ECOLOGY LETTERS

Ecology Letters, (2018) **21**: 1244–1254 doi: 10.1111/ele.13092

METHODS

Animal Sound Identifier (ASI): software for automated identification of vocal animals

Otso Ovaskainen,[1,2]* Ulisses Moliterno de Camargo[1,3] and Panu Somervuo[1]

[1]*Faculty of Biological and Environmental Sciences, University of Helsinki, P.O. Box 65, Helsinki FI-00014, Finland*

[2]*Centre for Biodiversity Dynamics, Department of Biology, Norwegian University of Science and Technology, N-7491, Trondheim, Norway*

[3]*The Helsinki Lab of Ornithology, The Finnish Museum of Natural History, University of Helsinki, Helsinki, Finland*

**Correspondence: E-mail: otso.ovaskainen@helsinki.fi*

Abstract

Automated audio recording offers a powerful tool for acoustic monitoring schemes of bird, bat, frog and other vocal organisms, but the lack of automated species identification methods has made it difficult to fully utilise such data. We developed Animal Sound Identifier (ASI), a MATLAB software that performs probabilistic classification of species occurrences from field recordings. Unlike most previous approaches, ASI locates training data directly from the field recordings and thus avoids the need of pre-defined reference libraries. We apply ASI to a case study on Amazonian birds, in which we classify the vocalisations of 14 species in 194 504 one-minute audio segments using in total two weeks of expert time to construct, parameterise, and validate the classification models. We compare the classification performance of ASI (with training templates extracted automatically from field data) to that of monitoR (with training templates extracted manually from the Xeno-Canto database), the results showing ASI to have substantially higher recall and precision rates.

Keywords

Automated vocal identification, autonomous audio recording, joint species distribution modelling, species classification, species identification, vocal communities.

Ecology Letters (2018) **21**: 1244–1254

INTRODUCTION

Acquiring adequately replicated large-scale and long-term data remains a major challenge in ecological research and biodiversity monitoring, especially for species-rich taxa in remote areas as well as taxa that require expert input for identification (Ferraz *et al*. 2008). For vocal taxa such as mammals (Payne *et al*. 2003; Enari *et al*. 2017; Suter *et al*. 2017), birds (Aide *et al*. 2013; Campos-Cerqueira & Aide 2016; Frommolt 2017), bats (MacSwiney *et al*. 2008; Armitage & Ober 2010), frogs (Crouch & Paton 2002; Measey *et al*. 2017) and insects (Fischer *et al*. 1997; Newson *et al*. 2017), automated audio recording offers a powerful tool for acoustic monitoring schemes (Aide *et al*. 2013). The application of bioacoustics monitoring is growing rapidly, both due to technical advances in data collection and management, and due to the rapid build-up of reference audio databases (Ribeiro *et al*. 2017; Wrege *et al*. 2017).

At present, the bottleneck with acoustic monitoring is not so much the data collection, but the process of extracting species detections from extensive recordings covering e.g. tens of thousands of hours (Stowell *et al*. 2016). Several methods have been proposed to semi-automatically identify species from audio recordings, e.g. in the context of LifeCLEF classification challenges (Goëau *et al*. 2015, 2016, 2017; Knight *et al*. 2017). Many of the available methods feed spectral features of sound to various kinds of classifiers, such as decision trees (Acevedo *et al*. 2009; Digby *et al*. 2013), random forests (Ross & Allen 2014; Lasseck 2015b), hidden Markov models (Aide *et al*. 2013) and convolutional neural networks (Salamon & Bello 2017). All available methods require some extent of manual work (Knight *et al*. 2017) and only a few are currently implemented in readily available software (e.g. Arbimon from Sieve-analytics, Raven from Cornell Lab of Ornithology, Sound Scope and Kaleidoscope from Wildlife Acoustics; Shonfield & Bayne 2017). However, automated identification algorithms that would be capable to process continuous audio data from the field and that would have classification accuracy even close to that of an expert observer are still lacking (Stowell *et al*. 2016; Camargo *et al*. 2017; Venier *et al*. 2017).

There are three reasons why automated identification is difficult. First, there is a high diversity of animal vocalisations, both in the types of the basic elements, called syllables (Brandes 2008), and in the way they are combined in e.g. complex vocalisations of songbirds (Brandes 2008; Kroodsma 2015). Second, real field data are complex, as vocalisations of the target species overlap with each other and with background noise, the elimination of which is a challenging task per se (Pacifici *et al*. 2008; Luther 2009). Third, vocalisations in reference databases (e.g. Xeno-Canto; http://www.xeno-canto.org) are typically based on targeted recordings, and they thus lack both biological and technical variation present in field data to be classified, potentially leading to biased results.

Here, we overcome the above-mentioned challenges by developing Animal Sound Identifier (ASI), a software for probabilistic classification of animal sounds directly from field data. We use a case study of crepuscular and nocturnal tropical birds to illustrate that the ASI framework is able to perform reliable species classification based on automatically localised training vocalisations, with minimised user effort for training the classification models. We describe a six-step procedure, which results in a probabilistic classification of the presences or absences of the vocalisations of the target species

(Fig. 1). We then compare the classification performance of ASI to that of monitoR (Katz *et al.* 2016) trained with templates extracted from the Xeno-Canto database. Finally, we illustrate the utility of acoustic monitoring data by deriving ecological inferences from the ASI-based classifications through a joint species distribution modelling approach. We provide MATLAB code and manual to allow users to process their own audio with ASI.

MATERIALS AND METHODS

We illustrate the use of ASI with data on crepuscular and nocturnal birds in the Amazon rainforest. Our data originate from 224 sampling sites where autonomous recorders were set to record for 3 hours during dusk and night for each of five consecutive nights, with a total of 1120 recording nights (see Figueira *et al.* 2015 for further details about data collection). We split these field recordings into one-minute segments, totalling 194 504 segments. Our aim was to apply ASI to these segments to first identify which birds vocalise in them, and then classify all segments for the presence-absences of the vocalisations of the identified species.

Step 1. Identifying letter candidates from field recordings

In the first step we asked ASI to provide 1000 letter candidates from the field recordings, where "letter" stands for a part of animal vocalisation that can be useful for its identification, possibly including one or more syllables, or only a part of a syllable (Brandes 2008). ASI can be used to search for candidate letters either in an unsupervised manner, or using pre-defined templates. The unsupervised search, which is one of the key novelties of ASI, is based on randomly generated letter candidates. This is done by selecting one of the segments, randomising a letter candidate (a rectangular part of the spectrogram, i.e. time-frequency representation of the audio signal) from the segment, and scanning through the other parts of the same segment or of other segments to locate the best match to the letter candidate (Fig. 2; for technical details see Supporting Information). The match between the letter candidate and the segment is measured by cross-correlation using the MATLAB function normxcorr2 (Haralick & Shapiro 1992; Lewis 1995). If the correlation exceeds a threshold value (with 0.9 as default value), ASI includes the located rectangle as a letter candidate, unless the area of high intensity is confined to a few pixels only, which is typical for noise (see Supporting Information for details). ASI then stochastically adjusts the boundaries of the rectangle to improve the correlation to the best match, and to locate the area with the signal to the middle of the rectangle. In each refinement attempt, ASI moves the lower-left and upper-right edges of the box defining the letter by adding to the x- and y-coordinates uniformly distributed random values (see Supporting Information for details).

Step 2. Choosing, improving and annotating letters

The automated search made in Step 1 extracts from possibly thousands of hours of field recordings a set of letter candidates, i.e. subsets of spectrograms that are likely to include vocalising species and be useful for their identification. To further minimise user input, the candidate letters are clustered based on their similarity, so that the user can process letter candidates representing the same vocalisation in a batch (Fig. 2d; see Supporting Information for more details). In the second step, the user first scans through the letter candidates to identify those that represent species of interest, and applies the visual and acoustic tools implemented in ASI to refine the letter boundaries, as necessary. In the example of Fig. 2, we have defined four letters for *Nyctidromus albicollis*, which represent two different vocalisation types (song and call). Multiple letters per species are recommended to be included to provide complementary information to identification from noisy data, and to include biological variation among and within individuals in their vocalisations.

Step 3. Fitting letter-specific models

In the third step, the user constructs letter-specific models that predict the probability by which the letter is present in each audio segment. To do so, ASI first computes for all audio segments the highest correlations to all letters. To explore fast the relationship between highest correlation and letter presence, ASI selects for each letter (a particular vocalisation type of a particular species) ten segments for which the highest correlation is approximately 0.1, 0.2, 0.3, . . ., 1.0. The user classifies these training data to positive and negative matches based on whether the training data actually contains the letter or not. ASI then fits a letter-specific model (probit regression) to the training data to convert the highest correlation into a classification probability. After this starts the adaptive refinement of letter-specific models. In this phase, ASI selects new training data adaptively based on the letter-specific model fitted so far, thus minimising user input by focusing on audio segments that are likely to provide especially high information gain (Fig. 3a–c). To do so, ASI samples a target classification probability uniformly from the range [0,1], uses the current fitted model to determine the corresponding target correlation, and then selects the audio segment for which the highest correlation with the letter is closest to the target correlation. The user is recommended to continue training the model until the mapping from correlation to classification probability (Fig. 3b) converges.

Step 4. Combining multiple letters to construct species-level predictors

In the fourth step ASI combines information from multiple letters to construct predictors for the presences or absences of the target species vocalisations at the level of the audio segments to be classified. These predictors are derived from the letter-specific probabilities that are the outcome of Step 3, and they characterise the temporal patterns at which the letters belonging to the focal species appear in the segment. To compute the species-level predictors, ASI first uses the letter-specific models to predict for each time frame the probability of presence for each letter. In our case study, the segments are one minute long, and we used as the time frame overlapping

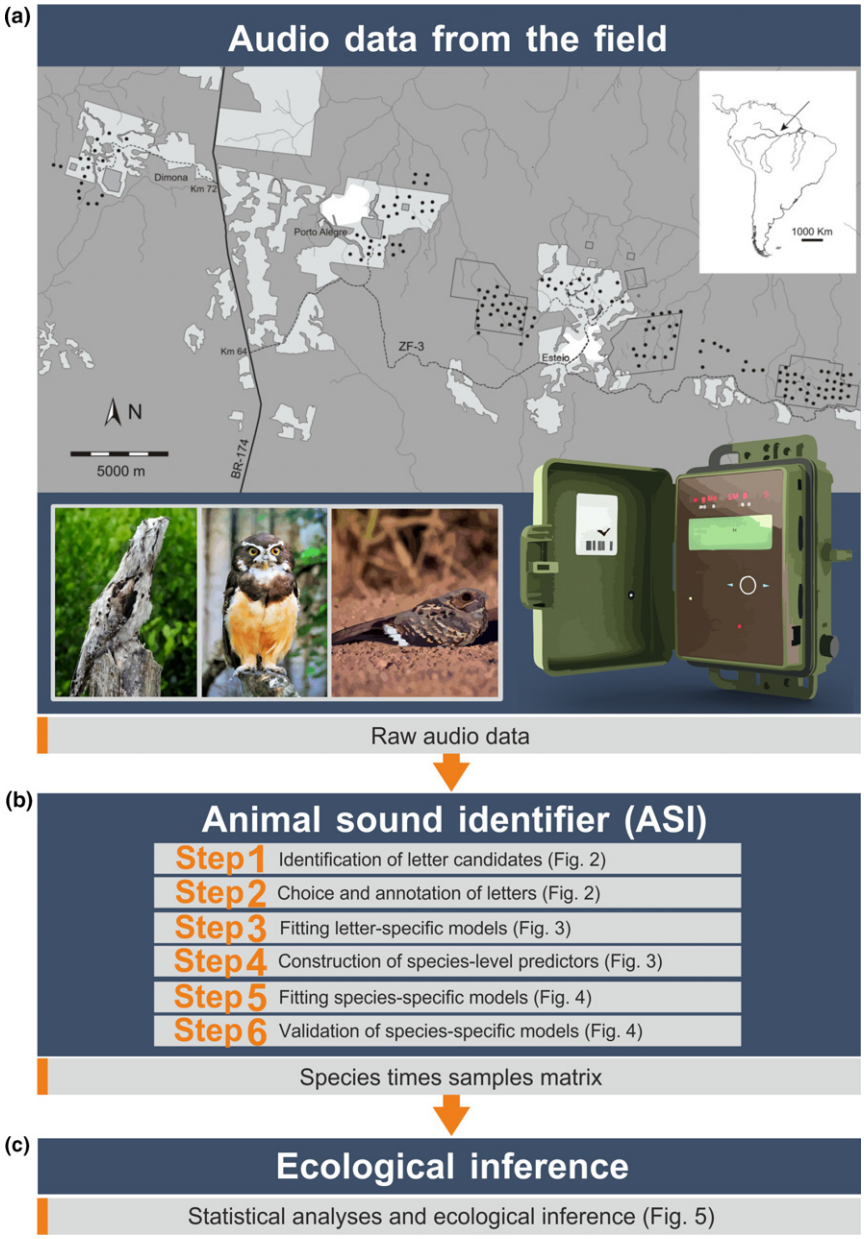

Figure 1 Overview of the Animal Sound Identifier (ASI) use in drawing ecological inference from autonomous-recorder audio data. (a) We acquired audio data from 224 Amazon rain forest sites using autonomous recorders. The data consisted of 194 504 one-minute segments that we wanted to classify for the detection of 14 crepuscular and nocturnal species. (b) ASI consists of a six-step pipeline that takes as input the raw audio data and provides as output the detection probabilities of the target species for the audio segments to be classified. (c) The data provided by ASI works as a starting point for downstream analyses, e.g. for ecological inference. The steps outlined here are further illustrated in Figs. 2–5.

16 ms time windows computed at every 10 ms. Therefore, the dimension of the matrix **A** of letter-specific probabilities (illustrated in the bottom half of Fig. 3d for one audio segment) is $n \times 6000$, where n is the number of letters constructed for the focal species, and 6000 is the number of time windows within the segment. ASI extracts from the matrix **A** a set of summaries that are organised as a vector **b**, called the vector of raw predictors. The elements of the vector **b** are the highest probability for each letter, the fraction of time frames for which each letter exceeds multiple probability thresholds (i.e. letter prevalence, counting vocalisations that are classified with at least e.g. 95, 90 or 50% probability), and the temporal autocorrelation structure of the letters described over logarithmically spaced time intervals (Fig. 3e). The vector **b** of raw-predictors is computed for each audio segment, and these are combined to form the matrix **B.** As the raw predictors are high-dimensional and correlated among the audio segments, they are further processed to produce the matrix of final predictors **C** to be used for parameterising species-level models. Each column of the matrix **C** consists of the vector **c** of predictors for one audio segment, the number of columns equalling the number of segments to be classified. The first element of the vector **c** is defined as the maximum of the letter-specific probabilities. The remaining elements of the vector **c** are

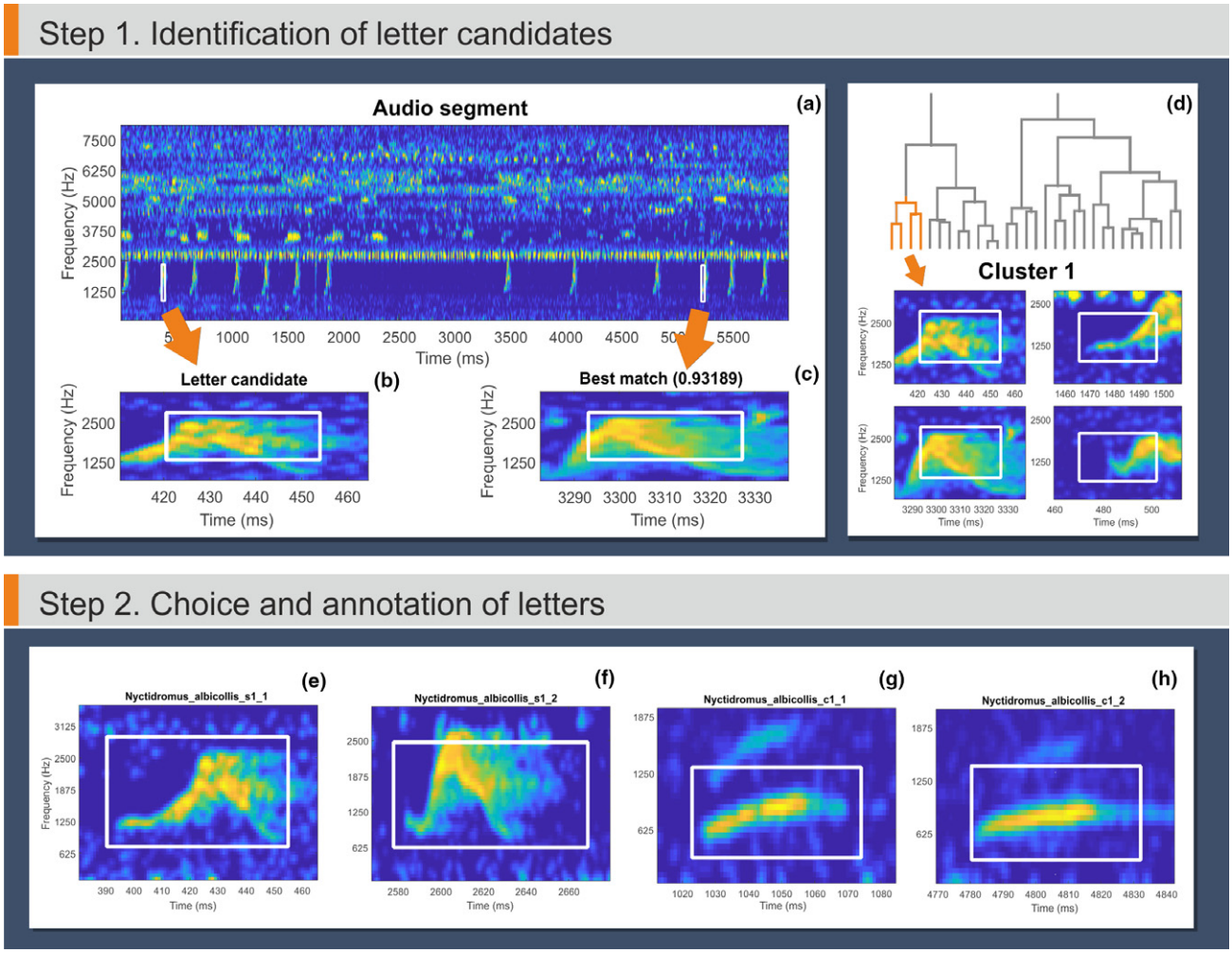

Figure 2 An illustration of Step 1 (identification of letter candidates) and Step 2 (choice and annotation of letters) of the ASI pipeline. (a) shows a one-minute segment of the raw data, from which ASI has identified a letter candidate (b) based on the fact that the same pattern repeats later in the same segment with a sufficiently high correlation (c). (d) ASI clusters the letter candidates to facilitate the selection and annotation of the letters to be done by the user. (e–h) exemplify four letters annotated by the user to represent songs (ef) and calls (gh) of *Nyctidromus albicollis*.

aimed to capture the dominating part of residual variation (after accounting for the maximum probability) in the raw-predictors **B** across the audio segments. To do so, ASI summarises the residual variation by modified principal components (MPCAs), where the modifications include shrinking of outliers and removal of dependency between axes (see Supporting Information for technical details).

Step 5. Fitting species-specific models

In the fifth step ASI guides the user to fit statistical models that predict the presence-absence of vocalisation of each species in each audio segment. The species-specific models are fitted following a similar adaptive approach as for the letter-specific models, thus first covering the entire predictor space and then focusing on the part of the parameter space where the present model involves a high amount of uncertainty (Fig. 4a). Typically it is sufficient to use the maximal probability and the first MPCA1 as the predictors (as done in Fig. 4a), but the user may explore also the discrimination power of the others MPCAs as needed. ASI informs the user about the quality of the present model in terms of its discrimination power, measured by Tjur (2009) R^2, and the user may thus decide when the model is of sufficient quality to cease the training phase. In the example shown in Fig. 4a, the model constructed for *Nyctidromus albicollis* was based on 193 user-classified training segments (the black and red dots in the figure corresponding to presences and absences, respectively), its R^2 equalled 0.36 for the training data and its predicted R^2 equalled 0.48 for all data. The reason why the predicted R^2 for all data often exceeds that of the training data is that the training data are specifically selected to involve cases that are especially difficult to classify.

Step 6. Validation of species-specific models

The models parameterised in the Step 5 estimate the probabilities by which each audio segment contains the vocalisations of each species – which is the main output of ASI. To validate these predictions against independent data that

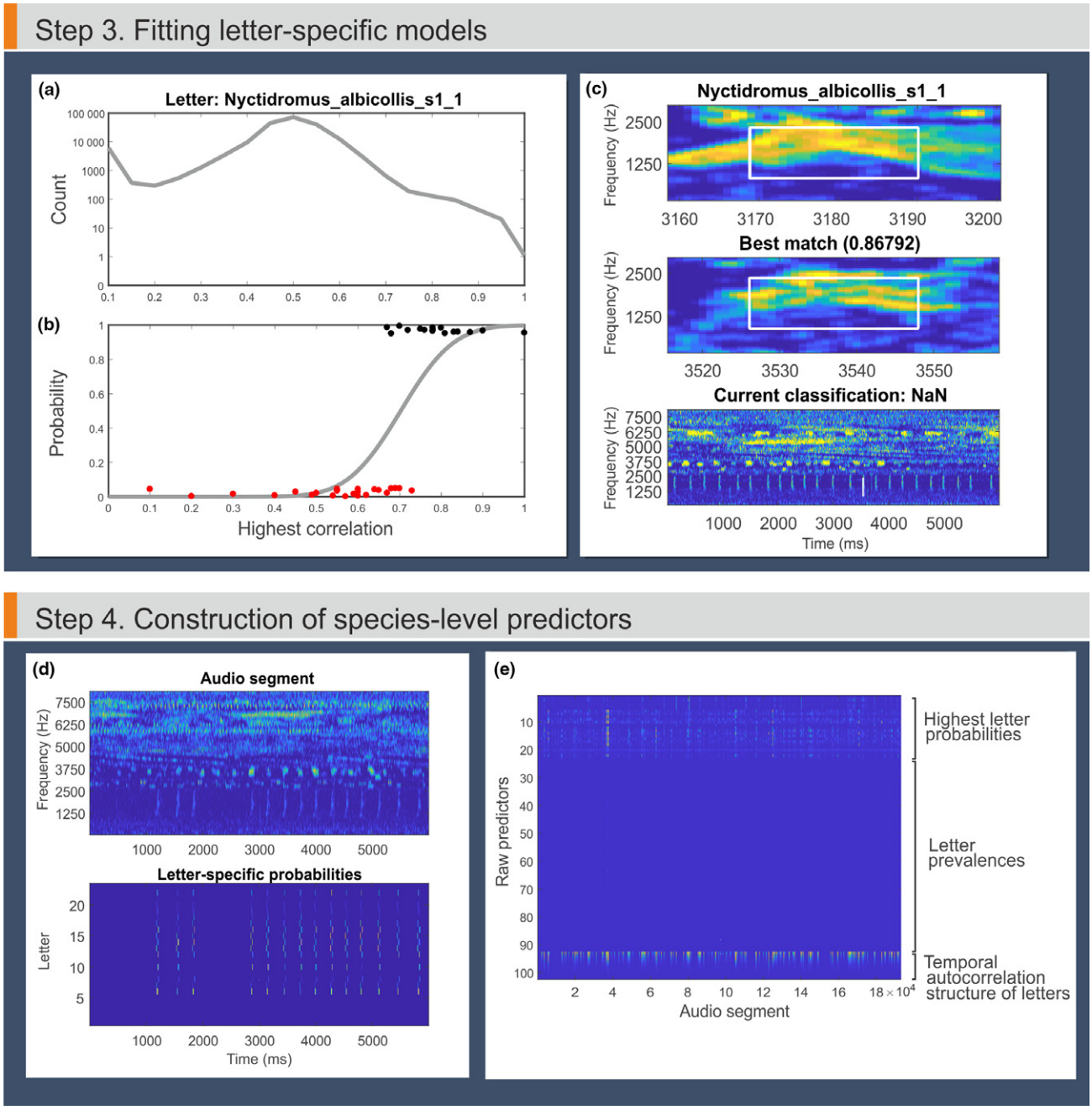

Figure 3 An illustration of Step 3 (fitting letter-specific models) and Step 4 (construction of species-level predictors) of the ASI pipeline. (a) ASI first scans through all the audio segments to compute the highest correlation between each segment and the focal letter, the density of the highest correlations being shown in logarithmic scale. (b) The user classifies training data as positive (black) and negative (red) matches, and ASI subsequently uses the data to model the probability that the best match in each segment is the focal letter. (c) The classification window shows the audio segment, the focal letter and the best match, providing the user with tools for listening to selected parts of the time-frequency space. (d) ASI then scans again through the audio segments to compute the letter-specific probabilities for each segment and time-frame, forming the matrix **A** of letter-specific probabilities (shown in panel d for a single audio segment). The information in matrix **A** is summarised as a set of species-level raw predictors, forming a vector **b** for each audio segment, which vectors are combined in the matrix **B** for the collection of all audio segments (e). The raw predictors consist of highest probabilities of the letters, the prevalence of each letter (proportion of time frames for which the letter is present, based on multiple probability thresholds), and the temporal autocorrelation structure of letter presences.

involves both presences and absences, we randomly sampled for each species 50 segments where the predicted probability was <0.5 and 50 segments where the predicted probability was >0.5. To ensure independence of validation data, we excluded those segments that were used as training data. In the validation phase, ASI provides the segments in a random order to the user, who then classifies them without knowledge about the model-predicted classification

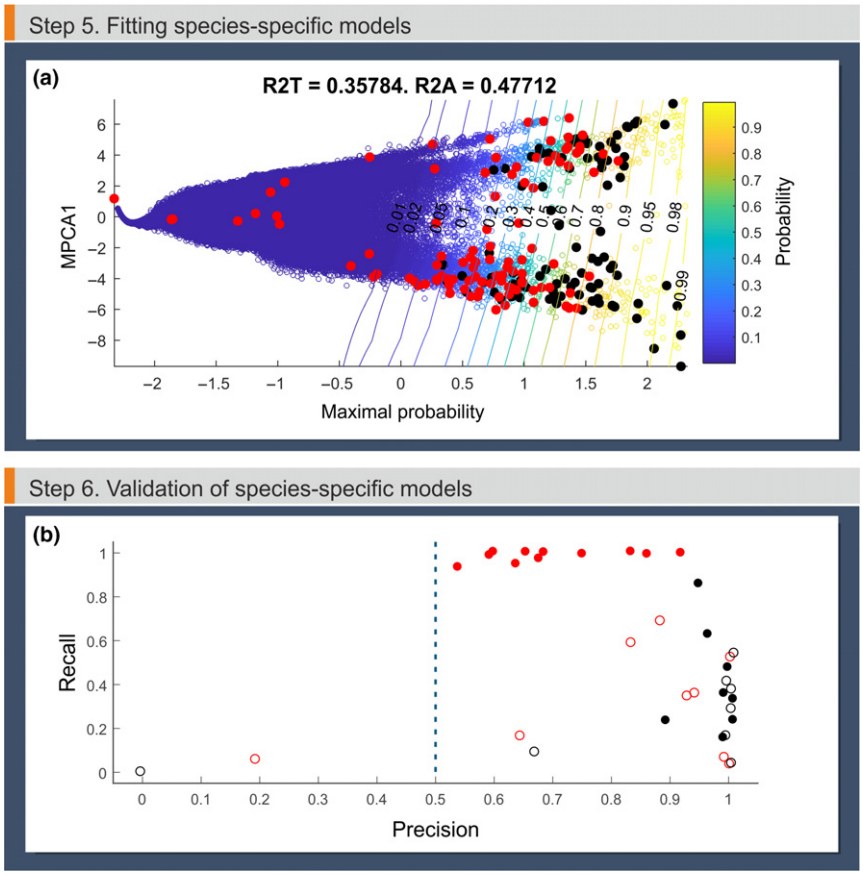

Figure 4 An illustration of Step 5 (fitting species-specific models) and Step 6 (validation of species-specific models) of the ASI pipeline. (a) shows a model fitted to the species *Nyctidromus albicollis*. The predictors used are the first two rows of the matrix **C** (see text) that consist of the maximal probability among the letter (shown at the probit scale) and the first modified principal component (MPCA1). Black and red dots show segments for which the user has classified the focal species to be, respectively, present or absent, while the remaining dots are coloured according to the probability predicted by the model. (b) shows the results of model validation, where the classification probabilities are evaluated against independent validation data in terms of their precision and recall (see text on how these were defined), for both of which 1 is the best and 0 the worst value. The solid dots show the results for ASI, with colours corresponding to 50% (red) and 90% (black) probability thresholds. For comparison, the open dots show the results for monitoR, with colours corresponding to cross-correlation thresholds defined using the greedy (red) and conservative (black) strategies (see text on how these were defined). Cases that are on the right-hand side of the vertical line have a higher precision than expected by random (precision > 0.5).

probabilities. We compared the model-predicted probabilities to the actual classifications in terms of recall and precision, as recommended by Knight *et al.* (2017). Recall is the proportion of target species vocalisations that are detected as hits by the classifier, whereas precision is defined as the proportion of classifier hits that are true detections of the target species. While the output of ASI is probabilistic, calculation of recall and precision requires that the classification probabilities are converted simply to 'hits' or 'non-hits'. We did such a conversion using both 50 and 90% as probability thresholds. We note that as the validation data are constructed so that about half of the segments are expected to contain the species, a classifier that performs 'by random' is predicted to have a precision of 0.5.

Comparison of classification performance between ASI and monitoR

We compared the classifications performed by ASI to those performed by monitoR (Katz *et al.* 2016). MonitoR is a template matching based method for classifying audio data for the presence/absence of target species, and it performed well in a recent methodological comparison (Knight *et al.* 2017). We applied monitoR by following its user manual as closely as possible by performing the following four steps. First, we downloaded for each of the 14 target species five Xeno-Canto reference audio files and used monitoR to extract one template from each file, resulting in a total of 70 templates. Second, to calibrate cross-correlation thresholds, we downloaded five additional Xeno-Canto reference audio files for each species (except three for one of the species for which no more were available). Third, we calculated the maximal cross-correlation between all template and calibration files, and used these data to define an optimal threshold value for each template. To do so, we computed how many hits a candidate threshold value would give to the target species (n) and to the non-target species (m) in the calibration data. We then defined the optimal threshold value as the one that maximised the ratio of hits to target vs. non-target species, as measured by the score $n/(m + 1)$. The template

and calibration data as well as correlation thresholds used in the monitoR analysis is provided in Supporting Information. We then applied monitoR to the same 100 species-specific validation segments that were used to validate the ASI models, following both a greedy and a conservative strategy. In the greedy strategy we considered monitoR to classify the species being present if any of the five templates exceeded the threshold. In the conservative strategy we considered monitoR to classify the species being present if at least two of the five templates exceeded the threshold (we made this choice as only seldom more than two templates exceeded the threshold). We then evaluated the performance of monitoR in terms of precision and recall.

HMSC analyses of the case study on Amazonian crepuscular and nocturnal birds

We derived ecological inferences from the classified data provided by ASI by applying Hierarchical Modelling of Species Communities (HMSC) (Ovaskainen *et al.* 2017). HMSC is a joint species distribution model that models the vector of species occurrences or abundances as a function of environmental, spatial or temporal predictors, and that estimates residual species co-occurrences (not explained by the predictors) at different spatial or temporal levels. We truncated the data to presences ($y = 1$) or absences ($y = 0$) using 0.5 as a probability threshold, and omitted from the analyses all audio segments that were classified as microphone problem. As described in more detail in Supporting Information, we used HMSC with probit regression to model the presence of a detection at the level of day-location pairs, including only those day-location pairs for which at least 50 min had been classified (there was variation in the sampling effort due to occasional microphone failure). In the HMSC model, we used as fixed effects a classification of sampling locations to primary and secondary forests (FOREST; a categorical covariate), the phase of the moon measured by luminosity (MOON; a continuous covariate) and the log-transformed number of minutes of sampling (EFFORT; a continuous covariate). As community-level random effects that model random variation in species occurrences as well as patterns of species co-occurrence, we included the sampling location (LOCATION), the day (DAY), and the location-day pair (LOCATION × DAY). Out of these, LOCATION was assumed to be spatially explicit and it models permanent spatial variation in occurrence and co-occurrence, whereas DAY models synchronised temporal variation in vocalisation activity due to e.g. weather. Our particular interest was LOCATION × DAY, as that models spatiotemporal co-occurrence, thus allowing us to ask whether two species vocalize in the same place at the same day more or less often than expected at random, potentially related to ecological interactions among the species.

RESULTS

We used ASI to classify $n = 194\,504$ sampling units (1-min segments of the data) for the occurrences of the vocalisations of $m = 14$ Amazonian crepuscular and nocturnal bird species, as well as to pinpoint audio tracks with microphone failure. In this case study, ASI tested *c.* 36 000 potential letter candidates to identify 1000 promising ones, which it then clustered into 700 clusters. The largest cluster consisted of > 100 very similar letter candidates that represented technical microphone failure (e.g. intense rainfall generates a transitory short-circuit on microphones, which generates a repetitive beat in the audio), whereas more variable bird sounds were found from smaller clusters of 1–5 candidates. As an example, Fig. 2d illustrates one cluster that contains four similar vocalisations of the species *Nyctidromus albicollis*. As a side product of the unsupervised letter search, we identified also vocalisations of frogs, lizards, crickets and jaguars (data not shown). The supervised approach in which ASI seeks for letter candidates with the help of user provided templates was not applied in the case study, but is illustrated in the Supporting Information.

We selected and annotated from the letter candidates 2–23 letters for each species, yielding in total 110 letters. To construct the letter-specific models, we classified on average 43 audio segments for each letter, thus performing 4757 manual classifications, out of which 1667 were positive and 3090 negative matches. To parameterise the species-specific models, we classified on average 225 audio segments for the 14 bird species, thus performing 3150 manual classifications, out of which 1233 were positive and 1917 negative matches.

Model validation was possible for 11 out of the 14 species as two of the species (*Micrastur mirandollei* and *Nyctibius bracteatus*) were so rare in the audio data that all of their detected vocalisations were used for training data, and we had identified one species (*Penelope* sp.) to genus level only so that comparison to monitoR would not have been possible. With 50% probability threshold, ASI located almost all true vocalisations, whereas with 90% probability threshold it did not provide almost any false positives (Fig. 4b). Averaging over the species, the mean recall rate of ASI was 0.99 (respectively, 0.30) for 50% (respectively, 90%) probability threshold, and its mean precision was 0.71 (respectively, 0.98) for 50% (respectively, 90%) probability threshold. The performance of monitoR was inferior to that of ASI as it resulted in lower recall-precision combinations (Fig. 4b). Averaging over the species, the mean recall rate of monitoR was 0.26 (respectively, 0.17) for greedy (respectively, conservative) strategy, and its mean precision was 0.82 (respectively, 0.83) for the greedy (respectively, conservative) strategy. The average optimal cross-correlation threshold for monitoR was 0.37 (range 0.18. . .0.58).

To produce the final ASI-based classifications for all segments, we re-parameterised the species-specific models by utilising also the validation data in model fitting. As incorporating more data in model fitting will likely improve its performance, the above reported measures of model performance (which are based on models fitted solely to training data) are likely to be conservative. With the final models, the average power of the model-predicted probabilities to discriminate presences and absences of vocalisations was $R^2 = 0.42$ (range from 0.08 to 0.85) (Fig. 5a). To use the full information when deriving ecological inference, we utilised the manually classified data (i.e. the training and validation data) directly by replacing the predicted detection probabilities by

the actually known presences or absences. Utilising the manually conducted classifications improved the average discrimination power to $R^2 = 0.60$ (range from 0.21 to 0.99) (Fig. 5a). The substantial increase is explained by the fact that many of the species were very rare in the audio data, and thus in some cases the manually confirmed vocalisations represented a large proportion of all vocalisations, even if the amount of manual classifications represent only a tiny fraction (*c.* 0.2%) of the data to be classified.

The fraction of sampling units in which the species were predicted to vocalise ranged from 0.04 to 1.3% (Fig. 5a). In the HMSC model, the variances explained by the fixed effects (averaged over the species) were FOREST: 9%, MOON: 5% and EFFORT: 11%. Among the 14 species, four were found to vocalise especially often in primary forests and two in secondary forests. The phase of the moon did not appear to have a strong effect: vocal activity increased with luminosity for one species and decreased for another species (Fig. 5a). The proportions of variance explained by the random effects (averaged over the species) were LOCATION: 41%, DAY: 17%, and LOCATION × DAY: 16%. The associations among the species were predominantly positive both at the level of location (Fig. 5b) and at the level of day (Fig. 5c). The analysis further pinpointed several species pairs that show positive associations at the level of location-day (Fig. 5d), possibly suggesting the presence of ecological interactions among them.

DISCUSSION

As illustrated by our results, ASI provides a powerful tool for classifying autonomous field recordings for the species occurrences. Compared to alternative approaches (Briggs *et al.* 2012; Potamitis 2014; Lasseck 2015a,b), the key novelties of ASI are the following five. First, ASI does not require any *a priori* templates of the target vocalisations, but it finds them

(a)

	Species	R2	R2*	P > 0.5	M	PF
1	*Crypturellus variegatus*	0.85	0.94	0.2	1	1
2	*Glaucidium hardyi*	0.27	0.32	0.4	0	0
3	*Lophostrix cristata*	0.47	0.55	0.5	0	0
4	*Micrastur mirandollei*	0.08	0.66	0.02	0	0
5	*Nyctibius aethereus*	0.40	0.52	0.2	0	0
6	*Nyctibius bracteatus*	0.61	0.99	0.04	0	0
7	*Nyctibius griseus*	0.37	0.58	0.2	0	–1
8	*Nyctibius leucopterus*	0.41	0.62	0.2	0	1
9	*Nyctidromus albicollis*	0.51	0.64	0.2	0	1
10	*Megascops watsonii*	0.27	0.29	1.3	0	0
11	*Penelope sp*	0.27	0.65	0.04	–1	–1
12	*Pulsatrix perspicillata*	0.13	0.21	0.06	0	0
13	*Ciccaba huhula*	0.48	0.58	0.3	0	0
14	*Tinamus major*	0.46	0.55	0.7	0	1
	Microphone problem	0.82	0.83	4.5		

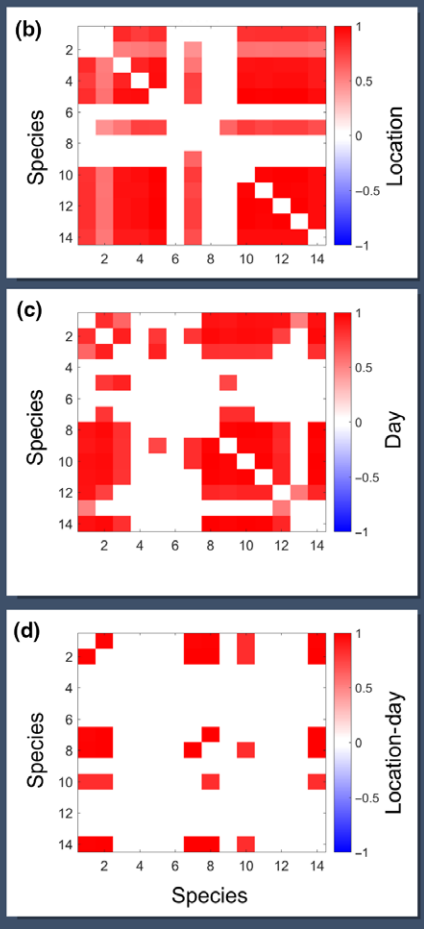

Figure 5 Ecological inference derived for the crepuscular and nocturnal bird case study. The first three columns in (a) show descriptive statistics for each target species: the R^2 value of the ASI models without (*R*2) or with (*R*2*) use of manually classified data, and the percentage (%) of audio segments for which the probability of occurrence is at least 0.5 ($P > 0.5$), The last two columns in (a) show whether the species vocalisation activity increases with moon luminosity (M) or is higher in primary than secondary forests (PF), the values 1 (respectively, -1) indicating a positive (respectively, negative) response with at least 95% posterior probability based on the fitted HMSC model. Panels (bcd) show species associations at the levels of LOCATION, DAY, and LOCATION x DAY. The red and blue squares indicate species pairs that co-occur or co-vocalise respectively more or less often than expected at random with at least 95% posterior probability based on the fitted HMSC model.

directly from field recordings. Second, as the letters are identified by a random search through the field recordings, they are fully representative of the relevant technical and biological variation in the data (Alldredge *et al.* 2007; Pacifici *et al.* 2008). Third, ASI generates training data adaptively, thus asking the user to classify only such training data for which classification by the present model would be uncertain, which data are thus especially valuable for improving classification accuracy. Fourth, ASI combines information across multiple letters (e.g. syllables) in a generic manner, avoiding the need to manually construct species-specific models, in contrast to e.g. the approach of Aide *et al.* (2013). Fifth, as ASI is based on a calibrated statistical model, it quantifies classification uncertainty with the help of probabilities, thus providing the user a quantitative mean to account for classification uncertainty in any downstream analyses.

We found ASI to yield higher recall and precision rates than monitoR, suggesting that ASI is able to detect species occurrences from extensive amounts of noisy field recordings in a computationally efficient and reliable manner. One key reason behind the performance difference between ASI and monitoR is that ASI's classification models were based directly on the field data, whereas in case of monitoR they were based on reference audio files from the Xeno-Canto database. The field recordings differ from Xeno-Canto reference audio files in many ways, including technical recording quality, the type of background noise, and the geographic region from which the vocalisations originate, all of which factors reduce classification accuracy. We note that also monitoR could be applied to templates extracted from field data, which could possibly improve its performance. However, doing so would require subsampling and manual listening of the field data in order to extract regions with target vocalisations to be used to generate templates. Covering enough data to provide multiple templates for all species, however, would be very time consuming for the users, especially if the interest is in rare species. This is exactly what the unsupervised search by ASI was designed for.

While ASI provides a major step forward in semi-automated classification of animal vocalisations, it clearly involves several limitations that we hope future research efforts to improve on. First of all, as the classification models are based on training data provided by the user, an upper limit for the performance of ASI is clearly set by the level of expertise of the user. Most obviously, if the user is not able to identify the species behind a certain vocalisation type, ASI will not be able to classify those vocalisation types either. One possibility to help users with varying levels of expertise to overcome this limitation might be to scan the letter candidates against existing databases, such as Xeno-Canto, in order to provide automated suggestions of species annotations. However, in the end the validity of such annotations needs always be checked by an expert user. A second limitation of ASI is that its core statistical approach is generically applied to all species, thus assuming that the same kinds of features (e.g. letter prevalence and autocorrelation structure) are relevant for all of them. Furthermore, the current implementation of ASI utilises cross-correlation as the basis of comparison between query and reference audio files, whereas also many other kinds of acoustic features could be applied (Bardeli 2009; Lasseck 2015b). Tailoring the modelling approach and choosing the acoustic features in a species-specific manner would likely improve the recall and precision rates especially for those species for which they were the lowest in our case study. A third limitation is that, in spite of the high recall and precision rates, the predicted classifications will always involve some uncertainty. Whether or not removing such uncertainty by post-classification validation is possible or necessary depends on the type of the data and the purpose of the study. For example, if the aims is to verify the occurrence of a rare species, it is clearly both necessary and possible to manually scan through the most likely detections and thus to confirm the occurrence of the target species. For another example, if the aim is to use the classifications in statistical analyses aimed for ecological inference, post-validation of both positive and negative classifications would surely be beneficial, but it may be very tedious to do in practice. In our case study, some of the species were so rare that the training data covered most occurrences, and thus we were able to use the manual classifications instead of the model predicted probabilities. In contrast, as the most common species (*Megascops watsonii*) was predicted to vocalize in *c.* 2500 one minute segments, a manual post-classification validation just for this one species would require extensive work, in particular for the validation of the absences which are equally informative as presences from the viewpoint of statistical modelling. As the key benefit of ASI is that it is able to classify massive amounts of data rather than a small sample of it, the disadvantage of having some level of classification error is likely to be more than compensated by the ample supply of data, as long as the recall and precision rates are sufficiently high for the signal to dominate the noise.

While our focus here was primarily in developing a method for automated species identification, the case study of crepuscular and nocturnal birds provided also some ecological insights on Amazonian birds. A previous analysis of nocturnal birds from the same study area (Sberze *et al.* 2010) reported generally similar habitat preferences as we observed here, most species showing no strong preference, but e.g. *Nyctibius griseus* being secondary forest specialist and *Nyctibius leucopterus* being primary forest specialist. However, the results of Sberze *et al.* (2010) are not directly comparable to our results because of differences in the field methods (their study was based on manual point counting with play back and thus the sample size was much lower) and in the statistical analyses (their study focused on occupancy patterns). Of particular interest are the co-occurrence patterns that we identified at three spatiotemporal levels. As we had accounted for the difference between primary and secondary forests in the fixed effects of the HMSC model, the positive associations related to permanent spatial variation (Fig. 5b) are likely to be related to more subtle habitat quality variation, many species favouring the same kinds of microhabitats. The positive associations among the species identified at the level of day (Fig. 5c) suggest that vocalisation activity was generally synchronised among the species over the days. As we accounted for luminosity in the fixed effect, the synchronisation is not likely to be related to the phase of the moon, but e.g. variation in weather conditions. While the positive associations identified at the level of location-day (Fig. 5d) may partly be

due to environmental covariates not included in the model, they provide an interesting array of data-driven hypotheses of interspecific interactions, and thus an exciting starting point for more detailed analyses (Ferraz *et al*. 2010).

A previous analysis of diurnal birds from the same study area found that while parrots inhabit both primary and secondary forests, their perching activity is higher in primary forests (Figueira *et al*. 2015). The study of Figueira *et al*. (2015) was based on autonomous recording, but they identified the species by manual identification. As the study focused solely on parrots, candidate locations for their vocalisations could be fast found by visual scanning of the data before confirming the identifications by listening. In spite of this, performing the manual identifications took several months of expert time, with the size of the data set being comparable with the one considered here. In the present study, the total amount of work by the user consisted of 14 working days, of which 4 days were spent in identifying and annotating letters (Step 2), 4 days in training the letter models (Step 3), 4 days in training the species models (Step 5), and 2 days in validating the species-level classifications (Step 6). This illustrates how ASI not only provides accurate classifications, but also makes an efficient use of human time. Consequently, we expect ASI to become widely adopted as a tool for utilising autonomous field recordings to acquire community-level data on the occurrences and abundances of vocal species.

ACKNOWLEDGEMENTS

Field recordings were obtained with the support of grants from the Smithsonian Institution Center for Tropical Forest Science and the Amazonas State Science Foundation, awarded to Gonçalo Ferraz and Cintia Cornelius. We thank Gonçalo Ferraz and three anonymous reviewers for many insightful comments on the manuscript. The research was funded by the Academy of Finland (grants 1273253, 250444 and 284601 to OO), the Research Council of Norway (CoE grant 223257), and the LUOVA graduate school of the University of Helsinki (PhD grant for UC).

AUTHORSHIP

OO came up with the original idea and developed the ASI software, with contributions from UC and PS. UC acquired the data and applied ASI to the case study of crepuscular and nocturnal birds. OO performed the HMSC analyses. OO and UC wrote the first version of the manuscript and all authors contributed to the final version.

DATA ACCESSIBILITY STATEMENT

Data are available in the Dryad Digital Repository: http://doi.org/10.5061/dryad.221mq23

REFERENCES

Acevedo, M.A., Corrada-Bravo, C.J., Corrada-Bravo, H., Villanueva-Rivera, L.J. & Aide, T.M. (2009). Automated classification of bird and amphibian calls using machine learning: a comparison of methods. *Ecol. Inform.*, 4, 206–214.

Aide, T.M., Corrada-Bravo, C., Campos-Cerqueira, M., Milan, C., Vega, G. & Alvarez, R. (2013). Real-time bioacoustics monitoring and automated species identification. *PeerJ*, 1, e103.

Alldredge, M.W., Simons, T.R. & Pollock, K.H. (2007). Factors affecting aural detections of songbirds. *Ecol. Appl.*, 17, 948–955.

Armitage, D.W. & Ober, H.K. (2010). A comparison of supervised learning techniques in the classification of bat echolocation calls. *Ecol. Inform.*, 5, 465–473.

Bardeli, R. (2009). Similarity search in animal sound databases. *IEEE Trans. Multimedia*, 11, 68–76.

Brandes, T.S. (2008). Automated sound recording and analysis techniques for bird surveys and conservation. *Bird Conserv. Int.*, 18, S163–S173.

Briggs, F., Lakshminarayanan, B., Neal, L., Fern, X.Z., Raich, R., Hadley, S.J. *et al*. (2012). Acoustic classification of multiple simultaneous bird species: a multi-instance multi-label approach. *J. Acoust. Soc. Am.*, 131, 4640–4650.

Camargo, U.M.D., Somervuo, P. & Ovaskainen, O. (2017). PROTAX-Sound: a probabilistic framework for automated animal sound identification. *PLoS ONE*, 12, e0184048.

Campos-Cerqueira, M. & Aide, T.M. (2016). Improving distribution data of threatened species by combining acoustic monitoring and occupancy modelling. *Methods Ecol. Evol.*, 7, 1340–1348.

Crouch, W.B. & Paton, P.W.C. (2002). Assessing the use of call surveys to monitor breeding anurans in Rhode Island. *J. Herpetol.*, 36, 185–192.

Digby, A., Towsey, M., Bell, B.D., Teal, P.D. & Giuggioli, L. (2013). A practical comparison of manual and autonomous methods for acoustic monitoring. *Methods Ecol. Evol.*, 4, 675–683.

Enari, H., Enari, H., Okuda, K., Yoshita, M., Kuno, T. & Okuda, K. (2017). Feasibility assessment of active and passive acoustic monitoring of sika deer populations. *Ecol. Ind.*, 79, 155–162.

Ferraz, G., Marinelli, C.E. & Lovejoy, T.E. (2008). Biological monitoring in the Amazon: recent progress and future needs. *Biotropica*, 40, 7–10.

Ferraz, G., Sberze, M. & Cohn-Haft, M. (2010). Using occupancy estimates to fine-tune conservation concerns RESPONSE. *Anim. Conserv.*, 13, 19–20.

Figueira, L., Tella, J.L., Camargo, U.M. & Ferraz, G. (2015). Autonomous sound monitoring shows higher use of Amazon old growth than secondary forest by parrots. *Biol. Cons.*, 184, 27–35.

Fischer, F.P., Schulz, U., Schubert, H., Knapp, P. & Schmöger, M. (1997). Quantitative assessment of grassland quality: acoustic determination of population sizes of orthopteran indicator species. *Ecol. Appl.*, 7, 909–920.

Frommolt, K.-H. (2017). Information obtained from long-term acoustic recordings: applying bioacoustic techniques for monitoring wetland birds during breeding season. *J. Ornithol.*, 158, 659–668.

Goëau, H., Glotin, H., Vellinga, W.-P., Planqu, R.É., Rauber, A. & Joly, A. (2015). Lifeclef bird identification task 2015. In: *CLEF working notes 2015*. Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016. (http://ceur-ws.org/Vol-1609/)

Goëau, H., Glotin, H., Vellinga, W.-P., Planqu, R.É. & Joly, A. (2016). LifeCLEF bird identication task 2016: the arrival of deep learning. In: *CLEF working notes 2016*. Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016. (http://ceur-ws.org/Vol-1609/)

Goëau, H., Glotin, H., Vellinga, W.-P., Planqu, R.É. & Joly, A. (2017). LifeCLEF bird identication task 2017. In: *CLEF working notes 2017*. Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Dublin, Ireland, September 11-14, 2017. (http://ceur-ws.org/Vol-1866/)

Haralick, R.M. & Shapiro, L.G. (1992). *Computer and robot vision*. Addison-Wesley, Boston, MA.

Katz, J., Hafner, S.D. & Donovan, T. (2016). Tools for automated acoustic monitoring within the R package monitoR. *Bioacoustics*, 25, 197–210.

Knight, E., Hannah, K.C., Foley, G.J., Scott, C.D., Brigham, R. & Bayne, E. (2017). Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs. *Avian Conserv. Ecol.*, 12, Article 14. http://www.ace-eco.org/vol12/iss2/art14/.

Kroodsma, D. (2015). *The singing life of birds: the art and science of listening to birdsong*. Houghton Mifflin Harcourt, Boston, MA.

Lasseck, M. (2015a). Improved automatic bird identification through decision tree based feature selection and bagging. In *The cross language image retrieval track (CLEF) conference*. (eds Cappellato, L., Ferro, N., Jones, G.J.F. & San Juan, E.). France, CEUR, Toulouse, 12pp.

Lasseck, M. (2015b). Towards automatic large-scale identification of birds in audio recordings. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 6th International Conference of the CLEF Association, CLEF'15, Toulouse, France, September 8-11, 2015, Proceedings* (eds Mothe, J., Savoy, J., Kamps, J., Pinel-Sauvagnat, K., Jones, J.F.G., SanJuan, E., Cappellato, L. & Ferro, N.). Springer International Publishing, Cham, pp. 364–375.

Lewis, J.P. (1995). Fast Template Matching. Vision Interface 95, Canadian Image Processing and Pattern Recognition Society: 120–123.

Luther, D. (2009). The influence of the acoustic community on songs of birds in a neotropical rain forest. *Behav. Ecol.*, 20, 864–871.

MacSwiney, G.M.C., Clarke, F.M. & Racey, P.A. (2008). What you see is not what you get: the role of ultrasonic detectors in increasing inventory completeness in Neotropical bat assemblages. *J. Appl. Ecol.*, 45, 1364–1371.

Measey, G.J., Stevenson, B.C., Scott, T., Altwegg, R. & Borchers, D.L. (2017). Counting chirps: acoustic monitoring of cryptic frogs. *J. Appl. Ecol.*, 54, 894–902.

Newson, S.E., Bas, Y., Murray, A. & Gillings, S. (2017). Potential for coupling the monitoring of bush-crickets with established large-scale acoustic monitoring of bats. *Methods Ecol. Evol.*, 8, 1051–1062.

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D. *et al.* (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecol. Lett.*, 20, 561–576.

Pacifici, K., Simons, T.R. & Pollock, K.H. (2008). Effects of vegetation and background noise on the detection process in auditory avian point-count surveys. *Auk*, 125, 600–607.

Payne, K.B., Thompson, M. & Kramer, L. (2003). Elephant calling patterns as indicators of group size and composition: the basis for an acoustic monitoring system. *Afr. J. Ecol.*, 41, 99–107.

Potamitis, I. (2014). Automatic classification of a taxon-rich community recorded in the wild. *PLoS ONE*, 9, e96936.

Ribeiro, J.W., Sugai, L.S.M. & Campos-Cerqueira, M. (2017). Passive acoustic monitoring as a complementary strategy to assess biodiversity in the Brazilian Amazonia. *Biodivers. Conserv.*, 26, 2999–3002.

Ross, J.C. & Allen, P.E. (2014). Random Forest for improved analysis efficiency in passive acoustic monitoring. *Ecol. Inform.*, 21, 34–39.

Salamon, J. & Bello, J.P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Process. Lett.*, 24, 279–283.

Sberze, M., Cohn-Haft, M. & Ferraz, G. (2010). Old growth and secondary forest site occupancy by nocturnal birds in a neotropical landscape. *Anim. Conserv.*, 13, 3–11.

Shonfield, J. & Bayne, E.M. (2017). Autonomous recording units in avian ecological research: current use and future applications. *Avian Conservation and Ecology*, 12(1):14. https://doi.org/10.5751/ACE-00974-120114.

Stowell, D., Woody, M., Stylianouz, Y. & Glotin, H. (2016). Bird detection in audio: a survey and a challenge. In: *2016 IEEE International Workshop on Machine Learning for Signal Processing*, Salerno, Italy.

Suter, S.M., Giordano, M., Nietlispach, S., Apollonio, M. & Passilongo, D. (2017). Non-invasive acoustic detection of wolves. *Bioacoustics*, 26, 237–248.

Tjur, T. (2009). Coefficients of determination in logistic regression models —A new proposal: the coefficient of discrimination. *The American Statistician*, 63, 366–372.

Venier, L.A., Mazerolle, M.J., Rodgers, A., McIlwrick, K.A., Holmes, S. & Thompson, D. (2017). Comparison of semiautomated bird song recognition with manual detection of recorded bird song samples. *Avian Conservation and Ecology*, 12, 2.

Wrege, P.H., Rowland, E.D., Keen, S. & Shiu, Y. (2017). Acoustic monitoring for conservation in tropical forests: examples from forest elephants. *Methods Ecol. Evol.*, 8, 1292–1301.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article:

Editor, Joshua Lawler
Manuscript received 12 February 2018
First decision made 1 April 2018
Manuscript accepted 3 May 2018