Article:

Lewis, T.C. orcid.org/0000-0001-8787-3581, Gutierrez Vargas, I., Beckerman, A.P. orcid.org/0000-0002-4797-9143 et al. (1 more author) (2024) An inadequate sampling of the soundscape leads to over-optimistic estimates of recogniser performance: a case study of two sympatric macaw species. Bioacoustics, 33 (4). pp. 311-331. ISSN 0952-4622

https://doi.org/10.1080/09524622.2024.2353647

Bioacoustics

The International Journal of Animal Sound and its Recording

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tbio20

An inadequate sampling of the soundscape leads to over-optimistic estimates of recogniser performance: a case study of two sympatric macaw species

Thomas C. Lewis, Ignacio Gutierrez Vargas, Andrew P. Beckerman & Dylan Z. Childs

To cite this article: Thomas C. Lewis, Ignacio Gutierrez Vargas, Andrew P. Beckerman & Dylan Z. Childs (2024) An inadequate sampling of the soundscape leads to over-optimistic estimates of recogniser performance: a case study of two sympatric macaw species, Bioacoustics, 33:4, 311-331, DOI: 10.1080/09524622.2024.2353647

To link to this article: https://doi.org/10.1080/09524622.2024.2353647

Published online: 02 Jul 2024.

BIOACOUSTICS
2024, VOL. 33, NO. 4, 311–331
https://doi.org/10.1080/09524622.2024.2353647

Taylor & Francis
Taylor & Francis Group

An inadequate sampling of the soundscape leads to over-optimistic estimates of recogniser performance: a case study of two sympatric macaw species

Thomas C. Lewis[a,b], Ignacio Gutierrez Vargas[c], Andrew P. Beckerman[a] and Dylan Z. Childs[a]

[a]Department of Biosciences, University of Sheffield, Sheffield, UK; [b]Bird Programme, Macaw Recovery Network, Guanacaste, Costa Rica; [c]Programa de Posgrado en Bioloǵ ıa, Sistema de Estudios de Posgrado, Universidad de Costa Rica, San Pedro, San José, Costa Rica

ABSTRACT

Passive acoustic monitoring (PAM) – autonomously recording ambient sound – could dramatically increase the scale and robustness of species monitoring in rainforest ecosystems. PAM generates large volumes of data that require automated methods of target species detection. Species-specific recognisers, which often use supervised machine learning, can achieve this goal. However, they require a large training dataset of target and non-target signals, which is time-consuming and challenging to create. Unfortunately, very little information about creating training datasets for supervised machine learning recognisers is available, especially for tropical ecosystems. Here, we show an iterative approach to creating a training dataset that improved recogniser precision from 0.12 to 0.55. By sampling background noise using an initial small recogniser, we can address one of the significant challenges of training dataset creation in acoustically diverse environments. Our work demonstrates that recognisers will likely fail in real-world settings unless the training dataset size is large enough and sufficiently representative of the ambient soundscape. We outline a workflow that can provide users with an accessible way to create a species-specific PAM recogniser that addresses these issues for tropical rainforest environments. Our work provides important lessons for PAM practitioners wanting to develop species-specific recognisers for acoustically diverse ecosystems.

ARTICLE HISTORY
Received 12 January 2023
Accepted 2 April 2024

KEYWORDS
Passive acoustic monitoring; machine learning; training dataset; iteration; species-specific

Introduction

Effective monitoring of wildlife populations is required to mitigate rapid and widespread environmental change (Gibbs et al. 1999; Pereira and David Cooper 2006; Nichols et al. 2015). Long-term, standardised monitoring provides data on the presence or abundance of target species, which is necessary to identify the factors affecting population growth, abundance, and persistence (Pollock et al. 2002; Fedy and Aldridge 2011; Nuttall et al. 2022). These data can be challenging to acquire for

CONTACT Thomas C. Lewis tomclewis29@gmail.com

a wide-ranging species of conservation concern, as they are often found at low density in inaccessible environments such as tropical forests or marine ecosystems (Barnes 2001; Guschanski et al. 2009; Dénes et al. 2018). The challenge is to collect sufficient volumes of ecologically relevant data at appropriate spatial scales. However, traditional survey methods are not well suited to meet this challenge because they are often impractical and labour intensive.

Passive Acoustic Monitoring (PAM) has emerged as a cost-effective method to address this challenge. It is one of several advances in high-throughput sensing technologies, such as remote sensing, LIDAR, and camera traps, that can scale up data collection while maintaining or minimising work effort on the ground (Gibb et al. 2019). PAM is a rapidly expanding field, benefitting from developments and cost reductions tied to hardware such as Automated Recording Units (Snaddon et al. 2013; Hill et al. 2019; Teixeira et al. 2019). These developments dramatically increase the quality and quantity of ecological data collected (Gibb et al. 2019). Automated Recording Units provide an efficient and non-invasive data collection platform to inform a wide variety of ecological metrics, including community composition (Pillay et al. 2019; Bradfer-Lawrence et al. 2020), abundance (Marques et al. 2013; Pérez-Granados et al. 2019), occupancy (Wood et al. 2019) and individual breeding biology (Marin-Cudraz et al. 2019).

Like any technology, PAM engenders several practical challenges. Critically, PAM requires expertise in post-collection data processing to collect meaningful information, such as detections of target species from raw audio files. For example, users can extract data manually by labelling target signals (Campos-Cerqueira et al. 2016; Abrahams and Geary 2020), though this approach is time-consuming and requires expert knowledge. Due to the large quantities of data produced, there is increasing interest in developing machine learning classifiers to automate target species identification from raw audio files. In addition, deep-learning classifiers can be highly accurate in bioacoustic tasks, though they require a large amount of data to train (Bermant et al. 2019; Stowell et al. 2019; Zhong et al. 2020). Such classifiers are increasingly used to track individual species in the wild, for example, sperm whales (*Physeter macrocephalus*; Bermant et al. 2019) and the Northern grey gibbon (*Hylobates funereus*; Clink et al. 2020). More generalised deep-learning classifiers to identify multiple species have also been developed, though currently, most are insufficient to classify all species of interest in many ecological applications (Ventura et al. 2015; Stowell et al. 2019; Zhong et al. 2020). BirdNet was the first generalised bioacoustic classifier for avian species. This deep artificial neural network (DNN) can identify over 3000 North American and European bird species (Kahl et al. 2021) with a mean recall of 92.2% when a broadcaster is within 50 m (Pérez-Granados 2023)

Developing effective classifiers for use with PAM can be time-consuming and prohibitively complex for non-expert users (Gibb et al. 2019). Proprietary software such as Kaleidoscope (Wildlife Acoustics, USA) and cloud-based platforms such as Arbimon (Aide et al. 2013; Bravo et al. 2017) offer a potential solution. However, such tools are often limited in their options for species identification tasks and, in some cases, incur a prohibitive cost for many conservation projects. Thus, where relevant off-the-shelf classifiers or suitable platforms are unavailable, custom classifiers must still be developed for individual applications on a case-by-case basis. These domain-specific classifiers are often created using supervised machine learning methods. These have a lower technical barrier to entry than deep learning classifiers and can be more suited to smaller datasets.

In broad terms, constructing a recogniser involves two stages: detecting regions of interest (ROI) and classifying potential signals. The number and complexity of these steps within a pipeline will vary depending on the methodology used (Lasseck 2014; Sebastián-González et al. 2015; Knight et al. 2020). Region of Interest detection identifies potential target signals. One simple and accessible ROI technique is template matching, which can be used as both the ROI identification and classification (Katz et al. 2016b). Template matching involves using a measure such as spectral cross-correlation to assess the similarity between one or more reference call patterns and a set of unknown call patterns. Using template matching for classification relies on creating a sufficiently representative call library (Aide et al. 2013; Gibb et al. 2019) and is therefore sensitive to variation in signal structure, i.e. call type and background noise (Brandes 2008; Katz et al. 2016a). Combining template matching with machine learning methods reduces the false-positive rate of a classifier compared to using template matching alone (Balantic and Donovan 2020). In this use case, template matching extracts regions of interest that are then classified by a supervised machine learning algorithm. For example, suppose the template library sufficiently represents intra-specific call-type variation with an appropriate cross-correlation score threshold. In that case, template matching will improve the quality of data input into a supervised machine learning classifier. Several different supervised machine learning approaches exist; random forest (Brieman 2001) is one of the most widely used (Tachibana et al. 2014; Noda et al. 2016; Raghuram et al. 2016) and performs well at bioacoustic tasks (Weerasena et al. 2018; Ayala-Berdon et al. 2020; Smith-Vidaurre et al. 2020).

An essential part of developing a recogniser is selecting the size of the dataset used to train the supervised machine learning. There are few concrete guidelines on what constitutes an adequate training dataset size. Sebastián-González et al. (2015) tested the effect of training dataset size on accuracy. They found that a 50% reduction in dataset size resulted in a loss of less than 1% balanced accuracy metric (BAC), suggesting that their chosen supervised machine learning (a Support Vector Machine) copes well with small datasets ($n = 642$). However, it is not easy to generalise these results because classifier performance varies on a species-to-species basis. For example, Digby et al. (2013) achieved a recall of 39.8% and precision of 98.1% with 3411 little spotted kiwis (*Apteryx owenii*) calls and 3072 negative cases. In contrast, Sebastián-González et al. (2015) used a maximum of 1285 ʻAmakihi (*Hemignathus viren virens*) and 2785 negative cases and achieved a BAC of 86.5%. Similarly, when attempting to find the best training dataset size to optimise their relative sound level (RSL) method, Knight et al. (2020) found that the common nighthawk (*Chordeiles minor*) had an optimal training dataset size of 10,590 cases. In contrast, the Ovenbird (*Seiurus aurocapilla*) needed only 5,540 cases to achieve a similar performance.

A challenge in developing classifiers for PAM studies is the spatiotemporal variation in background noise. For example, various species vocalise concurrently in tropical rainforests in overlapping frequency ranges (Slabbekoorn 2004). The soundscape also varies as biotic factors, such as breeding season, and abiotic factors, such as weather change over time. Plant and animal community composition can also vary spatially across heterogeneous tropical landscapes (Cintra and

Naka 2011; Wardhaugh et al. 2014; Ioki et al. 2016) these factors make ecosystems like tropical rainforests challenging for PAM (e.g. Heinicke et al. 2015). One way to improve classifier performance, which may be especially important when not denoising, is to ensure the training dataset captures the variation in background noise. A sufficiently representative dataset is likely to be very large. Creating such a training dataset is challenging because it requires identifying the many sources of background noise that may be a problem for the classifier. Moreover, determining this kind of background noise would need an *a priori* knowledge that is not available.

Aims and objectives

We present a case study of two sympatric macaw species: the critically endangered Great Green Macaw *Ara ambiguus* (BirdLife International 2020) and the regionally endangered Scarlet Macaw *Ara macao* (Monge et al. 2016) in the tropical rainforest of northern Costa Rica. Parrots are one of the most endangered orders of birds, with 34% of species classified as threatened (IUCN 2024). They are widely distributed globally and native to every continent in the tropics. For many species of conservation concern, there is a lack of data on factors important to conservation planning and policy, such as their distribution and abundance. This limitation is especially concerning in the case of the Great Green Macaw because they were uplisted to critically endangered in 2020 (BirdLife International 2020). Parrots represent a significant challenge for classification tasks due to their wide vocabulary (Taylor and Perrin 2005; Zdenek et al. 2015; Montes-Medina et al. 2016) and context-dependent calls (Bradbury 2003). The two focal species present an additional challenge because their calls are highly similar and difficult to distinguish, even for experts (TL *pers. obs.*). The study region is also highly varied, from highly anthropogenically disturbed areas to mature primary forests. This presents a challenge of spatial variation in background noise.

Using this case study, we aim to demonstrate that not adequately sampling the spatiotemporal variation in background noise will lead to inaccurate estimations of recogniser performance. We then give one approach to developing a recogniser that can handle spatiotemporal variation in background noise.

Materials and methods

Macaw call characteristics

The calls of the two focal species of macaw are similar (Figure 1). They occupy the same frequency band (0.5 kHz to 4 kHz) and are very similar in length (0.2 sec to 1.0 sec). This frequency band overlaps significantly with that of anthropogenic noise, meaning that bandpass filtering can only offer minimal help to remove background noise (Slabbekoorn 2004; Slabbekoorn and Ripmeester 2008). The primary difference between the calls of the two species is the lack of clear harmonic structure of the Scarlet Macaw call, when compared to the Great Green Macaw call (Figure 1).

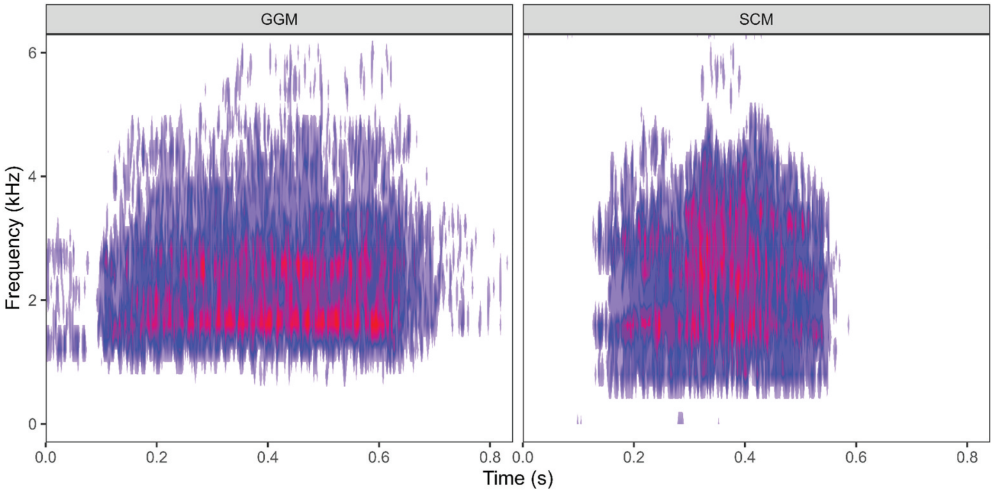

Figure 1. Spectrograms of a Great Green Macaw (GGM) and Scarlet Macaw (SCM) contact call. They are very similar apart from the blurring of the harmonics in the centre of the SCM call. This is the characteristic difference between the calls of the two species.

The audio

Survey recordings

We used 42 AudioMoth 1.2 (Hill et al. 2019; LabMaker, Germany) Automated Recording Units, which we installed on the tallest accessible trees (5–20 m) in a 10 km grid across the northeast of Costa Rica (Figure 2). They recorded four 30-min slots throughout the day (7:00–7:30, 10:00–10:30, 13:00–13:30, 16:00–16:30). Recording took place between 27 January 2020 and 30 June 2020. The standard sampling frequency was 48 kHz; however, the sampling frequency was not consistent as there was an error in configuring some devices for a portion of the deployment period. These devices sampled at 96 kHz rather than 48 kHz and therefore recordings were down sampled before they were used.

High-quality recordings

Template matching requires high-quality recordings of the target species to create reference templates that can be used for spectrogram cross-correlation. We did this with recordings made with a directional microphone, Sennheiser ME 67 (Sennheiser electronic GmbH & Co., United Kingdom) and a Roland *R*-05 Wave/MP3 digital recorder (Roland Corporation, United Kingdom). Recordings were digitised with a 16-bit sampling depth and 48 kHz sampling frequency; recordings were saved as WAV files. Recording occurred around known Great Green Macaw nest sites in northeastern Costa Rica between January and March 2019 (Figure 1). All recordings were of individuals within 200 m of the recorder. Initially, we selected four recordings containing a total of 157 calls from three different nest sites, as they included single individuals calling to their mate and groups of Great Green Macaw calling. This meant we could capture calls in different contexts (group and pair), increasing the call variation captured in our initial

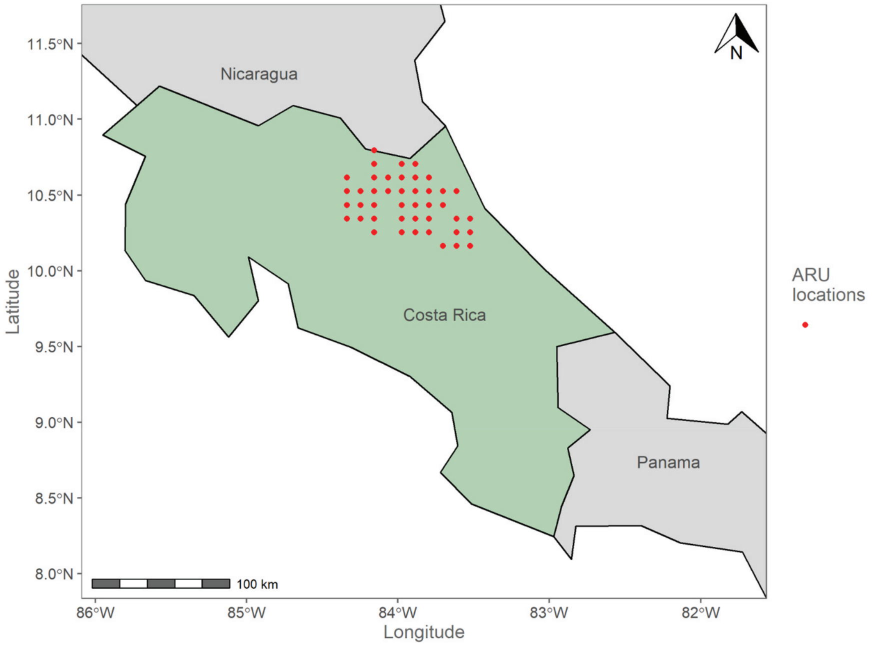

Figure 2. Automated recording unit locations in northeastern Costa Rica. Forty-two automated recording units were set out on a 10 km grid; holes in the grid are areas that were inaccessible due to dense forest or lack of permission from landowners.

dataset. These were selected by visually screening non-target signals that overlap with target calls.

We only used Great Green Macaw calls as high-quality recordings as we could not access Scarlet Macaw nest sites to record Scarlet Macaw calls. This is because within the study area, Scarlet Macaws are uncommon, and, therefore, their nest sites are infrequently discovered. We did not consider this an issue as the way we developed the templates ensured we were testing the templates against calls from both species.

Acoustic features

We extracted a total of 113 acoustic features using the warbleR package in R (R Core Team 2021): 20 measurements of frequency, time, and amplitude parameters, and 93 Mel-frequency cepstral coefficients (MFCCs) (Araya-Salas et al. 2017). MFCCs were initially designed for human speech recognition but have been widely used in bioacoustics (Loh et al. 2013; Colonna et al. 2016; Noda et al. 2016; Salamon et al. 2016). MFCCs reduce any signal to a set of coefficients (Colonna et al. 2016) while minimising any loss of biologically relevant information (Davis and Mermelstein 1980).

Recogniser workflow elements

Region of interest detection: template matching

The manually labelled calls from four high-quality recordings were used to construct templates ($n = 157$) following Hafner and Katz (2018) (Figure 3). We did not screen the

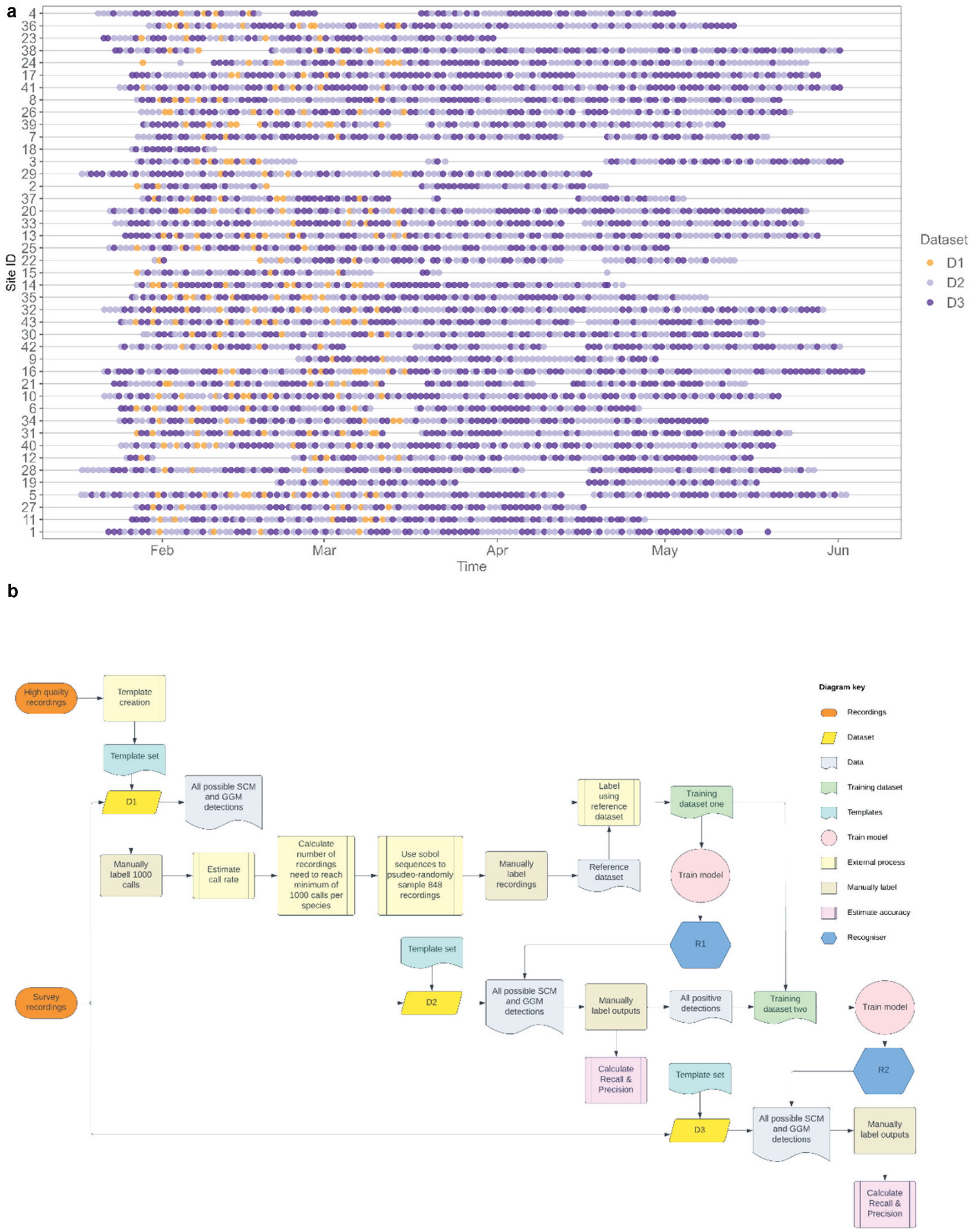

Figure 3. Schematic of classifier development. Panel (a) represents how all recordings were divided to create the three datasets. D1 was selected from recordings up to the middle of March when TL had to leave the field due to the pandemic. After April, recording devices were left in the field until June and so any failures were not rectified, which resulted in fewer recordings taking place towards the end of the study period. Most sites had issues with recording devices failing at some point; at site 43, the recording device failed and then was stolen, so it was not re-installed. (b) The workflow used to develop each classifier, starting from raw sound files (light green) and resulting in classifiers and performance metrics (red).

calls for background noise before converting them into templates. Once they had been converted into templates, we visually checked them and removed any if they contained any non-target signal ($n = 94$). With the remaining templates, we evaluated them on five 30 min, randomly selected Automated Recording Unit recordings that included Great Green Macaw and Scarlet Macaw calls. To sample randomly, we used dplyr:slice_sample() (Wickham et al. 2023) across the data frame containing target file names arranged by filename. Spectrogram cross-correlation is used to score the similarity of a detection to the template. We set a low default threshold (0.2) to allow flexibility in matching call types and amplitudes. Templates that detected less than 10% ($n = 28$) or over 200% ($n = 17$) of the total number of calls were removed. We removed templates with a true-positive-to-false-positive ratio above 1:5 or if 90% of a template's detections were identical to another's. In this case, to determine which template to keep, we manually screened the detections and kept the template that had the most unique detections. To test their accuracy, we ran the final group of templates ($n = 4$) over 10 randomly selected Automated Recording Unit recordings containing Scarlet Macaw and Great Green Macaw calls and 10 randomly selected recordings without any calls. We set the window length used in the template matching to 1 s to reflect the maximum call length (1 s).

Signal classification: supervised ML

We used a tidymodels workflow (Kuhn and Wickham 2020) in R (R Core Team 2021) to train each random forest algorithm. We used a 75:25 split of the training dataset to create the training and test data, meaning that 75% of the training dataset is used to train the classifier, and 25% is withheld to estimate performance. A random forest has three hyperparameters to be tuned: the number of trees, nodes, and variables per node. We set the number of trees to 1000 and used a tuning grid to tune the number of nodes (2–10 with five levels) and the number of variables per node (10–30 with five levels). Finally, we used k-fold cross-validation with 10 equal-size subsamples to estimate the area under the ROC curve, which was used to select the best value for the tuning parameters.

Recogniser one training dataset

To create the training dataset for recogniser one, we used recordings from between 27 January 2020 and 15 March 2020. We used recordings from this period because the global COVID-19 pandemic forced us out of the field and meant recogniser development started before all the data had been collected.

We chose an initial target of 1000 calls per species to create the dataset to match the smallest published training dataset size (Sebastián-González et al. 2015). To calculate how many recordings we would need to screen manually to reach 1000 calls per species, we first manually labelled 100 recordings to estimate the call rate per 30-min recording. The recordings were randomly selected using dplyr:slice_sample()(Wickham et al. 2023), as above. These were 3.13 Great Green Macaw and 1.18 Scarlet Macaw calls per recording. Therefore, to reach at least 1000 calls per species, we sampled 848 (1000/1.18) recordings using Sobol sequences to create a pseudo-random sample (Sobol 1967; Antonov and Saleev 1979) across sites and survey times (Figure 2). Sobol sequences are a type of quasi-random low-discrepancy sequence that are used in numerical methods for integrating and for generating pseudo-random samples (Renardy et al. 2021). Unlike traditional random sampling methods that use purely random numbers, Sobol sequences are designed to fill the space more uniformly (Sobol 1967; Antonov and Saleev 1979). They are also useful because they can handle dimensionality, and in this case we wanted to randomly sample across space and time. We manually labelled calls using Raven Lite (Bioacoustics Research Program 2016) to generate a reference dataset of known true cases: 2773 Great Green Macaw calls and 843 Scarlet Macaw calls.

We ran our template set over the 848 recordings and used the reference dataset to label all target cases; the remaining detections were labelled negative. We used Sobol sequences to create a pseudo-random sample (Sobol 1967; Antonov and Saleev 1979) of the negative cases equal to the total target cases of both classes (Great Green Macaw and Scarlet Macaw). Therefore, recogniser one's training dataset consisted of 4247 Great Green Macaw calls, 1393 Scarlet Macaw calls, and 5710 negative cases. The number of target cases was higher than that of manually labelled reference data because overlapping calls were only recorded once when manually labelling them. However, we designed the template matching detections not to overlap because the multiple templates could detect the same call. Therefore, these continuous calls would be detected as multiple cases rather than a single manually labelled case.

Build & validation datasets

When we received the second set of recordings from the field (16 March 2020 to 30 June 2020), we combined them with the recordings not used to create recogniser one's training dataset. We then divided these recordings into two sets, one that we used to evaluate the accuracy of the recogniser one and build a subsequent recogniser. We used the second to assess the accuracy of any other recognisers. To ensure an even spatial and temporal sampling of the recordings in each set, we again used Sobol sequences to create a pseudo-random sample (Sobol 1967; Antonov and Saleev 1979). This resulted in two equally sized datasets (build: $n = 8084$/validation: $n = 8084$ - Figure 2).

Recogniser performance

We used recall and precision as our performance metrics. Recall uses the number of false negatives (FN) to give a metric of how often a recogniser labels target signal as non-target signal (Recall = TP/TP + FN), and precision uses the number of false positives (FP) to create a metric for how often a recogniser labels non-target signal as target signal (Precision = TP/TP + FP).

We used two approaches to estimating the performance of the recognisers: 1) Evaluation: using 25% of the training data withheld from the model during training. 2) Validation: we ran the recogniser over the validation dataset and then manually checked all positive detections to count the number of true positives (TP) and false positives (FP). We did the same with a sample of the negative cases to count the number of true negatives (TN) and false negatives (FN) made by the recogniser. The second approach allowed us to get an accurate measure of performance.

As we expected the number of negative cases to be high and manually screening all of them would not be practical, we conducted a power analysis to determine how many negative cases we needed to check manually. For this, we needed the false-negative effect size:

$$h_{FN} = \frac{n_{FN}}{N} \tag{1}$$

Where n_{FN} is the number of false-negative cases, and N is the total number of all cases. We set the significance level to 0.05 and power to 0.95.

Sources of error

To understand how variation in background noise impacts performance, we labelled each false-positive case to species or genera level where possible. This meant we could investigate spatiotemporal variation in background noise and how this affects recogniser precision.

Recogniser two training dataset

We used all of the manually checked positive detections made by recogniser one. We combined these with the recogniser ones' training dataset to form the training dataset for recogniser two. Once trained, we applied the newly trained recogniser to the validation dataset and estimated performance in the same way as recogniser one.

Results

Region of interest detection: template matching

Across the whole dataset, the template set made 2,072,080 detections, a mean of 121.75 detections per recording ($n = 17020$). In the 848 recordings used to create the training dataset for recogniser one, the mean detection rate was 119.98 per recording. The mean number of Great Green Macaw calls found when we manually labelled the same dataset was 3.27 ($n = 848$), demonstrating that the template matching step captured a lot of non-target signals.

Recogniser one

The performance of recogniser one, estimated using the 25% withheld test data, was high in both recall (Great Green Macaw = 0.92/Scarlet Macaw = 0.85) and precision (Great Green Macaw = 0.93/Scarlet Macaw = 0.96 - Figure 4).

We applied it to the build dataset to more rigorously evaluate recogniser performance. Recogniser one made 37,639 positive detections (35814 Great Green Macaw and 1825 Scarlet Macaw), 4161 were true Great Green Macaw and 201 true Scarlet Macaw when we manually checked them. We checked 41,514 negative cases, determined by a power analysis using the Scarlet Macaw false-negative effect size, as it was the lowest of either target species (Great

Green Macaw = 0.0327, Scarlet Macaw = 0.0177). Finally, we reviewed all positive detections made by recogniser one (*n* = 37639) to estimate its performance. Compared to the test performance metrics, precision degraded (Great Green Macaw = 0.12/Scarlet Macaw = 0.11) significantly for both species when evaluated on the build dataset. Recall declined for the Scarlet Macaw (0.04), whereas it increased for the Great Green Macaw (0.98 - Figure 4).

Recogniser two

We combined the 4362 (4161 Great Green Macaw + 201 Scarlet Macaw) detections made by recogniser one in the build dataset with the 5640 (4247 Great Green Macaw + 1393 Scarlet Macaw) detections made by recogniser one's training dataset. Recogniser two was trained with these (8408 Great Green Macaw + 2112 Scarlet Macaw) and all non-target positive detections from recogniser one (*n* = 31999) plus the negative cases used to train recogniser one (*n* = 5710). Therefore, the final dataset was 8408 Great Green Macaw/ 2112 Scarlet Macaw/37709 negative.

Initial performance on the test data was lower than the performance of recogniser one: recall (Great Green Macaw = 0.69/Scarlet Macaw = 0.75) and precision (Great Green Macaw = 0.86/Scarlet Macaw = 0.78) (Figure 4). This is likely due to the size and spatiotemporal scale of the training dataset being larger and capturing more background noise, therefore being more representative. Hence, performance estimates are closer to reality.

We manually checked all positive (*n* = 6390) and 119,759 negative detections. Scarlet Macaw false-negative effect size was again used to calculate the number of negative detections to review, as it was the smallest (Great Green Macaw = 0.0498, Scarlet Macaw = 0.0104). Recall for the Great Green Macaw did not change significantly (0.66), whereas precision did drop (0.56), but it was not by as much as for recogniser one. Scarlet Macaw performance dropped for both metrics (Recall = 0.66/Precision = 0.24), but the decline was less severe than for recogniser one. The discrepancy between the two species' performance is likely due to the large difference in the training dataset size for each species.

Sources of error

The sources of error varied over time (Figure 5) and space (Figure 6). However, we can see a noticeable change in the dominant source of error after the period used to create the recogniser one (Figure 5b). This suggests that, even though performance is poor, the temporal variation in background noise is driving the worse performance later in the season (Figure 5a). Unidentified passerines stay a constant proportion of the error, whereas Amazona *spp*. decrease after this point, and clay-coloured thrush (*Turdus grayii*) and chickens (*Gallus domesticus*) increase. This is probably because the breeding season of the clay-coloured thrush begins after March, whereas the breeding season for Amazona *spp*. ends around this time.

Overall, false-positive detections from all sources were reduced by recogniser two compared to recogniser one, apart from the misclassification of Great Green Macaw as Scarlet Macaw (Figure 6). Recogniser one's false-positive detections at many sites are dominated by one or two sources of error (e.g. site 5, 1 and 29 - Figure 6). In contrast,

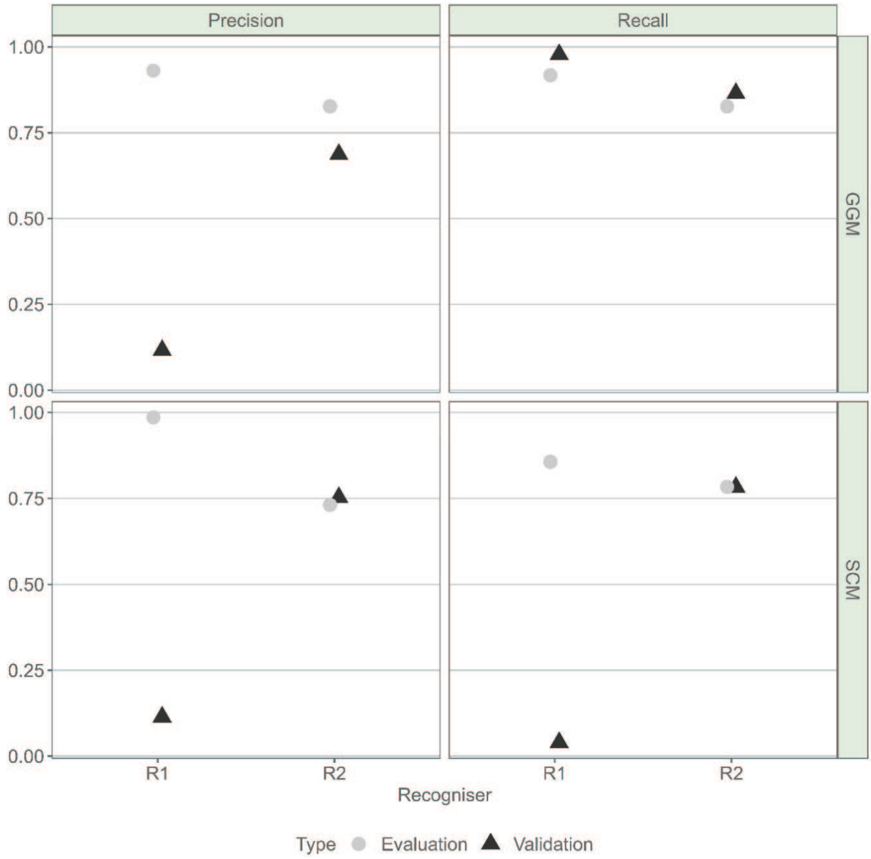

Figure 4. Classifier performance metrics show the improvement between R1 and R2. There is a trade-off between precision and recall: when precision increases, recall decreases and *visa-versa*. There is a constant large difference between evaluation and validation precision metrics, and this demonstrates the potential pitfalls of using default validation datasets to estimate the performance of classifiers.

there is a smaller number with a wide variety of sources of error (e.g. site 16, 20 and 25). However, there is a distinct spatial variation in the types of sources of error. For example, misclassifications are caused mainly by other wild bird species in the northern sites, whereas there is more anthropogenic noise (chainsaw, industry) in the south and east. Although a wild bird, the clay-coloured thrush is often strongly associated with areas of human disturbance (Dyrcz 1983), and our results support this association. Spatial variation in the source of error demonstrates that capturing this is critical in creating a representative training dataset.

Discussion

We have demonstrated that users must be careful when preparing a training dataset for a PAM recogniser. A training dataset that captures relevant background noise variation across space and time is required to construct a robust

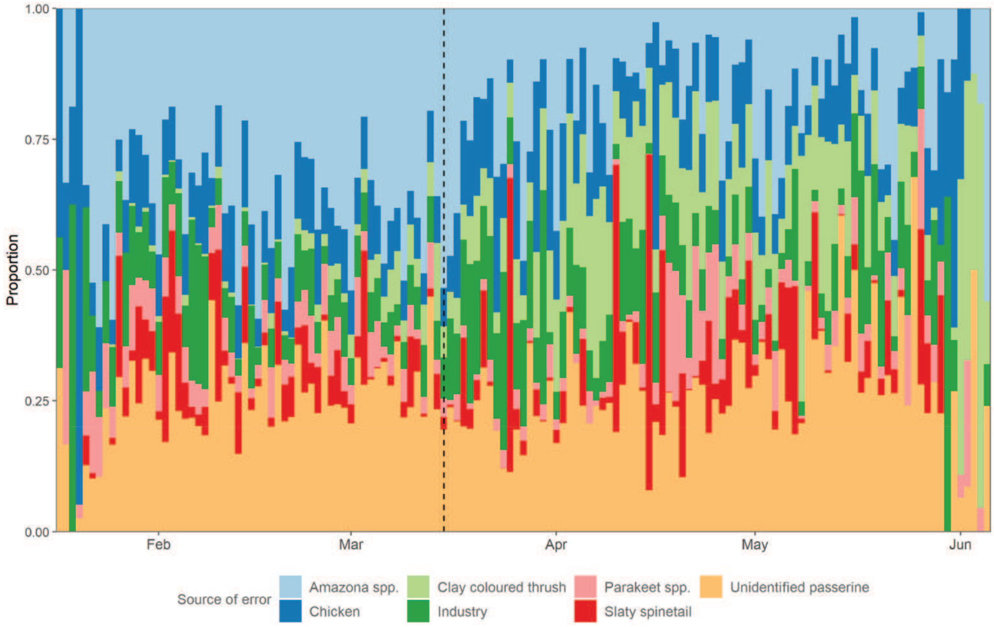

Figure 5. The temporal change in source background noise in the false-positive detection of recogniser one in the evaluation dataset across the study period. There is a distinct change in the proportions of *Amazona* spp., clay-coloured thrush and chicken after the time period that the recogniser one's training dataset was selected (dashed line).

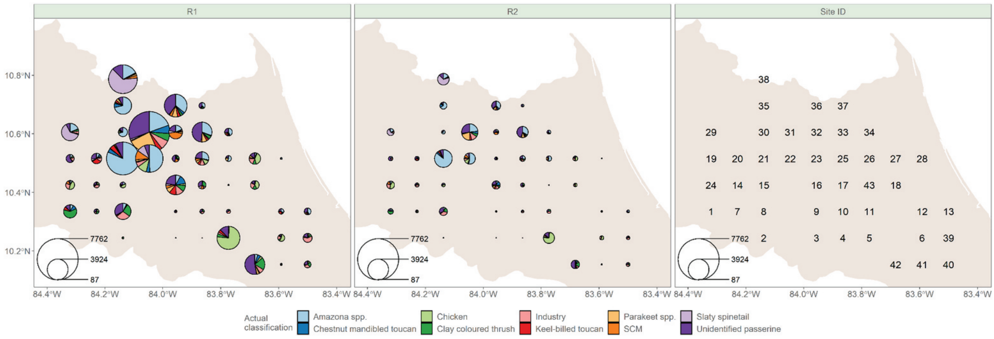

Figure 6. Site-level variation in the overall top 10 sources of error; Scarlet Macaw and Great Green Macaw detections from recogniser one (R1) and recogniser two (R2). There is a clear difference in the sources of error across the study sites, with northern sites being dominated by *Amazona* spp., unidentified passerines and at sites 29 and 38, the slaty spinetail (*Synallaxis brachyura*). Overall false positives were reduced by 90% from R1 to R2 and at each site, and the variation in sources of error was also generally reduced.

recogniser. Our workflow addresses this by using an initial small recogniser to create a training dataset that captures non-target signals driving false-positive detections. This method is similar to the 'basic recognition model' approach used by Buxton and Jones (2012), whereby a small amount of data is used to train a basic model, which is then used to search for more positive training data in other recordings. The main difference is that we extended this approach to gather positive and negative cases.

We initially assumed Great Green Macaw and Scarlet Macaw calls were sufficiently distinctive that using recordings from a narrow window in the breeding season would allow us to construct a robust recogniser. However, the performance of our initial breeding season recogniser was very poor when applied to data from outside the breeding season due to the increased diversity of the soundscape. Importantly, this variation is difficult to capture without using a large, manually labelled dataset. This would have required us to label over 9000 30-min recordings. Incorporating the positive detections and previous training datasets to train a second classifier significantly increases precision whilst losing only a small level of recall performance. This method is similar to the workflow of Balantic and Donovan (2020), using template matching and machine learning together to reduce false positives. The main benefit of our approach is the relatively simple pipeline structure. Nonetheless, we found that iterative labelling and training were required to create a useable recogniser when the target species exhibits considerable call-type variation and high spatiotemporal variation in background noise.

Recogniser two, which included seasonal variation, still performed poorly compared to many other machine learning PAM studies (Bravo et al. 2017; Jahn et al. 2017; Knight and Bayne 2019; Balantic and Donovan 2020; Knight et al. 2020; Gillings and Scott 2021). However, our results are comparable to similar studies in tropical environments (Swiston and Mennill 2009; Heinicke et al. 2015). This reduced performance likely reflects the highly diverse soundscape arising from the high diversity of species with overlapping signal frequencies (Slabbekoorn 2004) and anthropogenic noise (Slabbekoorn and Peet 2003) in tropical settings. Both factors were apparent in our study region in the northeast of Costa Rica, which encompasses primarily cropland and urban in the south and mainly cattle pasture and forest land use types in the north (Fagan et al. 2013; Jadin et al. 2016; Karra et al. 2021).

When we combine all our manually checked positive detections, we found that macaws are only found in 5.5% of the recordings ($n = 592$). This demonstrates another issue with creating a recogniser for a rare species: a lack of training data. Macaws' behaviour means that when they are at a site, it is likely they will call multiple times. The most calls we found in one recording were over 200. This can then be biased as sites where macaws are found less often will not be as represented in the training dataset. When this issue is considered alongside the high variation in spatio-temporal background noise, it demonstrates the challenge of creating a recogniser in this type of environment.

Another important finding was that the performance of the breeding season recogniser was hugely over-estimated. Practitioners should consider this when creating classifiers for large-scale rollout. Here, we have explicitly described the data used to assess model performance, but this is not always the case (Heinicke et al. 2015). It is unclear how many published recognisers would perform in a real-world PAM study, especially if studies use high-quality recordings to create their classifier (Bardeli et al. 2010; Buxton and Jones 2012; Priyadarshani et al. 2018). Manually checking positive detections and using power analysis to determine how many negative cases need to be reviewed is a more labour-intensive task but gives a more accurate assessment of real-world performance.

We did not succeed in creating a recogniser that can be used without manually checking outputs. This is not uncommon. Even when classifiers have high-performance metrics, manually checking positive classifications to filter out false positives is often done before downstream analyses of species abundance and distribution (Buxton and Jones 2012; Zwart et al. 2014; Colbert et al. 2015; Kalan et al. 2015; Sidie-Slettedahl et al. 2015). A way to deal with this need for manual checking is to use statistical methods that can reduce the amount of data needed for validation (Knight et al. 2020), account for false positives (Banner et al. 2018), false negatives (MacKenzie et al. 2002) or both (Chambert et al. 2018; Wright et al. 2020). Our method reduced the number of false positives that must be manually validated by 80%, from 3.95 per recording ($n = 33243$) in R1 to 0.75 per recording ($n = 6309$) in R2. Therefore, although it is still necessary to manually check all positive detections when the final recogniser runs on new data, this represents a significant reduction in the effort needed to clean the data.

Denoising can increase classifier accuracy (Stowell et al. 2016). Interestingly, denoising is not often used, even when studies report the high performance of recognisers. This supports the argument made by Priyadarshani et al. (2018) that many published PAM studies are done in low-noise environments, using species with simple calls. The primary barrier to denoising is that few simple, user-friendly techniques are available, so many classifiers do not use them (e.g. Sebastián-González et al. 2015; Balantic and Donovan 2020; Knight et al. 2020). We provide an alternative way to tackle the issue of background noise without using denoising. However, it does not entirely deal with the problem as our recogniser performance is not comparable to the best-published recognisers.

Future work

We could improve our classifier in several ways. Effective template matching requires a sufficiently representative call library (Aide et al. 2013; Gibb et al. 2019). We set the matching threshold very low to help our classifier deal with intra- and inter-call type variation of the Great Green Macaw and Scarlet Macaw. Although we did not miss any target signal, the random forest classified a large volume of data. Investing extra time in developing and refining the template set would likely reduce the time needed to review calls and improve performance by reducing the size of the initial dataset.

Training dataset imbalance is common in machine learning, not just in bioacoustics (Salamon and Bello 2017). Methods such as data augmentation can provide an excellent solution to this problem (Stowell et al. 2019). Our training dataset was very imbalanced, and the quantity of Scarlet Macaw training data was only ~25% of the Great Green Macaw training dataset. This resulted in the performance of Scarlet Macaw being significantly lower than that of the Great Green Macaw. Our current workflow uses a standard window size of 1 s, within which the call may occupy only a fraction. Therefore, we would have to change how we structure our pipeline to accommodate data augmentation, as many data augmentation techniques need to have the start and end of the target signal (Salamon and Bello 2017). Being able to find the beginning and end of a target signal automatically would be a massive step towards facilitating data augmentation. It would also help reduce the effect of background noise on random forest accuracy by reducing the amount of background noise present in each window to be classified.

Conclusions

When developing a recogniser there is a trade-off between recall and precision. If the target species are rare, then high recall, so no potential target signal is missed, might be preferred. However, this could come at the expense of precision. We have demonstrated that by using the iterative approach to creating a training dataset, we can increase precision significantly, whilst recall only decreases slightly. Importantly, we have shown that performance needs to be done over a large dataset to get an accurate estimation. This is particularly important if practitioners are using performance metrics to determine how many detections need to be screened to validate a survey.

Passive acoustic monitoring has great potential to enable the scaling up of biodiversity monitoring to inform policy and conservation strategies. There is an increasing need for simple methods to automate or semi-automate data extraction from PAM surveys (Marques et al. 2013; Stowell et al. 2016). The development of generalised deep-learning classifiers will make PAM a simple and accessible tool for biodiversity monitoring. However, their development may still be far off, and species-specific recognisers can be more accurate (Priyadarshani et al. 2018). Custom recognisers also allow practitioners to tailor them to their specific needs. Developing such recognisers is challenging, and although there is an increasing amount of literature on different methods to create classifiers, these are often small-scale studies in temperate environments (Sugai et al. 2019) that rely on high-quality recordings of species with simple calls (Priyadarshani et al. 2018). We have demonstrated a simple workflow that can provide users with an accessible but time-consuming way to create a PAM recogniser for acoustically diverse environments.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The work was supported by the Natural Environment Research Council [NE/L002450/1].

ORCID

Thomas C. Lewis http://orcid.org/0000-0001-8787-3581

References

Abrahams C, Geary M. 2020. Combining bioacoustics and occupancy modelling for improved monitoring of rare breeding bird populations. Ecol Indic. 112:106131. doi: 10.1016/j.ecolind.2020.106131.

Aide TM, Corrada-Bravo C, Campos-Cerqueira M, Milan C, Vega G, Alvarez R. 2013. Real-time bioacoustics monitoring and automated species identification. PeerJ. 1:e103. doi: 10.7717/peerj.103.

Antonov IA, Saleev VM. 1979. An economic method of computing LPτ-sequences. USSR Comput Math Math Phys. 19(1):252–256. doi: 10.1016/0041-5553(79)90085-5.

Araya-Salas M, Smith-Vidaurre G, Golding N. 2017. warbleR: an r package to streamline analysis of animal acoustic signals. Methods Ecol Evol. 8(2):184–191. doi: 10.1111/2041-210X.12624.

Ayala-Berdon J, Medina-Bello KI, López-Cuamatzi IL, Vázquez-Fuerte R, MacSwiney GMC, Orozco-Lugo L, Iñiguez-Dávalos I, Guillén-Servent A, Martínez-Gómez M. 2020. Random forest is the best species predictor for a community of insectivorous bats inhabiting a mountain ecosystem of central Mexico. Bioacoustics. 30(5):608–628. doi: 10.1080/09524622.2020.1835539.

Balantic CM, Donovan TM. 2020. Statistical learning mitigation of false positives from template-detected data in automated acoustic wildlife monitoring. Bioacoustics. 29 (3):296–321. doi: 10.1080/09524622.2019.1605309.

Banner KM, Irvine, K. M., Rodhouse, T. J., Wright, W. J., Rodriguez, R. M., Litt, A. R. 2018. Improving geographically extensive acoustic survey designs for modeling species occurrence with imperfect detection and misidentification. Ecol Evol. 8(12):6144–6156. doi: 10.1002/ece3.4162.

Bardeli R, Wolff, D., Kurth, F., Koch, M., Tauchert, K. H., Frommolt, K. H. 2010. Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. Pattern Recognit Lett. 31(12):1524–1534. doi: 10.1016/j.patrec.2009.09.014.

Barnes RFW. 2001. How reliable are dung counts for estimating elephant numbers? Afr J Ecol. 39 (1):1–9. doi: 10.1111/j.1365-2028.2001.00266.x.

Bermant PC, Bronstein MM, Wood RJ, Gero S, Gruber DF. 2019. Deep machine learning techniques for the detection and classification of sperm whale bioacoustics. Sci Rep. 9 (1):12588. doi: 10.1038/s41598-019-48909-4.

Bioacoustics Research Program. 2016. Raven Lite: interactive sound analysis software. Ithaca (NY): The Cornell Lab of Ornithology. http://www.birds.cornell.edu/raven.

BirdLife International. 2020. Great green macaw (ara ambiguus) - BirdLife species factsheet. [accessed 2021 Jan 5]. http://datazone.birdlife.org/species/factsheet/great-green-macaw-ara-ambiguus.

Bradbury JW. 2003. Vocal communication in wild parrots. In: Animal social complexity: intelligence, culture, and individualized societies. Cambridge (MA), US: Harvard University Press; p. 293–316.

Bradfer-Lawrence T, Bunnefeld, N., Gardner, N., Willis, S. G., Dent, D. H. 2020. Rapid assessment of avian species richness and abundance using acoustic indices. Ecol Indic. 115:106400. doi: 10.1016/j.ecolind.2020.106400.

Brandes TS. 2008. Feature vector selection and use with hidden Markov models to identify frequency-modulated bioacoustic signals amidst noise. IEEE Trans Audio Speech Lang Processing. 16(6):1173–1180. doi: 10.1109/TASL.2008.925872.

Bravo CJC, Berríos RÁ, Aide TM. 2017. Species-specific audio detection: a comparison of three template-based detection algorithms using random forests. PeerJ Comput Sci. 3:e113. doi: 10.7717/peerj-cs.113.

Brieman L. 2001. Random forests. J Mach Learn Res. 45(1):5–32. doi: 10.1023/A:1010933404324.

Buxton RT, Jones IL. 2012. Measuring nocturnal seabird activity and status using acoustic recording devices: applications for island restoration. J Field Ornithol. 83(1):47–60. doi: 10.1111/j.1557-9263.2011.00355.x.

Campos-Cerqueira M, Aide TM, Jones K. 2016. Improving distribution data of threatened species by combining acoustic monitoring and occupancy modelling. Methods Ecol Evol. 7 (11):1340–1348. doi: 10.1111/2041-210X.12599.

Chambert T, Waddle JH, Miller DAW, Walls SC, Nichols JD. 2018. A new framework for analysing automated acoustic species detection data: occupancy estimation and optimization of recordings post-processing. Methods Ecol Evol. 9(3):560–570. doi: 10.1111/2041-210X.12910.

Cintra R, Naka LN. 2011. Spatial variation in bird community composition in relation to Topographic Gradient and Forest Heterogeneity in a Central Amazonian Rainforest. Int J Ecol. 2012:e435671. doi: 10.1155/2012/435671.

Clink DJ, Klinck H, Zamora-Gutierrez V. 2020. Unsupervised acoustic classification of individual gibbon females and the implications for passive acoustic monitoring. Methods Ecol Evol. 12 (2):328–341. doi: 10.1111/2041-210X.13520.

Colbert DS, Ruttinger, J. A., Streich, M., Chamberlain, M., Conner, L. M., Warren, R. J. 2015. Application of autonomous recording units to monitor gobbling activity by wild turkey. Wildl Soc Bull. 39(4):757–763. doi: 10.1002/wsb.577.

Colonna JG, Gama J, Nakamura EF. 2016. How to correctly evaluate an automatic bioacoustics classification method. In: Luaces O, editor. Advances in Artificial Intelligence. Cham: Springer International Publishing (Lecture Notes in Computer Science); p. 37–47. 10.1007/978-3-319-44636-3_4.

Davis S, Mermelstein P. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoust Speech Signal Process. 28 (4):357–366. doi: 10.1109/TASSP.1980.1163420.

Dénes FV, Tella JL, Beissinger SR. 2018. Revisiting methods for estimating parrot abundance and population size. Emu - Austral Ornithol. 118(1):67–79. doi: 10.1080/01584197.2017.1401903.

Digby A, Towsey M, Bell BD, Teal PD. 2013. A practical comparison of manual and autonomous methods for acoustic monitoring. Methods Ecol Evol. 4(7):675–683.

Dyrcz A. 1983. Breeding ecology of the Clay-coloured Robin Turdus grayi in lowland Panama. Ibis (Lond 1859). 125(3):287–304. doi: 10.1111/j.1474-919X.1983.tb03115.x.

Fagan ME, DeFries, RS, Sesnie, SE, Arroyo, JP, Walker, W, Soto, C, Chazdon, RL, Sanchun, A. 2013. Land cover dynamics following a deforestation ban in northern Costa Rica. Environ Res Lett. 8(3):034017. doi: 10.1088/1748-9326/8/3/034017.

Fedy BC, Aldridge CL. 2011. The importance of within-year repeated counts and the influence of scale on long-term monitoring of sage-grouse. J Wildl Manage. 75(5):1022–1033. doi: 10.1002/jwmg.155.

Gibb R, Browning E, Glover-Kapfer P, Jones KE. 2019. Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. Methods Ecol Evol. 10(2):169–185. doi: 10.1111/2041-210X.13101.

Gibbs JP, Snell HL, Causton CE. 1999. Effective monitoring for adaptive wildlife management: lessons from the Galápagos Islands. J Wildl Manage. 63(4):1055–1065. doi: 10.2307/3802825.

Gillings S, Scott C. 2021. Nocturnal flight calling behaviour of thrushes in relation to artificial light at night. Ibis (Lond 1859). 163(4):1379–1393. doi: 10.1111/ibi.12955.

Guschanski K, Vigilant, L., McNeilage, A., Gray, M., Kagoda, E., Robbins, M. M. 2009. Counting elusive animals: comparing field and genetic census of the entire mountain gorilla population of Bwindi Impenetrable National Park, Uganda. Biol Conserv. 142(2):290–300. doi: 10.1016/j.biocon.2008.10.024.

Hafner SD, Katz J. 2018. A short introduction to acoustic template matching with monitoR. [accessed 2021 Mar 3]. https://cran.r-project.org/web/packages/monitoR/vignettes/monitoR_QuickStart.pdf .

Heinicke S, Kalan, A. K., Wagner, O. J., Mundry, R., Lukashevich, H., Kühl, H. S. 2015. Assessing the performance of a semi-automated acoustic monitoring system for primates. Methods Ecol Evol. 6(7):753–763. doi: 10.1111/2041-210X.12384.

Hill AP, Prince P, Snaddon JL, Doncaster CP, Rogers A. 2019. AudioMoth: a low-cost acoustic device for monitoring biodiversity and the environment. HardwareX. 6:e00073. doi: 10.1016/j.ohx.2019.e00073.

Ioki K, Tsuyuki S, Hirata Y, Phua M-H, Wong WVC, Ling Z-Y, Johari SA, Korom A, James D, Saito H, et al. 2016. Evaluation of the similarity in tree community composition in a tropical rainforest using airborne LiDAR data. Remote Sens Environ. 173:304–313. doi: 10.1016/j.rse.2015.07.024.

IUCN. 2024. The IUCN red list of threatened Species, IUCN red list of threatened species. [accessed 2024 Mar 14]. https://www.iucnredlist.org/en.

Jadin I, Meyfroidt P, Lambin EF. 2016. International trade, and land use intensification and spatial reorganization explain Costa Rica's forest transition. Environ Res Lett. 11(3):035005. doi: 10.1088/1748-9326/11/3/035005.

Jahn O, Ganchev TD, Marques MI, Schuchmann K-L. 2017. Automated sound recognition provides insights into the behavioral ecology of a tropical bird. PLOS ONE. 12(1):e0169041. doi: 10.1371/journal.pone.0169041.

Kahl S, Wood, C. M., Eibl, M., Klinck, H. 2021. BirdNET: a deep learning solution for avian diversity monitoring. Ecol Inf. 61:101236. doi: 10.1016/j.ecoinf.2021.101236.

Kalan AK, Mundry, R., Wagner, O. J., Heinicke, S., Boesch, C., Kühl, H. S. 2015. Towards the automated detection and occupancy estimation of primates using passive acoustic monitoring. Ecol Indic. 54:217–226. doi: 10.1016/j.ecolind.2015.02.023.

Karra K, Kontgis, C, Statman-Weil, Z, Mazzariello, JC, Mathis, M, Brumby, SP. 2021. Global land use/land cover with sentinel 2 and deep learning. In 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. IEEE. p. 4704–4707.

Katz J, Hafner SD, Donovan T. 2016a. Assessment of error rates in Acoustic Monitoring with the R package monitoR. Bioacoustics. 25(2):177–196. doi: 10.1080/09524622.2015.1133320.

Katz J, Hafner SD, Donovan T. 2016b. Tools for automated acoustic monitoring within the R package monitoR. Bioacoustics. 25(2):197–210. doi: 10.1080/09524622.2016.1138415.

Knight EC, Bayne EM. 2019. Classification threshold and training data affect the quality and utility of focal species data processed with automated audio-recognition software. Bioacoustics. 28 (6):539–554. doi: 10.1080/09524622.2018.1503971.

Knight EC, Sòlymos, P., Scott, C., Bayne, E. M. 2020. Validation prediction: a flexible protocol to increase efficiency of automated acoustic processing for wildlife research. Ecol Appl. 30(7): e02140. doi: 10.1002/eap.2140.

Kuhn M, Wickham H. 2020. 'Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles'. https://www.tidymodels.org.

Lasseck M. 2014 September. Large-scale identification of birds in audio recordings. CLEF (Working Notes). 643–653.

Loh C, Yuan T, Ramli DA. 2013. Frog sound identification system for frog Species Recognition.

MacKenzie DI, Nichols JD, Lachman GB, Droege S, Andrew Royle J, Langtimm CA. 2002. Estimating site occupancy rates when detection probabilities are less than one. Ecology. 83 (8):2248–2255. doi: 10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2.

Marin-Cudraz T, Muffat-Joly B, Novoa C, Aubry P, Desmet J-F, Mahamoud-Issa M, Nicolè F, Van Niekerk MH, Mathevon N, Sèbe F, et al. 2019. Acoustic monitoring of rock ptarmigan: a multi-year comparison with point-count protocol. Ecol Indic. 101:710–719. doi: 10.1016/j.ecolind.2019.01.071.

Marques TA, Thomas, L., Martin, S.W., Mellinger, D.K., Ward, J.A., Moretti, D.J., Harris, D., Tyack, P.L. 2013. Estimating animal population density using passive acoustics. Biol Rev Camb Philos Soc. 88(2):287–309. doi: 10.1111/brv.12001.

Monge O, Schmidt K, Vaughan C, Gutiérrez-Espeleta G. 2016. Genetic patterns and conservation of the Scarlet Macaw (Ara macao) in Costa Rica. Conserv Genet. 17(3):745–750. doi: 10.1007/s10592-015-0804-3.

Montes-Medina AC, Salinas-Melgoza A, Renton K. 2016. Contextual flexibility in the vocal repertoire of an Amazon parrot. Front Zool. 13(1):1–13. doi: 10.1186/s12983-016-0169-6.

Nichols JD, Johnson FA, Williams BK, Boomer GS. 2015. On formally integrating science and policy: walking the walk. Journal Of Applied Ecology. 52(3):539–543. doi: 10.1111/1365-2664.12406.

Noda JJ, Travieso CM, Sánchez-Rodríguez D. 2016. Methodology for automatic bioacoustic classification of anurans based on feature fusion. Expert Syst Appl. 50:100–106. doi: 10.1016/j.eswa.2015.12.020.

Noda JJ, Travieso, CM, Sánchez-Rodríguez, D, Dutta, MK, Singh, A. 2016. Using bioacoustic signals and support vector machine for automatic classification of insects. In: 2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN). IEEE. p. 656–659.

Nuttall MN, Griffin, O., Fewster, R.M., McGowan, P.J., Abernethy, K., O'Kelly, H., Nut, M., Sot, V., Bunnefeld, N. 2022. Long-term monitoring of wildlife populations for protected area management in Southeast Asia. Conserv Sci Pract. 4(2):e614. doi: 10.1111/csp2.614.

Pereira HM, David Cooper H. 2006. Towards the global monitoring of biodiversity change. Trends Ecol Evol. 21(3):123–129. doi: 10.1016/j.tree.2005.10.015.

Pérez-Granados C, Bota G, Giralt D, Barrero A, Gómez-Catasús J, Bustillo-De La Rosa D, Traba J. 2019. Vocal activity rate index: a useful method to infer terrestrial bird abundance with acoustic monitoring. Ibis (Lond 1859). 161(4):901–907. doi: 10.1111/ibi.12728.

Pérez-Granados C. 2023. A first assessment of birdnet performance at varying distances: a playback experiment. Ardeola. 70(2):257–269. doi: 10.13157/arla.70.2.2023.sc1.

Pillay R, Fletcher RJ, Sieving KE, Udell BJ, Bernard H. 2019. Bioacoustic monitoring reveals shifts in breeding songbird populations and singing behaviour with selective logging in tropical forests. Journal Of Applied Ecology. 56(11):2482–2492. doi: 10.1111/1365-2664.13492.

Pollock KH, Nichols JD, Simons TR, Farnsworth GL, Bailey LL, Sauer JR. 2002. Large scale wildlife monitoring studies: statistical methods for design and analysis. Environmetrics. 13(2):105–119. doi: 10.1002/env.514.

Priyadarshani N, Marsland S, Castro I. 2018. Automated birdsong recognition in complex acoustic environments: a review. Journal Of Avian Biology. 49(5):jav–01447. doi: 10.1111/jav.01447.

Raghuram MA, Chavan NR, Belur R, Koolagudi SG. 2016. Bird classification based on their sound patterns. Int J Speech Technol. 19(4):791–804. doi: 10.1007/s10772-016-9372-2.

R Core Team. 2021. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Renardy M, Joslyn LR, Millar JA, Kirschner DE. 2021. To Sobol or not to Sobol? The effects of sampling schemes in systems biology applications. Mathematical Biosciences. 337:108593. doi: 10.1016/j.mbs.2021.108593.

Salamon J, Bello JP. 2017. Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Process Lett. 24(3):279–283. doi: 10.1109/LSP.2017.2657381.

Salamon J, Bello JP, Farnsworth A, Robbins M, Keen S, Klinck H, Kelling S. 2016. Towards the Automatic Classification of Avian Flight Calls for bioacoustic monitoring. PLOS ONE. 11(11): e0166866. doi: 10.1371/journal.pone.0166866.

Sebastián-González E, Pang-Ching J, Barbosa JM, Hart P. 2015. Bioacoustics for species management: two case studies with a Hawaiian forest bird. Ecology And Evolution. 5(20):4696–4705. doi: 10.1002/ece3.1743.

Sidie-Slettedahl AM, Jensen KC, Johnson RR, Arnold TW, Austin JE, Stafford JD. 2015. Evaluation of autonomous recording units for detecting 3 species of secretive marsh birds. Wildl Soc Bull. 39(3):626–634. doi: 10.1002/wsb.569.

Slabbekoorn H. 2004. Habitat-dependent ambient noise: consistent spectral profiles in two African forest types. J Acoust Soc Am. 116(6):3727–3733. doi: 10.1121/1.1811121.

Slabbekoorn H, Peet M. 2003. Birds sing at a higher pitch in urban noise. Nature. 424 (6946):267–267. doi: 10.1038/424267a.

Slabbekoorn H, Ripmeester EAP. 2008. Birdsong and anthropogenic noise: implications and applications for conservation. Mol Ecol. 17(1):72–83. doi: 10.1111/j.1365-294X.2007.03487.x.

Smith-Vidaurre G, Araya-Salas M, Wright TF, Naguib M. 2020. Individual signatures outweigh social group identity in contact calls of a communally nesting parrot. Behav Ecol. Edited by M. Naguib, 31(2):448–458. doi: 10.1093/beheco/arz202.

Snaddon J, Petrokofsky G, Jepson P, Willis KJ. 2013. Biodiversity technologies: tools as change agents. Biol Lett. 9(1):20121029. doi: 10.1098/rsbl.2012.1029.

Sobol IM. 1967. On the distribution of points in a cube and the approximate evaluation of integrals. USSR Comput Math Math Phys. 7(4):86–112. doi: 10.1016/0041-5553(67)90144-9.

Stowell D, Petrusková T, Šálek M, Linhart P. 2019. Automatic acoustic identification of individuals in multiple species: improving identification across recording conditions. J R Soc Interface. 16 (153):20180940. doi: 10.1098/rsif.2018.0940.

Stowell D, Wood, M, Stylianou, Y, Glotin, H. 2016. Bird detection in audio: a survey and a challenge. In: 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP). IEEE. p. 1–6.

Sugai LSM, Silva TSF, Ribeiro JW, Llusia D. 2019. Terrestrial passive acoustic monitoring: review and perspectives. BioScience. 69(1):15–25. doi: 10.1093/biosci/biy147.

Swiston KA, Mennill DJ. 2009. Comparison of manual and automated methods for identifying target sounds in audio recordings of pileated, pale-billed, and putative Ivory-billed woodpeckers. J Field Ornithol. 80(1):42–50. doi: 10.1111/j.1557-9263.2009.00204.x.

Tachibana RO, Oosugi N, Okanoya K, Bolhuis JJ. 2014. Semi-automatic classification of birdsong elements using a linear support vector machine. PLOS ONE. 9(3):e92584. doi: 10.1371/journal.pone.0092584.

Taylor S, Perrin MR. 2005. Vocalisations of the Brown-headed parrot, Poicephalus cryptoxanthus: their general form and behavioural context. Ostrich. 76(1–2):61–72. doi: 10.2989/00306520509485474.

Teixeira D, Maron M, van Rensburg BJ. 2019. Bioacoustic monitoring of animal vocal behavior for conservation. Conserv Sci Pract. 1(8):e72. doi: 10.1111/csp2.72.

Ventura TM, de Oliveira AG, Ganchev TD, de Figueiredo JM, Jahn O, Marques MI, Schuchmann K-L. 2015. Audio parameterization with robust frame selection for improved bird identification. Expert Systems With Applications. 42(22):8463–8471. doi: 10.1016/j.eswa.2015.07.002.

Wardhaugh CW, Stork NE, Edwards W. 2014. Canopy invertebrate community composition on rainforest trees: different microhabitats support very different invertebrate communities. Austral Ecology. 39(4):367–377. doi: 10.1111/aec.12085.

Weerasena H, Jayawardhana M, Egodage D, Fernando H, Sooriyaarachchi S, Gamage C, Kottege N. 2018 October. Continuous automatic bioacoustics monitoring of bird calls with local processing on node level. TENCON 2018-2018 IEEE Region 10 Conference. IEEE. p. 0235–0239.

Wickham H, François, R, Henry, L, Müller, K, Vaughan, D. 2023. Posit Software. dplyr: a grammar of data manipulation. CRAN R Project. https://github.com/tidyverse/dplyr.

Wood CM, Popescu VD, Klinck H, Keane JJ, Gutiérrez RJ, Sawyer SC, Peery MZ. 2019. Detecting small changes in populations at landscape scales: a bioacoustic site-occupancy framework. Ecol Indic. 98:492–507. doi: 10.1016/j.ecolind.2018.11.018.

Wright WJ, Irvine KM, Almberg ES, Litt AR. 2020. Modelling misclassification in multi-species acoustic data when estimating occupancy and relative activity. Methods Ecol Evol. 11(1):71–81. doi: 10.1111/2041-210X.13315.

Zdenek CN, Heinsohn R, Langmore NE. 2015. Vocal complexity in the palm cockatoo (probosciger aterrimus). Bioacoustics. 24(3):253–267. doi: 10.1080/09524622.2015.1070281.

Zhong M, LeBien J, Campos-Cerqueira M, Dodhia R, Lavista Ferres J, Velev JP, Aide TM. 2020. Multispecies bioacoustic classification using transfer learning of deep convolutional neural networks with pseudo-labeling. Applied Acoustics. 166:107375. doi: 10.1016/j.apacoust.2020.107375.

Zwart MC, Baker A, McGowan PJK, Whittingham MJ. 2014. The use of automated bioacoustic recorders to replace human wildlife surveys: an example using Nightjars. PLOS ONE. 9(7):e102770. doi: 10.1371/journal.pone.0102770.