Procedia Computer Science

Volume 51, 2015, Pages 640–649

ICCS 2015 International Conference On Computational Science

Clustering Acoustic Events in Environmental Recordings for Species Richness Surveys

Philip Eichinski, Laurianne Sitbon and Paul Roe
Queensland University of Technology, Brisbane, Australia
philip.eichinski@qut.edu.au

Abstract

Environmental acoustic recordings can be used to perform avian species richness surveys, whereby a trained ornithologist can observe the species present by listening to the recording. This could be made more efficient by using computational methods for iteratively selecting the richest parts of a long recording for the human observer to listen to, a process known as "smart sampling". This allows scaling up to much larger ecological datasets.

In this paper we explore computational approaches based on information and diversity of selected samples. We propose to use an event detection algorithm to estimate the amount of information present in each sample. We further propose to cluster the detected events for a better estimate of this amount of information. Additionally, we present a time dispersal approach to estimating diversity between iteratively selected samples.

Combinations of approaches were evaluated on seven 24-hour recordings that have been manually labeled by bird watchers. The results show that on average all the methods we have explored would allow annotators to observe more new species in fewer minutes compared to a baseline of random sampling at dawn.

Keywords: Ecoacoustics, environmental monitoring, species richness, cluster analysis, acoustic events

1 Introduction

Monitoring of bird species in the natural environment is important because changes in bird populations can act as an indicator of changes in the health of the ecosystem. Avian richness surveys record which species are present at a site. These surveys traditionally rely on trained ornithologists making observations on-site at the location of the survey, an undertaking that is very time consuming and subject to availability of ornithologists with local knowledge.

These problems can be solved by having the ornithologists listen to audio recordings instead of having to travel to the location of the survey. Not only is just as effective as an on-site survey (Haselmayer & Quinn, 2000; Wimmer, Towsey, Roe, & Williamson, 2013), but also affords a number of advantages: it provides a permanent objective record of the health of the environment, allows future

Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2015

doi:10.1016/j.procs.2015.05.178

verification of the species observed, and alleviates constraints created by limited availability of ornithologists.

The greatest advantage however is the potential for increased scalability through computer-assisted analysis of the audio. By deploying numerous acoustic sensors for weeks or months, potentially much more valuable information about the health of the environment could be learned compared to what standard survey techniques allow. Collecting large amounts of such environmental data is feasible, but this brings computational challenges to ensure greater outcomes with the same amount of human annotation effort.

More specifically, the annotators need to be presented with the most informative minutes of the survey in order to minimize their effort. The goal of a richness survey is to record the presence of a species only once, and therefore including samples containing vocalizations of previously observed species is of no value, even if the samples are rich. In this context, the most informative samples are those likely to contain vocalizations of species not observed until a given point in the annotation process. In other words, both informativeness and diversity must be considered when ranking all available minutes of samples to present them to the annotators in the order that would maximize their chances to detect all species present while minimizing the time they spend annotating. Such approaches are also referred to as smart sampling.

In this paper, we propose to characterize informativeness with acoustic event detection. We propose two techniques to rank the minutes in this way. The first is a ranking based on the number of acoustic events present in the minutes. This is based on the intuition that the number of acoustic events is correlated with the number of bird vocalizations, which in turn should have some relationship to the number of species present. This method does not use any information about the content or nature of the events. The second technique proposes to use clustering to take into account the informative content (features) of the events, while maintaining scalability.

In addition, we introduce diversity in terms of time difference. This is to account for the fact that birds often vocalize over multiple minutes, and therefore consecutive minutes are more likely to be similar in content.

Acoustic events are detected using an algorithm that finds the time and frequency bounds of any sound which rises more than a given level above background noise. In normal remote environmental recordings taken for this purpose, birds produce the vast majority of such acoustic events.

Our evaluation is based on 7 daylong recordings from 3 different sites, and we observe the cumulative number of species present at each of the 120 best-ranked one-minute samples in each recording.

2 Related Research

There is increasing interest in using audio recordings for species richness surveys of birds as an alternative to the traditional point-count, which require an ornithologist to visit the site of the survey numerous times (Gregory, Gibbons, & Donald, 2004). Previous research has compared traditional point-counts with surveys conducted from audio recordings, and has demonstrated that the audio recording method generally performs better (Haselmayer & Quinn, 2000). Although a point count is conducted using both sight and sound, the vast majority of species vocalize to some degree, and additionally the recorded audio can be replayed at leisure to ensure accuracy.

"Smart sampling" aims to improve upon this by selecting the best samples from a long recording). Wimmer, et al. (2013) compared traditional point-counts with different methods of selecting the best samples from 480 hours of recording based only on the time of day. They found that randomly selecting 120 minutes during the "dawn chorus" - the three hours immediately after civil dawn – found the most species, finding 62% of total species compared to the 34% found using traditional point counts with the same human effort.

Towsey, Wimmer, Williamson, and Roe (2014)) used acoustic indices as a measure of informativeness for smart sampling. The indices, such as Acoustic Complexity Index (Pieretti, Farina, & Morri, 2011) and spectral entropy index (Sueur, Pavoine, Hamerlynck, & Duvail, 2008), are calculated over audio of arbitrary length, resulting in either a single index or index per frequency bin. The acoustic indices show some relationship to the number of species vocalizing in the audio over which they are calculated. The results using this relationship for smart sampling were promising: 67% of the total number of species present could be found in the 60 top ranked one-minute samples However a limitation of using indices calculated over entire minutes of audio is that they offer no information about individual events.

There has also been much research into fully automated species identification, however applying classifiers trained on specific species to richness survey is not yet feasible, due to the need to train and tune parameters on the large variety of potential species, and this is further confounded by the variety of local dialects within species. Our focus on representing the acoustic variety in a more generalized way avoids these problems.

3 Acoustic Event Detection

Our proposed techniques make use of acoustic event detection. An acoustic event is a short duration rise in amplitude above the background noise level, and is delimited by the start time, top and bottom frequency bounds and duration. Segmenting these events is quite challenging due to the diverse nature bird vocalizations, the presence of noise such as wind and insects, and the difficulty in distinguishing and, even defining, when one event stops and another begins. The algorithm employed for this task follows these steps (Towsey & Planitz, 2011):

1. A spectrogram is generated through a Short Time Fourier Transform (STFT) using hamming window with a frame width of 512 samples and overlap of 50%, and converted to a decibel scale.
2. Noise removal is performed using a wiener filter with a 5x5 neighborhood.
3. The spectrogram is converted to binary, whereby all values less than 4 decibels above the background noise level take the value 0 and all values above this threshold take the value 1.
4. Each connected area is treated as an event, and its frequency and time bounds recorded
5. Very large events (greater than 3000 pixels in area) are split into smaller events
6. Very small events are discarded (less than 150 pixels in area), as they are assumed to be noise.

The thresholds set for steps 3, 5 and 6 were determined empirically by applying various combinations of thresholds to a small number of randomly selected minutes. These were manually checked to determine which had the subjectively best delimitation of events, specifically:

- minimal number of missed vocalizations,
- minimal merging of vocalizations from different sources into single events, and
- minimal merging of repeated vocalizations from a single source into one event.

The number of acoustic events detected was close to 700 000 in all 168 hours analyzed.

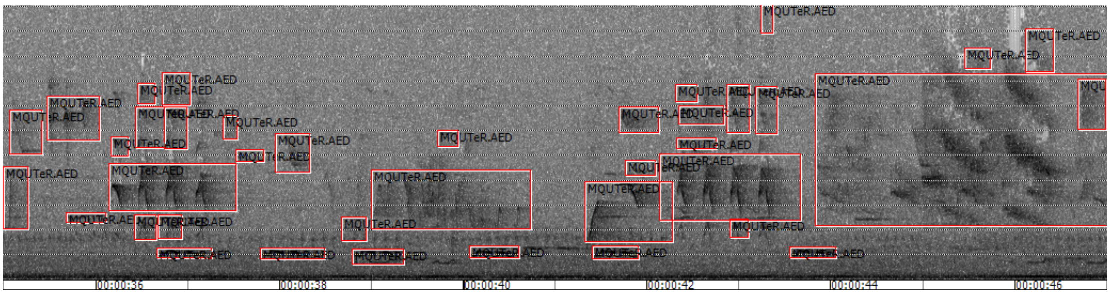

Figure 1. Acoustic event detection results on a spectrogram (frequency range is 11KHz, time range is 12s).

4 General Framework

In our proposed framework, the one-minute samples are selected incrementally in order to be able to account for diversity in the ranking process.

Given the set *M* of all minutes available in the survey, and a set *S* of the *n* minutes selected so far in the ranking, the selection of a new minute s_{n+1} is as follows:

$$s_{n+1} = Arg \max_{m_i \in M \setminus S} [Info\ (M_i) \times \min_{s_j \in S} [Dist(m_i, s_j)]]$$

where *Info(x)* is a function providing the informativeness of a minute *x* and *Dist(x, y)* is a function of the difference between two minutes *x* and *y*.

In this paper, the term *one-minute sample* is often abbreviated to either *minute* or *sample*.

4.1 Temporal dispersal of samples

Species often tend to vocalize for more than one minute at a time, and as such it follows that avoiding selecting samples in close proximity in time to those already ranked will increase diversity and thus the chance of observing new species.

In this context, the *Dist* function is a transformation from the time-difference in minutes between two samples. The function we propose here is a quadratic function of the distance between the minutes, which takes two parameters, the threshold and the amount. The threshold can be interpreted as the maximum difference in time between two samples to be considered as having potential similarity of content, while the amount *a* scales the function. The *Dist* function is then formalized as follows.

$$Dist(m_i, m_j) = \begin{cases} \frac{a}{t}\sqrt{t^2 - (|i-j| - t)^2} + (1-a), & if\ |i-j| > t \\ 1, & otherwise \end{cases}$$

where m_i and m_j are two samples where $|i-j|$ is the difference in time between the two samples, t is the threshold and a is the amount in the range [0,1]

We empirically determined and appropriate threshold t at 30 minutes and the amount a at 1. The resulting shape of the *Dist* function is as shown in **Figure 2**.

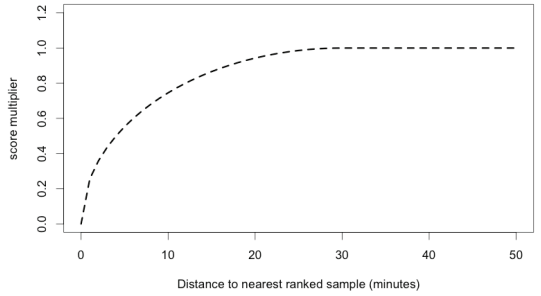

Figure 2. Temporal Dispersal score transformation function

4.2 Clustering Acoustic Events

We propose that clustering acoustic events can provide information about the nature of the events thus better estimating the richness of information of a given minute beyond the amount of information it contains.

Features for clustering

The clustering algorithm uses the following features of acoustic events:

- Duration
- Frequency range, defined as the difference between the maximum and minimum frequency of the event
- Mean peak frequency

Mean peak frequency is described as follows. The spectrogram represents time on the horizontal axis, with a resolution of around 23 milliseconds per time- frame, and frequency on the vertical axis, with a resolution of approximately 43 Hz per frequency bin (with a total of 256 frequency bins). Considering the frequency bins that fall within the frequency bounds of the event, the bin with the highest value in each time-frame is the peak frequency for that time-frame. The mean peak frequency is the average frequency of these peaks across the time frames of the event, weighted by their decibel values. Mean peak frequency gives a truer representation of the main frequency of the event than the midpoint between the frequency bounds, although in practice there may not be one main frequency.

In pilot experiments we found that additional features describing the contents of the event are not helpful (e.g. change in frequency over the course of the event, amplitude oscillation). We believe that this is due to the nature of the acoustic event detection: the delimitation of events is quite susceptible to errors caused by overlapping or nearby vocalizations being merged into single large events.

Clustering algorithm

Features are first scaled to their z-scores (i.e. the number standard deviations above the mean). K-means clustering was performed using the stats package of the R language (R-Core-Team, 2013), with each day-long recording treated separately.

We first experimented with various values for k (number of clusters) on a single day of audio and determined 240 clusters to be appropriate, although we found that the results are not significantly affected by this choice. We also experimented with hierarchical agglomerative clustering and found the result to be similar to k-means. K-means was chosen due to its lower complexity.

Cluster-based information function

We propose that the informativeness of a minute is determined by the sum of the weights of the clusters it contains, these weights being initially set to 1 and then decaying by a factor δ each time a minute containing the cluster is selected in the ranking. More formally, we can write the cluster-based information function as follows:

$$Info(m_i) = \sum_{C_k \in m_i} \delta^{|\{m \in S | C_k \in m\}|}$$

Because the set of minutes already selected in the ranking $m_j \in S$ changes each time a minute is selected, the ranking process is done iteratively, calculating $Info(m_i)$ for each unranked minute, selecting the minute with the highest value, then updating $m_j \in S$ and repeating. The algorithm allows clusters that appear in previously ranked minutes to be weighted less than clusters that do not yet appear in any previously ranked minutes.

Setting the decay value δ to 1 means that there is no preference towards choosing minutes that contain as yet unseen clusters. This results in minutes being ranked solely on the total number of clusters they contain. Setting δ less than 1 will cause the ranking process to prefer minutes with as yet unseen clusters, which will cause it to arrive more quickly at a point where all clusters are represented in the ranked minutes. The more often a cluster has appeared in higher ranked minutes, the lower contribution to $Info(m_i)$.

Setting the decay value δ very low, for example 0.1, ensures that samples are ranked only on the number of as yet unseen clusters, with almost no weight given to clusters already contained in higher ranked minutes. This small weight will add very little to the total score for the minutes containing that cluster, compared to the influence of a cluster that has not been included in any ranked minute yet. However, it is still non-zero, ensuring that once all clusters have been included, samples can still be selected intelligently based on the number of clusters they contain and the number of times those clusters have already been included.

5 Evaluation

5.1 Explored methods

In this paper we have presented 2 approaches to estimate the *Info* function as well as one temporal method to estimate the *Dist* function in addition to a uniform at 1. We then pair up these into four combinations detailed in Table 1, which follow the exploratory methodology presented in Figure 3:

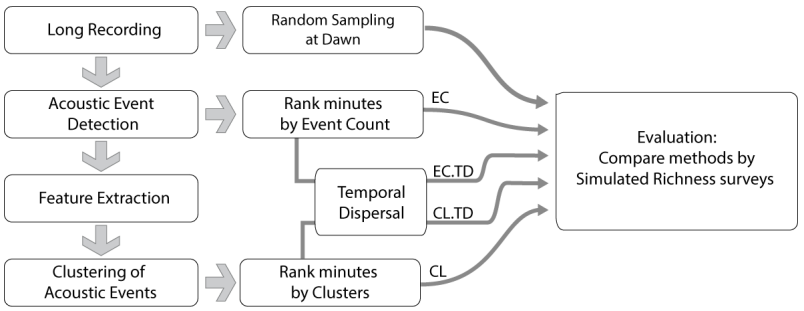

Figure 3. Experiment Design

	Info	*Dist*
EC	Number of events	Always 1
EC.TD	Number of events	Temporal dispersal
CL	Function of number of clusters seen and unseen	Always 1
CL.TD	Function of number of clusters seen and unseen	Temporal dispersal

Table 1 Combinations of Dist and Info functions for minute ranking

5.2 Data Sets

Seven 24-hour recordings (1440 minutes each) are used for the evaluation, recorded in a non-urban area of open woodland in Australia. Each minute of this audio has been labeled with the all species appearing in that minute. The labels were provided by bird-watchers with local knowledge, and serve as the ground truth. The recordings are listed in Table 2.

Recording	Site	Date	Number of species
1	NE	Oct 13	64
2	NE	Oct 14	57
3	NE	Oct 17	60
4	NW	Oct 13	60
5	NW	Oct 14	54
6	SE	Oct 13	61
7	SE	Oct 17	61
Total			98

Table 2: List of Recordings.

Using this labeled audio, a species richness survey can be simulated by sequentially checking the database for the species present in each ranked minute and adding new species to the running total. This can be plotted as a *species accumulation curve*. The species accumulation curve is the number of species found after x minutes, with x from 1 to 1440.

5.3 Baselines

Random sampling throughout the day

Random selection of minutes throughout the day is repeated 100 times to obtain a mean and standard deviation for each minute. In the plot, the standard deviation is shown as a shaded area surrounding the species accumulation curve.

Random sampling at dawn

By first selecting samples from three hours after sunrise, the performance of random sampling can be significantly improved, as the "dawn chorus" is the most active time for bird vocalizations. For the recordings in this study, this period was 5:15am to 8:15am. This comparison is useful because random sampling from dawn has proven to be an effective and simple method of sampling (Wimmer, et al., 2013). Again, the species accumulation curve is shown as the mean of 100 runs.

Optimal sampling

Using the labels in our test data, a greedy algorithm repeatedly searches the database of labeled minutes for the minute containing the most species that have not previously been found. This is repeated until all species known to be present have been found. Note that although this is not guaranteed to find the absolute optimal solution for including all the species present in the fewest minutes, it serves the purpose of a comparison to what is theoretically possible.

6 Results

The plots in Figure 4 show species accumulation curves for two out of the seven 24-hour recordings (due to space constraints, not all seven recordings could be included) using the approaches described in section 5.1, which are: *event count only* (EC), *event count with temporal dispersal* (EC.TD), *clustering* (CL) and *clustering with temporal dispersal* (CL.TD).

Also plotted are species accumulation curves achieved using the three baselines for comparison presented in section 5.3. Only the first 120 minutes are shown, as this period is the most relevant in demonstrating increased efficiency of a species richness survey.

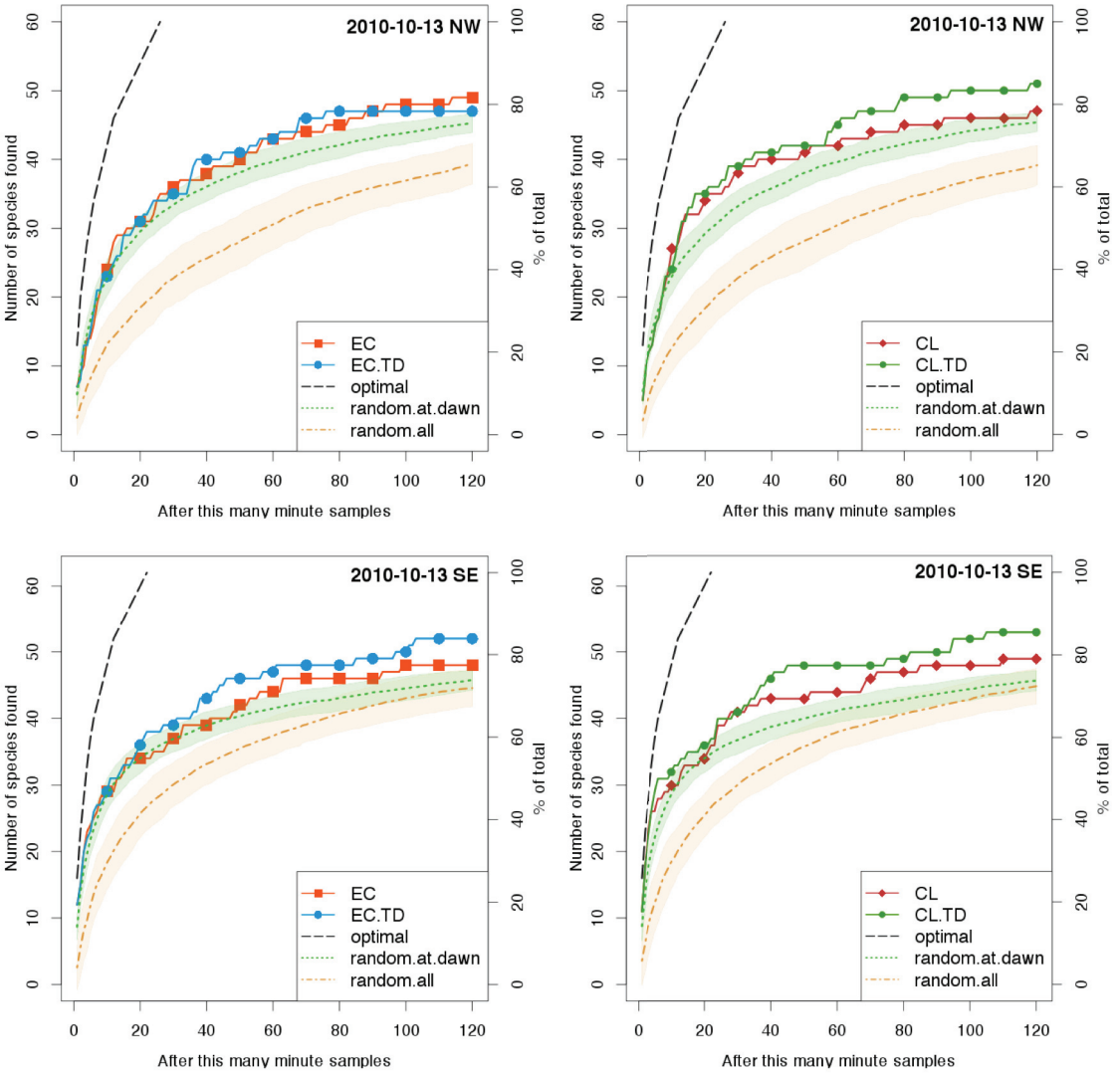

Figure 4 Species Accumulation Curves for two out of the seven days. On the left are results of using event count only (EC) and event count with temporal dispersal (EC.TD). On the right are results of using clustering (CL) and clustering with temporal dispersal (CL.TD).

The species accumulation curves reported for CL and CL.TD are those with the decay factor yielding the best results, which was δ=1, although the other values tested also performed above random sampling at dawn.

Species accumulation curves tend to rise steeply at first when there are many unobserved species remaining, then level off as is becomes harder to select a minute containing a new species. The difference between our methods and the baselines are of interest, not the absolute numbers. Therefore, the performance of each of these methods as a percent improvement over random sampling at dawn (R.D) was calculated by dividing the number of species in each ranked sample by the corresponding

value for R.D, then subtracting 1 and multiplying by 100. This was done for each of the seven recordings and averaged. Figure 5 shows that on average the performance of using *event count* for the *Info* function with temporal dispersal as the *Dist* function (EC.TD) starts to out-perform the baseline after 40 minutes. Further improvements were gained by sampling using cluster content as the *Info* function (Figure 6). The best performance was achieved by combining this with the temporal dispersal (CL.TD), achieving a 9% improvement over the random sampling at dawn by sample 60. This is also an improvement over the results reported by Towsey, et al. (2014)).

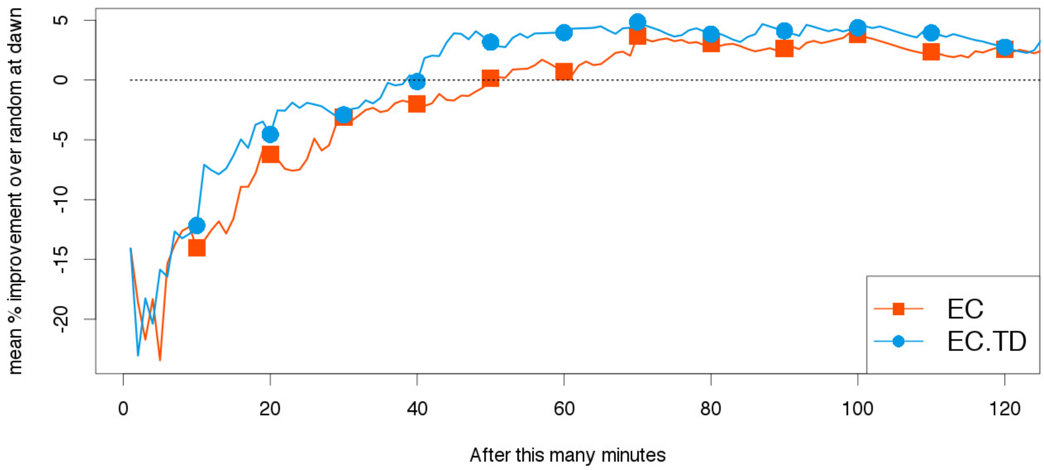

Figure 5 Average % improvement over random sampling at dawn across the seven 24-hour recordings evaluated for event count only (EC) and event count with temporal dispersal (EC.TD).

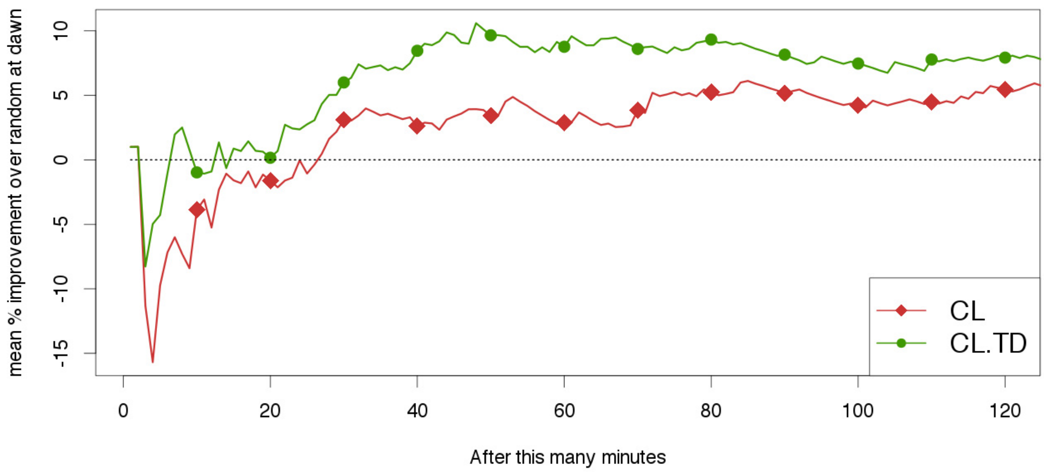

Figure 6 Average % improvement over random sampling at dawn across the seven 24-hour recordings evaluated for clustering (CL) and clustering with temporal dispersal (CL.TD).

7 Conclusion and future work

The results indicate that the estimates of informativeness of samples by the proposed iterative computational approaches have been successful in improving sample selection over the current state of the art. Smart sampling by acoustic event count and by cluster analysis of acoustic events are both effective methods of smart sampling. Including temporal dispersal further improved the results.

Ranking by clusters was able to perform as well as the state of the art, despite applying it with simple features and a simple decision algorithm. This is very promising because it can be improved in many ways, which will be the focus for future work. First of all, the acoustic event detection algorithm can be improved. A better detection of overlapping events would for example support the extraction of more discriminating features. Additional future work experimenting with different clustering algorithms may also provide better results.

8 Acknowledgements

The acoustic event detection algorithm used for this work was developed by Birgit Planitz, Michael Towsey and Anthony Truskinger

References

Gregory, R. D., Gibbons, D. W., & Donald, P. F. (2004). Bird census and survey techniques. In W. J. Sutherland, I. Newton & R. Green (Eds.), *Bird ecology and conservation* (pp. 17-56). Oxford: Oxford University Press.

Haselmayer, J., & Quinn, J. S. (2000). A comparison of point counts and sound recording as bird survey methods in Amazonian southeast Peru. *The Condor, 102*(4), 887-893.

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*(3), 868-873. doi:http://dx.doi.org/10.1016/j.ecolind.2010.11.005

R-Core-Team. (2013). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Sueur, J., Pavoine, S., Hamerlynck, O., & Duvail, S. (2008). Rapid Acoustic Survey for Biodiversity Appraisal. *PLoS ONE, 3*(12), e4065. doi:10.1371/journal.pone.0004065

Towsey, M., & Planitz, B. (2011). *Technical Report: acoustic analysis of the natural environment*. Technical Report. Retrieved from http://eprints.qut.edu.au/41131/4/41131.pdf

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*(0), 110-119. doi:http://dx.doi.org/10.1016/j.ecoinf.2013.11.007

Wimmer, J., Towsey, M., Roe, P., & Williamson, I. (2013). Sampling environmental acoustic recordings to determine bird species richness. *Ecological Applications, 23*(6), 1419-1428.