gradient of the predictor variable to be sensitive to variance in error. Consequently, the second test set (Test Set 2) was subsampled from the study dataset after classification, and consisted of 50 15 s files from each survey point per target species, with sampling based on the probability distribution of the classification score of the target species, using the `createDataPartition` function in the `CARET` R package (Kuhn, 2021). All files were manually assessed for the presence or the absence of the predicted target species. Optimal thresholds were determined in the same manner, and Precision, Recall and $F^{0.5}$ -Score were calculated. However, while this approach allows a more accurate assessment of the number of true and false positives, it comes at the expense of precise estimation of true and false negatives, which are necessarily excluded from the test dataset. To compensate, the recall scores for each species generated for Test Set 2 were multiplied by the recall score for Test Set 1 prior to the application of a threshold (hereafter 'corrected recall').

To see if heterogeneity of precision error detected by Test Set 2 varied from Test Set 1, we used Levene's test with Benjamini-Hochberg correction to compare the variance in precision score of each survey point for each species.

2.5 | Reducing heterogeneity of error

We used two methods to reduce both the absolute number of false positives and for comparability across study treatments—heterogeneity of precision error. First, we incorporated a measure of variance—CV, in the threshold selection process (hereafter

'CV-optimised threshold'). We normalised both the $F^{0.5}$ -score and CV values from each threshold interval to within the range zero to one. Instead of using the maximum possible $F^{0.5}$ -score, we included a term to favour threshold intervals with lower CVs, using the maximum of $(F^{0.5}\text{-Score} \times 2) + (1 - \text{CV})$, but weighted in favour of $F^{0.5}$ score. We recalculated accuracy metrics for Test Sets 1 and 2, comparing the CV scores for the two thresholding approaches.

Second, we built a second random forest model—a 'contextual classifier', trained on predictions from the Tadarida algorithm, time, date and acoustically derived environmental data. This secondary classification process was explicitly designed to reflect the contextual and environmental information used by experienced field observers to support identifications. This includes the environmental conditions, such as background noise levels (Simons et al., 2007), the presence of certain indicative species or groups which increases or decreases the likelihood of other species being present, and the observer's own capacity to overlook or ignore certain species (Kepler & Scott, 1981). We argue that an experienced observer uses an awareness of all these factors and adjusts identification confidence accordingly (Robinson et al., 2018). We have attempted to artificially replicate this process by providing a random forest both with the initial confidence scores made by the first classifier, and a wide array of contextual data which can be used to modify that initial prediction.

As we were primarily concerned with rectifying problems with precision, we designed the contextual classifier to operate only on those 15 s files already classified by Tadarida as having a target species present, similar to the secondary classification method used by Balantic and Donovan (2020) to reduce overall false-positive rates for template-matching. We took a random, stratified sample of files ($n = 2,900$, henceforth 'Training Set 2') in which Tadarida had classified the target species as present. We stratified the sample, taking 100 sound files from each location, further stratified into uneven quintiles of confidence score: 0–0.29, 0.3–0.49, 0.5–0.69, 0.7–0.84 and 0.85–1. These ranges were chosen to include a full range of confidence scores, while taking most samples from scores that were most likely to have a mix of true and false positives. When there were not enough samples within a quintile, which occurred mostly at high confidence ranges, additional samples were taken randomly. We manually checked for vocalisations of the target species in each sampled file and calculated the specificity of the classifier for each species at each survey location.

We built individual contextual classifiers for each of our seven target species using the stratified sample as training data. From each manually checked 15 s file, we calculated a series of variables to be used to train a new random forest. This included environmental data about each 15 s file; time, date, root mean square of the sound envelope calculated utilising the `SEEWAVE` package (Sueur et al., 2008) as a measure of background noise levels and the 'rainQ2' and 'rain_min' prediction of rainfall from the `HARDRAIN` package (Metcalf et al., 2020). We also used Tadarida confidence scores for each 15 s file as predictors. These included the maximum Tadarida confidence score of the target species, and for every class

in the Tadarida classifier ($n = 59$), the minimum, maximum, mean, 90th and 95th quantiles of the confidence scores. We also included the summed confidence score of each class per 15 s file, the ratio of classified sound events to the target species, and the three species most commonly detected in the file. In addition, we calculated the same confidence score variables across both 10 min and 1-hr periods, with the time centred around the file being classified. For the latter, we also calculated the 98th percentile of the classifier score for each class. This gave us a feature set of 716 predictors for each target species.

We used this feature set to build a distributed random forest classifier in the H2O package (LeDell et al., 2020), splitting the data into training (70%) and test (30%) datasets. Although random forests can handle a large number of predictor variables (Ross & Allen, 2014), we used the H2O variable importance function to ascertain relative variable importance, and rebuilt a final model with variables of an importance greater than 0.05, to avoid overtraining on unimportant predictors. Final models used a mean of 214 ± 140 variables (range 51–399)—see SOM Appendix 3 for more details of selected variables. Every 15 s file in which the Tadarida classifier had predicted the presence of a target species was then reclassified with a contextual classifier. We used a CV-optimised thresholding approach and calculated precision, corrected recall and F -score at the optimal threshold for each species.

We used pairwise Wilcoxon Signed-Rank tests with Benjamini–Hochberg correction to compare precision scores for each of the 29 points by species for the Tadarida classification with F -score optimised threshold, the Tadarida classification with a CV-optimised threshold and contextual classification with a CV-optimised threshold. We used Levene's test with Benjamini–Hochberg correction to compare the variance of precision between the three methods, to test if they had resulted in a significant reduction in the spatial heterogeneity of precision. Finally, we used Wilcoxon Signed-Rank tests with Benjamini–Hochberg corrections to compare the precision, adjusted recall, F -score and CV across all species for the Tadarida classification with CV-optimised thresholds and the contextual classification.

3 | RESULTS

3.1 | Classification performance

In general, the Tadarida classifier performed poorly prior to thresholding (Table 2). For most species, precision was low (0.27 ± 0.16 [Mean \pm SD])—as expected in a process that is classifying every sound event. Recall fared better (0.65 ± 0.2), with a minimum 0.42 for *P. perspicillata*. PR-AUC scores should better account for a large number of false positives at low threshold scores, but these were also low (0.54 ± 0.23). We found the classifier performed well on two species, *L. cristata* and *A. sericocaudatus* (PR-AUC scores = 0.78 and 0.84, respectively), while *P. perspicillata* had a PR-AUC of just 0.17. However, PR-AUC scores weight precision and

TABLE 2 Tadarida classification accuracy metrics without thresholds. $F^{0.5}$ -score is weighted twice as heavily in favour of precision than recall. These accuracy metrics are based on a randomly sampled test set of $n = 2,900$ 15 s files stratified to sample 100 files per survey point

Species	Precision	Recall	F -score	PR-AUC
<i>M. usta</i>	0.25	0.47	0.28	0.46
<i>L. cristata</i>	0.52	0.79	0.56	0.78
<i>P. perspicillata</i>	0.04	0.42	0.05	0.17
<i>G. hardyi</i>	0.40	0.45	0.41	0.52
<i>N. ocellatus</i>	0.29	0.75	0.33	0.61
<i>A. sericocaudatus</i>	0.26	0.92	0.30	0.84
<i>N. albicollis</i>	0.16	0.72	0.19	0.42

recall equally, and as here we prioritised precision over recall, it can still be possible to find thresholds that allow for high precision at the expense of recall. We found a dramatic increase in the precision and the $F^{0.5}$ -score of the classifier once an $F^{0.5}$ -score based threshold has been applied (Figure 2a in blue). Precision increases from 0.27 ± 0.16 to 0.83 ± 0.13 , while recall decreases to a mean of 0.38 ± 0.16 . $F^{0.5}$ -Score, reflecting the weighting of precision over recall, also increases substantially to 0.64 ± 0.17 from 0.54 ± 0.23 . However, some of the recall scores are particularly low—for example, just 0.12 for *P. perspicillata*.

3.2 | Detecting heterogeneity of error

We found considerable heterogeneity in precision when applying optimised $F^{0.5}$ -score thresholds to classification metrics derived from both test sets. Precision ranged from zero to one between survey sites for every species except *G. hardyi* with Test Set 2, which had a maximum precision of 0.95 (Figure 2b). Both test sets had high coefficients of variation in precision for many species, 0.57 ± 0.42 for Test Set 1 and 0.61 ± 0.17 for Test Set 2 (Figure 2c). High precision scores still masked much variation, for instance in the case of *A. sericocaudatus* which has precision scores of 0.96 and 0.89 for Test Sets 1 and 2, respectively, but precision CV of 0.43 and 0.50, respectively.

Levene's tests revealed a significantly higher estimate of precision variance with Test Set 2 compared to Test Set 1 for *P. perspicillata* ($F = 12.54$, $p = 0.006$). No significant difference was found between the precision variance in the other six target species, although Test Set 2 showed higher standard deviation of precision for four of the remaining six species, the exceptions being *M. usta* and *G. hardyi* (Figure 2b). Additionally, precision CV was higher with Test Set 2 than Test Set 1 for five species, the exceptions being *M. usta* and *N. albicollis* (Figure 2c). In addition, Test Set 1 produced the highest and lowest CV estimates, 0.28 for *P. perspicillata* and 1.50 for *N. albicollis*, perhaps indicative of low sample sizes causing relatively extreme estimations. A comparison of the other estimated accuracy metrics from Test Set 1 and Test Set 2 can be found in Figure 2a. Results hereafter refer to metrics derived from Test Set 2.

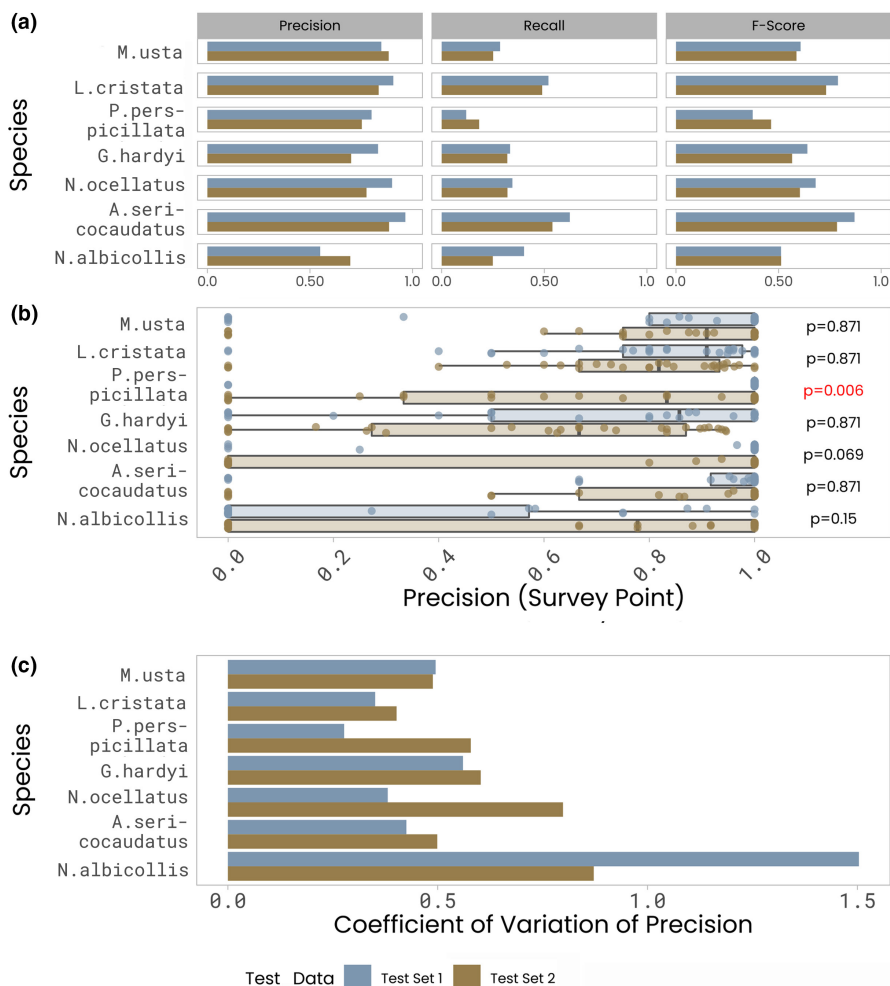


FIGURE 2 (a) A comparison of classification accuracy metrics for the Tadarida classifier, derived from Test Set 1 (randomly sampled prior to classification) and Test Set 2 (sampled post-classification based on the probability distribution of positive classifications), respectively, with $F^{0.5}$ -score optimised thresholds applied. Recall scores for Test Set 2 were multiplied by the recall score for Test Set 1 prior to the application of a threshold to compensate for Test Set 2 being sampled from Tadarida positive predictions. (b): Variation in precision scores by survey point, p -values (significance set at $p = 0.05$) show the result of Levene's test of homogeneity of variance between the two test sets. (c) Coefficient of variation (CV) in precision score calculated per species for each of the two test sets.

3.3 | Reducing heterogeneity of error

The use of CV-optimised thresholds resulted in the selection of higher thresholds by a mean of 0.08 ± 0.05 , with no thresholds decreasing, the threshold for *N. albicollis* stayed the same and a maximum increase of 0.13 for *A. sericocaudatus* (Figure 3). Importantly, CV-optimised thresholds reduced estimations of precision CV considerably, by an average of 0.2 ± 0.16 per species, with a maximum decrease of 0.43 for *A. sericocaudatus*, 0.39 for *N. ocellatus* and 0.2 for *M. usta*. A Levene's test showed a significant reduction in the variance of precision at the 29 survey points for *A. sericocaudatus* ($F = 9.58$, $p = 0.01$; Figure 4). The application of CV-optimised thresholds also resulted in increases in average precision to a mean of 0.86 ± 0.09 from 0.83 ± 0.13 , with a precision estimate of 0.97 for *A. sericocaudatus*, up from 0.88. This did come at some considerable cost to recall, with the mean decreasing from 0.34 ± 0.13 to 0.27 ± 0.09 .

The contextual classification reduced heterogeneity in precision across the 29 survey points (Figure 4). The variance in precision was less than the variance for the Tadarida classifier with $F^{0.5}$ -score optimised thresholds for all species, and significantly so (adjusted p -values ≤ 0.05) for all species except *L. cristata*. Variance in precision also reduced in comparison to the Tadarida classification with

CV-optimised thresholds for all species except *L. cristata*, which had a higher variance with the contextual classification by 0.001. Both *G. hardyi* and *N. albicollis* had significantly lower variances with the contextual classification than the Tadarida classification with CV-optimised thresholds. We found a significant decrease in the CV of precision scores, decreasing from a mean of 0.41 ± 0.25 (SD) with Tadarida classification and CV-optimised thresholds to 0.16 ± 0.11 (SD) with contextual classification (Wilcoxon Signed-Rank test with Benjamini-Hochberg correction, $F = 7$, adjusted p -value = 0.04; Figure 4).

The contextual classification also improved the overall precision of classification from a mean of 0.86 ± 0.09 (SD) with the Tadarida classifier and CV-optimised thresholds to 0.91 ± 0.05 (SD) (Figure 5). Although mean precision values across species did not differ significantly, contextual classification resulted in significantly higher precision scores than Tadarida with F -score optimised thresholds for all species (adjusted p -values ≤ 0.05) when precision scores are calculated at each site (Figure 4), and Tadarida with CV-optimised thresholds for *L. cristata*, *G. hardyi* and *N. albicollis*.

Additionally, we found significant improvements in corrected recall and $F^{0.5}$ -score with the contextual classifier compared to the Tadarida classifier with CV-optimised thresholds (Figure 5). $F^{0.5}$ -score also significantly improved, from 0.59 ± 0.11 to 0.75 ± 0.09

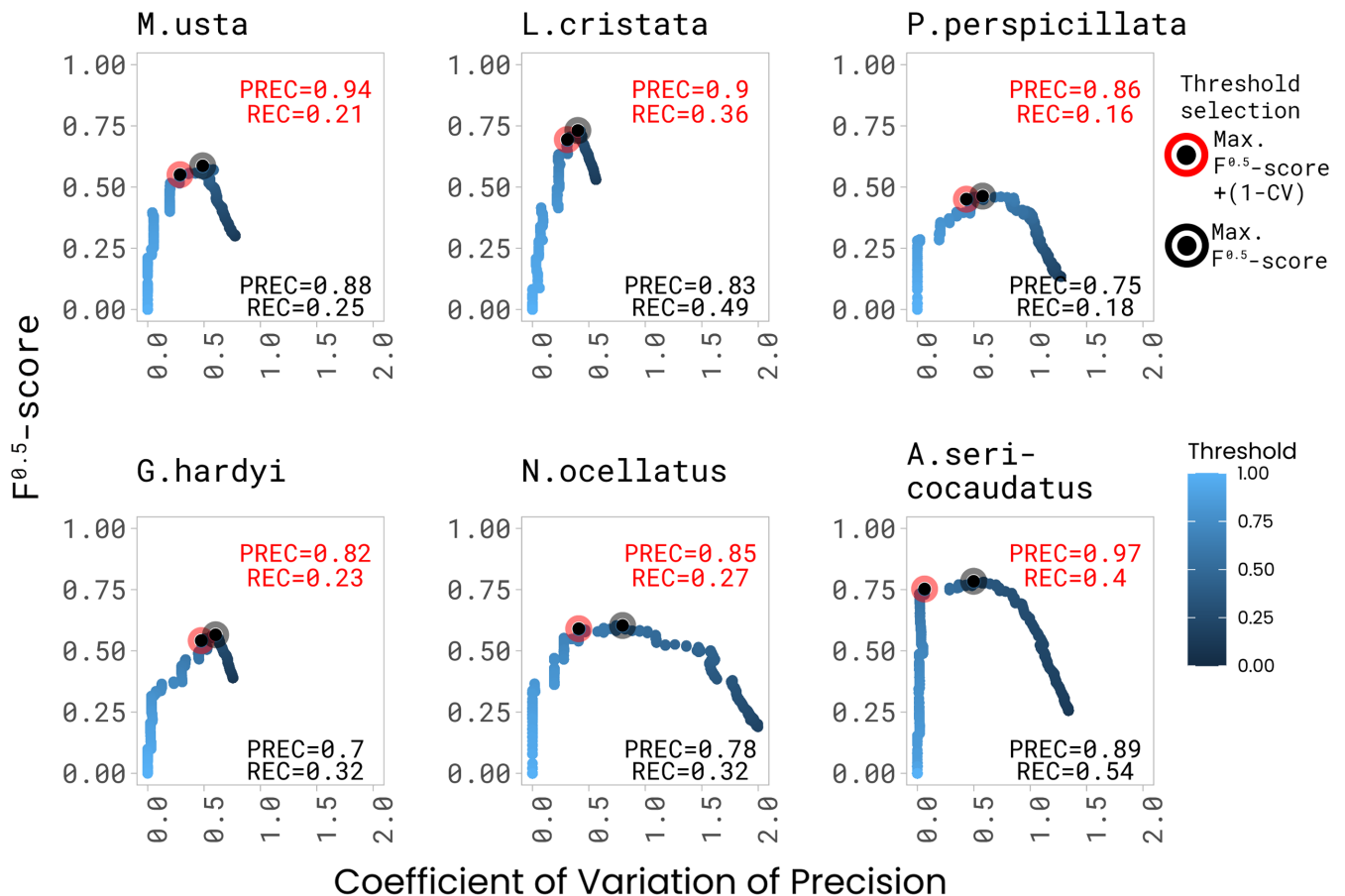


FIGURE 3 Threshold selection approaches with (red circles) and without (black circles) including the coefficient of variation (CV) of precision. Scores shown within each plot are precision (PREC) and recall (REC) for the CV-optimised threshold selection based on Test Set 2. *Nyctidromus albicollis* is not shown, but both threshold approaches selected the same threshold score giving a precision score of 0.7 and recall of 0.25.

($F = 43$, adjusted p -value = 0.04). Another benefit of the contextual classifier was a considerable increase in corrected recall, from 0.27 ± 0.09 to 0.46 ± 0.13 ($F = 44$, adjusted p -value = 0.04). All species showed higher corrected recall with contextual classification than Tadarida with CV-optimised thresholds, with a maximum increase of 0.28 observed for *N. ocellatus*.

4 | DISCUSSION

Heterogeneity of error is rarely tested for in ecoacoustic studies using automated classification (Wright et al., 2020) and could potentially confound ecological interpretation. We found strong evidence that classification of sound events with a random forest-based Tadarida algorithm exhibited strong indications of heterogeneous false-positive error. Six of seven target species had precision ranging from zero to one across 29 survey sites, and mean CV across all species was greater than 0.5 regardless of the test set used to estimate accuracy metrics. This highlights the need for accuracy metrics that better reflect the performance of machine learning classification under field conditions, and that do

not rely solely on those metrics optimised for machine learning use in 'laboratory conditions' (Wearn et al., 2019). In particular, we emphasise the need to include metrics to detect error variance across prediction variables, here successfully undertaken using precision CV.

Additionally, we found that using post-classification sampling of files from only positive classifications can provide a more reliable estimate of precision error heterogeneity, while providing better estimations of standard accuracy metrics. In comparison, standard approaches to drawing test sets (Knight et al., 2017) using stratified pre-classification sampling, resulted in both the largest and smallest estimates for each of Precision, Recall and $F^{0.5}$ -score. This is probably due to a low sample size of positive predictions when species have low call density, resulting in a less representative test set. However, the additional subsampling required for creating a post-classification test set may only be appropriate if classification is being conducted on large datasets, for species with very low call densities, or if those conducting the study have strong a priori reasons for expecting a high level of heterogeneity within error rates. In other circumstances, ensuring a large enough pre-classification sample to obtain a good number of positive predictions may suffice.

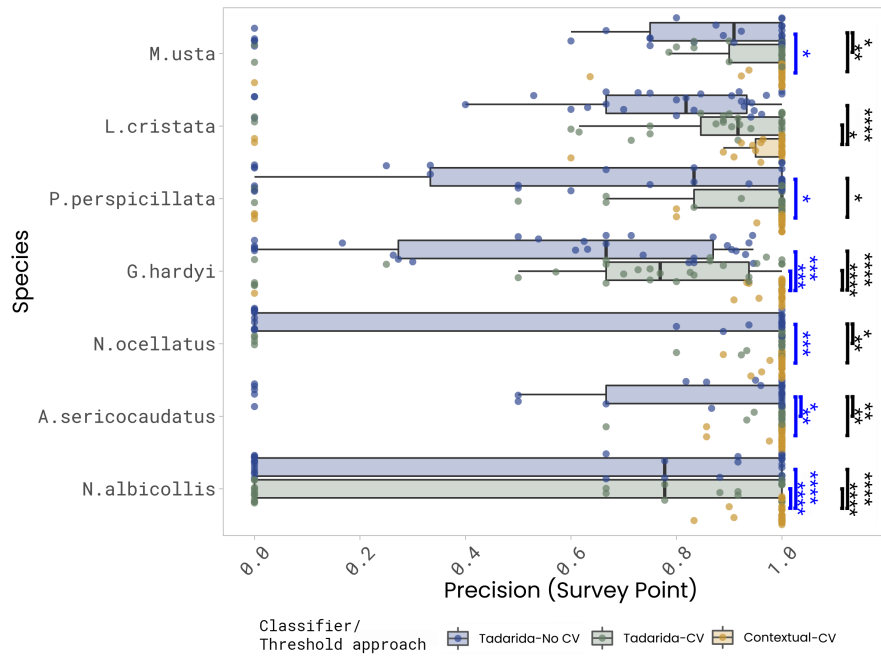


FIGURE 4 A comparison of the variance in precision at 29 survey points by classification method and threshold optimisation approach. Blue brackets and stars show the significant results of pairwise Levene's tests with Benjamini-Hochberg correction on the homogeneity of variance. Black brackets and stars show the significant results of pairwise Wilcoxon signed-rank tests with Benjamini-Hochberg correction on precision values. * $p < 0.05$ and > 0.01 , ** $p < 0.01$ and > 0.001 , *** $p < 0.001$ and > 0.0001 , **** $p < 0.0001$. Yellow boxplots show results for the contextual classifier with a CV-optimised threshold, green for the Tadarida classifier alone with a CV-optimised threshold and blue results for the Tadarida classifier with an $F^{0.5}$ -score optimised threshold.

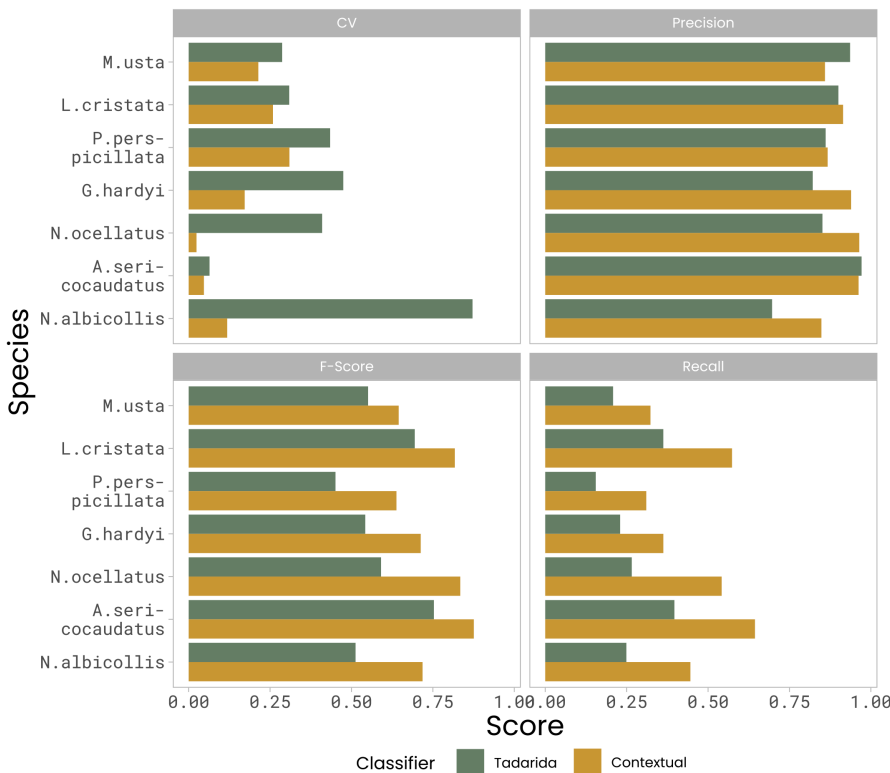


FIGURE 5 A comparison of accuracy metrics for Tadarida classification with CV-optimised thresholds applied, and contextual classification with CV-optimised thresholds applied. $F^{0.5}$ -score is weighted twice as heavily in favour of precision and recall is corrected based on the recall values from Test Set 1.

We also demonstrate two methods of reducing precision error heterogeneity. First, by incorporating a measure of heterogeneity of error in the thresholding process. This is straightforward to implement and resulted in substantial reductions of CV by 0.2 ± 0.16

per species. However, Levene's tests only found *A. sericocaudatus* to have a significantly lower precision variance than scores derived with a threshold optimised for $F^{0.5}$ -score alone. In addition, this method reduces error heterogeneity by increasing the confidence

score threshold, thus also requiring a greater reduction in recall. This suggests that incorporating heterogeneity of error in the thresholding process should only be used on its own for classifiers that already have high traditional performance metrics, in particular a high *F*-score and PR-AUC, requiring a limited reduction in precision error heterogeneity. This is further emphasised by the incorporation of CV into threshold selection not resulting in an increase in threshold for *N. albicollis* despite this species showing the highest precision error heterogeneity, because the classifier performed poorly enough in general that it did not justify the decrease in $F^{0.5}$ -scores.

The second method to reduce precision error heterogeneity was a secondary contextual classifier. In contrast, this required considerably more effort to incorporate into a classification workflow, but substantially reduced CV of precision for all species, and significantly reduced the variance of precision scores for all species except *L. cristata*. In keeping with previous studies that also used a secondary classifier (Balantic & Donovan, 2020), it also improved overall classification performance, significantly improving precision for all target species and simultaneously increasing recall in comparison to only using a CV-optimised threshold. This suggests that users of all but the best performing classification models, and those with particular concerns about error heterogeneity confounding ecological results, should consider using an additional contextual classification to redress variance in precision.

Concern around precision error heterogeneity led us to train the contextual classifier on positive predictions. This is because we believe variance in precision is more likely to confound ecological findings and be harder to mitigate against in other ways, such as summarising results over longer time periods, or the use of occupancy models, than heterogeneity of recall. However, for some uses of audio classification, homogeneity of recall may be as, or more, important—and previous research has suggested that a minimum level of recall is required for studies to be reliable (Knight et al., 2020), for instance recall can impact detectability in occupancy models, which could bias occupancy estimates (MacKenzie et al., 2017). In these cases, the use of a contextual classifier that allows for a lower threshold and increased recall is clearly beneficial but may be improved further by also training it on negative predictions. However, this does entail a higher degree of effort to implement, especially in cases of low call density due to the class imbalances inherent in detecting bird calls. In such cases, there are always likely to be many more true negatives than either false negatives or true positives, so finding sufficient instances of false negatives to train a secondary classifier may prove challenging without a priori knowledge of when and where positive instances are likely to occur.

There is no reason to think heterogeneity of error is unique to this dataset, or even to random forest classifications. It is likely to occur broadly in supervised learning classification methods, with parallels to image classification in camera-trapping (Wearn et al., 2019), including convolutional neural networks currently producing the best classification accuracy. Covariate shift from training datasets can be caused by underlying ecological factors impacting the soundscape varying across a range of gradients including space, time, light and

temperature (Royle, 2018; Yip et al., 2017). We therefore strongly recommend that future studies explicitly test for, and take measures to reduce variance in error.

The methods proposed here are just two possibilities, and are not necessarily optimal. Consideration should be afforded to study objectives and the ecology of the target species—for example the use of contextual classification here could bias against rarer species or events, so may only be appropriate for species with regular calling bouts, and for occupancy or abundance estimations rather than presence/absence of extremely rare species. Other approaches could instead be used to reduce heterogeneity of error, for instance with a suitably large initial training set, a naive classifier could be trained at short time-scales and then the naive classification scores used as features in a contextual classifier at longer time-scales with bootstrapping to maintain independence. This approach would remove the need to generate a second training dataset, although would probably make the collection of a suitable initial dataset more challenging. Other machine learning methods, or even ensembles, could outperform Distributed Random Forest algorithms, and other useful contextual variables could be used—for instance acoustic indices to better characterise and contextualise the soundscape. For those wishing to work within an occupancy model framework, it would be useful to undertake further research to compare the efficacy of methods to resolve precision and recall error within the occupancy model (e.g. Rempel et al., 2019) to reduce error heterogeneity during the classification process itself. Nonetheless, by revealing for the first time the potential importance and a solution for dealing with error heterogeneity we hope to stimulate further research, and to encourage those who use machine learning classification in ecoacoustics to carefully consider the implications of classification error on ecological inference.

AUTHOR CONTRIBUTIONS

All authors made substantial contributions. Oliver C. Metcalf, Jos Barlow, Yves Bas and Alexander C. Lees made substantial contributions to conception and design. Oliver C. Metcalf, Christian Devenish and Yves Bas contributed to code development and testing. Oliver C. Metcalf, Erika Berenguer and Filipe França contributed to acquisition of data. All authors contributed to analysis and interpretation of data, and drafting the article. All authors had final approval of the version to be published, agree to be accountable for the aspects of the work that they conducted and ensuring that questions related to the accuracy or integrity of any part of their work are appropriately investigated and resolved.

ACKNOWLEDGEMENTS

We would like to thank the RAS field and laboratory assistants: Marcos Oliveira, Gilson Oliveira, Renilson Freitas and Josué Jesus de Oliveira for their hard work and assistance, without whom this would not be possible. We are also grateful to Joice Ferreira and Liana Chessini Rossi for logistical field support in Brazil. Additional thanks to the Cornell Lab of Ornithology, particularly Matthew Medler and Jay McGowan for advising and helping with data from the Macaulay Library. Fieldwork in Brazil and later analysis was

supported by research grants ECOFOR (NE/K016431/1), and AFIRE (NE/P004512/1), PELD-RAS (CNPq/CAPES/PELD 441659/2016-0) and the BNP Paribas Foundation's Climate and Biodiversity Initiative (Project Bioclimate).

CONFLICT OF INTERESTS

The authors have no conflicts of interest to declare.

PEER REVIEW

The peer review history for this article is available at <https://publons.com/publon/10.1111/2041-210X.13967>.

DATA AVAILABILITY STATEMENT

All training and test datasets used for this article are available on Dryad (Metcalf et al., 2022) <https://doi.org/10.5061/dryad.69p8cz94j>.

ORCID

Oliver C. Metcalf  <https://orcid.org/0000-0003-3441-2591>

Jos Barlow  <https://orcid.org/0000-0003-4992-2594>

Erika Berenguer  <https://orcid.org/0000-0001-8157-8792>

Christian Devenish  <https://orcid.org/0000-0002-5249-0844>

Filipe França  <https://orcid.org/0000-0003-3827-1917>

Stuart Marsden  <https://orcid.org/0000-0002-0205-960X>

Charlotte Smith  <https://orcid.org/0000-0002-3767-6587>

Alexander C. Lees  <http://orcid.org/0000-0001-7603-9081>

REFERENCES

- Balantic, C. M., & Donovan, T. M. (2020). Statistical learning mitigation of false positives from template-detected data in automated acoustic wildlife monitoring. *Bioacoustics*, 29(3), 296–321. <https://doi.org/10.1080/09524622.2019.1605309>
- Barré, K., Le Viol, I., Julliard, R., Pauwels, J., Julien, J. F., Claireau, F., Kerbiriou, C., & Bas, Y. (2019). Accounting for automated identification errors in acoustic surveys. *Methods in Ecology and Evolution*, 10(8), 1171–1188. <https://doi.org/10.1111/2041-210X.13198>
- Bas, Y., Bas, D., & Julien, J.-F. (2017). Tadarida: A toolbox for animal detection on acoustic recordings. *Journal of Open Research Software*, 5(1), 6. <https://doi.org/10.5334/jors.154>
- Center for Conservation Bioacoustics. (2019). *Raven pro: Interactive sound analysis software (version 1.4) [computer software]*. The Cornell Lab of Ornithology. Retrieved from <http://www.birds.cornell.edu/raven>
- Chambert, T., Grant, E. H. C., Miller, D. A. W., Nichols, J. D., Mulder, K. P., & Brand, A. B. (2018). Two-species occupancy modelling accounting for species misidentification and non-detection. *Methods in Ecology and Evolution*, 9(6), 1468–1477. <https://doi.org/10.1111/2041-210X.12985>
- Clare, J. D. J., Townsend, P. A., & Zuckerberg, B. (2021). Generalized model-based solutions to false-positive error in species detection/nondetection data. *Ecology*, 102(2) e03241. <https://doi.org/10.1002/ecy.3241>
- Darras, K., Batáry, P., Furnas, B., Celis-Murillo, A., Van Wilgenburg, S. L., Mulyani, Y. A., & Tschirntke, T. (2018). Comparing the sampling performance of sound recorders versus point counts in bird surveys: A meta-analysis. *Journal of Applied Ecology*, 55(6), 2575–2586. <https://doi.org/10.1111/1365-2664.13229>
- Frontier Labs. (2015). *Bioacoustic Audio Recorder User Guide*. Retrieved from http://www.frontierlabs.com.au/web_documents/BARUserGuidev1.3.pdf
- Gibb, R., Browning, E., Glover-Kapfer, P., & Jones, K. E. (2019). Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. *Methods in Ecology and Evolution*. *British Ecological Society*, 10, 169–185. <https://doi.org/10.1111/2041-210X.13101>
- Juodakis, J., Castro, I., & Stephen, M. (2021). Precision as a metric for acoustic survey design using occupancy or spatial capture-recapture. *Environmental and Ecological Statistics*, 28, 587–608. <https://doi.org/10.1007/s10651-021-00513-4>
- Kepler, C. B., & Scott, J. M. (1981). Reducing bird count variability by training observers. *Studies in Avian Biology*, 6, 366–371.
- Knight, E. C., Eter, P., Olymos, S., Scott, C., & Bayne, E. M. (2020). Validation prediction: A flexible protocol to increase efficiency of automated acoustic processing for wildlife research. <https://doi.org/10.1002/eap.2140>
- Knight, E. C., Hannah, K. C., Foley, G. J., Scott, C. D., Brigham, R. M., & Bayne, E. (2017). Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs. *Avian Conservation and Ecology*, 12(2), art14. <https://doi.org/10.5751/ACE-01114-120214>
- Knight, E. C., Poo Hernandez, S., Bayne, E. M., Bulitko, V., & Tucker, B. V. (2019). Pre-processing spectrogram parameters improve the accuracy of bioacoustic classification using convolutional neural networks. *Bioacoustics*, 29, 1–19. <https://doi.org/10.1080/09524622.2019.1606734>
- Kuhn, M. (2021). caret: Classification and regression training. Retrieved from <https://cran.r-project.org/package=caret>
- Lapp, Rhinehart, Freeland-Haynes, Khilnani, & Kitzes. (2022). "OpenSoundscape v0.7.0". Retrieved from <https://github.com/kitzeslab/opensoundscape/blob/master/README.md>
- LeBien, J., Zhong, M., Campos-Cerqueira, M., Velev, J. P., Dodhia, R., Ferres, J. L., & Aide, T. M. (2020). A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network. *Ecological Informatics*, 59, 101113. <https://doi.org/10.1016/j.ecoinf.2020.101113>
- LeDell, E., Gill, N., Aiello, S., Fu, A., Candel, A., Click, C., Kraljevic, T., Nykodym, T., Aboyoun, P., Kurka, M., & Malohlava, M. (2020). *h2o: R Interface for the 'H2O' Scalable Machine Learning Platform*. R package version 3.28.0.4. Retrieved from <https://CRAN.R-project.org/package=h2o>
- Lees, A. C., de Moura, N. G., Andretti, C. B., Davis, B. J. W., Lopes, E. V., Pinto Henriques, L. M., Aleixo, A., Barlow, J., Ferreira, J., & Gardner, T. A. (2013). One hundred and thirty-five years of avifaunal surveys around Santarém, central Brazilian Amazon. *Revista Brasileira de Ornitologia*, 21(1), 16–57. Retrieved from <https://www.alice.cnptia.embrapa.br/handle/doc/976067>
- MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L. L., & Hines, J. E. (2017). *Occupancy estimation and modeling: Inferring patterns and dynamics of species occurrence: Second edition*. Elsevier. <https://doi.org/10.1016/C2012-0-01164-7>
- Mennill, D. J., & Vehrencamp, S. L. (2008). Context-dependent functions of avian duets revealed by microphone-Array recordings and multispeaker playback. *Current Biology*, 18(17), 1314–1319. <https://doi.org/10.1016/J.CUB.2008.07.073>
- Metcalf, O., Barlow, J., Bas, Y., Berenguer, E., Devenish, C., França, F., Marsden, S., Smith, C., & Lees, A. (2022). Data from: Detecting and reducing heterogeneity of error in acoustic classification. *Methods in Ecology and Evolution*. <https://doi.org/10.5061/dryad.69p8cz94j>
- Metcalf, O. C., Barlow, J., Marsden, S., Gomes de Moura, N., Berenguer, E., Ferreira, J., & Lees, A. C. (2021). Optimizing tropical forest bird surveys using passive acoustic monitoring and high temporal resolution sampling. *Remote Sensing in Ecology and Conservation*, 8, 45–56. <https://doi.org/10.1002/rse2.227>
- Metcalf, O. C., Ewen, J. G., McCready, M., Williams, E. M., & Rowcliffe, J. M. (2019). A novel method for using ecoacoustics to

- monitor post-translocation behaviour in an endangered passerine. *Methods in Ecology and Evolution*, 10(5), 626–636. <https://doi.org/10.1111/2041-210X.13147>
- Metcalfe, O. C., Lees, A. C., Barlow, J., Marsden, S. J., & Devenish, C. (2020). hardRain: An R package for quick, automated rainfall detection in ecoacoustic datasets using a threshold-based approach. *Ecological Indicators*, 109, 105793. <https://doi.org/10.1016/j.ecoli.2019.105793>
- Millard SP (2013) *EnvStats: An R package for environmental statistics*. Springer. ISBN 978-1-4614-8455-4, <<https://www.springer.com>>
- Newson, S. E., Bas, Y., Murray, A., & Gillings, S. (2017). Potential for coupling the monitoring of bush-crickets with established large-scale acoustic monitoring of bats. *Methods in Ecology and Evolution*, 8, 1051–1062. <https://doi.org/10.1111/2041-210X.12720>
- Priyadarshani, N., Marsland, S., & Castro, I. (2018). Automated birdsong recognition in complex acoustic environments: A review. *Journal of Avian Biology*, 49, jav.01447. <https://doi.org/10.1111/jav.01447>
- Priyadarshani, N., Marsland, S., Juodakis, J., Castro, I., & Listanti, V. (2020). Wavelet filters for automated recognition of birdsong in long-time field recordings. *Methods in Ecology and Evolution*, 11(3), 403–417. <https://doi.org/10.1111/2041-210X.13357>
- R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
- Rempel, R. S., Jackson, J. M., Van Wilgenburg, S. L., & Rodgers, J. A. (2019). A multiple detection state occupancy model using autonomous recordings facilitates correction of false positive and false negative observation errors. *Avian Conservation and Ecology*, 14(2), 1. <https://doi.org/10.5751/ACE-01374-140201>
- Robinson, W. D., Lees, A. C., & Blake, J. G. (2018). Surveying tropical birds is much harder than you think: A primer of best practices. *Biotropica*, 50, 846–849. <https://doi.org/10.1111/btp.12608>
- Ross, J. C., & Allen, P. E. (2014). Random Forest for improved analysis efficiency in passive acoustic monitoring. *Ecological Informatics*, 21, 34–39. <https://doi.org/10.1016/j.ecoinf.2013.12.002>
- Royle, J. A. (2018). Modelling sound attenuation in heterogeneous environments for improved bioacoustic sampling of wildlife populations. *Methods in Ecology and Evolution*, 9(9), 1939–1947. <https://doi.org/10.1111/2041-210X.13040>
- Royle, J. A., & Link, W. A. (2006). Generalized site occupancy models allowing for false positive and false negative errors. *Ecology*, 87(4), 835–841. [https://doi.org/10.1890/0012-9658\(2006\)87\[835:GSOMAF\]2.0.CO;2](https://doi.org/10.1890/0012-9658(2006)87[835:GSOMAF]2.0.CO;2)
- Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2), 227–244. [https://doi.org/10.1016/S0378-3758\(00\)00115-4](https://doi.org/10.1016/S0378-3758(00)00115-4)
- Simons, T. R., Alldredge, M. W., Pollock, K. H., & Wettröth, J. M. (2007). Experimental analysis of the auditory detection process on avian point counts. *The Auk*, 124(3), 986–999. <https://doi.org/10.1093/auk/124.3.986>
- Stowell, D., Petrusková, T., Šálek, M., & Linhart, P. (2019). Automatic acoustic identification of individuals in multiple species: Improving identification across recording conditions. *Journal of the Royal Society Interface*, 16(153). <https://doi.org/10.1098/rsif.2018.0940>
- Stowell, D., Wood, M. D., Pamuča, H., Stylianou, Y., & Glotin, H. (2019). Automatic acoustic detection of birds through deep learning: The first bird audio detection challenge. *Methods in Ecology and Evolution*, 10(3), 368–380. <https://doi.org/10.1111/2041-210X.13103>
- Sueur, J., Aubin, T., & Simonis, C. (2008). Seewave: A free modular tool for sound analysis and synthesis. *Bioacoustics*, 18, 213–226.
- Thieurmel, B., & Elmarhraoui, A. (2019). Suncalc: Compute sun position, Sunlight Phases, Moon Position and Lunar Phase. Retrieved from <https://cran.r-project.org/package=suncalc>
- Tobias, J. A., Planqué, R., Cram, D. L., & Seddon, A. N. (2014). Species interactions and the structure of complex communication networks. *Proceedings of the National Academy of Sciences of the United States of America*, 111(3), 1020–1025. <https://doi.org/10.1073/pnas.1314337111>
- Towsey, M., Planitz, B., Nantes, A., Wimmer, J., & Roe, P. (2012). A toolbox for animal call recognition. *Bioacoustics*, 21(2), 107–125. <https://doi.org/10.1080/09524622.2011.648753>
- Wearn, O. R., Freeman, R., & Jacoby, D. M. P. (2019). Responsible AI for conservation. *Nature Machine Intelligence*, 1, 72–73. <https://doi.org/10.1038/s42256-019-0022-7>
- Wright, W. J., Irvine, K. M., Almborg, E. S., & Litt, A. R. (2020). Modelling misclassification in multi-species acoustic data when estimating occupancy and relative activity. *Methods in Ecology and Evolution*, 11(1), 71–81. <https://doi.org/10.1111/2041-210X.13315>
- Yip, D. A., Bayne, E. M., Sólmos, P., Campbell, J., & Proppe, D. (2017). Sound attenuation in forest and roadside environments: Implications for avian point-count surveys. *The Condor*, 119(1), 73–84. <https://doi.org/10.1650/condor-16-93.1>
- Zhong, M., LeBien, J., Campos-Cerqueira, M., Dodhia, R., Lavista Ferres, J., Velev, J. P., & Aide, T. M. (2020). Multispecies bioacoustic classification using transfer learning of deep convolutional neural networks with pseudo-labeling. *Applied Acoustics*, 166, 107375. <https://doi.org/10.1016/j.apacoust.2020.107375>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Metcalf, O. C., Barlow, J., Bas, Y., Berenguer, E., Devenish, C., França, F., Marsden, S., Smith, C., & Lees, A. C. (2022). Detecting and reducing heterogeneity of error in acoustic classification. *Methods in Ecology and Evolution*, 13, 2559–2571. <https://doi.org/10.1111/2041-210X.13967>