EXPLORING ECO-ACOUSTIC DATA WITH K-DETERMINANTAL POINT PROCESSES

Mohamed Outidrarine[1]*,Pierre Baudet*[1]*,Vincent Lostanlen*[1*]*,*
Mathieu Lagrange[1]*, Juan Sebastián Ulloa*[2]

[1] Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
[2] Instituto de Investigación de Recursos Biológicos Alexander von Humboldt, Bogotá, Colombia
vincent.lostanlen@ls2n.fr

ABSTRACT

The deployment of acoustic sensor networks in a natural environment contributes to the understanding and the conservation of biodiversity. Yet, the sheer size of audio data which result from these recordings prevents listening them in full. In order to skim through an eco-acoustic corpus, one may typically draw K snippets uniformly at random. In this article, we present an alternative method, based on K-determinantal point processes (K-DPP). This method weights the sampling of K-tuples according to a two-fold criterion of relevance and diversity. To study the eco-acoustics of a tropical dry forest in Colombia, we define relevance in terms of time–frequency second derivative (TFSD) and diversity in terms of scattering transform. Hence, we show that K-DPP offers a better tradeoff than K-means clustering. Furthermore, we estimate the species richness of the K selected snippets by means of the BirdNET birdsong classifier, which is based on a deep neural network. We find that, for $K > 10$, K-DPP and K-means tend to produce a species checklist that is richer than sampling K snippets independently without replacement.

Index Terms— Acoustic sensor networks, Bird species classification, Scattering transform, Time-frequency second derivative

1. INTRODUCTION

The science of eco-acoustics aims to analyze outdoor acoustic scenes so as to characterize certain ecological processes, such as population dynamics, species assemblages, and the emergence of a sonic landscape or "soundscape" [1]. This young field of research contributes to the overarching goal of biodiversity conservation at the planetary scale. Yet, a supervised classifier of eco-acoustic events may only be deployed once defined a research hypothesis, a taxonomy of sounds of interest, and a training set with expert or crowdsourced annotation.

Hence, the mass collection of eco-acoustic signals raises a problem in exploratory data analysis: how to quickly skim through a given corpus of N audio snippets without listening to it in full? In this context, the naive approach consists in drawing a subcorpus $\mathcal{X}$ of $K \ll N$ equiprobable samples. One may refine this method by weighting the probability of drawing each signal $\boldsymbol{x}_i$ in the corpus by a "relevance" prior q_i. Intuitively, the audio feature q_i is designed so as to quantify the saliency of spectrotemporal modulations in the time–frequency domain, such as animal vocalizations. However, this reweighted approach incurs a form of selection bias in terms of species richness: the top-K values of the saliency measure q_i typically belong to much fewer than K distinct species. This is because few species will exhibit calls with high saliency while most species will only exhibit calls with low-to-moderate saliency. The former tend to be over-represented in $\mathcal{X}$, at the detriment of the latter.

In this article, we propose a method for sampling audio signals according to a probabilistic tradeoff between relevance and diversity. The key idea is to represent each $\boldsymbol{x}_i$ by a vector $\boldsymbol{\Phi}_i$ of norm $\sqrt{q_i}$ and such that the angles $\angle(\boldsymbol{\Phi}_i, \boldsymbol{\Phi}_j)$ approximate the auditory dissimilarity of the pair $(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Thus, we draw the subcorpus $\mathcal{X}$ with a probability which is proportional to the determinant of the indexed family $\boldsymbol{\Phi}_i$ of vectors $\boldsymbol{x}_i \in \mathcal{X}$. Because of this proportionality, the proposed method is known as *K-determinantal point process*, or K-DPP for short.

There is a growing body of literature on the topic of applying (K-)DPP to various machine learning problems, such as image search and graph threading in a document collection: we refer to [2] for an introduction. However, no application of K-DPP to machine listening has been published until today. Hence, the novel contribution of our paper is to offer a first proof of concept which demonstrates its interest for diverse sampling in eco-acoustics. A recent article [3] has employed a K-DPP in order to extract a subcorpus of urban acoustic scenes as part of a human annotation campaign; but the authors used a constant relevance term $q_i = 1$ and did not evaluate their approach.

As a first illustration, Figure 1 illustrates different sampling methods with $K = 3$ for eco-acousticc signals from a dry tropical forest. Qualitatively speaking, we notice that uniform independent draws (row 1) tend to lack in relevance, with some samples containing only high-frequency insect stridulations. Relevance weighting (row 2) mitigates this problem at the cost of a reduction in diversity: the most salient species tend to recur disproportionately. Clustering with K-means (row 3) restores diversity at the detriment of relevance. Lastly, K-DPP (row 4) seems to offer an interesting tradeoff between relevance and diversity.

2. DETERMINANTAL POINT PROCESS

2.1. Time–frequency second derivative (TFSD)

We decompose each signal $\boldsymbol{x}_i$, of same duration T, by means of a constant-Q wavelet filters $\boldsymbol{\psi}_{\lambda_1}$, with $\lambda_1 \in \Lambda$. We apply pointwise complex modulus and denote by $\mathbf{U}_1\boldsymbol{x}_i(t, \lambda_1) = \left|\boldsymbol{x}_i * \boldsymbol{\psi}_{\lambda_1}\right|(t)$ the resulting scalogram. representation. Then, we compute the first-order difference of $\mathbf{U}_1\boldsymbol{x}_i$ over both variables t and λ_1. Hence:

$$\begin{aligned}\mathrm{TFSD}(\boldsymbol{x}_i)(t, \lambda_1) = \mathbf{U}_1\boldsymbol{x}_i(t + \tau, \lambda_1 + \delta) + \mathbf{U}_1\boldsymbol{x}_i(t, \lambda_1) \\ - \mathbf{U}_1\boldsymbol{x}_i(t + \tau, \lambda_1) - \mathbf{U}_1\boldsymbol{x}_i(t, \lambda_1 + \delta), \quad (1)\end{aligned}$$

*This work is supported by an Atlanstic 2020 award on "Trainable Acoustic Sensors" (TrAcS).

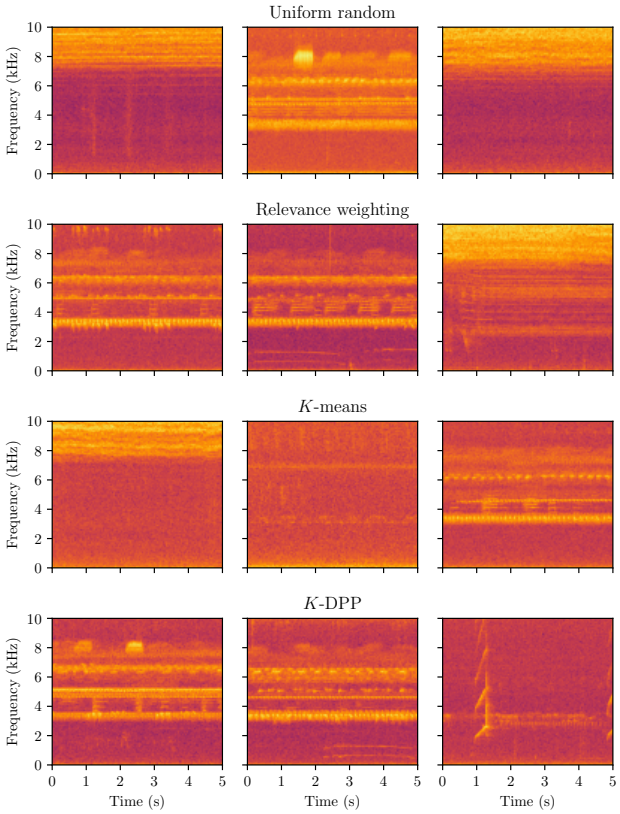

Figure 1: Short-term Fourier spectrograms of three samples from the dataset under study, drawn from four different random sampling methods: uniform, relevance-based, K-means, and K-DPP.

where the acronym TFSD stands for Time–Frequency Second Derivative. In practice, we set the hop size τ to 23 ms, the quality factor to $Q = 3$, the relative frequency interval δ to one third of an octave. We define a region of interest $\Lambda' \subset \Lambda$, corresponding to the third-octave bands λ_1 in $\mathbf{U}\boldsymbol{x}_i$ which range between 2 and 8 kHz. This region corresponds to the vocal range of most singing species of birds. Lastly, we sum the absolute values of the $\mathrm{TFSD}(\boldsymbol{x}_i)$ matrix over both regions Λ and Λ'. Their ratio provides a measure of relevance

$$q_i = \frac{\int_0^T \int_{\Lambda'} \left|\mathrm{TFSD}(\boldsymbol{x}_i)(t,\lambda_1)\right| \,\mathrm{d}t\,\mathrm{d}\lambda_1}{\int_0^T \int_{\Lambda} \left|\mathrm{TFSD}(\boldsymbol{x}_i)(t,\lambda_1)\right| \,\mathrm{d}t\,\mathrm{d}\lambda_1} \tag{2}$$

between zero and one. A recent study in an urban environment [4] has shown that the descriptor q_i (defined with a slight variation in spectrotemporal parameters) significantly correlates with the perceived time of presence of bird vocalizations.

2.2. Time scattering transform

The time scattering transform, also known as deep scattering spectrum [5], is a nonlinear convolutional operator in the time–frequency domain whose role is to extract amplitude modulations at various time scales in each wavelet subband. First, we integrate the scalogram $\mathbf{U}_1\boldsymbol{x}_i$ along time t, which gives the averaged scalogram:

$$\mathbf{S}_1\boldsymbol{x}_i(\lambda_1) = \int_0^T \mathbf{U}_1\boldsymbol{x}_i(t,\lambda_1)\,\mathrm{d}t, \tag{3}$$

also known as first-order scattering. The transformation from $\mathbf{U}_1$ to $\mathbf{S}_1$ guarantees a property of global invariance, which comes at the cost of a loss in discriminability: $\mathbf{S}_1\boldsymbol{x}_i(\lambda_1)$ ignores the amplitude modulations in each wavelet subband $\mathbf{U}_1\boldsymbol{x}_i(t,\lambda_1)$ around its average value, and so for every fixed λ_1. The key idea behind the scattering transform is to recover these amplitude modulations by means of a second filter bank of constant-Q wavelet filters $\boldsymbol{\psi}_{\lambda_2}$, except with a quality factor of $Q = 1$ now. We obtain the so-called amplitude modulation spectrum of $\boldsymbol{x}_i$:

$$\mathbf{U}_2\boldsymbol{x}_i(t,\lambda_1,\lambda_2) = \left|\mathbf{U}_1\boldsymbol{x}_1 * \boldsymbol{\psi}_{\lambda_2}\right|(t,\lambda_1), \tag{4}$$

where the convolution between scalogram and second-order wavelet is performed over time t, and broadcasted across all frequencies λ_1. Symmetrically to $\mathbf{U}_1$ et $\mathbf{S}_1$, we integrate $\mathbf{U}_2$ along time to obtain the matrix:

$$\mathbf{S}_2\boldsymbol{x}_i(\lambda_1,\lambda_2) = \int_0^T \mathbf{U}_2\boldsymbol{x}_i(t,\lambda_1,\lambda_2)\,\mathrm{d}t. \tag{5}$$

We concatenate the first-order coefficients $\mathbf{S}_1\boldsymbol{x}_i$ with the flattened matrix $\mathbf{S}_2\boldsymbol{x}_i$ so as to obtain a high-dimensional vector $\mathbf{S}\boldsymbol{x}_i$, which is generically indexed by the "scattering path" multiindex p, i.e., either the singleton λ_1 or to the pair (λ_1,λ_2) depending on order.

Time scattering approximately verifies a property of energy conservation, similarly to the Parseval identity for the Fourier transform. Therefore, dividing the vector $\mathbf{S}\boldsymbol{x}_i$ by its ℓ^2 norm is tantamount to normalizing the waveform $\boldsymbol{x}_i$ itself.

2.3. Likelihood kernel

We define $\phi_i = \mathbf{S}\boldsymbol{x}_i/\|\mathbf{S}\boldsymbol{x}_i\|_2$ the renormalized vector, and $\boldsymbol{\Phi}_i = \sqrt{q_i}\phi_i$ the vector that is parallel to ϕ_i and has ℓ^2 norm equal to $\sqrt{q_i}$. Our working hypothesis is that constructing a K-DPP with time scattering as the descriptor of choice will preserve the auditory diversity across eco-acoustic samples. We repeat the same operation for every signal $\boldsymbol{x}_i \in \mathcal{X}$.

The likelihood kernel of the associated K-DPP is defined as

$$\mathbf{L}_{i,j} = \langle \boldsymbol{\Phi}_i | \boldsymbol{\Phi}_j \rangle = \sqrt{q_i q_j}\,\langle \phi_i | \phi_j \rangle. \tag{6}$$

Given a set of distinct indices $\boldsymbol{\sigma} = \{\sigma_1 \dots \sigma_K\}$ between 1 and N, we denote by $\mathbf{L}_{\boldsymbol{\sigma}}$ the restriction of the matrix $\mathbf{L}$ to the rows and columns whose indices belong to $\boldsymbol{\sigma}$. Thus, the K-DPP with kernel $\mathbf{L}$ is a random variable over the K-uplets of $1 \dots N$, in which the probability of drawing a specific K-uplet $\boldsymbol{\sigma}$ is proportional to the determinant of the matrix $\mathbf{L}_{\boldsymbol{\sigma}}$:

$$\mathbb{P}[\mathcal{X} = (\boldsymbol{x}_{\sigma_1} \dots \boldsymbol{x}_{\sigma_K})] \propto \det \mathbf{L}_{\boldsymbol{\sigma}}. \tag{7}$$

3. APPLICATION TO ECO-ACOUSTICS

3.1. Protocol

We apply our protocol to a dataset of $N = 432$ audio snippets, which we recorded between February 14th and February 16th in the dry tropical forest of San Jacinto (Bolívar, Colombia) by means

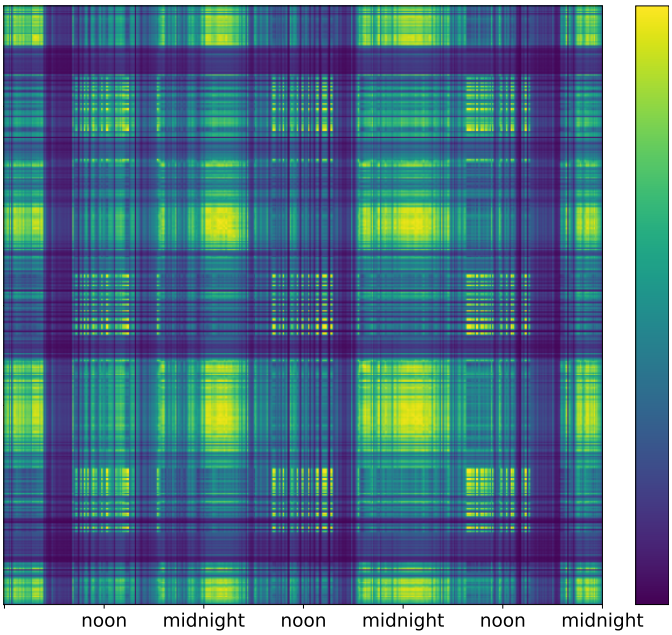

Figure 2: Likelihood kernel $\mathbf{L}$ for the audio snippets under study, arranged in a chronological order over three days. Darker colors denote a greater joint probability of sampling the row snippet and the column snippet as part of the same DPP draw.

of a Wildlife Acoustics "Song Meter 2" acoustic sensor. This sensor is equipped with an omnidirectional microphone and records intermittently at a rate of one five-second snippet every ten minutes, twenty-four hours a day. Such data acquisition campaign belongs to a larger endeavor for biodiversity coordination, which is being coordinated by the Alexander von Humboldt Biological Resources Research Institute [6]. For our experiment, we rely on a Python implementation of K-DPP via an open-source library named DPPy [7] by using the aforementioned criteria of relevance and diversity. We extract the TFSD via the scikit-maad package [8] and the time scattering transform via the Kymatio package [9].

Figure 2 presents the likelihood kernel $\mathbf{L}$ as computed over this eco-acoustic dataset. Note that the diagonal of the matrix $\mathbf{L}$ gives the relevance of observations: for every i, $\mathbf{L}_{i,i} = q_i$. We find that the relevance term, as described by the TFSD eco-acoustic indicator, roughly follows a daily periodicity: it is highest at dawn and dusk, lowest at midnight, and takes irregular values around noon. Future research is necessary to indicate whether this temporal pattern aligns with the chronobiology of vocalizing animals in the site under study.

In Figure 2, we also observe that $\mathbf{L}$ has a block-wise structure, which again, aligns with the daily cycle of animal vocalizations in the forest. Relatedly, we find that the soundscape under study seems to exhibit greater acoustical diversity, as measured by the scattering transform, during the night than during the day.

Mathematically speaking: for snippets $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ which are acoustically similar (e.g., because they are one day apart), their ℓ^2-normalized scattering transforms ϕ_i and ϕ_j are almost colinear. As a result, the inner product $\langle \phi_i | \phi_j \rangle$ is close to one, and the determinant associated to $\sigma = (i, j)$ is:

$$\begin{aligned} \det \mathbf{L}_\sigma &= q_i q_j - \sqrt{q_i q_j} \langle \phi_i | \phi_j \rangle \sqrt{q_j q_i} \langle \phi_j | \phi_i \rangle \\ &= q_i q_j \left(1 - \langle \phi_i | \phi_j \rangle^2\right), \end{aligned} \tag{8}$$

which is nonnegative but close to zero. The subtractive term $-\langle \phi_i | \phi_j \rangle^2$ has a repulsive effect over the pair of snippets $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, of the order of the cosine similarity between features ϕ_i and ϕ_j.

We compare our K-DPP to a naive baseline, which we name "uniform random". The baseline verifies, for each snippet i, the proportionality rule: $\mathbb{P}[\boldsymbol{x}_i \in \mathcal{X}] \propto 1/N$, where N is the total number of audio snippets. In addition, we refine the naive baseline so that the probability of drawing the snippet $\boldsymbol{x}_i$ is proportional to its relevance q_i: $\mathbb{P}[\boldsymbol{x}_i \in \mathcal{X}] \propto q_i$. Under both "uniform random" and "relevance weighting" methods, the probabilistic sampling is made K times independent without replacement. Thirdly, we evaluate a well-known method for unsupervised clustering: namely, K-means. This method produces a partition of the full corpus into K disjoint clusters $\mathcal{C}_1, \ldots, \mathcal{C}_K$ so as to minimize the intra-class variance:

$$\sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in \mathcal{C}_k} \left\| \mathbf{S}\boldsymbol{x}_i - \frac{1}{\operatorname{card} \mathcal{C}_k} \sum_{\boldsymbol{x}_j \in \mathcal{C}_k} \mathbf{S}\boldsymbol{x}_j \right\|^2. \tag{9}$$

We build the subcorpus $\mathcal{X}$ by selecting, for each cluster $\mathcal{C}_k$, the snippet $\boldsymbol{x}$ whose scattering transform $\mathbf{S}\boldsymbol{x}$ is closest to the Euclidean centroid of all points in $\mathcal{C}_k$. Unlike the first two naive baselines, K-means does not perform independent sampling without replacement: instead, the cluster assignment of every snippet affects the global cost function in Equation 9, and thus indirectly conditions the probability of sampling every other snippet. At the same time, we note that K-means clustering becomes impractical if repeated draws of the subcorpus $\mathcal{X}$ are needed, or if the number of elements K needs to be adjusted dynamically.

3.2. Relevance–diversity tradeoff

We run all four methods with $K = 3$ and repeat them 200 times independently. Figure 3 illustrates our findings. As expected, the uniform random method fares poorly on both metrics of relevance and diversity. Relevance weighting improves relevance but still lacks in diversity. Clustering with K-means is outperformed by relevance weighting on both metrics. Last but not least, we observe that the K-DPP reaches a favorable tradeoff between relevance and diversity: it offers a greater diversity than relevance-weighted sampling while guaranteeing a better relevance than K-means. More precisely, K-DPP triples the diversity of relevance-weighted sampling while retaining about 90% of its relevance on average.

3.3. Species inventory

In the previous subsection, we have verified that the K-DPP method yields diverse and relevant subcorpora, by our predefined measure of relevance (TFDS) and diversity (scattering transform). It remains to be seen whether these definitions are useful in practice for conservation science. For this purpose, we run every snippet $\boldsymbol{x}_i$ through BirdNET, a pretrained convolutional network for species identification [10]. In this way, our evaluation metric is the total number of distinct species in the K-uplet $\mathcal{X}$, also known as "species richness" in ecology. Moreover, we measure the "precision at K"; that is, the proportion of snippets in $\mathcal{X}$ which contain at least one bird vocalization, whatever be its species.

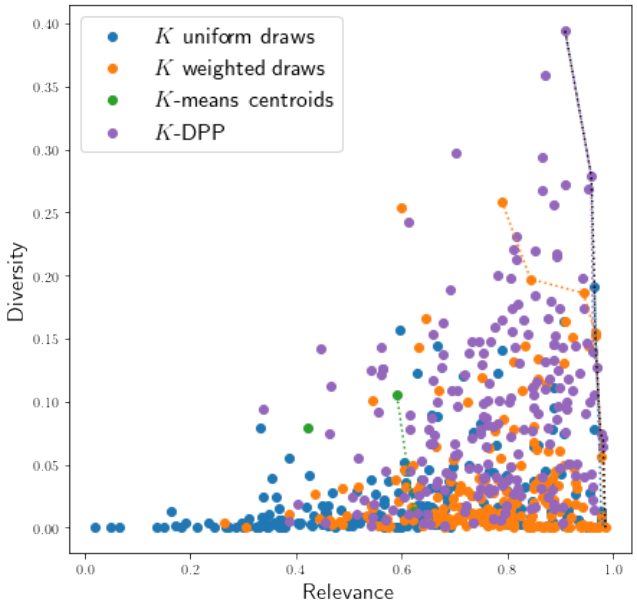

Figure 3: Scatter plot of relevance (x-axis) and diversity (y-axis) of K-uplets from our eco-acoustic dataset, as drawn from various sampling techniques (see legend). Dashed lines indicate the Pareto front for each method. The best K-uplets are in the top-right corner.

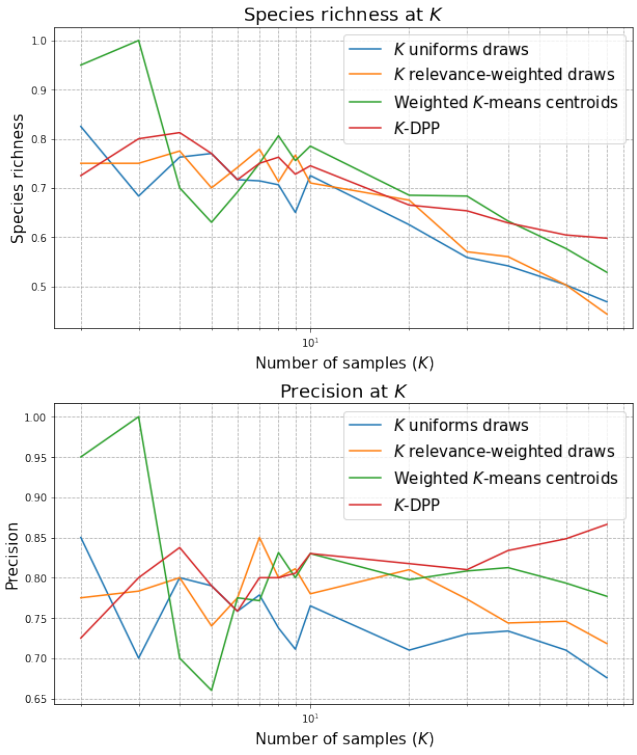

Figure 4: Species richness (top) and precision (bottom) of K-uplets drawn from all four presented methods, for a set cardinal ranging between $K = 3$ and $K = 80$. Higher is better.

Figure 4 illustrates our findings. We see that the naive baseline has a poor species richness and a poor precision. Meanwhile, sampling with K-DPP leads to better results: its K-uplets contain a richer species inventory and fewer false positives. Yet, a surprising result, deserving of further inquiry, is that K-means clustering matches K-DPP in richness, and even outperforms it for very small values of K.

4. CONCLUSION

We have shown that the use of K-determinantal point processes (K-DPP) in eco-acoustics allows to explore and summarize large volumes of audio data while satisfying an interesting tradeoff between relevance and diversity. By means of an off-the-shelf classifier of bird species (BirdNET), we have shown that K-DPP tend to enrich the species inventory of the subcorpus compared to random uniform sampling; and so, particularly for $K > 10$, when there a substantial risk of accidentally drawing near-duplicate snippets.

We note that the choice of a likelihood kernel plays a large role in the success of K-DPP. In our article, this choice was motivated by domain-specific knowledge in eco-acoustics and psycho-acoustics, and was later confirmed by means of a species classifier. Our paper calls attention on the risks underlying the random subsampling of a dataset, especially in the early phase of forming a research hypothesis. Relevance weighting reduces the risk of sampling false positives, yet at the cost of biasing the subcorpus $\mathcal{X}$ towards a narrow range of extremely salient events. Hence, we have advocated for a balanced approach, which takes both relevance and diversity into account. We have presented a first application of K-DPP for exploring biodiversity in a Colombian dry forest, with the hope to encourage more applications of this tool in the future.

5. ACKNOWLEDGMENT

We thank Guillaume Gautier and Rémi Bardenet for their help with the DPPy package.

6. REFERENCES

[1] J. Sueur and A. Farina, "Ecoacoustics: the ecological investigation and interpretation of environmental sound," *Biosemiotics*, vol. 8, no. 3, pp. 493–502, 2015.

[2] A. Kulesza, B. Taskar, *et al.*, "Determinantal point processes for machine learning," *Foundations and Trends in Machine Learning*, vol. 5, no. 2–3, pp. 123–286, 2012.

[3] M. Cartwright, J. Cramer, A. E. M. Mendez, Y. Wang, H.-H. Wu, V. Lostanlen, M. Fuentes, G. Dove, C. Mydlarz, J. Salamon, *et al.*, "SONYC-UST-V2: An urban sound tagging dataset with spatiotemporal context," in *Proc. DCASE*, 2020.

[4] F. Gontier, C. Lavandier, P. Aumond, M. Lagrange, and J.-F. Petiot, "Estimation of the perceived time of presence of sources in urban acoustic environments using deep learning techniques," *Acta Acustica united with Acustica*, vol. 105, no. 6, pp. 1053–1066, 2019.

[5] J. Andén and S. Mallat, "Deep scattering spectrum," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4114–4128, 2014.

[6] C. Pizano and H. García, Eds., *El bosque seco tropical en Colombia*. Bogotá, D.C., Colombia.: Instituto de Investigación de Recursos Biológicos Alexander von Humboldt (IAvH), 2014.

[7] G. Gautier, G. Polito, R. Bardenet, and M. Valko, "DPPy: DPP Sampling with Python," *Journal of Machine Learning Research - Machine Learning Open Source Software (JMLR-MLOSS)*, vol. 20, no. 180, pp. 1–7, 2019.

[8] J. S. Ulloa, S. Haupert, J. F. Latorre, T. Aubin, and J. Sueur, "scikit-maad: An open-source and modular toolbox for quantitative soundscape analysis in Python," *Methods in Ecology and Evolution*, vol. 12, no. 12, pp. 2334–2340, 2021.

[9] M. Andreux, T. Angles, G. Exarchakis, R. Leonarduzzi, G. Rochette, L. Thiry, J. Zarka, S. Mallat, J. andén, E. Belilovsky, J. Bruna, V. Lostanlen, M. Chaudhary, M. J. Hirn, E. Oyallon, S. Zhang, C. Cella, and M. Eickenberg, "title=Kymatio: Scattering transforms in Python," *Journal of Machine Learning Research*, vol. 21, no. 60, pp. 1–6, 2020.

[10] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, "Birdnet: A deep learning solution for avian diversity monitoring," *Ecological Informatics*, vol. 61, p. 101236, 2021.