École Pratique des Hautes Études
Sciences de la vie et de la terre

PSL

Niveau 2 : renforcement

Du 13/03/2023 au 17/03/2023

Formateur : Jean-Yves Barnagaud (jean-yves.barnagaud@ephe.psl.eu ; tél : 06 16 33 41 07)

Lieu : Maison des Sciences de l'Homme, 54 boulevard Raspail 75006 PARIS
Salle 15 sauf mercredi : salle 33 (sous-sol du bâtiment)

Pour s'y rendre :

Métro : lignes 10-12, Sèvres-Babylone ; ligne 12, Rennes ; ligne 4, Saint-Placide
Bus : lignes 68-94, Rue du Cherche-Midi (*le plus proche*) ; 70-86, Sèvres-Babylone
A pied : 800 mètres, 10 min de la Gare Montparnasse

Horaires :

9h-12h, 13h30-17h (sauf lundi : début à 10h et vendredi : fin à 15h30).

Pause méridienne : diverses possibilités de restauration existent à proximité immédiate, en particulier rue du Cherche-Midi. Il n'y a pas de service de restauration dans la Maison des Sciences de l'Homme, mais vous y trouverez des tables en intérieur et extérieur.

Matériel à prévoir :

- Un ordinateur portable (de préférence sous Windows, mais les Mac sont acceptés) avec MS Excel ou tout autre logiciel de tableur.
- Installer au préalable les logiciels gratuits R et R-studio, téléchargeables à ces adresses *(contacter le formateur en cas de difficultés)* :
https://cran.r-project.org/
https://rstudio.com/products/rstudio/download/

- Depuis R-Studio, allez dans l'onglet "Packages" (en bas à droite) et installez les packages suivants : ggplot2, formattable, cowplot, sjPlot, effects, jtools, ggeffects, visreg, ggpubr, car, nlme, mgcv, arm, questionr, MASS, corrplot, pscl, interactions, investr, lmtest, boot, polynom, biostat3, lubridate, ade4, vegan, adegraphics, FactoMineR

Bienvenue à la formation de l'École Pratique des Hautes Études en analyse de données pour l'écologie et la gestion de la biodiversité. Cette formation ambitionne de vous donner des bases généralistes solides afin de vous permettre de construire et d'analyser vos propres jeux de données, sans prérequis initial. Elle vous fournira les compétences nécessaires afin de vous approprier des méthodes plus spécialisées, lorsque vous en aurez besoin : suivis de populations, modèles de distributions d'espèces, analyses génétiques…

Ce deuxième niveau vous permettra de renforcer vos capacités d'analyse et de synthèse de données, tout en vous apportant la flexibilité nécessaire pour vous adapter à la diversité des questions et des données que vous rencontrerez en écologie. Il vise à vous donner une réelle maîtrise autonome du flux d'analyse de données, de leur exploration à l'inférence. Nous nous focaliserons sur une méthode particulière, le modèle linéaire. S'il est loin d'être le seul chemin possible pour vos analyses et certainement pas le plus à la pointe ou à la mode (comme le sont les méthodes de classification non-supervisées ou l'intelligence artificielle), le modèle linéaire est sans aucun doute la méthode statistique la plus pratiquée par les écologues, qu'ils soient académiques ou issus des structures de gestion. Sa simplicité d'implémentation, sa robustesse et son cadre unifié y sont sans doute pour quelque chose.

A l'issue de ce niveau, vous serez en mesure de construire des modèles opérationnels répondant à des objectifs d'analyses variés. Vous saurez interpréter les résultats, les présenter, mais vous connaîtrez aussi les limites au-delà desquelles des méthodes plus complexes deviennent indispensables. Nous utiliserons essentiellement des jeux de données réels, prêtés pour cette formation par une diversité de chercheurs et d'organismes que vous connaissez : CNRS, OFB, INRAE, … Nous accorderons les après-midi à des cas concrets implémentés sous le logiciel R. Ce module est orienté vers la pratique telle que vous la connaîtrez en contexte réel. Nous irons peu vers la théorie statistique : en tant qu'utilisateur, votre but premier doit être avant tout de savoir manipuler correctement les concepts et les outils, plus que de les décortiquer dans toute leur complexité. Parce que cette formation est résolument orientée vers la pratique, soyez-en acteurs : questionnez, échangez, critiquez, afin que cette semaine soit la plus riche possible.

Niveau 1 - découverte	**Niveau 2 - renforcement**	**Niveau 3 – approfondissement**
06-03-2023 au 10-03-2023	*13-03-2023 au 17-03-2023*	*20-03-2023 au 24-03-2023*
Une introduction à l'analyse de données et à l'environnement logiciel R	Une exploration plus poussée des techniques de modélisation statistique afin de répondre à des questions classiques d'étude et de suivi de la biodiversité	Une introduction aux techniques avancées d'analyse pour les jeux de données complexes, classiques des suivis écologiques à long terme ou large échelle

Objectifs du deuxième niveau

Objectif	Quand ?
Développer une démarche d'exploration de vos données en exploitant les approches multivariées *Une introduction à l'exploration graphique de jeux de données écologiques formés de nombreuses variables*	Lundi
Construire un modèle pour tester une hypothèse écologique *Formaliser une hypothèse biologique sous forme d'équation, implémenter, vérifier et interpréter ce modèle via le logiciel R*	Mardi
Modéliser des données d'occurrences, de fréquences et de comptages en utilisant un modèle linéaire généralisé *Exploiter la flexibilité du GLM pour tester des hypothèses sur des types de données communs en écologie, savoir en déterminer les limites*	Mercredi
Modéliser des optimums de réponse et des interactions entre variables *Traduire une question écologique complexe en un modèle statistique pertinent et interprétable ; présenter les résultats d'un modèle*	Jeudi
Questions personnalisées et bilan de compétences *Exercices et notions supplémentaires à la demande, examen facultatif de fin de formation.*	Vendredi

Organisation de la formation :

La formation se divise en deux types de séquences :

Les séquences de cours : elles visent à développer le principe et l'usage des méthodes d'analyse de données à partir de cas concrets. Si elles suivent un plan et un support prédéfinis, il ne s'agit pas de conférences ou de cours magistraux : nous adapterons leur contenu et le format en fonction de la progression du groupe et de vos questions.

Les cas pratiques : leur objectif est d'appliquer les méthodes vues en cours à des cas d'étude concrets issus de véritables jeux de données écologiques. Les cas pratiques seront structurés sous forme de problèmes concrets à travailler en groupes : la collaboration est souvent la clé d'une analyse de données réussie. Afin de vous permettre de vous approprier le sujet et les méthodes, vous serez autonomes sur la réalisation des analyses, le formateur étant présent en appui pour répondre de manière personnalisée à vos questions.

Supports de formation :

Les supports de formation vous sont fournis au format Power-Point ou PDF. Vous disposerez aussi des jeux de données et des scripts R commentés, nécessaires à la réplication des cas pratiques. N'hésitez pas à contacter le formateur en cas de questions sur ces scripts. L'ensemble des supports de cours vous sera fourni le lundi par le formateur.

Validation

La validation des acquis se fera à travers un examen sur table de deux heures vendredi après-midi, qui vous confrontera à des cas pratiques sur lesquels vous devrez formuler des réponses courtes. Elle repose sur l'ensemble de la formation. La validation est acquise pour une note de 10/20. Cet examen n'est pas conditionnant pour votre inscription éventuelle au troisième niveau de formation, mais il vous permettra de vous situer au regard des compétences essentielles à conforter ou acquérir avant d'aller plus loin.

Et ensuite ?

La modélisation statistique par régression linéaire est un vaste domaine qui a pour lui la flexibilité et la robustesse, mais aussi certaines limites. En fin de semaine, vous saurez analyser en autonomie des questions complexes, mais avec des données relativement simples : échantillonnage aléatoire ou systématique non stratifié, sans structuration spatiale ou temporelle, sans erreurs de détection. Cette étape d'appropriation des méthodes de modélisation linéaire est essentielle et vous permettra déjà de répondre à nombre de vos objectifs, mais vous constaterez rapidement que les données écologiques sont complexes : effets régions, tendances interannuelles, hétérogénéités liées à l'observateur, espèces rares… Le niveau 3 vous procurera une véritable compétence pratique afin d'analyser en autonomie ces jeux de données complexes, souvent structurés sur le long terme ou à large échelle spatiale. Il vous donnera aussi la capacité de vous approprier en autonomie des méthodes statistiques nouvelles à partir de la littérature, et de communiquer ces méthodes à d'autres – en somme, devenir une personne-ressources en analyse statistique de données écologiques.

Jour	Heures	Thèmes	Méthodes abordées	Compétences
Lundi *Salle 15*	**10h-12h**	Explorer des données écologique par les méthodes d'ordinations	Indices de dissimilarité pour l'écologie, analyses factorielles des correspondances, analyses en composantes principales, analyses en coordonnées principales, analyse discriminante	Explorer et synthétiser un jeu de données constitué de nombreuses variables. Comprendre et interpréter des dissimilarités entre sites ou espèces.
	13h30-17h	Travaux pratiques : les ordinations sous ade4		Construire et comprendre une analyse multivariée sous R, s'en servir pour explorer un jeu de données, communiquer les résultats
Mardi *Salle 15*	**9h-12h**	Formaliser une question biologique en un modèle statistique à plusieurs variables, le vérifier et l'interpréter	Modèle linéaire	Comprendre ce qu'est un modèle statistique et ses usages, écrire et exploiter son formalisme en fonction d'une question écologique, comprendre la méthode d'estimation des paramètres et leur signification.
	13h30-17h	Travaux pratiques : construire un modèle linéaire et en présenter les résultats		Définir une stratégie de modélisation, l'implémenter sous R, représenter graphiquement les résultats, expliquer et justifier la structure d'un modèle à un public averti
Mercredi *Salle 33*	**9h-12h**	Modéliser des comptages et des présences/absences	Modèle linéaire généralisé binomial et de Poisson, traitement de comptages surdispersés (modèles quasi-Poisson, GLM négatif-binomial, zero-inflated poisson)	Adapter le modèle linéaire à des variables non gaussiennes fréquentes en écologie. Modéliser des occurrences d'espèces rares
	13h30-17h	Travaux pratiques : implémenter un GLM sous R		Identifier les contraintes d'un jeu de données écologiques et les transcrire en un modèle linéaire généralisé approprié. Implémenter, représenter et interpréter un GLM, communiquer les résultats à un public non averti
Jeudi *Salle 15*	**9h-12h**	Modéliser des hypothèses écologiques complexes	Modélisation linéaire de structures quadratiques (polynômes du second degré), termes d'interaction, sélection de modèles, modèles non-linéaires	Identifier le compromis entre complexité et parcimonie dans un cadre de modélisation statistique.
	13h30-17h	Travaux pratiques : modéliser au plus près de l'hypothèse		Interpréter les résultats d'un modèle linéaire complexe, connaître ses limites et identifier des solutions, exposer ses résultats afin de valoriser leur dimension appliquée
Vendredi *Salle 15*	**9h-12h** **13h30-15h30**	Questions personnalisées Évaluation du niveau 2	Tous les thèmes abordés dans la semaine	