Un processus ponctuel déterminantal fini pour l'exploration de données éco-acoustiques

Pierre BAUDET[1], Mohamed OUTIDRARINE[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, Colombie

vincent.lostanlen@ls2n.fr

Résumé – Le déploiement de réseaux de capteurs acoustiques en milieu naturel contribue à la compréhension et à la sauvegarde de la biodiversité. Pourtant, la masse de données audio recueillies par ces capteurs ne peut être écoutée dans sa totalité. Afin de se donner un rapide aperçu auditif du contenu d'un corpus éco-acoustique, il est donc courant d'en tirer K extraits uniformément au hasard. Dans cet article, nous présentons une méthode alternative fondée sur un K-processus ponctuel déterminantal (K-DPP). Cette méthode pondère l'échantillonnage des K-uplets d'après un double critère de pertinence et de diversité. Pour l'étude éco-acoustique d'une forêt tropicale sèche en Colombie, nous définissons la pertinence en termes de dérivée seconde spectrotemporelle (TFSD) et la diversité en termes de diffusion en ondelettes. Dès lors, nous montrons que K-DPP offre un meilleur compromis qu'un partitionnement par K-moyennes. De plus, nous estimons la richesse spécifique des K extraits sélectionnés à l'aide du classifieur de chants d'oiseaux BirdNET, fondé sur un réseau de neurones profond. Pour $K > 10$, K-DPP et K-moyennes tendent à produire un inventaire d'espèces plus riche que K extraits tirés indépendamment.

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. For $K > 10$, K-DPP and K-means tend to produce a species checklist that is richer than sampling K snippets independently without replacement.

1 Introduction

L'éco-acoustique vise à analyser les sons naturels afin de caractériser certains processus écologiques tels que la dynamique des populations animales, l'assemblage d'espèces en communauté ou l'émergence d'un "paysage sonore".Cette discipline émergente répond à l'enjeu contemporain de préservation de la biodiversité à l'échelle planétaire. Or, une classification éco-acoustique par apprentissage supervisé ne peut être déployée qu'après avoir défini une hypothèse de recherche, une taxonomie de sons d'intérêt et une base d'entraînement.

La collecte massive de signaux éco-acoustiques pose donc un problème d'analyse de données exploratoire. Comment se donner un rapide aperçu d'un corpus de N extraits sonores, sans l'écouter en intégralité ? La méthode naïve consiste à tirer un sous-corpus $\mathcal{X}$ de $K \ll N$ échantillons de façon équiprobable. On peut raffiner cette méthode en pondérant la probabilité de tirage d'un signal $\boldsymbol{x}_i$ par un *a priori* de "pertinence" q_i : celui-ci est construit pour être d'autant plus élevé que $\boldsymbol{x}_i$ contient des évènements saillants dans le plan temps–fréquence, tels que des vocalisations animales. Mais cette approche crée un biais de sélection en matière de biodiversité : la mesure de pertinence q_i est typiquement concentrée dans un petit nombre d'espèces produisant des cris très saillants. Ces espèces ont donc tendance à être sur-représentées dans $\mathcal{X}$ au détriment de toutes les autres.

Dans cet article, nous proposons une méthode d'échantillonnage de signaux audio selon un double critère probabiliste équilibrant pertinence et diversité. Cette méthode requiert, au préalable, de représenter chaque signal $\boldsymbol{x}_i$ par un vecteur $\boldsymbol{\Phi}_i$ de norme $\sqrt{q_i}$ et tel que les angles $\angle(\boldsymbol{\Phi}_i, \boldsymbol{\Phi}_j)$ approximent la dissimilarité acoustique perçue de la paire $(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Dès lors, nous tirons le sous-corpus $\mathcal{X}$ avec une probabilité proportionnelle au déterminant de la famille de vecteurs $\boldsymbol{\Phi}_i$ associés aux $\boldsymbol{x}_i \in \mathcal{X}$. On parle ainsi de K-processus ponctuel déterminantal ou K-DPP pour *determinantal point process*.

Les (K-)DPP ont déjà été appliqués à diverses tâches d'indexation telles que la recherche d'images ou la génération d'une "revue de presse" thématique [7]. En revanche, aucune application à l'éco-acoustique n'est répertoriée à ce jour ; notre article en propose une première preuve de principe. Un article récent

Ce travail est financé par la bourse Atlanstic 2020 TrAcS.

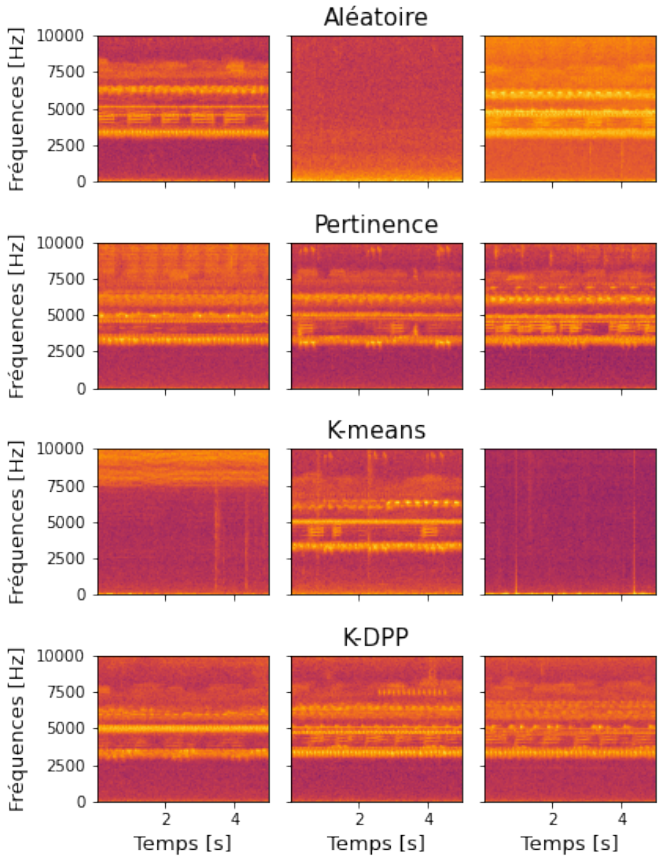

FIGURE 1 – Spectrogrammes de trois tirages donnés par chacune des quatre méthodes de sélection. On notre que K-DPP, en équilibrant pertinence et diversité, permet de proposer uniquement des chants d'oiseaux mais de contenu spectral diversifié.

[3] a utilisé un K-DPP pour extraire un sous-corpus de scènes sonores urbaines en vue d'une campagne d'annotation humaine ; mais cet usage informel n'a pas fait l'objet d'une évaluation. Pour notre cas d'étude, la figure 1 illustre différentes méthodes de tirage avec $K = 3$ pour des signaux $\boldsymbol{x}_i$ provenant d'une forêt tropicale sèche.

2 Processus ponctuel déterminantal

2.1 Critère de pertinence : *Time–Frequency Second Derivative* (TFSD)

On décompose chaque signal $\boldsymbol{x}_i$ au moyen d'un banc de filtres $\boldsymbol{\psi}_{\lambda_1}$ dont la largeur de bande vaut un tiers d'octave et l'on note $\mathbf{U}_1\boldsymbol{x}_i(t,\lambda_1) = \left|\boldsymbol{x}_i * \boldsymbol{\psi}_{\lambda_1}\right|(t)$ le scalogramme résultant. On calcule ensuite une discrétisation de la dérivée seconde de $\mathbf{U}_1\boldsymbol{x}_i$ selon le temps t et la (log-)fréquence $\lambda_1 \in \Lambda$:

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

En pratique, on choisit τ égal à 23 ms et δ un tiers d'octave. On définit une région d'intérêt Λ' correspondant aux bandes λ_1 de $\mathbf{U}\boldsymbol{x}_i$ comprises entre 2 et 8 kHz. Cette région correspond à l'ambitus vocal de la plupart des oiseaux chanteurs. Enfin, on somme les valeurs absolues de la matrices $\mathrm{TFSD}(\boldsymbol{x}_i)$ sur les régions Λ et Λ'. Leur rapport donne une valeur de pertinence

$$q_i = \frac{\iint_{\Lambda'} \left|\mathrm{TFSD}(\boldsymbol{x}_i)(t,\lambda_1)\right| \,\mathrm{d}t\,\mathrm{d}\lambda_1}{\iint_{\Lambda} \left|\mathrm{TFSD}(\boldsymbol{x}_i)(t,\lambda_1)\right| \,\mathrm{d}t\,\mathrm{d}\lambda_1} \quad (2)$$

comprise entre zéro et un. Une étude en milieu urbain [5] a montré que le descripteur q_i (muni de paramètres spectrotemporels légèrement différents) corrèle significativement avec le temps perçu de présence vocale d'oiseaux.

2.2 Descripteur : diffusion en ondelettes

La diffusion en ondelettes, implémentée par la bibliothèque Kymatio [2], permet de modéliser le signal sonore de manière pertinente pour les tâches de recherche. On intègre le scalogramme $\mathbf{U}_1\boldsymbol{x}_i$ selon le temps t, ce qui donne le scalogramme moyen ou *diffusion d'ordre 1* de $\boldsymbol{x}_i$:

$$\mathbf{S}_1\boldsymbol{x}_i(\lambda_1) = \int \mathbf{U}_1\boldsymbol{x}_i(t,\lambda_1)\,\mathrm{d}t. \quad (3)$$

Le passage de $\mathbf{U}_1$ à $\mathbf{S}_1$ permet de garantir une propriété d'invariance à la translation, au prix d'une perte de discriminabilité : $\mathbf{S}_1\boldsymbol{x}_i(\lambda_1)$ ignore les modulations d'amplitude de chaque canal fréquentiel $\mathbf{U}_1\boldsymbol{x}_i(t,\lambda_1)$ autour de sa valeur moyenne, à λ_1 fixé. L'idée-clé de la diffusion en ondelettes [1] consiste à recouvrer ces modulations d'amplitude au moyen d'un second banc de filtres passe-bande $\boldsymbol{\psi}_{\lambda_2}$ dont la largeur de bande vaut cette fois une octave. On obtient des coefficients de diffusion d'ordre 2 (en anglais *second-order scattering transform*) :

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

où la convolution est opérée sur la variable t. Symétriquement à $\mathbf{U}_1$ et $\mathbf{S}_1$, on intègre $\mathbf{U}_2$ en temps pour donner une matrice $\mathbf{S}_2\boldsymbol{x}_i(\lambda_1,\lambda_2) = \int \mathbf{U}_2\boldsymbol{x}_i(t,\lambda_1,\lambda_2)\,\mathrm{d}t$. On concatène les coefficients du premier ordre $\mathbf{S}_1\boldsymbol{x}_i$ avec ceux du second ordre $\mathbf{S}_2\boldsymbol{x}_i$ afin d'obtenir un vecteur en haute dimension $\mathbf{S}\boldsymbol{x}_i$ indexé génériquement par le "chemin de diffusion" *(scattering path)* p, valant le singleton λ_1 ou la paire (λ_1,λ_2) selon l'ordre considéré.

Nous posons comme hypothèse de travail que ce descripteur préserve la diversité des échantillons tirés lorsqu'il est associé aux propriétés de répulsion des DPP,

2.3 Noyau de vraisemblance

La diffusion en ondelettes vérifie approximativement une propriété de conservation de l'énergie, semblable à l'égalité de Parseval pour la transformée de Fourier. Par conséquent, diviser le vecteur $\mathbf{S}\boldsymbol{x}_i$ par sa norme ℓ^2 revient à normaliser le signal $\boldsymbol{x}_i$ lui-même. On définit $\phi_i = \mathbf{S}\boldsymbol{x}_i / \|\mathbf{S}\boldsymbol{x}_i\|_2$ le vecteur renormalisé, et $\boldsymbol{\Phi}_i = \sqrt{q_i}\phi_i$ le vecteur colinéaire à ϕ_i de norme $\sqrt{q_i}$. On répète l'opération pour tous les signaux $\boldsymbol{x}_i$ du corpus de départ. On calcule alors le noyau de vraisemblance du K-DPP [4]

$$\mathbf{L}_{i,j} = \langle \boldsymbol{\Phi}_i | \boldsymbol{\Phi}_j \rangle = \sqrt{q_i q_j}\,\langle \phi_i | \phi_j \rangle. \quad (5)$$

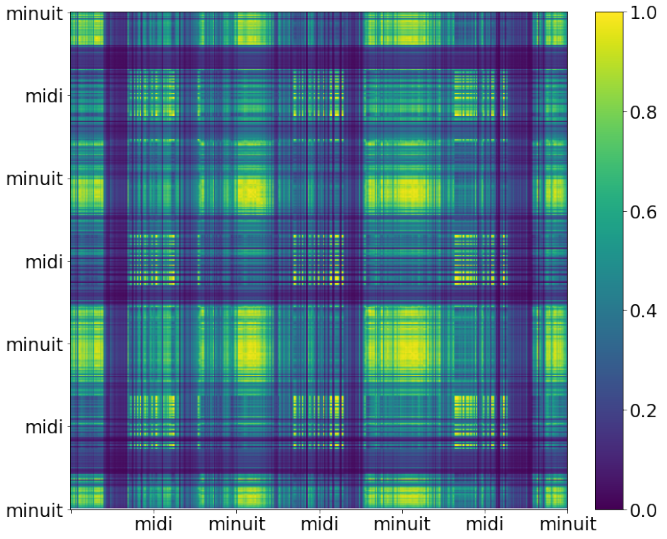

FIGURE 2 – Noyau de vraisemblance $\mathbf{L}$ pour les enregistrements étudiés.

Observons que la diagonale du noyau donne la pertinence des observations : $\mathbf{L}_{i,i} = q_i$. Étant donné un ensemble d'indices $\boldsymbol{\sigma} = \{\sigma_1 \dots \sigma_K\}$ distincts entre 1 et N, on note $\mathbf{L}_{\boldsymbol{\sigma}}$ la restriction de la matrice $\mathbf{L}$ aux lignes et colonnes indicées par $\boldsymbol{\sigma}$. On définit le K-DPP comme une variable aléatoire sur l'ensemble des K-uplets de $1 \dots N$ et dont la probabilité de tirage d'un K-uplet $\boldsymbol{\sigma}$ est proportionnelle au déterminant de la matrice $\mathbf{L}_{\boldsymbol{\sigma}}$:

$$\mathbb{P}[\mathcal{X} = (\boldsymbol{x}_{\sigma_1} \dots \boldsymbol{x}_{\sigma_K})] \propto \det \mathbf{L}_{\boldsymbol{\sigma}}. \tag{6}$$

3 Application à l'éco-acoustique

3.1 Protocole

Nous implémentons la méthode d'échantillonnage par K-DPP en utilisant la bibliothèque DPPy [4] et en utilisant les critères de diversité et pertinence décrits plus haut. Les données acoustiques proviennent d'une forêt tropicale sèche près de San Jacinto (Bolívar, Colombie) enregistrés avec un capteur Wildlife Acoustics "SongMeter 2" équipé d'un microphone omnidirectionnel qui enregistre par intermittence à raison de cinq secondes toutes les dix minutes. Notre étude comprend 432 enregistrements sur une période de 3 jours, du 14 au 16 février 2016, et s'inscrit dans un programme plus large de suivi de la biodiversité coordonné par l'Institut de recherche sur les ressources biologiques Alexander von Humboldt [8].

La figure 2 présente le noyau de vraisemblance calculé sur les sons de cette base de données : on identifie sur celle-ci des motifs jour–nuit ainsi qu'une similarité élevée entre les mêmes heures pour des jours différents. Dans l'ensemble, il apparaît que la diversité des coefficients de diffusion $\mathbf{S}\boldsymbol{x}_i$ est plus grande la journée que la nuit.

Nous comparons K-DPP à une méthode naïve, "K tirages uniformes", vérifiant pour tout i la relation de proportionnalité : $\mathbb{P}[\boldsymbol{x}_i \in \mathcal{X}] \propto 1/N$. Afin de prendre en compte la pertinence,

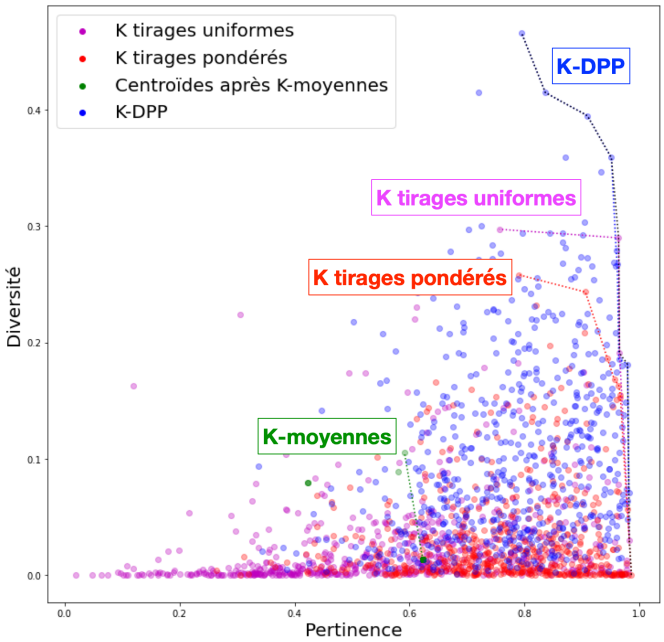

FIGURE 3 – Répartition des 3-échantillons en fonction de leur pertinence et diversité moyenne

nous la comparons également au résultat des "K tirages pondérés", effectués tels que : $\mathbb{P}[\boldsymbol{x}_i \in \mathcal{X}] \propto q_i$. Nous la comparons également à une sélection effectuée grâce à une méthode de référence : l'algorithme des K-moyennes. Celui-ci produit une partition du corpus en K groupes *(clusters)* $\mathcal{C}_1, \dots, \mathcal{C}_K$ qui minimise la variance intra-groupe :

$$\sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in \mathcal{C}_k} \left\| \mathbf{S}\boldsymbol{x}_i - \frac{1}{\operatorname{card} \mathcal{C}_k} \sum_{\boldsymbol{x}_j \in \mathcal{C}_k} \mathbf{S}\boldsymbol{x}_j \right\|^2. \tag{7}$$

On constitue $\mathcal{X}$ en sélectionnant, pour chaque groupe $\mathcal{C}_k$, l'élément le plus proche du centroïde de $\mathcal{C}_k$ dans l'espace $\mathbf{S}$.

3.2 Compromis pertinence–diversité

Comme on le voit sur 3, la méthode d'échantillonnage par K-DPP permet de réaliser un bon compromis entre diversité et pertinence : elle offre une plus grande diversité entre échantillons qu'un tirage seulement basé sur la pertinence tout en garantissant une meilleure pertinence que celle des échantillons obtenus par K-moyennes. La diversité des tirages K-DPP est en moyenne trois fois plus élevée que celle des tirages pondérés par la pertinence tout en atteignant environ 90% de la pertinence de cette dernière.

La figure 3 montre bien le compromis que permet de réaliser K-DPP : les échantillons $\boldsymbol{x}_i$ tirés seulement selon leur pertinence q_i présentent une diversité faible, les échantillons tirés à l'aide de l'algorithme K-moyenne sont plus diversifiés mais bien moins pertinents alors que les tirages par K-DPP dominent avec une pertinence et une diversité plus élevées.

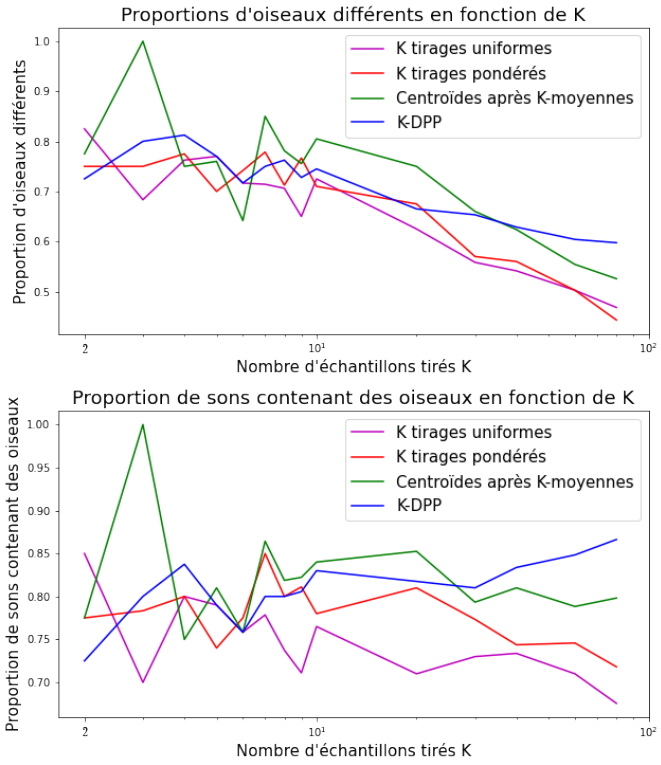

FIGURE 4 – Courbes d'oracle pour le nombre de oiseaux et le nombre de sons contenant des oiseaux

3.3 Inventaire d'espèces

Pour chaque signal $\boldsymbol{x}_i$, nous identifions l'espèce qui maximise la sortie du réseau BirdNET, un réseau de neurones convolutif qui détecte la présence de chants d'oiseau et tâche de les classifier en termes d'espèces. Ce traitement nous permet d'évaluer la proportion de chant d'oiseau dans tout K-uplet, ainsi que la richesse spécifique (nombre d'espèces) correspondante.

La figure 4 met en évidence la corrélation entre la pertinence q_i et la présence d'oiseaux dans un échantillon $\boldsymbol{x}_i$. La méthode "K tirages pondérés" (rouge) conduit à une faible proportion d'oiseaux ainsi qu'à une faible richesse spécifique. Le tirage par K-DPP (bleu) donne de meilleurs résultats : les tirages effectués par cette méthode présentent plus d'oiseaux que toutes les autres méthodes tout en privilégiant des espèces différentes. Cependant, un résultat surprenant est que la méthode basée sur les K-moyennes présente une richesse spécifique égale, voire légèrement supérieure, à K-DPP.

4 Conclusion

L'application du K-DPP en éco-acoustique permet une sélection de paysages sonores du milieu naturel présentant un bon compromis entre pertinence et diversité. BirdNet [6], un réseaux de neurones convolutionnel dans le domaine temps-fréquence, a permis de dresser un inventaire des espèces dans les échantillons sonores. Celui-ci met en évidence que la méthode K-moyennes, grâce à sa tendance à réduire la variance intra-groupe, permet d'obtenir un inventaire plus divers en termes d'espèces d'oiseaux mais les sélections proposées par cette méthode restent moins intuitives et interprétables que celles proposées par K-DPP qui équilibre par construction pertinence et diversité [4].

Pour cette dernière méthode, le noyau de vraisemblance, et plus précisément le choix d'un critère de pertinence et de la méthode de description du signal, jouent un rôle important. Dans notre contexte, le choix de ces deux critères, motivé par des considérations détaillées dans la section 2, a été confirmé par un critère bio-acoustique : l'inventaire de BirdNET. Par ailleurs, ces résultats montrent également que des tirages uniformes ne peuvent pas produire de tirages représentatifs que ce soit en terme de pertinence, de diversité ou même de bilan éco-acoustique. Quant aux tirages pondérés par la pertinence, ils donnent de meilleurs résultats que les tirages uniformes du point de vue de la pertinence mais au prix d'un manque de diversité des espèces d'oiseaux sélectio nnées. Ce manque est dû au biais de cette méthode qui consiste à maximiser le nombre d'évènement très saillants au sein de ses tirages. Cela montre bien l'importance d'équilibrer les deux critères dans ce problème et donc l'intérêt des K-DPP pour l'exploration de données éco-acoustiques.

Références

[1] J. ANDÉN et S. MALLAT. « Deep scattering spectrum ». In : *IEEE Transactions on Signal Processing* 62.16 (2014), p. 4114-4128.

[2] M. ANDREUX et al. *Kymatio: Scattering Transforms in Python*. 2018. DOI : 10.48550/ARXIV.1812.11214. URL : https://arxiv.org/abs/1812.11214.

[3] M. CARTWRIGHT et al. « SONYC-UST-V2: An urban sound tagging dataset with spatiotemporal context ». In : *Proc. DCASE*. 2020.

[4] G. GAUTIER et al. « DPPy: DPP Sampling with Python ». In : *Journal of Machine Learning Research - Machine Learning Open Source Software (JMLR-MLOSS)* 20.180 (2019), p. 1-7.

[5] F. GONTIER et al. « Estimation of the perceived time of presence of sources in urban acoustic environments using deep learning techniques ». In : *Acta Acustica united with Acustica* 105.6 (2019), p. 1053-1066.

[6] S. KAHL et al. « BirdNET: A deep learning solution for avian diversity monitoring ». In : *Ecological Informatics* 61 (2021), p. 101236.

[7] A. KULESZA, B. TASKAR et al. « Determinantal Point Processes for Machine Learning ». In : *Foundations and Trends in Machine Learning* 5.2–3 (2012), p. 123-286.

[8] C. PIZANO et H. GARCÍA, éd. *El bosque seco tropical en Colombia*. Bogotá, D.C., Colombia. : Instituto de Investigación de Recursos Biológicos Alexander von Humboldt (IAvH), 2014.