Automated audio recording as a means of surveying Tinamous (Tinamidae) in the Peruvian Amazon

Reid Rumelt[1], Arianna Basto[2], and Carla Mere Roncal[3]

[1]Cornell University
[2]Colorado State University
[3]University of Florida

April 14, 2021

Abstract

1. The use of machine learning technologies to process large quantities of remotely-collected audio data is a powerful emerging research tool in ecology and conservation. 2. We applied these methods to a field study of tinamou (Tinamidae) biology in Madre de Dios, Peru, a region expected to have high levels of interspecies competition and niche partitioning as a result of high tinamou alpha diversity. We used autonomous recording units to gather environmental audio over a period of several months at lowland rainforest sites in the Los Amigos Conservation Concession and developed a Convolutional Neural Network-based data processing pipeline to detect tinamou vocalizations in the dataset. 3. The classified acoustic event data are comparable to similar metrics derived from an ongoing camera trapping survey at the same site, and it should be possible to combine the two datasets for future explorations of the target species' niche space parameters. 4. Here we provide an overview of the methodology used in the data collection and processing pipeline, offer general suggestions for processing large amounts of environmental audio data, and demonstrate how data collected in this manner can be used to answer questions about bird biology.

1. INTRODUCTION

Recent reductions in the size and cost of autonomous data collection equipment have allowed ecologists to better and more efficiently survey their study sites (Acevedo &

Villanueva-Rivera, 2006). Much work has been done to examine the benefits of using camera trap networks to detect shy and retiring species whose detection probabilities greatly decrease in the presence of human researchers (O'Connell et al., 2010). However, many species for which remote surveying techniques are optimal are difficult to properly monitor with camera traps due to their small body sizes and / or preference for heavy vegetative cover (Newey et al., 2015). A number of these species, particularly interior forest birds, are much easier to detect via acoustic monitoring techniques due to their frequent, far-carrying vocalizations, and battery-operated automated recording units (ARUs) have recently become a cost-effective option for researchers working with these species (Brandes, 2008). ARUs can operate in the field for much longer periods of time than humans observers can (several days or weeks in many cases), efficiently and safely survey remote areas early in the morning and late at night, and, like camera traps, minimize disturbance to sensitive species. An additional benefit of audio recorders relative to camera traps is that audio recorders have a wider range of detectability than camera traps since they do not require direct line of sight, which increases the area of coverage and the likelihood of detecting rare species. However, in order to use audio recordings from ARUs, vocalizations from the target species must be detected among large quantities of survey audio. Efficiently and reliably identifying these detections presents a major challenge when developing a data-processing pipeline. Although this task is still a non-trivial consideration in developing a study design, recent advancements in machine learning (ML) classification techniques, coupled with dramatic increases in the availability and accessibility of powerful hardware, have made this process easier than ever (Kahl et al., 2019). We strongly believe that the application of machine learning techniques to the processing of large quantities of automated acoustic event detection data will prove to be a transformative development in the fields of ecology and conservation, allowing researchers to tackle biological questions that have previously been impractical to answer.

Several life history characteristics of tinamous (Tinamidae), a group of terrestrial birds that occur widely in the Neotropics, make them superb candidates for field-testing this type of audio processing pipeline. Although a few species in this family occupy open habitats, most show a high affinity for interior forest areas with thick vegetative cover (Bertelli & Tubaro, 2002). They are far more often heard than seen, and some species vocalize prolifically as part of the dawn and dusk choruses (Pérez-Granados et al., 2020). This preference for interior forest, along with their large body sizes and terrestrial nature, makes tinamous inordinately susceptible to the effects of anthropogenic habitat change, both in terms of outright habitat loss and to increased human hunting pressure in fragmented forest patches near populated areas (Thornton et al., 2012). Intensive life history research in the coming years will be critical to conservation of tinamous and their habitats, and autonomous recording has the potential to revolutionize this line of inquiry.

Here we present the preliminary results of an ongoing field study that involves deploying ARUs at lowland Amazonian forest sites in Madre de Dios, Peru. Although this region has tentatively among the highest levels of tinamou alpha diversity in the Neotropics (11 co-occurring species: eBird, 2017), there is currently a lack of research into which biological and ecological factors allow such high degrees of alpha diversity. We collected environmental audio of each day's dawn and dusk choruses and designed a data pipeline that uses a machine learning (ML) audio classifier to identify tinamou vocalization events in the audio data and organize the detections into a spatiotemporal database for future use in producing occupancy models for the target species. To our knowledge, this technology has not previously been used to conduct community-level surveying for tinamous and represents a promising alternative to camera traps and more traditional point-count surveying as a means of studying elusive yet highly vocal bird taxa.

2. MATERIALS AND METHODS

1. Data collection

Data collection was conducted under the auspices of the Amazon Conservation Association at the Los Amigos Conservation Concession (LACC), in the lowland rainforest of Madre de Dios, Peru. This site, which protects ˜145,000 ha of forest along the Rio Los Amigos basin, is one of the most biodiverse lowland rainforest sites in the Amazon basin with close to 600 bird species, eleven of which are tinamous in the genera *Tinamus* and *Crypturellus* (eBird, 2017; Table 1). The station's biological diversity is due in part to its diversity of terrestrial microhabitats, which include terra firme and floodplain primary forest, secondary and edge forest, Guadua bamboo stands, and Mauritia flexuosa palm swamps (Larsen et al., 2006). As studies at this site (Mere Roncal et al., 2019) and elsewhere in the Neotropics have demonstrated that tinamou species differ in their specific habitat utilization characteristics (Guerta & Cintra, 2014), LACC is an exemplary site for detecting tinamous across a variety of habitat gradients.

Acoustic monitoring was conducted using ten SWIFT ARUs (Kahl et al., 2019), provided by the Cornell Lab of Ornithology, from mid-July to early October of 2019. This period overlaps with the latter half of the dry season at LACC. The SWIFT units were deployed on rotating 14 day deployment periods at terra firme and floodplain forest sites (Figure 1, S1), 10 sites at a time, over three deployments from mid-July to late August. A fourth deployment, duration 27 days, was conducted as a follow-up at five of the 30 sites from late September to early October. As the chosen sites are part of the station's existing camera trap system (approximately a 1 km^2 grid located along the edge of open trails), we were able to merge our detection set with previously-collected site-level habitat data as well as to compare our tinamou detection rates to those calculated using camera trap detections. Recorders were tied to trees at a height of approximately 1.5 m from the ground with the microphone facing downwards. Each unit was programmed to record for five hours a day, from 5:00 to 7:30 in the morning and 16:00 to 18:30 in the afternoon to early evening, in order to cover periods of high vocal activity for tinamous (Dias et al., 2016). The SWIFT unit firmware allows for control of microphone gain and sampling frequency; we set these values to -33 dB (the default) and 16 kHz, respectively. Setting the sampling frequency to 16 kHz is a tradeoff that limits the acoustic frequency bandwidth to 0-8 kHz (Landau, 1967) in exchange for smaller file sizes and lower power demands than the default value of 32 kHz. The SWIFT firmware writes data as 30 min-long WAV files (˜58 MB). Each unit was intended to collect data for the shorter of (a) the entire 14 or 27 day recording period or (b) until battery power was exhausted. In practice, battery life was always the limiting factor, with a mean time-to-shutdown of 7.81 days (5.12 days for deployments 1-3 and 21.8 days for deployment 4). Due to supply limitations, we were forced to use a different brand of battery for deployments 1-3 than for deployment 4, which we suspect is at least partially responsible for the longer per-recorder run times in the latter deployment. At the end of each deployment period, all units were removed from the field, loaded with fresh recording media and batteries, and deployed to their next assigned site on the following day. All audio data was backed up to rugged solid state storage media for transport out of the field.

Our chosen classification procedure is a type of supervised machine learning, which requires a significant amount of training audio to produce a working model (Kotsiantis et al., 2007). We used a set of ˜3100 audio files of 2s duration (the typical phrase length in tinamou calls) to train an initial classifier. These files were coded as one of twelve classes: one class for each tinamou species, and a "junk" class containing audio of other bird species, non-bird organismal audio, and assorted environmental audio (Table 1). The training dataset was derived from audio downloaded from the Macaulay Library of Natural Sounds (https://macaulaylibrary.org) and Xeno-Canto (http://www.xeno-canto.org) databases (S2) as well as from exemplar cuts in the audio we collected in the field.

2. Data processing and classification

A series of preprocessing steps were applied to the audio after collection, beginning with normalizing all survey audio to -2 dB maximum gain. The SWIFT recorder firmware writes a high amplitude audio spike at the beginning of the first file recorded after the unit wakes from standby (e.g., the beginning of the 5:00 and 16:00 audio files); therefore, we chose to overwrite the first five seconds of audio on each of these files to prevent this spike from impacting the gain normalization step. As our chosen audio classifier architecture operates on fixed-length samples, we split each 30 min audio file into 7197 2s-long overlapping audio "windows" that advance forward by 0.25 seconds per window. The classifier operates on the log-Mel-weighted spectrogram (Knight et al. 2017) of each window, which is created dynamically during classification using STFT utilities in the TensorFlow python module (Table 2) at a native resolution of 512x512 px.

Audio event detection was conducted using a set of Convolutional Neural Network classifiers. The chosen classifier architecture is adapted from the multiclass single-label classifier called "Model 1" in Kahl et al. (2017). Our decision to use a multiclass single-label classifier architecture was driven by a desire for reduced learning complexity; however we feel there is merit to introducing a multilabel classifier in future analyses as existing ML techniques are capable of dealing with this task with minor modifications (Kahl et al., 2017). For similar reasons, we reduced the number of neurons per hidden layer by half to account for limitations in available processing power, and also down-sampled the 512x512 px spectrogram images to 256x256 px before training and classification. The full classifier architecture is described in Table 3. All data processing was performed either in Python, using a combination of TensorFlow 2.0 (Abadi et al. 2016) and other widely-used Python modules, or, in the case of later statistical testing, in R (R Core Team, 2019). During training, we applied the same STFT algorithm as used for the survey data to dynamically convert the training audio to log-Mel-weighted spectrograms, and implemented data augmentation to improve model generalization (Ding et al., 2016). These augmentation parameters, along with general model hyperparameters (Table 4), were chosen using a Bayesian hyperparameter search module in Python (Nogueira, 2014) that was driven to optimize the calculated multiclass F1-score (Sokolova & Lapalme, 2009) ($\beta = 1$) on a set of known good clips (hereafter the "validation set") created using a sample of clips not used in the training data (Table 1). F1-score was calculated as a macro-average of the 12 classes in order to give equal weight to rare classes.

Although the goal of hyperparameter search techniques is typically to identify an optimal set of parameters, we observed two apparent local optima that we chose to incorporate into our classification pipeline as two submodels: a. submodel 1, which added artificial gaussian noise to training spectrograms as part of the augmentation process and b. submodel 2, which did not. The set of class probabilities returned for each clip was the mean of the probabilities reported by the two models (hereafter the "ensemble"). We validated each submodel, as well as the ensemble, on the same validation set used in hyperparameter search.

When deployed on survey data, our classification pipeline yields classifications as a sequence of probability vectors of size 12, where each vector corresponds to one window in the sequence of overlapping windows. Raw class probabilities for windows that contain only the very beginning or the very end of tinamou vocalizations are often classified incorrectly, which we believe results from the fact that different tinamou species often share structural similarities with one another in those regions of their vocalizations. To reduce the impact of this pattern on our overall classification accuracy, we applied a "smoothing" post-processing to the class probabilities where each probability value was replaced by the weighted average of that value (weight = 1) and the values immediately before and after it in the time sequence (weight = 0.5). Windows with a maximum class probability < 0.85 were removed, and the remainder assigned the label with the highest class probability. All windows detected as positive were manually checked for accuracy and relabeled if incorrect.

We assessed the degree of marginal improvement in classifier performance due to increased training dataset size and increased structural uniformity between training clips and survey audio by running a second "pass" of the acoustic classifier on the survey data with a set of models that had been trained using a larger training dataset. To generate this dataset, the original training dataset was supplemented with all known good positive windows from the initial classification (the first "pass"). We sampled from this dataset to produce a new training dataset (n = 18,480) with the larger of 2000 randomly selected clips (4000 for the "junk" class), or as many clips as were available, per class (Table 1). We trained new submodels on this data using the same model architecture and hyperparameters that were used for models in the first pass. The sole change made to the training process between classifications 1 and 2 was to alter the batch generation code to produce batches with balanced class frequencies to offset the greatly increased degree of class imbalance in the supplemented dataset. Each submodel was validated using a new validation set that contained known-good survey audio whenever possible in order to ensure that the calculated metrics would be more indicative of each submodel's real world performance.

The survey data was classified with these new models, and the resulting class predictions were processed to extract probable detections as described previously. In order to decrease labor time, all positive windows from the initial classification were "grandfathered in" as correctly identified due to having been manually checked previously, which allowed us to only check positive detections that were newly identified during the second pass. Finally, all sequences of windows with a particular species classification that were >= 0.75s apart from any other sequence were grouped as a single vocal event.

For the purposes of quantifying model performance and generalizability, we calculated a precision, recall, F1-score, and precision-recall area under the curve (AUC) performance metrics for the primary and secondary models, presented on a per-class basis or as macro-averages across classes, after Sokolova & Lapalme (2009). All metrics were calculated based on classifier performance on a set of known good clips (hereafter the "validation set"), using data from the survey audio whenever possible in order to ensure that the performance metrics would be more indicative of each submodel's real world performance.

As a point of comparison for our audio detection counts, we also examined community science observation data for tinamous from eBird (Sullivan et al., 2009; Sullivan et al. 2015). We used stationary and traveling checklists containing tinamous that were submitted at the LACC hotspot between the months of July and October, removing stationary checklists with durations > 150 min and traveling checklists with lengths > 0.5 km in order to constrain the sampling effort parameter space of the eBird data such that it was more comparable to our 2.5 h morning and afternoon recording periods. Despite these filtering steps, the final eBird dataset still contained all locally-occurring tinamou species. However, it was clear that our acoustic data density for *C. strigulosus* vastly outstripped eBird data density, so we excluded this species from our analysis as we feel it warrants separate discussion. We produced estimates of occurrence probabilities by averaging the results of random samples from the eBird data (n = 1,000) and averaging the results of the same number of samples from the acoustic event dataset using the same underlying sampling effort density distribution as the eBird checklist durations. Audio frequency estimates were calculated separately for terra firme and floodplain habitat types on a site-level presence-absence frequencies and then averaged. In addition, we compared our audio detection counts to camera trap capture rates reported by Mere Roncal et al. (2019), also at LACC. Camera trap capture rates suggest seasonally-driven differences in tinamou activity rates, so we only considered detection rates from the dry season, which limited our comparison to the five tinamou species reported by Mere Roncal et al. (2019) for which dry season camera trap data is available. Occurrence frequencies were again calculated as the average of the distributions from terra firme and floodplain sites.

3. RESULTS

1. Model performance

The performance of all models is summarized in Figures 2-4 and Table 5-6. At the macro-averaged level, the ensemble model performed better than either submodel individually within each classification pass (Table 5). The addition of random artificial noise in submodel 2 improved both precision and recall in pass 1, though only recall in pass 2 (Table 5). The ensemble model of pass 2 performed substantially better than the corresponding model in pass 1 (Figure 2), likely both due to the larger training dataset used for this pass and the fact that the training and validation datasets used during this pass were both comprised of audio collected by us in the field, thus being more similar to one another than they were in pass 1. This increased similarity between training and validation datasets in pass 2 is also a potential explanation for the observed decrease in recall score with added artificial noise during this pass, though we did not perform further analysis of this specific result. Per-class performance was generally good, with visible improvements from pass 1 to pass 2 in most, though species with subjectively more variable vocalizations (e.g. *T. major*) performed less well (Figure 3, Table 6). Intriguingly the increase in classification accuracy we observed at the macro-averaged level did not hold uniformly true at a class level, with submodel 1 or 2 often yielding better results (Table 6). An analysis of classifier score distributions for positive detections showed increased score separation between true positive and false positive detections in pass 2 relative to pass 1 (Figure 4), indicating better overall predictive power in the case of the latter model (Knight et al. 2017). We also observed that our chosen score threshold yielded precision and recall values that were close to the inflection point of the precision-recall curve, indicating this value was an appropriate choice for ensuring a good balance of the two metrics.

2. Collection data and ecological analyses

We collected a total of 1216.5 h of audio, of which 544.5 hs (45%) came from deployment 4 and 225.5 (19%), 201.0 (17%), and 245.5 hs (20%) came from deployments 1, 2, and 3, respectively (S3). The total number of recording hours per habitat type (899.0 hs in terra firme, 317.5 hs in floodplain) was roughly proportional to the number of site-deployment combinations in each habitat type (24 vs 13). We detected a total of 15,891 tinamou vocalization events, 2,189 of which were added after the second classification pass. Our detections represent nine of the 11 species present at LACC, with data densities ranging from 4,468 detections for *C. strigulosus* to 26 for *T. tao* (S3). Two species were not detected: *C. atricapillus* and *C. obsoletus* . Both species are uncommon at Los Amigos (eBird, 2017; personal obs.), are known to have affinities for brushy edge habitats that were located away from most of the recorders (Cabot et al., 2020; Anjos, 2006), and were entirely absent from the camera trap dataset. Therefore, we suspect that their lack of detection indicates true absence from the dataset rather than poor class performance. The relative occurrence frequency of the tinamou species as measured by our audio detection pipeline differs significantly from the observation frequencies reported by eBird ($\chi^2 = 567.4$, $p < 2.2e\text{-}16$, Figure 5b), but notably there was no significant difference between these frequencies and camera trap capture rates for the five species represented in both datasets ($\chi^2 = 0.037102$, $p > 0.1$, Figure 5a).

4. DISCUSSION

The machine learning pipeline we used in this study appears to be an effective tool for collecting occurrence data across a range of habitat types at our target site. The lack of statistically significant differences in relative detection frequencies between the audio and camera trap data conflicted slightly with our expectation that acoustic sampling would yield more accurate occurrence metrics than camera trap sampling. However, the sample size of our audio dataset was much smaller than that collected by the camera trap network over a roughly similar period of time (n = 122 before filtering for season), which suggests that acoustic monitoring is capable of yielding much higher data densities per unit surveying time, at least for vocal species. Similarly, increasing the sample size of the camera trap dataset and collecting audio samples from the wet season may yet allow us to identify true underlying differences in detection probabilities for tinamous when surveyed acoustically versus visually.

The significant differences in detection frequency we observed between our data and the eBird data is likely a result of non-random spatial sampling. An example of this spatial non-randomness with a clear causative explanation is the relatively higher eBird detection frequency for *C. undulatus* , a species that is present widely in floodplain and transitional forest but is also extremely common in edge habitat near the station dwellings where ecotourists and birders visiting the station spend time when not hiking on trails (eBird, 2017; personal obs). We chose not to include *C. strigulosus* in frequency analyses as it is represented in our audio dataset mainly by detections at sites east of the Río Los Amigos that birders and ecotourists visiting the station are rarely if ever able to access, therefore heavily limiting its sampling density in the eBird dataset (personal obs). However, even in the absence of quantitative assessment, we nonetheless believe this is another clear case of spatially non-random eBird sampling patterns relative to the more structured audio and camera trap data. We therefore advise caution when using eBird data to generate site-level relative occurrence frequencies for tropical forest birds, as doing so properly requires a substantially better-informed set of sample bias corrections than we chose to use for this illustratively naïve approach. eBird's own Status and Trends methods are a classic example of how this can be done analytically, though the relatively low eBird data density across the Neotropics has meant that analyses using these methods have mainly been focused on the temperate zone (Sullivan et al., 2009; Sullivan et al., 2014; Fink et al., 2018). Employing study designs that use eBird data as an adjunct to more structured surveying techniques is another possible strategy (Reich et al., 2018), as this strategy reduces the proportion of overall bias due to eBird on ecological modeling efforts in this region while retaining the benefits of using multiple independent datasets to address the same question.

A common question posed by research scientists in the pursuit of an efficient but effective machine learning platform is "how much training data is enough data." Our two-pass classification strategy demonstrated clear classification accuracy improvements over a single pass, though the degree to which our ensemble modeling strategy improved classification performance varied substantially between classes. We suspect that most of the performance improvements that could be gained beyond what we saw in our analysis would come from gathering additional survey data, iterating the data collection and training processes to increase sample sizes, and further improving the model architecture and hyperparameters. It is important to note that the main limiting factor for our use of machine learning classification has been the amount of computational power available to us, which required us to decrease the complexity of our neural networks and the resolution of our spectrograms relative to those mentioned in the literature (Knight et al., 2017; Kahl et al., 2019). While doing so allowed us to produce classification results within acceptable time constraints, this speed benefit potentially came at the cost of reduced classification accuracy. An important future goal for our analyses is to securing sufficient computational power to run the classification at full resolution to quantify improvements in accuracy, as we strongly believe that understanding the minimum acceptable resolution necessary to achieve a given level of accuracy is a crucial logistical consideration for researchers seeking to build hardware systems to support similar data processing pipelines.

Acoustic monitoring represents a promising method for studying bird biology and life history. We are particularly excited by the prospect of being able to use this SWIFT survey data in future analyses to identify the life history and microhabitat characteristics that result in niche partitioning in the tinamou community of lowland Madre de Dios. We anticipate that additional data collection, particularly during the wet season, and further refinement of this machine learning pipeline will allow us to build occupancy models for these species using elevation maps and vegetation structure datasets that were collected for use with the associate camera trap grid as environmental covariates (Royle & Nichols, 2003).

5. ACKNOWLEDGEMENTS

This study was funded by a Franzen Fellowship grant from the Amazon Conservation Association. We thank the staff and *promotores* at Los Amigos Biological Station and Los Amigos Conservation Concession, respectively, for helping to ensure the success of our field work. RBR would especially like to thank *promotore* Don Hernan, for his sense of humor despite numerous language barriers and difficult working conditions. Special thanks goes to Bioacoustics at the Cornell Lab of Ornithology for their wonderful technical support and assistance with acquiring recording equipment, and to Robert Guralnick, for allowing us to use his personal SD cards. We are further indebted to Larry Chen, Andrew Dreelin, and Izkak Shafran for their editorial assistance with the manuscript, and to Adrian Forsyth, without whose support this work would not have been possible.

6. AUTHOR CONTRIBUTIONS

RBR, AB, and CMR developed sampling protocol, site map, and recorder rotation schedule using insights from an existing project being conducted by CMR. RBR and AB deployed and retrieved recorders (RBR exclusively for deployment 2, AB exclusively for deployment 4, joint effort for deployments 1 and 3). RBR constructed data processing pipeline and developed occurrence dataset. All authors contributed to writing and editing of manuscript as well as approving final draft for submission.

7. DATA AVAILABILITY

All training and validation data is publicly available from Xeno-Canto (https://www.xeno-canto.org) and (upon request) from the Macaulay Library of Natural Sounds (https://www.macaulaylibrary.org); all audio ID numbers are listed in supporting information document S2. Acoustic event database will be archived in the Dryad Digital Repository.

8. REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., … & Ghemawat, S. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467.

Acevedo, M. A., & VILLANUEVA-RIVERA, L. J. (2006). From the field: Using automated digital recording systems as effective tools for the monitoring of birds and amphibians. *Wildlife Society Bulletin* , *34* (1), 211-214.

Anjos, L. D. (2006). Bird Species Sensitivity in a Fragmented Landscape of the Atlantic Forest in Southern Brazil 1. Biotropica: The Journal of Biology and Conservation, 38(2), 229-234.

Bertelli, S., & Tubaro, P. L. 2002. Body mass and habitat correlates of song structure in a primitive group of birds. *biological Journal of the Linnean Society* , *77* (4), 423-430.

Brandes, T. S. (2008). Automated sound recording and analysis techniques for bird surveys and conservation. Bird Conservation International, 18(S1), S163-S173.

Cabot, J., D. A. Christie, F. Jutglar, P. F. D. Boesman, and C.J. Sharpe (2020). Black-capped Tinamou (*Crypturellus atrocapillus*), version 1.0. In Birds of the World (J. del Hoyo, A. Elliott, J. Sargatal, D. A. Christie, and E. de Juana, Editors). Cornell Lab of Ornithology, Ithaca, NY, USA. https://doi.org/10.2173/bow.blctin1.01

Dias, L. C. S., Bernardo, C. S. S., & Srbek-Araujo, A. C. (2016). Daily and seasonal activity patterns of the Solitary Tinamou (*Tinamus solitarius*) in the Atlantic Forest of southeastern Brazil. The Wilson Journal of Ornithology, 128(4), 885-894.

Ding, J., Chen, B., Liu, H., & Huang, M. (2016). Convolutional neural network with data augmentation for SAR target recognition. IEEE Geoscience and remote sensing letters, 13(3), 364-368.

eBird (2017). eBird: An online database of bird distribution and abundance [web application]. eBird, Cornell Lab of Ornithology, Ithaca, New York. Available: http://www.ebird.org. (Accessed: March 3, 2020).

Fink, D., T. Auer, A. Johnston, M. Strimas-Mackey, O. Robinson, S. Ligocki, B. Petersen, C. Wood, I. Davies, B. Sullivan, M. Iliff, S. Kelling. 2020. eBird Status and Trends, Data Version: 2018; Released: 2020. Cornell Lab of Ornithology, Ithaca, New York. https://doi.org/10.2173/ebirdst.2018

Guerta, R., & Cintra, R. (2014). Effects of habitat structure on the spatial distribution of two species of Tinamous (Aves: Tinamidae) in a Amazon terra-firme forest. Ornitol Neotrop, 25(1), 73-86.

Kahl, S., Wilhelm-Stein, T., Hussein, H., Klinck, H., Kowerko, D., Ritter, M., & Eibl, M. (2017, September). Large-Scale Bird Sound Classification using Convolutional Neural Networks. In CLEF (Working Notes).

Kahl, S., Stoter, F. R., Goeau, H., Glotin, H., Planque, R., Vellinga, W. P., & Joly, A. (2019, September). Overview of birdclef 2019: Large-scale bird recognition in soundscapes.

Katz, J., Hafner, S. D., & Donovan, T. (2016). Assessment of error rates in acoustic monitoring with the R package monitoR. Bioacoustics, 25(2), 177-196.

Knight, E., Hannah, K., Foley, G., Scott, C., Brigham, R., & Bayne, E. (2017). Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs. Avian Conservation and Ecology, 12(2).

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 160, 3-24.

Landau, H. J. (1967). Sampling, data transmission, and the Nyquist rate. Proceedings of the IEEE, 55(10), 1701-1706.

Larsen, T. H., Lopera, A., & Forsyth, A. (2006). Extreme trophic and habitat specialization by Peruvian dung beetles (Coleoptera: Scarabaeidae: Scarabaeinae). The Coleopterists Bulletin, 60(4), 315-324.

Mere Roncal, C., Middendorf, E., Forsyth, A., Caceres, A., Blake, J. G., Almeyda Zambrano, A. M., & Broadbent, E. N. (2019). Assemblage structure and dynamics of terrestrial birds in the southwest Amazon: a camera-trap case study. Journal of Field Ornithology, 90(3), 203-214.

Newey, S., Davidson, P., Nazir, S., Fairhurst, G., Verdicchio, F., Irvine, R. J., & van der Wal, R. (2015). Limitations of recreational camera traps for wildlife management and conservation research: A practitioner's perspective. *Ambio* , *44* (4), 624-635.

Nogueira, F. (2014). Bayesian Optimization: Open source constrained global optimization tool for Python.

O'Connell, A. F., Nichols, J. D., & Karanth, K. U. (Eds.). (2010). Camera traps in animal ecology: methods and analyses. Springer Science & Business Media.

Perez-Granados, C., Schuchmann, K. L., & Marques, M. I. (2020). Vocal behavior of the Undulated Tinamou (*Crypturellus undulatus*) over an annual cycle in the Brazilian Pantanal: New ecological information. Biotropica, 52(1), 165-171.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reich, B. J., Pacifici, K., & Stallings, J. W. (2018). Integrating auxiliary data in optimal spatial design for species distribution modelling. Methods in Ecology and Evolution, 9(6), 1626-1637.

Royle, J. A., & Nichols, J. D. (2003). Estimating abundance from repeated presence–absence data or point counts. Ecology, 84(3), 777-790.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information processing & management, 45(4), 427-437.

Sullivan, B.L., C.L. Wood, M.J. Iliff, R.E. Bonney, D. Fink, and S. Kelling. 2009. eBird: a citizen-based bird observation network in the biological sciences. Biological Conservation 142: 2282-2292.

Sullivan, B. L., Aycrigg, J. L., Barry, J. H., Bonney, R. E., Bruns, N., Cooper, C. B., . . . & Fink, D. (2014). The eBird enterprise: an integrated approach to development and application of citizen science. Biological Conservation, 169, 31-40.

Thornton, D. H., Branch, L. C., & Sunquist, M. E. (2012). Response of large galliforms and tinamous (Cracidae, Phasianidae, Tinamidae) to habitat loss and fragmentation in northern Guatemala. Oryx, 46(4), 567-576.

9. FIGURES AND TABLES

Figure 1: Map of survey points (subset of Los Amigos camera trap network). Group A: deployment 1; group B: deployments 2 & 4; group C: deployment 3.

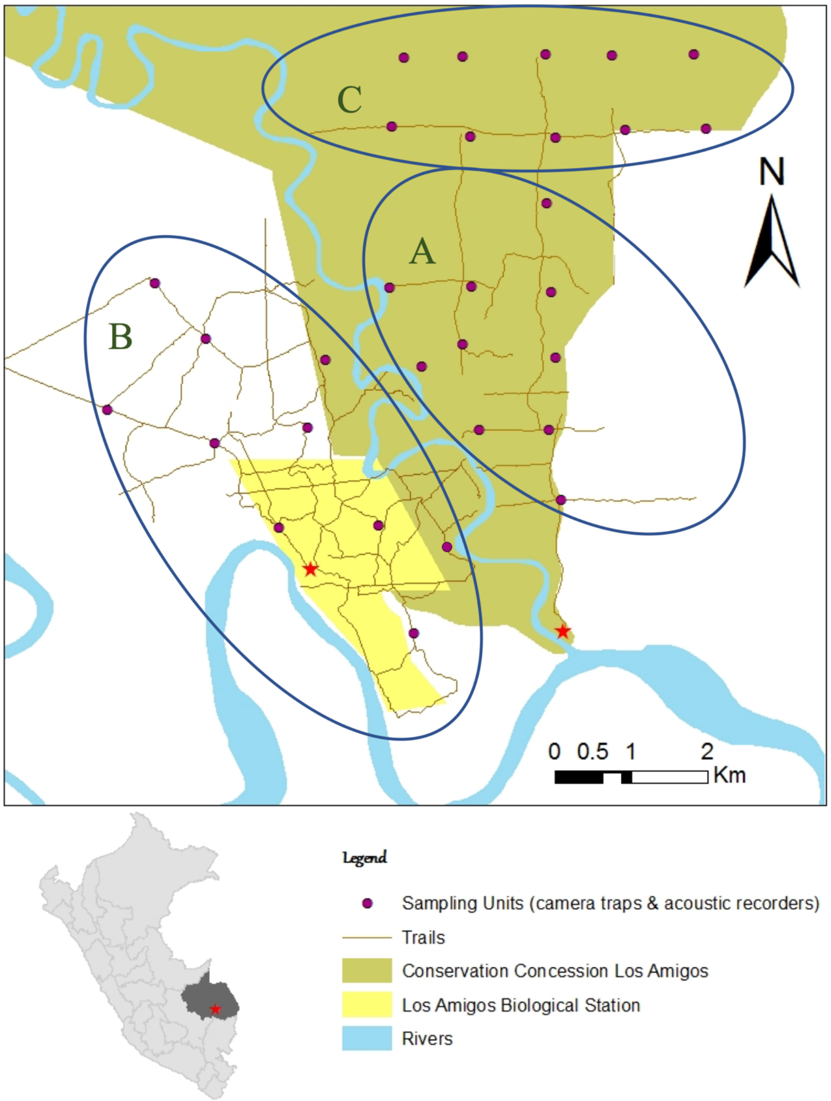
C
A
B
N
0 0.5 1 2
Km
Legend
Sampling Units (camera traps & acoustic recorders)
Trails
Conservation Concession Los Amigos
Los Amigos Biological Station
Rivers

Figure 2: Precision-recall curves for submodels 1 (solid line), 2 (dashed line), and ensemble (dotted line) for classification passes 1 (gray) and 2 (black), macro-averaged metrics. Dark blue lines indicate recall and precision measured at the chosen probability value for positive detections (p = 0.85).

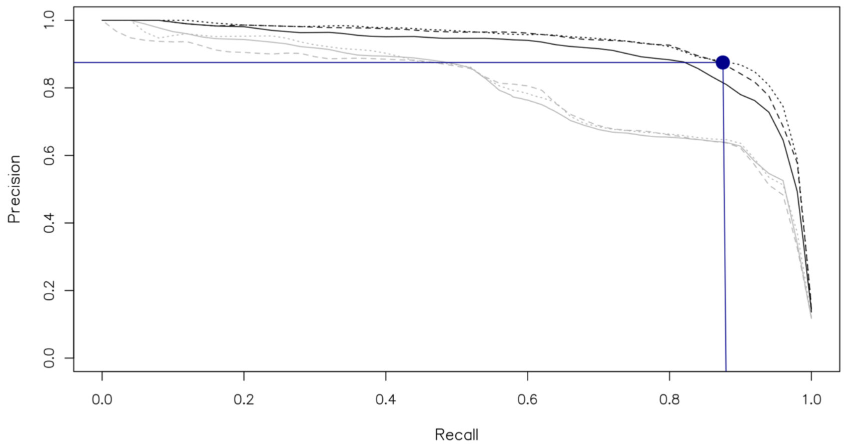

Figure 3: Precision-recall curves for submodels 1 (solid line), 2 (dashed line), and ensemble (dotted line) for classification passes 1 (gray) and 2 (black), per class.

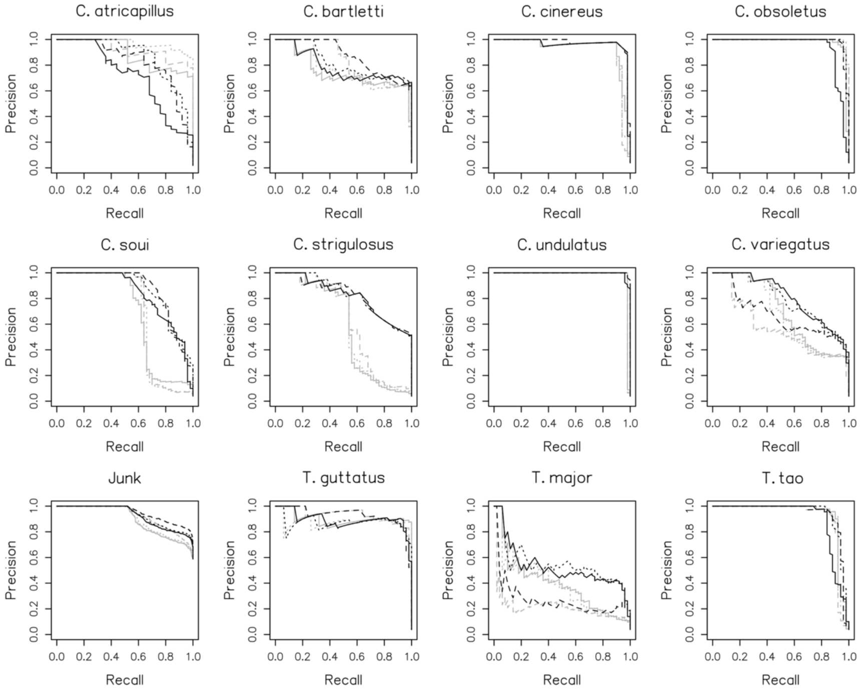

Figure 4: Relative density of classifier true positive (TP) and false positive (FP) detections on the validation set by submodel and training pass. “Final, Ensemble” (lower right) was the model used to classify survey data.

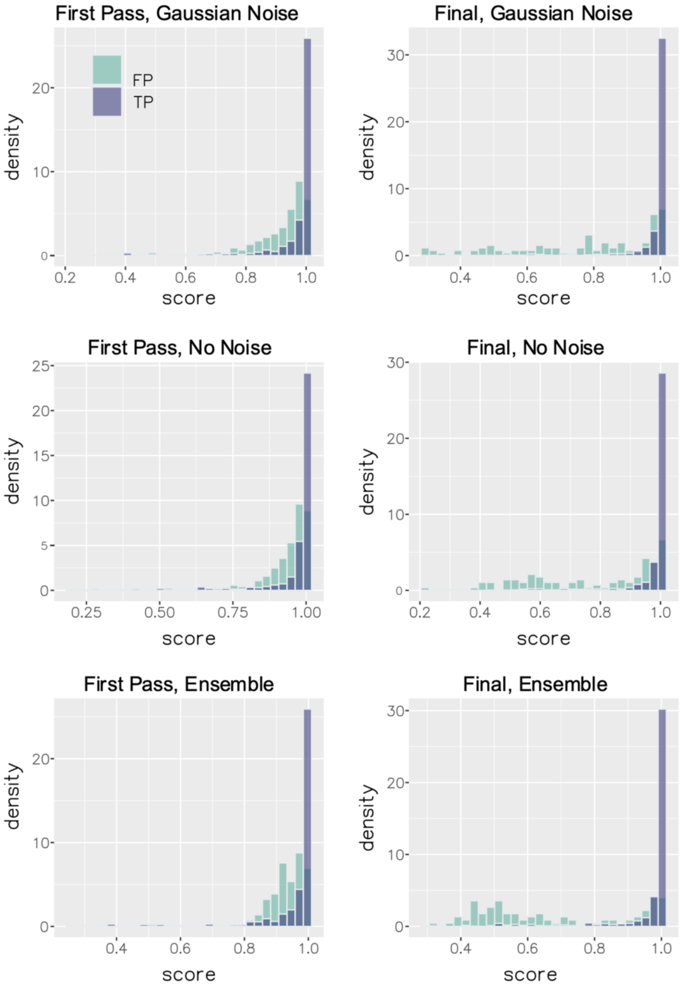

Figure 5: Relative audio event detection frequency versus relative detection frequencies from eBird and camera trap data

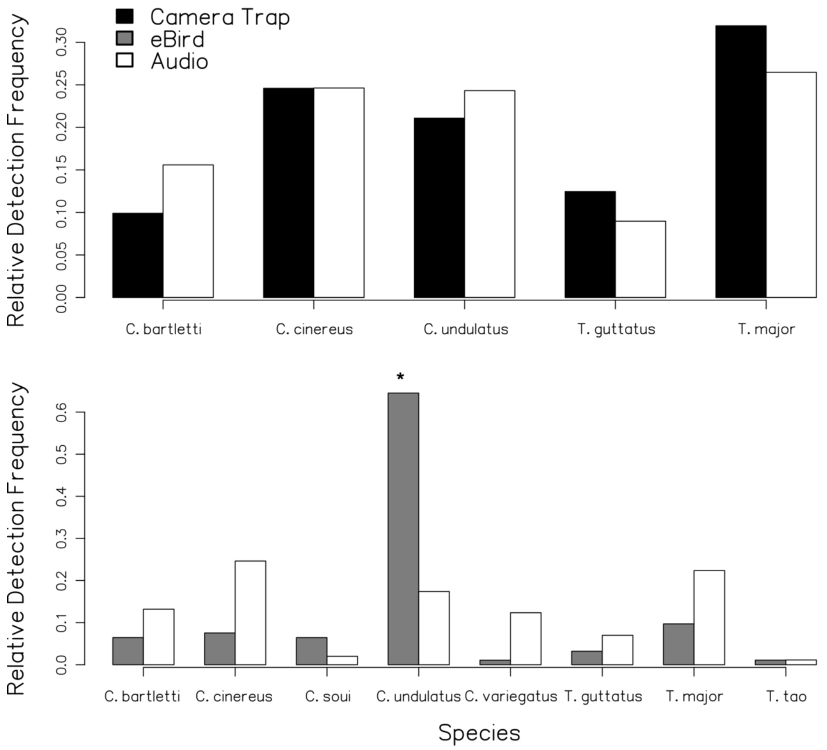

Table 1: List of tinamou species at Los Amigos Biological Station, with source audio totals (first classification totals / second classification totals).

Common name	Scientific name	Training audio	Validation audio
Gray Tinamou	*Tinamus tao*	370/419	50/50
Great Tinamou	*Tinamus major*	252/2000	50/50
White-throated Tinamou	*Tinamus guttatus*	263/1266	50/50
Cinereous Tinamou	*Crypturellus cinereus*	252/2000	50/50
Little Tinamou	*Crypturellus soui*	276/461	50/50
Undulated Tinamou	*Crypturellus undulatus*	311/2000	50/50
Brown Tinamou	*Crypturellus obsoletus*	255/255	50/50
Brazilian Tinamou	*Crypturellus strigulosus*	242/2000	50/50
Black-capped Tinamou	*Crypturellus atrocapillus*	79/79	25/25
Variegated Tinamou	*Crypturellus variegatus*	200/2000	50/50
Bartlett's Tinamou	*Crypturellus bartletti*	320/2000	44/50
Non-tinamou audio	—	294/4000	750/750

Table 2: STFT settings

Setting	Value
Dimensions	512x512 px

Setting	**Value**
Channels	1 (grayscale)
Window size	1024
Stride	64
Frequency band	0 Hz – 8000 Hz
Sampling rate	16 kHz
Mel bin number	256

Table 3: CNN architecture. L2 kernel and activity regularization (1e-06, with default biases turned off) were applied to each Conv2D layer, with batch normalization (momentum = 0.01) applied between the Conv2D and MaxPooling2D layers. ReLU activation was used for all Conv2D layers and the first Dense layer, with Softmax activation applied to the output layer.

Layer Type	**Details**
Input	Size 256x256x1
Conv2D	Size 32x7x7, Stride 2
MaxPooling2D	Size 2
Conv2D	Size 64x5x5, Stride 1
MaxPooling2D	Size 2
Conv2D	Size 128x3x3, Stride 1
MaxPooling2D	Size 2
Conv2D	Size 256x3x3, Stride 1
MaxPooling2D	Size 2
Conv2D	Size 512x3x3, Stride 1
Flatten	(None)
Dense	256 Units
Dropout	0.5
Dense	4 Units

Table 4: Ranges and chosen values for hyperparameters and augmentation parameters. Single value ranges indicate that the parameter was held constant. Zero-values for continuous parameters indicate that the parameter was not used. Values that are ranges indicate that values were randomly chosen from this range on a per-spectrogram basis (augmentation only).

Parameter	**Type**	**Range**	**Value**
Batch size	Hyperparameter	[16, 32, 64, 128]	64
Dropout	Hyperparameter	[0.2, 0.5]	0.5
Epochs	Hyperparameter	20	20
L2 amount	Hyperparameter	[1e-6, 1e-2]	1e-6
Learning rate	Hyperparameter	[1e-9, 1e-2]	0.0075
Network size scale	Hyperparameter	[1, 2, 3, 4]	2
Gaussian noise intensity	Augmentation	[0, 20]	0.8
Gaussian blur	Augmentation	[0, 3]	0
Horizontal shift	Augmentation	[0, 50 px]	[0, 20 px]
Random dB offset	Augmentation	[0, -40 dB]	[0, -40 dB]
Vertical shift	Augmentation	[0, 5 px]	[0, 2 px]

Table 5: Overall classifier performance (after Sokolov and Lapalme 2007 and Knight et al. 2017). prAUC = Precision-Recall AUC

Classification pass	**Submodel**	**Precision**	**Recall**	**F1-Score**	**prAUC**
1	*No noise*	0.777	0.803	0.735	0.797
	Artificial noise	0.783	0.827	0.754	0.788
	Ensemble	0.803	0.845	0.773	0.803
2	*No noise*	0.878	0.846	0.820	0.909
	Artificial noise	0.853	0.890	0.856	0.933
	Ensemble	0.875	0.882	0.861	0.938

Table 6: prAUC scores for each submodel type, per class. Filled-in cells indicate the model with the maximum prAUC value for each class within the two passes.

Classification pass	**Species**	**No noise**	**Artificial Noise**	**Ensemble**
1	*C. atricapillus*	0.891	0.921	0.968
	C. bartletti	0.748	0.826	0.770
	C. cinereus	0.934	0.930	0.926
	C. obsoletus	0.981	0.992	0.996
	C. soui	0.680	0.683	0.684
	C. strigulosus	0.577	0.614	0.597
	C. undulatus	0.982	0.981	0.981
	C. variegatus	0.703	0.576	0.671
	Junk	0.896	0.925	0.902
	T. guttatus	0.903	0.931	0.882
	T. major	0.386	0.216	0.387
	T. tao	0.933	0.915	0.932
2	*C. atricapillus*	0.536	0.705	0.725
	C. bartletti	0.995	0.964	0.994
	C. cinereus	0.981	0.984	0.982
	C. obsoletus	0.928	0.985	0.983
	C. soui	0.886	0.918	0.912
	C. strigulosus	0.965	0.960	0.965
	C. undulatus	0.996	0.999	0.999
	C. variegatus	0.919	0.939	0.938
	Junk	0.991	0.995	0.995
	T. guttatus	0.953	0.951	0.955
	T. major	0.913	0.899	0.920
	T. tao	0.892	0.944	0.936

Table 7: Confusion matrix for ensemble validation. Columns are true species labels, rows are predicted species labels. Species with asterisks were ultimately not detected in the survey data.

	C. atricapillus	C. bartletti	C. cinereus	C. obsoletus	C. soui	C. strigulosus	C. undulatus	C. va
C. atricapillus*	6	0	0	2	4	0	0	2
C. bartletti	0	49	1	0	0	0	0	0
C. cinereus	0	1	48	0	0	0	0	0
C. obsoletus*	0	0	0	45	0	0	0	0

	C. atricapillus	C. bartletti	C. cinereus	C. obsoletus	C. soui	C. strigulosus	C. undulatus	C. va
C. soui	1	0	0	0	41	0	0	0
C. strigulosus	0	0	0	0	0	50	0	0
C. undulatus	0	0	0	0	0	0	49	0
C. variegatus	0	0	0	0	0	0	0	47
Junk	0	3	0	0	3	8	1	12
T. guttatus	0	0	0	0	1	0	0	0
T. major	0	0	0	0	0	0	0	1
T. tao	0	0	0	0	0	1	0	0