Received: 22 January 2018 | Accepted: 5 October 2018

DOI: 10.1111/2041-210X.13114

RESEARCH ARTICLE

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

CityNet—Deep learning tools for urban ecoacoustic assessment

Alison J. Fairbrass[1,2,3*] | Michael Firman[4*] | Carol Williams[3] | Gabriel J. Brostow[4] | Helena Titheridge[1] | Kate E. Jones[2,5]

[1]Department of Civil, Environmental and Geomatic Engineering, Centre for Urban Sustainability and Resilience, University College London, London, UK

[2]Department of Genetics, Evolution and Environment, Centre for Biodiversity and Environment Research, University College London, London, UK

[3]Bat Conservation Trust, London, UK

[4]Department of Computer Science, University College London, London, UK

[5]Zoological Society of London, Institute of Zoology, London, UK

Correspondence
Alison J. Fairbrass
Email: alison.fairbrass@gmail.com
Michael Firman
Email: mdfirman@gmail.com
Kate E. Jones
Email: kate.e.jones@ucl.ac.uk

Funding information
Engineering and Physical Sciences Research Council, Grant/Award Number: EP/G037698/1 and EP/K015664/1

Handling Editor: Nick Isaac

Abstract

1. Cities support unique and valuable ecological communities, but understanding urban wildlife is limited due to the difficulties of assessing biodiversity. Ecoacoustic surveying is a useful way of assessing habitats, where biotic sound measured from audio recordings is used as a proxy for population abundance and/or activity. However, existing algorithms systematically over and underestimate measures of biotic activity in the presence of typical urban non-biotic sounds in recordings.
2. We develop CityNet, a deep learning system using convolutional neural networks (CNNs), to measure audible biotic (CityBioNet) and anthropogenic (CityAnthroNet) acoustic activity in cities. The CNNs were trained on a large dataset of annotated audio recordings collected across Greater London, UK. Using a held-out test dataset, we compare the precision and recall of CityBioNet and CityAnthroNet separately to the best available alternative algorithms: four Acoustic Indices: Acoustic Complexity Index, Acoustic Diversity Index, Bioacoustic Index, and Normalised Difference Soundscape Index, and a state-of-the-art bird call detection CNN (bulbul). We also compare the effect of non-biotic sounds on the predictions of CityBioNet and bulbul. Finally we apply CityNet to describe acoustic patterns of the urban soundscape in two sites along an urbanisation gradient.
3. CityBioNet was the best performing algorithm for measuring biotic activity in terms of precision and recall, followed by bulbul, whereas the Acoustic Indices performed worst. CityAnthroNet outperformed the Normalised Difference Soundscape Index, but by a smaller margin than CityBioNet achieved against the competing algorithms. The CityBioNet predictions were impacted by mechanical sounds, whereas air traffic and wind sounds influenced the bulbul predictions. Across an urbanisation gradient, we show that CityNet produced realistic daily patterns of biotic and anthropogenic acoustic activity from real-world urban audio data.
4. Using CityNet, it is possible to automatically measure biotic and anthropogenic acoustic activity in cities from audio recordings. If embedded within an autonomous sensing system, CityNet could produce environmental data for cites at

*Denotes joint first authorship.

large-scales and facilitate investigation of the impacts of anthropogenic activities on wildlife. The algorithms, code and pretrained models are made freely available in combination with two expert-annotated urban audio datasets to facilitate automated environmental surveillance in cities.

KEYWORDS
acoustic indices, anthropogenic, biodiversity assessment, convolutional neural networks, ecoacoustics, machine learning, soundscapes, urban ecology

1 | INTRODUCTION

Over half of the world's human population now live in cities (UNDESA, 2016) and urban biodiversity can provide people with a multitude of health and well-being benefits including improved physical and psychological health (Crouse et al., 2017; Natural England 2016). Cities can support high biodiversity including native endemic species (Aronson et al., 2014), and act as refuges for biodiversity that can no longer persist in intensely managed agricultural landscapes surrounding cities (Hall et al., 2016). However, our understanding of urban biodiversity remains limited (Beninde, Veith, & Hochkirch, 2015; Faeth, Bang, & Saari, 2011). One reason for this is the difficulties associated with biodiversity assessment, such as gaining repeated access to survey sites and the resource intensity of traditional methods (Farinha-Marques, Lameiras, Fernandes, Silva, & Guilherme, 2011). This inhibits our ability to conduct the large-scale assessment that is necessary for understanding urban ecosystems.

Ecoacoustic surveying has emerged as a useful method of large-scale quantification of ecological communities and their habitats (Sueur & Farina, 2015). Passive acoustic recording equipment facilitates the collection of audio data over long time periods and large spatial scales with fewer resources than traditional survey methods (Digby, Towsey, Bell, & Teal, 2013). A number of automated methods have been developed to measure biotic sound in the large volumes of acoustic data that are typically produced by ecoacoustic surveying (Sueur & Farina, 2015). For example, Acoustic Indices use the spectral and temporal characteristics of acoustic energy in sound recordings to produce whole community measures of biotic and anthropogenic sound (Sueur, Farina, Gasc, Pieretti, & Pavoine, 2014). However, several commonly used Acoustic Indices have been shown to be biased by non-biotic sounds (Fuller, Axel, Tucker, & Gage, 2015; Gasc, Pavoine, Lellouch, Grandcolas, & Sueur, 2015; Towsey, Wimmer, Williamson, & Roe, 2014), and are not suitable for use in the urban environment without the prior removal of certain non-biotic sounds from recordings (Fairbrass, Rennett, Williams, Titheridge, & Jones, 2017).

Machine learning (ML) is being increasingly applied to biodiversity assessment and monitoring because it facilitates the detection and classification of ecoacoustic signals in audio data (Acevedo, Corrada-Bravo, Corrada-Bravo, Villanueva-Rivera, & Aide, 2009; Stowell & Plumbley, 2014; Walters et al., 2012). Using annotated audio datasets of soniferous species, a ML model can be trained to recognise biotic sounds based on multiple acoustic characteristics, or features, and to associate these features with taxonomic classifications, and can then assign a classification to sounds within recordings. Acoustic Indices only use a limited number of acoustic features in their calculations, such as spectral entropy within defined frequency bands (Boelman, Asner, Hart, & Martin, 2007; Kasten, Gage, Fox, & Joo, 2012; Villanueva-Rivera, Pijanowski, Doucette, & Pekin, 2011) or entropy changes over time (Pieretti, Farina, & Morri, 2011). Additionally, the relationship between the features and the algorithm outputs are chosen by a human, rather than learned automatically from an annotated dataset. In contrast, ML algorithms can utilise many more features in their calculations, and the relationship between inputs and outputs is determined automatically based on the annotated training data provided. Convolutional Neural Networks, CNNs (or Deep learning) (LeCun, Bengio, & Hinton, 2015) can even choose, based on the annotations in the training dataset, the features that discriminate different classes in datasets without being specified a priori, and can take advantage of large quantities of training data where their ability to outperform human defined algorithms increases as more labelled data become available.

Species-specific ML algorithms have been developed to automatically identify the sounds emitted by a range of soniferous organisms including birds (Stowell & Plumbley, 2014), bats (Walters et al., 2012; Zamora-Gutierrez et al., 2016), amphibians (Acevedo et al., 2009), and grasshoppers (Chesmore & Ohya, 2004). However, these algorithms are focussed on a small number of species limiting their usefulness for broad classification tasks across communities. More recently, algorithms that detect whole taxonomic groups are being developed, for example, bird sounds in audio recordings from the UK and the Chernobyl Exclusion Zone (Grill & Schlüter, 2017), but these algorithms remain untested on noisy audio data from urban environments. There are currently no algorithms that produce whole community measures of biotic sound that are known to be suitable for use in acoustically complex urban environments.

Here, we develop the CityNet acoustic analysis system, which uses two CNNs for measuring audible (0–12 kHz) biotic (CityBioNet) and anthropogenic (CityAnthroNet) acoustic activity in audio recordings from urban environments. We use this frequency range as

it contains the majority of sounds emitted by soniferous species and anthropogenic activity in the urban environment (Fairbrass et al., 2017). The CNNs were trained using CitySounds2017, an expert-annotated dataset of urban sounds collected across Greater London, UK that we develop here. Each CNN predicted the presence or absence of each sound type at each moment in time, and these per-time predictions were aggregated to provide a measure of acoustic activity. We compared the performance of CityNet using a held-out dataset by comparing the algorithms' precision and recall to four commonly used Acoustic Indices: Acoustic Complexity Index (ACI) (Pieretti et al., 2011), Acoustic Diversity Index (ADI) (Villanueva-Rivera et al., 2011), Bioacoustic Index (BI) (Boelman et al., 2007), Normalised Difference Soundscape Index (NDSI) (Kasten et al., 2012), and to bulbul, a state-of-the-art algorithm for detecting bird sounds in order to summarise avian acoustic activity (Grill & Schlüter, 2017). As the main focus of the study was the development of algorithms for ecoacoustic assessment of biodiversity in cities, we conducted further analysis on the two best performing algorithms for measuring biotic sound, CityBioNet and bulbul, by investigating the effect of non-biotic sounds on the accuracy of the algorithms. Finally, we applied CityNet to investigate daily patterns of biotic and anthropogenic sound in the urban soundscape.

2 | MATERIALS AND METHODS

We developed two CNN models, CityBioNet and CityAnthroNet within the CityNet system to generate measures of biotic and anthropogenic acoustic activity respectively. The CityNet pipeline (Figure 1) consisted of seven main steps as follows:

1. *Record audio*: Audible frequency (0–12 kHz) .wav recordings were made using a passive acoustic recorder at a sample rate of 24 kHz.
2. *Audio conversion to Mel spectrogram:* Each audio file was automatically converted to a Mel spectrogram representation with 32 frequency bins, represented as rows in the spectrogram, using a temporal resolution of 21 columns per second of raw audio. Each column in the spectrogram was computed by running the fast Fourier transform on a section of the audio time signal. Each spectrogram column was computed from 0.0928 s of audio (which corresponds to a window size of 2,048 samples), and has a Hann window applied. The columns were extracted from the audio signal at a frequency of 21.53 Hz (or equivalently with a hop length of 1,024 audio samples, on our 22,050 Hz audio). Before use in the classifier, the values of the spectrogram S was converted to a log-scale representation, using the formula $\ln(A + B * S)$. For CityBioNet the parameters $A = 0.001$ and $B = 10.0$ were used, while for CityAnthroNet the parameters $A = 0.025$ and $B = 2.0$ were used. These parameters were chosen manually to emphasise biotic and anthropogenic sounds by visually inspecting the transformed spectrograms on a selection of spectrograms taken from $\text{CitySounds2017}_{\text{train}}$.

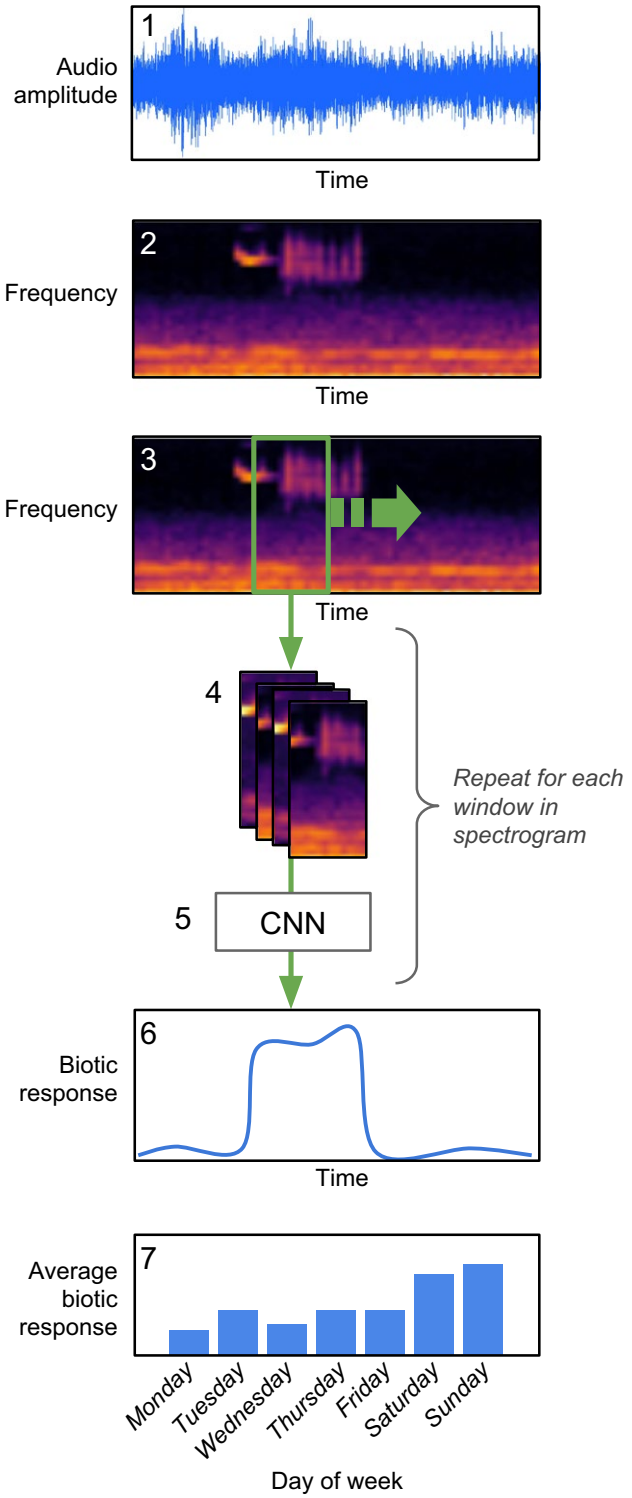

FIGURE 1 The CityNet analysis pipeline for measuring biotic and anthropogenic acoustic activity. Raw audio (1), recorded in the field, is converted to a spectrogram representation (2). A sliding window is run across the time dimension, and a window of the spectrogram extracted at each step (3). This spectrogram window is preprocessed with four different normalisation strategies, and the results concatenated. This stack of spectrograms is passed through a CNN (5), which was trained on $\text{CitySounds2017}_{\text{train}}$. The CNN gives, at each 1-s time step, a prediction of the presence/absence of biotic or anthropogenic acoustic activity (6). Finally, these per-time-step measures can be aggregated to give summaries over time or space (7)

3. *Extract window from spectrogram:* A single input to the CNN comprised a short spectrogram chunk *Ws*, 21 columns in width, representing 1 s of audio.
4. *Apply different normalisation strategies:* There are many different methods for preprocessing spectrograms before they are used in ML; for example, whitening (Lee, Pham, Largman, & Ng, 2009) and subtraction of mean values along each frequency bin (Aide et al., 2013). CNNs are able to accept inputs with multiple channels of data, for example, the red, green, and blue channels of a colour image. We exploited the multiple input channel capability of our CNN by providing as input four spectrograms each preprocessed using a different normalisation strategy (see Supplementary Methods), which gave considerable improvements to network accuracy above any single normalisation scheme in isolation. After applying different normalisation strategies, the input to the network consisted of a 32 × 21 × 4 tensor.
5. *Apply CNN classifier:* As described above, classification was performed with a CNN, whose parameters were learnt from training data. The CNN comprised a series of layers, each of which modified its input data with parameterised mathematical operations which were optimised to improve classification performance during training (see Supplementary Methods for details). The final layer produced the prediction of presence or absence of biotic or anthropogenic sound. To increase performance we trained an ensemble of five CNNs for each task. The final prediction was an average of the predictions from each network in the ensemble.
6. *Make prediction for each moment in time:* At test time, steps (3–5) were repeated independently for CityBioNet and CityAnthroNet to predict the presence/absence of biotic and anthropogenic sound in every 1 s chunk throughout the audio file, allowing each chunk to be categorised into one of four states (Figure 2).
7. *Summarise:* Where appropriate, the chunk-level predictions were summarised to gain insights into trends over time and space. For example, predicted activity levels for each half-hour window could be averaged to inspect the level of biotic and anthropogenic activity at different times of day.

The ML pipeline was written in PYTHON v.2.7.12 (Python Software Foundation, 2016) using THEANO v.0.9.0 (The Theano Development Team, et al. 2016) and LASAGNE v.0.2 (Dieleman et al., 2015) for ML and LIBROSA v.0.4.2 (McFee et al., 2015) for audio processing.

2.1 | Acoustic dataset

We selected 63 green infrastructure (GI) sites in and around Greater London, UK to collect audio data to train and test the CityNet

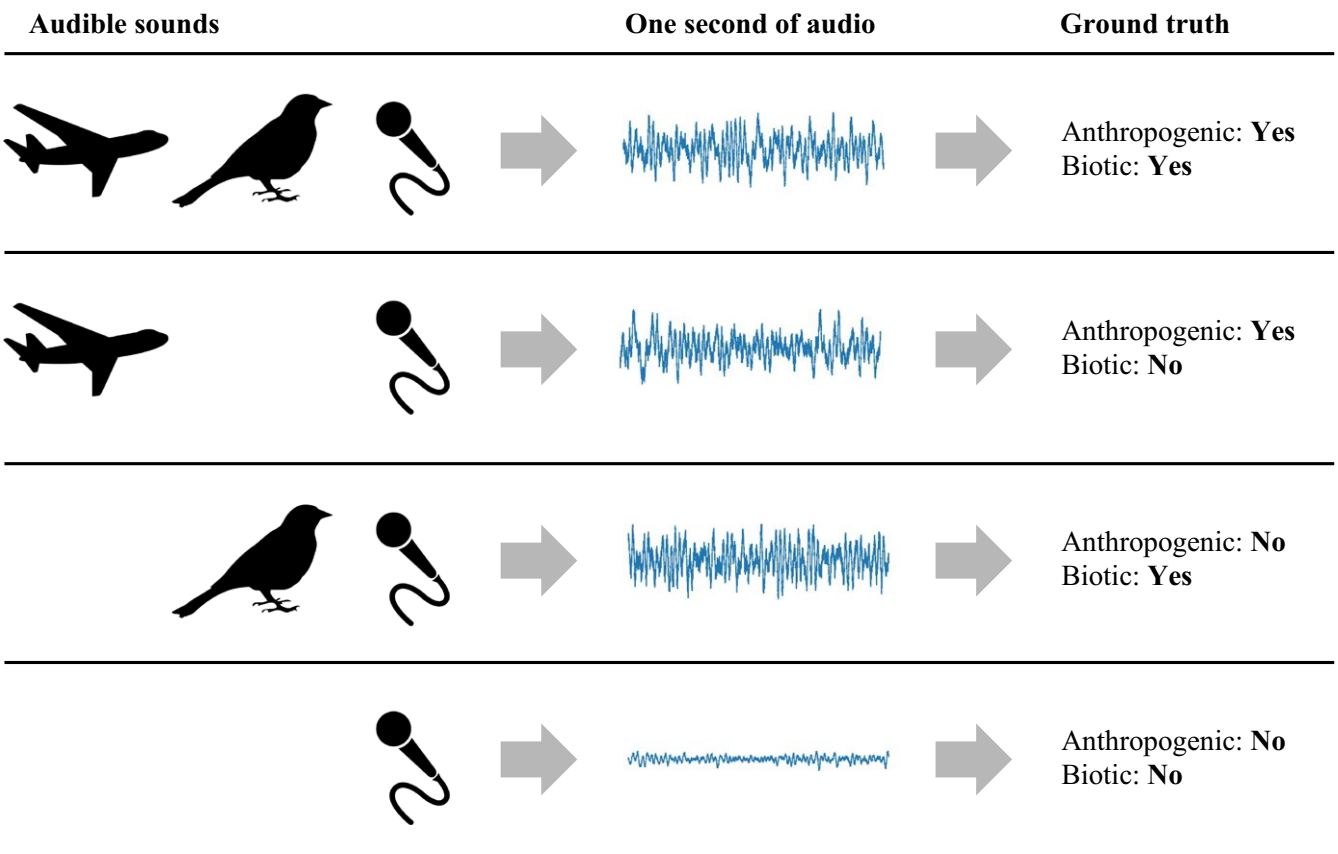

FIGURE 2 The four acoustic states predicted by the CityNet algorithms. Each 1 s chunk of audio may contain anthropogenic and biotic sound (top row), just anthropogenic sound (second row), just biotic sound (third row), or neither biotic nor anthropogenic sound (final row). CityBioNet and CityAnthroNet were independently used to detect presence or absence of biotic and anthropogenic sounds, allowing each chunk of audio to be categorised into one of four states

algorithms. These sites represent a range of GI in and around Greater London in terms of GI type, size and urban intensity. Each site was sampled for seven consecutive days across the months of May to October between 2013 and 2015 (Figure 3, Supporting Information Table S1). Sampling was conducted to ensure that each urban intensity class was surveyed within each month between May and October. At each location, a Song Meter SM2+ digital audio field sensor (Wildlife Acoustics, Inc., Concord, MA, USA) was deployed, recording sound between 0 and 12 kHz at a 24 kHz sample rate. The sensor was equipped with a single omnidirectional microphone (frequency response: −35 ± 4 dB) oriented horizontally at a height of 1 m. Files were saved in .wav format onto a SD card. Audio was recorded in computationally manageable chunks of 29 min of every 30 min (23.2 hr of recording per day), which were divided into 1-min audio files using Slice Audio File Splitter (NCH Software Inc. 2014), leading to a total of 613,872 discrete minutes of audio recording (9,744 min for each of the 63 sites). This constituted the CitySounds2017 dataset.

2.2 | Acoustic training dataset

To create our training dataset ($\text{CitySounds2017}_{\text{train}}$) we randomly selected 1,100 1-min recordings from a random 70% of the study sites (44 sites, 25 recordings from each site). A.F. manually annotated the spectrograms of each recording, computed as the log magnitude of a discrete Fourier transform (non-overlapping Hamming window size = 720 samples = 10 ms), using AudioTagger (available at https://github.com/groakat/AudioTagger). Spectrograms were annotated by localising the time and frequency bands of discrete sounds by drawing bounding boxes as tightly as visually possible within spectrograms displayed on a Dell UltraSharp 61 cm LED monitor. Types of sound, such as "invertebrate", "rain", and "road traffic", were identified by looking for typical patterns in spectrograms (Supporting Information Figure S1), and by listening to the audio samples represented in the annotated parts of the spectrogram. Categories of sounds were then grouped into biotic, anthropogenic and geophonic classes following Pijanowski et al. (2011), where we define biotic as sounds generated by non-human biotic organisms, anthropogenic as sounds associated with human activities, and geophonic as non-biological ambient sounds, for example, wind and rain (see Supporting Information Table S2 for sound categories and sample sizes). There were not enough examples to create a classifier for these separate categories ("invertebrate", "rain", etc.) with current ML techniques. However, advances in low-shot learning (e.g., Wang, Girshick, Hebert, & Hariharan, 2018) may allow our annotations to be used to create such a fine-grained classifier in the future.

2.3 | Acoustic testing dataset and evaluation

To evaluate the performance of the CityNet algorithms, we created a testing dataset ($\text{CitySounds2017}_{\text{test}}$) by selecting 40 1-min recordings from CitySounds2017 from the remaining 30% of sites (19 sites, average 2 ± 1 recordings per site). The testing dataset was randomly selected from the remaining sites so that the four potential acoustic states that CityNet algorithms can predict (Figure 2) were represented. $\text{CitySounds2017}_{\text{test}}$ was sampled from different recording sites to $\text{CitySounds2017}_{\text{train}}$ to demonstrate that the CityNet algorithms generalise to sounds recorded at new site locations (Figure 3, Supporting Information Table S1). Choosing the testing dataset from the same, rather than different, sites made little impact on the performance of the algorithms (see further details in Supporting Information Section S1: Supplementary Methods), suggesting that the results generalised well. To optimise the quality of the annotations in $\text{CitySounds2017}_{\text{test}}$, we selected five human labellers to separately annotate the sounds within the audio recordings (using the same methods as above) to create a single annotated test dataset. Conflicts were resolved using a majority rule, and in cases where there was no majority, we used our own judgement on the most suitable classification. Overall, we found the labellers to be reasonably consistent—at least four labellers agreed on the classification 86.0% of the time, and all five labellers agreed 64.7% of the time. Most of the disagreements occurred when labelling quieter anthropic sounds such as distant aeroplanes. We used multiple labellers to produce $\text{CitySounds2017}_{\text{test}}$ to optimize the quality of the annotations. Due to the resource intensity of this technique it was not used to create the larger $\text{CitySounds2017}_{\text{train}}$ dataset. Our CitySounds2017 annotated training and testing datasets are available at https://doi.org/10.6084/m9.figshare.c.3904006.v1.

Using the $\text{CitySounds2017}_{\text{test}}$ dataset, we separately assessed the performance of the two CityNet algorithms, CityBioNet and CityAnthroNet, using two measures: precision and recall. The CityBioNet and CityAnthroNet algorithms give an estimate of the level of biotic or anthropogenic acoustic activity for each 1-s audio chunk as a number between 0 and 1. Different thresholds could be used to convert these activations into sound category assignments (e.g., "sound present" or "sound absent"). At each threshold, a value of precision and recall was computed. Precision is the fraction of the 1-s chunks that contained the sound (according to the annotations in $\text{CitySounds2017}_{\text{test}}$) which were also correctly identified as containing the sound under that threshold. Recall is the fraction of 1-s chunks labelled as containing the sound that were retrieved by the algorithm (Supporting Information Figure S2). The threshold was swept between 0 and 1 and the resulting values of precision and recall were plotted as a precision-recall curve. Summary statistics were computed for the average precision under all the threshold values and the recall when the threshold chosen gave a precision of 0.95. The fraction of true positives, false positives, true negatives and false negatives were also computed, using the same threshold. These analyses were conducted in PYTHON v.2.7.12 (Python Software Foundation, 2016) using SCIKIT-LEARN v.0.18.1 (Pedregosa et al., 2011) and MATPLOTLIB v.1.5.1 (Hunter, 2007). The experiments were run on a machine running Ubuntu 16.04 with a 3.60 GHz Xeon processor, 64 GB of RAM and a 2 GB Nvdia GPU. With that processing speed, 60 s of audio can be classified with both CityAnthroNet and CityBioNet in 0.977s, with 0.14 s used for computing spectrograms while the remaining 0.86 s is spent running the networks.

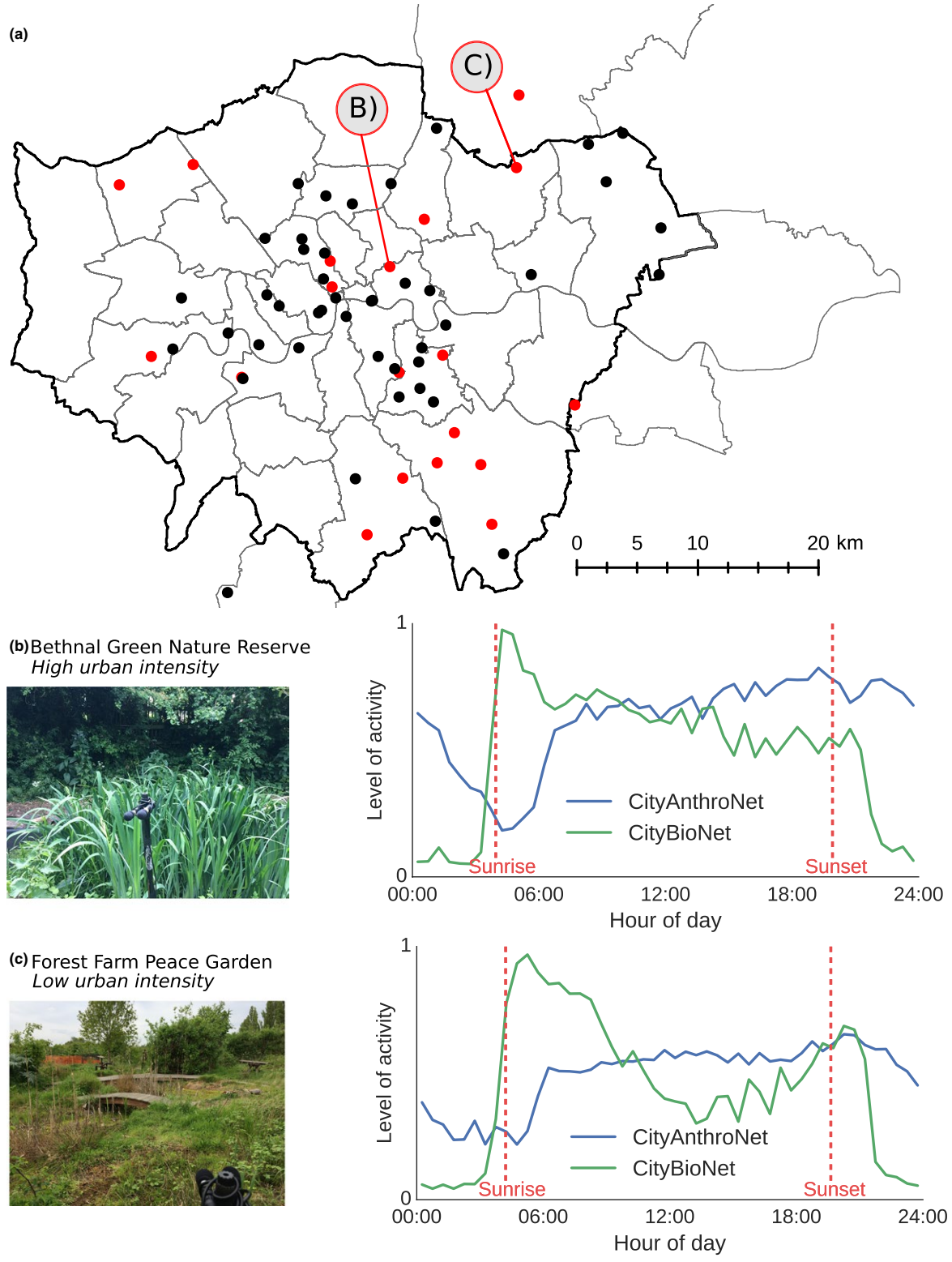

FIGURE 3 Location of study sites and average daily acoustic patterns at two sites along an urbanisation gradient. Points in (a) represent locations used for the training dataset, CitySounds2017$_{train}$ (black) and testing dataset, CitySounds2017$_{test}$ (red). Here CityNet was run across the entire 7 days of recording at two sites of high (b) and low (c) urban intensity to predict the presence/absence of biotic and anthropogenic sound at each second of the week using a threshold of 0.5. The predicted number of seconds containing biotic and anthropogenic sound for each half-hour period was averaged over the week to produce average daily patterns of acoustic activity. Greater London boundary indicated with bold line. Boundary data from the UK Census (http://www.ons.gov.uk/, accessed 04/11/2014)

2.4 | Competing algorithms

We also compared the precision and recall of the CityNet algorithms to acoustic measures produced by four Acoustic Indices: Acoustic Complexity Index (ACI) (Pieretti et al., 2011), Acoustic Diversity Index (ADI) (Villanueva-Rivera et al., 2011), Bioacoustic Index (BI) (Boelman et al., 2007), and Normalised Difference Soundscape Index (NDSI) (Kasten et al., 2012). The NDSI generates a measure of anthropogenic disturbance according to the formula:

$$\mathrm{NDSI} = \frac{\mathrm{NDSI_{bio}} - \mathrm{NDSI_{anthro}}}{\mathrm{NDSI_{bio}} + \mathrm{NDSI_{anthro}}} \quad (1)$$

where $\mathrm{NDSI_{bio}}$ and $\mathrm{NDSI_{anthro}}$ are the total estimated power spectral density for the largest 1 kHz biotic sound bin (2–8 kHz) and the anthropogenic sound bin (1–2 kHz) respectively. Rather than compare CityNet to the NDSI, we compared the biotic ($\mathrm{NDSI_{bio}}$) and anthropogenic ($\mathrm{NDSI_{anthro}}$) elements of the NDSI to the measures produced by CityBioNet and CityAnthroNet, respectively, as these were more comparable. As the Acoustic Indices are all designed to give a summary of acoustic activity for an entire file, they were analysed on the $\mathrm{CitySounds2017_{test}}$ dataset by treating each 1-s chunk of audio as a separate sound file to enable direct comparisons to CityNet. The Acoustic Indices' measures do not have a natural threshold for classification into biotic/non-biotic sound, meaning we could not calculate confusion matrices. However, a threshold between their lowest value and their highest value was used in combination with the range of precision and recall values to form precision-recall curves. All Acoustic Indices were calculated in R v.3.4.1 (R Core Team, 2017) using the SEEWAVE v.1.7.6 (Sueur, Aubin, & Simonis, 2008) and SOUNDECOLOGY v.1.2 (Villanueva-Rivera & Pijanowski, 2014) packages.

The precision and recall of CityBioNet was also compared to bulbul (Grill & Schlüter, 2017), an algorithm for detecting bird sounds in entire audio recordings in order to summarise avian acoustic activity which was the winning entry in the 2016–2017 Bird Audio Detection challenge (Stowell, Wood, Stylianou, & Glotin, 2016). Like CityNet, bulbul is a CNN-based classifier which uses spectrograms as input. However, it does not use the same normalisation strategies as CityNet, and it was not trained on data from noisy, urban environments. Bulbul was applied to each second of audio data in $\mathrm{CitySounds2017_{test}}$, using the pretrained model provided by the authors together with their code.

2.5 | Impact of non-biotic sounds

We conducted additional analysis on the non-biotic sounds that affect the predictions of CityBioNet and bulbul, as these were found to be the best performing algorithms for measuring biotic sound. To do this, we created subsets of the $\mathrm{CitySounds2017_{test}}$ dataset comprising all the seconds that contained a range of non-biotic sounds, for example, a road traffic data subset containing all of the seconds in $\mathrm{CitySounds2017_{test}}$ where the sound of road traffic was present. We then used a Chi-square test to identify significant differences in the proportion of seconds in which the presence/absence of biotic sound at threshold 0.5 was correctly predicted in the full and subset datasets by each algorithm, and the Cramer's *V* statistic was used to assess the effect size of differences as this is unbiased by sample sizes (Cohen, 1992). These analyses were conducted in R v.3.4.1 (R Core Team, 2017).

2.6 | Ecological application

We used CityNet to generate daily average patterns of biotic and anthropogenic acoustic activity for two study sites across an urbanisation gradient (sites E29RR and IG62XL with high and low urbanisation, respectively, Supporting Information Table S1). To control for the date of recording; both sites were surveyed between May and June 2015. CityNet was run over the entire 7 days of recordings from each site to predict the presence/absence of biotic and anthropogenic sound for every 1-s audio chunk using a threshold of 0.5. Measures of biotic and anthropogenic activity were created for each half hour window between midnight and midnight by averaging the predicted number of seconds containing biotic or anthropogenic sound within that window over the entire week.

3 | RESULTS

3.1 | Acoustic performance

CityBioNet had an average precision of 0.934 and recall of 0.710 at 0.95 precision, while CityAnthroNet had an average precision of 0.977 and recall of 0.858 at 0.95 precision (Table 1, Figure 4). In comparison the ACI, ADI, BI and $\mathrm{NDSI_{bio}}$ had a lower average precision (0.663, 0.439, 0.516, and 0.503 respectively) and failed to achieve 0.95 precision at any threshold value. CityBioNet also outperformed bulbul which had an average precision of 0.872 and recall at 0.95 of 0.398 (Table 1). In comparison to CityAnthroNet, the $\mathrm{NDSI_{anthro}}$ had a comparable average precision (0.975 vs. 0.977), but a lower recall at 0.95 precision (0.815 vs. 0.858). CityBioNet correctly predicted the presence of biotic sound (True Positives) in a greater proportion of audio data than bulbul (33.2% in comparison with 18.5% for CityBioNet and bulbul respectively) (Table 2). However, CityBioNet failed to correctly predict the presence of biotic sound (False Negatives) in 13.5% of recordings in comparison with 28.0% incorrect predictions by bulbul. CityBioNet correctly predicted the absence of biotic sound (True Negatives) in 51.6% of the audio data in comparison with 52.6% for bulbul, and CityBioNet failed to correctly predict the absence of biotic sound (False Positives) in 1.7% of audio data in comparison with 1.0% incorrect predictions by bulbul (Table 2).

3.2 | Impacts of non-biotic sounds

CityBioNet was strongly (Cramer's *V* effect size >0.5) negatively affected by mechanical sound (the presence/absence of biotic sound was correctly predicted in 28.60% less of the data when mechanical sounds were also present) (Table 3). Bulbul was moderately (Cramer's *V* effect size 0.1–0.5) negatively affected by the sound of

TABLE 1 Average precision and recall results for CityNet and competing algorithms for each 1-s audio chunk in the $CitySounds2017_{test}$ dataset. Recall results are presented at 0.95 precision, however, some methods did not achieve 0.95 precision under any threshold. Recall values for these are methods (in parentheses) are given as the recall that gave the highest precision. Higher values are better for both metrics. The highest values in each section are shown in bold. ACI represents Acoustic Complexity Index, ADI Acoustic Diversity Index, BI Bioacoustic Index, and $NDSI_{bio}$ and $NDSI_{anthro}$ biotic and anthropogenic Normalised Difference Soundscape Index respectively

Acoustic measures	Recall at 0.95 precision	Average precision
Biotic		
CityBioNet	**0.710**	**0.934**
Bulbul	0.398	0.872
ACI	(0.000)	0.663
ADI	(0.001)	0.439
BI	(0.002)	0.516
$NDSI_{bio}$	(0.000)	0.503
Anthropogenic		
CityAnthroNet	**0.858**	**0.977**
$NDSI_{anthro}$	0.815	0.975

air traffic and wind (the presence/absence of biotic sound was correctly predicted in 5.34% and 6.93% less of the data when air traffic and wind sounds were also present in recordings respectively).

3.3 | Ecological application

CityNet produced realistic patterns of biotic and anthropogenic acoustic activity in the urban soundscape at two study sites of low and high urban intensity (Figure 3b,c). At both sites, biotic acoustic activity peaked just after sunrise and declined rapidly after sunset. A second peak of biotic acoustic activity was recorded at sunset at the low urban intensity site but not at the high urban intensity site. At both sites anthropogenic acoustic activity rose sharply after sunrise, remained constant throughout the day and declined after sunset.

4 | DISCUSSION

Both CityBioNet and CityAnthroNet outperformed the competing algorithms on the $CitySounds2017_{test}$ dataset. CityBioNet performed better than bulbul on noisy recordings from the urban environment; it was robust to more non-biotic sounds, including road traffic, air traffic, and rain. Being robust to the sound of road traffic supports the suitability of CityBioNet for use in cities, as the urban soundscape is dominated by the sound of road traffic (Fairbrass et al., 2017) which has been shown to bias several of the Acoustic Indices tested here (Fairbrass et al., 2017; Fuller et al., 2015). The sound of rain has also been shown to bias several Acoustic Indices (Depraetere et al., 2012; Fairbrass et al., 2017; Gasc et al., 2015) and the development of a method that is robust to this sound is a considerable contribution to the field of ecoacoustics. The urban biotic soundscape is dominated by the sounds emitted by birds (Fairbrass et al., 2017), and the good performance of bulbul, an algorithm for measuring exclusively bird sounds, on the $CitySounds2017_{test}$ dataset, confirms this. Birds are used as indicator species in existing urban biodiversity monitoring schemes (Kohsaka et al., 2013) using data collected from traditional forms of biodiversity survey. The algorithms developed here could be used to support such existing schemes by making it easier to collect data.

CityNet is the only method currently available for measuring both biotic and anthropogenic acoustic activity using a single system in noisy audio data from urban environments. There is increasing evidence that anthropogenic noise affects wildlife in a variety of ways including altering communication behaviour (Gil & Brumm, 2014) and habitat use (Deichmann, Hernández-Serna, Delgado C, Campos-Cerqueira, & Aide, 2017). However, these investigations are limited in scale by the use of resource intensive methods of measuring biotic and anthropogenic sound in the environment or from audio data. Others rely on Acoustic Indices (Pieretti & Farina, 2013) which have been shown to be unreliable in acoustically disturbed environments (Fairbrass et al., 2017). CityNet could facilitate the investigation of the impacts of anthropogenic activities on wildlife populations at scales not currently possible with traditional acoustic analysis methods. The detection space of soniferous species is determined by a number of factors including habitat characteristics, calling frequency (kHz), animal height, direction, and speed of travel (Darras, Pütz, Fahrurrozi, Rembold, & Tscharntke, 2016). These and other factors related to the audio recording equipment used should be taken into consideration when designing acoustic biodiversity investigations.

CityBioNet clearly outperformed all the Acoustic Indices tested, but the difference in performance between CityAnthroNet and the competing algorithm for measuring anthropogenic acoustic activity ($NDSI_{anthro}$) was much less marked. These results suggest that the measurement of biotic sound in noisy audio data from urban environments requires more sophisticated algorithms than the measurement of anthropogenic sound. Possibly anthropogenic sounds are more easily separable from other sounds in frequency space, a theory which is the basis of a number of Acoustic Indices (Boelman et al., 2007; Kasten et al., 2012), facilitating the use of human defined algorithms such as $NDSI_{anthro}$. Whereas, because biotic sounds occur in a frequency space shared with anthropogenic and geophonic sounds (Fairbrass et al., 2017), algorithms such as Acoustic Indices which only use a small number of features to discriminate sounds are not sufficient for use in cities. Therefore, ML algorithms which are able to utilise larger numbers of features to discriminate sounds, such as the CNNs implemented in the CityNet system, are better able to detect biotic sounds in recordings that also contain non-biotic sounds. Noise reduction algorithms can be used to process audio recordings prior to signal detection and a recent unsupervised method developed by Lin, Fang, and Tsao (2017) to separate biological sounds from long recordings could be used as a preprocessing step to further improve

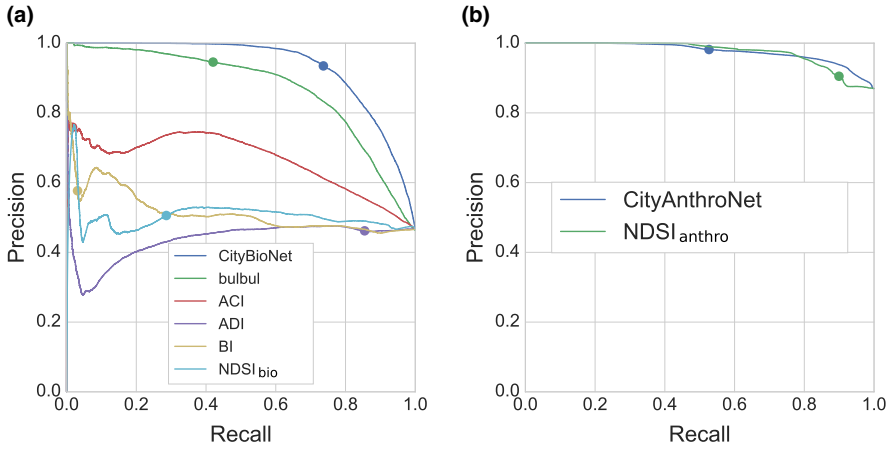

FIGURE 4 Precision-recall curves for CityNet and competing algorithms predicting (a) biotic and (b) anthropogenic acoustic activity for each 1-s audio chunk in the CitySounds2017$_{test}$ dataset. Dots indicate the precision and recall values at a threshold value of 0.5. ACI, Acoustic Complexity Index; ADI, Acoustic Diversity Index; BI, Bioacoustic Index; NDSI$_{bio}$ and NDSI$_{anthro}$, biotic and anthropogenic Normalised Difference Soundscape Index respectively

TABLE 2 Comparison of the predicted acoustic performance of the CityBioNet and bulbul algorithms for each 1-s audio chunk in the CitySounds2017$_{test}$ dataset. Numbers report the percentage of 1-s audio clips in the CitySounds2017$_{test}$ dataset predicted either correctly (True Positives and True Negatives) or incorrectly (False Positives and False Negatives) as containing biotic (rows 1 and 2) or anthropogenic (row 3) sound. To create these measures, the predictions from the classifiers were converted to binary classifications using a threshold that gives a precision of 0.95

	True Positive (%)	True Negative (%)	False Negative (%)	False Positive (%)
CityBioNet	33.16	51.59	13.52	1.74
Bulbul	18.47	52.59	27.96	0.97
CityAnthroNet	74.57	9.09	12.41	3.93

CityNet's performance. Using test data drawn from the same dataset as the training data may have biased our results in favour of CityNet. Future work could compare the methods assessed here using alternative test data drawn from an independent dataset.

Low-cost acoustic sensors and algorithms for the automatic measurement of biotic sound in audio data are facilitating the assessment and monitoring of biodiversity at large temporal and spatial scales (Sueur & Farina, 2015), but to date this technology has only been deployed in nonurban environments (e.g. Aide et al., 2013). In cities, the availability of mains power and Wifi connections is supporting the development of the urban Internet of Things (IoT) using sensors integrated into existing infrastructure to monitor environmental factors including air pollution, noise levels, and energy use (Zanella, Bui, Castellani, Vangelista, & Zorzi, 2014). The CityNet system could be integrated into an IoT sensing network to facilitate large-scale urban environmental assessment. Large-scale deployment of algorithms such as CityNet requires low power usage and fast running times. One way to help to achieve this aim would be to combine the two networks (CityBioNet and CityAnthroNet) into one CNN which predicts both biotic and anthropogenic acoustic activity simultaneously.

An expansion of CityNet to ultrasonic frequencies would increase the generality of the tool as it could be used to monitor species in cities that emit sounds at frequencies higher than 12 kHz such as bats and some invertebrates. Bats are frequently used as ecological indicators because they are sensitive to environmental changes (Walters et al., 2013). Acoustic methods are commonly used to monitor bat populations using passive ultrasonic recorders meaning bat researchers and conservationists are faced with the challenge of extracting meaningful information from large volumes of audio data. The development of automated methods for measuring bat calls in ultrasonic data has focused to date on the identification of bat species calls and many algorithms are proprietary (e.g., Szewczak, 2010; Wildlife Acoustics, 2017). The development of an open-source algorithm that produces community-level measures of bats would be a valuable addition to the toolbox of bat researchers and conservationists.

Retraining CityNet with labelled audio data from other cities would make it possible to use the system to monitor urban biotic and anthropogenic acoustic activity more widely. However, as London is a large and heterogeneous city, CityNet has been trained using a dataset containing sounds that characterise a wide range of urban environments. Our data collection was restricted to a single week at each study site, which limits our ability to assess the ability of CityNet system to detect environmental changes. Future work should focus on the collection of longitudinal acoustic data to assess the sensitivity of the algorithms to detect environmental changes. Our use of human labellers would have introduced subjectivity and bias into our dataset. The task of annotating large audio datasets from acoustically complex urban environments is highly resource intensive, a problem

TABLE 3 Impact of non-biotic sounds on the CityBioNet and bulbul predictions. Values represent differences in the proportion of 1-s audio chunks in the full CitySounds2017$_{test}$ dataset (40 min) and the subset datasets (size in time indicated in left-hand column) in which the presence/absence of biotic sound was correctly predicted by both algorithms (Chi-square test statistic for difference in proportions of successes in each dataset, and Cramer's *V* effect size measure). Bold type indicates 95% significance Chi-square test statistic

Sound type	CityBioNet	Bulbul
Anthropogenic		
Air traffic (9 m 4 s)	**−2.11 (30.35, 0.05)***	**−5.34 (162.73, 0.12)****
Mechanical (11 s)	**−28.60 (134.38, 0.77)*****	0.02 (0.01, 0.01)*
Road traffic (29 m 15 s)	**0.79 (10.15, 0.02)***	**1.41 (27.67, 0.03)***
Siren (1 m 21 s)	2.28 (5.73, 0.06)*	**3.70 (12.95, 0.09)***
Geophonic		
Rain (2 m 44 s)	−0.77 (1.29, 0.02)*	−1.51 (4.17, 0.04)*
Wind (53 s)	0.76 (0.47, 0.02)*	**−6.93 (33.11, 0.17)****

Effect sizes indicated as *<0.1, **0.1–0.5, ***>0.5.

which has been recently tackled with citizen scientists to create the UrbanSound and UrbanSound8k datasets using audio data from New York city, USA (Salamon, Jacoby, & Bello, 2014). These comprise short snippets of 10 different urban sounds such as jackhammers, engines idling, and gunshots. These datasets do not fully represent the characteristics of urban soundscapes for three reasons. First, they assume only one class of sound is present at each time, while in fact multiple sound types can be present at one time (consider a bird singing while an aeroplane flies overhead). Second, they only include anthropogenic sounds, while CityNet measures both anthropogenic and biotic sounds. Finally, each file in these datasets has a sound present, whereas urban soundscapes contain many periods of silence or geophonic sounds, two important states which are not present in UrbanSound and UrbanSound8k. Due to these factors, these datasets are unsuitable for the purpose of this research project, although recent work has overcome a few of these shortcoming through the annual Detection and Classification of Acoustic Scenes and Events challenges, and using synthesised soundscape data (Mesaros et al., 2017; Salamon, MacConnell, Cartwright, Li, & Bello, 2017). This highlights the need for an internationally coordinated effort to create a consistently labelled audio dataset from cities to support the development of automated urban environmental assessment systems with international application. There were a number of sounds that occurred too infrequently to analyse their impact on CityNet's predictions. As more labelled data become available it will be possible to investigate the impact of these rarer sounds on CityNet's predictions and also to generate more complex acoustic measurements, such as acoustic activity of specific sound types or acoustic diversity.

5 | CONCLUSIONS

The CityNet system for measuring biotic and anthropogenic acoustic activity in noisy urban audio data outperformed the state-of-the-art algorithms for measuring biotic and anthropogenic sound in entire audio recordings. Integrated into an IoT network for recording and analysing audio data in cities it could facilitate urban environmental assessment at greater scales than has been possible to date using traditional methods of biodiversity assessment. We make our system available open source in combination with two expertly annotated urban soundscape datasets to facilitate future research development in this field.

ACKNOWLEDGEMENTS

We thank multiple site owners and managers for supporting the study by providing access to recording sites, and multiple acoustic annotators and a transport expert for help creating the CitySounds2017 dataset. We were financially supported by a BHP Billiton Sustainable Resources for Sustainable Cities Catalyst Grant and by the Engineering and Physical Sciences Research Council (EPSRC) through a doctoral training grant (EP/G037698/1) to H.T., and EPSRC grant (EP/K015664/1) to K.E.J, G.J.B., and M.F.

AUTHORS' CONTRIBUTIONS

A.J.F., M.F., H.T., and K.E.J. conceived ideas and designed methodology; A.J.F. collected the data; A.J.F. and M.F. analysed the data and led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

DATA ACCESSIBILITY

All recordings and annotations in the CitySounds2017 dataset are available on Figshare (https://doi.org/10.6084/m9.figshare.c.3904006.v1) and all Python code underlying the CityNet algorithms are available on GitHub (https://github.com/mdfirman/CityNet) and has a Zenodo https://doi.org/10.5281/zenodo.1463057.

ORCID

Alison J. Fairbrass http://orcid.org/0000-0002-4907-9683
Gabriel J. Brostow http://orcid.org/0000-0001-8472-3828
Helena Titheridge http://orcid.org/0000-0003-2194-1531
Kate E. Jones http://orcid.org/0000-0001-5231-3293

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. *Ecological Informatics*, *4*, 206–214. https://doi.org/10.1016/j.ecoinf.2009.06.005

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. https://doi.org/10.7717/peerj.103

Aronson, M. F. J., La Sorte, F. A., Nilon, C. H., Katti, M., Goddard, M. A., Lepczyk, C. A., … Winter, M. (2014). A global analysis of the impacts of urbanization on bird and plant diversity reveals key anthropogenic drivers. *Proceedings of the Royal Society B: Biological Sciences*, *281*, 20133330. https://doi.org/10.1098/rspb.2013.3330

Beninde, J., Veith, M., & Hochkirch, A. (2015). Biodiversity in cities needs space: A meta-analysis of factors determining intra-urban biodiversity variation. *Ecology Letters*, *18*, 581–592. https://doi.org/10.1111/ele.12427

Boelman, N. T., Asner, G. P., Hart, P. J., & Martin, R. E. (2007). Multi-trophic invasion resistance in Hawaii: Bioacoustics, field surveys, and airborne remote sensing. *Ecological Applications*, *17*, 2137–2144. https://doi.org/10.1890/07-0004.1

Chesmore, E., & Ohya, E. (2004). Automated identification of field-recorded songs of four British grasshoppers using bioacoustic signal recognition. *Bulletin of Entomological Research*, *94*, 319–330.

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science*, *1*, 98–101. https://doi.org/10.1111/1467-8721.ep10768783

Crouse, D. L., Pinault, L., Balram, A., Hystad, P., Peters, P. A., Chen, H., … Villeneuve, P. J. (2017). Urban greenness and mortality in Canada's largest cities: A national cohort study. *The Lancet Planetary Health*, *1*, e289–e297. https://doi.org/10.1016/S2542-5196(17)30118-3

Darras, K., Pütz, P., Fahrurrozi, , Rembold, K., & Tscharntke, T. (2016). Measuring sound detection spaces for acoustic animal sampling and monitoring. *Biological Conservation*, *201*, 29–37. https://doi.org/10.1016/j.biocon.2016.06.021

Deichmann, J. L., Hernández-Serna, A., Delgado C, J. A., Campos-Cerqueira, M., & Aide, T. M. (2017). Soundscape analysis and acoustic monitoring document impacts of natural gas exploration on biodiversity in a tropical forest. *Ecological Indicators*, *74*, 39–48. https://doi.org/10.1016/j.ecolind.2016.11.002

Depraetere, M., Pavoine, S., Jiguet, F., Gasc, A., Duvail, S., & Sueur, J. (2012). Monitoring animal diversity using acoustic indices: Implementation in a temperate woodland. *Ecological Indicators*, *13*, 46–54. https://doi.org/10.1016/j.ecolind.2011.05.006

Dieleman, S., Schlüter, J., Raffel, C., Olson, E., Sønderby, S. K., Nouri, D., … Degrave, J. (2015). *Lasagne*, https://doi.org/10.5281/zenodo.27878

Digby, A., Towsey, M., Bell, B. D., & Teal, P. D. (2013). A practical comparison of manual and autonomous methods for acoustic monitoring. *Methods in Ecology and Evolution*, *4*, 675–683. https://doi.org/10.1111/2041-210X.12060

Faeth, S. H., Bang, C., & Saari, S. (2011). Urban biodiversity: Patterns and mechanisms. *Year in Ecology and Conservation Biology*, *1223*, 69–81.

Fairbrass, A. J., Rennett, P., Williams, C., Titheridge, H., & Jones, K. E. (2017). Biases of acoustic indices measuring biodiversity in urban areas. *Ecological Indicators*, *83*, 169–177. https://doi.org/10.1016/j.ecolind.2017.07.064

Farinha-Marques, P., Lameiras, J., Fernandes, C., Silva, S., & Guilherme, F. (2011). Urban biodiversity: A review of current concepts and contributions to multidisciplinary approaches. *Innovation: The European Journal of Social Science Research*, *24*, 247–271.

Fuller, S., Axel, A. C., Tucker, D., & Gage, S. H. (2015). Connecting soundscape to landscape: Which acoustic index best describes landscape configuration? *Ecological Indicators*, *58*, 207–215. https://doi.org/10.1016/j.ecolind.2015.05.057

Gasc, A., Pavoine, S., Lellouch, L., Grandcolas, P., & Sueur, J. (2015). Acoustic indices for biodiversity assessments: Analyses of bias based on simulated bird assemblages and recommendations for field surveys. *Biological Conservation*, *191*, 306–312. https://doi.org/10.1016/j.biocon.2015.06.018

Gil, D., & Brumm, H. (2014). Acoustic communication in the urban environment: Patterns, mechanisms, and potential consequences of avian song adjustments. In D. Gil, & H. Brumm (Eds.), *Avian urban ecology* (pp. 69–83). Oxford, UK: Oxford University Press.

Grill, T., & Schlüter, J. (2017). *Two convolutional neural networks for bird detection in audio signals*. 25th European Signal Processing Conference (EUSIPCO2017), Kos, Greece. https://doi.org/10.23919/EUSIPCO.2017.8081512

Hall, D. M., Camilo, G. R., Tonietto, R. K., Smith, D. H., Ollerton, J., Ahrné, K., … Fowler, R. (2016). The city as a refuge for insect pollinators. *Conservation Biology*, *31*, 24–29.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing In Science & Engineering*, *9*, 90–95. https://doi.org/10.1109/MCSE.2007.55

Kasten, E. P., Gage, S. H., Fox, J., & Joo, W. (2012). The remote environmental assessment laboratory's acoustic library: An archive for studying soundscape ecology. *Ecological Informatics*, *12*, 50–67. https://doi.org/10.1016/j.ecoinf.2012.08.001

Kohsaka, R., Pereira, H. M., Elmqvist, T., Chan, L., Moreno-Peñaranda, R., Morimoto, Y., … da Luz Mathias, M. (2013). Indicators for management of urban biodiversity and ecosystem services: City biodiversity index. In T. Elmqvist, M. Fragkias, J. Goodness, B. Güneralp, P. J. Marcotullio, R. I. McDonald, S. Parnell, M. Schewenius, M. Sendstad, K. C. Seto, & C. Wilkinson (Eds.), *Urbanization, biodiversity and ecosystem services: Challenges and opportunities* (pp. 699–718). Dordrecht, The Netherlands: Springer. https://doi.org/10.1007/978-94-007-7088-1

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444. https://doi.org/10.1038/nature14539

Lee, H., Pham, P., Largman, Y., & Ng, A. Y. (2009). *Unsupervised feature learning for audio classification using convolutional deep belief networks* (pp. 1096–1104). Proceedings of the 22nd International Conference on Neural Information Processing Systems, Istanbul, Turkey.

Lin, T.-H., Fang, S.-H., & Tsao, Y. (2017). Improving biodiversity assessment via unsupervised separation of biological sounds from long-duration recordings. *Scientific Reports*, *7*, 4547. https://doi.org/10.1038/s41598-017-04790-7

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). *librosa: Audio and music signal analysis in python* (pp. 18–25). Proceedings of the 14th python in science conference, Austin, TX.

Mesaros, A., Heittola, T., Diment, A., Elizalde, B., Shah, A., Vincent, E., … Virtanen, T. (2017). *DCASE 2017 challenge setup: Tasks, datasets and baseline system*. DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events.

Natural England. (2016). Links between natural environments and mental health: evidence briefing. Retrieved from http://publications.naturalengland.org.uk

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Perrot, M. D. E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pieretti, N., & Farina, A. (2013). Application of a recently introduced index for acoustic complexity to an avian soundscape with traffic noise. *The Journal of the Acoustical Society of America*, *134*, 891–900. https://doi.org/10.1121/1.4807812

Pieretti, N., Farina, A., & Morri, D. (2011). A new methodology to infer the singing activity of an avian community: The Acoustic Complexity Index (ACI). *Ecological Indicators*, *11*, 868–873. https://doi.org/10.1016/j.ecolind.2010.11.005

Pijanowski, B. C., Villanueva-Rivera, L. J., Dumyahn, S. L., Farina, A., Krause, B. L., Napoletano, B. M., … Pieretti, N. (2011). Soundscape ecology: The science of sound in the landscape. *BioScience*, *61*, 203–216. https://doi.org/10.1525/bio.2011.61.3.6

Python Software Foundation. (2016). Python Language Reference. Retrieved from http://www.python.org

R Core Team. (2017). R: A language and environment for statistical computing. Retrieved from http://www.R-project.org

Salamon, J., Jacoby, C., & Bello, J. P. (2014). *A dataset and taxonomy for urban sound research* (pp. 1041–1044). ACM MM'14. Association for Computing Machinery, Orlando, FL.

Salamon, J., MacConnell, D., Cartwright, M., Li, P., & Bello, J. P. (2017). *Scaper: A library for soundscape synthesis and augmentation*. 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY.

Stowell, D., & Plumbley, M. D. (2014). Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. *PeerJ*, *2*, e488. https://doi.org/10.7717/peerj.488

Stowell, D., Wood, M., Stylianou, Y., & Glotin, H. (2016). *Bird detection in audio: A survey and a challenge* (pp. 1–6). 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing. IEEE, Vietri sul Mare, Italy.

Sueur, J., Aubin, T., & Simonis, C. (2008). Equipment review: Seewave, a free modular tool for sound analysis and synthesis. *Bioacoustics*, *18*, 213–226. https://doi.org/10.1080/09524622.2008.9753600

Sueur, J., & Farina, A. (2015). Ecoacoustics: The Ecological Investigation and Interpretation of Environmental Sound. *Biosemiotics*, *8*, 493–502. https://doi.org/10.1007/s12304-015-9248-x

Sueur, J., Farina, A., Gasc, A., Pieretti, N., & Pavoine, S. (2014). Acoustic indices for biodiversity assessment and landscape investigation. *Acta Acustica United With Acustica*, *100*, 772–781. https://doi.org/10.3813/AAA.918757

Szewczak, J. M. (2010). SonoBat. Retrieved from www.sonobat.com

The Theano Development Team, Al-Rfou, R., Alain, G., Almahairi, A., Angermueller, C., Bahdanau, D., … Belikov, A. (2016). Theano: A Python framework for fast computation of mathematical expressions. arXiv. Retrieved from https://arxiv.org/abs/1605.02688

Towsey, M., Wimmer, J., Williamson, I., & Roe, P. (2014). The use of acoustic indices to determine avian species richness in audio-recordings of the environment. *Ecological Informatics*, *21*, 110–119. https://doi.org/10.1016/j.ecoinf.2013.11.007

UN-DESA. (2016). The World's Cities in 2016. Data Booklet. Retrieved from http://www.un.org/en/development/desa/population/

Villanueva-Rivera, L. J., & Pijanowski, B. C. (2014). Package 'soundecology'. Soundscape ecology. Retrieved from http://cran.r-project.org/web/packages/soundecology/index.html

Villanueva-Rivera, L. J., Pijanowski, B. C., Doucette, J., & Pekin, B. (2011). A primer of acoustic analysis for landscape ecologists. *Landscape Ecology*, *26*, 1233–1246. https://doi.org/10.1007/s10980-011-9636-9

Walters, C. L., Collen, A., Lucas, T., Mroz, K., Sayer, C. A., & Jones, K. E. (2013). Challenges of using bioacoustics to globally monitor bats. In R. A. Adams & S. C. Pedersen (Eds.), *Bat evolution, ecology, and conservation* (pp. 479–499). Dordrecht, The Netherlands: Springer.

Walters, C. L., Freeman, R., Collen, A., Dietz, C., Brock Fenton, M., Jones, G., … Jones, K. E. (2012). A continental-scale tool for acoustic identification of European bats. *Journal of Applied Ecology*, *49*, 1064–1074. https://doi.org/10.1111/j.1365-2664.2012.02182.x

Wang, Y.-X., Girshick, R., Hebert, M., & Hariharan, B. (2018). Low-shot learning from imaginary data. arXiv preprint arXiv:1801.05401.

Wildlife Acoustics. (2017). Kaleidoscope analysis software. Retrieved from https://www.wildlifeacoustics.com/products/kaleidoscope-software-ultrasonic

Zamora-Gutierrez, V., Lopez-Gonzalez, C., MacSwiney Gonzalez, M. C., Fenton, B., Jones, G., Kalko, E. K., … Jones, K. E. (2016). Acoustic identification of Mexican bats based on taxonomic and ecological constraints on call design. *Methods in Ecology and Evolution*, *7*, 1082–1091. https://doi.org/10.1111/2041-210X.12556

Zanella, A., Bui, N., Castellani, A., Vangelista, L., & Zorzi, M. (2014). Internet of things for smart cities. *IEEE Internet of Things Journal*, *1*, 22–32. https://doi.org/10.1109/JIOT.2014.2306328

SUPPORTING INFORMATION

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

How to cite this article: Fairbrass AJ, Firman M, Williams C, Brostow GJ, Titheridge H, Jones KE. CityNet—Deep learning tools for urban ecoacoustic assessment. *Methods Ecol Evol*. 2019;10:186–197. https://doi.org/10.1111/2041-210X.13114