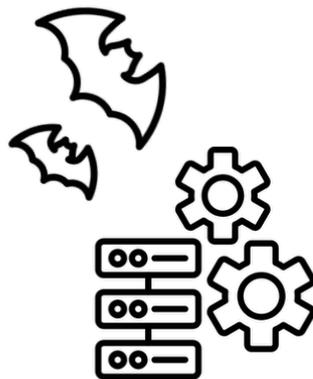
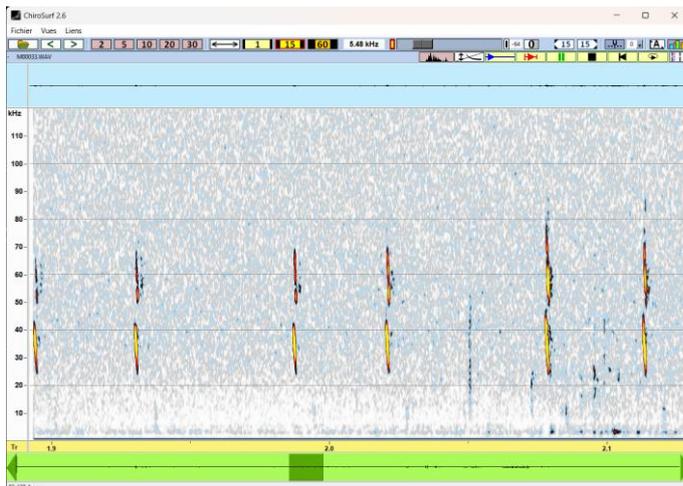




Ecofirst

Biodiversity & Forest Engineering

Recommandations générales en matière de gestion des données acoustique pour les professionnels



Références administratives

Note technique rédigée par

Ecofirst SC

Société coopérative

BE 0692.806.959

www.ecofirst.eu

Adresse postale : Grand-Rue, 12 – B-6870 Awenne - BELGIQUE

Rédacteurs de la note :

Pierrette Nyssen / + 32 473 265 264 / pierrette.nyssen@ecofirst.eu

Jean-François Godeau / +32 472 94 48 47 / jf.godeau@ecofirst.eu



Ecofirst

Biodiversity & Forest Engineering

Illustrations : Couverture : Pierrette Nyssen (Spectrogramme d'Oreillard), icônes www.thenounproject.com
Version 1.0, 2024

Contexte

Ecofirst a la chance de travailler dans le cadre d'une série de projets, sur des jeux de données bruts issus de différents types d'étude, par différents partenaires et collaborateurs (bureaux d'étude, associations, administrations, partenaires académiques, volontaires). Nous avons de ce fait l'occasion de côtoyer différentes manières de travailler, de bien comprendre les pratiques et méthodes appliquées par chacun dans un cadre général de collecte et d'analyse de données acoustiques, en particulier pour l'étude des chiroptères.

Cette note, rédigée sans prétention, vise simplement à formuler, au regard de cette expérience, une série de recommandations en vue d'un traçage aisé des opérations réalisées, et pour éviter une série d'écueils et de problèmes qui peuvent survenir dans les données à différentes étapes de leur analyse. Ces recommandations peuvent être des pistes pour les bureaux d'étude et autres professionnels du secteur pour améliorer leurs pratiques, mais constituent aussi pour les commanditaires des études, des guidelines permettant de s'assurer (que ce soit en amont ou en aval de l'étude) de la qualité du travail réalisé.

Recommandations

1. Organisation des dossiers en interne selon une logique constante

Les enregistrements acoustiques génèrent, dans la plupart des études, de gros volume de données. Un premier élément de bonne gestion est d'adopter des habitudes de classement et d'archivage de ces données d'une manière logique, constante dans le temps et reproductible. Les noms des dossiers et sous-dossiers doivent dans l'idéal être clairs, compréhensibles pour tous (sans équivoque) et suivre une structure invariable. Ceci permettra à chacun (que de soit le porteur du projet, un de ses collègues ou un intervenant externe) de s'y retrouver rapidement, même 10 ans plus tard. De même, un double archivage (avec une version de travail et une version backup) permet d'éviter des déconvenues (telle que la perte accidentelle de tout ou partie des données).

2. Traçage continu des opérations réalisées

Le traitement des données, depuis leur récolte, jusqu'à la production d'un rapport final, implique de multiples opérations. Nous recommandons de créer un fichier de métadonnées associé à chaque étude, qui consigne clairement les différentes opérations menées, la date de l'opération, la personne qui s'en est chargée, le résultat obtenu. Ceci devrait s'accompagner d'une sauvegarde et d'un archivage automatique des paramètres utilisés pour les machines (détecteurs par exemple), des résultats des tests des micros, des versions des logiciels utilisés et des instructions données aux logiciels (options, paramètres, zone géographique, etc). Ceci peut se faire par sauvegarde des fichiers de paramètres au moment de l'opération, par une retranscription dans les métadonnées des codes utilisés pour traiter les données ou encore par une capture d'écran des interfaces de certains logiciels de traitement. Toute correction des données, problème constaté, écart par rapport au protocole habituel, ... devrait être clairement consigné dans les métadonnées également au fur et à mesure du processus de traitement.

Les métadonnées détaillées devraient également expliquer le contexte de la collecte, les paramètres utilisés, la méthode d'analyse, etc. Ces informations permettent aux utilisateurs (que ce soit en interne de la structure ou par rapport à l'extérieur) de comprendre comment les données ont été obtenues et comment elles doivent être interprétées.

3. Processus de validation : traçage des validations réalisées, homogénéisation des pratiques

De nombreux professionnels utilisent des algorithmes de pré-tri des enregistrements. L'outil utilisé varie selon les habitudes de chaque structure et selon le groupe taxonomique (ex : SonoChiro, Tadarida, BatDetect2, BTO acoustic pipeline pour les chauves-souris, BirdNet pour les oiseaux, etc). Bien que les fichiers d'output des algorithmes soient différents d'un outil à l'autre, ils sortent tous un tableau avec une (ou parfois plusieurs) lignes par enregistrement et une proposition d'identification, quasi toujours associée à un indice de confiance sur cette identification.

Quel que soit l'outil utilisé, une validation par un opérateur expérimenté est indispensable pour vérifier la solidité des résultats, chasser les erreurs et permettre une interprétation correcte des résultats. Ce travail de validation doit être tracé lui aussi. Il est de pratique courante de vérifier certaines séquences en les ouvrant dans un logiciel de visualisation, suite à quoi l'opérateur décide de garder l'identification proposée par l'algorithme ou de la modifier. Nous recommandons de garder dans le fichier de validation la trace de l'identification initialement proposé par l'algorithme, que celle-ci soit validée ou invalidée par l'opérateur. L'identité de l'opérateur en charge de la validation doit également être consignée soit dans les métadonnées de l'étude, soit dans le fichier de validation. Si plusieurs personnes travaillent sur le même jeu de données, l'identité du validateur doit être consignée pour chaque séquence validée. Avoir vérifié une série de séquences à la main, l'opérateur peut parfois valider une série de séquences de manière semi-automatique, c'est-à-dire accepter l'identification proposée par l'algorithme sans ouvrir les sons ou modifier l'identification proposée par l'algorithme, à nouveau sans ouvrir les

sons, sur base de l'expérience acquise avec les sons vérifiés. Nous recommandons dans ce cas de garder une trace claire dans le fichier de validation des séquences qui ont effectivement été vérifiées à la main (fichier ouvert) et des séquences qui ont été validées de manière semi-automatique. Enfin, nous recommandons de ne pas consigner d'information (sur l'état de validation par exemple) uniquement via un code couleur dans un tableur mais s'y associer une valeur écrite dans une cellule afin de pouvoir y faire appel lors d'un traitement ultérieur (ce que la couleur ne permet en général pas).

Une homogénéisation des pratiques entre opérateur d'une même structure (voire idéalement entre structures) est importante pour garantir un traitement uniforme des résultats et une comparabilité des résultats des différentes parties d'étude. Ceci comprend le processus de sélection des séquences à valider, la manière de consigner le travail de validation dans le tableur, les seuils de validation semi-automatique, les conditions dans lesquelles le recours à un taxon supra-spécifique est appliqué, les paramètres d'identification utilisés lors de la validation, etc.

L'ensemble de ces pratiques devraient au final permettre de visualiser clairement le travail de validation réalisé, par qui et sur base de quel output de l'algorithme de pré-tri.

4. Vérifications de la durée des contacts

La notion de contact de chauves-souris, à savoir un enregistrement d'un passage de chiroptère, d'une durée maximum de 5 secondes est un élément central de la quantification de l'activité chiroptérologique mesurée. Le nombre de contacts est en effet un élément de comparaison important de l'activité d'une (ou plusieurs) espèce(s) donnée(s), sur un point donné et pendant une durée de temps donnée. Si les détecteurs travaillent de telle sorte que la durée des enregistrements est potentiellement de plus de 5 secondes, il est nécessaire de scinder ces enregistrements plus longs en « tranches » de maximum 5 s, avec un outil tel que le logiciel Kaleidoscope par exemple. En effet, ne pas scinder les enregistrements plus longs que 5 s conduit à une sous-évaluation de l'activité des chauves-souris. Des fichiers de longueurs distinctes (suite à un paramétrage non constant des détecteurs ou suite à un découpage des fichiers sur des durées variables) rend bancaire toute comparaison des activités (entre sites, entre périodes, entre études), que ce soit au sein du même bureau ou entre études produites par bureaux différents.

Nous recommandons une vérification systématique de la durée des enregistrements traités afin de déceler un éventuel problème ou oubli. Ceci peut se faire soit en vérifiant la durée des quelques fichiers les plus lourds, soit en vérifiant le pas de temps qui sépare deux fichiers successifs dans l'horodatage des fichiers.

5. Vérification des dates (horodatage fichiers)

Certains détecteurs nomment les enregistrements selon un horodatage : la date et l'heure du début du fichier sont compris dans le nom de celui-ci. Nous recommandons, pour les détecteurs qui proposent une autre structure de nom de fichier, de commencer le processus de traitement des données par une opération qui renomme tous les fichiers avec un horodatage.

Lorsque l'algorithme de pré-tri reprend dans son tableau de résultat des informations liées à la date et l'heure d'enregistrement, nous recommandons de vérifier qu'elles correspondent effectivement à l'horodatage des fichiers. Une non-correspondance identifie un problème dont il faut chercher la source le plus tôt possible.

NB : une attention sera portée à identifier les problèmes de transformation des heures UTC / UTC+2, qui sont des sources fréquentes d'erreur. Nous recommandons de travailler toujours avec le même référentiel horaire.

6. Systématiser le recours à des taxons supra-spécifique en cas de doute sur l'identification

L'analyse des séquences sonores de chauves-souris est une discipline difficile qui demande de l'expérience et du temps. La capacité d'identification à l'espèce augmente avec la longueur des séquences (nombre de cris de l'individu) et leur qualité. Dans certaines situations, le sonar de plusieurs espèces est très similaire lorsque le

comportement de ces espèces sur le terrain est identique. Il est de ce fait tout à fait normal d'avoir dans les résultats d'études une certaine proportion de séquences identifiées à un niveau taxonomique supra-spécifique (identification au genre ou attribution à une guild, groupes de genres ou d'espèces au sonar proche). Une identification d'une très grande proportion de contacts à l'espèce n'est pas un gage de sérieux mais au contraire signe un taux d'erreur probable plus important. Nous recommandons la prudence et le recours à un taxon supra-spécifique pour les séquences où l'identification n'est pas suffisamment sûre.

7. Systématiser le recours à un contrôle de validation extérieur en cas de détection d'une espèce très rare ou improbable dans la zone géographique concernée

Lorsqu'une espèce peu probable est détectée au cours d'une étude, nous recommandons de confronter l'identification à celle réalisée par un tiers. Les contacts qui doivent attirer notre attention sont les espèces détectées en dehors des limites de leur aire de répartition connue (que ce soit à échelle locale ou internationale), les espèces très rares, les espèces détectées dans des milieux qu'elles ne fréquentent pas habituellement (selon la littérature) ou à des périodes inhabituelles. Ces données doivent être validées avec soin et attention car elles influencent, plus que d'autres, l'état de la connaissance sur ces espèces, leurs habitats et leur aire géographique.

A ce jour, il n'existe aucun comité de validation officiel pour les chauves-souris (ça pourrait être mis en place à l'échelle régionale ou fédérale). Dans ce cas, un partage de la séquence avec un autre expert en acoustique, idéalement issu d'une autre structure, en Belgique ou à l'étranger, pourrait permettre de confirmer l'identification. Dans ce cas, les personnes consultées pour la validation devraient être consignées également dans les métadonnées et/ou le fichier de validation.

8. Rendre les données brutes accessibles au public

Nous recommandons que les données brutes, y compris les enregistrements acoustiques et les fichiers de données brutes validés et les métadonnées de l'étude, soient rendues accessibles au public en même temps que les rapports d'étude ou les résultats produits (rapports par exemple) à partir de ces données. Cette mise à disposition peut se faire par l'upload sur une plateforme dédiée ou en les annexant directement aux rapports d'étude, accompagnées d'une description des métadonnées nécessaires pour une bonne compréhension et utilisation des données.

Pourquoi rendre les données brutes publiques ?

1. Favoriser la transparence et la confiance : La mise à disposition des données brutes permet de garantir la transparence des méthodes utilisées et des résultats obtenus. Cela permet aux parties prenantes, y compris le public, les chercheurs et les autorités, de vérifier l'intégrité des données et de mieux comprendre les conclusions des études. Cela renforce la confiance dans les travaux réalisés et dans les décisions qui en découlent.
2. Soutenir la reproductibilité des études et de ce fait assurer la validité scientifique des conclusions. En mettant les données brutes à disposition, on permet à d'autres de reproduire les études et de vérifier si les résultats sont robustes et fiables, contribuant ainsi à la rigueur scientifique.
3. Encourager la réutilisation des données : En rendant les données brutes accessibles, on permet à d'autres chercheurs, bureaux d'étude ou institutions d'exploiter ces données à des fins d'analyse complémentaire, de validation ou de recherche. Cela peut conduire à de nouvelles découvertes, à l'amélioration des méthodes ou à la réévaluation des conclusions initiales, contribuant ainsi à l'enrichissement collectif des connaissances scientifiques et techniques.

9. Mettre en place un système de contrôles croisés en interne

Toute manipulation de données est susceptible de générer des erreurs, souvent involontaires, mais qui entachent potentiellement la robustesse des résultats. Nous recommandons la mise en place de contrôles, réalisés en interne de la structure par une autre personne que celle qui a pris en charge la manipulation des données. Cette procédure devrait permettre de réduire le risque d'erreurs. Ces contrôles doivent être menés dans le but de tester la cohérence des résultats, la concordance de ceux-ci avec les informations synthétiques présentées dans le rapport (tableaux, graphes), l'exactitude des calcul effectués et des valeurs présentées sous forme de graphiques, de vérifier l'absence de doublons.

Conclusions

Ces recommandations sont loin d'être exhaustives et complètes, mais elles pointent déjà une série de points d'attention utiles. Nous sommes convaincus qu'un travail de qualité ne peut être mené qu'à partir de données manipulées avec soin et rigueur. Chaque structure est responsable de mettre en œuvre les moyens nécessaires pour tendre vers cet idéal. Bon travail à tous !