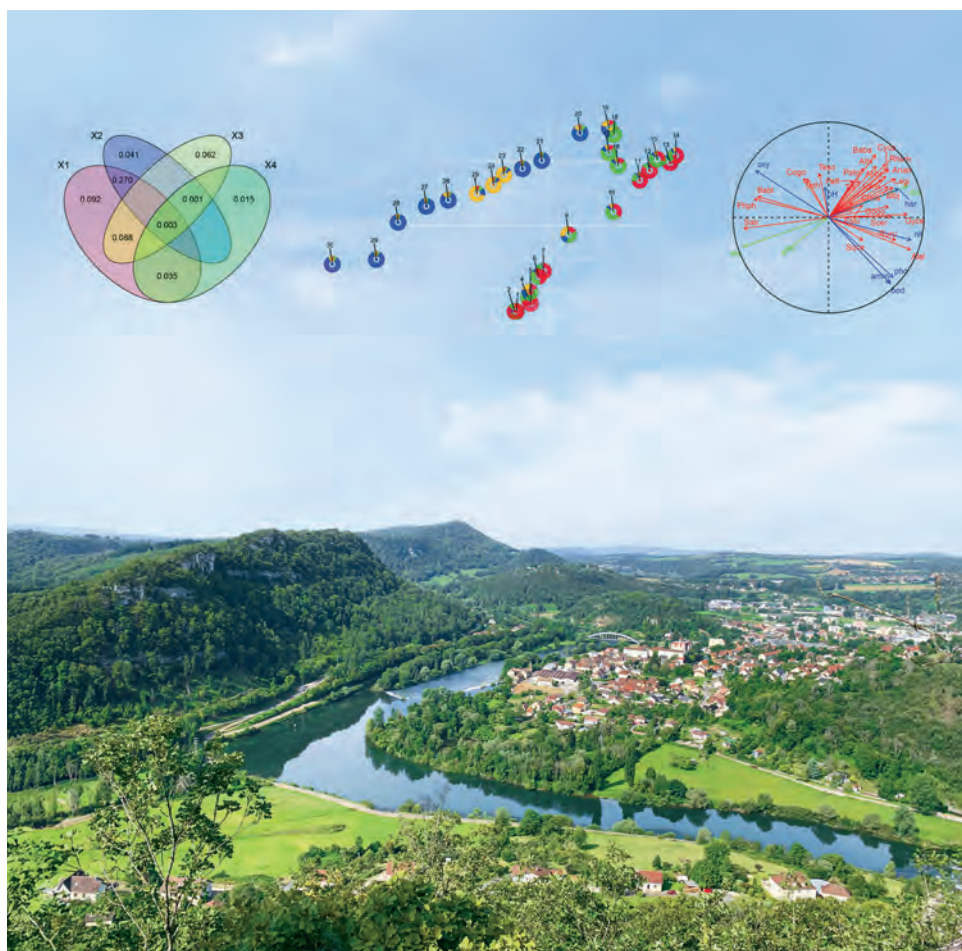




Daniel Borcard, François Gillet et Pierre Legendre

L'écologie numérique en R



Les Presses de l'Université de Montréal

Table des matières

Préface	i
Introduction	1
1.1 Pourquoi l'écologie numérique ?	1
1.2 Pourquoi R ?	2
1.3 Lectorat et structure du livre	3
1.4 Comment utiliser ce livre	4
1.5 Les jeux de données	5
1.5.1 Les données de poissons du Doubs	6
1.5.2 Les données d'acariens oribates	8
1.6 Noms et sigles des méthodes	9
1.7 Petit rappel sur les sources d'aide	13
1.8 Il est maintenant temps... ..	14
Analyse exploratoire des données	15
2.1 Objectifs	15
2.2 Exploration des données	15
2.2.1 Extraction des données	15
2.2.2 Données d'espèces : premier contact	17
2.2.3 Données d'espèces : un examen plus approfondi	19
2.2.4 Transformation des données écologiques	25
2.2.5 Données environnementales	32
2.3 Conclusion	39
Mesures et matrices d'association	41
3.1 Objectifs	41
3.2 Les principales catégories de mesures d'association (un bref aperçu)	42
3.2.1 Mode Q et mode R	42
3.2.2 Coefficients symétriques ou asymétriques en mode Q : le problème du double zéro	42
3.2.3 Mesures d'association pour données qualitatives ou quantitatives	44
3.2.4 En résumé... ..	44
3.3 Mode Q : calcul de matrices de dissimilarité entre objets	44
3.3.1 Mode Q : données d'espèces quantitatives	46
3.3.2 Mode Q : occurrences d'espèces (présence-absence)	49
3.3.3 Mode Q : données quantitatives (excluant les abondances d'espèces) ..	54

3.3.4 Mode Q: données binaires (excluant les présences-absences d'espèces)	56
3.3.5 Mode Q: types mixtes comprenant des variables catégorielles (qualitatives multiclassées, facteurs)	57
3.4 Mode R: calcul de matrices de dépendance entre variables	59
3.4.1 Mode R: données d'abondance d'espèces	59
3.4.2 Mode R: données de présence-absence d'espèces	60
3.4.3 Mode R: données quantitatives et ordinales (autres que les abondances d'espèces)	60
3.4.4 Mode R: données binaires (autres que les données d'espèces)	63
3.5 Prétransformations de données d'espèces	63
3.6 Conclusion	64
Le groupement	67
4.1 Objectifs	67
4.2 Le groupement: survol	67
4.3 Groupement hiérarchique basé sur les liens	71
4.3.1 Groupement agglomératif à liens simples	71
4.3.2 Groupement agglomératif à liens complets	72
4.4 Groupement agglomératif moyen	74
4.5 Groupement de Ward selon la variance minimum	76
4.6 Groupement flexible	77
4.7 Interprétation et comparaison de résultats de groupements hiérarchiques	79
4.7.1 Introduction	79
4.7.2 Corrélation cophénétique	80
4.7.3 La recherche de groupes interprétables	83
4.7.3.1 Graphique des niveaux de fusion	83
4.7.3.2 Comparer deux dendrogrammes pour mettre en évidence les branches communes	87
4.7.3.3 Rééchantillonnage bootstrap multi-échelle	88
4.7.3.4 Largeurs de silhouette moyennes	90
4.7.3.5 Comparaison entre la matrice de dissimilarité et des matrices binaires représentant des partitions	92
4.7.3.6 Analyse de fidélité des espèces	94
4.7.3.7 Diagramme de silhouette de la partition finale	96
4.7.3.8 Dendrogramme final avec options graphiques	98
4.7.3.9 Diagramme spatial du résultat du groupement	102
4.7.3.10 Diagramme de sériation et tableau de communauté ordonné	102
4.8 Partitionnement non hiérarchique	106
4.8.1 Partitionnement <i>k</i> -means	107
4.8.1.1 <i>k</i> -means avec départs aléatoires	107
4.8.1.2 Utilisation du partitionnement <i>k</i> -means pour optimiser une classification obtenue de manière indépendante	111
4.8.2 Partitionnement autour des médoïdes PAM (<i>partitioning around medoids</i>)	114

4.9 Comparaison avec des données environnementales	117
4.9.1 Comparaison d'une typologie avec des données externes : approche ANOVA	118
4.9.2 Comparaison de deux typologies : approche par tableau de contingence	122
4.10 Assemblages d'espèces	122
4.10.1 Statistiques simples sur le contenu des groupes	122
4.10.2 Le coefficient de concordance <i>W</i> de Kendall	123
4.10.3 Assemblages d'espèces dans des données de présence-absence	127
4.10.4 Réseau de cooccurrence d'espèces	128
4.11 Espèces diagnostiques ou indicatrices	132
4.11.1 Introduction	132
4.11.2 IndVal : valeurs indicatrices d'espèces	132
4.11.3 Indices de type corrélation	137
4.12 Arbre de régression multivariable (ARM) : groupement sous contrainte	139
4.12.1 Introduction	139
4.12.2 Principe du calcul	140
4.12.2.1 Partitionnement des données sous contrainte	140
4.12.2.2 Validation croisée des partitions et élagage de l'arbre	141
4.12.2.3 Calcul de la validation croisée	142
4.12.3 Application à l'aide des paquets mvpart et MVPARTwrap	143
4.12.4 Combinaison de l'ARM et d'IndVal	149
4.13 L'ARM comme méthode de groupement monothétique	151
4.14 Groupement sous contrainte de contiguïté spatiale ou temporelle	153
4.15 Une approche très différente : le groupement flou	157
4.15.1 Le groupement <i>c</i> -means par la fonction fanny() du paquet cluster	158
4.15.2 Groupement flou avec objets non classés (NC, <i>Noise clustering</i> en anglais), à l'aide de la fonction vegclust()	163
4.16 Conclusion	167
Ordination en espace réduit d'un seul tableau de données	169
5.1 Objectifs	169
5.2 Vue d'ensemble de l'ordination	170
5.2.1 Espace multidimensionnel	170
5.2.2 Ordination en espace réduit	170
5.3 Analyse en composantes principales ACP (<i>principal component analysis</i> , PCA)	172
5.3.1 Survol	172
5.3.2 ACP des variables environnementales du Doubs avec rda()	173
5.3.2.1 Préparation des données	173
5.3.2.2 ACP d'une matrice de corrélation	174
5.3.2.3 Extraction, interprétation et représentation graphique des résultats dans la sortie d'une ordination de vegan	177

5.3.2.4	Projection de variables supplémentaires dans un biplot d'ACP	181
5.3.2.5	Projection d'objets supplémentaires dans un biplot d'ACP	183
5.3.2.6	Combinaison de résultats d'un groupement et d'une ordination	185
5.3.3	ACP sur des données d'espèces transformées	186
5.3.3.1	Application aux données de poissons transformées par Hellinger	187
5.3.3.2	Explication passive (post hoc) des axes par des variables environnementales	189
5.3.4	Domaine d'application de l'ACP	190
5.3.5	L'ACP par la fonction PCA.newr()	191
5.3.6	Imputation de valeurs manquantes en ACP	191
5.4	Analyse factorielle des correspondances AFC (<i>correspondence analysis</i> , CA)	196
5.4.1	Introduction	196
5.4.2	L'AFC par la fonction cca() du paquet vegan	198
5.4.2.1	Exécution de l'analyse et traçage des biplots	198
5.4.2.2	Projection de sites ou d'espèces supplémentaires dans un diagramme d'AFC	201
5.4.2.3	Ajustement a posteriori de courbes de variables environnementales	202
5.4.2.4	Réorganisation du tableau de données sur la base d'un axe d'ordination	203
5.4.3	AFC par la fonction CA.newr()	204
5.4.4	Effet d'arche et analyse des correspondances détendancée ACD (<i>detrended correspondence analysis</i> , DCA)	205
5.4.5	Analyse des correspondances multiples ACM (<i>multiple correspondence analysis</i> , MCA)	206
5.4.5.1	ACM des variables environnementales du jeu de données des oribates	207
5.5	Analyse en coordonnées principales ACoP (<i>principal coordinate analysis</i> , PCoA)	211
5.5.1	Introduction	211
5.5.2	Application de l'ACoP aux données du Doubs par cmdscale() de stats et wascores() de vegan	213
5.5.3	Application de l'ACoP aux données du Doubs par la fonction pcoa()	214
5.6	Positionnement multidimensionnel non métrique PMnM (<i>Nonmetric multidimensional scaling</i> , nMDS)	218
5.6.1	Introduction	218
5.6.2	Application aux données de poissons du Doubs	219
5.6.3	ACoP ou PMnM?	222
5.7	Une fonction d'ACP simplifiée à programmer étape par étape	225

Ordination canonique	229
6.1 Objectifs	229
6.2 Survol de l'ordination canonique	230
6.3 Analyse canonique de redondance, ACR (<i>redundancy analysis, RDA</i>)	231
6.3.1 Introduction	231
6.3.2 ACR des données du Doubs	233
6.3.2.1 Préparation des données	233
6.3.2.2 ACR par vegan	234
6.3.2.3 Récupérer, interpréter et visualiser les résultats d'un objet de sortie d'ACR de vegan	239
6.3.2.4 Tests par permutation des résultats d'une ACR	246
6.3.2.5 ACR partielle	249
6.3.2.6 Sélection de variables explicatives	252
6.3.2.7 Reconstitution environnementale : projection de nouveaux sites dans une ACR pour estimer les valeurs de variables explicatives	261
6.3.2.8 Partitionnement de la variation	262
6.3.2.9 L'ACR en tant qu'outil pour l'ANOVA multivariable	268
6.3.2.10 Relations non linéaires en ACR	275
6.3.3 Analyse de redondance basée sur la distance, ACR-bd (<i>distance-based redundancy analysis, db-RDA</i>)	281
6.3.4 Une fonction d'ACR écrite à la main	286
6.4 Analyse canonique des correspondances, ACC (<i>canonical correspondence analysis, CCA</i>)	288
6.4.1 Introduction	288
6.4.2 ACC des données du Doubs	290
6.4.2.1 ACC par vegan	290
6.4.2.2 Triplot d'ACC	290
6.4.2.3 Tests par permutation en ACC, sélection ascendante et ACC parci- monieuse	294
6.4.2.4 Graphiques interactifs tridimensionnels	295
6.5 Analyse discriminante linéaire, ADL (<i>linear discriminant analysis, LDA</i>) .	297
6.5.1 Introduction	297
6.5.2 Analyse discriminante à l'aide de lda ()	298
6.6 Autres analyses asymétriques	303
6.6.1 Courbes de réponse principales CRP (<i>principal response curves, PRC</i>)	303
6.6.2 Analyse de co-correspondance, ACoC (<i>co-correspondence analysis, CoCA</i>)	306
6.7 Analyse symétrique de deux ou plusieurs tableaux de données	309
6.8 Analyse des corrélations canoniques, ACoC (<i>canonical correlation analysis, CCorA</i>)	310
6.8.1 Introduction	310
6.8.2 Analyse des corrélations canoniques à l'aide de la fonction CCorA ()	311

6.9	Analyse de co-inertie, ACOI (<i>co-inertia analysis, CoIA</i>)	313
6.9.1	Introduction	313
6.9.2	Analyse de co-inertie par la fonction coinertia() d' ade4	314
6.10	Analyse factorielle multiple, AFM (<i>multiple factor analysis, MFA</i>)	318
6.10.1	Introduction	318
6.10.2	Analyse factorielle multiple par FactoMineR	320
6.11	Mise en relation des traits fonctionnels des espèces avec l'environnement	325
6.11.1	La méthode du quatrième coin	327
6.11.2	Analyse RLQ	328
6.11.3	Application en R	329
6.12	Conclusion	335
	Analyse spatiale des données écologiques	339
7.1	Objectifs	339
7.2	Structures spatiales et analyse spatiale : un bref aperçu	340
7.2.1	Introduction	340
7.2.2	Dépendance spatiale induite et autocorrélation spatiale	341
7.2.3	Échelles spatiales	342
7.2.4	Hétérogénéité spatiale	344
7.2.5	Fonctions de corrélation ou d'autocorrélation spatiale et corrélogrammes spatiaux	344
7.2.6	Test statistique de la présence d'une corrélation spatiale : conditions d'application	349
7.2.7	Modélisation des structures spatiales	351
7.3	Analyse de données multivariées par surfaces polynomiales	352
7.3.1	Introduction	352
7.3.2	L'analyse des surfaces polynomiales en pratique	353
7.4	Vecteurs propres spatiaux et modélisation spatiale	357
7.4.1	Introduction	357
7.4.2	Coordonnées principales de matrices de voisinage (PCNM) et cartes des vecteurs propres de Moran basées sur la distance (dbMEM)	358
7.4.2.1	Introduction	358
7.4.2.2	Variables dbMEM sur plans d'échantillonnage réguliers	360
7.4.2.3	Analyse dbMEM des données des acariens	362
7.4.2.4	Analyse dbMEM simplifiée : la fonction quickMEM()	371
7.4.2.5	Combiner l'analyse dbMEM et le partitionnement de la variation	372
7.4.3	Les MEM dans un contexte plus large : poids des liens autres que les distances géographiques	378
7.4.3.1	Introduction	378
7.4.3.2	Analyse MEM généralisée des données des acariens	381
7.4.3.3	Autres types de matrices de connectivité	388
7.4.3.4	Contrôle de la corrélation spatiale à l'aide de MEM	394

7.4.3.5 MEM dans les plans d'échantillonnage à échelles spatiales imbriquées	394
7.4.4 Variables MEM avec corrélation spatiale positive ou négative : lesquelles utiliser ?	395
7.4.5 Cartes de vecteurs propres asymétriques (AEM) : quand la directionnalité est importante	396
7.4.5.1 Introduction	396
7.4.5.2 Principe et application de la méthode	397
7.5 Une autre façon d'examiner les structures spatiales : l'ordination multi-échelle OMÉ (<i>multiscale ordination</i> , <i>MSO</i>)	402
7.5.1 Principe	402
7.5.2 Application aux données d'acariens - Approche exploratoire	403
7.5.3 Application aux données détendancées des acariens et de l'environnement	407
7.6 Test de l'interaction spatio-temporelle en ANOVA multivariable sans réplication des observations	411
7.6.1 Introduction	411
7.6.2 Test de l'interaction espace-temps avec les fonctions stimodels() et quicksti()	415
7.7 Conclusion	418
Diversité des communautés	419
8.1 Objectifs	419
8.2 Les multiples facettes de la diversité	420
8.2.1 Introduction	420
8.2.2 La diversité spécifique résumée à un simple nombre	421
8.2.2.1 Diversité spécifique et raréfaction	421
8.2.2.2 Les composantes de la diversité de l'abondance des espèces : richesse et équitabilité	422
8.2.3 Les indices de diversité taxonomique en pratique	424
8.2.3.1 Indices de diversité spécifique des communautés de poissons de la rivière Doubs	424
8.2.3.2 Analyse de raréfaction des communautés d'oribates	426
8.3 Quand l'espace est important: les diversités alpha, bêta et gamma	430
8.4 La diversité bêta	430
8.4.1 La diversité bêta mesurée par un simple nombre	430
8.4.2 La diversité bêta comme variance de la matrice de composition des communautés : indices SCBD et LCBd	434
8.4.2.1 Introduction	434
8.4.2.2 Calcul de la diversité bêta comme $\text{Var}(\mathbf{Y})$	435
8.4.2.3 Propriétés des mesures de dissimilarité en rapport avec la diversité bêta	436
8.4.2.4 Contributions locales à la diversité bêta (LCBD)	437
8.4.2.5 Contributions des espèces à la diversité bêta (SCBD)	438

8.4.2.6 Calcul des LCBD et SCBD avec beta.div() du paquet adespatial	438
8.4.3 Partitionnement de la diversité bêta en composantes de remplacement, de différence de richesse et d'imbrication	441
8.4.3.1 La famille d'indices de Podani	443
8.4.3.2 La famille d'indices de Baselga	445
8.4.3.3 Les saveurs de l'imbrication	447
8.4.3.4 Calcul du remplacement, de la différence de richesse et de l'imbrication à l'aide de beta.div.comp() du paquet adespatial	450
8.4.3.5 Explication des indices de remplacement et de différence de richesse	456
8.5 Diversité fonctionnelle, composition fonctionnelle et diversité phylogénétique des communautés	460
8.5.1 Diversité fonctionnelle alpha	460
8.5.2 Diversités taxonomique, phylogénétique et fonctionnelle bêta	464
8.6 Conclusion	469
Références	471
Références – Paquets R (par ordre alphabétique)	483
Index	489

Préface

L'écologie a la cote. Par conséquent, enseigner l'écologie consiste en l'art de présenter une matière fascinante à un auditoire conquis d'avance. Ce n'est pas facile : la complexité de la science écologique moderne s'étend bien au-delà des chapitres d'introduction enseignés au collège et des merveilleux documentaires présentés dans les médias. Mais un auditoire bien disposé est prêt à consentir à cet effort. L'écologie numérique, c'est une autre histoire. Pour des raisons obscures une majorité des personnes attirées par l'écologie sont étrangement réticentes à l'aborder sous une forme quantitative et à utiliser des outils mathématiques pour aider à la comprendre. C'est comme si la nature était intrinsèquement non mathématique, ce qu'elle n'est certainement pas : les mathématiques sont le langage commun à toutes les sciences. Les enseignants en biostatistique et en écologie numérique doivent par conséquent surmonter cette réticence. Avant même d'aborder le sujet lui-même ils doivent convaincre leur auditoire de son intérêt et de sa nécessité.

Durant de nombreuses décennies les écologues, qu'ils soient étudiants ou chercheurs (dans les sphères privée, académique ou gouvernementale), ont planifié leurs recherches et collecté leurs données sans ou avec très peu de considérations statistiques, pour ensuite confier l'analyse « statistique » de leurs résultats à des personnes engagées dans ce but. Ces personnes ont parfois été des statisticiens compétents et dans de nombreux cas elles ont participé à l'intégration progressive de la statistique dans l'ensemble du processus de la recherche. Dans d'autres cas en revanche, le résultat consistait en une grande quantité de données résumées par une poignée de statistiques de base et de tests de signification qui étaient très loin de révéler toute la richesse des structures cachées dans ces tableaux. La séparation des mondes de la statistique et de l'écologie présentait de nombreux problèmes. Le plus important était que les écologues n'avaient pas connaissance de l'éventail de méthodes disponibles à cette époque, et que les statisticiens n'étaient pas conscients des hypothèses écologiques à tester ni des contraintes spécifiques aux données écologiques (le problème du double zéro est un bon exemple). En plus d'empêcher le traitement adéquat des données, cette double incompréhension freinait le développement de méthodes taillées sur mesure pour l'écologie.

La réponse à cette situation consiste à former des écologues avec une inclination vers les mathématiques. Heureusement, ces derniers ont été nombreux à apparaître au cours des décennies récentes. Le résultat de leurs travaux est un immense développement de l'écologie statistique, la publication de quelques excellents manuels et la prise de conscience croissante de la responsabilité des écologues en ce qui a trait à la planification et à l'analyse appropriées de leurs recherches. Cette réalisation rend aussi la tâche plus facile aux enseignants.

Jusqu'aux premières années de ce millénaire, un ingrédient essentiel manquait toutefois pour que l'enseignement soit efficace et que la pratique de la statistique se généralise parmi les écologues : un ensemble de programmes standardisés disponible à tous et partout. Un cours de biostatistique ou d'écologie numérique sans exercices pratiques ne signifie rien. Un cours lié à un programme disponible dans le commerce est mieux, mais son application par le chercheur risque d'être limitée ou compromise si ce dernier perd l'accès à ce programme. De plus, les programmes commerciaux sont en général conçus pour un éventail d'utilisateurs plus large que les écologues et peuvent ne pas contenir les méthodes nécessaires à l'analyse des données écologiques. Le langage **R** a résolu ce problème, grâce au dévouement de nombreux chercheurs qui ont gratuitement créé et mis à disposition des paquets de fonctions complets, bien conçus et bien documentés. Désormais, l'enseignant n'a plus à dire : « voici comment fonctionne l'ACP... sur le papier ». Il peut maintenant dire : « voici comment fonctionne l'ACP. Je vous montre maintenant à l'écran comment en exécuter une, et dans quelques minutes vous serez en mesure d'exécuter la vôtre, et de le faire n'importe où dans le monde sur vos propres données ! »

Le langage **R** est conçu comme un environnement d'auto-apprentissage, une autre propriété fondamentale. Un livre sur **R** se doit donc d'adhérer à cette philosophie et de fournir le soutien nécessaire à toute personne souhaitant explorer le sujet par elle-même. Nous avons écrit ce livre pour fournir un pont entre la théorie et la pratique de l'écologie numérique, pont que tout le monde peut franchir.

Comme le langage **R** est une entité vivante, nous profitons de cette traduction pour mettre à jour le code fourni dans l'édition anglaise de 2018. Nous restons fidèles à la présentation adoptée dans cette édition, générée par R Markdown dans RStudio®, avec des couleurs différentes pour les fonctions, les objets, les arguments et les commentaires.

Notre souhait est que notre livre, déjà publié en anglais (deux éditions), en chinois mandarin (deux éditions) et en japonais, devienne un instrument de travail utile aux étudiants, aux enseignants et aux chercheurs francophones.

Par ailleurs, l'un des auteurs (PL) a produit un cours en ligne, en français, comprenant 18 enregistrements vidéo qui couvrent bon nombre des sujets exposés aux chapitres 3 ainsi que 5 à 8 de ce livre. Ce cours, intitulé *L'écologie numérique en 8 leçons*, est disponible sur le site du Centre de développement professionnel (PRAXIS) de l'Université de Montréal (<https://tinyurl.com/mvuputum>).

Vous pouvez aussi trouver la page Web du cours en cherchant l'expression *Écologie numérique en 8 leçons* à l'aide d'un fureteur Web. Le cours est également disponible en anglais sous le titre *Numerical Ecology in 8 lessons*, à la même adresse Web.



Ce livre présente l'écologie numérique, branche de l'écologie quantitative consacrée à l'analyse numérique des données écologiques dont celles portant sur les communautés naturelles. L'accent y est mis sur les méthodes multivariées, devenues incontournables depuis les années 1970. L'objectif principal est de fournir aux chercheurs les outils nécessaires pour choisir et appliquer, de façon logique, les traitements statistiques les mieux adaptés à leurs données. Fondé sur le langage **R**, désormais essentiel dans ce domaine, l'ouvrage adopte une approche résolument pratique, centrée sur l'utilisateur déjà familier avec les bases de ce langage.

Destiné aux étudiants universitaires, chercheurs, enseignants et professionnels en écologie, biologie, sciences de l'environnement, géographie, géomatique, biostatistique, sciences du climat et autres disciplines connexes, il permet d'approfondir la maîtrise des méthodes statistiques appliquées aux données écologiques. Une introduction claire aux notions fondamentales associées aux objets multivariés complète l'ensemble.

Daniel Borcard est chercheur retraité du Département de sciences biologiques de l'Université de Montréal. Spécialiste en écologie numérique, il s'intéresse particulièrement à l'analyse spatiale des communautés écologiques et au développement de méthodes statistiques appliquées aux données environnementales.

François Gillet est professeur émérite au laboratoire Chrono-environnement de l'Université Marie et Louis Pasteur à Besançon, en France. Ses travaux portent principalement sur l'écologie des communautés végétales, ainsi que la modélisation statistique et dynamique des systèmes écologiques.

Pierre Legendre est professeur émérite au Département de sciences biologiques de l'Université de Montréal. Figure éminente de l'écologie numérique, il est reconnu internationalement pour ses contributions à la biostatistique appliquée à l'écologie, notamment dans le domaine de l'analyse multivariée des données écologiques.

105 \$ • 80 €

Couverture : © François Gillet (2021)

www.pum.umontreal.ca

ISBN 978-2-7606-5423-5



9 782760 654235