



CERTIFICAT EN ANALYSE DE DONNÉES POUR L'ÉCOLOGIE ET LA GESTION DE LA BIODIVERSITÉ

Cas pratique : construction de modèles sous R

Jean-Yves Barnagaud : jean-yves.barnagaud@ephe.psl.eu

Objectifs : *Savoir déterminer les limites dans lesquelles il est possible de répondre à une question écologique par une analyse statistique en fonction des données disponibles. Identifier les éléments de complexité dans une hypothèse biologique et déterminer dans quelles limites il est possible d'y répondre au regard des données disponibles, sans pour autant se laisser bloquer. A cette fin, réinterpréter la question comme un modèle statistique, l'écrire formellement sous forme d'une équation, vérifier les conditions d'application, l'implémenter sous R, interpréter les résultats avec un regard critique. Différencier significativité statistique et biologique.*

Cas d'étude 1: Nous analysons des données issues d'un programme de points d'écoute d'oiseaux mené dans les forêts du Perche normand par Christian Thiébaud, agent de l'Office National des Forêts, à la fin des années 2000. Nous posons la question suivante : quelle gestion des peuplements forestiers du Perche est optimale pour maximiser l'indicateur écologique « richesse spécifique en oiseaux ».

Hypothèse : Nous ne formulons volontairement pas d'hypothèse, afin que la question reste aussi vague qu'elle le serait dans le cas d'un problème d'analyse réellement posé par un gestionnaire d'espace naturel dans ce contexte. A vous de déterminer les hypothèses qu'il est possible ou non de formuler au regard des données disponibles.

Données : Les données sont stockées dans *oiseaux_perche.txt*. Les points d'écoute (« mpoint ») sont placés dans quatre massifs (« massif ») proches les uns des autres (~10km de distance entre massifs). A l'intérieur d'un massif, les points sont disposés sur une grille régulière de 500x250m. Chaque point est échantillonné une fois, par un relevé auditif et visuel de 5 min, durant lequel tout oiseau observé dans un rayon de 100m est identifié. Nous ne nous intéressons ici qu'à la richesse spécifique (« rs »), nombre d'espèces observées en 5 minutes. Plusieurs autres variables sont récoltées par l'observateur lors de son passage : l'essence dominante sur le point (CHE : chêne, DOU : douglas, EPC : épicéa, PS : pin sylvestre, SP : sapin), le nombre de strates (de la strate herbacée à la strate des arbres matures), la distance du point à la lisière la plus proche (dist_lisiere), la hauteur dominante des arbres matures sur le point (Hdom). Une extraction SIG de données de l'Inventaire Forestier National

a permis de quantifier quelques paramètres paysagers dans un rayon de 500 m autour du point : shannon_area, un indice de diversité de l'habitat, longueur_lisiere, la longueur totale de lisières (m), et mean_area, l'aire moyenne (ha) des patch d'habitat uniforme.

Cas d'étude 2: Les suivis acoustiques de chiroptères sont devenus un standard des études d'impact préalables aux projets d'aménagement du territoire (comme l'installation d'éolienne). L'objectif est de connaître les variations d'activité des chauves-souris sur les sites d'intérêt, au regard de variables temporelles ou spatiales. Dans notre cas d'étude, nous en sommes à une phase exploratoire qui vise à optimiser un futur protocole de suivi des chiroptères. Notre but est de savoir si les périodes d'activité maximale et minimale des chauves-souris sont identiques d'une espèce à l'autre. Nous nous intéresserons d'une part aux variations d'activité intra-saisonnière (mensuelles), d'autre part aux variations quotidiennes (horaires). L'idée est de savoir s'il est possible de resserrer le créneau horaire et de limiter le nombre de campagnes à faire chaque année, tout en continuant à échantillonner au mieux les espèces les plus communes.

Hypothèse : Nous ne formulons volontairement pas d'hypothèse, afin que la question reste aussi vague qu'elle le serait dans le cas d'un problème d'analyse réellement posé par un bureau d'étude dans ce contexte. A vous de déterminer les hypothèses qu'il est possible ou non de formuler au regard des données disponibles.

Données : Les données sont stockées dans « *activite_chiropteres.txt* ». Les données ont été acquises par le bureau d'études Biotope à d'autres fins qui ne seront pas développées ici, et les sites ne sont pas communiqués car ils sont confidentiels par convention : sachez seulement qu'il y a des sites le long de pistes forestières, en bordure de forêt, près de peuplements en régénération – c'est-à-dire dans divers contextes. Les données en revanche sont bien réelles : nous avons supprimé des variables, mais pas des individus. Les champs sont : « point » = l'identifiant de chaque point de relevé, « Date_nuit » = la date de la nuit du relevé, « julian_date » : la même date, réexprimée en jours depuis le 1^{er} janvier pour faciliter l'analyse, « Mois », « Heure », « Minute » : colonnes explicites ; « heure.rel » = heure relative centrée sur minuit afin de ne pas couper la nuit en deux (de 12h à 00h : heures négatives, de 00h à 12h : heures positives) ; « Espece_validee » = les espèces identifiées (nous avons supprimé les données correspondant à des espèces non identifiées), « NbCris » = le nombre de cris compté sur chaque séquence enregistrée ; on fait l'hypothèse qu'ils reflètent l'activité des chauves-souris, mélange indiscernable entre le nombre d'individus et le niveau d'excitation des individus. (NB : il existe d'autres mesures plus fines de l'activité des chauves-souris, mais par souci de simplicité nous en resterons à celle-ci).

Tâches à effectuer :

- Confronter la question posée aux données disponibles afin de la reformuler et le cas échéant la recentrer sur ce qu'il est réellement possible de faire. Argumenter ce choix et ses limites.
- Reformuler la question sous forme d'un modèle statistique, en écrire l'équation et l'implémenter sous R.

- Effectuer les vérifications nécessaires, interpréter le modèle et si besoin, l'ajuster en fonction des contraintes imposées par la question et les données.
- Préparer une restitution orale de 10 minutes maximum. Il s'agira de proposer une solution au problème posé par le gestionnaire, lui expliquer les enjeux méthodologiques qui vous ont conduit à prendre vos décisions, rendre compte de la méthodologie afin qu'elle soit compréhensible et expliquer les résultats. Enfin, proposer une solution au problème posé en identifiant ses limites et en proposant des priorités argumentées pour un prochain échantillonnage.

Conseils : Les questions posées sont aussi vagues que celles auxquelles vous aurez à répondre dans un cas réel – elles sont d'ailleurs parfaitement classiques. La difficulté vient du fait que les données disponibles ne permettent pas, comme souvent, de répondre à la question de manière exhaustive – voire ne permettent de la traiter que sous un angle très particulier. A vous de ne pas vous laisser bloquer, en identifiant comment répondre au problème de manière informative pour le gestionnaire sans sacrifier la rigueur statistique. Mettez-vous à la place d'un directeur de travaux qui a besoin d'une réponse opérationnelle afin de prendre une décision sur des mesures à mettre en place.