Guide pratique pour la valorisation et le post-traitement de données naturalistes en vue d'assurer la qualité des résultats produits

-

Décembre 2019

La robustesse d'une analyse réalisée à partir de données va dépendre de la qualité des données disponibles et de l'adéquation entre les données sélectionnées et la problématique/question posée. Bien connaitre les données utilisées permettra ainsi d'améliorer la qualité des résultats.

LISTE DES PRECONISATIONS (non exhaustive)

La sélection des données pour répondre à un besoin, une question particulière, est une étape indispensable. Avoir une réflexion sur les données à sélectionner pour répondre à une question précise permettra de ne pas biaiser les analyses et de répondre le plus efficacement possible à la question posée. Certaines données doivent en effet parfois être écartées (trop imprécises, doublons, ou ne correspondant pas au périmètre du projet, données trop anciennes par exemple). Cela sous-entend de bien définir la problématique posée au préalable à la réalisation des traitements. Des phases de calibrage et de test afin de définir la meilleure sélection possible sont parfois nécessaire pour atteindre ensuite l'objectif fixé.
Chaque question doit générer une réflexion sur les données susceptibles d'y répondre. On peut donner à titre d'exemple d'usage simple et classique de données naturalistes :

- Identification/actualisation de zones à enjeux (exemple mise à jour des espèces déterminantes d'une ZNIEFF) : sélection de toute donnée correctement déterminée (y compris opportuniste) ;
- Publication de la répartition connue d'une espèce (exemple pour un article naturaliste) : sélection de toute donnée, selon le stade biologique (nicheur/hivernant…), en estimant éventuellement l'absence par le fait que d'autres espèces du même groupe ont été observées ;
- Modélisation de l'aire de distribution probable : selon la technique de modélisation, des données opportunistes peuvent être utilisées ou il faut uniquement s'appuyer sur des données d'inventaire structuré ;
- Comparaison d'intérêt de secteurs : sélection de données d'inventaire structuré (effort de prospection standardisé dans l'espace) ;
- Tendance de présence, d'abondance, de taux de détection : sélection de données des suivis protocolés, répétés dans le temps.

Les métadonnées (lorsqu'elles ont été bien renseignées) permettent de faciliter la recherche de données adaptées aux usages.

Lors des phases d'analyse, il est important de considérer les potentielles erreurs des données et de s'assurer de leur prise en compte lors des traitements. Selon les analyses, l'importance des erreurs de détermination est plus ou moins problématique (parfois peu impactant sur un indice de tendance d'espèce commune, très fort pour la détermination de secteur de conservation pour des espèces rares…).
Connaitre dans la mesure du possible les traitements appliqués en amont sur les données permettra de calibrer sa méthode en fonction. Notamment, il peut s'avérer important de bien considérer la précision géographique des données et les traitements de reprojection (opération fréquente) qui auraient pu être appliqués.

Décrire, documenter le processus utilisé (la méthode) et garder les scripts pour la constitution d'un indicateur, en incluant les règles de sélection des données d'entrée permettra notamment d'assurer la traçabilité des opérations réalisées, mais également de rendre la méthode reproductible. Dans la documentation, être transparent sur les limites de l'analyse au regard des limites identifiées dans les données disponibles.

Bien veiller à citer les sources utilisées lors de la production des résultats, notamment en utilisant les métadonnées. Selon le nombre de contributeurs, on peut recommander de citer tous les contributeurs (peu de contributeurs) ou tous les jeux de données mobilisés (citer ainsi les programmes, leur coordinateurs et financeurs). Cela permet de valoriser les producteurs, mais aussi d'être transparent sur la méthode et les sources utilisées pour l'analyse.

Toujours respecter les droits accordés par les producteurs de données en matière de diffusion en respectant les licences associées. En particulier, veiller à bien respecter les règles concernant la diffusion des données sensibles. C'est-à-dire par exemple de ne pas diffuser dans l'analyse un résultat permettant de localiser la présence d'une espèce sur un niveau plus fin que celui prévu par la liste de sensibilité du SINP :
https://inpn.mnhn.fr/programme/donnees-observations-especes/references/sensibilite