Detecting Bird Loss in Forest Ecosystems

Sushmita Azad
University of Illinois
Urbana-Champaign, IL
sazad2@illinois.edu

Sogol Bazargani
University of Illinois
Urbana-Champaign, IL
sogolb2@illinois.edu

Aabhas Chauhan
University of Illinois
Urbana-Champaign, IL
aabhasc2@illinois.edu

Zuzana Burivavlova
University of Wisonsin
Madison, WI
burivavlova@wisc.edu

***Abstract*—Loss of biodiversity is notoriously hard to detect. Most organizations, including governments, usually rely on satellite imagery to detect forest cover and use that to extrapolate biodiversity and ecosystem health.**

However, threats like selective logging, wildlife baits/traps, and hunting, that don't necessarily reduce canopy cover are impossible to detect by analyzing canopy cover.

In order to address this problem, in this research we utilize the soundscapes of natural ecosystems, and build supervised and weakly supervised models for what healthy ecosystems with an abundance of any particular part of the fauna sound like.

Our study provides a quantitative analysis of the bird loss as a result of logging, across three different locations in the rain forests of Bornea, Indonesia.

***Index Terms*—Machine learning, Audio Analysis, Bornea, Indonesia**

I. Introduction

Primary tropical forests are being degraded by selective logging, although, detecting this degradation as a result of man-made activities like road and camp construction is a difficult task.

Analyzing the health of forests is not only determined by their cover but also by the level of degradation in those forests left standing. Remote sensing analyses using high resolution satellite imagery is a popular technique to analyze forest cover loss but it fails to monitor forest degradation levels. Repeated on-ground surveys are a viable but expensive option. The newly emerging field of bio-acoustics - recording and analysis of entire soundscapes - is turning out to be a successful alternative for the same.

By using the soundscapes of natural ecosystems, we can build models for what healthy ecosystems with an abundance of any particular part of the fauna sound like, and corresponding models for ecosystems with loss of this species, and use these models to arrive at more accurate and non-invasive estimates of ecosystem health.

For our study, we use sound data collected from 4 different sites in Indonesian Borneo forests.

- **Site A** – Logged for an extended time
- **Site B & C** – Fairly pristine
- **Site D** – Logging begins midway through recordings.

By analyzing data from these different sites, we want to build representative models for what 1) healthy, 2) newly logged and 3) extensively logged ecosystems sound like, and use these models to be able to gauge the human-impact on similar tropical rain forests without relying on satellite imagery.

In the future, we would like to extend this to be an analysis over time, and determine how long it takes for an ecosystem to regain its bird population after being selectively logged.

Through all our experiments, we use dawn call, i.e. birdsong at daybreak as our indicator of bird population and in turn – health of biodiversity.

We have conducted extensive experiments on audio sounds recorded 4 hours a day, over a span of 3 months. We have analyzed these audios using supervised and weakly supervised models. We compare the performance of our model against multiple baselines.

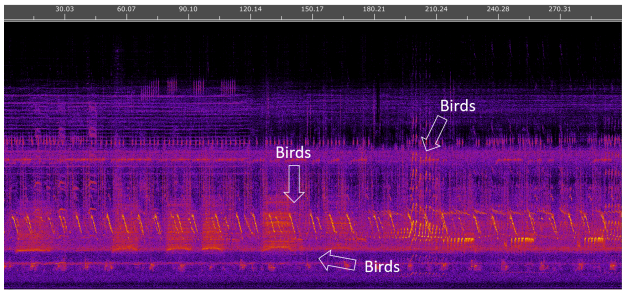
II. Related Work

Since bio-acoustics is a newly emerging field bridging biological and acoustics science, majority of the research in this area analyze the biodiversity through simple methods. To the best of our knowledge, biodiversity loss has not been significantly studied using machine learning based methods. As a related study in this area, [1] analyzes the intensities of logging through a pantropical meta-analysis and using an information-theoretic approach. Their framework is based on modeling relative species richness based on the logging intensity, time since last logging event happened, area of logged forest region, taxonomic group and others. According to the study, all taxonomic groups (mammals, amphibians, invertebrates) other than birds exhibit a decline in species richness with increase in logging.

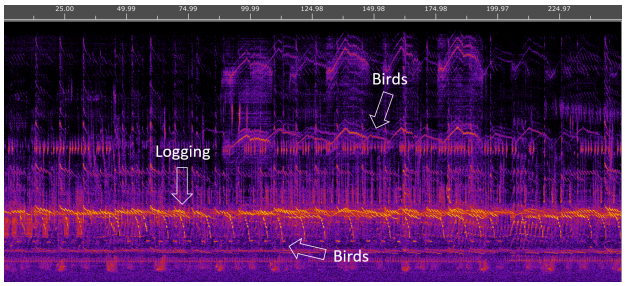
Other related work have attempted to detect individual species using expert help, specific algorithms [2] [3] and deep learning techniques [4]. However, most of these works use abundant manually and crowd-sourced annotated data that is not openly available for use.

Traditional approaches in this field use HMM and GMM to classify the bird sounds. More specifically [2] uses a combination of Principal Component Analysis, Vector Quantization, and Hidden Markov Models for classification. It extracts the features using Mel Frequency Cepstral Coefficient (MFCC) Analysis, and then applies GMM to perform the classification.

Recent approaches in this area are mostly based on Convolutional Neural Networks (CNNs). More specifically, Authors in [5] use the image classification model Inception V4 and extend it with a time-frequency attention mechanism. This method has proved to be significantly effective. As other recent methods,



(a) Pristine Site



(b) Logged Site

Fig. 1: Input Spectrograms

we can point out to [6], which uses a more classical CNN architecture, and [7] which is based on a multi-modal deep neural network.

III. Methodology

In order to learn a universal model for each site in the forest, we trained a classifier on manually annotated data. Due to the unavailability of such labels for our audio samples, we hand annotated several minutes of audio from each site and used that to train the model.

Training a classifier on audio data is generally composed of two steps:

1) Pre-processing and Feature Extraction
2) Model Training

A. Pre-processing and Feature Extraction

Initially in order to extract relevant features from the audio signal $s(n)$, it was divided into $3/4$ overlapping frames by the application of hamming window function and analyzed using Short-term Fourier Transform (STFT).

$$S(l,k) = \sum_{n=0}^{N-1} s(n+lM)w(n)exp(-j(\frac{2\pi}{N})nk) \qquad (1)$$

where N is the size of STFT, $k(0 \leq k \leq K-1)$ is the frequency bin index, l is the time frame index, w is an analysis window of size l_w, and M is the hop size.

Fig. 1 shows the two input spectrograms generated from the pre-processed STFTs for the pristine site and the logged site.

To extract features from the audio frames, we projected our high dimensional audio frames to low dimensional features while keeping the underlying structure of spectro-temporal modulations, to be able to use them as the input to our model. In this research we took advantage of popular feature extraction methods like Principal Component Analysis (PCA) and Non-Negative Matrix Factorization (NMF) which have proved to be successful in various audio classification problems.

1) Principal Component Analysis: PCA proposed by [8] is a dimensionality reduction technique that aims to model the total variance of the data via a number of uncorrelated principal components. In other words, the first component is the axis passing the data with the highest variance. The second component tries to maximize the variance which is not explained by the first component, so components are orthogonal and independent.

In this research, we experimented with different number of components in order to find the optimal one. We used 8 to 12 components based on characteristics of acoustics at different sites.

2) Non-negative Matrix Factorization: NMF proposed by [9] has proved to be very effective in decomposing mixture of sources in an audio signal in a time frequency domain. In fact based on [10] the most accurate source separation results can be obtained using NMF. Therefore since we aim to separate bird sound from logging and noise, NMF can be an improvement over PCA.

Given a matrix $V \in R_+^{M \times N}$, where each column is a data point (N data points), NMF would decompose this into two non-negative low rank matrices H and W, through an iterative process where,

$$V \approx H \times W \qquad (2)$$

where $H \in R_+^{M \times k}$ and $H \in R_+^{k \times N}$ and $k \leq min(M,N)$. As a result, NMF reduces the dimensionality of a M-dimensional data to k dimensions. Furthermore, since the audio signals always contain non-negative values, factorizing this data into non-negative matrices makes much more sense.

In this research, experiments were conducted to find the optimal value for k, the results shown in Figure 2 indicate that 6-8 features for each data point achieves accuracy of approximately 90%, which is already an improvement over PCA.

This was followed by chunking the data into one second frames for ease in evaluating accuracy at a later stage.

B. Supervised Model Training

For the purpose of this research project, first we trained a binary classifier for each site to detect the presence or

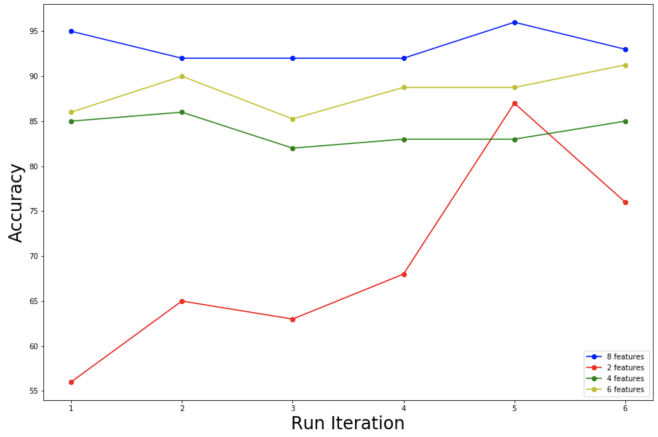


Fig. 2: Accuracy of the model based on number of components chosen for NMF.

absence of birds in one second chunks. As the next step, we perform a multi-way classification on the data obtained from the combination of all four sites.

Various approaches have been used for audio classification, like Gaussian Classifiers, SVM, etc. Based on the research conducted in [11], SVM and Gaussian classifier perform equivalently accurate in audio classification, although Gaussian classifiers are known to perform more accurately when the volume of data available for training is significant.

Gaussian Classifier is mainly based on the assumption that each class can be modeled under a Gaussian or normal distribution. Providing the labeled data, the mean and variance for each class can be estimated. Later the probability of a test sample x belonging to a class w_i can be calculated using

$$P(x|w_i) = N(x|\mu_i, \sigma_i) \tag{3}$$

Where μ_i and σ_i are the mean and co-variance of the class i respectively. Final class prediction can be determined considering the class prior and the value of the aforementioned probability density function.

After separately training models on each site, we trained a multi-way classifier on the entire data from different sites.

C. Scalability and Semi-supervision

Previous approach can successfully classify the data if the annotated sound frames are available. However, hand-annotating hours of sound data is extremely time-consuming, expensive and requires professional resources. Consequently, lack of labeled data makes the supervised approach impractical and unscalable. To tackle this problem in this section we study the effect of using clustering results as assigned labels. Then we train the aforementioned Gaussian classifier on this weakly labeled data, and use our formerly-used hand annotated frames to test the weakly supervised model. The accuracy of the later model can be a measurement of how accurate the cluster labels are in comparison to the true labels.

We attempted two clustering techniques to auto-cluster the data, independently for all sites, namely K-means and Gaussian Mixture Models (GMM).

1) **K-means**: K-means is a popular hard clustering approach which assigns each data point to the closest centroid in the dataset. This process happens through multiple iterations and continues until the algorithm converges. Formally, the objective function for K-means can be given by

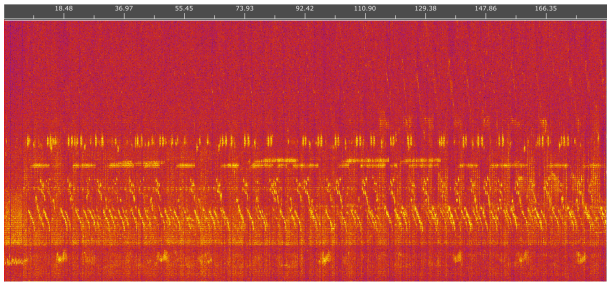
$$J(V) = \sum_{i=0}^{m} \sum_{j=0}^{K} w_{ij} ||x_i - \mu_k||^2 \tag{4}$$

where $V = \{x_0, x_1, ..., x_m\}$ is the set of m data points, K is the number of clusters and w_{ij} is equal to 1 if x_i belongs cluster j, 0 otherwise.
K-means attempts to minimize the objective function $J(V)$. The accuracy of the clustering approach is extremely dependent on the initialization step.
For our project, we used 2 clusters for K-Means, one to detect birds, the other to detect noise (includes insect sounds, logging sounds, etc.). For different sites, different initialization were considered. The number of data points differed based on the sample size considered for each site. The observations are discussed in more detail in the next section. However, the clustering algorithm severely under-performed when there were logging or rain sound in the background.
2) **Gaussian Mixture Models (GMM)**: As opposed to K-means, GMM is a soft clustering approach that predicts the probability of each data point belonging to each cluster, assuming that a cluster can be represented by a mixture of Gaussian distributions. The GMM training consists of an initialization step followed by an optimization step. The optimization step uses the Expectation Maximization (EM) algorithm to update parameters - mean, variance and probability of a point belonging to a specific cluster (gaussian mixture), along with the latent hidden variable. The first three parameter updation constitutes the maximization step, while the later constitutes the expectation step.
In our experiments, we decided to use one gaussian per class to achieve a much higher accuracy on the classifier. The decision was made based on clustering accuracy results achieved after running it on various possible numbers for gaussians per class. Based on the variability in data from site to site, it was difficult to adjust to a specific value across all models.
In this project, similar to K-means, 2 clusters were considered for all sites.

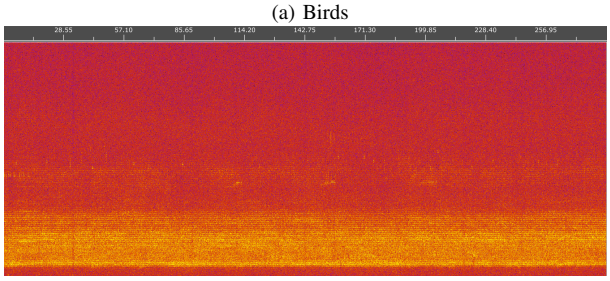
EXPERIMENTS

Dataset Description and Parameter Setting

In this project, we studied the effect of logging on audio recordings collected from rain forests of Borneo, Indonesia. Sounds were recorded from 3 types of fairly distinguishable

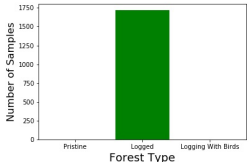


(a) Birds

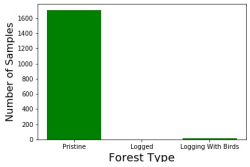


(b) Noise

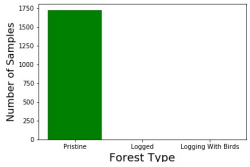
Fig. 3: Spectrograms after auto-clustering using GMM



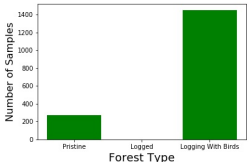
(a) Extensively Logged



(b) Unlogged Forest



(c) UnLogged Forest



(d) Newly Logged

Fig. 4: Determining Type of Forest From Model Match

sites explained below along with the cluster types considered for each:

1) **Site B and C**
   These sites were fairly pristine, and the recordings only contained bird sound, insects and environmental noises. Therefore, the frames were classified into two classes 1- birds, 2- no birds. The spectrogram analysis of these two sites after auto-clustering is shown in Figure 3. As it can be observed, the bird sound frequencies are fairly stable and recognizable through the spectrogram.
2) **Site D**
   This audio is recorded in a healthy forest in which the logging begins halfway through the process. The sound frames contain either only birds or logging along with birds. As it can be observed in Figure 1(b), the bird sounds are not stable and limited to a specific frequency band anymore. This can be due to the immediate impact of logging on birds. The classifier trained on this site categorized a frame as containing 1- only birds, or 2- logging and bird at the same time.
3) **Site A**
   This site has been logged extensively, and after manually analyzing the data as well as looking at the clustering results, it seems clear that most of the birds have left the site, and only logging is heard.

Model Performance Comparison

The following models, both supervised and weakly supervised were considered for the experiments:

- **Supervised Models**
  1) PCA as feature extraction method with 8-12 components, Gaussian classifier trained on hand annotated data and tested with hand annotated data
  2) NMF as feature extraction method with 6-8 components, Gaussian classifier trained on hand annotated data and tested with hand annotated data
- **Weakly Supervised models**
  1) PCA as feature extraction method with 8-10 components, K-means providing large scale training data, Gaussian classifier trained on K-means results and tested with hand annotated data
  2) NMF as feature extraction method with 6-8 components, K-means providing large scale training data, Gaussian classifier trained on K-means results and tested with hand annotated data
  3) PCA as feature extraction method with 8-10 components, GMM providing large scale training data, Gaussian classifier trained on GMM results and tested with hand annotated data
  4) NMF as feature extraction method with 6-8 components, GMM providing large scale training data, Gaussian classifier trained on GMM results and tested with hand annotated data

Results

As it can be observed in figure 5 and table I, for a healthy forest (Site B and C), using Gaussian classifier on hand-annotated data, we achieved the best performance, about 83.4% and 88.3% accuracy using PCA and NMF respectively. For a forest in danger, applying PCA and NMF on hand-labelled data resulted in 71.77% and 77.41% respectively.

The fully supervised model performed accurate enough, although the model wasn't scalable and about 2 hours of data from each site was under experiment. Therefore the next set of experiments were conducted on larger scale data considering weak supervision.

By using weak supervision in training process, we expected to achieve lower accuracy since clustering could not be as precise as hand annotating data. Although through clustering we can train our model on the sound data recorded over 3 months, 4 hours a day which was more valuable for the purpose of this project.

TABLE I: HA: Hand-annotated, fully supervised. Auto-Class: Automatically classified with no supervision. MV: Multi-variate

	Model Accuracy		
Model	Site C	Site D (Birds vs Birds+ Logging)	Site B and C
HA + PCA (8 f) + Gaussian Classifier	89.6%	71.77%	83.4%
HA + NMF(6 features) + Gaussian Classifier	95.45%	77.41%	88.3%
K-means Auto-Class + PCA + Gaussian Classifier	81.37%	50%	79.41%
MV-GMM + NMF + Gaussian Classifier	85%	86.25%	82.5%

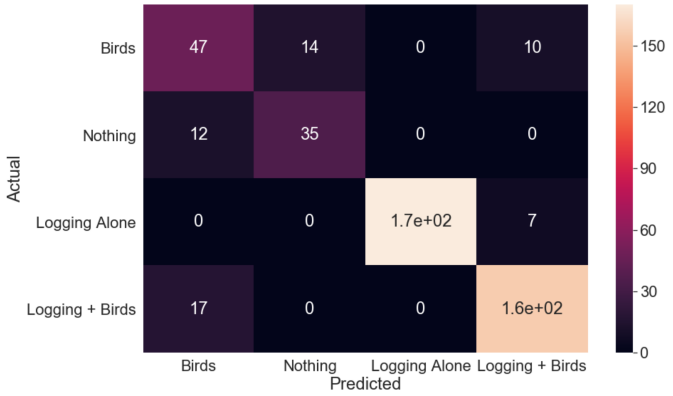


Fig. 5: Confusion Matrix of Gaussian Classification Results, Hand-Annotated training

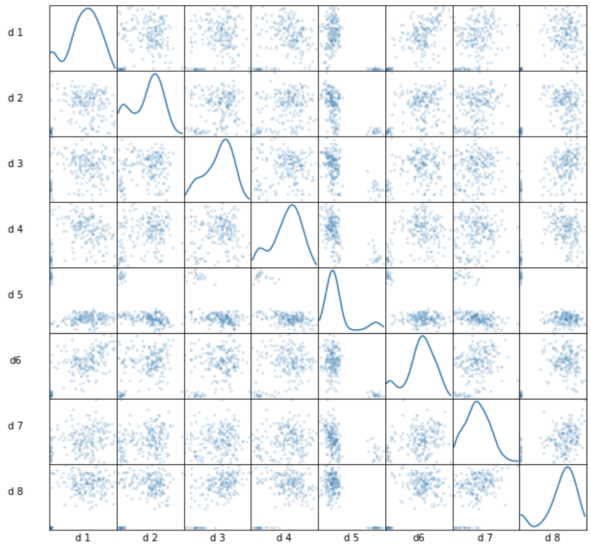


Fig. 6: Correlation between 8 dimensions resulted from NMF on Site D

By using K-means we achieved about 79.41% accuracy for site B and C, and 50% for site D. Due the hard clustering nature of K-means, and having bird sounds in different frequencies, K-means could not perform well enough on site D. As a solution GMM was used for clustering the data, and 86.25% accuracy was achieved. This can be attributed to the fact that the data points do not form uniform spherical clusters for K-means to generate good results but are captured well by an elliptically shaped Gaussian distribution. As shown in figure 6, dimension 5 using NMF captures certain attributes that help define the clusters better. In other dimensions, the data points are overlapping and difficult to segregate. With hard clustering algorithms like K-means, this turns out to be a major problem. Hence, the results with GMM with NMF are much better than K-Means.

As a final validation, we took random samples of data from each of our sites, and per minute, plotted which model (Birds vs Logging Alone vs Birds + Logging) it fit most closely with. From Figure 4, we can see that we are able to very accurately predict the type of forest, with Pristine Forests (Sites B C) matching the Birds model, Extensively Logged Forest (Site A) being recognized, and Mixed Forest (site D) mapping primarily to the Bird+Logging bucket, with some seconds of only birds without logging, which is in line with the data.

This gives us encouragement that the models can be used to acoustically determine the health of a forest ecosystem, and by training it on larger amounts of data in the future, might be a beneficial tool to researches in Indonesian Boreno.

The major drawback of both the clustering methods was visible when there was rain or logging sounds in the background. With overlapping frequency bands, the method struggled to achieve high accuracy.

Conclusion

This research aimed to develop a universal bird model that could detect the presence of birds in diverse forest environments. We are able to robustly detect the presence of birds across sites, with and without external factors like logging, which can help detect when birds stop inhabiting a habitat. We are also able to categorize the type of forest based on their adherence to each of the models. There is more to be done to deploy this in the wild, but we are encouraged by our initial results.

Acknowledgment

Our heartiest thanks to Zuzana Burivavlova of U.Wisconsin, Madison, for generously sharing the data collected for Burivalova, Zuzana, Edward T. Game, and Rhett A. Butler. "The sound of a tropical forest." Science 363.6422 (2019): 28-29.

References

[1] Z. Burivalova, Ç. H. Şekercioğlu, and L. P. Koh, "Thresholds of logging intensity to maintain tropical forest biodiversity," *Current Biology*, vol. 24, no. 16, pp. 1893–1898, 2014.

[2] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[3] A. P. Hill, P. Prince, E. Piña Covarrubias, C. P. Doncaster, J. L. Snaddon, and A. Rogers, "Audiomoth: Evaluation of a smart open acoustic device for monitoring biodiversity and the environment," *Methods in Ecology and Evolution*, vol. 9, no. 5, pp. 1199–1211, 2018.

[4] D. Stowell, M. D. Wood, H. Pamuła, Y. Stylianou, and H. Glotin, "Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge," *Methods in Ecology and Evolution*, vol. 10, no. 3, pp. 368–380, 2019.

[5] A. Sevilla and H. Glotin, "Audio bird classification with inception-v4 extended with time and time-frequency attention mechanisms." in *CLEF (Working Notes)*, 2017.

[6] S. Kahl, T. Wilhelm-Stein, H. Hussein, H. Klinck, D. Kowerko, M. Ritter, and M. Eibl, "Large-scale bird sound classification using convolutional neural networks." in *CLEF (Working Notes)*, 2017.

[7] B. Fazeka, A. Schindler, T. Lidy, and A. Rauber, "A multi-modal deep neural network approach to bird-song identification," *arXiv preprint arXiv:1811.04448*, 2018.

[8] H. Hotelling, "Analysis of a complex of statistical variables into principal components." *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.

[9] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, p. 788, 1999.

[10] F. Weninger, J. L. Roux, J. R. Hershey, and S. Watanabe, "Discriminative nmf and its application to single-channel source separation," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[11] J. A. Arias, J. Pinquier, and R. André-Obrecht, "Evaluation of classification techniques for audio indexing," in *2005 13th European Signal Processing Conference*. IEEE, 2005, pp. 1–4.