Développement de workflows en R pour le nettoyage, la validation et la visualisation de données spatiales en écologie

Jordan Benrezkallah[1], Natasha de Manincor[2], Ahlam Sentil[3], Kévin Tougeron[4] & Denis Michez[5]

Résumé

Avec l'augmentation des volumes de données en sciences, il est essentiel de disposer de workflows reproductibles pour le nettoyage, l'exploitation et l'interopérabilité des données. En écologie, ces enjeux sont particulièrement cruciaux pour le suivi des espèces et la conservation. Dans ce cadre, nous avons développé un workflow en R pour traiter de grands jeux de données sur les pollinisateurs européens, en lien avec les projets **PULSE** et **SAFEGUARD**. Ces initiatives visent à alimenter la Liste Rouge européenne des abeilles et analyser les tendances des pollinisateurs à partir de bases de données consolidées. Nous avons ainsi normalisé 69 jeux de données issus de toute l'Europe, représentant plus de 5,5 millions d'enregistrements et couvrant **2081 espèces d'abeilles sauvages.**

Le **workflow** développé en **R** inclut les étapes suivantes :

1. **Importation et correction** des formats (encodage, caractères spéciaux, dates, coordonnées).
2. **Contrôle qualité** (unicité des identifiants, formats des coordonnées, cohérence temporelle).
3. **Vérification des noms d'espèces**, sur base d'une liste d'espèces notamment via des recherches approximatives (*fuzzy search*).
4. **Assignation de la classification taxonomique** en attribuant les rangs taxonomiques.
5. **Attribution des divisions administratives** aux enregistrements.
6. **Génération automatisée de cartes spatiales** pour validation par des experts taxonomistes.

Ce processus permet d'obtenir des données nettoyées et validées pour estimer l'aire de répartition des espèces et évaluer leur statut de conservation. Les données sont structurées selon le **standard Darwin Core**, garantissant leur interopérabilité avec d'autres bases de données et facilitant leur réutilisation. L'ensemble du workflow ainsi que les scripts sont accessibles sur un **dépôt GitHub**, suivant ainsi les principes **FAIR** (*Findable, Accessible, Interoperable, Reusable*).

Cette présentation mettra en avant les défis du traitement de données écologiques massives et illustrera la puissance de R pour leur standardisation et leur visualisation.

Mots-clefs : Biologie - Statistique Spatiale - Data - Reproductibilité

1 Laboratoire de Zoologie et laboratoire d'Écologie des Interactions et Changements globaux, Université de Mons, jordan.benrezkallah@umons.ac.be
2 Laboratoire de Zoologie, Université de Mons, natasha.demanincor@umons.ac.be
3 Laboratoire de Zoologie, Université de Mons, ahlam.sentil@umons.ac.be
4 Laboratoire d'Écologie des Interactions et Changements globaux, Université de Mons, kevin.tougeron@umons.ac.be
5 Laboratoire de Zoologie, Université de Mons, denis.michez@umons.ac.be