N° d'ordre: 2005-264 Année 2005

THESE
présenté devant l'UNIVERSITE CLAUDE BERNARD - LYON I

pour l'obtention du
DIPLOME DE DOCTORAT
(*arrêté du 25 avril 2002)*

Soutenue le 14 décembre 2005

par

Clément CALENGE

Des outils statistiques pour l'analyse des semis de points dans l'espace écologique

Composition du jury

Directeurs :	Mme Anne-Béatrice DUFOUR
	M. Daniel MAILLARD
Rapporteurs :	M. Antoine GUISAN
	M. Patrick DUNCAN
Examinateurs :	M. Christian GAUTIER
	M. Francis LALOE

UMR CNRS 5558
Laboratoire de Biométrie et Biologie Evolutive
Université Claude Bernard LYON I - Bât G. Mendel
43, boulevard du 11 novembre 1918
69622 Villeurbanne

R

Je voudrais tout d'abord remercier tous les membres du jury, qui ont accepté de juger ce travail, et qui m'ont fait part de leurs remarques pertinentes lors de la soutenance.

Je remercie également l'Office national de la chasse et de la faune sauvage qui a financé ce travail.

Je voudrais exprimer toute ma gratitude à Anne-Béatrice Dufour, qui m'a souvent prodigué ses conseils dans de nombreux domaines (collaborations, rédaction d'articles, etc.), qui a souvent su apporter une structure à un esprit pas toujours très organisé, et qui m'a toujours soutenu lors de ces trois années. Merci pour ces nombreux conseils.

Je suis également extrêmement reconnaissant à Daniel Maillard qui m'a suivi et soutenu dans ma formation en tant que biométricien depuis 1998. Il m'a toujours encouragé à suivre cette voie, et c'est lui qui m'a convaincu de me lancer dans cette aventure.

Je remercie aussi Daniel Chessel pour m'avoir fait profiter de sa grande expérience et de ses connaissances en biométrie, lors des nombreuses discussions que nous avons pu avoir.

Je voudrais également exprimer toute ma gratitude aux membres des Conservatoires et Jardins Botaniques de Genève. Ils m'ont permis de prendre du recul vis-à-vis de ma position de biométricien. Je voudrais remercier en particulier Rodolphe Spichiger, Bastian Bise et Cyrille Chatelain, avec qui j'ai étroitement collaboré.

Je remercie également Gaëlle Darmon pour sa gentillesse et surtout sa patience lors de notre collaboration. Merci aussi pour m'avoir accueilli en Bauges et m'avoir montré que les chamois et les mouflons n'étaient pas que des points sur des cartes.

Je tiens également à remercier les nombreuses personnes qui m'ont permis d'améliorer directement ou indirectement la qualité des outils de la programmathèque adehabitat, et notamment Mathieu Basille, qui y a non seulement implémenté l'ENFA, mais qui a aussi relu toutes les fiches d'aides une à une, en m'indiquant plusieurs incohérences et erreurs dans les fonctions.

Je voudrais également remercier Paolo Cavallini, Sander Oom, et tous les membres de Faunalia (Italie), qui ont effectué un travail considérable pour la création et l'entretien du groupe de travail Animove.

Un grand merci à tous les membres du laboratoire de Biométrie et Biologie Evolutive pour m'avoir accueilli pendant ces trois années.

Je voudrais enfin exprimer toute ma gratitude à ma famille qui m'a soutenu au cours de mes études et m'a permis d'arriver jusque-là.

A Sandrine...

R´ ´

La mise en relation d'un ou de plusieurs semis de points avec des cartes de variables environnementales est centrale en Ecologie. Les points représentent en général des localisations d'individus d'une ou de plusieurs espèces. Cette thèse présente une démarche pour identifier les variables environnementales qui sont le plus structurantes de la distribution des points. Elle repose sur l'exploration des données dans l'espace géographique et dans l'espace défini par les variables environnementales (espace écologique). L'analyse d'un seul semis et celle de plusieurs semis de points sont considérées. Des collaborations ont permis de développer ou d'améliorer des outils d'analyse spatiale (analyse discriminante sur vecteurs propres du graphe de voisinage) ou écologique (ENFA, analyse K-select, analyse factorielle des rapports de sélection), reposant sur la théorie de l'analyse factorielle. Une bibliothèque de fonctions pour le logiciel R, adehabitat, a été programmée pour faciliter cette démarche d'analyse.

Mots-clés: adehabitat, autocorrélation, analyse discriminante, Biométrie, distribution spatiale, espace écologique, espace géographique, logiciel R, marginalité, niche écologique, processus de points, schéma de dualité, sélection de l'habitat, semis de points, spécialisation

A

Relating one or several point patterns with maps of environmental variables is at the very core of Ecology. The points are generally the locations of individuals belonging to one or several species. This thesis describes an analysis approach to identify the variables that most strongly affect the point distribution. We stress the importance of data exploration in the geographical space and in the space defined by the environmental variables (ecological space). The analysis of one pattern and that of several point patterns are considered. Collaborations have allowed the development or the improvement of tools for spatial (discriminant analysis of the eigenvectors of neighbourhood operator) or ecological analysis (ENFA, K-select analysis, factorial analysis of selection ratios). These methods rely on the theory of factorial analysis. A package of functions for the R software, adehabitat, has been programmed to facilitate this analysis approach.

Keywords: adehabitat, autocorrelation, discriminant analysis, biometry, spatial distribution, ecological space, geographical space, R software, marginality, ecological niche, point process, duality diagram, habitat selection, point pattern, specialisation

Table des matières

Résumé **II**

Abstract **III**

Introduction **1**

Chapitre 1 Problématique **4**

1.1 Problématique écologique 5

1.1.1 Importance de l'étude de la sélection de l'habitat en gestion de la faune sauvage 5

1.1.2 Problématiques similaires dans d'autres domaines de l'Ecologie 6

1.2 Les données 7

1.2.1 Echantillonnage spatial d'un ou plusieurs semis de points 7

1.2.2 Mesure de leur environnement 7

1.3 Un semis de points 8

1.3.1 Distribution des points dans l'espace géographique 9

1.3.2 Le modèle de la niche écologique 9

1.4 Plusieurs semis de points 10

1.5 Le logiciel R 10

1.5.1 Plan du mémoire 11

La démarche biométrique

Chapitre 2 L'étude de la sélection de l'habitat du point de vue du biologiste 14
2.1 Définition des concepts . . . 14
2.1.1 les concepts d'habitat et de niche . . . 14
2.1.1.1 le concept d'habitat . . . 14
2.1.1.2 Le concept de niche écologique . . . 15
2.1.1.3 Notre position dans ces débats . . . 16
2.1.2 La sélection de l'habitat . . . 17
2.2 Les mesures . . . 18
2.2.1 La mesure de l'utilisation de l'habitat . . . 18
2.2.2 La mesure de la préférence/qualité de l'habitat . . . 19
2.2.3 La mesure de la disponibilité de l'habitat . . . 20
2.3 Les analyses . . . 21
2.3.1 Les protocoles de type I . . . 22
2.3.1.1 Les prémisses : une seule variable qualitative d'habitat . . . 22
2.3.1.2 Plusieurs variables d'habitat . . . 24
2.3.2 Les suivis d'individus par balises radios (types II et III) . . . 26
2.4 Discussion . . . 27
2.4.1 Le problème de l'inférence . . . 28
2.4.2 La modélisation "aveugle" . . . 29
2.4.3 Les études exploratoires et les études confirmatoires . . . 30
2.4.4 Quelle est la meilleure méthode ? . . . 31

Chapitre 3 Structures spatiales et autocorrélation spatiale : vers une autre démarche 33
3.1 L'autocorrélation spatiale : un défaut des données ? . . . 33
3.1.1 Définitions de l'indépendance et de l'autocorrélation . . . 33
3.1.2 Deux types de structures spatiales dans les études de sélection de l'habitat 34
3.1.3 Un cas d'étude : la distribution des dégâts de sangliers sur vignobles . . . 35
3.1.4 l'autocorrélation spatiale : un nouveau paradigme . . . 37
3.2 Choix d'une démarche d'analyse . . . 38
3.2.1 La démarche biométrique . . . 38
3.2.1.1 Synthèse . . . 38
3.2.1.2 La construction d'un modèle . . . 39
3.3 Les outils de l'exploration . . . 40
3.3.1 Les outils mathématiques . . . 40

3.3.1.1 L'analyse dans l'espace géographique 40
3.3.1.2 L'analyse dans l'espace écologique 41
3.3.2 Les outils informatiques 41
3.3.2.1 les bibliothèques de fonctions existantes 42
3.3.2.2 La bibliothèque de fonctions **adehabitat** 42

Les outils de l'analyse dans l'espace géographique

Chapitre 4 L'analyse de semis de points d'un seul type **45**
4.1 Bases de l'analyse des semis de points 45
4.2 Analyse exploratoire des données 47
4.2.1 Outils graphiques communément utilisés 47
4.2.2 Méthode du noyau 47
4.2.2.1 Principe de la méthode 48
4.2.2.2 Utilisation en pratique 50
4.2.3 Mesure du caractère aléatoire du semis 51
4.2.3.1 Distance à l'événement le plus proche (fonction G) 52
4.2.3.2 Distance entre un point arbitraire et l'événement le plus proche (Fonction F) 53
4.2.4 Les tests de la CSR 53
4.3 Modélisation des processus de points 54
4.3.1 Définitions 54
4.3.2 La fonction $K(t)$ de Ripley 55
4.3.2.1 Présentation et propriétés 55
4.3.2.2 La fonction $K(t)$ comme méthode exploratoire 56
4.3.2.3 Utilisation en pratique 56
4.3.3 Quelques modèles utilisés 57
4.3.3.1 Le processus de Poisson 57
4.3.3.2 Le processus inhomogène de Poisson 58

4.3.3.3 Le processus de Neyman-Scott 59
4.4 Conclusion 60

Chapitre 5 L'analyse de plusieurs semis de points **61**
5.1 Nécessité de l'analyse multivariée 61
5.2 Principe mathématique de l'analyse discriminante 62
5.3 L'Analyse Factorielle des Correspondances 63
5.3.1 Principe mathématique 64
5.3.2 L'AFC est une analyse discriminante 64
5.4 Analyses discriminantes spatiales 65
5.4.1 L'utilisation des polynômes des coordonnées géographiques 65
5.4.2 L'utilisation des vecteurs propres de voisinage 66
5.5 Conclusion 69

Les outils de l'analyse dans l'espace écologique

Chapitre 6 L'analyse d'une niche écologique **72**
6.1 Les paramètres descriptifs de la niche écologique 72
6.2 Principe de l'analyse de la niche écologique 74
6.3 l'analyse factorielle de la niche écologique (ENFA) 77
6.3.1 Historique de l'analyse 77
6.3.2 Principe mathématique de l'analyse 77
6.3.3 Interprétation des résultats 81
6.4 Les cartes de qualité de l'habitat 81
6.4.1 L'algorithme BIOCLIM 83
6.4.2 L'algorithme DOMAIN 84
6.4.3 Cartographie reposant sur l'utilisation de l'ENFA 85
6.4.4 Les distances de Mahalanobis 86
6.4.4.1 Principe mathématique 86

6.4.4.2 Parenté avec l'ENFA 87
6.4.5 Autres méthodes 88
6.5 Conclusion 88

Chapitre 7 Plusieurs niches écologiques **92**
7.1 La discrimination 92
7.1.1 Introduction 92
7.1.2 L'analyse canonique des correspondances 93
7.1.2.1 Principe mathématique 93
7.1.2.2 Relations avec l'analyse discriminante 95
7.2 La sélection de l'habitat 95
7.2.1 L'analyse de la marginalité 96
7.2.1.1 Les protocoles de type II : l'analyse OMI 96
7.2.1.2 Extensions aux designs de type III : l'analyse K-select 98
7.2.2 L'analyse factorielle des rapports de sélection 99
7.2.2.1 Principe Mathématique 100
7.2.2.2 Relations avec l'AFC 101
7.2.2.3 Relations avec l'Analyse OMI 103
7.3 Conclusion 104

Une application de la démarche

Chapitre 8 La distribution des mouflons dans le massif des Bauges **107**
8.1 Problématique biologique 107
8.1.1 Cadre de l'étude : le programme herbivorie 107
8.1.2 L'étude de l'utilisation de l'espace par le mouflon 109
8.1.3 Les recensements du mouflon 109
8.2 Modélisation de la distribution des mouflons 110
8.2.1 Etablissement des données 110

8.2.2 Analyse de la structure sociale 111
8.2.3 Modélisation spatiale 114
8.2.3.1 Etude des causes de l'agrégation 114
8.2.3.2 Modélisation du processus 117
8.2.4 Analyse de la sélection de l'habitat 121
8.2.4.1 Mise en forme et exploration préliminaire 121
8.2.4.2 La sélection de l'habitat par le mouflon 122
8.2.5 Discussion 123
8.3 Conclusion 124

Chapitre 9 Discussion **128**
9.1 Importance de l'approche systémique 128
9.2 L'outil informatique 129
9.3 La consultation 130
9.4 Perspectives : Les données GPS 130
9.5 Apports en termes de gestion de la faune sauvage 132

Conclusion **134**

Références **136**

Introduction

Ces trois années de thèse m'ont permis d'apprendre la démarche biométrique. J'avais initialement une formation de biologiste (Maîtrise de Biologie des Populations et des Ecosystèmes), complétée ensuite par un DESS de Biostatistique. J'ai appris le métier d'ingénieur en Biostatistique, métier qui exige des compétences techniques en Mathématique et en Informatique pour pouvoir appliquer, de façon presque répétitive, un certain nombre de méthodes statistiques "standards" afin de valoriser les données des biologistes. Six mois de stages, puis deux mois de vacations, suivis de 2 ans d'objection de conscience à l'Office national de la chasse et de la faune sauvage (ONCFS) m'ont permis d'adapter ces compétences techniques aux questions posées dans le domaine de la gestion de la faune sauvage. Le temps passé dans cet organisme s'est concrétisé par la rédaction de plusieurs articles dans des revues scientifiques (M *et al.* 2001, C *et al.* 2002a, M *et al.* 2002, C *et al.* 2002b, 2003, 2004, J *et al.* 2004, C *et al.* 2005b, B *et al.* 2005).

Mais lorsque s'est posée la question de la sélection de l'habitat par le sanglier en milieu méditerranéen, je me suis trouvé face à un problème pour lequel les solutions techniques étaient à ma connaissance inexistantes dans la littérature scientifique. L'analyse des données de radiopistage collectées sur le sanglier posait la question suivante : Comment déterminer ce qu'il peut y avoir de commun ou de différent dans la sélection des meilleurs types d'habitats par plusieurs animaux *au sein de leur domaine vital*, lorsque l'habitat est décrit par plusieurs variables environnementales ? C'est cette question qui a constitué la motivation principale de la thèse. J'ai donc amorcé ce travail avec pour objectif principal d'affiner un savoir-faire technique dans une équipe dont la réputation de solidité scientifique n'était plus à faire. Mon objectif était ambitieux : il s'agissait de développer une démarche statistique permettant l'analyse de la sélection de l'habitat par la faune sauvage, de façon à disposer *en toutes circonstances, quelles que soient les données*, d'une solution technique qui permette leur analyse, et qui renvoie des résultats immédiatement interprétables par le biologiste (objectif décrit dans C 2002). La mise en évidence des interactions entre la faune sauvage et ses habitats étant un problème situé au cœur même de l'Ecologie, on peut mesurer l'importance de l'enjeu.

J'ai donc commencé cette thèse avec cette vision "technique" de la Biométrie, en ayant deux objectifs prioritaires : (i) développer une méthode qui permette de répondre à la question qui avait motivé cette thèse, et (ii) apporter aux biologistes des outils informatiques permettant de mettre en œuvre des techniques déjà existantes, grâce à la programmation d'une bibliothèque de fonctions pour le logiciel R. Le développement de l'analyse K-select (C *et al.* 2005a) m'a permis d'atteindre le premier objectif. J'ai ensuite programmé la bibliothèque **adehabitat**

pour atteindre le second (C soumis). Cette bibliothèque est aujourd'hui disponible sur le réseau qui distribue le logiciel R.

Mais c'est au cours de ma deuxième année de thèse que j'ai vraiment pris conscience de l'écart pouvant exister entre le métier d'ingénieur et celui de chercheur. Daniel C (Laboratoire de Biométrie, Université Lyon 1) m'a en effet mis en relation avec les Conservatoires et Jardins Botaniques de Genève (CJB), afin d'initier une collaboration dont le but était d'analyser la distribution de plusieurs espèces arborées paraguayennes à l'échelle continentale, en se basant sur une liste d'occurrences d'espèces établie d'après les localisations de prélèvements des échantillons des herbiers du Missouri et de Genève. Les données, à première vue, semblaient en effet présenter une certaine parenté avec les données de radio-pistage. Dans les deux cas, on dispose de semis de points de différents types (animaux ou espèces) distribués dans l'espace géographique, et l'on cherche à caractériser cette distribution (distribution des domaines vitaux ou structuration de la composition végétale). Dans le cadre de cette collaboration, la mise en évidence de cette distribution devait permettre de tirer des conclusions sur le mode d'organisation de la végétation au Paraguay, et d'aider les biologistes à construire un modèle décrivant l'évolution de cette organisation depuis la dernière grande glaciation, au Pléistocène.

Lorsque je travaillais pour l'ONCFS, ma connaissance de la bibliographie de l'espèce étudiée et un contact quotidien avec les personnes chargées de sa gestion me permettaient de savoir *a priori* les résultats que j'étais censé trouver, et ce savoir guidait les analyses indépendamment des données, une stratégie par ailleurs courante en biologie, comme l'indiquent T *et al.* (1993). En revanche, le domaine de la Biogéographie sud-américaine m'était complètement étranger, et je ne possédais aucune connaissance de la phytosociologie paraguayenne sur laquelle m'appuyer pour tirer ces conclusions biologiques. Mon savoir-faire technique ne m'était pas d'une grande utilité face à ces données radicalement nouvelles que sont les données d'herbiers, qui soulèvent des problèmes originaux (pression d'échantillonnage non uniforme sur la zone d'étude, recueil des données sur plus d'un siècle, etc.). Il m'est ainsi apparu qu'un maçon habile n'est pas nécessairement un bon architecte.

C'est grâce à de longues discussions avec les biologistes des CJB, à la fois de vive voix et par messages électroniques, et grâce à des analyses exploratoires, que la base de données a pu être établie (suppression de certaines espèces trop rares, de nombreuses données d'origine incertaine, de sources de données dont la validité était remise en question, etc.). C'est au fur et à mesure de ces analyses préliminaires que le problème, initialement posé en termes biologiques, a pu être traduit sous la forme de questions statistiques et résolu par *les* analyses. Cette collaboration s'est traduite par la rédaction d'un article identifiant les grandes structures végétales en Amérique du Sud (S *et al.* 2004) et par un autre article en deux parties soulignant le caractère complètement original des données d'herbier, et la nécessité des analyses exploratoires pour comprendre leur structure. La première partie, purement méthodologique, illustre les liens existant entre trois méthodes permettant ce type d'analyse, dont une développée dans le cadre de cette thèse, l'analyse discriminante sur vecteurs propres de voisinage (C *et al.* 2006). La seconde partie illustre la démarche d'analyse utilisée pour analyser les occurrences d'espèces issues d'herbier (S *et al.* 2006b). Une synthèse de ces trois articles a été présentée au colloque d'Edimbourg en septembre 2004, et a fait l'objet d'un chapitre dans un livre traitant

des savanes tropicales (S *et al. in press*). Enfin, un article méthodologique introduisant la méthode du noyau comme outil exploratoire des semis de points en Biogéographie a été rédigé et soumis à *Applied Vegetation Science* (C *et al.* soumis).

Cette collaboration m'a permis de prendre conscience de la nécessité de définir une position sur le plan scientifique. La Biométrie n'est pas une simple étape technique, c'est un véritable métier, comme je le montrerai dans ce mémoire. J'ai, parallèlement au travail précédent, initié une autre collaboration avec Gaëlle Darmon (Laboratoire de Biométrie, Université de Lyon). Cette collaboration a pour objectif de mettre en évidence les interactions entre deux espèces d'ongulés de montagne : le chamois et le mouflon. A l'heure de la rédaction de ce mémoire, le travail est encore en cours, et la soumission prochaine d'un article sur le sujet est prévue. La collaboration avec les CJB m'a permis d'aborder le problème avec une toute autre approche que celle que j'aurais utilisée à mon entrée en thèse (approche que j'avais d'ailleurs en premier lieu abordée, mais que je ne présente pas ici, pour des raisons que le lecteur comprendra).

Les nombreuses collaborations effectuées dans le cadre de cette thèse m'ont amené à utiliser des outils relevant de différents champs de la Statistique (processus de points, schéma de dualité, modéle linéaire généralisé) et de l'Ecologie (gestion de la faune sauvage, Biogéographie). Ce mémoire est organisé de façon à présenter ces outils dans le cadre de la démarche biométrique que je me suis appropriée. Celle-ci est avant tout *scientifique*, et non plus simplement technique. Mes recherches prennent donc place dans un contexte interdisciplinaire incluant l'Ecologie, les Mathématiques et l'Informatique, contexte qui fait l'objet de la partie suivante.

Chapitre 1

Problématique

Les nombreux facteurs qui influencent l'utilisation de l'espace par la faune sauvage peuvent être classés en deux grandes catégories. La première comprend ceux qui traduisent les interactions avec les autres animaux, tels que la prédation et la compétition inter ou intra-spécifique. La seconde correspond à la combinaison des caractéristiques de l'environnement (climat, altitude, composition de la végétation, etc.) qui permet l'occupation d'une zone, incluant la survie et la reproduction des animaux. Réussir à isoler l'effet particulier de l'environnement est un enjeu majeur aujourd'hui dans de nombreux domaines de l'Ecologie, car cette influence reflète les exigences écologiques de l'espèce étudiée.

Les études visant à expliquer la distribution d'êtres vivants sur une zone par les structures environnementales sont donc très fréquentes dans la littérature scientifique. Elles peuvent être conduites à plusieurs échelles (J 1980) : il peut s'agir de mettre en évidence les facteurs bio-climatiques qui influencent l'aire de répartition d'une espèce (e.g. H *et al.* 2002), de déterminer les variables qui affectent les variations de densité d'une population d'animaux sur une zone (e.g. M *et al.* 2002), ou d'isoler les caractéristiques de l'environnement activement recherchées par un animal au sein de son domaine vital (e.g. A *et al.* 1993).

Des données sont alors recueillies par les biologistes afin de répondre à ces questions. Les moyens mis en œuvre peuvent être très variés en fonction de l'échelle à laquelle l'étude est effectuée. Par exemple, la distribution des animaux d'une espèce sur une zone de quelques centaines d'hectares peut être échantillonnée grâce au parcours par des observateurs de transects placés aléatoirement sur la zone ; l'ensemble des localisations des animaux détectés lors de ces opérations constitue un semis de points qui reflète les variations de densité de la population sur la zone (D 1983). Si l'échelle de travail est plus fine, par exemple si l'objectif est de mettre en évidence les zones dans lesquelles un animal recherche sa nourriture, d'autres techniques telles que le radio-pistage peuvent être utilisées ; l'ensemble des localisations des animaux suivis constitue un ou plusieurs semis de points qui reflète la distribution du temps passé par les animaux dans les différentes zones qui leur sont accessibles. Ces deux exemples ne sont qu'un maigre échantillon de la très vaste diversité des méthodes utilisables par les biologistes pour étudier l'utilisation de l'espace par la faune sauvage.

Une fois ces données collectées, c'est l'analyse statistique qui met en évidence l'influence

particulière de l'environnement sur la distribution des animaux. Notre but est précisément de développer une approche statistique permettant de caractériser la distribution spatiale d'un ou plusieurs ensembles d'organismes à partir d'un ou plusieurs semis de points, puis de mettre en relation cette distribution avec les caractéristiques de l'environnement afin d'en déterminer les variables les plus structurantes. Nous concentrerons notre attention sur la possibilité de dissocier les effets de l'environnement des autres influences pouvant affecter cette distribution.

Ce chapitre décrit le cadre théorique dans lequel cette thèse prend place. L'accent est mis sur ses implications en termes de problématique écologique, et notamment dans le domaine de la gestion de la faune sauvage. Nous décrivons également le type de données qui est au centre de ce mémoire, c'est-à-dire les semis de points. Enfin, nous introduisons la démarche choisie pour les analyser.

1.1 P ´ ´

1.1.1 Importance de l'étude de la sélection de l'habitat en gestion de la faune sauvage

Le but de la gestion de la faune sauvage est de permettre la coexistence entre l'Homme et la faune sauvage, pour le bénéfice des deux, et en minimisant les problèmes pouvant survenir entre eux (e.g. les dégâts qu'occasionne la faune sur les cultures). La gestion durable de la faune atteint ces objectifs par le contrôle ou la manipulation des populations d'animaux sauvages et de leurs habitats.

Elle nécessite la connaissance des facteurs pouvant influencer le fonctionnement démographique de ces populations, ainsi que des échelles spatiales pertinentes auxquelles ils doivent être étudiés. Les mécanismes de régulation des populations ne sont par exemple pas forcément les mêmes à des échelles fines, avec l'existence possible de sous-populations hétérogènes démographiquement, qu'à des échelles plus vastes. Identifier à quels niveaux de fonctionnement se déroulent les différents processus à même d'entraîner des variations démographiques est notamment essentiel pour la mise en place des schémas départementaux de gestion de la faune sauvage.

Répondre à ces besoins de gestion et de conservation des populations naturelles nécessite en outre d'établir des modèles prédisant l'impact des modifications environnementales susceptibles d'intervenir à différentes échelles sur la dynamique des populations. Pendant longtemps, les écologues ont considéré l'habitat comme homogène et ont étudié la dynamique des populations à travers le temps avec peu d'attention pour sa composante spatiale. Or pour un même effectif, cette dynamique sera différente selon sa distribution (variabilité de la nourriture, interférence entre les individus, etc..). L'hétérogénéité spatiale modifie les modalités de mouvements des animaux, leurs taux de dispersion, et crée des habitats de différentes qualités influençant ainsi la dynamique des populations. La plupart des individus ne se distribuent pas aléatoirement dans l'espace et les populations présentent ainsi une structuration spatiale bien définie. Analyser la dynamique spatiale de ces populations nécessite d'identifier précisément le type de structure

émergeant et les échelles auxquelles se placer pour l'étudier. Une fois ces structures définies, il est alors nécessaire de mesurer les flux qui les caractérisent pour pouvoir ensuite modéliser la dynamique du système et proposer en fonction des objectifs définis les mesures de gestion adaptées.

Son principal outil pour contrôler le niveau de la population d'animaux est la chasse. Ainsi, les moyens d'action les plus courants du gestionnaire sont la définition de plans de chasse quantitatifs (quotas d'individus à prélever) et qualitatifs (sexe et âge des individus à prélever), de la date des périodes de chasse (en fonction de la biologie de l'espèce). Des mesures d'accompagnement peuvent aussi être mises en place comme des méthodes de protection des cultures (C *et al.* 2004). Si l'espèce est menacée d'extinction dans une zone donnée, une interdiction de tir pourra être décidée, éventuellement conjuguée à des opérations de réintroductions.

Mais si la gestion de la faune sauvage s'effectue directement sur la population, elle doit aussi agir indirectement sur les habitats qu'elle utilise. Ainsi, la définition d'une politique de gestion efficace pour une population implique de pouvoir comprendre la nature des interactions que les individus qui la composent peuvent entretenir avec leur environnement. M (2001) souligne "*il serait illusoire d'espérer parvenir à un bon équilibre entre la faune et la flore en agissant uniquement par le plan de chasse sur la gestion quantitative et qualitative de l'espèce. Le forestier, responsable de la gestion d'un écosystème dans sa globalité, doit également intervenir pour favoriser la capacité du milieu à accueillir les espèces animales, qui doivent pouvoir y trouver facilement abri et nourriture.*". La nouvelle législation sur la chasse va dans ce sens avec la création des schémas départementaux de gestion cynégétique (chapitre V de la loi du 23 février 2005 sur le développement des territoires ruraux). L'élaboration de ces schémas, confiée aux Fédérations Départementales des chasseurs (article L 425-4 de la section 2), a entre autres choses pour mission de définir les aires dont les caractéristiques naturelles apparaissent répondre le mieux aux exigences des espèces et évaluer les conséquences de la présence de ces espèces vis-à-vis des contraintes humaines (agriculture, forêt, tourisme).

Il devient alors essentiel de connaître les caractéristiques de l'environnement qui sont nécessaires à – ou recherchées par – les individus de la population gérée, et une bonne gestion ne pourra s'effectuer que si, parallèlement, des études sont effectuées pour déterminer ces besoins.

Un certain nombre d'outils permettent au gestionnaire de suivre l'évolution de l'impact de ces mesures de gestion sur l'équilibre du système animal-environnement. Par exemple sur les ongulés le suivi annuel d'un faisceau d'indicateurs comme l'indice kilométrique (V *et al.* 1991), le poids des individus (M *et al.* 1989) et l'indice de consommation (M *et al.* 2001) permettent d'établir annuellement un diagnostic de l'évolution de la population et du milieu qu'elle utilise. Ce diagnostic permet de faire évoluer les règles de gestion pour atteindre les objectifs fixés au préalable.

1.1.2 Problématiques similaires dans d'autres domaines de l'Ecologie

Des problèmes similaires peuvent être trouvés dans tous les domaines de l'Ecologie. Par exemple, l'Ecologie des communautés a pour objectif l'étude de la distribution, de l'abondance,

de la démographie et des interactions entre les populations de deux espèces coexistantes ou plus. L'ensemble des espèces étudiées - la communauté - est au centre de l'étude des scientifiques de cette discipline, et un de ses principaux buts est de mettre en évidence les variables environnementales qui structurent le plus ces communautés (T B 1986).

La Biogéographie est un autre exemple de domaine dans lequel l'étude de l'influence de l'environnement sur la distribution des individus est centrale. Cette discipline a en effet pour objectif d'identifier la distribution actuelle des espèces animales et végétales, et de déterminer l'histoire des variations de leur aire de distribution. En connaissant les besoins écologiques des espèces à grande échelle, le biologiste peut tirer un certain nombre de conclusions du plus grand intérêt. Par exemple, si l'histoire de la colonisation d'une zone par une espèce est connue, il est alors possible d'inférer sur les variations passées du climat (S *et al.* 2004). De même, la construction d'atlas biogéographiques repose sur ces connaissances (H 1995). En déterminant les exigences écologiques d'une espèce, il est possible d'estimer les cartes de sa distribution potentielle. Dans les zones où la présence de l'espèce est prédite mais non observée, déterminer les raisons de son absence (mauvais échantillonnage, présence d'espèces compétitrices, etc.) peut être d'un intérêt majeur pour sa conservation (M *et al.* 1992), ou pour déduire l'histoire paléoécologique de la zone étudiée (S *et al.* 1995).

Ces deux exemples montrent l'importance de la mise en relation de la distribution d'un ou plusieurs organismes avec la distribution de variables environnementales sur une zone. Cela permet de déterminer les facteurs qui ont le plus d'influence sur la distribution de ces organismes, pour ensuite pouvoir la prédire dans d'autres zones ou pour prédire les changements susceptibles de survenir après des modifications du milieu (embroussaillement, coupe forestière, construction d'infrastructures humaines, changements climatiques, etc.).

Pour atteindre cet objectif, le biologiste doit tout d'abord établir un protocole d'échantillonnage de la distribution des individus sur la zone d'étude. Comme nous l'avons indiqué en introduction de ce chapitre, les moyens mis en œuvre pour cet échantillonnage peuvent être très variés en fonction de l'échelle à laquelle l'influence de l'environnement est étudiée. En conséquence, les données recueillies présentent également une très grande diversité.

1.2 L ´

1.2.1 Echantillonnage spatial d'un ou plusieurs semis de points

La très vaste diversité d'informations pouvant être recueillies pour mettre en évidence l'influence de l'environnement sur la distribution des individus peut être organisée, en fonction de l'échelle de l'étude, en deux grandes classes de données : les données populationnelles et les données individuelles.

– **Les données populationnelles**. Ce sont des échantillonnages ou des recensements d'une population d'individus sur une zone *à un instant t*. Les variations d'intensité d'utilisation des différentes zones étudiées sont mesurées par les variations de *densité* des animaux sur

la zone (M *et al.* 2002). La collecte des données implique par exemple le parcours de transects, des échantillonnages par quadrats, etc. Chaque individu détecté est alors localisé dans l'espace à l'aide de ses coordonnées géographiques, et les attributs qui le caractérisent (espèce, sexe, âge, etc.) sont également relevés. Les données correspondent alors à un semis de points distribué sur une zone, éventuellement discrétisé sur une grille de quadrats.

– **Les données individuelles**. L'étude se focalise sur une composante du système précédent : l'individu. Le travail est réalisé à une échelle spatiale plus fine, et corrélativement à une échelle temporelle plus longue. La mesure de l'intensité d'utilisation des différentes zones correspond à *la proportion du temps passé par l'animal dans les différentes zones qui lui sont accessibles*. Les outils utilisés pour ce type d'étude sont variés, mais le radiopistage reste la technique la plus utilisée (S 1994) : des animaux sont capturés, puis équipés de colliers émetteurs qui permettent leur suivi. Les données sont alors constituées d'un semis de points par animal reflétant la distribution spatiale de son activité.

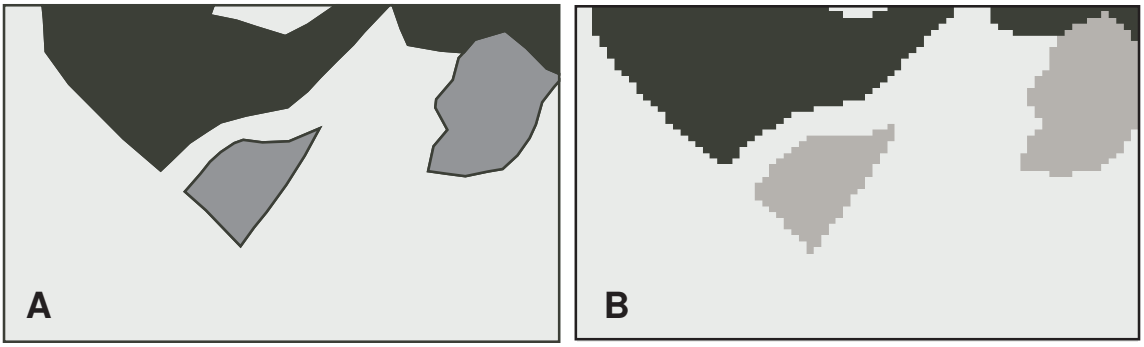
Qu'il s'agisse de données populationnelles ou de données individuelles, nous nous intéresserons toujours, au cours de notre travail, à des "*semis de points*".

1.2.2 Mesure de leur environnement

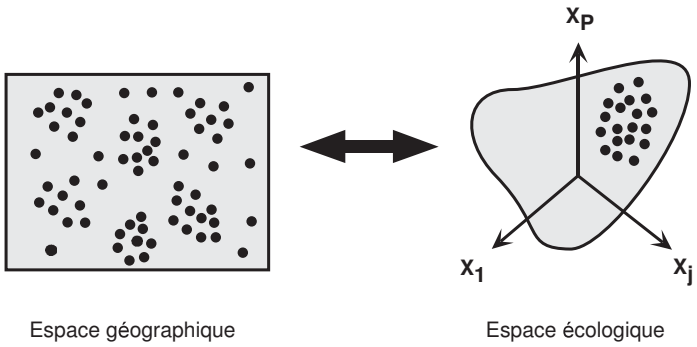
La deuxième étape indispensable à ce type d'étude est la mesure des variables de l'environnement. Les points du semis étant référencés dans l'espace, il est nécessaire de disposer de cartes géoréférencées des variables d'environnement (altitude, végétation, etc.) avec lesquelles ils peuvent être mis en relation.

Le développement récent des Systèmes d'Information Géographiques (SIG) a facilité cette association entre les deux types de données. D (2003) définit un SIG comme "*un système informatisé (matériels et logiciels) capable de stocker, gérer, manipuler, analyser, modéliser, représenter des données à références spatiales*". Les premiers SIG ont été développés dans les années 1970, mais l'expansion de leur utilisation en Ecologie est beaucoup plus récente. C'est seulement depuis le début des années 1990 que ces logiciels sont accessibles aux biologistes.

L'information spatiale peut être représentée selon deux grands modes (D 2003) :
– **le mode vecteur** (figure 1A) : il s'agit d'un système de représentation orienté objet. Les cartes vecteurs stockent de l'information concernant des objets dont les coordonnées sont définies précisément. Ces objets peuvent être des points (e.g. localisations d'animaux), des polylignes (e.g. fleuves, routes) ou des polygones (patchs d'habitat). A chaque objet stocké sous forme vecteur est associé un certain nombre d'attributs (surface, périmètre, type d'habitat rencontré dans le patch, etc.).
– **le mode raster** (figure 1B) : il s'agit d'un système de représentation orienté image. La zone cartographiée est découpée selon une grille de pixels, et la variable cartographiée prend une valeur pour chacun de ces pixels.



F . 1 – *Les deux grands modes de stockage de l'information spatiale. Ici la même carte est représentée : (A) en mode vecteur ; (B) en mode raster.*



F . 2 – *Le couplage entre espace géographique et espace écologique. Chaque occurrence de l'espèce étudiée possède des coordonnées dans l'espace géographique (e.g. longitude et latitude) et des coordonnées dans l'espace écologique : un espace multidimensionnel défini par les P variables environnementales X_j mesurées à chaque occurrence (ici, seulement trois variables sont représentées). Dans les deux cas, la zone grisée définit l'espace disponible aux individus.*

Comme nous le montrons par la suite, les structures de l'environnement présentes sur une même zone sont en général décrites par *plusieurs* variables (altitude, type de végétation, distance aux points d'eau, etc.). Pour cette raison, le mode raster est souvent préféré au mode vecteur. En effet, l'unité de base de cartographie - le pixel - est identique pour toutes les variables cartographiées, ce qui facilite grandement les analyses (H et G 2002).

1.3 U

Quel que soit l'objet biologique au centre de l'étude (la population ou l'individu), les écologues disposent de deux types de données qu'ils tentent de mettre en relation : (i) le ou les semis de points représentant son utilisation de l'espace, et (ii) les cartes de variables environnementales qui permettent de déterminer les besoins écologiques de cet objet (figure 2).

1.3.1 Distribution des points dans l'espace géographique

Il est indispensable de commencer par étudier la distribution des points *dans l'espace géographique*, afin de déterminer, en collaboration avec le biologiste, quelle est la nature des facteurs qui déterminent la distribution étudiée. Les contraintes liées à l'environnement jouent bien sûr un rôle important dans cette distribution. Ainsi, une espèce arboricole aura probablement la forêt comme habitat de prédilection. Mais d'autres facteurs peuvent expliquer cette distribution. Ainsi, des facteurs sociaux peuvent expliquer un regroupement des individus en "clusters" sur une zone. Inversement, une territorialité des individus peut aboutir à une distribution régulière de la population sur la zone. Des raisons historiques peuvent expliquer l'absence de l'espèce dans des milieux qui lui sont favorables (e.g. cas du chevreuil en milieu méditerranéen, C - et C 1992).

Bien que l'objectif principal de ce mémoire soit de mettre en évidence les variables environnementales qui affectent la distribution spatiale des espèces ou des individus, il devient évident qu'on ne peut dissocier l'aspect environnemental des autres aspects. Or l'espace géographique est constitué par un nombre bien plus faible de dimensions (seulement deux) que l'espace défini par les variables environnementales. Il est donc beaucoup plus aisé de déterminer quels peuvent être ces aspects par l'exploration des semis de points dans l'espace géographique, et par la modélisation des processus spatiaux qui les ont générés. De nombreux outils sont disponibles dans le domaine de l'analyse des processus de points pour prendre en compte le caractère spécifiquement spatial de ce type de données (D 1983, C 1991, B et G 1995). La plupart d'entre eux ont été développés dans le domaine de la foresterie, afin de modéliser la distribution d'arbres sur une zone (H 1995, S et P 2000, D *et al.* 2001). Ceux-ci sont en revanche plus rarement utilisés dans le domaine de l'étude de la faune sauvage, mais peuvent être d'une grande utilité pour ce type d'analyse (e.g. K 2001).

1.3.2 Le modèle de la niche écologique

A chaque point de l'espace géographique peut être associé un certain nombre de mesures environnementales (pente, type de végétation, etc.). Chacune de ces variables environnementales définit une dimension dans un espace multidimensionnel appelé *espace écologique* (figure 2). La distribution de l'espèce dans cet espace écologique va permettre de déterminer ses besoins ; cet hypervolume dans lequel l'espèce peut maintenir une population viable est appelé *niche écologique* de l'espèce.

Ce concept de niche écologique, formalisé par H (1957), a fourni aux écologues un excellent modèle pour l'étude des relations entre une espèce et son environnement (M - *et al.* 1992). Dans la mesure où la niche est définie dans un espace multivarié, les méthodes statistiques utilisées pour l'étudier doivent impérativement prendre en compte cet aspect des données (e.g G 1971). Ainsi, les méthodes d'analyse factorielle, telles que l'Analyse en Composantes Principales, l'Analyse Factorielle des Correspondances, l'Analyse Discriminante ou l'Analyse Canonique des Correspondances ont rencontré un très large succès chez les écologues. Ces techniques permettent de déterminer les variables qui peuvent avoir une influence sur la forme de la niche et sa position dans l'espace écologique. Les méthodes permettant l'in-

férence statistique sur la niche sont en revanche plus complexes à mettre en œuvre, tels que les modèles linéaires généralisés ou les modèles additifs généralisés, mais sont de plus en plus utilisés aujourd'hui (L 1998, P et F 2001, A 2002, C *et al.* 2002, H 2002, L *et al.* 2002a, Z *et al.* 2002).

1.4 P

De nombreuses études en Ecologie se focalisent sur plusieurs types d'objets biologiques (individus, populations ou espèces), et l'échantillonnage résulte donc en plusieurs semis de points distribués sur la zone d'étude. Il peut par exemple s'agir de la distribution spatiale de plusieurs espèces végétales ou animales (J 1971, H 1991, B et L 1994, J et G 2001), ou de la distribution du temps passé par plusieurs animaux suivis par radiopistage sur une zone donnée (P et P 1984, W et M 1997, R *et al.* 2000).

Comme pour le cas d'un seul semis de points, il est important d'étudier la distribution des semis dans les espaces géographique et écologique. Les questions posées par les biologistes dans ce cas de figure sont en revanche différentes. Il peut par exemple s'agir de discriminer au maximum les types de points, c'est-à-dire de déterminer leurs différences du point de vue écologique ou géographique. Il est difficile de fournir une liste exhaustive des questions posées dans ce type d'étude, étant donnée la variabilité des problématiques. Mais encore une fois, ce sont les méthodes d'analyse factorielles qui sont le plus fréquemment utilisées (e.g. H 1991, J ¨ et S 1998, J et G 2001).

1.5 L R

Du fait de la variabilité des questions posées et des données récoltées par les biologistes dans les études sur l'influence de l'environnement, le responsable de l'analyse doit être capable de combiner des méthodes statistiques très diverses, permettant à la fois la prise en compte des caractères spatial et écologique des données. La démarche biométrique suppose alors l'utilisation d'outils informatiques adaptés. Le développement récent du logiciel R (H et L 2001) répond à ces attentes.

R est une implémentation d'un langage de programmation interactif appelé S, développé aux laboratoires AT&T Bell Laboratories par John C et ses collègues au milieu des années 1970. Grâce à ce langage, l'utilisateur peut facilement construire ses propres analyses. Ross I et Robert G (université d'Auckland) ont développé le logiciel R à partir de ce langage de programmation, et l'ont rendu libre en 1995 (H et L 2001). Les sources de ce logiciel peuvent donc être téléchargées sur Internet et modifiées par tous les utilisateurs. R est donc au centre d'un environnement de développement coopératif très actif appelé *Comprehensive R Archive Network* (CRAN), qui comprend de nombreux sites miroirs sur Internet.

L'utilisateur peut aborder le langage S sans la moindre compétence en programmation, car ce logiciel est conçu de telle manière à l'encourager à glisser sur la pente de la programmation,

sans même qu'il s'en rende compte (C 1998). Ainsi, quand il devient expérimenté, il programme plus facilement les fonctions dont il a besoin pour analyser ses données, et encore un peu plus d'expérience lui permet d'organiser ses fonctions en *bibliothèques de fonctions*. Ces bibliothèques peuvent ensuite être soumises à CRAN, afin de les rendre disponibles aux autres utilisateurs de R.

Ainsi, de nombreuses bibliothèques de fonctions ont été développées pour fournir des outils spécifiques dans de nombreux domaines de l'analyse statistique (séries temporelles, statistiques circulaires, réseaux de neurones, analyse des processus de points, etc.), mais aussi des méthodes utilisées dans d'autres domaines scientifiques (biologie moléculaire, épidémiologie, psychométrie, finance, géographie, géologie, etc.). En Ecologie, quelques bibliothèques de fonctions ont été développées, et notamment la bibliothèque **ade4**, qui permet l'analyse de données écologiques et environnementales dans le contexte des méthodes euclidiennes exploratoires (C *et al.* 2004). C'est également en suivant cette stratégie que j'ai pu programmer la bibliothèque de fonctions **adehabitat**, qui permet l'utilisation des outils développés dans le cadre de cette thèse (§ 3.3.2.2).

La très large diversité d'analyses disponibles, l'environnement de développement, la gratuité du programme et la facilité d'apprentissage du langage de programmation font de R l'outil biométrique par excellence.

1.5.1 Plan du mémoire

Compte tenu de tout ce qui précède, quatre grandes parties ont été organisées. La première a pour objectif de présenter une réflexion sur la démarche biométrique construire au fur et à mesure de ces trois années d'étude. Elle présente le point de vue du biologiste sur les méthodes d'étude de la sélection de l'habitat. Nous y discutons brièvement des polémiques autour des concepts biologiques centraux dans ce type d'étude, des méthodes utilisées pour récolter les données, des outils développés pour les analyser, et de la marche à suivre pour utiliser ces outils (chapitre 2). Le problème de l'autocorrélation spatiale, très discuté dans la littérature, est également présenté (chapitre 3). Les structures spatiales d'une population sur une zone peuvent être causées par l'environnement ou être intrinsèques à la population. Cette caractéristique fait de l'autocorrélation spatiale un nouveau paradigme, c'est-à-dire un objet d'étude en soi. A partir de cette discussion, nous essayons de construire une démarche d'étude de la distribution spatiale de points dans l'espace écologique et dans l'espace géographique.

La seconde et la troisième parties présentent certains des outils mathématiques et informatiques pouvant être utilisés pour mener à bien les études de la distribution d'un ou plusieurs semis de points dans les espaces écologiques et géographiques. Comme il est impossible de donner une liste exhaustive de tous ces outils, ces parties se concentrent uniquement sur ceux que j'ai utilisés ou développés au cours des collaborations menées dans le cadre de cette thèse.

La seconde partie se focalise sur l'analyse d'un ou plusieurs semis de points dans l'espace géographique. Elle présente des outils pour l'exploration et la modélisation des processus d'un seul type de points dans l'espace géographique (chapitre 4), ainsi que les méthodes explora-

toires permettant la discrimination spatiale de plusieurs catégories de points (chapitre 5).

La troisième partie présente des outils exploratoires de la niche écologique. Le cas de l'analyse d'une seule niche écologique est étudié en premier lieu, ainsi que les algorithmes permettant la cartographie de la qualité des habitats (chapitre 6). Lorsque plusieurs niches écologiques sont étudiées, deux aspects sont essentiels : la discrimination entre les niches, et la sélection de l'habitat (chapitre 7). Cette partie joue un rôle charnière dans ce mémoire. Son but principal est de montrer les étroites relations de parenté entre les différentes méthodes d'analyse factorielles présentées dans la seconde et la troisième parties, et la très grande cohérence du modèle sur lequel la plupart d'entre elles reposent : le schéma de dualité.

La dernière partie, qui ne contient qu'un chapitre (chapitre 8), présente une application de la démarche développée dans la première partie à l'aide des outils présentés dans les seconde et troisième parties. L'intérêt principal de ce chapitre est qu'il présente en détail tout le travail d'analyse de la distribution dans l'espace géographique et écologique des mouflons dans le massif des Bauges. Le but de cette partie est de montrer que chaque jeu de données possède des caractéristiques particulières, qui appellent nécessairement la construction d'une analyse particulière.

La démarche biométrique

Chapitre 2

L'étude de la sélection de l'habitat du point de vue du biologiste

Ce chapitre constitue une étude bibliographique des méthodes ordinairement utilisées pour analyser la sélection de l'habitat par la faune sauvage. Nous montrons que les définitions des concepts biologiques utilisés dans ce type d'étude ne sont pas toujours très claires, et que le problème de leur définition est au cœur de violentes polémiques entre écoles de pensées. Puis, nous nous concentrons sur un domaine en particulier, la gestion de la faune sauvage, et nous effectuons une revue des méthodes de statistiques ordinairement utilisées dans ce domaine pour mettre en évidence la sélection de l'habitat par la faune sauvage. Nous discutons enfin de la validité de ces méthodes, et du contexte dans lequel elles doivent – ou devraient – être appliquées.

2.1 D´

2.1.1 les concepts d'habitat et de niche

2.1.1.1 le concept d'habitat

Toute étude visant à mettre en évidence la sélection de l'habitat par un organisme ou par un ensemble d'organismes doit être claire sur la terminologie employée. Plusieurs auteurs ont souligné la rareté des définitions de termes tels que *habitat, sélection de l'habitat, qualité de l'habitat, utilisation, disponibilité,* malgré leur usage fréquent dans les publications scientifiques (K 1981, H *et al.* 1997, M 2001). Il est en général sous-entendu que ces termes sont compris par les lecteurs.

Pourtant, la définition du concept d'*habitat* est sujette à polémique. Il ne s'agit pas ici seulement d'une question de vocabulaire. Chaque domaine de l'Ecologie utilisant ce terme possède sa propre vision de ce qu'est l'habitat. Or, étant donnée la fréquence de l'utilisation de ce mot en Ecologie, une définition précise est requise (M 2001).

Pour la majorité des auteurs, l'habitat est simplement l'endroit où un animal réside (M - 2001). Cette définition courante a été critiquée, car elle couvre une très grande variabilité de

significations similaires, mais pas identiques, et beaucoup soulignent la nécessité de la préciser (pour une revue, voir H *et al.* 1997). Ainsi, très fréquemment, le terme habitat se réfère au type d'associations végétales rencontrées sur une zone (K 1981). On parle alors de "types d'habitat", et ce concept est souvent considéré comme synonyme de "types de végétation" (e.g. pelouses, forêts de résineux, etc.). Plusieurs auteurs ont pourtant souligné que cette vision des choses est beaucoup trop réductrice. H *et al.* (1997) précisent alors cette définition : l'habitat correspond aux ressources et conditions présentes sur une zone qui produisent son occupation – incluant la survie et la reproduction – par un organisme donné.

Mais M (2001) insiste sur la dimension spatiale du concept d'habitat. L'habitat étant généralement compris comme l'endroit où l'animal réside, cet auteur définit l'habitat comme la surface physique occupée par un animal durant une période donnée (une définition proche de celle du concept de *domaine vital*, B 1998). Les facteurs que l'on considèrent habituellement comme des composantes de l'habitat sont inclus dans cette zone. D'après lui, ces facteurs relèvent plus de la définition de la *niche écologique*.

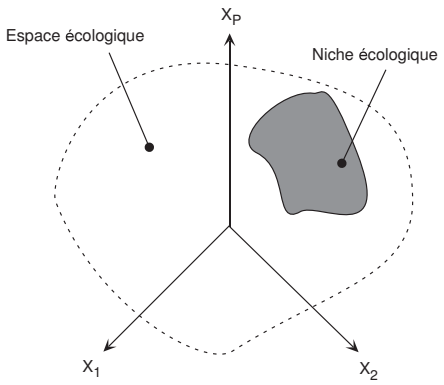
2.1.1.2 Le concept de niche écologique

Le concept de niche écologique a une longue histoire en Ecologie, et sa définition est également sujette à polémique. G (1917) a introduit ce terme pour décrire *l'ensemble ou l'étendue des caractéristiques environnementales qui permet aux individus d'une espèce de survivre et de se reproduire*. Dix ans plus tard, E (1927) donne une autre définition de la niche écologique, et lui donne un aspect plus fonctionnel : pour lui, la niche est *la fonction qu'occupe l'espèce dans la communauté dont elle est un membre*.

G a mis l'accent sur les facteurs limitant le développement d'une espèce. E , lui, ignore ces facteurs et se concentre sur l'impact de l'espèce sur son environnement. Selon ce point de vue, la hyène et le renard polaire occupent des niches écologiques similaires car ces deux espèces occupent des positions semblables dans les écosystèmes auxquels elles appartiennent (tous deux sont des prédateurs), même si ces écosystèmes sont différents (L 1995).

La niche Grinnellienne met en évidence la façon dont l'environnement affecte l'espèce. La niche Eltonienne souligne la façon dont l'espèce affecte l'environnement. La différence entre ces deux écoles est profonde, et le débat n'est toujours pas clos aujourd'hui (L 1995, M 2001). Alors que le concept de niche Grinellienne trouve sa place dans une optique autécologique (elle décrit la position d'une seule espèce dans son environnement), la pensée d'E est synécologique (elle décrit la position d'une espèce au sein d'une communauté d'espèces).

Trente ans plus tard, dans un article maintenant célèbre, H (1957) formalise le concept de niche écologique par un modèle géométrique qui aura un impact déterminant sur l'Ecologie de ces cinquante dernières années (figure 3). Les *P* variables environnementales de l'étude définissent un espace multidimensionnel – chacune des variables correspondant à une



F . *3 – Le modèle de la niche écologique formalisé par* H *(1957).*

dimension dans cet espace – et la niche correspond à *l'hypervolume dans cet espace dans lequel l'espèce peut maintenir une population viable*. H a transformé un concept biologique en un objet mathématique pouvant être analysé grâce au développement des méthodes d'analyse multivariée.

Mais cette définition de la niche n'a pas tranché le débat entre les écoles Grinellienne et Eltonienne. Si la définition de la niche par H correspond *a priori* à une optique Grinellienne, c'est-à-dire à une définition de la niche centrée sur les besoins de l'espèce, il utilise aussi la niche pour décrire l'impact d'une espèce sur son environnement, à travers les notions de *niche fondamentale* et de *niche réalisée*. La niche fondamentale correspond à la niche qu'occuperait l'espèce en l'absence de compétiteurs, alors que la niche réalisée correspond à l'hypervolume effectivement occupé par l'espèce dans un écosystème, donc en présence d'autres espèces compétitrices. En pratique, mettre en évidence la niche fondamentale d'une espèce ne peut se faire par la collecte de données sur un écosystème, mais par la modélisation mathématique des contraintes physiologiques théoriques (G et Z 2000). La collecte de données sur une espèce dans un écosystème ne peut aboutir qu'à une modélisation de la niche réalisée, c'est-à-dire à une modélisation du résultat de l'effet des interactions biotiques et de l'exclusion compétitive sur la niche fondamentale. Les deux écoles ne diffèrent donc pas seulement par leur vision des choses, mais aussi par leur *modus operandi* (L 1995).

Retenons que les termes de *niche* et d'*habitat* sont étroitement liés. Pour certains auteurs, l'habitat décrit les composantes physiques et biologiques de l'environnement qui permettent son occupation par l'espèce étudiée, et la niche correspond à la position de l'espèce étudiée au sein de l'écosystème. Pour d'autres, c'est la niche qui décrit les caractéristiques de l'environnement permettant l'occupation d'une zone par une espèce correspond à cette définition, et l'habitat correspond à la surface occupée par une espèce, dans l'espace géographique. Pour ces derniers, l'impact d'une espèce sur son environnement s'étudie à l'aide de la distinction entre niche fondamentale et niche réalisée.

2.1.1.3 Notre position dans ces débats

Les termes *habitat* et *niche* sont loin d'être clairement définis, comme c'est le cas pour de nombreux concepts centraux en Ecologie (H *et al.* 1997, M 2001). Chaque école de pensée a sa vision propre de ce à quoi peuvent correspondre ces termes. Le chercheur d'un domaine particulier de l'Ecologie utilise sa vision de ce que sont pour lui l'habitat ou la niche d'une espèce pour conduire des études et ainsi valoriser ce point de vue.

Notre rôle, en tant que biométriciens, n'est pas de trancher entre ces différentes visions, ni de donner une meilleure définition des termes impliqués. Il est plutôt d'analyser les données collectées par les biologistes, et de les aider à valoriser ainsi leurs connaissances. Mais nous devons garder à l'esprit que derrière les concepts biologiques se cachent des polémiques et, par le dialogue, essayer de cerner au mieux la vision des choses de notre interlocuteur.

Le biométricien peut également avoir un rôle unificateur. En effet, chaque domaine de l'Ecologie développe des méthodes d'analyse qui lui sont propres. Ces méthodes sont le reflet des techniques d'échantillonnage utilisées et de la vision du concept par cette spécialité. Or, la communication entre les différents domaines de l'Ecologie est en général assez rare (A 1999). En multipliant les collaborations avec des scientifiques des différents champs de l'Ecologie travaillant sur des concepts voisins, nous pouvons faire le lien entre eux, et introduire de nouvelles approches qui contribueront à l'enrichissement de la vision des interlocuteurs. Un bon exemple de ce point de vue est donné par la méthode du noyau que nous avons introduite en Biogéographie dans le cadre de cette thèse (Annexe 7). Le biométricien a ici une position stratégique.

2.1.2 La sélection de l'habitat

Chaque espèce possède une niche écologique *sensu* G & H . Cette niche décrit l'ensemble des conditions environnementales en dehors desquelles aucune population viable de l'espèce ne peut se développer. La niche est très étroitement liée au concept d'habitat, lequel possède une dimension spatiale. A cette portion de l'espace écologique qu'est la niche correspond une distribution des animaux dans l'espace géographique. Or, les animaux sont mobiles ; ils peuvent se déplacer et éviter ainsi les zones ne correspondant pas à leurs exigences écologiques.

Cette sélection de l'habitat par les animaux est en conséquence un mécanisme évolutif qui assure que les individus recherchent et restent dans l'environnement particulier auquel ils sont adaptés (R 1981). L'animal va avoir des "*images mentales*" de ce que constitue pour lui un bon milieu, et va rechercher les milieux qui correspondent à ces images mentales. Ainsi, chaque espèce a un monde perceptif caractéristique, c'est-à-dire un jeu d'images de recherche prédéterminées. Cette idée que chaque individu répond a son monde perceptuel comme un tout organisé a été formalisée par J (1971) grâce au concept de *niche gestalt*.

Le terme *gestalt* est un terme allemand utilisé en psychologie humaine (G et R - 1999). Gestalt signifie la forme, la figure perçue. Cette théorie de psychologie se penche

sur les tenants et les aboutissants de la perception, dont le produit fini est une “forme”. Elle traite de l’organisation d’un système composé d’éléments biologiques (les sensations), psychologiques (les perceptions), bio-psychologiques (les émotions) et de leur résultantes et traitement (les pensées, les actions et les symptômes). Dans la théorie de la niche-gestalt, les perceptions de l’animal vont constituer un ensemble structuré où chaque élément, chaque processus ne peut s’envisager que dans son rapport au tout. Cet ensemble d’images de recherches constitue l’image mentale de la niche écologique de l’espèce (M *et al.* 1992). Sous de nombreux aspects la niche-gestalt est donc le jeu de facteurs qui provoquent une sélection de l’habitat.

La sélection de l’habitat est alors *un processus hiérarchique impliquant une série de décisions comportementales innées ou apprises prises par un animal concernant l’habitat qu’il utilise aux différentes échelles spatiales de l’environnement* (H *et al.* 1997). Le résultat de la sélection de l’habitat, c’est-à-dire l’association d’une espèce particulière à un habitat particulier va refléter des aspects majeurs de l’Ecologie et du comportement d’une espèce.

Notons également l’importance du concept d’échelle dans ce type d’étude. La définition de la sélection de l’habitat implique en effet la notion de processus *hiérarchique* qui se produit *aux différentes échelles spatiales de l’environnement*. En effet, nombreux sont ceux qui ont indiqué que les facteurs qui influencent la distribution des animaux peuvent être différents selon qu’on s’intéresse à l’aire de répartition géographique d’une espèce ou à la sélection des sites alimentaires par un individu au sein de son domaine vital (J 1980, R *et al.* 1981, M *et al.* 1992, L 1992, A *et al.* 1993, O 1997, 1998, P *et al.* 1998, G - et Z 2000, L *et al.* 2002, G et H 2003).

Bien que l’idée que les phénomènes écologiques puissent être différents à différentes échelles spatiales soit connue depuis longtemps (e.g. G 1968), ce n’est que récemment que l’accent a été mis sur son importance pour l’étude de la sélection de l’habitat. L’échelle de sélection de J (1980) fut la première à être définie. Cette échelle comprend quatre niveaux de sélection : (i) la sélection de premier ordre correspond à la sélection de l’aire de répartition géographique par l’espèce, (ii) la sélection de second ordre correspond à la sélection du domaine vital sur une zone au sein de cette aire de répartition, (iii) la sélection de troisième ordre identifie les sites alimentaires utilisés par l’individu au sein de son domaine vital, et (iv) la sélection de quatrième ordre correspond, au sein d’un site utilisé par l’animal, aux ressources alimentaires que l’animal consomme. Cette échelle de sélection sert de référence absolue à toutes les études traitant de la sélection de l’habitat (T et T 1990, W et G 1990, A - *et al.* 1993, M *et al.* 2002).

2.2 L

L’Office national de la chasse et de la faune sauvage étant le financeur de ce travail, nous utilisons ici la signification du terme *habitat* que lui donnent les biologistes de cet organisme : l’habitat décrit les combinaisons de variables environnementales disponibles sur une zone donnée. Ces combinaisons peuvent être utilisées ou non par l’espèce étudiée. Les études de la

sélection de l'habitat ont alors pour objectif de déterminer quelles sont les combinaisons recherchées ou évitées par les animaux. Nous utilisons à l'occasion le terme *type d'habitat*. Les types d'habitats correspondent à une discrétisation des combinaisons des caractéristiques environnementales qui, d'après le biologiste, ont une signification importante pour l'espèce étudiée (types de végétation, zonation altitudinale en montagne, etc.).

2.2.1 La mesure de l'utilisation de l'habitat

Avant même de pouvoir mettre en évidence la sélection de l'habitat par une espèce, il faut pouvoir déterminer les habitats qu'elle utilise. Mais comment mesurer l'utilisation de l'habitat ? Dans le chapitre 1, nous avons souligné que les biologistes différencient deux types d'approche pour mesurer l'utilisation de l'espace par la faune sauvage, en fonction de l'objet de l'étude considéré. Soit la population est au centre de l'étude et les variations d'utilisation de l'espace sont mesurées par les variations de densité de la population sur une zone à un instant *t* (études populationnelles), soit c'est l'individu qui est l'objet de l'étude, et l'utilisation de l'espace est mesurée par le pourcentage de temps qu'il passe dans les différentes zones de son domaine vital (études individuelles).

Dans le cas d'études populationnelles, il faut collecter des données qui reflètent la distribution des animaux sur la zone. Ainsi, un échantillonnage systématique de la zone peut être entrepris, en plaçant une grille de quadrats sur la zone, et en notant la présence/absence ou l'abondance de l'espèce au sein de chaque quadrat. Mais, en général les données de présence/absence sont préférées aux données d'abondance (M 1998, P et F 2001), car elles sont plus économiques à collecter – d'un point de vue temps comme d'un point de vue personnel – sans qu'il y ait de réelle perte d'information. D'autres études impliquent le parcours de transects ou de circuits par des observateurs et le positionnement des individus détectés, afin d'obtenir une carte de la distribution des individus de l'espèce détectés sur la zone (chapitre 8 pour un exemple). Le principal inconvénient de ce type d'étude est qu'il repose sur l'hypothèse d'une égale détectabilité des animaux sur toute la zone échantillonnée (T et T 1990, M *et al.* 2002). Or la détectabilité d'un animal peut varier en fonction de son âge, du type d'habitat que l'on échantillonne, des conditions météorologiques, de la saison, de l'heure de la journée, et des compétences de l'observateur (P et F 2001). Un autre moyen consiste à se focaliser sur les indices de présence de l'espèce (e.g. des fécès ou des traces, L et K 1988, M *et al.* 2002), et à supposer que la distribution de ces indices reflète celle de la population (hypothèse qui doit alors être vérifiée par les biologistes).

Les études individuelles font appel à la capture et au marquage d'individus, par exemple à l'aide de balises radio, qui permettent de les localiser sans les déranger (W et G 1990). Les animaux sont ensuite localisés à intervalles réguliers pendant une période donnée, et à la fin de l'étude, on dispose d'un ensemble de localisations pour chaque animal suivi, qui reflète son utilisation de l'espace. L'utilisation de l'espace est alors mesurée par le pourcentage de temps passé par l'animal dans chaque type d'habitat. La grande difficulté de ce type d'étude est d'arriver à capturer un échantillon représentatif de la population étudiée. Parce que plusieurs animaux peuvent présenter des stratégies différentes d'utilisation de l'espace, le nombre d'animaux suivis doit être important. Or, la capture des animaux est souvent une étape difficile dans

l'étude (M 1996).

En outre, plusieurs auteurs utilisent des données de type individuel dans des études populationnelles (e.g. A *et al.* 1993, C *et al.* 1999, M *et al.* 2000). En effet, les animaux suivis par radio-pistage sont en général sédentaires, c'est-à-dire qu'ils restreignent leur activité sur une surface réduite appelée *domaine vital*. Plusieurs méthodes permettent d'estimer la forme et la position de ce domaine vital à l'aide de données de radio-pistage (e.g. M 1947, W 1989, 1995a, G et W 2004). La distribution des animaux sur la zone peut alors être représentée par la distribution des domaines vitaux des animaux suivis. Pourtant, une telle étude suppose que tous les animaux présents sur la zone ont une égale probabilité d'être capturés et marqués, et que l'échantillon d'animaux suivis est suffisamment grand pour permettre une telle étude (M *et al.* 2002). Or, ces hypothèses sont rarement vérifiées en réalité : le suivi des individus est en effet une opération très coûteuse en temps comme en personnel, ce qui limite nécessairement le nombre d'animaux suivis.

2.2.2 La mesure de la préférence/qualité de l'habitat

Une fois que les données qui mesurent l'utilisation de l'espace par la faune ont été récoltées, il faut pouvoir en tirer des conclusions sur les habitats qu'ils recherchent, autrement dit mesurer la qualité des habitats pour ces animaux. Les études populationnelles comme les études individuelles supposent que la présence d'individus à un endroit donné est en soi une preuve que les conditions écologiques y sont adéquates pour la survie de l'espèce – ou de l'individu – et donc font partie de sa niche écologique (G 1971). Pourtant, nombreux sont ceux qui soulignent que ce n'est pas parce qu'un habitat est très utilisé qu'il est recherché par les animaux, ou même nécessaire à l'animal (H et H 1990, W et G 1990, N et R 1996, P *et al.* 1998). En théorie, pour pouvoir mesurer la qualité de l'habitat, il faudrait pouvoir mesurer la fitness des individus (i.e. leur survie et leur reproduction) et la corréler aux caractéristiques de l'environnement. En pratique, on suppose le plus souvent que la présence d'individus implique une fitness positive pour l'espèce (G 1971, P et F 2001).

Mais un point mérite d'être noté. Les animaux ne passent pas une durée égale à effectuer les différentes activités qui leur sont nécessaires (recherche de nourriture, repos, défense du territoire, etc.). Le temps passé à un type d'activité donné n'est donc pas nécessairement la meilleure mesure de son importance. Comme M *et al.* (1992) l'indiquent : "*drinking may take only a few minutes of each day, but without water the animal is unlikely to survive*". Récemment, B et M (2003) ont formalisé cet aspect à travers le concept de *devises d'utilisation* (*currencies of use*), un concept plus aisément mesurable dans les études individuelles. La préférence d'un habitat peut se mesurer de différentes façons : elle peut bien sûr être mesurée par la densité d'occurrences de l'individu à un site donné, qui reflète le temps qu'y a passé l'individu, mais aussi par la fréquence des visites de ce site (cas d'un habitat où l'individu est rarement localisé, mais dont la fréquentation est régulière), par la distance parcourue pour y accéder, etc.

Par ailleurs, même si l'on suppose que l'abondance des occurrences dans une zone est la preuve que cette zone est de bonne qualité, l'absence d'occurrences ne signifie pas forcément que l'habitat est de mauvaise qualité. Une absence en un point peut vouloir dire que (i) l'espèce ne peut pas vivre là (la niche n'inclut pas ce point), (ii) l'espèce pourrait vivre ici, mais n'en a jamais eu l'opportunité pour des raisons historiques, (iii) l'espèce vit ici, mais l'échantillonnage n'a pas réussi à la détecter. Ces trois limites ont été reportées par de nombreux auteurs (G 1971, J 1981b, P 1984, M *et al.* 1992, N et R 1996, S 2000, H *et al.* 2002). Ainsi, les zones où l'espèce n'a pas été détectée ne peuvent pas être considérées comme étant de mauvaise qualité. La surface sur laquelle l'échantillonnage a été effectué est considérée comme *disponible* à l'espèce, mais on ne sait rien de son utilisation (P 1979). L'étude de la sélection de l'habitat consiste alors à comparer les conditions environnementales disponibles à celles qui sont utilisées par l'espèce, pour déterminer les facteurs environnementaux qui affectent le plus la distribution de l'espèce (N *et al.* 1974, J 1980, P et C 1987, A *et al.* 1993, M *et al.* 2002, H *et al.* 2002).

2.2.3 La mesure de la disponibilité de l'habitat

L'étude de la sélection de l'habitat est donc effectuée en comparant l'utilisation des habitats à leur disponibilité. La question se pose alors de savoir comment mesurer la disponibilité des habitats. La disponibilité de l'habitat a été définie par H *et al.* (1997) comme *l'accessibilité et la procurabilité des composantes physiques et biologiques d'un habitat aux animaux*. Sa mesure présente donc deux aspects :

- Sur quelles variables environnementales doit porter l'étude (composantes biologiques) ?
- Quelle est la surface que l'on peut considérer comme disponible aux animaux étudiés (composantes physiques) ?

Encore une fois, on souligne ici la dualité des aspects spatiaux et écologiques de la notion d'habitat.

M *et al.* (1992) indiquent "*we must be able to measure what the animal sees*". En effet, il faut que les variables environnementales prises en compte dans l'analyse aient une signification pour l'animal (P 1979, W 1981, M 2001). L'animal perçoit-il la différence entre deux sous-espèces d'une plante donnée présente sur la zone ? L'étude de la distribution des animaux en fonction de l'altitude est-elle toujours judicieuse ? etc. Ce problème est souvent soulevée en ornithologie, car il est au cœur du concept de niche-gestalt développé dans ce domaine (J 1971). Trois types de variables environnementales peuvent être pris en compte dans ce type d'étude (G et Z 2000, G et H 2003) : les *ressources* (éléments consommés par les animaux), les *variables directes* (variables ayant une importance physiologique pour l'espèce ; e.g. température, précipitation) et les *variables indirectes* (variables qui traduisent des combinaisons des variables précédentes, e.g. l'exposition ou le type d'association végétale). G et Z (2000) notent que l'inconvénient majeur d'utiliser des variables indirectes est que le modèle qui résultera de l'analyse ne sera valable que sur une zone restreinte.

Du point de vue spatial, il est facile de définir la disponibilité aux animaux dans les études populationnelles. En effet, la zone considérée comme disponible est généralement la zone échantillonnée, c'est-à-dire la zone d'étude ; les frontières de la zone d'étude doivent alors être définies sur des bases biologiques, et non sur des bases logistiques (P et C 1987).

En revanche, la délimitation de la surface disponible est l'un de problèmes principaux des études individuelles (P 1979, A *et al.* 1993, M C *et al.* 1998, E - *et al.* 1998, O 1998). Il est en effet indispensable de prendre en compte la sédentarité des animaux (R et M K 1999) : l'animal restreint ses déplacements à l'intérieur de son domaine vital, et l'objectif de l'étude sera souvent de mettre en évidence les zones les plus recherchées au sein de cette surface (A *et al.* 1993). Ce type d'étude implique une définition de la surface disponible *pour chaque individu suivi*. Or, la définition du concept de domaine vital est encore problématique (e.g. B 1998). En outre, ce que le biologiste et ce que l'animal considèrent comme disponible peuvent être fondamentalement différents : la territorialité, le bruit, la présence d'autres espèces ou de conspécifiques peuvent interdire à l'animal l'accès à certaines zones (W et G 1990). Mais d'autre part, ces facteurs peuvent précisément être l'objet de l'étude ; il peut par exemple être intéressant pour le biologiste de déterminer si la présence d'une espèce A dans une zone interdit l'accès de cette zone à l'espèce B (voir chapitre 8 pour un exemple).

2.3 L

Les biologistes voient fréquemment l'analyse des données comme le choix de la meilleure méthode parmi toutes celles qui sont disponibles dans la littérature, les *standards du domaine*. Nous illustrons ce point de vue sur un exemple. Etant donnée la très grande diversité des domaines dans lesquels se pose la question de mettre en relation la distribution des individus d'une espèce avec les caractéristiques de l'environnement, il est difficile de faire un inventaire complet de toutes ces méthodes développées dans tous les domaines. Nous nous concentrons donc sur un champ en particulier, la recherche en gestion de la faune sauvage (*wildlife management*), et nous illustrons ces aspects à travers une présentation historique de quelques analyses utilisées pour mettre en évidence la sélection de l'habitat dans ce domaine.

Dans le paragraphe précédent, nous avons pu montrer que toutes les études de la sélection de l'habitat reposaient sur la comparaison entre l'utilisation et la disponibilité des habitats. T et T (1990) ont classé les types de données généralement utilisés en trois grands types de protocoles, ou *designs* :

– **les protocoles de type I** : dans ce type d'études, c'est à la fois l'utilisation et la disponibilité qui sont mesurées à l'échelle de la population. Les animaux ne sont pas identifiés (pas de suivi individuel). On mesure alors l'intensité d'utilisation d'une zone par la densité des animaux sur cette zone. Ces études populationnelles font appel à des échantillonnages par transects, par quadrats. Elles peuvent se focaliser sur les animaux eux-mêmes ou sur des indices de leur présence, tels que des fécès ou des traces ;

– **les protocoles de type II** : ce type d'études correspond également à une approche populationnelle de la sélection de l'habitat. La disponibilité des habitats est en effet mesurée à l'échelle de la population. En revanche, les animaux sont identifiés (par exemple à l'aide de balises radio), et l'utilisation est mesurée pour chacun. Par exemple, on se trouve dans ce cas de figure quand on cherche à mettre en évidence les caractéristiques de l'environnement qui déterminent l'emplacement des domaines vitaux des animaux au sein d'une zone donnée ;

– **les protocoles de type III** : ce type d'études correspond à une *approche individuelle* de la sélection de l'habitat. Les animaux sont identifiés (par exemple à l'aide de balises radio), mais cette fois c'est à la fois l'utilisation *et* la disponibilité des habitats qui sont mesurées pour chaque animal. Par exemple, ce type d'étude peut correspondre aux cas où l'on cherche à mettre en évidence la sélection des sites alimentaires au sein du domaine vital de chaque animal.

Cette classification des types de données est aujourd'hui une référence incontournable dans la littérature traitant de l'analyse de la sélection de l'habitat (A *et al.* 1993, O 1997, 1998, M *et al.* 2002).

Les analyses ordinairement utilisées dans l'étude de la sélection de l'habitat peuvent donc être classées selon ces trois grandes catégories. Ces analyses ont été développées afin de répondre aux quatre grandes questions habituellement posées dans ce type d'étude, identifiées par M *et al.* (1992) :

– l'utilisation de l'habitat est-elle sélective ou aléatoire ?
– si l'utilisation est sélective, quels types d'habitats sont utilisés plus ou moins qu'attendu sous l'hypothèse aléatoire ?
– si l'utilisation est sélective, le pattern de sélection est-il similaire entre les individus ?
– si l'utilisation est sélective, les patterns de sélection diffèrent-ils entre groupes d'animaux (e.g. sexes, classes d'âge), saisons, ou traitements expérimentaux ?

Notons dès à présent que l'on peut parfaitement vouloir étudier la sélection de l'habitat par un groupe d'animaux, sans avoir à supposer que tous les animaux sélectionnent l'habitat de la même façon (A *et al.* 1993).

2.3.1 Les protocoles de type I

2.3.1.1 Les prémisses : une seule variable qualitative d'habitat

Avant le développement des Systèmes d'Information Géographique dans les années 1990, l'acquisition de données sur plusieurs variables environnementales dans les études de sélection de l'habitat était difficile et coûteux en temps comme en personnel. Ainsi, bien que le concept d'habitat soit par essence un concept multivarié (H *et al.* 1997), pour des raisons pratiques, la zone d'étude était discrétisée en plusieurs types d'habitats, qui correspondaient souvent à des types d'associations végétales (pelouses, forêts, etc.). Cette stratégie est aujourd'hui encore très fréquente dans la littérature (H et V 1991, M C *et al.* 1998, J *et al.*

2000, R *et al.* 2000, M L et H 2001, L *et al.* 2003).

Lorsque cette approche est choisie pour mesurer la disponibilité de l'habitat, les études populationnelles permettent de disposer d'une distribution des individus dans les différents types d'habitats. On peut alors mettre en évidence la sélection de l'habitat en comparant la disponibilité des habitats à leur utilisation. Les données que l'on doit analyser sont très simples : On a I types d'habitats et N individus détectés. Soit u_i le nombre d'individus détectés dans le type d'habitat i (utilisation), et p_i le pourcentage de la zone d'étude recouverte par ce type d'habitat, avec i variant de 1 à I. Deux questions sont ordinairement posées dans ce type d'étude :

- La distribution des individus dans les différents types d'habitat est-elle significativement différente de celle qu'on obtiendrait sous l'hypothèse d'une utilisation aléatoire des habitats ?
- Comment établir un classement des habitats par ordre de préférence pour l'espèce considérée ?

De nombreuses méthodes ont été développées pour répondre à ces deux questions. Nous en présentons deux ici, qui nous servirons dans les chapitres suivants.

Le test de l'existence d'une sélection de l'habitat peut être effectué à l'aide d'un test du χ^2 (N *et al.* 1974, B *et al.* 1981, W et G 1990, M *et al.* 2002). On teste l'existence d'une sélection de l'habitat significative, en comparant la distribution des occurrences de l'espèce dans les différents types d'habitat avec la distribution théorique obtenue sous l'hypothèse d'une utilisation aléatoire de l'habitat :

$$\chi^2_{\text{obs}} = \sum_{i=1}^{I} \frac{(u_i - Np_i)^2}{Np_i} \tag{2.1}$$

Si N est suffisamment grand, que les occurrences de l'espèce sont indépendantes et que p_i n'est pas trop proche de 0, alors cette statistique converge vers une distribution du χ^2 à $I - 1$ degrés de liberté. Un test significatif indique la présence d'une sélection de l'habitat par l'espèce considérée. Le test du χ^2 est de loin le plus utilisé pour tester la sélection de l'habitat dans ce type d'études (B *et al.* 1981, A et R 1986, P et C 1987, A - et R 1992, S 1994, C 1996). De nombreuses autres méthodes ont été développées pour tester l'existence d'une sélection de l'habitat avec ce type de protocole, mais nous ne les détaillerons pas ici (voir A et R 1986, 1992)

La question qui se pose ensuite est de déterminer en quoi consiste cette sélection de l'habitat. La méthode la plus utilisée consiste à calculer un intervalle de confiance sur la proportion utilisée de chaque habitat, pour la comparer ensuite à la proportion disponible de cet habitat (N *et al.* 1974, W et G 1990, S 1994, M *et al.* 2002). Soit $\overline{o_i} = u_i/N$ l'estimation la proportion utilisée pour l'habitat i. Alors un intervalle de confiance sur la proportion théorique o_i est obtenue par l'équation :

$$\overline{o_i} - z_{1-\alpha/2I}\sqrt{\frac{\overline{o_i}(1-\overline{o_i})}{N}} \leq o_i \leq \overline{o_i} + z_{1-\alpha/2k}\sqrt{\frac{\overline{o_i}(1-\overline{o_i})}{N}}$$

où $z_{1-\alpha/2I}$ est la valeur de la loi normale pour le niveau de probabilité α modifié par la correction de Bonferroni, pour prendre en compte le fait que I intervalles de confiance sont construits simultanément (B et A 1995). La convergence vers une loi normale est assurée sous les mêmes hypothèses que le test du χ^2 présenté équation 2.1. On peut ensuite classer les habitats comme préférés, rejetés ou indifférents selon que la proportion disponible est respectivement inférieure, supérieure, ou entre les limites de ces intervalles de confiance. Il est à noter que plusieurs corrections de ces intervalles de confiance ont été proposées afin de prendre en compte le fait que l'approximation normale utilisée ne tient pas pour les petits échantillons (pour une revue, cf. C 1996).

Une autre approche pour déterminer la préférence des animaux pour les types d'habitat consiste à calculer des indices de sélectivité, qui mesurent l'intensité avec laquelle les types d'habitats sont recherchés ou évités (cf. M *et al.* 2002, pour une revue, et notamment leur tableau 1.1, p.10). Mais le plus utilisé et le plus recommandé est vraisemblablement l'indice de sélectivité de M *et al.* (1972) (C 1983, H 1985, M *et al.* 2002). Pour l'habitat i, cet indice est simplement calculé par $\widehat{w_i} = o_i/p_i$. Un intervalle de confiance sur l'indice théorique w_i peut être obtenu par l'équation :

$$\widehat{w_i} - z_{1-\alpha/2I}\sqrt{\frac{\overline{o_i}(1-\overline{o_i})}{Np_i^2}} \leq w_i \leq \widehat{w_i} + z_{1-\alpha/2I}\sqrt{\frac{\overline{o_i}(1-\overline{o_i})}{Np_i^2}}$$

Sous l'hypothèse d'une absence de sélection de l'habitat i, l'indice de sélectivité w_i devrait être égal à 1. w_i est supérieur à 1 si l'habitat est recherché, et inférieur à 1 s'il est évité. Cette hypothèse est testable grâce à ces intervalles de confiance. Cet indice présente en outre une propriété intéressante : il correspond à la probabilité d'utilisation par l'espèce du type d'habitat i multipliée par une constante inconnue (M *et al.* 2002). Ces indices de sélection se placent donc dans le cadre théorique des fonctions de sélection de ressources que nous décrivons dans le paragraphe suivant.

2.3.1.2 Plusieurs variables d'habitat

Mesurer l'habitat à l'aide d'une variable qualitative décrivant plusieurs types d'habitat est aujourd'hui encore une démarche courante, car elle est souvent suffisante pour résoudre les problèmes biologiques qui ont motivé l'étude. Mais les méthodes présentées dans le paragraphe précédent requièrent que le nombre d'habitats ne soit pas trop important, ce qui permet de s'assurer que la proportion disponible des types d'habitats ne soit pas trop proche de 0 (A et R 1992). En conséquence, le nombre d'habitats considérés dans ce type d'étude dépasse

rarement 10. Cette démarche est souvent critiquée (M 2001, K 2003). En effet, la structure de l'habitat varie souvent de façon continue, et découper artificiellement des gradients en types d'habitats discrets et homogènes peut masquer des effets importants.

L'habitat, considéré du point de vue des caractéristiques environnementales, est un concept multivarié. Or, la récente expansion des Systèmes d'Information Géographique a rendu possible la prise en compte de cette caractéristique dans les études de sélection de l'habitat (D *et al.* 1987). Ces logiciels permettent de gérer de l'information spatiale. Il est donc possible de disposer de plusieurs cartes décrivant les caractéristiques environnementales sur une zone donnée. Par ailleurs, l'échantillonnage effectué par les biologistes aboutit à une distribution d'occurrences géoréférencée sur la zone. Et c'est en couplant la distribution d'occurrences aux cartes que l'on pourra mettre en évidence la sélection de l'habitat par la faune sauvage (M *et al.* 2002).

L'analyse des données pourra avoir deux principaux objectifs :
- mettre en évidence les variables environnementales qui vont avoir une influence majeure sur la distribution des animaux
- modéliser la qualité de l'habitat en fonction des variables environnementales.

Pour atteindre ces deux objectifs, ce sont principalement des méthodes de modélisation statistiques qui sont utilisées, telles que le modèle linéaire généralisé (M C et N 1989). Ce type d'outils permet en effet l'ajustement de cartes de qualité de l'habitat qui rencontrent un grand succès chez les gestionnaires de la faune sauvage (C et C 1995, K et D 1997, E *et al.* 1998, B et M D 1999, M *et al.* 2002). En effet, lorsqu'un bon modèle est ajusté, on peut s'en servir pour prédire les effets de changements de l'environnement sur la qualité de l'habitat pour une espèce, ce qui peut avoir des implications importantes, par exemple pour la conservation d'espèces menacées.

Ceci explique le récent succès des *fonctions de sélection des ressources* (B et M D - 1999, M *et al.* 2002), une méthodologie de modélisation statistique développée par M *et al.* (1993). Les fonctions de sélection des ressources reposent sur le raisonnement suivant : la zone d'étude peut être partitionnée en un certain nombre d'unités de ressources (UR, e.g. les pixels d'une carte raster) ; on dispose par ailleurs d'informations sur ces UR, mesurées par P variables environnementales $X_1, ..., X_j, ..., X_P$. Les fonctions de sélection des ressources permettent de modéliser la probabilité *relative* d'utilisation d'une UR en fonction des variables environnementales. En effet, pour pouvoir modéliser la probabilité *réelle* d'utilisation, il faudrait pouvoir disposer d'un échantillon des zones non-utilisées par l'espèce, ce qui est malheureusement impossible, comme nous avons pu le montrer au paragraphe 2.2.2. En revanche, on peut supposer que la probabilité *réelle* d'utilisation d'une UR est liée aux variables environnementales par l'équation :

$$w(X_1, ..., X_j, ..., X_P) = \exp(\beta_0 + \beta_1 X_1 + ... + \beta_j X_j + ... + \beta_P X_P) \tag{2.2}$$

$w(X_1, ..., X_j, ..., X_P)$ est comprise entre 0 et 1 si le terme de l'exponentielle est négatif. Il est possible de déterminer la valeur des coefficients $\beta_1, ..., \beta_P$ grâce à une simple régression logistique, en utilisant l'échantillon d'UR utilisées et un échantillon d'UR disponibles. Soit Y une variable décrivant le statut de N UR, qui prend la valeur 1 pour une UR utilisée, et 0 pour une UR disponible qui ne fait pas partie de l'échantillon d'UR utilisée. En ajustant un modèle linéaire généralisé pour la loi binomiale, avec une fonction de lien logit, on peut estimer les paramètres de l'équation :

$$P(Y = 1) = \frac{\exp(\beta_0^* + \beta_1 X_1 + ... + \beta_j X_j + ... + \beta_P X_P)}{1 + \exp(\beta_0^* + \beta_1 X_1 + ... + \beta_j X_j + ... + \beta_P X_P)}$$

Il est prouvé que les coefficients des variables dans cette régression logistique sont identiques à ceux de l'équation 2.2 (cf. M *et al.* 2002, p. 99-100). En revanche β_0^* est différent de β_0. Autrement dit, on n'estime pas la probabilité d'utilisation d'une ressource, mais cette probabilité multipliée par une constante inconnue. En remplaçant les coefficients estimés par cette régression logistique dans l'équation 2.2, il est possible de cartographier les variations de la probabilité d'utilisation des pixels d'une carte raster.

L'ajustement d'une fonction de sélection des ressources suppose que l'on a indépendance entre les animaux, afin de permettre l'ajustement de la fonction par le maximum de vraisemblance (i.e. pas de territorialité, pas de structures grégaires, etc.). En outre, dans la mesure où l'on utilise le modèle linéaire généralisé pour ajuster la fonction, il faut que le nombre de variables incluses dans le modèle soit limité par rapport au nombre d'UR. B et M D (1999) indiquent en conséquence un point fondamental : "***we assume that the modeler knows the limiting factors that influence the distribution and abundance of the study organism and that data are available on key resource variables***". En d'autres termes, les fonctions de sélection des ressources sont inadéquates pour déterminer les variables qui ont une influence prépondérante sur la distribution de l'espèce.

D'autres méthodes de modélisation ont également été proposées dans ce cadre. Ainsi, le modèle additif généralisé (GAM, H et T 1990) peut être utilisé à la place du modèle linéaire généralisé. Cette méthodologie repose sur la même approche que le modèle linéaire généralisé, mais au lieu de modéliser une variable dépendante par une combinaison linéaire de variables explicatives $X_1, X_2, ..., X_P$, les GAM modélisent la variable dépendante par une combinaison linéaire de *fonctions* des variables explicatives $f(X_1)$, $f(X_2)$..., $f(X_P)$. L'effet de chaque variable explicative est lissé automatiquement lors de l'ajustement, et on peut déterminer pour quelles valeurs de ces variables explicatives l'effet sur la variable dépendante est le plus fort. Pour cette raison, les GAM ont rencontré un vif succès chez les écologues ces dernières années (L 1998, P et F 2001, A 2002, B et W 2002, C *et al.* 2002, G *et al.* 2002, L *et al.* 2002b, R *et al.* 2002, Z *et al.* 2002), bien que L *et al.* (2002c) indiquent que l'inconvénient majeur des GAM est la difficulté d'exporter un modèle en dehors du logiciel qui a permis de l'estimer.

Enfin, des méthodes telles que l'ENFA (*Ecological Niche Factor Analysis* (H *et al.* 2002, Z *et al.* 2002, R *et al.* 2003) ou les distances de Mahalanobis (C *et al.* 1993, K et D 1997, K et R 1998), fréquemment utilisées, reposent sur la formalisation mathématique de la niche effectuée par H (1957). Elles ont également pour but de permettre la cartographie de la qualité de l'habitat pour l'espèce. Nous ne les détaillons pas ici ; Une discussion détaillée de ces outils peut être trouvée dans le chapitre 6.

Ainsi, les outils permettant la prédiction de la qualité de l'habitat pour une espèce dans les études populationnelles ne manquent pas dans la littérature. En conséquence, le choix d'une méthode appropriée pour analyser les données est souvent difficile pour les écologues (M *et al.* 1999). Mais rappelons que la plupart des méthodes sont utilisées à des fins prédictives, et que peu de méthodes ont été développées pour permettre la mise en évidence des facteurs écologiques influençant la distribution des animaux sur la zone.

2.3.2 Les suivis d'individus par balises radios (types II et III)

Les études de la sélection de l'habitat reposant sur des données de radio-pistage soulèvent des problématiques très variées. Il peut s'agir de déterminer ce qu'il y a de commun dans la sélection de l'habitat par les animaux suivis, ou au contraire de déterminer ce qui les différencie le plus, pour établir une typologie d'animaux, ou enfin de mettre en évidence des différences entre catégories d'animaux (e.g. le sexe, l'âge, etc.).

Le radio-pistage est utilisé dans la majorité des études traitant de sélection de l'habitat (S 1994). Pourtant c'est sur ce type de données que le moins d'outils statistiques ont été développés. Plusieurs stratégies sont alors rencontrées dans la littérature écologique pour contourner cette carence en méthodes :

- Une stratégie courante consiste à regrouper les localisations de tous les animaux suivis, et d'effectuer les analyses sans prendre en compte la structuration par animal, à l'aide des méthodes développées pour les protocoles de type I. Ainsi, de nombreux articles utilisent le test de N *et al.* (1974) pour analyser la sélection de l'habitat, en calculant les pourcentages utilisés et disponibles sur l'ensemble des localisations de tous les animaux suivis (S 1983, P et P 1984, L *et al.* 1986, M *et al.* 1987, H et V 1991, P et H 1991). Or, l'unité expérimentale dans les études de radio-pistage n'est pas la localisation, mais l'animal (A et R 1992, A *et al.* 1993, A *et al.* 1996, O et W 1999). Travailler à l'échelle de la localisation peut donc conduire à des biais.

- La seconde stratégie correspond à l'approche que M *et al.* (2002, p.8) appellent "*first and second stage analysis*". Cette approche consiste à analyser séparément les données obtenues sur chaque animal à l'aide de méthodes initialement développées pour les protocoles de type I, puis à combiner les résultats de façon à en tirer des conclusions à l'échelle de la population. Par exemple, W et G (1990) indiquent comment adapter aux protocoles de type II la statistique du χ^2 proposée par N *et al.* (1974) , pour P animaux

suivis et J types d'habitats :

$$\chi^2_{WG} = \sum_{j=1}^{J} \sum_{i=1}^{P} \frac{(u_{ij} - p_i u_{\bullet j})^2}{p_i u_{\bullet j}}$$

où u_{ij} est le nombre de localisations de l'animal j dans l'habitat i, p_i est la proportion disponible de l'habitat i, et $u_{\bullet j}$ est le nombre total de localisations de l'animal j. Dans ce cas de figure, le test du χ^2 est calculé pour chaque animal, puis les différentes statistiques sont sommées sur tous les animaux, pour en tirer un test global. Une extension similaire de cette statistique existe pour les protocoles de type III (W et G 1990). De façon similaire, M *et al.* (2002) ont proposé une adaptation des indices de sélectivité aux protocoles de types II et III : ces indices sont calculés pour chaque animal, puis une moyenne et une erreur-type sont calculées sur l'ensemble des indices. De même, cette approche a été proposée pour étendre le principe des fonctions de sélection des ressources à la modélisation de la qualité de l'habitat avec des données de radio-pistage (M *et al.* 2002, B *et al.* 2002).

– Enfin, une dernière stratégie, la plus rare, consiste à développer des méthodes spécifiquement conçues pour les données de radio-pistage. Ces méthodes, qui prennent en compte les caractéristiques particulières de ces données, rencontrent en général un très grand succès auprès des écologues. Ainsi, l'analyse compositionnelle développée par A *et al.* (1993) fait figure de référence incontournable (T *et al.* 1996, P *et al.* 1998, C *et al.* 1999, J *et al.* 2000, L *et al.* 2003, J *et al.* 2004). Cette méthode prend en compte le fait que l'unité d'échantillonnage n'est pas la localisation, mais l'animal ; elle permet de tester la différence de l'utilisation de l'habitat entre catégories d'animaux, par exemple entre les mâles et les femelles ; et enfin, elle peut être appliquée sur des protocoles de type II comme sur des protocoles de type III.

2.4 D

Dans ce paragraphe, nous discutons principalement de l'utilisation des méthodes appartenant à la famille du modèle linéaire dans les études de sélection de l'habitat, telles que le modèle linéaire généralisé (M C et N 1989) ou le modèle additif généralisé (H - et T 1990). Cependant, les points que nous mettons en exergue soulignent que les statistiques sont souvent mal utilisées par les biologistes, et que la démarche sous-jacente aux méthodes qu'ils utilisent est à revoir.

2.4.1 Le problème de l'inférence

Quel que soit le protocole impliqué et l'échelle d'étude de la sélection de l'habitat, les méthodes de modélisation statistique sont celles qui rencontrent le plus large succès auprès des éco-

logues chargés de la gestion de la faune sauvage. En effet, ces méthodes permettent en théorie de construire des modèles prédictifs donnant la qualité de l'habitat en fonction des conditions de milieu rencontrées sur la zone, et donc de construire des scénarios de gestion potentiels. Ainsi, H *et al.* (2003) utilisent les fonctions de sélection des ressources pour minimiser l'impact d'une exploitation forestière sur la chouette tachetée du Nord (*Strix occidentalis caurina*) par la conservation des habitats les plus utilisés par cette espèce menacée.

Mais la principale question sous-jacente à toutes ces études est celle de la possibilité d'étendre les résultats à la population étudiée. Lorsqu'on travaille sur des données individuelles (i.e. suivis par radio-pistage), la validité des conclusions tirées dépendra bien sûr du nombre d'animaux suivis dans l'étude (W et G 1990, A et R 1992, A *et al.* 1993, O et W 1999). Mais au delà de la question du nombre d'animaux suivis, la question de l'inférence des résultats obtenus à la réalité du terrain pose de sérieux problèmes, ce qui est d'autant plus gênant que les écologues ont besoin de cette inférence pour aider à évaluer les effets de certaines modifications du milieu (coupes forestières, embroussaillement, etc.) sur la population (B *et al.* 2002, O et S -S 2002, L *et al.* 2002b).

Certains ont proposé de rajouter, à la fin du processus de modélisation, une étape de validation. Le moyen le plus utilisé pour ce but est la validation croisée (e.g. la *K-fold cross-validation* B *et al.* 2002). Cette approche consiste, avant l'analyse, à mettre de côté une fraction des localisations des animaux (e.g. 10%), puis à mesurer la qualité d'ajustement du modèle en vérifiant que la qualité de l'habitat prédite aux localisations du jeu de validation est importante. Des statistiques permettant cette mesure, telles que la courbe ROC (*Receiver Operating Characteristic*) sont parfois utilisées (B *et al.* 2002).

Pourtant, la validation croisée d'une étude ne permet pas à elle seule de nous assurer de sa validité d'un point de vue inférentiel. Il faut également s'assurer de la représentativité de l'échantillon en examinant la validité de toutes les hypothèses sous-jacentes à ce type d'étude, à savoir :

- l'indépendance entre les animaux (e.g. pas de territorialité, ni d'évitement entre les animaux, ni d'agrégation, etc.) ;
- le libre et égal accès de toutes les zones à tous les animaux présents sur la zone d'étude (e.g. pas de barrières physiques sur la zone) ;
- toutes les variables environnementales susceptibles d'affecter la distribution des animaux ont été identifiées et incluses dans le modèle ;
- l'échantillon d'animaux est représentatif de la population étudiée, ce qui implique dans les études populationnelles une détectabilité des animaux uniforme sur la zone d'étude, et dans les études individuelles, une probabilité de capture identique pour tous les animaux de la population étudiée ;
- dans le cas d'études individuelles, une disponibilité constante des ressources pendant la période d'étude.

Si l'une de ces hypothèses n'est pas vérifiée, ce qui est fréquent (e.g. voir chapitre 8), alors le modèle n'est plus valable, et l'inférence ne tient plus car l'échantillon présente des biais. Le seul moyen de valider l'étude est d'utiliser un jeu de données recueilli de façon totalement indépendante, une étape qui est rarement effectuée (M *et al.* 1992).

Pourtant, même lorsque les modèles sont validés sur des données indépendantes, l'inférence peut être un problème. Une illustration de ce problème est donnée dans les deux études de K et D (1997) et de K et R (1998) sur le lièvre de Californie (*Lepus californicus*). K et D (1997) ont construit une carte de qualité de l'habitat du lièvre de Californie à l'aide des distances de Mahalanobis sur un jeu de localisations recueillies de 1987 à 1989, et de 1992 à 1993. Puis, le modèle construit à été validé par le calcul des prédictions sur un jeu de localisations indépendant recueillies sur la même zone de 1993 à 1995. Ces localisations étaient toujours détectées dans des zones prédites comme de haute qualité. Ce modèle avait été ajusté dans le cadre de la conservation de l'aigle royal (*aquila chysaetos*) dans L'Idaho, dont la principale proie est le lièvre de Californie. En maximisant la qualité de l'habitat pour le lièvre, on maximise du même coup celle du rapace. Mais leur zone d'étude est soumise à de fréquents incendies, qui se traduisent par une réduction de l'embroussaillement de la zone. Le feu peut donc devenir un moyen de gestion des deux espèces. Par conséquent, K et R (1998) se sont posés la question de savoir s'il était possible de se servir de leur modèle pour prédire les réponses de la distribution des lièvres à différents scénarios de gestion, en utilisant des simulations. Ces auteurs ont répondu par la négative ; les simulations du modèle aboutissaient à des prédictions absurdes. **Lorsque la configuration de l'habitat change sur une zone, la validité du modèle peut être remise en question**. La question la plus importante est donc parfaitement exprimée par J (1981b) :

"*What is the universe to which our results are to pertain ? If it is a single study area, in a single year, with the measurements we have observed, there is no problem. If we want to generalize, let us be careful. Our study area must be representative of the area we want to extrapolate to, similarly the year and the habitat.*".

2.4.2 La modélisation "aveugle"

Le principal problème pratique des méthodes de modélisation statistique appartenant à la famille du modèle linéaire est de déterminer quelles variables environnementales inclure dans le modèle. Pour augmenter le pouvoir prédictif du modèle, on doit minimiser le nombre de variables explicatives (M C et N 1989). Par ailleurs, ces méthodes de modélisation supposent que les variables susceptibles d'influencer la distribution de l'espèce étudiée sont connues (B et M D 1999). Il est possible de sélectionner ces variables dans la littérature écologique ou en puisant dans le savoir du biologiste (B et A 1998), mais le plus souvent, déterminer quelles variables inclure dans le modèle est la tâche la plus difficile de la modélisation (L 1999, G et Z 2000).

Les méthodes de type "*régression pas-à-pas*" (*stepwise regression*) sont alors fréquemment utilisées pour sélectionner les variables à inclure dans le modèle (par exemple B et B 1984, L *et al.* 1986, H 1993, M *et al.* 2000, V *et al.* 2002). Ces méthodes permettent de sélectionner de façon automatique la combinaison de variables environnementales qui maximise un critère mesurant la qualité d'ajustement du modèle (e.g. le critère d'Akaiké, B et A 1998). Pourtant, ces outils ont été très lourdement critiqués par les statisticiens, qui les comparent à des "expéditions de pêche" se substituant au raisonne-

ment scientifique (J 1981b, W 1981, B et A 1998, H *et al.* 2002).

Les méthodes appartenant à la famille du modèle linéaire généralisé (M C et N - 1989) ont été développées dans un cadre bien défini (B et A 1998, C 1998), c'est-à-dire pour tester des hypothèses *précises*, et ne sont pas prévues pour l'exploration des données. En théorie, il faudrait donc : (i) poser des hypothèses sur la sélection de l'habitat par l'espèce étudiée, (ii) construire un protocole d'échantillonnage en accord avec la théorie écologique permettant de vérifier ces hypothèses, (iii) collecter les données en suivant *à la lettre* ce protocole et (iv) construire le modèle (R 1981).

Pour cette raison, B et A (1998), dont l'ouvrage sert souvent de référence pour l'utilisation de ce type de méthodes en Ecologie, indiquent : "*The process of analyzing data with few or no a priori questions, by subjectively and iteratively searching the data for patterns and 'significance', is often called by the derogatory term of 'data dredging'. Other terms include 'post hoc data analysis' or 'data mining'. Often the problem arises when data on many variables have been taken with little or no a priori motive without benefit of supporting science. (...) Example of data dredging includes the examination of crossplots of all the variables or the examination of a correlation matrix of the variables. These data-dependent activities can suggest apparent linear or nonlinear relationships in the sample and therefore lead the investigator to consider additional models. These activities should be avoided, as they probably lead to over-fitted models with spurious parameter estimates and nonimportant variables as regards the population. The sample may be well fit but the goal is to make valid inference from the sample to the population.*"

Il est donc hors de question de se servir de ce type de méthodes pour déterminer quelles variables environnementales inclure dans le modèle, puis de construire le modèle *sur les mêmes données*.

2.4.3 Les études exploratoires et les études confirmatoires

On peut identifier deux types d'études dans la littérature traitant de la sélection de l'habitat par la faune sauvage (J 1981a, M *et al.* 1992, C 1998) : les études exploratoires (*hindcasting studies*) qui ont pour but d'identifier les variables environnementales recherchées par l'espèce, la population ou l'animal, et les études confirmatoires (*forecasting studies*) qui ont pour but de prédire les conditions de milieu adéquates pour l'espèce en fonction des conditions environnementales. Ces deux types d'études doivent absolument être distingués.

En effet, M *et al.* (1992) soulignent que de nombreux auteurs utilisent les résultats des études exploratoires pour prédire les conditions futures pour l'espèce. **Ces prédictions ne peuvent pas être fiables si le protocole n'a pas été construit de façon à valider les résultats d'une pré-étude**, car les conditions environnementales, démographiques et écologiques peuvent varier de façon significative dans le temps et l'espace.

Ceci confirme ce qui a été affirmé dans le paragraphe précédent. Pour pouvoir effectuer des prédictions, il faut avoir un protocole d'échantillonnage correctement construit. Or, l'une des bases de la statistique est que pour pouvoir construire un protocole d'échantillonnage correct, il faut déjà disposer d'information sur les processus étudiés, tirée de la théorie écologique ou de pré-études (S et R 1981, T *et al.* 1993). Si l'on ne connaît que peu de choses de la population étudiée, il faudra en premier lieu conduire des études exploratoires. Et pour la construction du protocole de telles études, c'est le bon sens du biologiste qui domine. Mais en contrepartie, ces mêmes données ne pourront être utilisées à des fins confirmatoires.

Cette différence profonde entre les deux types d'études n'est souvent pas perçue par les biologistes. Une illustration de ce point de vue nous est donnée par un des référés de la revue *Ecology* chargé d'expertiser un article que nous avons écrit (C *et al.* 2005a, cf. Annexe 1). L'article présente une méthode exploratoire de la sélection de l'habitat utilisable avec des données de radio-pistage, l'analyse K-select, que nous détaillons au Chapitre 7. La principale remarque du référé concernant l'article était la suivante :

"*I agree in principle with your description of the 2 alternatives of exploratory and predictive (or confirmatory) objectives, but* ***I am skeptical that this distinction will be made in practice****. This discussion is incomplete, however, without reference to the principles of information-theoretic methods as espoused by Burnham and Anderson (1998). They emphasize the importance of a priori specification of a plausible set of models (multiple working hypotheses)* ***that are based on knowledge gained, perhaps from previous exploratory analyses****. They emphasize the need for the former approach, and caution against misinterpretation of exploratory results.*"

Nous avons décrit le point de vue de B et A (1998) dans le paragraphe précédent : ces auteurs, qui se placent dans une optique confirmatoire, déconseillent formellement toute exploration des données (qu'ils qualifient péjorativement de "*data-dredging*"), car une variable qui semble influente dans l'échantillon n'a pas nécessairement d'influence dans la population, ce qui peut conduire à la construction de modèles incluant des variables sans intérêt. Les différents modèles testés, qui formalisent les hypothèses que l'on cherche à tester, doivent être construits *a priori*, c'est-à-dire avant même la collecte des données. Et ces modèles doivent en théorie être construits, non pas sur la base des résultats d'*analyses* exploratoires, mais sur la base d'*études* exploratoires, donc sur des données différentes.

Or, conduire une étude sur le terrain, même exploratoire, est une opération coûteuse en temps, en argent et en personnel. En outre, les questions posées par ces études exploratoires sont en général très diverses. Par exemple, dans les études individuelles, ces questions concerneront aussi bien les surfaces occupées par les animaux (domaines vitaux), que les habitats qu'ils recherchent, les interactions entre les animaux, ou avec d'autres espèces compétitrices ou prédatrices, le processus de dispersion des juvéniles, etc. Il est pratiquement impossible de construire un protocole qui permette de répondre simultanément à toutes ces questions avec les mêmes données, et le plus souvent, la construction du protocole repose sur les compétences du biologiste qui recherchera le compromis permettant de répondre au mieux à toutes ces questions. Dans le cas d'études reposant sur l'utilisation du radio-pistage, cela impliquera de nombreux choix, tels que le choix du mode de capture (e.g. sites de captures ou battues), la position des

sites de captures sur la zone d'étude (dans quels habitats les capturer), éventuellement la surveillance des sites de capture, le choix des animaux à suivre en fonction de leur sexe, leur âge, leur statut de dominance dans le groupe, etc. (pour un exemple, cf. M 1996)

Les données recueillies dans ce type d'étude devront ensuite être analysées, ou plus exactement, *explorées* pour en tirer un modèle conceptuel du fonctionnement du système biologique étudié (R 1981, cf. chapitre suivant). Et c'est seulement après cette première étape que l'on pourra construire un protocole dont les données serviront à la prédiction, étape ordinairement occultée.

On comprend dès lors pourquoi T *et al.* (1993) constataient que beaucoup aimeraient que les analyses soient indépendantes des données. En effet, si les écologues pouvaient disposer, grâce aux mêmes données, des qualités informatives des études exploratoires et des qualités de prédiction des études confirmatoires, ils pourraient faire de considérables économies.

2.4.4 Quelle est la meilleure méthode ?

La question de savoir comment mettre en évidence la sélection de l'habitat par la faune sauvage se pose dans de nombreux champs de l'Ecologie (§ 1.1.2). Nous avons également montré que les biologistes espéraient à partir d'un seul jeu de données effectuer des analyses exploratoires et confirmatoires.

Cet espoir prend forme de la façon suivante. Plusieurs méthodes sont développées au sein de chacun des domaines de l'Ecologie. Lorsqu'une nouvelle méthode est développée, les auteurs mettent en général en exergue une contrainte particulière des données qu'ils jugent indispensable de prendre en considération, et qui n'est pas prise en compte par les méthodes existantes. Ainsi, M et I (1998) montrent que la sélection de l'habitat par un animal peut varier en fonction de la disponibilité des types d'habitat ; A *et al.* (1993) soulignent que la sélection de l'habitat peut être variable en fonction du sexe ou de l'âge de l'animal ; etc.

Ce nouvel aspect de la sélection de l'habitat peut être mis en évidence soit à partir des contraintes soulevées par un jeu de données particulier (cf. exemple chapitre 8), soit pour insister sur une idée qu'ils considèrent comme ayant une place importante dans la théorie, c'est-à-dire avec pour objectif de clarifier la définition des concepts impliqués (e.g. pour l'idée que la sélection de l'habitat puisse varier en fonction de l'échelle spatiale, J 1980, A *et al.* 1993). Ces auteurs développent alors des méthodes d'analyse qui permettent de prendre en compte ces nouveaux aspects.

Mais beaucoup de biologistes voient l'étape de l'analyse comme le choix, parmi le pool de méthodes disponibles dans la littérature, de *la* meilleure méthode pour l'analyse de leur jeu de données (M *et al.* 1999) ; C (1992) souligne que l'analyse des données leur apparaît souvent comme une étape technique. Mais aucune méthode ne permet de prendre en compte tous les aspects de la sélection de l'habitat à la fois (ce qui serait d'autant plus difficile, que, comme nous l'avons vu, le concept d'habitat n'est pas lui même très clairement défini).

Lorsque vient le moment de l'analyse, le biologiste est souvent perdu parmi la multitude des méthodes disponibles. Ainsi, si on lui demande *a priori* de choisir entre la prise en compte d'une éventuelle réponse fonctionnelle des animaux ou d'une éventuelle variation temporelle dans la sélection de l'habitat, le biologiste répondra probablement qu'il serait préférable de pouvoir prendre en compte les deux...

Il faut alors s'orienter vers une autre démarche pour analyser les données. Dans le chapitre suivant, nous reconstruisons cette démarche bien connue des biométriciens à travers l'étude d'un exemple, le problème de l'autocorrélation spatiale dans les études de sélection de l'habitat.

Chapitre 3

Structures spatiales et autocorrélation spatiale : vers une autre démarche

Dans le chapitre précédent, nous avons illustré le point de vue des biologistes sur la Biométrie dans le cadre de l'étude de la sélection de l'habitat par la faune sauvage. Le terme *habitat* recouvre plusieurs significations différentes, mais deux dimensions importantes peuvent être dégagées de ce concept, la dimension spatiale et la dimension écologique. Or, nous n'avons encore rien dit de l'analyse de l'aspect spatial de l'habitat.

Dans ce chapitre, nous montrons que les contraintes spatiales, lorsqu'elles sont négligées, peuvent conduire à des conclusions erronées concernant les habitats recherchés par une espèce. Dans un souci de simplicité, nous nous concentrons sur les études populationnelles pour décrire les aspects spatiaux sous-jacents à la notion d'habitat. En effet, c'est pour ce type d'étude que le plus grand nombre de méthodes ont été développées, et que le problème des structures spatiales est le plus discuté.

3.1 L' ´ : ´ ´ ?

3.1.1 Définitions de l'indépendance et de l'autocorrélation

Les méthodes d'analyse de la sélection de l'habitat reposent fréquemment sur l'hypothèse d'indépendance entre les observations (J 1981a, A *et al.* 1993, A et R 1992, C 1996, P *et al.* 1998). M *et al.* (2002) indiquent ce que l'on entend par *indépendance* dans ce type d'études : "*An animal's selection of a resource is assumed to be independent of selections made by all other animals. This assumption may be violated when animals are gregarious or territorial in habitat studies or when competition for food occurs in foraging studies*". L'hypothèse d'indépendance n'est pas spécifique aux études de sélection de l'habitat. Il s'agit en effet d'une des hypothèses les plus importantes de la statistique classique (C et O 1981, L et F 1989, L 1993).

D'un point de vue strictement statistique, on parle d'observations indépendantes lorsque les observations sont échantillonnées au sein d'une population statistique de telle sorte qu'une

valeur observée n'a d'influence sur aucune autre (L et L 1998). Dans le cas d'études populationnelles, l'indépendance implique que la détection d'un animal à un point A n'affecte pas la probabilité de détecter un autre animal à un point B, quels que soient A et B.

Le concept d'autocorrélation est lié à celui d'indépendance, à la différence que l'autocorrélation se réfère à un modèle. Etant donné un modèle de la distribution des localisations sur une zone, la composante de la variation spatiale qui n'est pas expliquée par le modèle est appelée autocorrélation spatiale (C et O 1981). L et L (1998) en donne également une définition : "*spatial autocorrelation may be loosely defined as the property of random variables which take values, at pairs of sites a given distance apart, that are more similar (positive autocorrelation) or less similar (negative autocorrelation) than expected for randomly associated pairs of observations. Autocorrelation only refers to the lack of independence among the* ***error components*** *of field data, due to geographic proximity*".

Lorsque des tests d'hypothèses sont effectués sur des données autocorrélées positivement, cela se traduit par une augmentation des erreurs de type I, c'est-à-dire à un rejet trop fréquent de l'hypothèse nulle (L 1993, L *et al.* 2002). Dans les études de la sélection de l'habitat, cela pourra par exemple conduire à inclure dans les modèles des variables environnementales qui n'ont pas de réel effet sur l'utilisation de l'espace par l'espèce étudiée.

3.1.2 Deux types de structures spatiales dans les études de sélection de l'habitat

L'autocorrélation spatiale est rarement prise en compte dans les études de la sélection de l'habitat, car ses effets sur les paramètres des modèles sont assez mal connus (L *et al.* 2002b). En réalité, comprendre l'origine de l'autocorrélation spatiale est fondamental, car elle tire son origine des structures spatiales présentes dans le jeu de données.

Ces structures spatiales peuvent avoir deux origines (L 1993, M 2001, L - *et al.* 2002, C 2005) :

- la distribution des variables d'habitat influençant la distribution des animaux : si les ressources ne sont pas distribuées aléatoirement sur la zone d'étude, alors il est probable que les animaux, qui consomment ces ressources, ne le sont pas non plus ;
- la distribution des animaux en elle-même : ce type de structure n'est pas liée aux variables d'habitat étudiées, mais tire son origine d'autres sources, par exemple de l'organisation sociale (O et S -S 2002), d'une réponse agrégative due aux indices de présence des conspécifiques (L *et al.* 2002), de stratégies de réponse à la prédation (C 2005).

Ces deux types de structures doivent absolument être distingués. Que les variables d'habitat présentent une structure spatiale n'est pas vraiment gênant pour l'étude de la sélection de l'habitat. Mais lorsque ce sont des facteurs liés à la biologie de l'espèce et non plus des facteurs environnementaux qui sont à l'origine des structures spatiales, cela pose un grave problème

pour les analyses.

Une discussion très édifiante sur ce sujet peut être trouvée dans la réponse de L (1999) à l'article de B et M D (1999) sur les fonctions de sélection des ressources. Ces derniers proposent en effet d'ignorer l'autocorrélation spatiale lors de la construction du modèle, mais de déterminer son importance *a posteriori*, par l'examen des résidus du modèle. Si une autocorrélation significative peut être mise en évidence dans les résidus du modèle, ces auteurs recommandent l'utilisation de la régression autologistique à la place de la régression logistique usuelle, pour l'ajustement de la fonction de sélection des ressources. La régression autologistique est une régression logistique dans laquelle on inclut, en plus des variables environnementales, une autocovariable. Pour une unité de ressource donnée, cette autocovariable est calculée par la moyenne de la variable réponse dans les unités de ressources voisines, pondérée par l'inverse de la distance avec ces cellules voisines (A *et al.* 1996, H *et al.* 2003).

Mais L (1999) critique cette démarche :

"*This methodological approach is common to other areas of ecology dealing with spatial data. Independence of the response variable across sites is the first major assumption of such models. As the authors emphasize, the second major assumption is that the important resources (explanatory variables) are included in the model. Unfortunately, because we usually do not know what these are, we may often be forced into doing some hypothesis testing. This is where violation of the first assumption can introduce systematic errors into our understanding of the relative importance of resources*".

En résumé, lorsqu'on travaille sur un semis de points représentant la distribution d'une population sur une zone, ce semis de points présente fréquemment des structures spatiales. La première question qui se pose est donc de **déterminer l'origine de ces structures spatiales**.

3.1.3 Un cas d'étude : la distribution des dégâts de sangliers sur vignobles

Dans le cadre de la fonction d'ingénieur en Biostatistique que j'occupais lorsque travaillais pour l'Office national de la chasse et de la faune sauvage, j'avais analysé les données d'une étude sur les facteurs influençant la présence de dégâts de sangliers sur les vignes. Ces données avaient été collectées par Daniel M pendant sa thèse (M 1996). Il avait interrogé tous les viticulteurs de Puéchabon (Hérault, à 30 km au nord-ouest de Montpellier) sur le niveau de dégâts que subissait chacune des vignes de ce village, de 1990 à 1992. Les analyses de ces résultats d'enquête ont fait l'objet d'un article aujourd'hui publié dans *European Journal of Wildlife Research* (C *et al.* 2004, Annexe 8). Nous illustrons ici les points de vue décrits dans le paragraphe précédent, à travers l'analyse critique de l'approche que j'avais utilisée à l'époque.

En milieu méditerranéen, le régime alimentaire du sanglier est principalement constitué de fruits forestiers. Les glands de chêne vert (*Quercus ilex*) qui tombent à l'automne constituent

la principale ressource alimentaire du sanglier. En été, les glands sont rares et la disponibilité alimentaire est minimale sur la zone (M 1996). Or, c'est à ce moment que les vignes fructifient, offrant aux sangliers une source de nourriture attractive. Or, toutes les vignes ne produisent pas du raisin en même temps. Trois périodes de maturation de vignes peuvent être identifiées : les vignes précoces fructifient entre mi-juillet et mi-août, les vignes normales entre mi-août et mi-septembre, et les vignes tardives entre mi-septembre et mi-octobre. En outre, les vignes de Puéchabon ne sont pas toutes localisées à la même distance du bois de chênes verts, le milieu fermé assure protection aux sangliers. Le problème posé par Daniel M était de déterminer si la distance au bois et la date de maturation du raisin avaient une influence sur la présence et l'importance des dégâts de sangliers sur le vignoble.

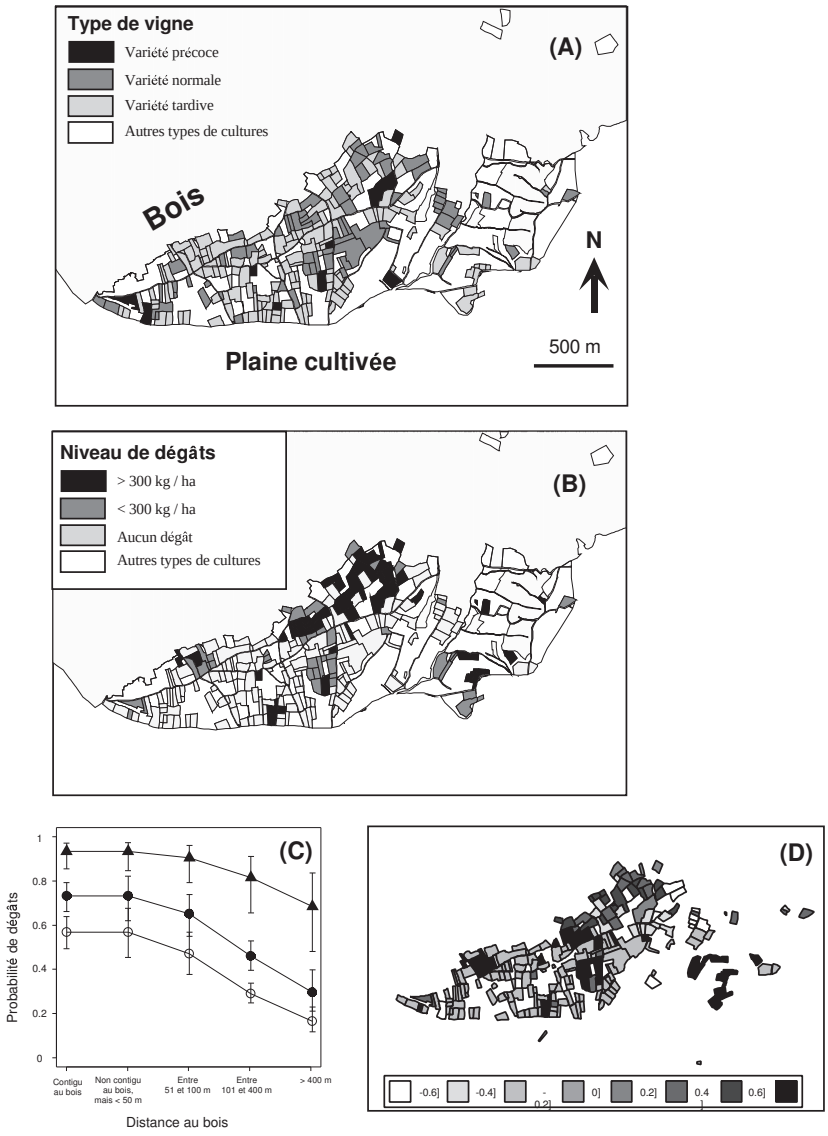
Cette étude avait donc pour objectif la mise en évidence d'une forme de sélection des ressources par le sanglier. Chaque vigne est considérée comme l'unité d'échantillonnage (unité de ressources). La présence et l'absence de dégâts peuvent être considérés respectivement comme présence et absence d'utilisation. J'avais alors utilisé la régression logistique afin de prédire la probabilité d'utilisation d'une vigne en fonction de sa période de maturation et de sa distance au bois (C *et al.* 2004).

Les résultats que j'avais obtenus sont résumés sur la figure 4. Les deux variables étudiées ont une influence significative sur la probabilité de dégâts. L'interprétation de ce modèle confirme les hypothèses des biologistes, à savoir que : (i) plus une vigne est éloignée du bois et moins elle a de chances d'être endommagée, et (ii) plus la date de maturation d'une vigne est précoce et plus elle a de chances d'être endommagée.

Or, je n'avais pas calculé les résidus de ce modèle à l'époque où j'avais écrit cet article. La cartographie de ces résidus apporte pourtant des informations intéressantes (figure 4D). En effet, ces valeurs semblent fortement structurées dans l'espace : il y a un regroupement des parcelles pour lesquels la probabilité de dégâts est sous-estimée. Cette impression est confirmée par un test de l'autocorrélation spatiale de Geary ($C = 0.5937$, $P < 0.0001$, cf. annexe 8 pour une description détaillée du principe du test). En d'autres termes, la distance au bois et la date de maturation du raisin ne suffisent pas à expliquer les structures spatiales présentes sur la zone d'étude. La probabilité qu'une parcelle soit endommagée dépend également des dégâts effectués sur les parcelles voisines.

Le caractère agrégé des dégâts de sanglier apparaît d'ailleurs de façon très nette sur la carte des dégâts (figure 4B) ; nous avions identifié la présence de ces patchs, et nous en expliquions la présence par l'effet conjugué de la distance au bois et de la date de maturation des vignes. En réalité, cette agrégation révèle une faille dans l'analyse. *L'erreur est de considérer la parcelle de vigne comme unité d'échantillonnage*. En effet, il n'y a aucune raison pour que des découpages ayant un sens pour le viticulteur en ait un pour le sanglier. Dans la mesure où il n'y a pas indépendance entre une parcelle de vigne et les vignes voisines, il aurait fallu travailler au niveau du patch de dégâts.

Il est bien connu des biologistes que les sangliers utilisent régulièrement les mêmes routes pour rechercher leur nourriture ; les passages répétés des sangliers forment ce qu'on appelle des



F . *4 – Les résultats de l'étude des dégâts de sangliers sur les vignobles de Puéchabon (Hérault) : (A) cartographie de la date de maturation des vignes ; (B) cartographie du niveau de dégâts subi par les vignes ; (C) modèle logistique prédisant la probabilité qu'une vigne soit endommagée en fonction de la distance au bois et de la date de maturation ; (D) cartographie des résidus de ce modèle*

coulées. Les vignes précoces sont les premières à être endommagées, ce qui peut expliquer le niveau plus important de dégâts qu'on y observe. Il est probable que par la suite les sangliers vont continuer à utiliser les mêmes coulées, même lorsque la maturation des vignes précoces sera finie. Les sangliers endommageront alors les vignes voisines à maturation normale ou tardive.

Comme on le voit dans cet exemple, la présence d'autocorrélation spatiale dans les résidus est révélatrice de la présence de structures spatiales non prises en compte, mais qui peuvent être

du plus grand intérêt pour la compréhension du système étudié.

3.1.4 l'autocorrélation spatiale : un nouveau paradigme

Plusieurs auteurs soulignent que l'autocorrélation spatiale n'est pas un défaut des données, mais qu'elle est au contraire révélatrice des structures spatiales du système étudié (e.g. L - et F 1989, L 1993, L 1999). Ainsi, L (1993) indique :

"*The observed variable of interest – for example species composition – are most often influenced, at any given locality, by the species assemblage structure at surrounding localities, because of the contagious biotic processes such as growth, reproduction, mortality, migration, and so on. In such a case, because the value at any one locality can be at least partly predicted by the values at neighboring points, these values are not stochastically independent from one another. This may come as a surprise to ecologists who have been trained in the belief that nature follows the assumptions of classical statistics, one of them being the independence of the observations. However, field ecologists know from experience that living beings in nature are distributed neither uniformly nor at random ; the same applies to the physical variables that we use to describe environment.*"

L souligne que la structuration spatiale est *fonctionnelle*. C'est exactement cette idée qui est mise en exergue par M (2001) lorsqu'il souligne l'importance de la dimension spatiale du concept d'habitat (§ 2.1.1.1). Il est alors difficile d'ignorer la dimension spatiale. L'autocorrélation n'est pas seulement une constatation qui doit être faite sur les données avec une grimace ; en comprendre l'origine est centrale dans les études de l'utilisation de l'espace. On comprend alors pourquoi, lorsqu'il y a autocorrélation dans un jeu de données, il ne faut donc pas se contenter d'un laconique "*Our solution to both of these problems (influential observations and spatial autocorrelation) was to use STATA's cluster option to calculate variances estimates which are robust to both influential observations and within-site correlation (StataCorp. 1999)*" (V *et al.* 2002), une stratégie malheureusement souvent retrouvée dans la littérature (B *et al.* 2002, N *et al.* 2004). Il est parfaitement légitime de vouloir intégrer l'autocorrélation au modèle, mais si l'autocorrélation est indicatrice de structures et qu'on ne sait rien des structures en question, c'est une démarche à éviter absolument. Il faut connaître ce que l'on modélise.

L'Ecologie est l'étude des interactions entre les êtres vivants, et avec leur environnement. La présence d'une structure spatiale dans les données est indicatrice de la présence d'interactions, soit entre les êtres vivants, soit avec leur environnement. L'autocorrélation spatiale est indicatrice de la présence de structures. Pour toutes ces raisons, L donne à l'autocorrélation le statut de nouveau paradigme. L'étude des sources de l'autocorrélation spatiale est aussi importante que l'étude des composantes biologiques qui affectent la distribution de la population, et peut aider à la compréhension du système.

3.2 C ' ´ '

3.2.1 La démarche biométrique

Nous avons pu montrer dans ce chapitre l'importance de la prise en compte de l'aspect spatial dans les études de sélection de l'habitat. Cette dualité entre l'espace géographique et l'espace écologique qui apparaissait dans les polémiques autour de la définition du concept d'*habitat* transparaît également au niveau des analyses : ces deux aspects sont indissociables.

3.2.1.1 Synthèse

Suite à l'intervention de J (1981a) lors d'un séminaire sur l'utilisation des statistiques multivariées dans les études de sélection de l'habitat, Jim W a posé la question "*is the distribution of plots systematically to space a random sample of plants ?*". Et J a répondu "*No, but it may be adequately represented by a model of randomness. The crucial issue is to define the population of which the sample is representative*". Aucun protocole d'échantillonnage, qu'il s'agisse d'un échantillonnage aléatoire ou systématique, ne permet dans l'absolu de s'assurer de l'indépendance entre les unités d'échantillonnage (S et R 1981, L et F 1989). Il faut pouvoir disposer d'informations suffisamment précises sur les processus en jeu pour construire un protocole qui autorise cette supposition (M 2001). En général, lorsque les données sont collectées, l'objectif du biologiste est précisément d'identifier ces processus, comme le souligne J . Mais cette réponse soulève également un aspect essentiel de la démarche biométrique que résume C (1992) :

"*Préciser en quoi le prélèvement est un échantillon, ou même de quoi il est un échantillon, de quelles structures spatiales il est extrait, de quelles stratégies de dispersion il témoigne, sous quelles influences immédiates ou lointaines de l'environnement il se trouve, de quels rythmes il dépend, de quels biais conjoncturels ou chroniques il souffre, tout aspect définissant son usage potentiel, voilà bien une question d'essence biométrique.*"

La démarche de modélisation "aveugle" proposée par B et A (1998) dans le chapitre précédent ne peut donc être utilisée sur les données ordinairement récoltées pour étudier la sélection de l'habitat. Cette démarche repose sur l'utilisation de méthodes qui réclament l'indépendance entre les individus. Or cette hypothèse est discutable parce que par définition, dans une population biologique, les individus interagissent entre eux. Si le protocole d'échantillonnage n'a pas été construit de façon à éliminer totalement les structures spatiales non causées par des facteurs autres que les variables environnementales mesurées, alors les aspects spatiaux et écologiques sont indissociables.

On peut donc résumer l'étude de la sélection de l'habitat à une simple figure (figure 2), qui représente un couplage entre l'espace géographique et l'espace écologique. Quelle que soit l'échelle à laquelle on travaille (un continent, un pays, une zone de quelques hectares), quel que soit l'objet biologique que l'on considère (l'espèce, la population, ou l'animal), l'étude de la sélection de l'habitat ne peut se faire qu'en considérant à la fois ces deux aspects. Dans la

mesure où l'espace géographique ne contient que deux dimensions, il semble plus logique de commencer par examiner la distribution spatiale des occurrences de l'objet biologique étudié (qu'il s'agisse de localisations d'animaux suivis par radio-pistage ou d'occurrences d'espèces), puis de déterminer, dans l'espace écologique de dimension plus élevée, les facteurs qui ont le plus d'influence sur la distribution observée. C'est à la fois par l'analyse des données et par le dialogue avec le biologiste que pourront être déterminées les parts relatives des contraintes spatiales et des exigences écologiques dans la distribution observée.

3.2.1.2 La construction d'un modèle

La population biologique est un système complexe, dont les éléments interagissent entre eux et avec leur environnement. Pour l'étudier, les biologistes collectent des données qui constituent une image de ce système. C (1992) souligne :

"*En collectant des données, et en les entrant dans l'ordinateur, on transforme l'objet d'étude, mais celui-ci reste un objet qui prend son sens dans une discipline biologique. La base de données prend alors le statut expérimental auquel on posera des questions partielles et précises.*"

Les données constituent une image floue et imprécise des processus qui agissent sur le système. Et l'analyse aura pour but d'améliorer la qualité de cette image, c'est-à-dire de déterminer quelle est cette information contenue dans les données et d'en supprimer les biais. Le point fondamental est ici que **l'analyse sera guidée par les données**. T *et al.* (1993) indiquent :

"*Faire de la biométrie, c'est étudier une partie du Monde Vivant pour répondre à une question. (...) Dans tous les cas, la biométrie s'appuie sur un ensemble de connaissances déjà connues ; elle essaie alors de préciser une question à laquelle elle est capable de répondre. Ensuite, elle construit une description plus ou moins schématique de la réalité afin de prévoir l'évolution du système étudié.*"

Cette représentation schématique de la réalité est appelée *modèle*. Un bon modèle doit être simple tout en permettant une bonne reproduction des observations. Il faut noter que, dans la mesure où ces modèles constituent seulement une approximation de la réalité, ils sont tous intrinsèquement faux. La qualité du modèle dépend donc du contexte dans lequel il a été construit et de son application. La qualité du modèle se mesure par la crédibilité (niveau de confiance subjectif du modèle) et par la qualification du modèle (domaine d'application du modèle G - et Z 2000).

Ainsi, T *et al.* (1993) soulignent : "*un modèle accepté n'est plus un instrument de connaissance, il devient outil opérationnel permettant de simuler un réel inaccessible (...). C'est un butoir à nos connaissances actuelles : son amélioration par contre est un moyen de les approfondir*".

3.3 L ’

La construction d'un modèle se fait au fur et à mesure de l'analyse, par l'identification des structures présentes dans les données. Pour y parvenir, le biométricien a besoin d'outils mathématiques pour identifier ces structures, et de programmes informatiques permettant de les utiliser.

3.3.1 Les outils mathématiques

Ordinairement, le choix des outils mathématiques par le biométricien se fait en fonction de ses préférences personnelles (T *et al.* 1993). Dans ce mémoire, nous nous basons principalement sur les méthodes factorielles multivariées pouvant être formulées dans le cadre du schéma de dualité (E 1987), d'une part parce que ces méthodes ont une grande importance au sein de l'école de Biométrie lyonnaise, mais aussi parce qu'elles ont déjà prouvé leur efficacité à de nombreuses reprises dans les études exploratoires (C 1992, D 2003).

Le schéma de dualité est une généralisation de l'analyse en composantes principales (ACP). Toutes les méthodes de cette famille reposent sur l'ACP d'un triplet de matrices. Le triplet est constitué par (i) un tableau à analyser, (ii) une matrice de pondération des colonnes de ce tableau et (iii) une matrice de pondération de ses lignes. L'ACP pondérée du tableau assigne des scores à ses lignes (composantes principales) et à ses colonnes (axes principaux). Cette analyse maximise la somme des carrés des scores des lignes (inertie), tout en assurant que les composantes principales successives retournées par l'analyse sont orthogonales (scores de lignes non-corrélés). Le tableau à analyser et les matrices de pondération dépendent de l'analyse qu'on souhaite appliquer ; ainsi, cette famille regroupe des méthodes géométriques simples, telles que l'ACP et l'analyse factorielle des correspondances, ou des méthodes plus complexes, telles que l'analyse discriminante et l'analyse de co-inertie (D 2003).

Les données auxquelles nous nous intéressons plus particulièrement dans ce mémoire sont les semis de points (*point pattern*). Les outils que nous utilisons ont pour but l'exploration des espaces géographique et écologique, pour identifier les processus qui ont généré les données.

3.3.1.1 L'analyse dans l'espace géographique

Des méthodes d'analyse des structures spatiales ont été utilisées dans de nombreux domaines scientifiques distincts tels que la géographie, la géologie ou la botanique (P *et al.* 2002). La géographie, où la question se pose le plus directement, répartit les méthodes d'analyse des structures spatiales en trois grandes catégories (L et F 1989) : l'analyse des semis de points, l'analyse des semis de lignes, et l'analyse des surfaces (étude de phénomènes spatialement continus ; une ou plusieurs variables sont mesurées à chaque point d'observation, et chaque point est supposé représenter la portion environnante de l'espace). Bien que les variables environnementales susceptibles d'expliquer la distribution des points dans l'espace géographique puissent être elles-mêmes structurées spatialement, nous ne nous concentrons ici

que sur l'analyse des semis de points. Nous renvoyons le lecteur à l'excellente synthèse effectuée par A (2000) pour plus de renseignements sur les deux autres catégories de méthodes.

De nombreux outils ont été développés dans la littérature pour permettre l'identification des structures des semis de points dans l'espace géographique (C et O 1981, D 1983, L et F 1989, L 1993, B et G 1995, D 1999). Le but de l'analyse des semis de points est de déterminer si la distribution géographique des points est aléatoire ou non et de décrire le type de pattern construit par les données. Il est ensuite possible d'inférer sur le type de processus qui a produit la structure observée. Nous décrivons ce type d'outils en détail dans le chapitre 4.

Nous nous posons également la question de la modélisation de plusieurs semis de points répartis sur la même zone. Ce problème se pose dans les études multi-spécifiques, ou lorsque plusieurs animaux sont suivis par radio-pistage et que les interactions entre les animaux suivis doivent être mises en évidence. Pour ce type de problèmes, les outils développés dans le domaine de l'analyse des semis de points sont rares (D 1999). Nous nous concentrons alors sur une approche multivariée reposant sur le schéma de dualité. Nous décrivons comment ces analyses peuvent être utilisées pour l'analyse de plusieurs semis de points au chapitre 5 et dans l'article que nous avons écrit pour la revue *Candollea* (C *et al.* 2006, Annexe 4).

3.3.1.2 L'analyse dans l'espace écologique

Nous nous concentrons ici sur la notion de niche écologique telle que l'a définie H (1957). En effet, c'est cette formalisation qui a vraiment permis le développement d'outils permettant l'exploration des exigences écologiques d'une espèce (M *et al.* 1992). G (1971) note : "*any statistical model of the Hutchinsonian niche must necessarily be a multivariate model*", une idée à laquelle M *et al.* (1992) souscrivent à 100% : "*The idea that numerous factors can influence whether an animal occupies a given area, a concept formally articulated by Hutchinson (1957), lends itself to mathematical analysis with multivariate statistical techniques*". L'utilisation des méthodes multivariées s'impose donc d'elle-même. Les méthodes qui reposent sur le schéma de dualité permettent la représentation graphique de la niche écologique dans un espace de dimension réduite choisi en fonction de critères biologiques, et sont donc idéales pour permettre son exploration.

Pourtant, l'utilisation de ce type de méthodes doit se faire dans une optique exploratoire. Il ne faut pas espérer que cette méthode construira à elle seule le modèle désiré. Ce n'est pas son rôle, comme le souligne M *et al.* (1992) : "*"In summary, multivariate methods such as PCA are designed to impart order to a set of data. (...) It is up to us to apply biological knowledge and common sense to the results*". Le biologiste reprend donc la place qui lui revient de droit, celle qui consiste à interpréter les résultats. Et c'est seulement à partir du dialogue entre le biométricien et le biologiste que peut se construire le modèle.

De même que pour l'aspect spatial, nous nous intéressons également au cas où plusieurs niches écologiques sont étudiées simultanément. Il peut par exemple s'agir de la position dans l'espace écologique de plusieurs animaux suivis par radio-pistage, ou de plusieurs espèces pré-

sentes simultanément sur la même zone. Nous utilisons alors le même type de méthodes, c'est-à-dire des méthodes multivariées reposant sur le schéma de dualité. Cette approche est décrite dans le chapitre 7.

3.3.2 Les outils informatiques

Nous avons décrit dans le chapitre 1 les qualités inestimables du logiciel R, *outil biométrique par excellence*. Ce langage très flexible est facile à utiliser, ce qui permet au biométricien de construire facilement et rapidement une très large diversité d'analyses. En outre, le grand nombre de fonctions graphiques disponibles le rend adéquat à l'analyse exploratoire des données écologiques.

3.3.2.1 les bibliothèques de fonctions existantes

De nombreuses bibliothèques de fonctions sont disponibles sous R pour analyser les aspects spatiaux et écologiques des données.

Plusieurs bibliothèques de fonctions sont disponibles pour analyser des semis de points. Ainsi, la bibliothèque **splancs** (R et D 1993) nous a été de la plus grande utilité, car les fonctions qu'elle contient permettent d'effectuer la majeure partie des analyses dont nous discutons dans le chapitre suivant. Nous nous sommes occasionnellement servis de la bibliothèque **spatstat** (B et T 2005) qui contient des fonctions permettant le même type d'analyses que **splancs**, mais qui permet en outre d'ajuster une plus vaste diversité de processus de points. Son architecture repose en effet sur la généralisation des *pairwise interaction processes* proposée par B et T (2000), une famille de processus de points qui peuvent être ajustés par la méthode du maximum de pseudo-vraisemblance. Notons qu'à l'heure où ce texte est écrit, un accord entre les auteurs de ces deux bibliothèques de fonctions a été conclu pour les fusionner en une bibliothèque nommée **Rasp** (R *et al.* 2003).

Pour l'analyse de données écologiques, l'école de Biométrie lyonnaise a développé une bibliothèque de fonctions appelée **ade4**, qui permet l'analyse exploratoire de données structurées dans un ou plusieurs espaces multidimensionnels (C *et al.* 2004). Cette bibliothèque rend accessible aux utilisateurs un grand nombre de méthodes reposant sur le schéma de dualité. Cette bibliothèque est l'outil idéal pour explorer la niche écologique d'une espèce.

3.3.2.2 La bibliothèque de fonctions **adehabitat**

Les données que nous voulons analyser sont des données à référence spatiale. Comme nous l'avons noté dans le paragraphe 1.2.2, ce sont les Systèmes d'Information Géographique (SIG) qui ont permis la récente explosion des études de la sélection de l'habitat. En effet, les informations sur les variables environnementales utilisées dans ce type d'études sont géoréférencées, ce qui permet un couplage avec les localisations des animaux, qu'elles aient été recueillies par

radio-pistage ou par d'autres modes d'échantillonnage. Or, si les SIG permettent cette association entre les deux types de données, ils ne possèdent en général que peu de fonctionnalités permettant une analyse statistique rigoureuse. D'autre part, comme le notent G et Z - (2000), "*most statistical packages cannot read GIS-maps directly, and the interchange files are generaly huge in size*". Les logiciels de statistiques, dont R, ne possèdent donc pas les fonctions d'un SIG. Les opérations spatiales courantes, telles que le calcul de pentes à partir d'une carte d'altitude, la jointure spatiale (i.e. déterminer la valeur d'une variable à une localisation donnée), la rastérisation de cartes vecteur ou le changement de systèmes de coordonnées ne sont pas disponibles dans ces programmes.

Il était donc nécessaire de pouvoir disposer d'une interface entre R et les SIG, et notamment Arcview (ESRI 1996), le plus utilisé des SIG en Ecologie (D 2003). J'ai donc programmé une bibliothèque de fonctions dans le but de fournir une telle interface (C soumis, cf. Annexe 12, 13, 14 et 15). Cette bibliothèque, nommée **adehabitat**, permet d'importer, d'exporter et de gérer des cartes raster sous R. Les cartes raster peuvent en outre être associées sous la forme de cartes *multicouches* ("*multi-layer maps*") constituées d'une seule et même grille de pixels prenant des valeurs pour différentes variables. Certaines fonctions d'**adehabitat** permettent la conversion de ces cartes multicouches vers des classes d'objet pouvant être gérées par le reste de l'environnement de R (la classe *data.frame*), et notamment par les fonctions de la bibliothèque **ade4**. Ceci facilite grandement les analyses de la niche dans l'espace écologique.

Cette bibliothèque offre en outre plusieurs fonctionnalités spatiales utiles à l'étude de la sélection de l'habitat. Il est possible de déterminer la composition de l'habitat à certaines localisations, de dénombrer les occurrences d'une espèce dans chaque pixel d'une carte raster, d'effectuer des opérations de rastérisation, ou au contraire de vectorisation, le calcul de zones tampon autour de points ou de lignes, etc. De nombreuses fonctions graphiques permettent l'exploration des données, tant au niveau spatial qu'au niveau écologique.

Enfin, **adehabitat** contient des fonctions plus spécifiques permettant par exemple l'estimation du domaine vital à partir de données de radio-pistage (méthode du noyau, polygone convexe minimum, etc.) ou l'application de certaines méthodes d'analyse de la sélection de l'habitat (analyse compositionelle, calcul de rapports de sélection, des distances de Mahalanobis, l'ENFA, etc.).

Cette bibliothèque, disponible sur le réseau qui distribue le logiciel R depuis septembre 2004, est accompagnée d'un didacticiel qui en facilite l'apprentissage (disponible en Annexe 13). **adehabitat** est un des principaux produits de ma thèse, qui permet de jouer à la fois sur les capacités de représentation des SIG et à la fois sur les potentialités d'analyses offertes par R.

Les outils de l’analyse dans l’espace géographique

Chapitre 4

L'analyse de semis de points d'un seul type

La première partie de ce mémoire a développé la démarche biométrique que nous avons adoptée. Nous nous concentrons maintenant sur les outils qui permettent l'application de cette démarche. Beaucoup d'outils utilisés au cours de ces trois années de thèse existaient déjà dans la littérature (la fonction $K(t)$ de Ripley, la méthode du noyau, etc.). D'autres ont dû être développés pour répondre à des questions soulevées par certains jeux de données (analyse K-select, analyse discriminante sur vecteurs propres de voisinage, l'analyse factorielle des rapports de sélection).

Dans ce chapitre, nous présentons quelques outils pour l'analyse des semis de points. Bien que ces outils soient couramment utilisés dans de nombreux domaines de l'Ecologie (C et O 1981, D 1983, C 1991, B et G 1995), ils sont encore peu utilisés dans les domaines relevant de la gestion de la faune sauvage ou de la Mammalogie (C - 2005). Cette méconnaissance justifie une introduction à ce champ de la statistique spatiale dans ce mémoire, bien que nous ne présentions dans ce chapitre que les outils qui nous ont été utiles en pratique pour analyser des données dans le cadre de cette thèse, c'est-à-dire une infime fraction des méthodes disponibles dans ce domaine.

L'objectif de ce chapitre est de décrire la démarche générale de l'analyse d'un semis de points, lorsque les points sont d'un seul type (e.g. occurrences d'une seule espèce). Ceci passe par l'utilisation de méthodes exploratoires permettant l'identification des structures du semis. Dans son célèbre article sur l'importance du concept d'échelle en Ecologie, L (1992) notait "*understanding patterns in terms of the processes that produce them is the essence of science, and is the key to the development of principles for management*". Nous décrivons donc également quelques solutions possibles pour modéliser les processus de points à l'origine du semis. Notre objectif n'est pas d'analyser des jeux de données pour en tirer des conclusions biologiques, mais de s'en servir comme matière à l'application des outils dont nous développons le principe. Une application concrète de certaines de ces méthodes peut être trouvée dans le chapitre 8.

4.1 B '

Les semis de points définissent les structures de données les plus simples de la statistique spatiale. En effet, ces données ne sont constituées que des coordonnées d'un ensemble de points sur la surface échantillonnée. Cependant, que la structure de ces données soit simple ne veut pas dire pour autant que les méthodes utilisées pour les analyser le soient aussi. B et G (1995) soulignent en effet que les méthodes utilisées pour analyser ces semis sont souvent plus complexes que celles qu'on peut rencontrer dans les autres champs de la statistique spatiale. Bien que ce type d'analyse soit rarement utilisé dans le domaine de l'Ecologie animale, beaucoup d'études de la sélection de l'habitat par la faune sauvage reposent sur des données qui peuvent être exprimées sous la forme d'un semis de points (P 1993, S 2000, S - *et al.* 2003, B 2004, C 2005).

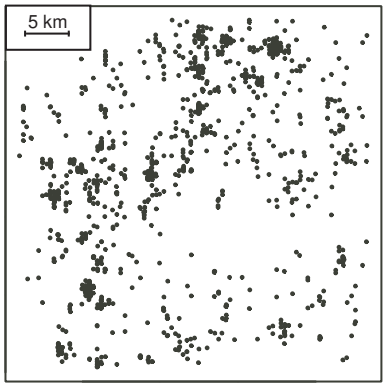
Mais avant de décrire les outils qui permettent d'identifier les structures des semis, nous devons définir quelques concepts et notations mathématiques utilisées tout au long de ce chapitre. Les points qui constituent le semis sont appelés *événements*, afin de ne pas les confondre avec d'autres points arbitraires du plan utilisés pour étudier les propriétés du semis. Un *processus de points* est un mécanisme stochastique qui génère un ensemble d'événements $\mathbf{x}_i$ (chaque vecteur $\mathbf{x}_i$ étant un vecteur de longueur 2 contenant les coordonnées x et y de l'événement), dénombrables sur une zone A du plan $\mathbb{R}^2$. Un *semis de points* $\mathbf{X}$ correspond à une réalisation de ce processus. $\mathbf{X}$ correspond alors à un ensemble de N événements sur A (D 1983, C 1991, B et G 1995, D 1999).

Nous utilisons dans la première partie de ce chapitre un jeu de données décrivant la distribution spatiale d'indices de présence du lynx (*Lynx lynx*) dans le massif du Jura (B 2004). Ce semis est constitué de 1319 indices de présence récoltés de 1980 à 1999 sur une zone de 216 000 hectares localisée dans le Jura Français. Ces indices de présence peuvent être directs (individus observés ou capturés) ou indirects (indices d'attaques sur troupeaux, traces, poils, cadavre de lynx, féces, indices d'attaques sur proies sauvages). La distribution de ces indices sur la zone d'étude est présentée sur la figure 5. Ce jeu de données est actuellement analysé par Mathieu B (Laboratoire de Biométrie, Université Lyon 1). Il a été choisi pour illustrer les outils présentés dans ce chapitre car : (i) il est disponible dans la bibliothèque de fonctions **adehabitat**, ce qui permet au lecteur de reproduire les analyses s'il le désire (cf. Annexe 10), (ii) la forme de la zone d'étude est très simple (rectangulaire), donc facile à gérer d'un point de vue informatique, (iii) Les événements du semis sont très fortement structurés. Le jeu de données doit être pris pour ce qu'il est : un jeu d'illustration.

4.2 A '

4.2.1 Outils graphiques communément utilisés

Une première exploration visuelle des cartes du semis permet déjà de se faire une idée de la structure des données. Très souvent, cette étape conduit à émettre des hypothèses sur les pro-



F . 5 – *Distribution des 1319 indices de présence du Lynx dans le Jura (*B *2004)*

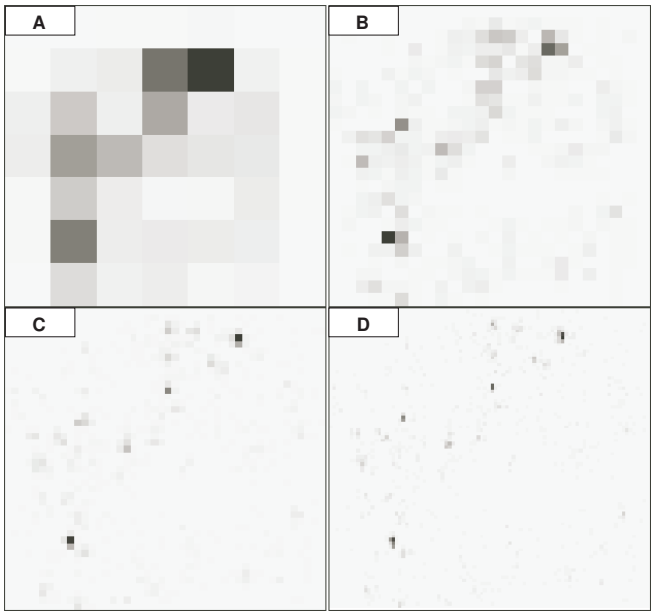
cessus qui ont généré les données. Le semis est-il aléatoire, agrégatif (événements regroupés en "clusters"), ou régulier (événements plus éloignés les uns des autres que sous l'hypothèse aléatoire) ? De tels graphiques sont appelés "*cartes ponctuelles*" (*dot map*) par B et G (1995). Ainsi, la carte ponctuelle des indices de présence du lynx nous renseigne immédiatement sur le caractère agrégatif du semis (figure 5).

Cependant, dans de nombreux cas, le caractère agrégatif du semis n'est pas aussi net. D (1983) souligne que l'oeil humain est un très mauvais outil pour détecter les structures d'un semis de points (cf. figure 13A pour un exemple de semis aléatoire), et insiste sur la nécessité d'utiliser d'autres outils graphiques. Ainsi, une autre méthode consiste à placer une grille sur la zone d'étude, et à compter le nombre d'événements tombant dans chaque quadrat. La carte ainsi obtenue permet de se faire une idée des variations de densité sur la zone étudiée, et d'identifier les clusters d'événements. Mais l'efficacité de cette méthode va dépendre fortement de la maille de la grille utilisée. Si la maille est trop large, toutes les structures fines du semis seront perdues. Inversement, si elle est trop fine, beaucoup de quadrats seront vides, et cette stratégie n'apportera pas grand-chose de plus à l'analyse exploratoire que la carte ponctuelle (figure 6).

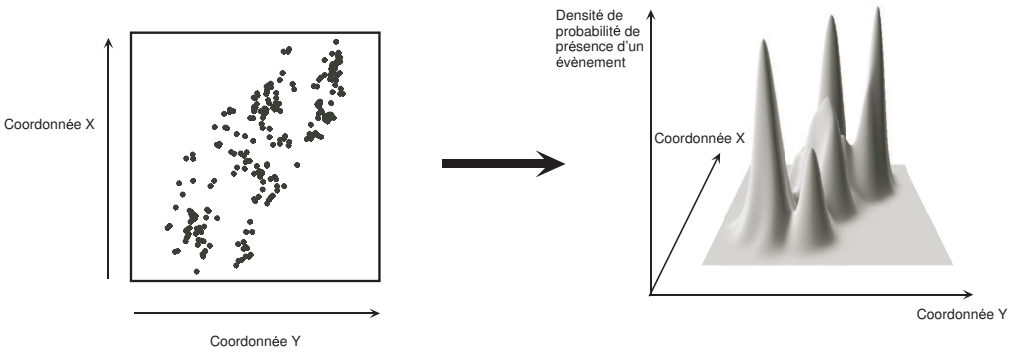
4.2.2 Méthode du noyau

La méthode du noyau (*kernel method*) est une autre méthode graphique très souvent recommandée en première approche, car elle évite les problèmes liés à la méthode des quadrats, tout en apportant une information visuelle nette sur la structure du semis (S 1986). Elle lisse le semis de points, et permet l'identification des zones de plus fortes densités en événements au premier coup d'œil (D 1983, B et G 1995).

Cette méthode repose sur l'idée que le semis de points est la réalisation d'un processus inconnu qui peut être décrit par une fonction donnant la densité de probabilité de présence d'un événement en tout point de la zone d'étude. Elle permet une estimation non paramétrique de cette fonction (objectif décrit sur un semis de points fictif sur la figure 7), dont l'exploration visuelle permet de suggérer des modèles paramétriques pour les processus qui ont généré les



F . 6 – *Dénombrement des indices de présence de lynx dans des grilles de différentes mailles : (A) quadrats de 8000 m de côté ; (B) quadrats de 2000 m de côté ; (C) quadrats de 1000 m de côté ; (D) quadrats de 500 m de côté.*



F . 7 – *Objectif de la méthode du noyau : ici, à partir d'un semis de points fictif, la méthode estime la fonction de densité de probabilité du processus qui l'a généré.*

données.

Etant donnée sa place centrale dans l'exploration de la structure des semis de points, nous avons effectué une étude détaillée de cette méthode dans le cadre de ma thèse qui s'est concrétisée par la rédaction d'un article en introduisant le principe général dans le domaine de la Biogéographie. Cet article a été écrit en collaboration avec les Conservatoires et Jardins Botaniques de Genève, et soumis pour publication dans la revue *Applied Vegetation Science* (cf. Annexe 7).

Notons que la méthode du noyau est très étudiée dans le domaine de l'analyse des données de radio-pistage, car elle est souvent utilisée pour estimer le domaine vital des animaux (W - 1989, 1995b, S *et al.* 1998, 1999). La fonction de densité de probabilité de présence de l'animal estimée sur les localisations de l'animal obtenues par radio-pistage est appelée *Distribution d'Utilisation*. Le *domaine vital* est alors défini comme la surface incluse par le contour de la distribution d'utilisation tel que le volume compris sous la distribution et à l'intérieur du contour représente un certain pourcentage du volume total. Ce pourcentage est habituellement fixé à 95% (W 1993), ce qui permet l'estimation de la plus petite surface à l'intérieur de laquelle l'animal a 95% de chances d'être présent.

4.2.2.1 Principe de la méthode

La fonction de densité de probabilité de présence des événements est la fonction f qui permet de calculer la probabilité $P(B)$ qu'un événement soit détecté dans une sous-région B, grâce à l'équation :

$$P(B) = \int_B f(\mathbf{u}) d\mathbf{u}$$

où $\mathbf{u}$ est un vecteur de longueur 2 donnant les coordonnées d'un point arbitraire sur le plan. Autrement dit, la probabilité de présence d'un événement sur B est calculée comme le volume compris sous la surface de la distribution et à l'intérieur des limites de la région B.

Par définition, une fonction de densité de probabilité bivariée doit respecter les deux propriétés suivantes :

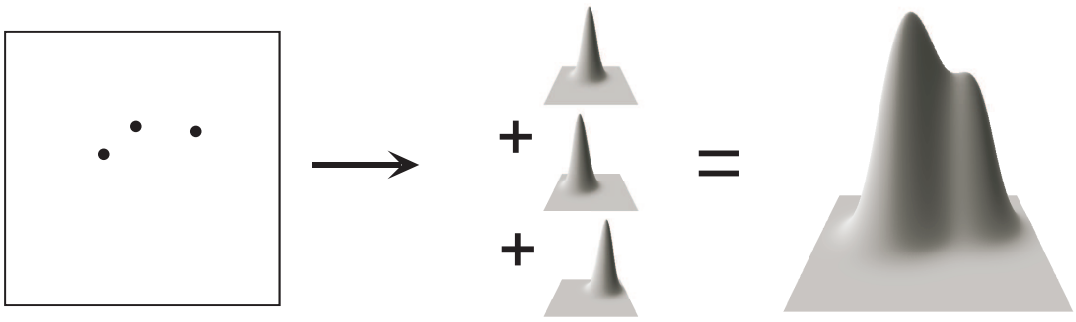
$$\int_{\mathbb{R}^2} f(\mathbf{u}) \mathrm{d}\mathbf{u} = 1 \tag{4.1}$$

et

$$f(\mathbf{u}) > 0 \text{ pour tout } \mathbf{u} \tag{4.2}$$

La méthode du noyau permet l'estimation non-paramétrique de cette fonction, à partir d'un ensemble $\mathbf{X}$ d'événements de coordonnées $\mathbf{x}_i$ ($i = 1, ..., N$). Le principe de la méthode est décrit dans la figure 8 à l'aide d'un exemple très simple constitué d'un ensemble de seulement trois événements. Une fonction de densité de probabilité unimodale et radialement symétrique, c'est-à-dire une fonction en forme de "cloche", est placée au dessus de chaque événement. Cette fonction de densité de probabilité, appelée *fonction de noyau* (*kernel function*), peut être n'importe quelle fonction qui respecte les deux conditions 4.1 et 4.2 décrites plus haut. Un choix courant est la fonction de densité de probabilité normale bivariée standard :

$$K(\mathbf{u}) = \frac{1}{2\pi} \exp\left(\frac{1}{2}\mathbf{u}^t\mathbf{u}\right)$$



F . 8 – *Principe du lissage d'un semis de points par la méthode du noyau. La fonction de densité de probabilité de présence d'un événement est calculée à partir de la somme des fonctions de noyaux placées au dessus de chaque événement observé.*

où $\mathbf{u}^t$ indique la transposée de **u**. La somme de toutes ces fonctions sur tous les événements permet l'estimation de la fonction de densité de probabilité de présence d'un événement en tout point **u** de $\mathbb{R}^2$. Exprimée de façon plus formelle, la fonction de densité de probabilité recherchée est estimée par l'équation :

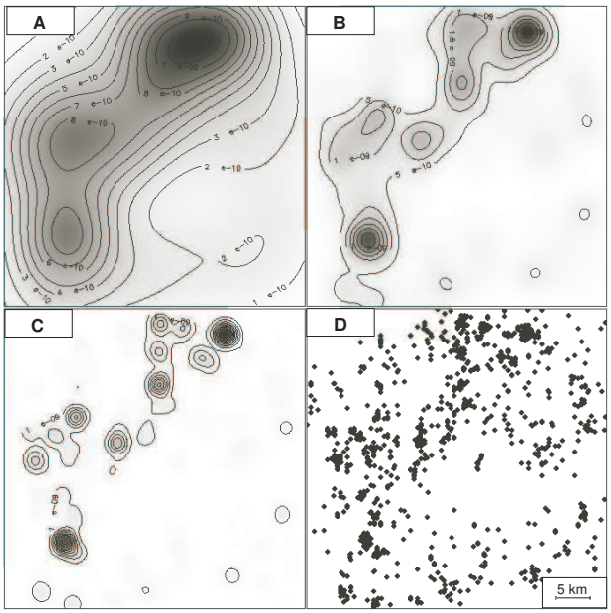
$$\hat{f}(\mathbf{u}) = \frac{1}{Nh^2} \sum_{i=1}^{N} K\left\{\frac{1}{h}(\mathbf{u} - \mathbf{x}_i)\right\} \tag{4.3}$$

où la fonction $K(\mathbf{x})$ est la fonction de noyau utilisée, et h est la valeur de l'écart-type de la fonction de noyau. Le paramètre h est donc un paramètre de lissage à déterminer par l'utilisateur qui contrôle la "largeur" de la fonction de noyau placée au dessus de chaque événement, et par voie de conséquence, la quantité de lissage que doit subir le semis.

4.2.2.2 Utilisation en pratique

En pratique, une grille de pixels est placée sur la zone d'étude et, pour une valeur de h donnée, la valeur de l'estimation de la fonction de densité de probabilité de présence d'un événement est calculée pour chaque pixel $\mathbf{u}_j$ grâce à l'équation 4.3. La maille de la grille utilisée n'a pas de réelle influence sur les estimations (S 1986, W 1989).

En revanche, la forme de la distribution obtenue est très fortement influencée par la valeur choisie pour le paramètre h. Trop grand, ce paramètre peut conduire à des estimations qui cachent les structures fines du semis ; trop petit, il tend à laisser trop de bruit dans les données et obscurcir l'information à grande échelle (figure 9). De nombreuses méthodes ont été proposées pour choisir le paramètre de lissage adéquat (e.g. *Least square cross validation*, cf. Annexe 7),



F . *9 – Lissage par la méthode du noyau de la distribution des indices de présence du lynx dans le Jura, avec différents paramètres de lissage : (A)* h *= 5000 m ; (B)* h *= 2000 m ; (C)* h *= 1000 m ; (D) Distribution des événements avant lissage.*

mais S (1986) recommande, lorsque cette méthode est utilisée à des fins exploratoires, de construire des cartes lissées du semis avec différentes valeurs pour le paramètre de lissage, afin de mettre en évidence les structures rencontrées à différentes échelles spatiales.

De nombreuses fonctions du logiciel R permettent d'appliquer la méthode du noyau. Ainsi, la bibliothèque **adehabitat** contient la fonction `kernelUD()`, avec laquelle la figure 9 a été réalisée. Mais des fonctions sont disponibles dans d'autres bibliothèques, telles que `kde2d()` de la bibliothèque **MASS**, `kernel2d()` de la bibliothèque **splancs**, `ksmooth.ppp()` de la bibliothèque **spatstat**, toutes les fonctions de la bibliothèque **KernSmooth**, etc. Chacune prend place dans le cadre théorique spécifique à chaque auteur de bibliothèque.

Notons que la figure 9 illustre bien les problèmes qui peuvent se poser au biométricien lors de l'analyse exploratoire des données : à quoi correspondent les noyaux avec une densité plus forte mis en évidence sur la zone ? Après examen des données, ces deux noyaux correspondent uniquement à des attaques de lynx sur troupeaux. Ces attaques étant indemnisées par l'Etat, lorsqu'elles se produisent, il y a de fortes chances pour qu'elles soient reportées par les éleveurs. En revanche, les attaques sur la faune sauvage ont moins de chances d'être reportées. La pression d'échantillonnage est plus forte à l'endroit où se trouvent les troupeaux. La distribution des indices de présence reflète autant la distribution de la pression d'échantillonnage que celle du lynx. Comme le souligne B (2004), il n'est sans doute pas judicieux de regrouper au sein d'un même jeu de données les indices de présence "domestiques" et "sauvages".

4.2.3 Mesure du caractère aléatoire du semis

Le processus de points décrivant la distribution aléatoire des événements dans l'espace est appelé *Complete Spatial Randomness* (CSR, D 1983, C 1991, B et G 1995). La CSR est le plus simple des modèles de processus de points (§ 4.3.3.1). Ce processus suppose que : (i) le nombre d'événements dans une région A de surface $|A|$ suit une loi de Poisson de paramètre $\lambda|A|$; (ii) sachant N événements $\mathbf{x}_i$ dans la région A, les $\mathbf{x}_i$ sont un échantillon aléatoire et indépendant de la loi uniforme sur A.

Comme nous l'avons montré dans le chapitre 3, les organismes ne sont jamais distribués de façon aléatoire dans leur environnement. Mais le test de la CSR est un point de départ aux analyses : "*Our interest in CSR is that it represents an idealized standard which, if strictly unattainable in practice may nevertheless be tenable as a convenient first approximation*" (D 1983). C'est seulement si les données permettent de rejeter la CSR que l'on peut se poser la question de la façon dont le semis en diffère.

Tous les ouvrages de référence traitant des processus de points présentent les mêmes outils pour effectuer ce test (D 1983, C 1991, B et G 1995, D 1999). La plupart reposent généralement sur des mesures de distances entre les événements, ou entre les événements et des points choisis aléatoirement sur la zone d'étude. Nous décrivons les plus courants dans cette partie (dont nous nous servons dans le chapitre 8). S'il y a rejet de la CSR, une étude plus en détail permet de déterminer le type de processus à l'origine des données.

4.2.3.1 Distance à l'événement le plus proche (fonction G)

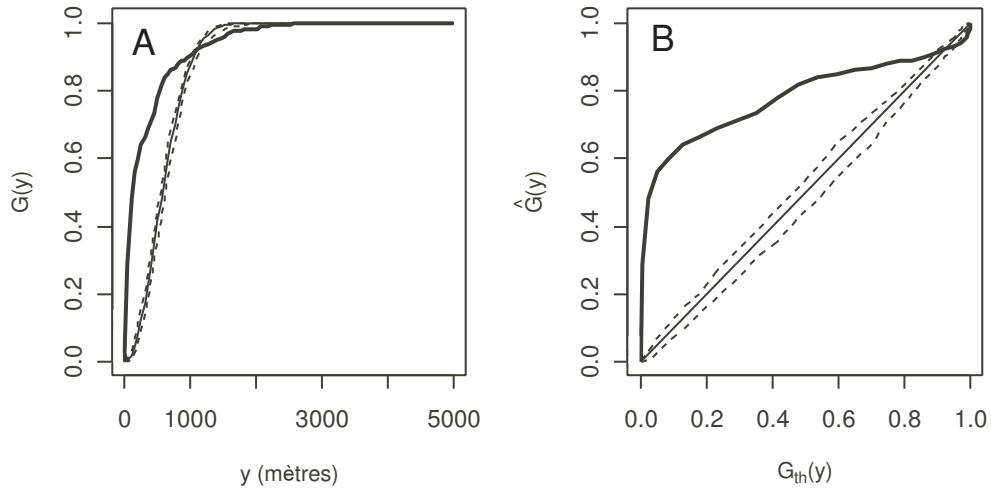
La distance y_i entre un événement et son plus proche voisin peut être utilisée comme mesure du caractère aléatoire du semis. La fonction $\hat{G}_{obs}(y)$ est définie par l'équation :

$$\hat{G}_{obs}(y) = N^{-1}\#(y_i \leq y)$$

où #(.) mesure un nombre de cas. La fonction $\hat{G}(y)$ mesure la proportion d'événements pour lesquels la distance au plus proche voisin est inférieure ou égale à la valeur de distance y. Des enveloppes de confiance peuvent être construites autour de cette courbe grâce à des simulations de la CSR. A l'étape k du processus de simulation, N points sont choisis aléatoirement sur la zone d'étude, et la fonction $G^*_k(y)$ est calculée pour le semis simulé. On peut alors calculer la fonction théorique $\hat{G}_{th}(y)$ comme la moyenne des s simulations effectuées.

Une représentation graphique peut alors être effectuée pour mettre en évidence les structures du semis (D 1983). Celle-ci consiste à comparer la distribution observée des distances à l'événement le plus proche avec la distribution théorique sous l'hypothèse de la CSR. Ces graphiques sont appelés "*Empirical Distribution Function plot*" (que nous traduirons par "courbe EDF").

La courbe EDF de la fonction $G(y)$ estimée sur les indices de présence du lynx est présentée figure 10. Il y a plus de couples d'événements voisins séparés par une distance comprise entre



F . *10 – Calcul de la fonction G(.) sur le semis des indices de présence du lynx (figure 5). (A) Valeurs observées et théoriques (sous l'hypothèse de la CSR) de la fonction G(y) en fonction de la distance y. (B) Valeur observée de $\hat{G}_{obs}(y)$ en fonction de sa valeur théorique $G_{th}(y)$ sous l'hypothèse de la CSR. Sur les deux graphiques, la courbe épaisse correspond à la courbe observée, la courbe fine correspond à la courbe théorique, et les courbes en pointillés indiquent les limites inférieures et supérieures de 100 simulations de la CSR.*

0 et 1000 mètres que sous l'hypothèse aléatoire. Inversement, on trouve moins d'événements voisins séparés par une distance de 1000 à 1500 mètres qu'attendus sous l'hypothèse aléatoire. Ce résultat est causé par le caractère très agrégé du semis.

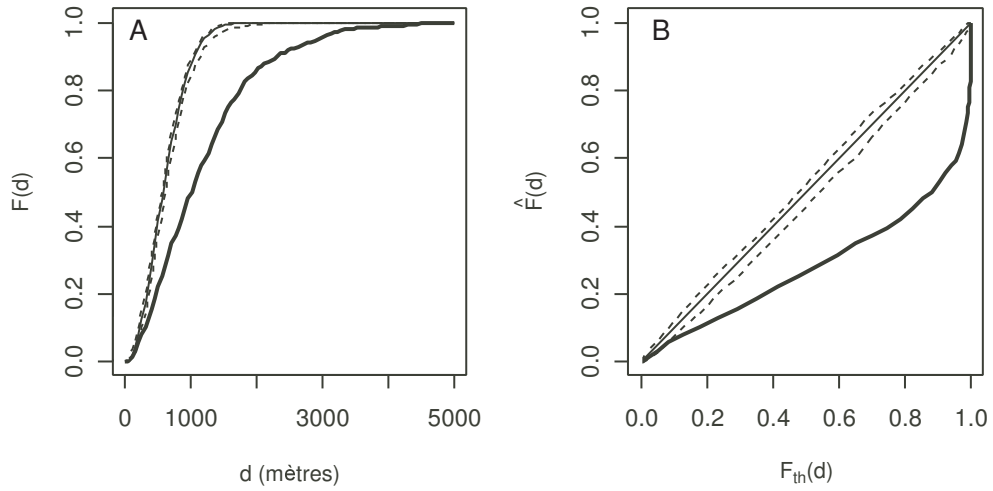
4.2.3.2 Distance entre un point arbitraire et l'événement le plus proche (Fonction F)

Une autre fonction couramment utilisée pour mesurer le caractère aléatoire d'un semis de points mesure la distance entre des points tirés aléatoirement sur la zone d'étude et les événements les plus proches. En pratique, une grille est superposée au semis de points, et pour chaque nœud i de la grille, on mesure la distance d_i à l'événement le plus proche. Puis on calcule la fonction $\hat{F}_{obs}(d)$ par l'équation :

$$\hat{F}_{obs}(d) = m^{-1}\#(d_i \leq d)$$

où m est le nombre de nœuds de la grille. Comme pour la fonction précédente, l'examen de la distribution et de la courbe EDF permet de déterminer en quoi le semis diffère de la CSR. D (1983) indique que la fonction $\hat{F}(d)$ est une mesure des "espaces vides" dans le semis : en effet, la fonction $1 - \hat{F}(d)$ est une estimation de la proportion de tous les points aléatoires localisés à une distance d'au moins d de chacun des événements du semis.

La courbe EDF de la fonction $F(d)$ estimée sur le semis des indices de lynx est présentée figure 11. On trouve plus de couples d'événements voisins séparés par une distance comprise entre 0 et 1000 mètres que sous l'hypothèse aléatoire. A toutes les échelles, la distance d ob-



F . *11 – Calcul de la fonction $F(.)$ sur le semis des indices de présence du lynx (figure 5). (A) Valeurs observées et théoriques (sous l'hypothèse de la CSR) de la fonction $F(d)$ en fonction de la distance d. (B) Valeur observée de $\hat{F}_{obs}(d)$ en fonction de sa valeur théorique $F_{th}(d)$ sous l'hypothèse de la CSR. Sur les deux graphiques, la courbe épaisse correspond à la courbe observée, la courbe fine correspond à la courbe théorique, et les courbes en pointillés indiquent les limites inférieures et supérieures de 100 simulations de la CSR.*

servée entre un point aléatoire et l'événement le plus proche est plus faible qu'attendu sous l'hypothèse aléatoire, ce qui souligne le caractère très agrégatif du semis.

4.2.4 Les tests de la CSR

Les méthodes que nous avons présentées jusqu'ici permettent une exploration des propriétés du semis afin de déterminer s'il est aléatoire et dans le cas contraire d'en identifier les structures. Toutefois, un test est parfois requis afin de confirmer les impressions produites par les graphiques.

Plusieurs solutions ont été proposées pour tester la CSR, et la majorité d'entre elles est liée aux méthodes que nous avons décrites précédemment. Nous en présentons deux ci-dessous :

– Un premier test repose sur la méthode des quadrats que nous avons décrite paragraphe 4.2.1 (D 1983). Le nombre d'événements présents dans chaque quadrat d'une grille superposée au semis peut être utilisé pour tester la CSR, par exemple par un test du χ^2. Si la grille est constituée de m quadrats, on utilise la propriété indiquant que sous l'hypothèse de la CSR, la distribution des N événements sur la zone devrait être un échantillon de la loi uniforme, c'est-à-dire que le nombre d'événements dans chaque quadrat devrait être en moyenne égal à $\overline{n} = N/m$. Le test est alors réalisé grâce à l'équation :

$$\chi^2 = \sum_{i=1}^{m} (n_i - \overline{n})^2 / \overline{n}$$

Notons également que sous l'hypothèse de la CSR, le nombre d'événements sur une surface A suit une loi de Poisson de paramètre $\lambda|A|$. Or, cette loi est caractérisée par une moyenne égale à sa variance. Ainsi, une mesure autrefois communément utilisée pour mesurer le caractère agrégé du semis est le rapport variance / moyenne du nombre d'événements présents dans chaque quadrat (D 1999, D *et al.* 2002). Or, la statistique du χ^2 est égale à ce rapport multiplié par $m - 1$ (D 1983).

– Un second test, recommandé par R et D (1993), repose sur l'observation que sous l'hypothèse de la CSR, la fonction F devrait être en moyenne égale à la fonction G (B et G 1995). Ce test est également un test de Monte Carlo basé sur s simulations de la CSR. Le critère utilisé proposé par R et D (1993) est *la valeur maximale de la différence en valeur absolue entre les fonctions F et G*. Ce critère est calculé sur le semis observé, puis sur s simulations de la CSR, et la comparaison de l'observation à la distribution des simulations permet de calculer la probabilité pour que la valeur observée soit le résultat d'une réalisation de la CSR.

4.3 M ´

4.3.1 Définitions

Nous avons présenté dans le paragraphe précédent quelques outils graphiques permettant d'explorer le semis étudié et de déterminer si la CSR est un bon modèle pour expliquer la distribution des événements. Ces outils basiques permettent également de déterminer en quoi et à quelle(s) échelle(s) les processus que l'on cherche à modéliser structurent les données.

Mais avant de décrire la démarche de modélisation, nous devons définir un certain nombre de notions clefs dans l'analyse des processus de points (D 1983, C 1991, B et G 1995, D 1999). Ces processus sont habituellement décrits par deux propriétés (D 1983) :

– L'*intensité de premier ordre*, qui mesure l'intensité du processus en tout point $\mathbf{u}$, est définie par l'équation :

$$\lambda(\mathbf{u}) = \lim_{|d\mathbf{u}| \to 0} \left\{ \frac{E\left[N(d\mathbf{u})\right]}{|d\mathbf{u}|} \right\}$$

– L'*intensité de second ordre* du processus, qui décrit les interactions entre les événements du semis en deux points $\mathbf{u}$ et $\mathbf{v}$, est définie par l'équation :

$$\lambda_2(\mathbf{u}, \mathbf{v}) = \lim_{|\mathrm{d}\mathbf{u}|, |\mathrm{d}\mathbf{v}| \to 0} \left\{ \frac{E\left[N(\mathrm{d}\mathbf{u}) N(\mathrm{d}\mathbf{v})\right]}{|d\mathbf{u}|\,|d\mathbf{v}|} \right\}$$

Plusieurs hypothèses sont en général supposées pour faciliter la modélisation du processus à l'origine du semis :

- Un processus est dit *stationnaire* si toutes les déclarations probabilistes sur le processus dans la région A sont invariantes en translation : si on fait glisser les limites de la zone d'étude, les propriétés du processus restent les mêmes.

- Un processus est dit *isotrope* si on observe également une invariance en rotation de ses propriétés.

Lorsque le processus est stationnaire et isotrope, alors $\lambda(\mathbf{u})$ prend une valeur constante λ, qui mesure la densité du processus. Notons que lorsque le processus n'est pas stationnaire, la méthode du noyau peut être utilisée pour obtenir des estimateurs non paramétriques pour $\lambda(\mathbf{u})$ § 4.3.3.2). De façon similaire, pour un processus stationnaire et isotrope, l'intensité de second ordre du processus ne dépend que de la distance t entre deux événements localisés aux points $\mathbf{u}$ et $\mathbf{v}$. Ainsi, $\lambda_2(\mathbf{u}, \mathbf{v})$ se simplifie en $\lambda_2(t)$. Les hypothèses de stationnarité et d'isotropie sont souvent formulées pour simplifier la modélisation des processus. Ces hypothèses peuvent être violées si un pattern systématique, par exemple un gradient, est présent dans les données. Mais ces hypothèses sont le plus souvent raisonnables comme première approximation, au moins sur une étendue géographique restreinte (D 1983).

C'est donc en se basant sur ces deux propriétés que l'on peut modéliser le processus à l'origine des données. Mais si l'intensité de premier ordre est facile à estimer dans le cas de processus stationnaires et isotropes – c'est la densité en événements sur la région étudiée – il est plus difficile de mesurer la seconde, d'autant que celle-ci varie en fonction de la distance entre deux événements.

4.3.2 La fonction $K(t)$ de Ripley

4.3.2.1 Présentation et propriétés

La fonction $K(t)$, développée par R (1977), permet de mesurer l'intensité de second ordre pour les processus stationnaires et isotropes. Pour cette raison, cette fonction connaît un succès considérable dans l'analyse des semis de points en Ecologie, et notamment en phytobiologie (H 1995, H *et al.* 1996, 1997, C et S 1999, W *et al.* 2003). On la rencontre aussi de plus en plus souvent dans les études sur la faune sauvage (R *et al.* 2000, L *et al.* 2003, L et D 2004). Elle est une des rares méthodes utilisées en Ecologie pour prendre en compte spécifiquement le caractère spatial du semis.

Nous décrivons ici le calcul et les propriétés de cette fonction. Soit $N(t)$ l'espérance du nombre d'événements localisés à une distance inférieure ou égale à t d'un événement arbitraire du semis. On calcule alors la fonction $K(t)$ grâce à l'équation :

$$K(t) = \lambda^{-1} N(t) \tag{4.4}$$

La fonction de R est donc une standardisation de $N(t)$ qui permet de comparer des semis ayant des densités différentes. Il existe un lien entre cette fonction et l'intensité de second ordre $\lambda_2(t)$ (D 1983), exprimé par l'équation :

$$K(t) = \lambda^{-2} \int_0^t \lambda_2(y) 2\pi y \mathrm{d}y$$

Chaque fois qu'un auteur développe un nouveau type de modèle de processus de points, celui-ci est caractérisé par une expression formelle de ses intensités de premier et de second ordres. Lorsque le processus est stationnaire, la densité en événements sur la zone suffit pour estimer l'intensité de premier ordre, et c'est la fonction $K(t)$ qui est utilisée pour mesurer l'intensité de second ordre. L'expression formelle de cette fonction pour un processus donné permet d'ajuster le modèle au semis étudié grâce à la méthode des moindres carrés, et donc de déterminer les paramètres inconnus du modèle. Cette approche est utilisée dans le paragraphe 4.3.3.3 pour le processus de Neyman-Scott.

4.3.2.2 La fonction $K(t)$ comme méthode exploratoire

Du fait des relations entre la fonction $K(t)$ et l'intensité de second ordre du processus, celle-ci est le plus souvent utilisée à des fins d'exploration du semis dans la littérature (H 1995, H *et al.* 1996, 1997, G et P 2000, D *et al.* 2001, K 2001, L *et al.* 2003). En effet, sous l'hypothèse de la CSR, le nombre d'événements attendus dans un rayon t autour d'un événement arbitraire est égal à $\lambda\pi t^2$. Donc, sous cette hypothèse, $K(t) = \pi t^2$.

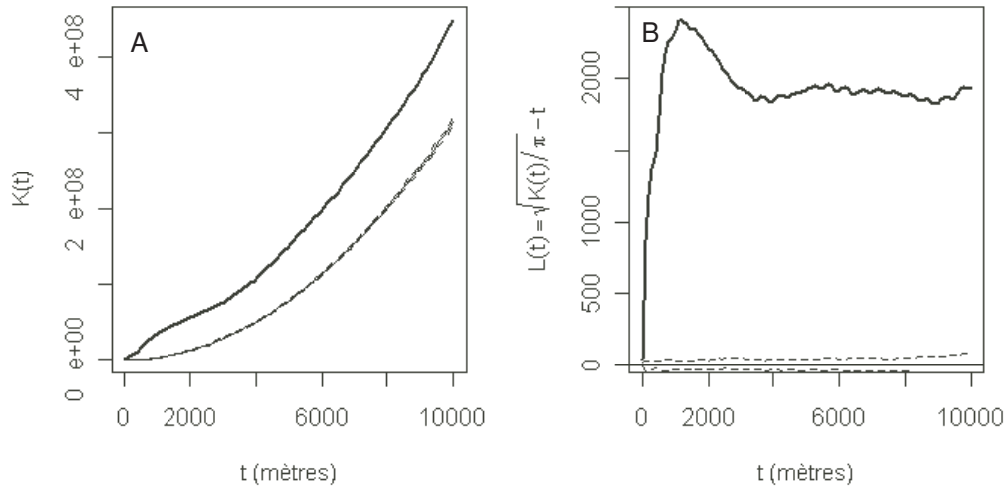
En conséquence, pour un processus agrégatif, la fonction $K(t)$ est supérieure à πt^2, et inversement pour un processus régulier. Pour cette raison, de nombreux auteurs préfèrent travailler avec sa forme linéarisée $L(t)$ définie par l'équation :

$$L(t) = \sqrt{K(t)/\pi} - t$$

Sous l'hypothèse de la CSR, $L(t) = 0$ quel que soit t. La fonction $L(t)$ est donc plus facile à interpréter. Grâce à elle, il est possible non seulement de déterminer si les événements du semis sont distribués de façon aléatoire, mais dans le cas contraire à quelle(s) échelle(s) se situent les structures. Pour ce, la CSR est simulée un grand nombre de fois, et pour chacun des semis simulés, une fonction $L^*(t)$ est calculée. On peut déterminer si le semis observé est aléatoire en traçant le graphique $\hat{L}(t)$ en fonction de t, puis en ajoutant les limites supérieures et inférieures de ces fonctions calculées sous l'hypothèse de la CSR.

4.3.2.3 Utilisation en pratique

En pratique, la fonction $K(t)$ est estimée de la façon suivante : pour une valeur de t donnée, les événements localisés dans un cercle de rayon t centré sur une événement $\mathbf{x}_i$ sont dénom-



F . *12 – Exemple de calcul de la fonction $K(t)$ sous sa forme classique (A) et sous sa forme linéarisée $L(t)$ (B) sur la distribution des indices de présence du lynx (figure 5). Sur les deux graphiques, la courbe épaisse correspond à la courbe observée, la courbe fine correspond à la courbe théorique, et les courbes en pointillés indiquent les limites inférieures et supérieures de 100 simulations de la CSR.*

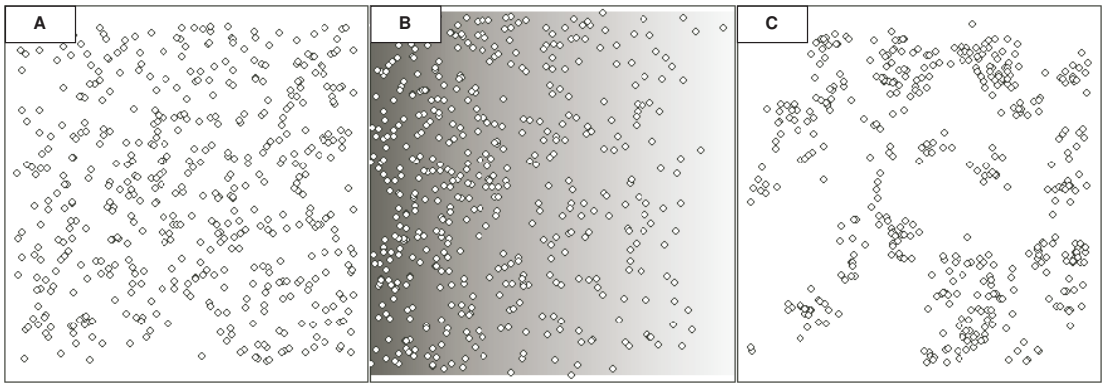
brés. On peut alors estimer $\hat{N}(t)$ comme la moyenne de ces nombres, et en déduire $\hat{K}(t)$ grâce à l'équation 4.4, en prenant comme estimateur de λ la densité du semis.

Notons que les effets de bord peuvent avoir un effet important sur la valeur de l'estimation de la fonction $K(t)$. Pour un événement $\mathbf{x}_i$ localisé à une distance inférieure à t de la bordure de la zone d'étude, le nombre d'événements localisés dans un cercle de rayon t centré sur $\mathbf{x}_i$ sera sous-estimé, car il ne comptabilise pas les événements localisés dans la zone du cercle située en dehors de la zone d'étude (D 1983, G et P´ 1999, H 1995, G *et al.* 1999, W et M 2004). Plusieurs méthodes de correction de ces effets de bord ont été développées dans la littérature, mais nous n'en discuterons pas ici. Remarquons seulement que la fonction K est fréquemment utilisée pour comparer le semis observé aux semis générés par un processus donné, le plus souvent par la CSR. Même si le biais est présent dans les données lorsque les effets de bord ne sont pas corrigés, le même biais est présent dans les simulations du processus que l'on cherche à ajuster, et les effets de bord ne posent pas vraiment de problèmes (L *et al.* 2003).

Dans le cas de l'étude sur le semis des indices de lynx, le caractère très agrégatif du semis de points apparaît de façon nette figure 12, en particulier à 1500 mètres, ce qui correspond à peu près au rayon des "noyaux" mis en évidence à l'aide de la méthode du noyau (figure 9).

4.3.3 Quelques modèles utilisés

Toutes les méthodes décrites précédemment sont essentiellement exploratoires et sont utilisées afin de suggérer des modèles pour les processus qui ont généré les données. Ces processus



F . *13 – Réalisation de trois types de processus de points sur une zone carrée de longueur unité : (A) le processus de Poisson, utilisé pour simuler la CSR ($\lambda = 500$) ; (B) le processus inhomogène de Poisson, un processus non stationnaire pour lequel l'intensité est variable sur la zone (ici, un gradient) ; (C) Un processus de Neyman-Scott, utilisé pour générer des semis agrégés (ici, 50 "parents" sont distribués aléatoirement sur la zone, lesquels génèrent à leur tour un nombre aléatoire de "descendants", issu d'une loi de Poisson de paramètre $\lambda = 10$. Les descendants sont distribués dans l'espace selon une loi normale bivariée centrée sur les parents et de variance $\sigma^2 = 0.0001$).*

peuvent être agrégatifs, réguliers ou aléatoires. Ils fournissent des modèles de référence qui permettent la simulation du processus, ce qui peut être utile par exemple pour définir la forme de l'espace écologique disponible dans les études de sélection de l'habitat (voir chapitres suivants). Nous présentons ici les trois processus les plus couramment utilisés pour ajuster un modèle au semis, à savoir le processus de Poisson, le processus inhomogène de Poisson, et le processus de Neyman-Scott.

4.3.3.1 Le processus de Poisson

Le processus de Poisson est le modèle utilisé pour simuler la CSR, dont il possède les deux propriétés (figure 13A) : le nombre d'événements dans une région A suit une loi de Poisson de paramètre $\lambda|A|$ (λ est la densité en événements sur la zone et $|A|$ est la surface de la zone), et les coordonnées des événements sur cette zone sont tirées dans une loi uniforme. La supposition implicite à ce modèle est la stationnarité. Il n'y a donc pas pour ce processus de variations de l'intensité de premier ordre (λ est constant) ni de l'intensité de second ordre ($L(t) = 0$ quel que soit t).

4.3.3.2 Le processus inhomogène de Poisson

Le processus inhomogène de Poisson occupe une place considérable dans les études de sélection de l'habitat. En effet, ce processus est non-stationnaire. En d'autres termes, le semis généré par ce processus est soumis aux variations de densité sur la zone. Ce processus est obtenu en remplaçant λ par $\lambda(\mathbf{x})$ dans le processus de Poisson (figure 13B).

C'est ce processus qui est supposé dans toutes les études reposant sur l'hypothèse d'indépendance entre les individus. La distribution des individus sur la zone est alors simplement le reflet des variations de la qualité de l'habitat. En effet, le processus inhomogène de Poisson fournit un cadre possible pour l'introduction de covariables grâce à une fonction d'intensité $\lambda(\mathbf{x}) = \lambda\{z_1(\mathbf{x}), z_2(\mathbf{x}), z_3(\mathbf{x}), ...\}$. Nous avons déjà discuté dans le chapitre 2 des outils permettant ce type de modélisation lorsque ce type de processus est à l'origine du semis.

L'ajustement de ce processus à un semis de points peut être simplement effectué à l'aide de la méthode du noyau (D 1983, C 1991). Notons que la simulation de ce processus est également très facile (B et G 1995). Il suffit simplement de simuler un processus de Poisson stationnaire classique dont l'intensité λ est égale au maximum de la valeur prise par l'intensité sur la zone ($\max(\lambda(x))$), puis à conserver les événements générés avec une probabilité égale à $\lambda(x)/\lambda_{\max}$.

4.3.3.3 Le processus de Neyman-Scott

Le processus de Neyman-Scott, ou *Poisson Cluster Process*, est un processus permettant la simulation de semis de points agrégatifs. Ce processus est simulé de la façon suivante. En premier lieu, un ensemble d'événements "*parents*" est généré par un processus de Poisson d'intensité ρ. Puis, chaque parent génère à son tour un nombre aléatoire S d'événements "*descendants*", nombre qui est issu d'une distribution de probabilité $\{p_s, s = 0, 1, ...\}$. Ces événements sont positionnés par rapport à leurs parents grâce à une fonction bivariée de densité de probabilité $h(.)$. Le semis de points est alors constitué uniquement des événements descendants.

Ce processus est stationnaire, avec une intensité de premier ordre égale à $\lambda = \rho E(S)$. D (1983) décrit les calculs qui permettent de déduire l'intensité de deuxième ordre de ce processus, et notamment la fonction $K(t)$:

$$K(t) = \pi t^2 + E\{S(S-1)\}H_2(t)/(\rho\mu^2)$$

Où $H_2(.)$ est fonction de distribution correspondant à la fonction de densité de probabilité $h_2(.)$.

L'ajustement d'un processus de Neyman-Scott implique donc deux choix : le choix d'une distribution de probabilité p_s et le choix d'une fonction de densité de probabilité $h(.)$. Un choix courant pour le nombre de descendants par parent est une distribution de Poisson de paramètre μ. La fonction de densité de probabilité $h(.)$ décrivant la position des descendants par rapport aux parents est souvent la fonction normale bivariée radialement symétrique, c'est-à-dire la densité de probabilité de présence d'un événement à la localisation de coordonnées (x, y) est donnée par l'équation

$$h(x, y) = (2\pi\sigma^2)^{-1}\exp\{-(x^2 + y^2)/(2\sigma^2)\} \tag{4.5}$$

Si la simulation d'un processus de Neyman-Scott est assez simple à réaliser, l'ajustement du processus au semis est quant à lui beaucoup plus complexe. Il faut en effet estimer deux

paramètres : (i) σ^2 (équation 4.5), qui mesure "l'étalement" des événements par rapport aux parents, et (ii) ρ, qui mesure la densité de parents sur la zone.

Comme nous l'avons indiqué plus haut, l'ajustement d'un modèle à un semis de points se fait souvent à l'aide de la fonction $K(t)$, sous l'hypothèse de stationnarité (D 1983, B et G 1995). Ainsi, la fonction `pcp()` de la bibliothèque de fonctions **splancs** pour le logiciel R minimise l'équation 4.6, en supposant que les paramètres σ et ρ sont stockés dans un vecteur θ :

$$k_1 = \int_0^{t_0} \left(\left[\hat{K}(t) \right]^c - [K(t;\theta)]^c \right) \mathrm{d}h \tag{4.6}$$

L'équation 4.6 permet l'ajustement du processus de Neyman-Scott par la méthode des moindres carrés (D 1983, C 1991). Dans cette équation, $\hat{K}(h)$ est l'estimation de la fonction K calculée sur les données, et $K(h, \theta)$ est la valeur théorique de la fonction K pour un vecteur de paramètres θ donné. c est une constante qui permet de limiter les fluctuations d'échantillonnage de $\hat{K}(t)$ qui ont tendance à augmenter avec t. D (1983) suggère une valeur de $c = 0.25$ pour les zones d'étude carrées en se fondant sur son expérience de ce type de processus. Cet auteur suggère sur les mêmes bases de poser t_0 égal au quart de la longueur de la zone d'étude. Mais il recommande également d'essayer différentes valeurs pour ces deux constantes avant de se prononcer sur un choix définitif. La valeur de k_1 peut alors servir de critère pour tester la qualité d'ajustement du modèle dans un test de Monte Carlo (après S simulations du processus et comparaison avec le semis observé).

4.4 C

Nous avons présenté dans ce chapitre les méthodes les plus couramment utilisées dans l'analyse des semis de points. Nous avons souligné l'importance fondamentale de l'exploration graphique des données, qu'il s'agisse de la construction de cartes ponctuelles, de lissage par la méthode du noyau, ou d'exploration des propriétés du semis en utilisant des méthodes basées sur les distances entre événements. La modélisation du processus est l'étape la plus complexe de l'analyse. Un grand nombre de modèles sont disponibles dans la littérature, mais ce n'est seulement que pour peu d'entre eux que l'ajustement et le calcul d'un critère de qualité d'ajustement est possible.

Ce chapitre ne donne pas une revue exhaustive de toutes les méthodes existant dans le champ de l'analyse des processus de points. Il a simplement pour objectif de décrire la démarche utilisée dans ce type d'étude. Dans le cadre des études biométriques de la sélection de l'habitat, l'analyse des semis de points est première, mais est conjuguée à l'utilisation d'autres types d'outils, que nous présentons dans les chapitres suivants. Une application concrète peut être trouvée dans le chapitre 8.

Les outils que nous avons présentés ici sont utiles pour décrire et modéliser un semis de points d'un seul type. Ils ne peuvent être utilisés que pour les protocoles de type I. Lorsque

plusieurs types d'événements sont rencontrés sur la zone d'étude (les localisations de plusieurs animaux, les occurrences de plusieurs espèces d'arbres, etc.), de nouvelles questions se posent, notamment sur les interactions entre les différents types. Sur le plan biologique, ces problèmes sont révélateurs d'un grand nombre de questions liées par exemple à l'étude de la compétition entre plusieurs espèces. Nous détaillons ces méthodes dans le chapitre suivant.

Chapitre 5

L'analyse de plusieurs semis de points

De nombreuses études en Ecologie reposent sur l'analyse de plusieurs semis de points. Ce type de données est par exemple utile à la mise en évidence de la zonation géographique de plusieurs espèces végétales (G -C 1999), de la ségrégation spatiale entre les animaux appartenant à différentes classes d'âge ou de sexe (C 1998), ou de la territorialité d'animaux suivis par radio-pistage (M 1992, D 1990). Etant donnée la diversité des questions posées avec ce type de données, il est impossible de donner une liste complète de toutes les méthodes utilisables pour les analyser.

En théorie, il serait préférable d'étudier séparément les propriétés des semis de points avant de songer à mettre en évidence les interactions entre eux, mais en pratique lorsque le nombre de semis augmente, cette opération devient fastidieuse. En outre, les solutions manquent dans la littérature pour modéliser simultanément les différents processus *et* les interactions pouvant exister entre eux lorsque le nombre de semis est important (D 1999). Notons que des solutions existent dans le cas de seulement deux semis (D 1983), qui utilisent par exemple des généralisations de la fonction $K(t)$ de Ripley (B et G 1995, H *et al.* 1996, 1997, C et S 1999, D 1999, W *et al.* 2003, C 2005).

Dans ce chapitre, nous décrivons une démarche plus générale pour étudier les interactions entre plusieurs semis. Cette approche, essentiellement exploratoire, est au cœur du travail mené dans le cadre de ma collaboration avec les conservatoires et jardins botaniques de Genève, qui avait pour but de mettre en évidence la zonation géographique d'espèces arborées en Amérique du sud et au Paraguay. Suite à cette collaboration, nous avons rédigé trois articles (S *et al.* 2004, 2006b, C *et al.* 2006), une synthèse de ces articles dans un chapitre de livre (S *et al. in press*), et nous avons présenté les résultats de ce travail lors du colloque international sur les savanes tropicales et forêts sèches qui s'est tenu a Edimbourg en septembre 2003. Toutes ces publications sont consignées en Annexe 3, 4, 5 et 6. Ce chapitre a donc surtout pour but de détailler les aspects théoriques liés à cette collaboration, à travers la question de la discrimination spatiale de plusieurs catégories de points.

5.1 N´ ´ ’ ´

Les méthodes exploratoires et descriptives sont privilégiées pour l'analyse de la distribution des semis. Ce ne sont plus les propriétés de chaque semis qui sont au centre de l'étude, mais la façon dont les types d'événements sont organisés les uns par rapport aux autres. L (1993) indique :

"*Ecologists who are studying community structure, or other multivariate datasets, may not be interested in mapping individual species or variables. The geographic variability of the whole dataset can be represented in any of three ways. The first is to compute a reduced space ordination (correspondence for species presence or abundance data ; principal component or other scalings for other data for which linear relationships can be assumed) and to map the first few component separately ; This method may fail to produce interesting maps, however, if the main components are not spatial. The second method, suggested by Legendre (1990) solves this problem ; a canonical ordination (...) of the multivariate dataset is computed, constrained by a high-order trend surface equation ; the resulting canonical variates are then mapped as above. The third method is to use clustering techniques, with or without constraint of spatial contiguity, to divide the map into more homogeneous subsets*".

Si L recommandait les méthodes multivariées pour l'analyse spatiale, il n'avait peut-être pas remarqué les liens de parenté étroits existant entre sa première méthode (l'analyse factorielle des correspondances) et la seconde (analyse canonique des tendances de surface), que nous décrivons dans les sections suivantes.

Dans un but pédagogique, nous nous concentrons uniquement ici sur l'étude de la zonation géographique de plusieurs espèces (végétales ou animales). Suite à un échantillonnage de la zone d'étude, chaque espèce étudiée est caractérisée par un certain nombre de points, chaque point correspondant à la localisation d'un spécimen de l'espèce sur une zone reportée par l'écologue. Ces points sont appelés des *occurrences* de l'espèce (G -C 1999). Chaque occurrence est caractérisée par trois variables : ses coordonnées X et Y, et l'espèce à laquelle elle appartient. L'objectif central de ce type d'étude est d'établir une typologie des espèces, ou plus exactement de *discriminer* les espèces, en fonction de la distribution spatiale de leurs occurrences.

Cette idée de discrimination appelle immédiatement l'utilisation de l'*analyse discriminante* (M 1994). Cette analyse repose sur le couplage entre un tableau de variables descriptives des occurrences et une variable qualitative décrivant les espèces auxquelles appartiennent les occurrences. Elle renvoie une *combinaison linéaire des variables descriptives* – appelée *variable discriminante* – qui maximise la séparation entre les espèces.

5.2 P ´ ’

Soit N le nombre total d'occurrences, toutes espèces confondues, S le nombre d'espèces étudiées et P le nombre de variables descriptives qui servent à discriminer les espèces. Soit $\mathbf{X}$ la matrice $N \times P$ contenant les valeurs prises par les variables descriptives pour chaque occurrence, et $\mathbf{f}$ la variable qualitative décrivant l'appartenance d'une occurrence à une espèce. La variable $\mathbf{f}$ peut être recodée sous la forme d'une matrice $\mathbf{Y}$ contenant les indicatrices des classes. A l'intersection de la ligne i et de la colonne j, cette matrice contient 1 si la i^{e} occurrence appartient à la j^{e} espèce, et 0 sinon. Soit $\mathbf{D}$ une matrice diagonale $N \times N$ de pondération des lignes de $\mathbf{X}$ et de $\mathbf{Y}$ contenant à l'intersection de la ligne i et de la colonne i le poids associé aux occurrences (e.g. $1/N$). Si le tableau $\mathbf{X}$ est centré par colonne, la matrice $\mathbf{T}$ de variances-covariances associée est calculée par :

$$\mathbf{T} = \mathbf{X}^t\mathbf{D}\mathbf{X}$$

Par ailleurs, la matrice $\mathbf{D}_m$ contient les pondérations associées aux espèces :

$$\mathbf{D}_m = \mathbf{Y}^t\mathbf{D}\mathbf{Y}$$

Cette matrice $S \times S$ est diagonale et contient à l'intersection de la ligne i et de la colonne i la proportion du nombre total d'occurrences appartenant à la catégorie i. La matrice $\mathbf{G}$ $(S \times P)$ contient les moyennes des variables de $\mathbf{X}$ pour chacune des classes définies dans $\mathbf{Y}$:

$$\mathbf{G} = \mathbf{D}_m^{-1}\mathbf{Y}^t\mathbf{D}\mathbf{X}$$

Enfin, on peut calculer la matrice $\mathbf{K}$ $(S \times P)$:

$$\mathbf{K} = (\mathbf{Y}^t\mathbf{D}\mathbf{Y})^{-1}\mathbf{Y}^t\mathbf{D}\mathbf{X}(\mathbf{X}^t\mathbf{D}\mathbf{X})^{-1} = \mathbf{G}\mathbf{T}^{-1}$$

L'analyse discriminante correspond alors à l'analyse du triplet :

$$(\mathbf{K}, \mathbf{T}, \mathbf{D}_m)$$

La matrice d'inertie associée à ce triplet est la matrice $\mathbf{WT}$ (E 1987), que nous supposerons de rang r :

$$\mathbf{WT} = \mathbf{K}^t\mathbf{D}_m\mathbf{K}\mathbf{T}$$

Cette matrice admet une base de r vecteurs propres $\mathbf{v}_k$ $(k = 1, ..., r)$ de longueur P, $\mathbf{T}$ normés $(\mathbf{v}_1^t\mathbf{T}\mathbf{v}_1 = 1)$ et deux à deux $\mathbf{T}$ orthogonaux $(\mathbf{v}_1^t\mathbf{T}\mathbf{v}_2 = 0)$. r valeurs propres λ_k sont associées à ces vecteurs. Par ailleurs, on peut calculer r vecteurs $\mathbf{o}_k$:

$$\mathbf{o}_k = \mathbf{X}\mathbf{v}_k$$

Les vecteurs $\mathbf{o}_k$ sont de longueur N et contiennent un score associé à chaque occurrence. La variance totale de ces scores est égale à 1, et la variance inter-classes est maximisée par l'analyse. L'analyse discriminante maximise donc le rapport de la variance inter-classes sur la variance totale, donc la discrimination des espèces par les variables descriptives. Et pour chaque vecteur $\mathbf{o}_k$, λ_k est une mesure du pouvoir discriminant de ce vecteur. Cette analyse est optimale

quand l'objectif est de séparer des groupes en fonction d'un jeu de mesures relevées sur les occurrences (G 1971, M 1994).

5.3 L'A F C

L'analyse factorielle des correspondances (AFC) est une méthode très communément utilisée en Ecologie (G 1984), qui y a été introduite indépendamment par plusieurs auteurs pour permettre l'analyse de tableaux de contingence issus d'échantillonnage systématique d'une zone donnée (R et R 1967, H 1971, H 1973).

Cette méthode peut aussi être utilisée pour analyser les interactions spatiales entre plusieurs classes d'occurrences. En effet, l'AFC repose sur l'utilisation d'une grille de quadrats, et sur le dénombrement des occurrences de chaque espèce dans chacun des Q quadrats de la grille. Il est possible de construire une table de contingence $\mathbf{N}$ $(Q \times S)$ contenant à l'intersection de la ligne i et de la colonne j le nombre d'occurrences de l'espèce j dans le quadrat i. L'AFC de $\mathbf{N}$ assigne un ensemble de scores à chaque quadrat et à chaque espèce, qui maximisent la discrimination des espèces par les quadrats et des quadrats par les espèces (T et C 1992). *Cette analyse est par conséquent optimale pour mettre en évidence l'organisation spatiale des espèces lorsqu'une grille de quadrats est utilisée* (T B 1985).

5.3.1 Principe mathématique

Soit $\mathbf{N}$ un tableau de contingence $Q \times S$, contenant à l'intersection de la ligne i et de la colonne j le nombre n_{ij} d'occurrences de l'espèce j dans le quadrat i. Soit $n_{\bullet j}$ le nombre total occurrences dans le quadrat j, $n_{i\bullet}$ le nombre total d'occurrences de l'espèce i, et N le nombre total d'occurrences (somme de toutes les cases de $\mathbf{N}$). On peut calculer les proportions associées à ces nombres :

$$p_{ij} = \frac{n_{ij}}{N} \qquad p_{i\bullet} = \frac{n_{i\bullet}}{N} \qquad p_{\bullet j} = \frac{n_{\bullet j}}{N}$$

On note $\mathbf{P}$ la matrice $Q \times S$ contenant les p_{ij}, $\mathbf{D}_Q$ et $\mathbf{D}_S$ les matrices diagonales

$$\mathbf{D}_Q = Diag(p_{1\bullet}, ..., p_{Q\bullet}) \qquad \mathbf{D}_S = Diag(p_{\bullet 1}, ..., p_{\bullet S})$$

Par ailleurs, soit $\mathbf{Z}$ la matrice définie par

$$\mathbf{Z} = \mathbf{D}_Q^{-1}\mathbf{P}\mathbf{D}_S^{-1}$$

L'AFC du tableau $\mathbf{N}$ est l'analyse du triplet $(\mathbf{Z}, \mathbf{D}_S, \mathbf{D}_Q)$ (T et C 1992). On diagonalise alors la matrice $\mathbf{C}$ (E 1987) :

$$\mathbf{C} = \mathbf{D}_S^{1/2}\mathbf{Z}^t\mathbf{D}_Q\mathbf{Z}\mathbf{D}_S^{1/2}$$

Si $\mathbf{C}$ est de rang q, alors cette diagonalisation permet d'obtenir q axes principaux $\mathbf{a}_k$ (scores des espèces) et q valeurs propres λ_k associées. On peut calculer les composantes principales associées (scores des quadrats) par :

$$\mathbf{b}_k = \frac{1}{\sqrt{\lambda_k}} \mathbf{Z} \mathbf{D}_S \mathbf{a}_k$$

Les vecteurs $\mathbf{a}_k$ et $\mathbf{b}_k$ sont tous deux de norme $\sqrt{\lambda_k}$. Les vecteurs $\mathbf{a}_k$ sont $\mathbf{D}_S$ orthogonaux, et les vecteurs $\mathbf{b}_k$ sont $\mathbf{D}_Q$ orthogonaux. Une propriété bien connue de l'AFC (G 1984) est que

$$\sum_{k=1}^{q} \lambda_k = \frac{1}{N} \chi^2_{obs}$$

où χ^2_{obs} est la statistique calculée pour tester les écarts à l'indépendance entre lignes et colonnes du tableau de contingence **N**. La statistique du χ^2 est utilisée pour tester l'existence d'interactions entre les distributions des différentes espèces, et l'AFC permet la meilleure représentation possible de ces écarts à l'indépendance.

5.3.2 L'AFC est une analyse discriminante

Or l'AFC peut être vue différemment. On peut en effet construire deux matrices **X** ($N \times S$) et **Y** ($N \times Q$) qui décrivent l'appartenance (0 ou 1) des occurrences (en lignes), respectivement aux espèces et aux quadrats (en colonnes).

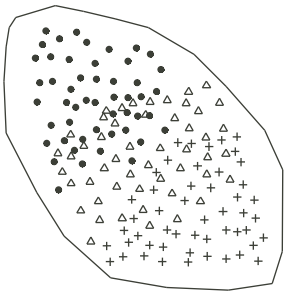
Notre objectif est de discriminer au mieux les espèces en fonction de leur distribution dans les quadrats. Il semble alors logique d'effectuer une analyse discriminante des espèces par les quadrats, donc de **Y** par **X**. Or, on peut montrer une équivalence entre l'AFC et cette analyse discriminante. En effet, si l'on recalcule les matrices utilisées en analyse discriminante, on peut montrer les identités suivantes :

$$\mathbf{D}_m = \mathbf{D}_S \tag{5.1}$$
$$\mathbf{T} = \mathbf{D}_Q \tag{5.2}$$
$$\mathbf{G} = \mathbf{Y}^t \mathbf{D} \mathbf{X} = \mathbf{P} \tag{5.3}$$

D'où on déduit la matrice

$$\mathbf{K} = \mathbf{D}_Q^{-1} \mathbf{P} \mathbf{D}_S^{-1} = \mathbf{Z}$$

L'analyse du triplet $(\mathbf{Z}, \mathbf{D}_S, \mathbf{D}_Q)$ est donc exactement identique à l'analyse du triplet $(\mathbf{K}, \mathbf{T}, \mathbf{D}_m)$. *L'AFC est donc une analyse discriminante*, ce qu'avaient déjà noté T et C (1992). En réalité, l'AFC est une double analyse discriminante ; c'est à la fois l'analyse discriminante des espèces par les quadrats et l'analyse discriminante des quadrats par les espèces.



F . *14 – Distribution de trois espèces virtuelles sur une zone d'étude (représentées par des cercles pleins, des croix et des triangles). Ces trois espèces sont distribuées selon un gradient nord-ouest / sud-est.*

5.4 A

L'AFC peut donc être considérée comme une analyse discriminante des espèces par un certain nombre de variables qui décrivent la position dans l'espace de ces occurrences (variables indicatrices des quadrats). Plusieurs occurrences voisines sont regroupées au sein d'un même quadrat, et l'AFC associe à chaque occurrence un score qui va maximiser la discrimination des espèces en fonction de leur répartition dans les quadrats. Or, l'utilisation d'une grille de quadrats implique une discrétisation abusive de la distribution des occurrences, qui peut cacher les structures fines des semis étudiés. En outre, même si des occurrences voisines sont recensées dans le même quadrat, l'AFC ne rend absolument pas compte de la proximité dans l'espace des quadrats.

En revanche, le principe d'un analyse discriminant les espèces en fonction de leur distribution spatiale peut être généralisé, en remplaçant le tableau **X**, qui décrit dans l'AFC l'appartenance des occurrences aux quadrats, par un autre tableau décrivant la position dans l'espace de ces occurrences. Nous décrivons ici deux choix possibles pour ce tableau, les polynômes des coordonnées cartésiennes et les vecteurs propres du graphe de voisinage. Les deux méthodes qui en résultent, respectivement l'analyse canonique des tendances de surface et l'analyse discriminante sur vecteurs propres de l'opérateur de voisinage sont illustrées figure 15, conjointement au principe de l'AFC. Cette illustration repose sur la distribution de trois espèces virtuelles sur une zone (figure 14).

5.4.1 L'utilisation des polynômes des coordonnées géographiques

Intuitivement, les coordonnées cartésiennes des occurrences sur la zone étudiée semblent être la meilleure description de leur position dans l'espace. Par conséquent les polynômes des coordonnées cartésiennes (i.e. les variables $x, y, x^2, y^2, x.y, x^3, y.x^2$ etc.) peuvent être utilisés pour prendre en compte l'espace de façon explicite dans les analyses (S 1914, H 1991, B *et al.* 1992). Soit $\mathbf{X}_1$ le tableau contenant F polynômes des coordonnées (colonnes) des occurrences (lignes). L'utilisation de ce tableau dans les analyses pouvant être formulées dans le cadre du schéma de dualité (§ 3.3.1) est courante en Ecologie (G 1968, W - 1985, L 1993, H 1991, G -C *et al.* 2003), où ce type d'analyse

prend le nom d'*analyse des tendances de surface* (*trend surface analysis*).

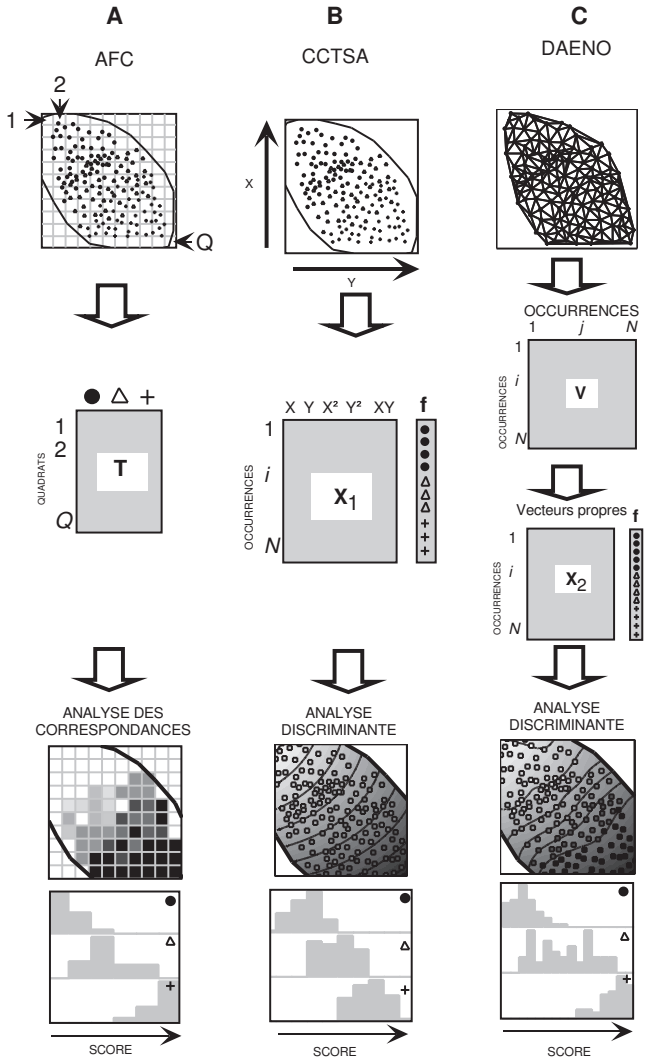
Ainsi, l'analyse discriminante de $\mathbf{Y}$ par le tableau $\mathbf{X}_1$ est utilisée en Ecologie sous le nom d'*analyse canonique des tendances de surfaces* (*canonical correlation trend surface analysis*, G -C *et al.* 2003). Cette méthode permet de trouver la combinaison linéaire des fonctions polynomiales qui va maximiser la discrimination des espèces, c'est-à-dire de maximiser la séparation spatiale des espèces, ce qui est l'objectif recherché (figure 15B).

Le défaut principal de cette analyse est que les fonctions polynomiales des coordonnées des occurrences sont souvent fortement corrélées entre elles (e.g. la coordonnée x est fortement corrélée à x^2), ce qui peut poser un problème car l'analyse discriminante requiert que les variables servant à la discrimination ne soient pas trop corrélées entre elles (M 1994). Afin de contourner ce problème, certains auteurs ont proposé l'utilisation de polynômes orthogonaux (calculés grâce à l'*orthogonalisation de Gram-Schmidt*, H 1997, p. 66) à la place des polynômes classiques (B et L 1994). Mais comme ils l'indiquent, les polynômes orthogonaux expliquent exactement la même quantité de variabilité que les polynômes classiques. Un grand nombre de fonctions polynomiales doit donc être inclus dans l'analyse pour prendre en compte une plus grande partie de la variation spatiale. Notons que même si aucune règle n'existe pour le choix du nombre de fonctions polynomiales, le degré maximum des polynômes dépasse rarement 3 en pratique, ce qui ne permet de modéliser que des structures spatiales simples et à grande échelle (L et L 1998). L'analyse des structures spatiales à petite échelle n'est pas possible avec ce type de mesures de la position dans l'espace.

5.4.2 L'utilisation des vecteurs propres de voisinage

M´ *et al.* (1993) ont proposé une alternative intéressante à l'utilisation de polynômes des coordonnées des occurrences dans les analyses spatiales. Les vecteurs propres de l'opérateur de voisinage fournissent en effet des indices de la position de chaque occurrence *relativement aux autres occurrences du semis*. Nous décrivons brièvement cette approche ici.

En premier lieu, il est nécessaire de calculer un réseau de relations de voisinages entre les occurrences du semis. Une revue récente des algorithmes possibles pour atteindre cet objectif peut être trouvée dans O (2004). La triangulation de Delaunay est un choix courant (figure 16, U et F 1985). Cette triangulation est construite à partir du pavage de Voronoi, lequel correspond à une partition du plan en un ensemble de polygones. Chaque polygone est associé à une occurrence, et recouvre la surface comprenant tous les points du plan plus proches de cette occurrence que des autres occurrences du semis. On peut alors tracer des lignes qui connectent les occurrences dont les polygones associés partagent un côté commun. Ce réseau de lignes constitue la triangulation de Delaunay. Nous nous servons ici de cette triangulation pour construire le graphe de voisinage associé au semis, c'est-à-dire le réseau de lignes reliant les occurrences voisines. Il est codé sous la forme d'une matrice carrée $\mathbf{V}$ ($N \times N$), qui contient à l'intersection de la ligne i et de la colonne j la valeur 1 si la i^{e} occurrence est voisine de la j^{e} occurrence, et 0 sinon. La matrice $\mathbf{V}$ est appelée *opérateur de voisinage*.



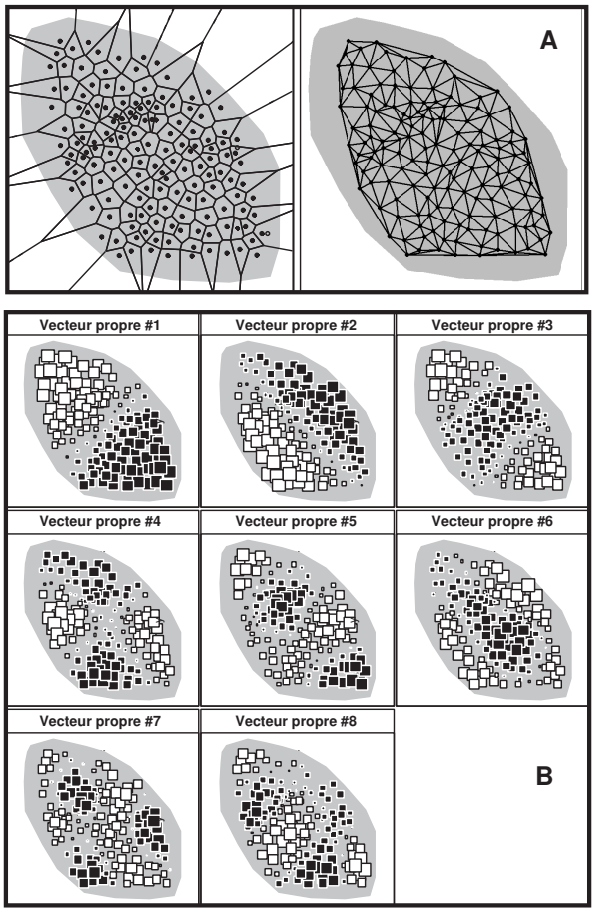
F . *15 – Trois possibilités pour les analyses discriminantes spatiales des trois semis de points présentés figure 14, distribués selon un gradient du nord-ouest au sud-est de la zone. (A) Analyse factorielle des correspondances ; (B) Analyse canonique des tendances de surface ; (C) Analyse discriminante des vecteurs propres du graphe de voisinage. Pour les analyses B et C, les scores des occurrences sont ensuite lissés sur la zone d'étude avec une régression lowess (*C et D *1988). Le lissage est présenté en niveaux de gris.*

On calcule alors la matrice diagonale $\mathbf{D}_N$, qui contient, à l'intersection de la ligne et de la colonne i, le nombre de voisins de l'occurrence i :

$$\mathbf{D}_N = Diag(\mathbf{V}\mathbf{1}_N)$$

où $\mathbf{1}_N$ est le vecteur de longueur N ne contenant que des 1. On peut enfin calculer la matrice $\mathbf{S}$ par :

$$\mathbf{S} = \frac{1}{m}\mathbf{D}_N - \frac{1}{m}\mathbf{V}$$



F . *16 – (A) Calcul du réseau de relations de voisinage du semis d'occurrences grâce à la triangulation de Delaunay. Le pavage de Voronoi est représenté à gauche. La triangulation de Delaunay est représentée à droite. (B) Carte des scores contenus dans les 8 premiers vecteurs propres de l'opérateur de voisinage associé à la triangulation de Delaunay. Chaque carré correspond à une occurrence du semis. Sa taille est proportionnelle à la valeur absolue du score, et sa couleur indique le signe du score (noir = positif ; blanc = négatif).*

où *m* est égal au nombre de paires de voisins (la somme de toutes les valeurs de la matrice **V**). La diagonalisation de **S** renvoie un ensemble de vecteurs propres orthogonaux. Le premier vecteur propre assigne un score à chaque occurrence tel que l'autocorrélation spatiale de ces scores sur la zone est maximisée. Ainsi, deux occurrences voisines auront des scores similaires, et deux occurrences éloignées (du point de vue du graphe de voisinage) auront des scores très différents. Le second vecteur maximise également l'autocorrélation spatiale sous contrainte d'orthogonalité avec le premier, etc. (voir figure 16B). Les justifications théoriques pour les formules énoncées ci-dessus peuvent être trouvées dans M´ *et al.* (1993).

En tant que tels, les vecteurs propres de l'opérateur de voisinage ne sont pas d'une grande utilité pour le biologiste. En revanche, ils peuvent être utilisés dans les analyses spatiales à la place des polynômes des coordonnées pour permettre la prise en compte d'une plus grande par-

tie de la variabilité spatiale (T *et al.* 1995, M 1998). En outre, alors que les coordonnées cartésiennes permettent de connaître la position d'une occurrence par rapport à un point de référence sans signification biologique (le point de coordonnées [0,0]), les vecteurs propres de l'opérateur de voisinage donnent la position d'un point relativement à l'ensemble du semis étudié, ce qui leur donne une signification biologique plus forte. Enfin, deux occurrences peuvent être proches dans l'espace, mais séparées par une barrière infranchissable. Cette barrière peut être prise en compte dans l'analyse en supprimant les relations de voisinage entre ces deux occurrences dans la matrice **V**.

Le principal inconvénient de cette approche est qu'elle est difficile à mettre en œuvre lorsque le nombre d'occurrences augmente. Si la matrice **S** est trop grande, sa diagonalisation devient alors une opération qui réclame un temps de calcul considérable sur ordinateur. En outre, les analyses effectuées sont liées uniquement au semis étudié : de nouvelles occurrences recueillies sur la même zone d'étude ne pourront être ajoutées *a posteriori* à l'analyse en tant qu'individus supplémentaires, car elles ne font pas partie du graphe de voisinage initial. Si ces nouvelles occurrences doivent être incluses dans les analyses, il est alors nécessaire de reconstruire un nouveau graphe de voisinage sur le nouvel ensemble d'occurrences. L'utilisation des vecteurs propres du graphe de voisinage doit donc être utilisée essentiellement à des fins exploratoires, pour aider les biologistes à comprendre les structures présentes sur la zone.

Ces vecteurs propres ont été utilisés dans une analyse discriminante (figure 15) pour la première fois dans le cadre de cette thèse, pour mettre en évidence la zonation géographique d'espèces arborées en Amérique du sud et au Paraguay. Nous avons présenté les résultats de cette application à deux échelles de l'analyse discriminante sur vecteurs propres du graphe de voisinage dans deux articles (S *et al.* 2004, 2006b), consignés en annexe 3 et 5.

5.5 C

Ce chapitre ne fait pas une revue exhaustive des méthodes disponibles pour étudier les interactions spatiales entre plusieurs semis de points. Plusieurs autres méthodes ont en effet été développées pour atteindre cet objectif dans certains champs de l'Ecologie. Ainsi, H et H (2002) présente une méthode permettant le test et la détection d'associations entre espèces, en se basant sur des indices de distances entre les aires de répartitions géographiques des espèces. Plusieurs auteurs ont développé des méthodes permettant le test des interactions statiques entre animaux suivis par radio-pistage (e.g. étude de la territorialité), qui reposent en général sur l'analyse des recouvrements entre domaines vitaux (D 1990, M 1992).

Parmi les méthodes exploratoires, l'analyse canonique des correspondances (ACC, T B 1986) est sans doute la méthode la plus fréquemment utilisée (B *et al.* 1992, B et L 1994). Cette méthode est en réalité une AFC sous contraintes, qui couple un tableau de contingence quadrat-espèces à un autre tableau contenant des variables descriptives des quadrats – appelées *variables instrumentales*. Nous détaillons le principe de cette méthode dans le chapitre 7, et nous montrons que cette méthode est liée de très près à l'analyse discriminante, dont elle est simplement un cas particulier.

Nous avons illustré comment l'analyse discriminante pouvait être utilisée pour mettre en évidence la zonation géographique de plusieurs espèces, mais il est important de souligner que cette approche peut également être utilisée sur d'autres types de données biologiques. Ainsi, D (2004) a utilisé l'AFC pour mettre en évidence les différences de l'utilisation de l'espace par sept éléphants suivis par balise Argos au parc national de Zakouma au Tchad, et a pu identifier une très nette opposition entre deux grandes zones de migrations. Cette approche permet donc l'exploration des semis et l'identification des structures spatiales. La connaissance de ces structures permet ensuite aux biologistes de tirer des conclusions sur l'organisation des différentes catégories d'événements, et les aide à construire des modèles du fonctionnement du système étudié.

L'analyse discriminante est donc une méthode très générale qui peut être utilisée pour explorer la distribution de plusieurs semis de points. La question qui se pose est de savoir comment mesurer la position des événements dans l'espace. Nous avons présenté trois méthodes possibles pour mesurer cette position dans l'espace, mais comme nous l'indiquons dans l'annexe 4 (C *et al.* 2006), les trois méthodes renverront des résultats similaires si les structures spatiales sont nettes dans les données.

Les outils de l’analyse dans l’espace écologique

Chapitre 6

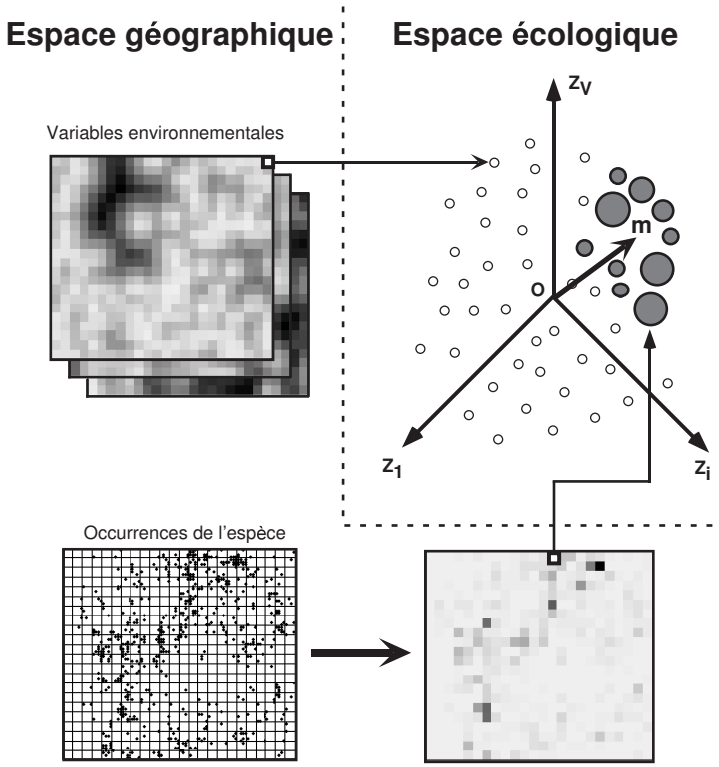
L'analyse d'une niche écologique

Dans ce chapitre, nous montrons comment le modèle de la niche écologique présenté au chapitre 2 peut servir à l'étude de l'influence des variables environnementales sur la distribution d'une espèce. Nous présentons le cas (presque) simple de l'étude d'un seul type de points (une seule espèce, individus d'une seule population, etc.). Les principaux paramètres descriptifs de la niche écologique ainsi que les outils disponibles permettant son exploration sont également décrits. Tous les outils présentés dans cette section sont disponibles – ou peuvent être très facilement programmés – avec la bibliothèque de fonctions **adehabitat**.

6.1 L ` ´

Nous définissons ici la niche écologique comme *la fonction donnant la densité de probabilité de présence de l'espèce pour une combinaison donnée des variables d'habitat*. Nous supposons que les *P* variables environnementales étudiées sont référencées spatialement sous la forme d'une carte raster *multicouches*, c'est-à-dire que chacun des *N* pixels de la carte est renseigné pour toutes les variables environnementales étudiées. Ce type de cartes peut être géré grâce à la classe de données "`kasc`" de **adehabitat** (cf. didacticiel de la bibliothèque en Annexe 13). Notons que la classe "`SpatialPixelsDataFrame`" de la bibliothèque **sp** permet également de construire de telles cartes. Chaque pixel possède donc une position dans l'espace géographique définie par ses coordonnées cartésiennes, et une position dans l'espace écologique définie par les valeurs des variables environnementales. Nous qualifions cette localisation dans l'espace écologique de *point disponible*, et la distribution de ces points dans ce même espace est appelée *espace disponible* (figure 17). Par convention, nous supposons que les variables environnementales sont centrées et réduites, c'est-à-dire que l'origine de l'espace correspond à la moyenne des conditions disponibles sur la zone et que la variabilité des points disponibles est identique pour toutes les variables.

Par ailleurs, on dispose d'un semis de localisations de l'espèce, qui reflète la distribution de la population sur la zone. Le dénombrement des localisations dans chaque pixel de la carte permet d'associer un *poids d'utilisation* à chaque point de l'espace disponible. Les points disponibles possédant un poids non-nul sont appelés *points utilisés* dans la suite de ce mémoire.



F . *17 – Le modèle pour l'étude de la niche écologique. Chaque pixel de la carte raster de la zone possède des valeurs pour les variables environnementales* $\mathbf{z}_i$*, qui définissent une position pour le point disponible associé dans l'espace écologique. Par ailleurs, en dénombrant les occurrences de l'espèce dans chacun des pixels, on peut associer à chacun de ces points un poids d'utilisation. Le vecteur de marginalité* **m** *indique la position de l'optimum pour l'espèce.*

La distribution dans l'espace écologique de ces poids d'utilisation constitue une discrétisation de la niche écologique de l'espèce pour la zone considérée. Le calcul de ces poids à partir d'une carte raster multicouches et d'un semis de localisations peut être effectué grâce à la fonction `count.points()` de **adehabitat**.

La niche écologique d'une espèce est en général caractérisée par une mesure de sa position et de son étendue dans l'espace disponible :

– La position de la niche écologique dans l'espace disponible est ordinairement décrite grâce à la position de son centre de gravité. Si la niche est normale multivariée, ce point correspond à l'*optimum* pour l'espèce sur la zone, c'est-à-dire aux conditions environnementales pour lesquelles la probabilité de présence de l'espèce est maximale. Les coordonnées de cet optimum sur chaque variable sont calculées par les moyennes pour chaque variable des points utilisés pondérée par leur poids d'utilisation. Ces moyennes sont stockées dans le vecteur **m**, appelé *vecteur de marginalité* (H 1995, D *et al.* 2000, H *et al.* 2002) ou *M-spécialisation* (P 1984). Comme les variables environnementales sont centrées, ce vecteur mesure l'écart entre les conditions rencontrées

en moyenne sur la zone et celles qui sont utilisées par l'espèce.

– La mesure de l'étendue de la niche est en général une mesure de l'inertie de la niche dans l'espace écologique. Lorsqu'une seule variable est étudiée, c'est la variance des pixels utilisés pondérée par leurs poids d'utilisation qui sert de mesure de l'étendue de la niche écologique (*S-spécialisation*, P 1984, H *et al.* 2002, H 1995). Lorsque plusieurs variables sont étudiées, la somme de ces variances mesure l'inertie associée à la niche écologique, c'est-à-dire la *tolérance* de l'espèce aux conditions présentes sur la zone (D *et al.* 2000).

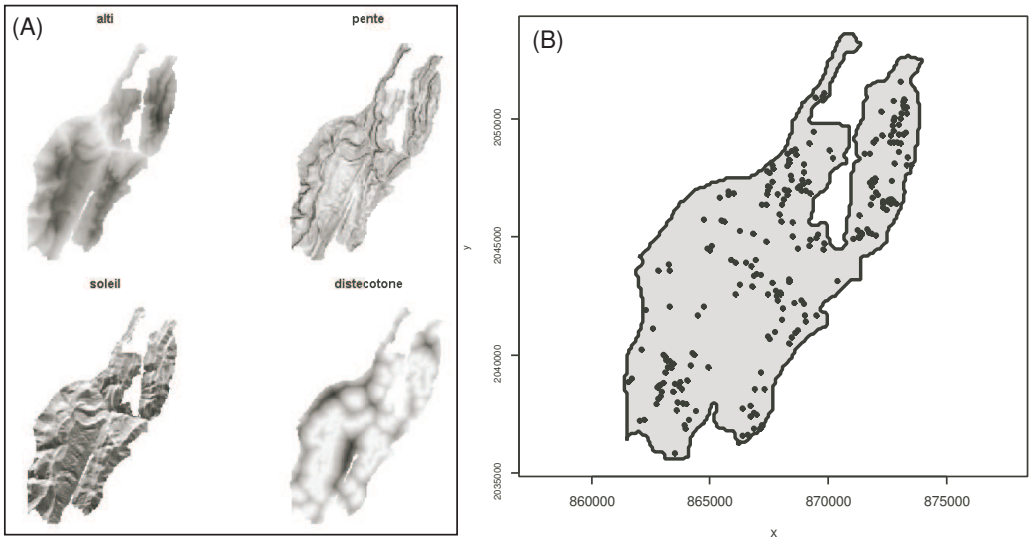
Notons que dans les années 1970, lorsque la majorité des études de la niche écologique se focalisait sur une seule variable qualitative décrivant les conditions environnementales (e.g. des types de végétation), la largeur de la niche était le plus fréquemment décrite grâce à des indices de diversité (P 1979). Or, les indices de diversité sont aussi des mesures de variabilité. Ainsi, C et G (1998) montrent que l'indice de diversité de Simpson est une généralisation de la mesure de la variance pour une variable qualitative. Ainsi, il y a dans la littérature écologique une très grande cohérence de la façon de mesurer l'amplitude de la niche écologique.

6.2 P ' ´

Nous nous servons ici du jeu de données utilisé par M (2003) pour analyser la distribution des chamois (*Rupicapra rupicapra*) dans le massif de Chartreuse (Alpes Françaises). En se basant sur des recensements effectués en novembre 1992 et 1997 par la Fédération départementale des chasseurs de l'Isère, David M a étudié la sélection de l'habitat par ces animaux. Nous nous concentrons ici sur la distribution de ces animaux en 1997. Comme dans le chapitre 4, ce jeu de données sert uniquement de matériau pour illustrer le fonctionnement des méthodes que nous présentons. Rappelons qu'une réelle analyse de ces données nécessiterait également la recherche de structures spatiales dans l'espace géographique, étape que nous évitons ici.

La distribution sur la zone des 235 groupes de chamois détectés cette année-là est présentée figure 18, ainsi que les cartes raster décrivant la distribution spatiale de quatre variables environnementales (altitude, pente, ensoleillement, et distance à l'écotone milieu ouvert/ milieu fermé). Notons que le jeu de données tel que nous le présentons dans ce chapitre n'est pas disponible dans **adehabitat**, mais que cette bibliothèque contient un sous-échantillon de ces données qui permet au lecteur intéressé de reproduire les analyses (jeu de données `chamchar`). Les commandes R utilisées pour les analyses de ce chapitre sont consignées en Annexe 11.

Comme pour les études spatiales, l'étude de la niche écologique d'une espèce doit commencer par l'exploration graphique des données. La difficulté est que l'espace à explorer n'est pas de dimension 2, mais de dimension bien supérieure. Bien sûr, la construction d'histogrammes décrivant la distribution des points disponibles et celle des points utilisés pour chaque variable environnementale (e.g. à l'aide de la fonction `histniche()` de **adehabitat**, figure 19B) peut



F . *18 – (A) Les variables d'habitat mesurées sur la zone d'étude : "alti" est l'altitude, "soleil" est l'ensoleillement et "distecotone" est la distance aux écotones milieux ouverts (e.g. pâturages) / milieux fermés (e.g. pinède). Les zones les plus claires correspondent aux valeurs des variables les plus faibles. (B) Distribution des chamois détectés sur la zone. Les coordonnées des détections sont ici données en mètres.*

apporter des informations sur les types d'habitats recherchés ou évités. Mais le caractère multidimensionel de l'espace écologique appelle l'utilisation de méthodes d'exploration multivariées. Parmi elles, les méthodes reposant sur le schéma de dualité occupent la première place (§ 3.3.1).

Nous avons indiqué dans le paragraphe précédent que la niche était généralement caractérisée par une mesure de sa position et une mesure de son étendue. La mesure de la position de la niche est généralement considérée comme l'optimum de la niche, c'est-à-dire la combinaison de variables environnementales pour laquelle la densité de probabilité de présence de l'espèce est maximale. Un grand nombre de méthodes reposent sur cette supposition. Cette hypothèse n'est acceptable que dans la mesure où la niche est normale multivariée, ou tout au moins unimodale (i.e. un seul optimum pour l'espèce). Or, comme plusieurs auteurs l'ont indiqué, ce cas de figure est loin d'être le cas général (A 1999, O et M 2002).

On comprend dès lors l'importance de l'exploration des données. Cette exploration est facilitée par toutes les possibilités du modèle euclidien d'analyse de données. On peut envisager toute sorte de méthodes exploratoires reposant sur ce modèle, et construire des analyses au cas par cas. Par exemple, on peut effectuer une analyse en composantes principales (ACP) sur les points disponibles, et représenter les poids d'utilisation sur le nuage qui en résulte, pour avoir une idée de la position et de la forme de la niche sur le plan exprimant la majeure partie de la variabilité environnementale dans l'espace disponible. On peut également appliquer l'ACP

uniquement sur les points utilisés, et ajouter les points disponibles en tant qu'individus supplémentaires à l'analyse, pour examiner les plans sur lesquels la niche est la plus étendue (ou au contraire, la plus restreinte). Il est aussi possible d'effectuer une ACP non-centrée du nuage de points utilisés, puis d'ajouter sur les plans factoriels les points disponibles en individus supplémentaires, afin de maximiser la marginalité sur le premier plan factoriel, et donc l'éloignement entre le barycentre de la niche et l'origine de l'espace disponible. Ces exemples ne sont naturellement qu'un maigre inventaire des analyses qu'il est possible d'effectuer lors de cette étape d'exploration, qui permet au biométricien, en collaboration avec le biologiste, de construire un modèle conceptuel du fonctionnement du système étudié dans cet espace.

Mais l'étude de la structuration spatiale des variables environnementales ne doit pas non plus être négligée. En effet, il peut exister des corrélations entre les variables environnementales, qui définissent des structures sur la zone d'étude. Ainsi, en montagne la végétation change avec l'altitude, la proximité des chemins touristiques est souvent être inversement corrélée à la pente, la présence de cours d'eau est souvent influencée par le relief, etc. L'examen des cartes des variables environnementales, la discussion avec les biologistes travaillant sur la zone, ainsi que des analyses multivariées sur les points disponibles (e.g. une analyse en composantes principales normée, figure 19), permettent de se familiariser avec la zone d'étude et de savoir à quels endroits des structures particulières de l'environnement sont rencontrées.

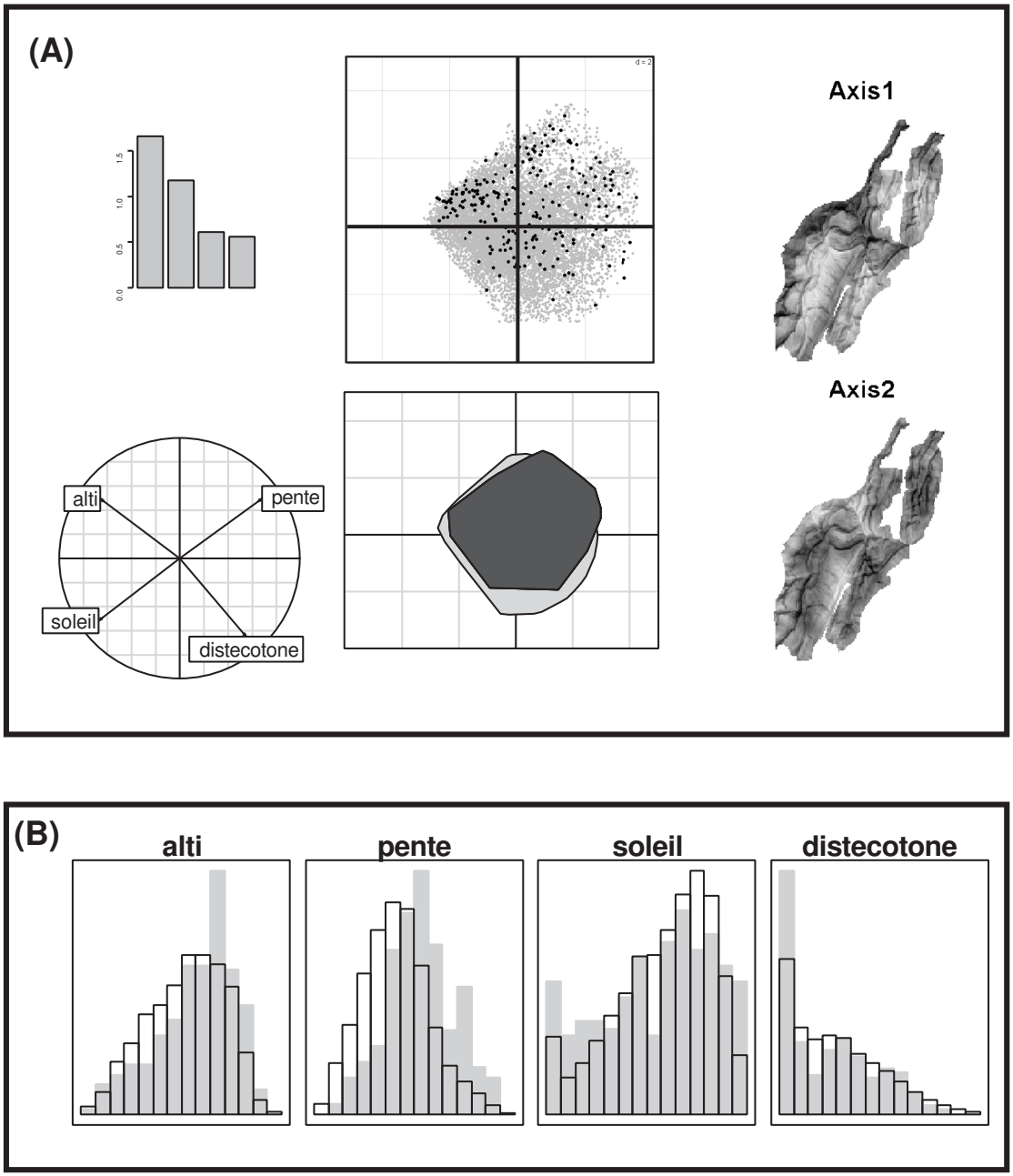
Il est donc impossible de donner une liste exhaustive de toutes les méthodes utilisables pour ces analyses. La bibliothèque de fonctions **ade4** est d'une très grande utilité pour cette étape, car elle possède une architecture facilitant ce type d'exploration. La fonction `kasc2df()` d'**adehabitat** permet de transformer les cartes raster multicouches en tableaux de données analysables par les fonctions d'**ade4**.

6.3 ' ´ (ENFA)

6.3.1 Historique de l'analyse

L'analyse factorielle de la niche écologique – aussi appelée ENFA (*ecological niche factor analysis*) – mérite une attention particulière. En effet, nous avons effectué une étude approfondie de cette méthode dans le cadre de cette thèse.

C'est à H (1995) que l'on doit les bases de cette analyse. Il est le premier à avoir noté que l'identification de facteurs limitants pour l'espèce serait obtenue en recherchant dans l'espace écologique un certain nombre d'axes pour lesquels la niche serait la plus étroite possible par rapport à l'espace disponible, c'est-à-dire pour lesquels la *restriction de niche* – aussi appelée *spécialisation* – serait maximisée. Les projections orthogonales des points disponibles et utilisés sur ces axes permettraient alors d'associer un score à chaque pixel. La spécialisation sur un axe donné peut être mesurée par *le rapport de la variance des scores des points disponibles sur la variance des scores des points utilisés*. Un rapport important sur une dimension de l'espace écologique implique qu'une faible étendue de l'espace disponible est effectivement



F . *19 – (A) Résultats de l'analyse en composantes principales normée effectuée sur le tableau donnant la valeur des variables environnementales (colonnes) pour chacun des pixels (lignes) de la carte du massif de Chartreuse. Sont présentés, de haut en bas et de gauche à droite, le diagramme des valeurs propres, le cercle des corrélations, le nuage de points sur le premier plan de l'analyse (les points noirs indiquent les pixels dans lesquels des chamois ont été détectés), une représentation de la niche sur ce premier plan (le polygone gris clair inclue 95% des points disponibles, et le polygone gris foncé, 95% des points utilisés), et la cartographie dans l'espace géographique des scores des pixels sur les deux premiers axes de l'analyse. (B) histogrammes univariés de la niche pour les différentes variables environnementales étudiées (gris = utilisé ; transparent = disponible).*

utilisée, et donc que la dimension considérée est peut-être un facteur limitant pour l'espèce.

H *et al.* (2002) ont repris ces bases théoriques pour développer l'ENFA. Cette analyse consiste tout d'abord à extraire l'axe de marginalité des données, puis à rechercher, orthogonalement à cet axe, un certain nombre d'*axes de spécialisation* qui maximisent la restriction de niche. H *et al.* avaient développé l'ENFA comme préliminaire à la construction de cartes d'habitat potentiellement utilisable par l'espèce (H et A 2003a, b). Le logiciel B - (H 2004), téléchargeable gratuitement sur internet (URL : http ://www2.unil.ch/biomapper/), permet l'application de la méthode, ce qui a facilité sa diffusion. L'ENFA a par conséquent rencontré un très vif succès chez les écologues (Z *et al.* 2002, P 2003, R *et al.* 2003, B *et al.* 2004, G *et al.* 2004, H *et al.* 2004, C *et al.* 2005).

6.3.2 Principe mathématique de l'analyse

Nous avons généralisé le principe mathématique de l'ENFA dans le cadre de cette thèse. En effet l'algorithme publié par H *et al.* (2002) présente deux inconvénients :

- Ces auteurs insistent sur le fait que les données analysées par l'ENFA doivent être des données de présence, et non des données d'abondance. Ainsi, un pixel dans lequel 10 individus ont été détectés a le même poids dans l'analyse qu'un pixel ne contenant qu'un seul individu. Cependant, il peut arriver que l'abondance ait une signification pour le biologiste, bien que ce soit rarement le cas en Biogéographie, domaine dans lequel cette analyse a été développée. Il est alors nécessaire d'étendre le principe de l'analyse de façon à pouvoir y intégrer des poids d'utilisation associés aux pixels.

- Les équations de H *et al.* ne permettent l'application de l'analyse que sur des variables environnementales quantitatives (e.g. altitude). Une variable qualitative (e.g. type de végétation) peut bien sûr être prise en compte dans l'analyse, après un recodage de la variable en un ensemble de variables indicatrices des classes (prenant la valeur 1 ou 0 pour indiquer l'appartenance ou non d'un pixel à une catégorie). Mais quand le nombre de catégories de la variable qualitative est grand, son poids dans l'analyse est très important par rapport au poids des variables quantitatives. Il est alors nécessaire de pondérer les indicatrices de classes de façon à ce que la somme des poids de toutes les variables indicatrices pour une variable qualitative soit identique au poids d'une variable quantitative (H et S 1976). Cela implique un changement de pondération des variables environnementales dans l'analyse, ce qui est malheureusement impossible avec l'approche proposée par H *et al.*

Pour ces deux raisons, nous avons généralisé le principe de l'ENFA de façon à autoriser la spécification de poids d'utilisation des pixels et/ou d'une matrice de pondération des variables. Nous décrivons ici cette généralisation.

Soit $\mathbf{Z}$ la matrice donnant la valeur des P variables d'habitat (colonnes) dans les N pixels de la carte (lignes). La matrice $\mathbf{Z}$ est centrée pour la pondération uniforme. Elle définit un nuage

de points dans un espace à P dimensions. A chacun des points de ce nuage est associé un poids d'utilisation. Ces poids sont stockés dans la matrice diagonale $\mathbf{D}_p$ ($N \times N$). La somme de tous les poids de cette matrice vaut 1. A chaque colonne de $\mathbf{Z}$ est associée une pondération, stockée dans la matrice diagonale $\mathbf{Q}$ ($P \times P$). Enfin, soit $\mathbf{D}$ la matrice diagonale $N \times N$ contenant les poids associés aux pixels disponibles (égaux à $1/N$ dans le cas d'une pondération uniforme).

On définit le vecteur de marginalité $\mathbf{m}$, qui définit les coordonnées du centre de gravité du nuage des points utilisés dans l'espace écologique ayant pour origine le barycentre des points disponibles :

$$\mathbf{m} = \mathbf{Z}^t\mathbf{D}_p\mathbf{1}_N$$

Le problème de l'ENFA est alors défini par les équations 6.1, 6.2 et 6.3 :

Problème 1 :

$$\begin{aligned} \mathbf{u}^t\mathbf{Q}\mathbf{u} &= 1 && (6.1)\\ \mathbf{u}^t\mathbf{Q}\mathbf{m} &= 0 && (6.2)\\ \mathbf{Z}\mathbf{Q}\mathbf{u} &= \mathbf{y} \\ R &= \frac{\mathbf{y}^t\mathbf{D}\mathbf{y}}{\mathbf{y}^t\mathbf{D}_p\mathbf{y}} \quad Max && (6.3) \end{aligned}$$

La condition 6.1 signifie que le vecteur $\mathbf{u}$ est normé pour la métrique $\mathbf{Q}$. La condition 6.2 implique que le vecteur $\mathbf{u}$ est orthogonal au vecteur de marginalité pour la même métrique. Comme ce vecteur $\mathbf{u}$ est normé, $\mathbf{y}$ est la projection des lignes de $\mathbf{Z}$ sur $\mathbf{u}$. $\mathbf{y}$ contient donc des scores pour les lignes de $\mathbf{Z}$. La condition 6.3 implique que le rapport de la variance des scores disponibles à la variance des scores utilisés doit être maximum. On maximise ainsi la spécialisation.

En fait, le problème de l'ENFA peut être exprimé d'une façon différente. On peut en effet le considérer comme la recherche d'un vecteur $\mathbf{w}$ qui respecte les conditions 6.4, 6.5 et 6.6 :

Problème 2 :

$$\begin{aligned} \mathbf{w}^t\mathbf{Q}\mathbf{m} &= 0 && (6.4)\\ \mathbf{Z}\mathbf{Q}\mathbf{w} &= \mathbf{a} \\ \mathbf{a}^t\mathbf{D}_p\mathbf{a} &= 1 && (6.5)\\ \mathbf{a}^t\mathbf{D}\mathbf{a} &\quad Max && (6.6) \end{aligned}$$

On démontre en annexe 9 cette équivalence entre les deux problèmes, avec :

$$\mathbf{u} = \frac{\mathbf{w}}{\|\mathbf{w}\|_\mathbf{Q}}$$

Si l'on trouve un vecteur **w** qui respecte les conditions du problème 2, le normer pour la matrice de pondération des variables environnementales **Q** permet de trouver une solution au problème 1. Or, le problème 2 est plus facile à résoudre, à condition d'effectuer quelques changements de variables. En effet, on peut poser :

$$\mathbf{Z}^* = \mathbf{Z}\mathbf{Q}^{\frac{1}{2}} \qquad \mathbf{w}^* = \mathbf{w}\mathbf{Q}^{\frac{1}{2}} \qquad \mathbf{m}^* = \mathbf{m}\mathbf{Q}^{\frac{1}{2}}$$

En outre, on définit :

$$\mathbf{S}^* = \mathbf{Z}^{*t}\mathbf{D}_\mathbf{p}\mathbf{Z}^* \qquad \mathbf{G}^* = \mathbf{Z}^{*t}\mathbf{D}\mathbf{Z}^*$$

Ce qui permet de reformuler le problème 2 :

Problème 2 bis :

$$\mathbf{w}^{*t}\mathbf{m}^* = 0 \tag{6.7}$$
$$\mathbf{w}^{*t}\mathbf{Z}^{*t}\mathbf{D}_\mathbf{p}\mathbf{Z}^*\mathbf{w}^* = \mathbf{w}^{*t}\mathbf{S}^*\mathbf{w}^* = 1 \tag{6.8}$$
$$\mathbf{w}^{*t}\mathbf{Z}^{*t}\mathbf{D}\mathbf{Z}^*\mathbf{w}^* = \mathbf{w}^{*t}\mathbf{G}^*\mathbf{w}^* \quad Max \tag{6.9}$$

On effectue encore un changement de variables pour simplifier la résolution du problème :

$$\mathbf{v} = \mathbf{S}^{*\frac{1}{2}}\mathbf{w}^* \quad \mathbf{x} = \mathbf{S}^{*-\frac{1}{2}}\mathbf{m}^* \quad \mathbf{b} = \frac{\mathbf{x}}{\sqrt{\mathbf{x}^t\mathbf{x}}} \quad \mathbf{W} = \mathbf{S}^{*-\frac{1}{2}}\mathbf{G}^*\mathbf{S}^{*-\frac{1}{2}}$$

Le problème de l'ENFA peut finalement être exprimé sous la forme du problème 3 :

Problème 3 :

$$\mathbf{v}^t\mathbf{x} = 0 \tag{6.10}$$
$$\mathbf{v}^t\mathbf{v} = 1 \tag{6.11}$$
$$\mathbf{v}^t\mathbf{W}\mathbf{v} = Max \tag{6.12}$$

On démontre que les vecteurs qui remplissent ces trois conditions sont les vecteurs propres de la matrice **H** :

$$\mathbf{H} = (\mathbf{I}_v - \mathbf{b}\mathbf{b}^t)\mathbf{W}(\mathbf{I}_v - \mathbf{b}\mathbf{b}^t)$$

Le vecteur **b** est vecteur propre de la matrice **H** pour la valeur propre 0. En effet, comme ce vecteur est normé,

$$\begin{aligned}\mathbf{H}\mathbf{b} &= (\mathbf{I}_v - \mathbf{b}\mathbf{b}^t)\mathbf{W}(\mathbf{I}_v - \mathbf{b}\mathbf{b}^t)\mathbf{b} \\ &= (\mathbf{I}_v - \mathbf{b}\mathbf{b}^t)(\mathbf{W}\mathbf{b} - \mathbf{W}\mathbf{b}\mathbf{b}^t\mathbf{b}) \\ &= (\mathbf{I}_v - \mathbf{b}\mathbf{b}^t)0 = 0\end{aligned}$$

Par ailleurs, si le vecteur **v** est vecteur propre de la matrice **H**, alors il l'est également de la matrice **W** :

$$\begin{aligned}\mathbf{v}^t\mathbf{H}\mathbf{v} &= \mathbf{v}^t(\mathbf{I}_v - \mathbf{b}\mathbf{b}^t)\mathbf{W}(\mathbf{I}_v - \mathbf{b}\mathbf{b}^t)\mathbf{v} \\ &= (\mathbf{v}^t - \mathbf{v}^t\mathbf{b}\mathbf{b}^t)\mathbf{W}(\mathbf{v} - \mathbf{b}\mathbf{b}^t\mathbf{v}) \\ &= \mathbf{v}^t\mathbf{W}\mathbf{v}\end{aligned}$$

Le vecteur **v** est donc bien vecteur propre de **W**, et les conditions 6.11 et 6.12 sont par conséquent respectées. En outre, comme le vecteur **b** est égal au vecteur **x** multiplié par une constante, et que les vecteurs **v** et **b** sont orthogonaux, cela assure le respect de la dernière condition (Eq. 6.10).

La solution du problème de l'ENFA peut donc être déduite des vecteurs propres **v** de la matrice **H**. Les vecteurs **u** respectant les conditions du problème 1 sont obtenus par

$$\mathbf{w} = \mathbf{Q}^{-\frac{1}{2}}\mathbf{S}^{*-\frac{1}{2}}\mathbf{v} \quad \text{et} \quad \mathbf{u} = \frac{\mathbf{w}}{||\mathbf{w}||_\mathbf{Q}}$$

Le vecteur **u** contient les coordonnées des variables environnementales considérées. Les scores associés aux pixels sont stockés dans le vecteur **y**, obtenu par :

$$\mathbf{ZQu} = \mathbf{y}$$

Si l'on a P variables environnementales dans l'étude, l'analyse renvoie $P - 1$ axes de spécialisation $\mathbf{u}_i$. Soit $\mathbf{y}_i$ le i^{e} axe de spécialisation de l'analyse, calculé à partir du i^{e} vecteur $\mathbf{u}_i$. Les vecteurs $\mathbf{y}_i$ sont deux à deux $\mathbf{D}_p$-orthogonaux. En d'autres termes, les scores des points utilisés contenus dans le vecteur $\mathbf{y}_i$ ne sont pas corrélés aux scores des mêmes points dans le vecteur $\mathbf{y}_j$. Les structures de corrélation internes de la niche sont ainsi "détruites" ; seule importe la spécialisation.

Enfin, on définit le vecteur **q** comme le vecteur de marginalité **Q**-normé à 1, c'est-à-dire $\mathbf{q} = \mathbf{m}/||\mathbf{m}||_\mathbf{Q}$. Les scores des points disponibles sur le vecteur de marginalité sont alors contenus dans le vecteur **f** :

$$\mathbf{ZQq} = \mathbf{f}$$

Notons que les scores des points utilisés dans le vecteur **f** peuvent être corrélés aux scores des mêmes points dans les vecteurs $\mathbf{y}_i$. L'axe de marginalité n'a pas le même statut que les axes de spécialisation, ce qui peut compliquer l'interprétation des résultats.

6.3.3 Interprétation des résultats

Lorsqu'on dispose d'un tableau décrivant la valeur de P variables environnementales dans N pixels, et d'un vecteur décrivant la proportion de localisations dans chacun de ces pixels, l'application de l'ENFA retourne un vecteur **q** donnant les coordonnées du vecteur de marginalité dans l'espace écologique (scores des variables environnementales) ainsi qu'un axe associé

f contenant les scores des pixels de la carte projetés sur ce vecteur. Elle retourne également $P-1$ vecteurs de spécialisation $\mathbf{u}_j$ donnant les scores des variables, $P-1$ vecteurs $\mathbf{y}_j$ associés donnant les coordonnées des pixels sur ces axes, et $P-1$ valeurs propres λ_j qui mesurent la spécialisation sur ces axes (figures 20 et 21).

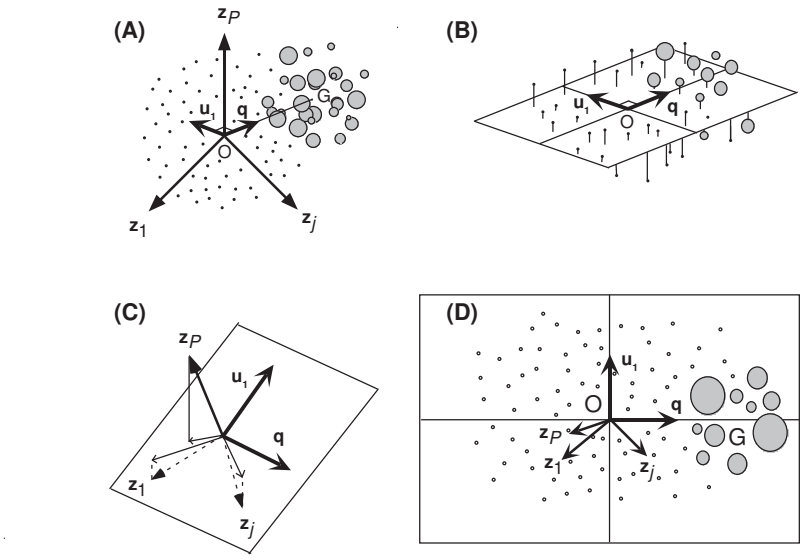
H *et al.* (2002) recommandent d'utiliser ensuite ces axes de spécialisation pour construire des cartes de qualité de l'habitat (§ 6.4), mais B (2004) préfère les utiliser pour dresser des cartes factorielles de la niche écologique. En effet, on peut représenter les coordonnées des variables et des individus sur le même plan, formé par l'axe de marginalité **q** et un des axes de spécialisation $\mathbf{u}_i$, pour obtenir une "photographie" de la niche dans l'espace écologique. Ce type de représentation est appelée *biplot* (G 1971, T B et L 1994, G 2003). L'orthogonalité existant entre le vecteur de marginalité et les axes de spécialisation assure que cette image de la niche n'est pas déformée (figure 20). Cette image est en outre la meilleure représentation possible de la niche, dans la mesure où l'éloignement entre le barycentre de l'espace disponible et le barycentre de la niche est maximisé sur l'axe de marginalité (ce qui donne une idée de la position de la niche dans l'espace écologique), et où la restriction de niche est maximisée sur le premier axe de spécialisation. La signification biologique des axes peut être déterminée grâce aux coordonnées des variables environnementales.

L'interprétation devient plus compliquée lorsqu'on considère le plan formé par deux axes de spécialisation successifs (e.g. $\mathbf{u}_1$ et $\mathbf{u}_2$). En effet, les contraintes de l'ENFA impliquent que la variance des scores utilisés est maintenue égale à 1 et qu'en même temps la variance des scores disponibles est maximisée. Il résulte de ces contraintes que ces axes sont deux à deux $\mathbf{S}^*$ orthogonaux, ce qui implique que $\mathbf{u}_1^t\mathbf{u}_2 \neq 0$. La représentation de la niche donnée par la carte factorielle des axes de spécialisation sera nécessairement une image déformée de la niche (figure 20). On préfèrera alors interpréter la signification des plans, non plus à l'aide des scores contenus dans les vecteurs $\mathbf{u}_j$, mais à l'aide des coefficients de corrélation entre les variables environnementales et les scores des points disponibles sur axes de l'analyse. Notons ici que ce problème est rencontré dans toutes les méthodes d'analyses factorielles qui ne renvoient pas des axes successifs orthogonaux pour la métrique identité, en particulier pour les analyses canoniques dont fait partie l'analyse discriminante (T B et L 1994).

En pratique, l'ENFA peut être appliquée grâce à la fonction `enfa()` de **adehabitat**, programmée par Mathieu B . Un certain nombre de fonctions graphiques associées permettent une interprétation facile et rapide des résultats (histogrammes des scores sur les axes de l'analyse avec `hist.enfa()`, *biplot* effectué à l'aide de la fonction `scatter.enfa`, figure 22).

6.4 L ' '

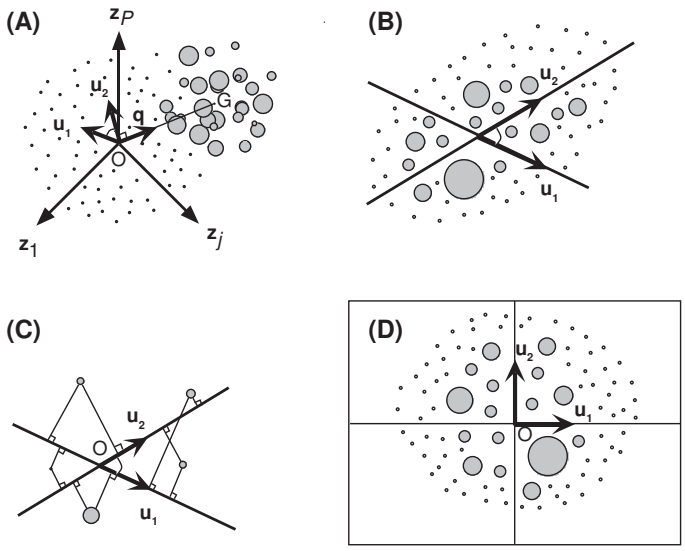
La construction de cartes de qualité de l'habitat pour l'espèce étudiée est souvent au centre des préoccupations des études sur la niche écologique d'une espèce. En effet, l'objectif final de l'étude de la sélection de l'habitat est souvent la prédiction des zones favorables à l'espèce étudiée. Mais ces cartes peuvent aussi être de formidables outils d'exploration de la niche éco-



F . *20 – Construction du biplot associé à l'ENFA : (A) l'espace écologique est ici défini par trois variables environnementales* $\mathbf{z}_1$, $\mathbf{z}_j$, *et* $\mathbf{z}_P$. *Le vecteur de marginalité* **m** *relie le barycentre de l'espace disponible O au barycentre de la niche G. Le vecteur* **q** *correspond au vecteur de marginalité normé à 1. Le vecteur* $\mathbf{u}_1$ *correspond au premier axe de spécialisation. (B) les vecteurs* **q** *et* $\mathbf{u}_1$ *sont orthogonaux, et définissent un plan sur lequel les points sont projetés. (C) De même, les variables d'habitat peuvent être projetées sur ce même plan. (D) la représentation des coordonnées des variables et des individus sur un même graphe permet permet d'obtenir la meilleure représentation de la niche de l'espèce, et une interprétation immédiate de cette niche.*

logique. Les zones prédites comme étant de bonne qualité, mais non utilisées le sont sûrement pour d'excellentes raisons (raisons historiques, non prise en compte d'une variable environnementale importante, négligence d'un aspect spatial important, cf § 1.3.1). La construction de telles cartes pourra donc être effectuée parallèlement à l'utilisation des méthodes exploratoires décrites précédemment, qu'il s'agisse des méthodes d'exploration de l'espace géographique ou de l'espace écologique.

L'estimation de ce type de cartes repose toujours sur un modèle de la niche. Le modèle le plus fréquemment utilisé est une fonction normale bivariée, qui présente l'avantage que l'optimum pour l'espèce est situé au barycentre de la niche. Deux grandes familles de méthodes sont distinguées pour la construction de ces cartes (R *et al.* 2001, E *et al.* 2004) : (i) les *techniques de discrimination de groupes* qui reposent sur des données de présence et d'absence de l'espèce, et (ii) les *techniques de construction de profils* (*profile techniques*) qui n'utilisent que les présences pour construire ces cartes. Comme nous l'avons indiqué dans le § 2.2.2, les données d'absence sont rares dans les domaines qui touchent à l'étude de la faune sauvage. Pour cette raison, nous ne nous concentrons ici que sur les techniques de construction de profils.

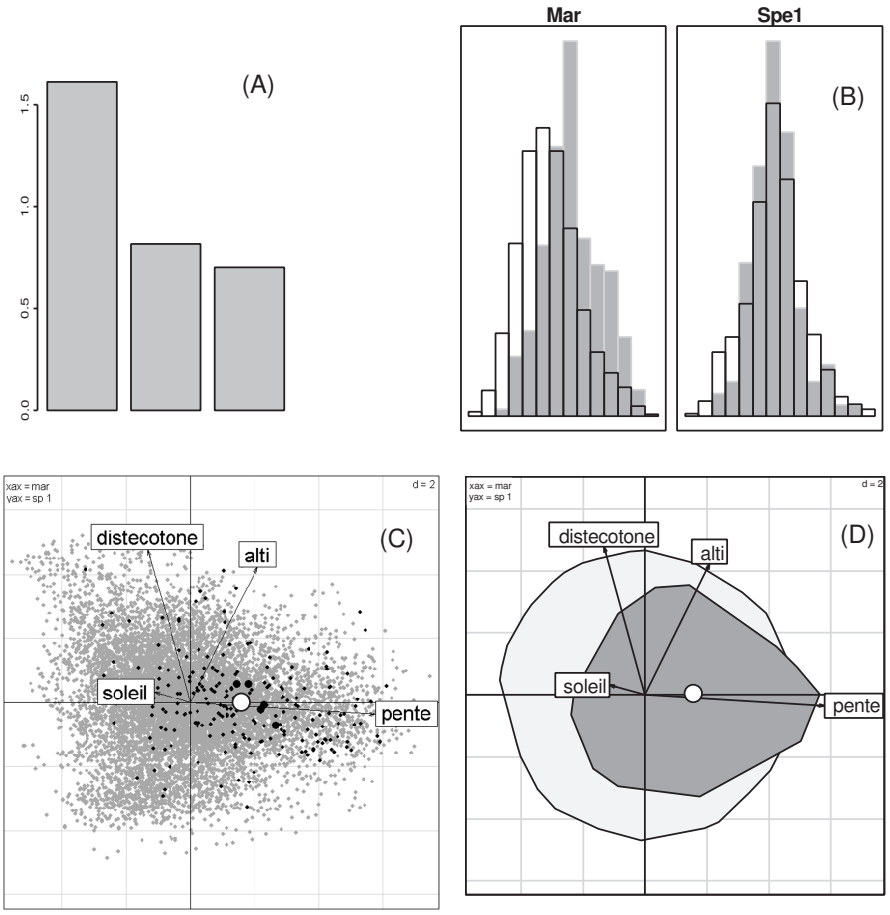


F . *21 – Construction des cartes factorielles de la niche basées uniquement sur les axes de spécialisation : (A) l'espace écologique est ici défini par trois variables environnementales* $\mathbf{z}_1$, $\mathbf{z}_j$, *et* $\mathbf{z}_P$. *Le vecteur de marginalité* **m** *relie le barycentre de l'espace disponible O au barycentre de la niche G. Le vecteur* **q** *correspond au vecteur de marginalité normé à 1. Le vecteur* $\mathbf{u}_1$ *correspond au premier axe de spécialisation, et le vecteur* $\mathbf{u}_2$ *au deuxième axe de spécialisation. (B) les vecteurs* $\mathbf{u_1}$ *et* $\mathbf{u}_2$ *ne sont pas orthogonaux. (C) En projetant les points orthogonalement sur chaque vecteur, et (D) en rendant ces vecteurs orthogonaux, on déforme l'espace écologique.*

Toutes les méthodes existant dans la littérature pour la construction de cartes "profils" peuvent être reformulées dans le cadre du concept de niche écologique (H et A 2003a). Ces méthodes ont en effet toutes pour objectif d'assigner une valeur numérique à chaque point disponible de l'espace écologique en fonction de sa position par rapport à la niche de l'espèce étudiée. Ces valeurs sont ensuite cartographiées dans l'espace géographique pour obtenir des cartes de qualité de l'habitat. Selon la méthode utilisée, la valeur que prendra un point disponible donné pourra correspondre soit à une mesure de la distance dans l'espace écologique entre le point disponible et l'optimum pour l'espèce, soit à une mesure directement liée à la probabilité de présence de l'espèce au point considéré. Nous effectuons ici une rapide présentation de quelques méthodes, en accordant une attention spéciale à la méthode des distances de Mahalanobis, dont nous démontrons ici qu'elle présente une certaine parenté avec l'ENFA.

6.4.1 L'algorithme BIOCLIM

L'un des tout premiers algorithmes développé pour construire ce type de cartes est l'algorithme BIOCLIM (B 1991). Le principe de cette méthode est très simple, voire simpliste. Pour chacune des variables environnementales étudiées, on peut déterminer les limites minimum et maximum des valeurs des pixels utilisés. L'ensemble de ces limites calculées sur toutes les variables définit une "boîte" dans l'espace écologique. Cette boîte constitue une estimation



F . *22 – Résultats de l'ENFA utilisée pour mettre en évidence les caractéristiques de la niche écologique du chamois dans le massif de Chartreuse. (A) Valeurs propres de spécialisation de l'analyse (un axe de spécialisation est ici suffisant pour expliquer la majeure partie de la spécialisation du jeu de données) ; (B) histogrammes des scores des points disponibles (transparent) et utilisés (gris) sur l'axe de marginalité et le premier axe de spécialisation ; (C) biplot du premier plan de l'analyse. L'abscisse correspond à l'axe de marginalité, et l'ordonnée au premier axe de spécialisation. Les points disponibles sont indiqués en gris, et les points utilisés sont indiqués en noirs (leur diamètre est proportionnel aux poids d'utilisation de chaque point). Le point blanc indique la position de la moyenne des scores utilisés (extrémité du vecteur de marginalité). (D) Même biplot, simplifié pour la représentation des résultats. Le polygone gris clair représente le nuage de points disponible, et le polygone gris foncé correspond à la niche de l'espèce. Chaque polygone correspond au polygone convexe minimum incluant 95% des points.*

de l'enveloppe contenant 100% des points utilisés dans l'espace écologique, donc de la niche. On peut de la même façon calculer la valeur des 2.5^{e} et 97.5^{e} percentiles de la distribution des points utilisés sur chaque variable pour calculer une enveloppe de volume plus restreint (*core bioclimate*, C *et al.* 1993). On procède ainsi récursivement, pour des quantiles délimitant des enveloppes de plus en plus restreintes, jusqu'à atteindre la médiane de chacune des variables.

On détermine ensuite l'enveloppe de dimension minimale contenant chacun des points disponibles, ce qui permet de leur assigner une valeur de probabilité. Ces valeurs peuvent alors être cartographiées dans l'espace géographique (figure 23). Les limites d'une telle méthode sont évidentes : considérer indépendamment les différentes variables environnementales revient à supposer que la niche est de forme cubique. Une telle hypothèse est rarement respectée en réalité. Pour cette raison, la qualité de l'habitat est très souvent surestimée avec cette méthode (C *et al.* 1993). Malgré ce défaut, elle est aujourd'hui encore très utilisée par les biogéographes, et en particulier en Australie où elle a été développée (e.g. B et H 2002, F *et al.* 2001, P *et al.* 2001).

6.4.2 L'algorithme DOMAIN

L'algorithme DOMAIN est également l'un des plus utilisés dans le domaine de la Biogéographie (C *et al.* 1993, S 2000). Celui-ci repose sur l'utilisation de la métrique de Gower, qui définit une façon de mesurer les distances dans l'espace écologique. La distance $d_{Gower}(i, j)$ entre deux points de coordonnées $\mathbf{z}_{i\bullet}$ et $\mathbf{z}_{j\bullet}$ dans un espace de dimension P est calculée par l'équation :

$$d_{Gower}(i, j) = \frac{1}{P} \sum_{k=1}^{P} \frac{|\mathbf{z}_{ik} - \mathbf{z}_{jk}|}{a_k}$$

où $\mathbf{z}_{ik}$ est la valeur de la variable k pour le point i, et a_k est l'étendue de la variable k (i.e. max - min). On peut alors calculer la similarité entre ces deux points par $R_{ij} = 1 - d_{Gower}$. En pratique, l'étendue a_k est calculée uniquement sur les points utilisés. Pour un point disponible i de l'espace écologique on calcule la similarité de Gower à chacun des points utilisés de la niche, et la proximité de ce point à l'ensemble de la niche est calculée par le maximum des similarités entre ce point et tous les points de la niche (C *et al.* 1993). La cartographie de ces similarités permet ensuite d'obtenir une carte de qualité de l'habitat. Le principal inconvénient de l'algorithme DOMAIN est qu'il ne prend pas en compte les variations de densité des points utilisés dans l'espace écologique, rendant les cartes très sensibles aux valeurs extrêmes (H et A 2003b). L'algorithme DOMAIN peut être appliqué sous R à l'aide de la fonction `domain()` de **adehabitat** (figure 23).

6.4.3 Cartographie reposant sur l'utilisation de l'ENFA

H *et al.* (2002) utilisent les axes de l'ENFA pour calculer un indice reflétant la probabilité d'utilisation d'un point disponible en fonction de sa position dans l'espace écologique. Pour chacun des axes de spécialisation, la distribution des scores des points utilisés est découpée en classes, de telle façon que la médiane sépare exactement deux classes. Pour un point disponible j dans l'espace écologique, on peut compter le nombre de points utilisés qui sont soit dans la même classe, soit dans une classe plus éloignée de la médiane que j. Ce nombre est ensuite multiplié par deux, puis divisé par le nombre total de points utilisés. Sur un axe donné, lorsque le point disponible est proche de la médiane, cet indice vaut 1. S'il est complètement en dehors

de la niche, cet indice vaut 0. On calcule ensuite une moyenne de ces indices pondérée par les valeurs propres de l'ENFA. Cette moyenne constitue une mesure de la qualité de l'habitat pour l'espèce.

Or l'axe de marginalité a un statut particulier dans l'analyse (§ 6.3.2). Les scores des points utilisés sur l'axe de marginalité sont toujours corrélés aux scores des mêmes points sur les axes de spécialisation, bien que pas nécessairement de façon importante. En outre, l'axe de marginalité explique toujours une certaine quantité de spécialisation. Ainsi, ne pas inclure l'axe de marginalité dans le calcul de ces cartes se traduira par la perte d'une dimension de l'espace écologique qui explique peut-être une quantité non négligeable de la sélection de l'habitat par l'espèce. Inversement, l'inclure se traduira par un biais dans le calcul des valeurs de qualité de l'habitat, puisque ce calcul suppose l'orthogonalité des vecteurs contenant les scores. H *et al.* tranchent le nœud gordien en proposant de calculer cet indice pour l'axe de marginalité et de l'intégrer au calcul de la moyenne pondérée. La moitié du poids est alors donnée à l'axe de marginalité, et l'autre moitié aux axes de spécialisation. Mais ce choix est intuitif, et ne repose sur aucune base théorique.

Plusieurs autres méthodes ont également été développées pour construire ce type de cartes après une utilisation de l'ENFA (e.g. H et A 2003a, b). Par exemple, puisque les corrélations entre les scores des points utilisés sur les axes de spécialisation de l'ENFA sont nulles, on peut utiliser les distances euclidiennes classiques pour mesurer la position d'un point disponible à la niche. Cette distance est définie, dans l'espace défini par les K axes de spécialisation, par l'équation :

$$\delta(i, j) = \sqrt{\sum_{k=1}^{K} w_k(\mathbf{z}_{ik} - \mathbf{z}_{jk})^2} \tag{6.13}$$

où w_k est le poids associé à chacun des K axes de spécialisation considérés. H et A (2003b) proposent d'utiliser comme poids les valeurs propres de l'ENFA. Ces auteurs effectuent ce choix intuitif en indiquant qu'un axe sur lequel la spécialisation est la plus importante doit avoir plus de poids dans l'analyse.

En pratique, pour un point disponible a donné, H et A (2003b) proposent de calculer comme la moyenne géométrique des distances euclidiennes entre ce point et les N_u points utilisés de l'espace écologique comme indice de la distance entre ce point et la niche :

$$\mu_G(a) = N_u \sqrt{\Pi_{i=1}^{N_u} \delta(a, i)}$$

La cartographie de ces distances permet d'obtenir une carte de qualité de l'habitat (figure 23). Notons que ces méthodes reposent sur l'hypothèse de non-corrélation des axes successifs de l'analyse. Encore une fois, l'axe de marginalité pose un vrai problème, à cause de sa corrélation potentielle avec les axes de spécialisation. Ce problème, pourtant important, est rarement discuté dans la littérature développant ce type de méthodes.

6.4.4 Les distances de Mahalanobis

La méthode des *distances de Mahalanobis* peut apporter des solutions à bien des problèmes soulevés par les méthodes précédentes. Cette approche est très utilisée dans les études de la sélection de l'habitat par la faune sauvage pour construire des cartes de qualité de l'habitat (C *et al.* 1993, K et D 1997, K et R 1998, M *et al.* 2002), mais curieusement, elle est assez peu connue en Biogéographie.

6.4.4.1 Principe mathématique

Soit **S** la matrice de variance-covariance des points utilisés, $\mathbf{z}_{i\bullet}$ le vecteur contenant la valeur des variables pour le pixel *i*, et **m** le vecteur de marginalité contenant les moyennes utilisées pour chacune des *P* variables étudiées. La distance de Mahalanobis associée au pixel *i* est calculée grâce à l'équation :

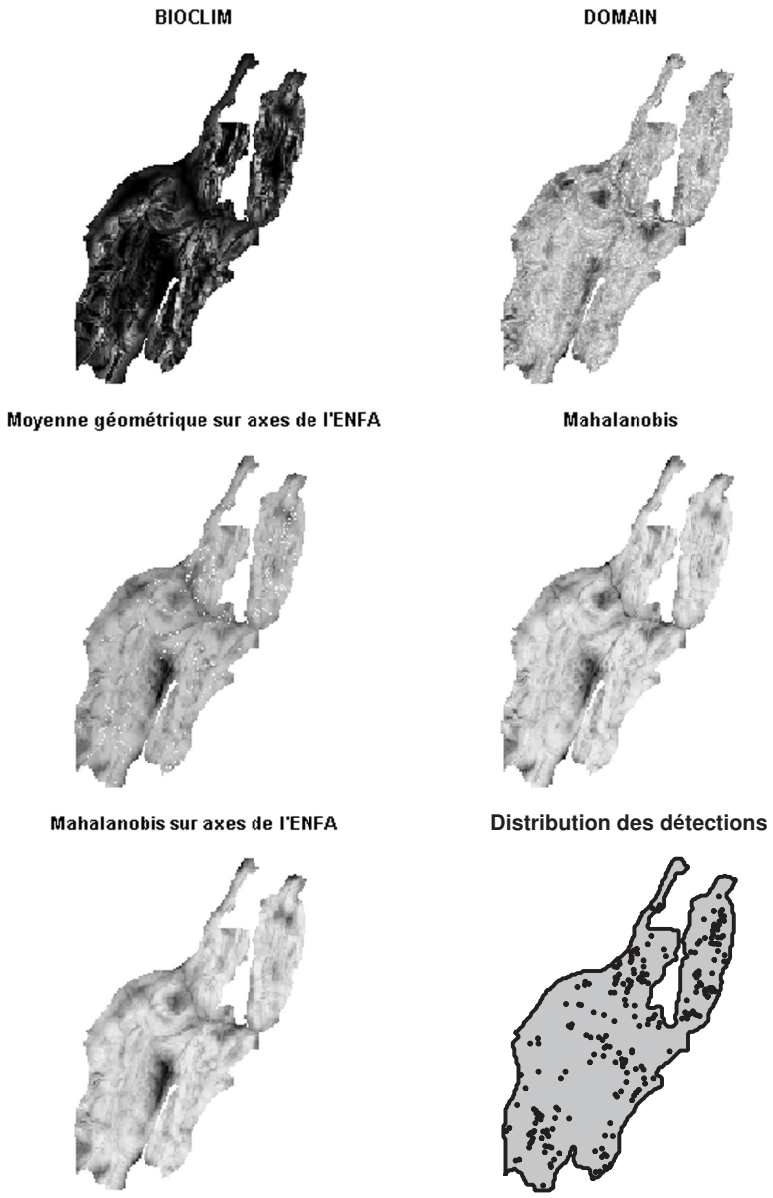
$$d_{Mahalanobis}(i) = (\mathbf{z}_{i\bullet} - \mathbf{m})^t \mathbf{S}^{-1} (\mathbf{z}_{i\bullet} - \mathbf{m})$$

Plus le point disponible sera proche du barycentre de la niche, et meilleure sera la qualité de l'habitat (voir figure 24). Les distances de Mahalanobis permettent la prise en compte des structures de variance-covariance de la niche (i.e. l'espace disponible est déformé de façon à "sphériciser" la niche). L'utilisation de cette méthode repose sur l'hypothèse que la niche ne présente qu'un seul optimum, hypothèse qui devra au préalable être vérifiée grâce aux méthodes d'exploration décrites dans le § 6.2. Dans le cas contraire, certaines zones exclues de la niche pourront être faussement prédites comme étant favorables à l'espèce.

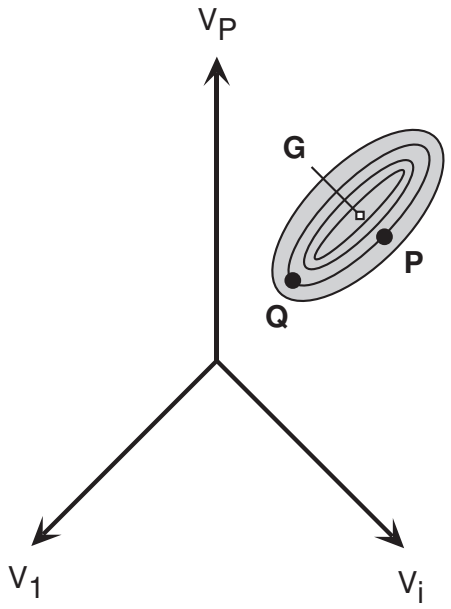
Notons que la méthode des distances de Mahalanobis peut aussi être utilisée pour estimer une carte de probabilité. En effet, sous l'hypothèse de normalité multivariée de la niche, ces distances sont approximativement distribuées selon une loi du χ^2 à $P - 1$ degrés de liberté (C *et al.* 1993). On peut alors cartographier les probabilités que les points disponibles appartiennent à la niche. En pratique, les cartes des distances de Mahalanobis, comme celles des probabilités associées peuvent être calculées grâce à la fonction `mahasuhab()` de **adehabitat** (figure 23).

6.4.4.2 Parenté avec l'ENFA

La méthode des distances de Mahalanobis reposent sur les mêmes bases mathématiques que l'ENFA exprimée sous la forme du problème 2 (§ 6.3.2, Equations 6.4, 6.5 et 6.6). En d'autres termes, à partir d'un tableau **Z** décrivant la valeur de *P* variables environnementales (en colonnes) pour *N* points disponibles (en lignes), d'une matrice diagonale **Q** contenant la pondération des variables, d'une matrice diagonale **D** contenant la pondération des lignes et d'une matrice diagonale $\mathbf{D}_p$ donnant le poids d'utilisation associé à chacun des points disponibles, l'ENFA renvoie un vecteur **f** contenant les scores des points disponibles sur le vecteur de marginalité, et $P - 1$ vecteurs $\mathbf{a}_i$ contenant les scores de ces points sur les axes de spécialisation.



F . *23 – Cinq exemples d'algorithmes utilisés pour construire des cartes de qualité de l'habitat. Ici, la qualité de l'habitat est calculée pour les chamois du massif de Chartreuse. Les zones les plus claires correspondent aux zones les plus favorables. Les cartes sont construites grâce aux algorithmes BIOCLIM (*B *1991), DOMAIN (*C *et al. 1993), au calcul de la moyenne géométrique des distances euclidiennes sur tous les axes de l'ENFA (*H et A *2003b), au calcul des distances de Mahalanobis sur les variables environnementales (*C *et al. 1993), et au calcul de ces mêmes distances sur l'axe de marginalité et le premier axe de spécialisation de l'ENFA appliquée sur ce jeu de données (figure 22).*



F . *24 – Principe du calcul des distances de Mahalanobis : La niche écologique (en gris) est supposée normale multivariée dans l'espace défini par les trois variables d'habitat. Le point* ***G*** *est l'optimum de la niche. La distance de Mahalanobis entre le point* ***P*** *et le point* ***G*** *est une mesure de la qualité de l'habitat pour l'espèce au point* ***P****. Les contours d'égale distance de Mahalanobis à* ***G*** *sont indiqués sur la figure. Cette distance prend en compte la structure de variance-covariance de la niche. Les points* ***P*** *et* ***Q*** *sont à la même distance de Mahalanobis de* ***G****, bien que la distance euclidienne entre* ***P*** *et* ***G*** *soit inférieure à la distance euclidienne entre* ***Q*** *et* ***G****.*

Dans un premier temps, considérons uniquement les axes de spécialisation, et laissons de côté l'axe de marginalité. Pour un axe de spécialisation donné, la variance des scores des points disponibles pondérée par les poids d'utilisation vaut 1 (Equation 6.5), et c'est la variance des scores des points disponibles qui est maximisée (Equation 6.6). En outre, les scores des points disponibles sur les axes de spécialisation successifs sont non-corrélés. Soit la matrice $\mathbf{A} = [\mathbf{a}_1|...|\mathbf{a}_{P-1}]$ contenant les scores des points disponibles sur les $P-1$ axes de spécialisation. Par définition, la moyenne des points utilisés est égale à 0 pour ces axes, c'est-à-dire

$$\mathbf{A}^t\mathbf{D}_p\mathbf{1}_N = 0$$

En outre, puisque la matrice $\mathbf{A}$ est $\mathbf{D}_p$-orthonormée, la matrice de variances-covariances $\mathbf{S} = \mathbf{A}^t\mathbf{D}_p\mathbf{A}$ est égale à la matrice identité. Les distances de Mahalanobis calculées sur ces axes sont donc identiques aux distances euclidiennes classiques des points à l'origine des axes, donc au barycentre de la niche :

$$d_{Mahalanobis}(i) = \sqrt{\sum_{k=1}^{P-1}(\mathbf{a}_{ik})^2}$$

Il est en effet logique, avec une niche sphérique (même variance dans toutes les dimensions), d'utiliser comme indice de distance à l'optimum la distance euclidienne classique.

Mais les axes de spécialisation sur lesquels on travaille habituellement sont plutôt les axes générés par les vecteurs $\mathbf{u}_i$ et non par les vecteurs $\mathbf{w}_i$. Ainsi, les distances de Mahalanobis calculées sur les axes $\mathbf{y}_i$ générés par les vecteurs $\mathbf{u}_i$ peuvent être calculées par les distances euclidiennes pondérées (Equation 6.13), en prenant pour chaque axe k un poids w_k égal à l'inverse de la variance des scores des points disponibles sur cet axe pondérée par les poids d'utilisation, c'est-à-dire $1/\mathbf{y}_k^t \mathbf{D}_p \mathbf{y}_k$. ***Cette remarque remet en question la justesse du choix des valeurs propres comme pondération dans le calcul de la moyenne géométrique des distances euclidiennes proposé par*** **H** et A ***(2003b)***.

La similarité entre distances euclidiennes et distances de Mahalanobis n'est valable que lorsqu'on considère les axes de spécialisation retournés par l'ENFA. Encore une fois, la différence de statut entre l'axe de marginalité et les axes de spécialisation pose un problème. Ce problème, décidément récurrent, remet en question l'utilisation de l'ENFA comme méthode de cartographie de la qualité des habitats. Nous soutenons ici l'opinion que cette analyse devrait être principalement utilisée que pour permettre la représentation de la niche sur un plan factoriel et identifier les variables environnementales ayant le plus d'importance pour l'espèce. Les distances de Mahalanobis sont un outil bien plus puissant pour cartographier la qualité des habitats. Cette méthode repose sur le même modèle de la niche que l'ENFA (un modèle normal multivarié), mais évite les problèmes liés à la marginalité de la niche, en centrant l'espace écologique sur le barycentre de la niche et uniquement sur lui, alors que l'ENFA centre l'espace écologique à la fois sur le barycentre de la niche, mais aussi sur le barycentre de l'espace disponible. Lorsqu'une cartographie des habitats est désirée, ce double centrage n'est pas nécessaire.

Mais il est aussi possible de cartographier la qualité de l'habitat en utilisant les résultats de l'ENFA. Par exemple, si suite à une ENFA, on peut identifier un petit nombre d'axes de spécialisations ayant une signification biologique, on peut cartographier la qualité de l'habitat en utilisant les distances de Mahalanobis calculées sur le tableau contenant ces axes *et* l'axe de marginalité. Il s'agit d'ailleurs de la stratégie que nous avons adoptée dans la fonction `predict.enfa()` de la bibliothèque **adehabitat**. Ceci conduit à des cartes de qualité de l'habitat avec beaucoup moins de "bruit" (figure 23), puisque construites sur un nombre plus faible de variables environnementales.

6.4.5 Autres méthodes

Toutes les méthodes présentées précédemment peuvent aboutir à une surestimation du volume de la niche lorsque le modèle supposé (pour la plupart le modèle normal multivarié) ne s'ajuste pas bien à la niche observée. Le moyen le plus sûr de calculer des enveloppes de probabilités pour la niche est de *modéliser* la niche, par exemple à l'aide du modèle linéaire généralisé (C *et al.* 2002, H 2002, G *et al.* 1999). Comme nous l'avons indiqué précédemment, de telles méthodes ne pourront être utilisées qu'à condition d'avoir identifié tous les facteurs susceptibles d'influencer la distribution spatiale de l'espèce étudiée (§ 2.4.3).

Les approches de cartographie qui font appel à cette méthodologie sont pour beaucoup des

méthodes de discrimination de groupes (E *et al.* 2004) qui réclament non seulement un échantillon des zones utilisées mais aussi un échantillon des zones non-utilisées. Pourtant, des méthodes de construction de profils permettent également cette modélisation. Nous avons déjà discuté des *fonctions de sélection des ressources* (§ 2.3.1.2), et du succès important que cette méthodologie rencontre chez les gestionnaires de la faune sauvage. Son grand avantage est qu'elle ne nécessite pas forcément de données d'absence, et peut être utilisée avec une simple connaissance de la disponibilité des habitats (e.g. B et M D 1999, M *et al.* 2002).

6.5 C

Les outils que nous avons présentés dans cette section facilitent l'exploration de la niche dans l'espace écologique. Les techniques multivariées reposant sur le schéma de dualité permettent de représenter la niche écologique en fonction de certains critères. Les cartes de la qualité de l'habitat doivent également être considérées comme des outils exploratoires. En effet, ces cartes sont construites à partir des données, et ne sont que des représentations "brutes", dans l'espace géographique de la position des points disponibles dans l'espace écologique. Les anglo-saxons parlent de *point estimate*, c'est-à-dire qu'on ne dispose d'aucune estimation de la variabilité associée à la qualité de l'habitat estimée pour un point. Pour pouvoir disposer d'une estimation de cette variabilité, il faudrait disposer d'un *modèle de tout le système étudié*, et pas seulement de la niche écologique. Il faudrait alors distinguer la part relative des variables environnementales de l'effet des contraintes spatiales, sociales, de prédation, etc., et aucune méthode statistique ne permet de le faire de façon automatique.

Ce modèle est également nécessaire pour pouvoir construire des tests de la sélection de l'habitat. Les méthodes de Monte Carlo sont de la plus grande utilité pour ce type de tests. Mais il est impossible de présenter une liste des tests permettant de répondre à toutes les questions pouvant être posées dans ce domaine. Un exemple peut éclairer ce point de vue. Un test possible de l'existence d'une sélection de l'habitat peut être effectuée en randomisant les localisations de l'espèce sur la zone, puis en effectuant une ENFA sur le résultat, et enfin, en récupérant la première valeur propre de spécialisation de l'analyse comme critère pour mesurer la sélection de l'habitat. Cette étape peut être répétée un grand nombre de fois. On simule ainsi la *Complete Spatial Randomness* (§ 4.2.3), c'est-à-dire une utilisation de l'habitat totalement aléatoire, sans aucune contrainte spatiale. Or, s'il est démontré que les individus sont regroupés en clusters, ce test n'est plus valide. Peut-être un processus de Neyman-Scott (§ 4.3.3.3) sera-t-il dans ce cas plus adapté pour simuler une utilisation aléatoire de l'habitat...

On voit donc que l'exploration de la niche écologique est indissociable de l'exploration des aspects spatiaux. L'objectif est de déterminer (i) la part relative de l'influence environnementale dans la distribution des localisations sur la zone, et (ii) construire un modèle, même conceptuel, du fonctionnement spatial de la population. La population est un système complexe, et l'analyse aura pour but de déterminer la nature de cette complexité, de délimiter ses différents compartiments, et de mettre en évidence les interactions entre ces différents compartiments, qu'ils soient environnementaux ou propres au système.

Chapitre 7

Plusieurs niches écologiques

Le présent chapitre présente quelques outils utiles à l'étude de plusieurs niches dans l'espace écologique. Nous y soulignons la très grande cohérence de ces outils du point de vue du modèle de la niche écologique. Etudier plusieurs niches dans l'espace écologique est une opération complexe, et les questions qui peuvent être posées par le biologiste ou par les données sont très diverses. Nous présentons ici deux grandes familles de méthodes : (i) les méthodes de discrimination qui ont pour objectif de rechercher les combinaisons de variables environnementales qui permettent de séparer au mieux les niches des espèces étudiées, et (ii) les méthodes d'étude de la sélection de l'habitat, qui ont au contraire pour but de mettre en évidence la similarité entre les niches de plusieurs espèces, et d'établir une typologie des espèces en fonction de la façon dont leur utilisation de l'habitat diffère de la disponibilité générale de ces habitats.

7.1 L

7.1.1 Introduction

Dans cette section, les catégories de points sont nommées "*espèces*" et les points en eux-mêmes "*occurrences*", comme dans le chapitre 5. Gardons à l'esprit qu'il s'agit d'un choix terminologique qui n'a pour but que de faciliter la compréhension des outils par le lecteur. Ces catégories pourraient tout aussi bien représenter un autre type d'objets biologiques (animaux suivis par radio-pistage, classes d'âge d'individus détectés sur une zone, etc.). On dispose d'un ensemble d'occurrences de différentes espèces localisées dans l'espace géographique, et d'un ensemble de cartes raster décrivant la structuration des variables environnementales sur la zone d'étude.

Certaines méthodes d'analyses exploratoires peuvent être envisagées en première approche pour extraire de l'information sur la séparation des espèces. Ainsi, l'analyse en composantes principales inter-classes peut apporter des informations sur les variables qui séparent le plus les espèces étudiées (e.g. M *et al.* 1999). Cette analyse est très simple à mettre en œuvre (fonction `between()` de la bibliothèque **ade4**). Elle correspond à l'ACP du tableau contenant les moyennes des variables environnementales (en colonnes) pour chacune des espèces (en

lignes). Cette analyse maximise donc la variance inter-classes, c'est à dire la séparation des centres de gravité des niches des espèces. Mais séparer les centres de gravité n'est pas synonyme de discrimination. Une telle question appelle naturellement l'utilisation de l'analyse discriminante (cf. § 5.2).

Soit **Z** un tableau donnant la composition des variables environnementales aux différentes occurrences des espèces étudiées, et **f** une variable qualitative décrivant l'appartenance de ces occurrences aux différentes espèces. L'analyse discriminante de **f** par **Z** renvoie une combinaison linéaire des colonnes de **Z** qui discrimine au maximum les espèces étudiées. Cette analyse est par conséquent optimale pour séparer les espèces du point de vue de leurs exigences écologiques (G 1971, P 1984).

7.1.2 L'analyse canonique des correspondances

Après la soumission de l'article sur la distribution des espèces arborées au Paraguay, qui introduit l'analyse discriminante sur vecteurs propres du graphe de voisinage (S *et al.* 2006b, § 5.4.2, Annexes 4 et 5), l'un des experts de la revue *Candollea* nous a fait la remarque suivante :

"The statistical analysis is new to me and I find the explanation difficult to understand. (...) As an alternative, I would suggest they adopt the standard methods used for this type of analysis, both of which are both quantitative and multidimensional. Those methods are called Detrended Correspondence Analysis (DCA) and Canonical Correspondence Analysis (CCA) (...). The CCA analyses species with distributions correlated with the environmental variables. The environmental variables are typically organized in a second matrix where the variables are rows and the columns are localities. (...) the axes can be interpreted with respect to the environmental variables, because the CCA generates an output matrix that shows the correlation coefficient between the environmental variable and each ordination axis. The CCA is essentially a gradient analysis and is ideal for evaluating how the distributions of species are related to environmental variables. DCA and CCA are standard methods used by dozens (hundreds) of community ecologists ; they are easy to use, easy to interpret and statistically robust."

En effet, l'analyse canonique des correspondances (ACC) mérite d'être développée en détail dans le cadre de cette thèse, du fait de sa très fréquente utilisation dans de nombreux domaines de l'Ecologie (T B 1986, 1987, C *et al.* 1987, T B et P 1988, L - *et al.* 1988a, b, H 1991, B *et al.* 1992, P 1993, B et L 1994, J ¨ et S 1998, D 1999, J et G 2001). En outre, son objectif est précisément de mettre en relation la distribution de plusieurs espèces sur une zone avec des variables environnementales. Une réflexion sur cette méthode permet de répondre aux critiques de cet expert.

Cette analyse est liée de très près à l'AFC dont elle est une extension. Certains auteurs la nomment même AFC sur variables instrumentales (L *et al.* 1991). On dispose d'une part d'un tableau donnant le nombre d'individus de chaque espèce (colonnes) dans chaque quadrat

(lignes) d'une grille superposée à la zone d'étude. D'autre part, on dispose d'un certain nombre de mesures environnementales pour chacun de ces quadrats. Ces mesures sont arrangées dans un tableau quadrats (lignes) × variables environnementales (colonnes). L'ACC permet de trouver des combinaisons linéaires des variables environnementales qui maximisent la séparation des espèces.

7.1.2.1 Principe mathématique

L'article d'origine ayant introduit l'ACC en Ecologie décrivait un algorithme récursif permettant l'application de cette méthode (T B 1986). Toutefois, C *et al.* (1987) ont pu montrer que l'ACC peut être exprimée dans le cadre des analyses reposant sur le schéma de dualité. C'est cette approche que nous présentons ici.

Reprenons les notations du § 5.3.1. Rappelons rapidement en quoi consistent ces notations. Soit **N** le tableau quadrats × espèces indiquant le nombre d'occurrences des S espèces dans chacun des Q quadrats d'une grille virtuelle placée sur la zone d'étude. On calcule alors le tableau $\mathbf{P} = \mathbf{N}/N$, où N est le nombre total d'occurrences dénombrées sur la zone. On note **P** la matrice $Q \times S$ contenant les p_{ij}, $\mathbf{D}_Q$ et $\mathbf{D}_S$ les matrices diagonales

$$\mathbf{D}_Q = Diag(p_{1\bullet}, ..., p_{Q\bullet}) \qquad \mathbf{D}_S = Diag(p_{\bullet 1}, ..., p_{\bullet S})$$

Avec $p_{i\bullet}$ la proportion du nombre total d'occurrences tombant dans le quadrat i et $p_{\bullet j}$ la proportion du nombre total d'occurrences appartenant à l'espèce j. Par ailleurs on dispose d'un tableau **Z** ($Q \times V$) donnant la valeur des V variables environnementales (colonnes) mesurées dans les Q quadrats (lignes). On suppose ces variables centrées et réduites. On peut alors calculer le tableau **L** ($S \times V$) :

$$\mathbf{L} = \mathbf{D}_S^{-1}\mathbf{P}^t\mathbf{Z}$$

Ce tableau contient, à l'intersection de la ligne j et de la colonne k la moyenne de la variable k pour les occurrences de l'espèce j. Enfin, on peut calculer la matrice de variances-covariances des variables environnementales par $\mathbf{V} = \mathbf{Z}^t\mathbf{D}_Q\mathbf{Z}$.

L'ACC correspond alors à l'analyse du triplet $(\mathbf{L}, \mathbf{V}^{-1}, \mathbf{D}_S)$ (C *et al.* 1987, D *et al.* 2000). On diagonalise alors la matrice **M**

$$\mathbf{M} = \mathbf{V}^{-1/2}\mathbf{L}^t\mathbf{D}_S\mathbf{L}\mathbf{V}^{-1/2}$$

Si cette matrice est de rang r, alors elle admet une base de r vecteurs propres $\mathbf{e}_k$ ($k = 1, ..., r$) de longueur V, $\mathbf{V}^{-1}$ orthonormés. r valeurs propres sont associées à ces vecteurs. On peut alors calculer r vecteurs $\mathbf{a}_k$:

$$\mathbf{a}_k = \mathbf{Z}\mathbf{V}^{-1}\mathbf{e}_k$$

Ces vecteurs sont stockés dans la matrice $\mathbf{A} = [\mathbf{a}_1, ..., \mathbf{a}_r]$. Ils contiennent des scores associés aux quadrats. Or, on peut calculer les moyennes conditionnelles de ces scores par espèce par :

$$\mathbf{g}_k = \mathbf{D}_S^{-1}\mathbf{P}^t\mathbf{a}_k$$

La variance de ces moyennes conditionnelles est obtenue par :

$$\begin{aligned} v_k &= \mathbf{g}_k^t\mathbf{D}_S\mathbf{g}_k \\ &= \mathbf{a}_k^t\mathbf{P}\mathbf{D}_S^{-1}\mathbf{P}^t\mathbf{a}_k \\ &= \mathbf{e}_k^t\mathbf{V}^{-1}\mathbf{Z}^t\mathbf{P}\mathbf{D}_S^{-1}\mathbf{P}^t\mathbf{Z}\mathbf{V}^{-1}\mathbf{e}_k \\ &= \mathbf{e}_k^t\mathbf{V}^{-1}\mathbf{L}^t\mathbf{D}_S\mathbf{L}\mathbf{V}^{-1}\mathbf{e}_k \end{aligned}$$

Or, l'ACC reposant sur le schéma de dualité, une des propriétés des vecteurs $\mathbf{e}_k$ est qu'ils maximisent la quantité $\|\mathbf{L}\mathbf{V}^{-1}\mathbf{e}_k\|^2_{\mathbf{D_S}}$, c'est-à-dire v_k (E 1987). Le maximum est atteint pour la première valeur propre de l'analyse. Ainsi, l'ACC permet de trouver une combinaison des variables environnementales (scores) dont la variance des moyennes conditionnelles par espèce est maximale. On maximise ainsi la séparation des espèces en fonction des variables de milieu.

7.1.2.2 Relations avec l'analyse discriminante

Plusieurs auteurs ont noté que l'ACC est une analyse discriminante (C *et al.* 1987, L *et al.* 1988a). Nous redémontrons cette équivalence ci-dessous. On peut, à partir des deux tableaux **N** et **Z** construire une matrice **Y** ($N \times S$) décrivant l'appartenance (0 ou 1) des N occurrences (en lignes) aux S espèces, et un tableau **X** ($N \times V$) donnant la valeur des variables environnementales mesurées dans le quadrat auquel appartient chacune des N occurrences. Soit **D** la matrice diagonale ($N \times N$) contenant les poids associés à chaque occurrence (par défaut, $1/N$ pour toutes les occurrences). Alors, on peut noter les égalités suivantes :

$$\begin{aligned} \mathbf{D}_S &= \mathbf{Y}^t\mathbf{D}\mathbf{Y} \\ \mathbf{V} &= \mathbf{X}^t\mathbf{D}\mathbf{X} \\ \mathbf{L} &= \mathbf{D}_S^{-1}\mathbf{Y}^t\mathbf{D}\mathbf{X} \end{aligned}$$

Par ailleurs, on peut montrer qu'il y a identité entre l'analyse du triplet $(\mathbf{L}, \mathbf{V}^{-1}, \mathbf{D}_P)$ et celle du triplet $(\mathbf{V}^{-1}\mathbf{L}, \mathbf{V}, \mathbf{D}_P)$. Dans les deux cas, la matrice diagonalisée est la matrice **M**, et de fait, les vecteurs propres des deux analyses sont identiques. Alors, si l'on reprend les notations utilisées dans le § 5.2 pour l'explication de l'analyse discriminante, on peut montrer les identités suivantes :

$$\begin{aligned} \mathbf{K} &= \mathbf{V}^{-1}\mathbf{L} \\ \mathbf{D}_m &= \mathbf{D}_S \\ \mathbf{T} &= \mathbf{V} \end{aligned}$$

L'ACC correspond donc à l'analyse du triplet $(\mathbf{K}, \mathbf{T}, \mathbf{D}_m)$. Le schéma de l'ACC est donc exactement identique à celui de l'analyse discriminante des espèces par les variables environnementales. L'ACC trouve la combinaison linéaire des variables environnementales qui maximise la discrimination des niches des espèces. Par conséquent, lorsqu'on analyse une liste d'occurrences d'espèces, il est parfaitement inutile de discrétiser la distribution à l'aide d'une grille de quadrats ; une telle opération se traduit nécessairement par une perte de précision. Bien sûr, cette opération est indispensable lorsque les données environnementales proviennent elles-mêmes d'une discrétisation de la zone d'étude, c'est-à-dire provenant d'une carte raster multicouches. Mais même dans ce cas-là, nous préfèrons voir l'ACC comme une analyse discriminante, à cause de ses propriétés plus générales.

7.2 L ´ ’

La question de la sélection de l'habitat est différente. Il ne s'agit plus de rechercher ce qu'il y a de différent entre les niches écologiques, mais au contraire de mettre en évidence ce qu'il peut y avoir de commun entre elles. Sont-elles situées dans la même partie de l'espace écologique ? sont-elles recouvrantes ? Sont-elles de même taille ?

Les études sur la sélection de l'habitat impliquant l'étude de plusieurs niches écologiques (au sens où nous l'avons définie dans le § 6.1) se focalisent en général sur une seule espèce. Ce type de données est généralement obtenu à la suite d'une étude de radio-pistage. Pour cette raison, nous modifions pour cette section la terminologie employée. Les différentes catégories de points sont appelées "*individus*" ou "*animaux*" et les points en eux-mêmes sont appelés "*localisations*". Mais ici encore, gardons à l'esprit que ce choix terminologique n'a pour but que de simplifier le discours, et que les individus peuvent tout aussi bien faire référence à d'autres types de catégories (espèces, classes d'âge, cf. chapitre 8 pour un exemple).

Par essence, l'étude de la sélection de l'habitat repose sur la comparaison des caractéristiques de l'environnement utilisées par les individus avec celles qui leur sont disponibles. Le concept d'échelle est essentiel dans ce type d'étude (§ 2.1.2), et tout particulièrement lorsque celle-ci est étudiée à l'aide de données de radio-pistage (§ 2.3.2). Dans cette section, nous détaillons des méthodes d'analyse factorielle qui permettent des représentations de l'espace écologique optimales selon certains critères.

Dans le cadre de l'étude d'une seule niche écologique, nous nous sommes servis de deux paramètres descriptifs de la niche écologique pour mettre en évidence cette sélection, la marginalité et la tolérance (§ 6.1). La marginalité mesure la position de la niche dans l'espace écologique par rapport au nuage de points disponible, et la tolérance mesure le "volume" de la niche. Dans cette section, nous ne nous servons que de la marginalité. En effet, un tel choix permet de s'appuyer sur les études précédentes menées sur plusieurs niches écologiques (D *et al.* 2000).

Comme l'ont noté A *et al.* (1993), dans les études reposant sur des données de

radio-pistage, l'unité d'échantillonnage n'est pas la localisation, mais l'individu. Alors, afin de mettre en évidence la sélection de l'habitat par les individus, nous avons choisi de rechercher ce qu'il y a de commun dans l'orientation et la taille de ces vecteurs de marginalité entre les individus. Tous les individus ne sélectionnent pas nécessairement les mêmes habitats, et il existe peut-être des "types" de sélection de l'habitat (les individus femelles qui élèvent des jeunes ont sans doute d'autres besoins que les mâles ; un individu suivi en période de disette alimentaire n'adopte peut-être pas le même comportement que pendant une période où les ressources sont abondantes). Cette variabilité de la sélection de l'habitat sera peut-être exprimée au niveau des vecteurs de marginalité, qui prendront des directions différentes. Nous décrivons ici trois approches qui reposent sur cette idée : l'analyse OMI, l'analyse K-select et l'analyse factorielle des rapports de sélection. Les deux premières sont explicitement des analyses de la marginalité des individus, et la troisième est liée de près aux deux autres.

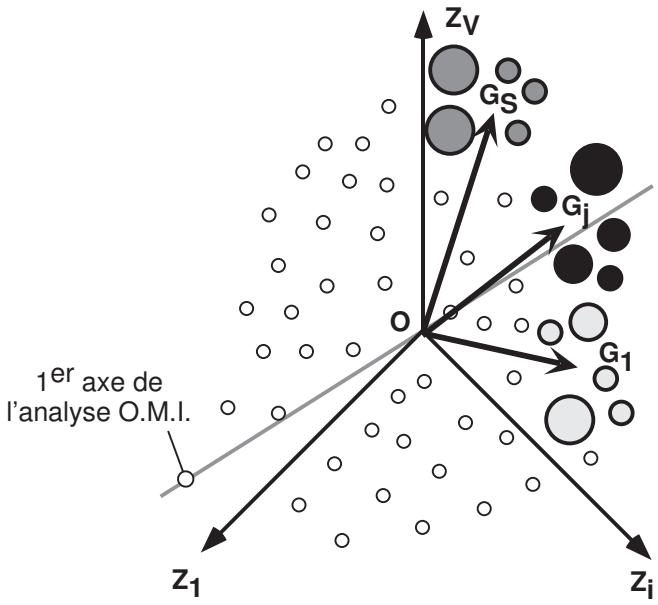
7.2.1 L'analyse de la marginalité

Deux approches peuvent être distinguées dans les études de radio-pistage, en fonction de l'échelle à laquelle est menée l'étude : les protocoles de type II et les protocoles de type III (§ 2.3.2). Nous décrivons comment l'analyse de la marginalité peut être effectuée dans ces deux types d'études.

7.2.1.1 Les protocoles de type II : l'analyse OMI

Dans les protocoles de types II, on dispose d'un nuage de points dans l'espace écologique et d'une mesure de l'utilisation pour chaque individu (figure 25). Nous décrivons ici l'analyse OMI (*Outlying Mean Index*, l'un des termes anglo-saxons pour désigner la marginalité), une analyse de la marginalité développée par D *et al.* (2000). Bien que développée à l'origine en Ecologie des communautés, cette analyse est parfaitement adaptée à l'analyse des protocoles de type II. Nous en décrivons le principe ci-dessous.

Nous reprenons les notations des paragraphes précédents. On dispose d'un échantillon de localisations de S individus sur une zone. On dispose par ailleurs d'une carte raster "multicouches" de cette zone donnant la valeur de V variables environnementales dans Q pixels. Soit $\mathbf{N}$ la matrice $Q \times S$ donnant le nombre de localisations de chaque individu dans chaque pixel de la carte. Soit $\mathbf{P} = \mathbf{N}/N$, où N est le nombre total de localisations collectées, tous individus confondus. Soit $\mathbf{Z}$ la matrice $Q \times V$ contenant les valeurs des variables environnementales pour chacun des pixels de la carte. On peut associer aux colonnes de $\mathbf{Z}$ une matrice de pondération des variables environnementales $\mathbf{Q}$ ($V \times V$). Dans le cas où les variables étudiées sont toutes des variables quantitatives centrées réduites, alors toutes les variables ont le même poids, et la matrice $\mathbf{Q}$ est tout simplement la matrice identité $\mathbf{I}_V$ ($V \times V$). Notons que cette matrice peut être différente si l'on dispose d'un mélange de variables quantitatives et qualitatives, ou si l'on ne travaille que sur des variables qualitatives (§ 7.2.2.3). Enfin, soit $\mathbf{D}_S$ la matrice diagonale $S \times S$ contenant les proportions du nombre total de localisations collectées pour chaque individu. On peut alors calculer la matrice $\mathbf{L}$:



F . *25 – Principe de l'analyse OMI (*D *et al. 2000). A chacun des points disponibles de l'espace écologique est associé un poids d'utilisation par l'individu* j *(avec* $j = 1, ..., S$*). Pour l'animal* j*, un vecteur de marginalité est calculé dans cet espace, reliant le barycentre de l'espace disponible* $\mathbf{O}$ *au barycentre de la niche* $\mathbf{G}_j$*. L'analyse OMI recherche un système d'axes qui maximise la marginalité moyenne sur les premiers axes.*

$$\mathbf{L} = \mathbf{D}_S^{-1}\mathbf{P}^t\mathbf{Z}$$

Cette matrice $S \times V$ contient à l'intersection de la ligne i et de la colonne j la moyenne des points utilisés par l'individu i pour la variable j, c'est-à-dire la coordonnée du vecteur de marginalité de l'individu i sur la variable j.

L'analyse OMI correspond alors à l'analyse du triplet $(\mathbf{L}, \mathbf{Q}, \mathbf{D}_S)$. L'analyse OMI est donc tout simplement l'analyse en composantes principales *non-centrée* du tableau donnant les coordonnées des vecteurs de marginalité des individus. Ce non-centrage est essentiel dans l'analyse, comme le souligne N -M (1973) :

"Mathematically and geometrically, centering involves the specification of the origin, the point of reference of the multivariate model. Ecologically too, it means the choice of a point of reference for the description of the vegetation. It is the 'point of zero information' ; anything that is at it, is trivial and uninteresting ; anything that deviates from it is information."

Dans le cas de l'étude de la sélection de l'habitat, il semble alors logique de choisir comme origine de l'espace écologique le barycentre des points disponibles. L'ACP non-centrée du tableau $\mathbf{L}$ renvoie donc des vecteurs contenant des scores pour les individus qui maximisent la marginalité expliquée. La signification de ces axes peut être interprétée grâce aux coordonnées des variables, comme dans une ACP classique. L'analyse OMI trouve donc des combinai-

sons linéaires des variables environnementales sur lesquelles l'éloignement entre ce qui est en moyenne disponible et ce qui est en moyenne utilisé par les individus est maximisé (figure 25). Cette analyse peut être effectuée grâce à la fonction `niche()` de la bibliothèque de fonctions **ade4** pour R.

7.2.1.2 Extensions aux designs de type III : l'analyse K-select

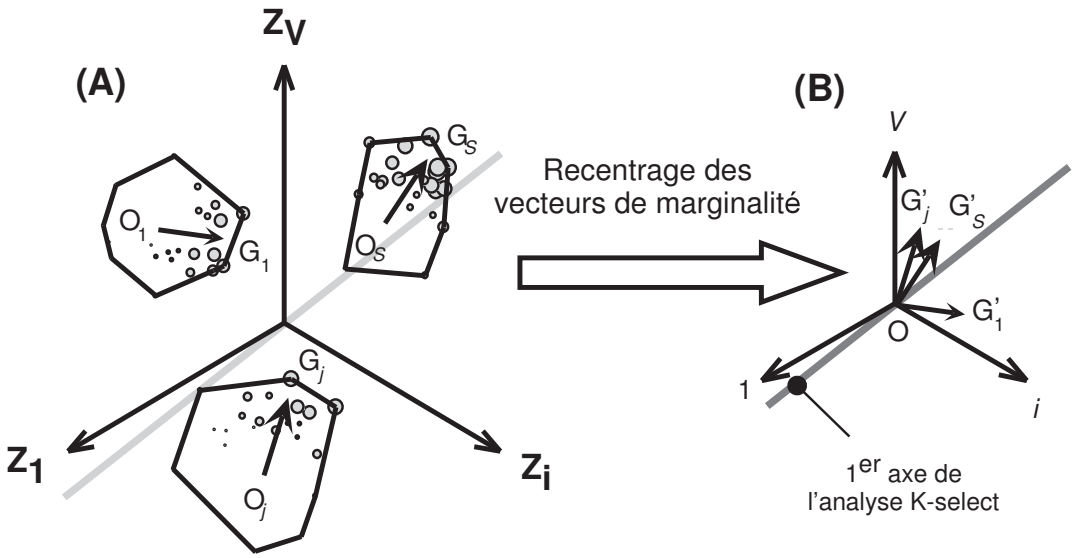
L'analyse de la marginalité dans les études de type III est plus compliquée. En effet, ce type de protocole suppose une mesure de l'utilisation *et* de la disponibilité *pour chaque individu*. Par exemple, pour un individu donné, la disponibilité peut être mesurée par les valeurs des variables environnementales dans les pixels appartenant à son domaine vital (Une stratégie par ailleurs souvent retrouvée dans la littérature, cf. par exemple A *et al.* 1993), et l'utilisation est mesurée par le nombre de localisations dans chacun de ces pixels.

Ainsi, on dispose dans l'espace écologique d'un nuage de points disponibles, ainsi que de poids d'utilisations *pour chaque individu*. La question est exactement identique à celle qui est posée par l'analyse OMI. Il s'agit d'établir une typologie des animaux en fonction de leur sélection de l'habitat, en se concentrant uniquement sur les vecteurs de marginalité des individus.

Pour répondre à cette question, nous avons développé une nouvelle méthode dans le cadre de cette thèse, l'analyse K-select, qui a fait l'objet d'un article publié dans *Ecological Modelling* (C *et al.* 2005a, Annexe 1). Cette méthode repose exactement sur le même principe que l'analyse OMI. La seule information intéressante pour nous ici est la taille et l'orientation des vecteurs de marginalité des animaux. On peut donc construire un tableau $\mathbf{L}^*$ $(S \times V)$, qui contient, à l'intersection de la ligne *i* et de la colonne *j* la différence entre la moyenne des points disponibles et des points utilisés pour la variable *j* et l'individu *i*, c'est-à-dire la coordonnée du vecteur de marginalité de l'individu *i* sur la variable *j, recentré de telle façon que tous les vecteurs de marginalité aient la même origine* (figure 26).

L'analyse K-select est alors l'analyse du triplet $(\mathbf{L}^*, \mathbf{Q}, \mathbf{D}_S)$, où la matrice $\mathbf{D}_S$ est calculée de la même façon que dans le paragraphe précédent (matrice diagonale des poids des individus), et **Q** est une matrice de pondération des variables. Comme pour l'analyse OMI, l'analyse K-select renvoie des combinaisons linéaires des variables environnementales qui maximisent la marginalité moyenne expliquée sur les premiers axes. Cette analyse peut être effectuée grâce à la fonction `kselect()` de la bibliothèque **adehabitat**. On peut ensuite projeter sur ces axes les vecteurs de marginalité *non-recentrés* (fonction `plot.kselect()`), afin de déterminer si la sélection des variables environnementales dépend de leur disponibilité (cas d'une une réponse fonctionnelle, cf. M et I 1998). Il est également possible de projeter sur ces axes les points disponibles et utilisés pour avoir une représentation graphiques de la niche des individus (fonction `kplot.kselect()`). Une illustration de cette analyse est fournie dans l'article introduisant la méthode (C *et al.* 2005a, Annexe 1).

Notons que pour le cas particulier des données de radio-pistage, la bibliothèque de fonctions **adehabitat** contient une classe d'objets appelée **"`sahrlocs`"** qui facilite l'application de l'analyse OMI et de l'analyse K-select. Un objet de la classe "`sahrlocs`" contient trois cartes



F . *26 – Principe de l'analyse K-select. (A) Chacun des S individus possède son propre espace disponible dans l'espace écologique défini par les variables* $\mathbf{z}_i$*, et chaque point disponible (non représentés ici dans un souci de lisibilité) est associé à un poids d'utilisation. Il est donc possible, pour un individu donné j, de calculer un vecteur de marginalité reliant le barycentre des points disponibles* $\mathbf{O}_j$ *au barycentre des points utilisés* $\mathbf{G}_j$*. La recherche de l'axe, dans l'espace écologique, qui explique le plus possible de la marginalité présente dans les données (en gris clair) passe par un recentrage. (B) Ce recentrage permet de donner une origine commune à tous les vecteurs de marginalité. L'analyse K-select consiste alors en une ACP non centrée du tableau des coordonnées des points* $\mathbf{G}'_k$*, c'est-à-dire une analyse OMI.*

"multicouches" possédant les mêmes attributs (couvrant la même zone, avec la même résolution) : (i) une carte décrivant la valeur des variables environnementales sur la zone d'étude, (ii) une carte décrivant la position des domaines vitaux sur cette même zone (e.g. calculés par les fonctions `mcp()` et `mcp.rast()`), et (iii) une carte donnant le poids d'utilisation associé à chacun des pixels (e.g. le nombre de localisations recensées dans chaque pixel à l'aide de la fonction `count.points.id()`). Les fonctions `sahrlocs2niche()` et `sahrlocs2kselect()` permettent de convertir cet objet au format nécessaire pour l'application respective de l'analyse OMI et de l'analyse K-select. Un certain nombre de fonctions sont associées à cette classe d'objet, qui permettent une exploration graphique des données (e.g. `image.sahrlocs()`, `plot.sahrlocs()`).

7.2.2 L'analyse factorielle des rapports de sélection

Une autre méthode mérite notre attention ici, l'analyse factorielle des rapports de sélection. Nous avons développé cette méthode dans le cadre de cette thèse pour analyser la sélection de l'habitat en se basant sur la distribution des localisations de plusieurs animaux sur une zone, *avec une description de l'environnement par une variable qualitative* (e.g. des types de végétation). Cette méthode a fait l'objet d'un article actuellement soumis dans la revue *Ecology*

(C et D soumis), consigné en annexe 2.

Cette analyse généralise deux approches communément utilisées pour étudier la sélection de l'habitat avec ce type de protocole, les rapports de sélection de M *et al.* (1972, § 2.3.1.1) et le test de N *et al.* (1974) modifié par W et G (1990, § 2.3.2), dans le cadre des analyses factorielles. Après une brève présentation du principe de cette méthode, nous montrons qu'elle est liée de près aux analyses OMI et K-select présentées dans le paragraphe précédent.

7.2.2.1 Principe Mathématique

L'analyse factorielle des rapports de sélection peut être effectuée sur des protocoles de type II et III, mais nous ne présentons ici que l'analyse des protocoles de type II. Une extension de ce raisonnement pour les protocoles de type III est développée dans l'article en annexe 2.

On dispose des localisations de S animaux suivis sur une zone comportant V types d'habitats. Soit p_j la proportion de la zone d'étude recouverte par le type d'habitat j. Nous nommons cette proportion "*proportion disponible*" dans la suite de cette section. Soit u_{ij} le nombre de localisations de l'animal i dans l'habitat j, $u_{i\bullet}$ le nombre total de localisations de l'animal i, $u_{\bullet j}$ le nombre de localisations dans l'habitat de type j, tous animaux confondus, et N le nombre total de localisations collectées sur tous les animaux. Pour un animal i et un type d'habitat j donnés, le rapport de sélection de M *et al.* (1972) est calculé par :

$$w_{ij} = \frac{u_{ij}}{u_{\bullet j} p_j} \tag{7.1}$$

Sous l'hypothèse d'une utilisation aléatoire de l'habitat par les animaux, ces rapports devraient être égaux à 1. Ils sont supérieurs à 1 lorsque les habitats sont recherchés, et inférieurs lorsque les habitats sont évités.

La statistique de N *et al.* (1974) modifiée par W et G (1990) permet de tester l'existence d'une sélection de l'habitat :

$$\chi^2_j = \sum_{j=1}^{V} \frac{(u_{ij} - p_j u_{i\bullet})^2}{p_j u_{i\bullet}}$$

En effet, sous l'hypothèse d'une utilisation aléatoire de l'habitat, cette statistique est distribuée selon une loi du χ^2 à $(S-1)$ degrés de liberté. On peut reformuler cette statistique en utilisant les rapports de sélection de M *et al.* définis par l'équation 7.1 :

$$\chi^2_{WG} = \sum_{j=1}^{V} \sum_{i=1}^{S} p_j u_{i\bullet} (w_{ij} - 1)^2$$

Pour effectuer l'analyse factorielle des rapports de sélection, nous avons besoin de trois matrices. La matrice $\mathbf{W}$ $(S \times V)$ contient les rapports de sélection centrés sous l'hypothèse d'une utilisation aléatoire de l'habitat, c'est-à-dire

$$\mathbf{W} = \left[w_{ij} - 1 \right]_{i=1\ldots S, j=1\ldots V}$$

Dans l'espace multidimensionnel défini par les types d'habitat, l'origine correspond donc à un animal fictif qui utiliserait tous les types d'habitat de façon totalement aléatoire. Dans l'espace défini par les individus, l'origine correspond à un type d'habitat fictif qui serait utilisé en proportion de sa disponibilité par tous les animaux. Par ailleurs, la matrice **D** contient les poids associés aux lignes de **W**, dans ce cas le nombre de localisations par animal :

$$\mathbf{D} = Diag(u_{1\bullet}...u_{i\bullet}...u_{S\bullet})$$

Enfin, la matrice **A** contient les poids associés aux colonnes de **W**, c'est-à-dire la disponibilité des types d'habitat :

$$\mathbf{A} = Diag(p_1...p_j...p_V)$$

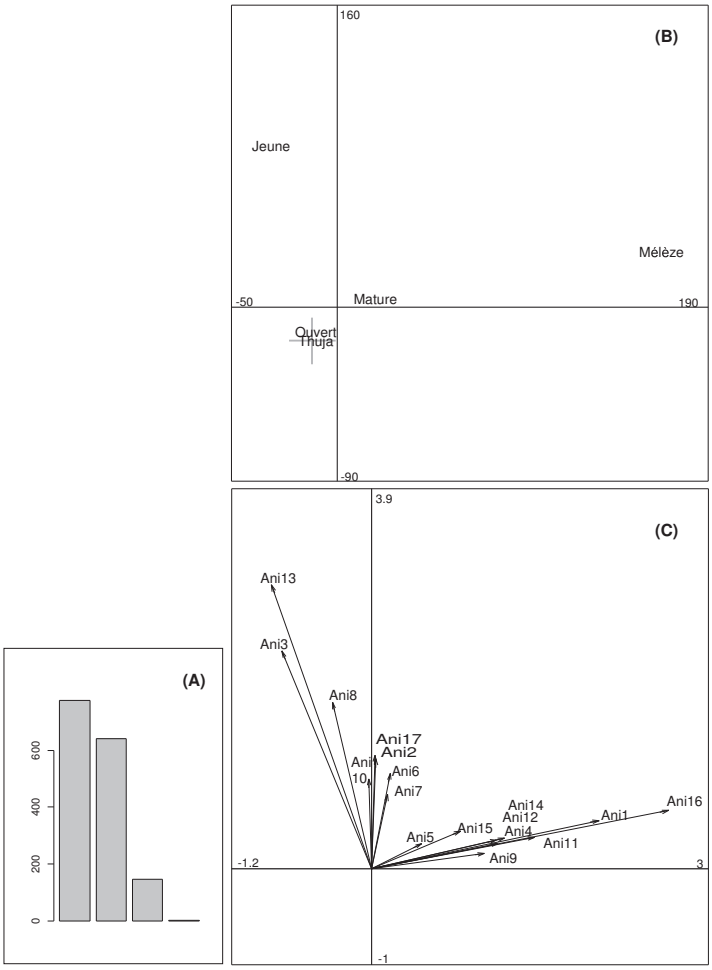
L'analyse factorielle des rapports de sélection est alors l'analyse du triplet (**W**, **A**, **D**). L'inertie de cette analyse est égale à la trace de la matrice $\mathbf{W}^t\mathbf{DWA}$, c'est-à-dire :

$$\text{Inertie} = \sum_{j=1}^{V}\sum_{i=1}^{S} p_j u_{i\bullet}(w_{ij}-1)^2 = \chi^2_{WG}$$

Ainsi, l'analyse en composantes principales du tableau contenant les rapports de sélection de M *et al.* pour chaque animal et chaque habitat, pondérés respectivement par le nombre de localisations des animaux et par la proportion disponible des habitats *maximise la statistique de* N *et al. (1974) modifiée par* W et G *(1990)*. Il y a donc une grande cohérence entre les deux approches, lorsqu'elles sont considérées du point de vue de l'analyse factorielle.

L'intérêt majeur de cette analyse est, comme pour les autres analyses factorielles présentées dans ce mémoire, qu'elle permet une exploration de la sélection de l'habitat sans aucune hypothèse préalable sur les données. L'application de cette analyse sur le très célèbre jeu de données d'A *et al.* (1993) permet de se rendre compte immédiatement des avantages de cette approche. Ce jeu de données contient la distribution des localisations de 17 écureuils dans 5 types d'habitat (disponible dans **adehabitat** sous le nom de `squirrel`).

La figure 27 présente les résultats de l'analyse factorielle sur ces données. Deux groupes d'animaux sont séparés sur le premier plan de l'analyse. Il existe donc deux stratégies de sélection de l'habitat par ces animaux. Cet aspect des données n'avait pas été mis en évidence A *et al.* (1993). En effet, l'analyse compositionelle, comme beaucoup de méthodes d'analyse de la sélection de l'habitat, suppose l'homogénéité de l'échantillon du point de vue de la sélection. Notons toutefois qu'il est possible de tester l'existence d'une différence de stratégie d'utilisation de l'habitat entre différents groupes à l'aide de l'analyse compositionelle, mais comme toutes les méthodes reposant sur le modèle linéaire, les groupes doivent être définis *a priori*. L'analyse factorielle des rapports de sélection permet d'établir une typologie après exploration des données, ce qui permet une meilleure compréhension des données (à ce sujet, voir le § 2.4.3 sur les méthodes exploratoires et confirmatoires).



F . *27 – Résultats de l'analyse factorielle des rapports de sélection appliquée sur le tableau donnant la distribution des localisations de 17 écureuils dans 5 types d'habitat. (A) Les valeurs propres de l'analyse ; deux axes expliquent la majeure partie de la sélection de l'habitat. (B) Coordonnées des types d'habitat. La croix grise indique la position d'un habitat fictif qui serait rejeté par tous les animaux (i.e. utilisation nulle). (C) Coordonnées des individus.*

7.2.2.2 Relations avec l'AFC

L'analyse factorielle des rapports de sélection a été initialement introduite en Ecologie sous le nom d'analyse factorielle des correspondances décentrée (AFCD), pour permettre l'analyse de la distribution de peuplements ichtyologiques à l'aide de campagnes de pêche électrique (D ´ *et al.* 1995). Lors de chaque campagne de pêche électrique j ($j = 1...V$), un nombre r_j relevés ponctuels est effectué (variant entre les campagnes de pêches), chacune des S espèces pêchées est notée. Un tableau de contingence **U** est alors construit, donnant le nombre de poissons appartenant à chaque espèce (en lignes) pêchés lors de chaque campagne (en colonnes).

Ce tableau de contingence appelle immédiatement l'utilisation de l'analyse factorielle des correspondances. Soit u_{ij} le nombre de poissons de l'espèce i pêchés lors de la campagne j,

$u_{i\bullet}$ le nombre total de poissons de l'espèce i, $u_{\bullet j}$ le nombre total de poissons pêchés lors de la campagne j, et N le nombre total de poissons pêchés lors de toutes les campagnes. On définit alors les matrices suivantes :

$$\begin{aligned} \mathbf{\Gamma} &= \frac{1}{N}\mathbf{U} \\ \mathbf{D}_V &= Diag(u_{\bullet 1}...u_{\bullet j}...u_{\bullet V}) \\ \mathbf{D}_S &= Diag(u_{1\bullet}...u_{i\bullet}...u_{S\bullet}) \end{aligned}$$

L'AFC "classique" correspond alors à l'analyse du triplet $(\mathbf{D}_S^{-1}\mathbf{\Gamma}\mathbf{D}_V^{-1}, \mathbf{D}_V, \mathbf{D}_S)$. Cependant, une telle analyse ne prend pas en compte la variabilité du nombre de relevés ponctuels entre les campagnes. Afin de corriger ce défaut de l'analyse, D ´ *et al.* ont simplement modifié la pondération associée aux relevés dans l'analyse. La matrice $\mathbf{D}_V$ est alors remplacée par la matrice diagonale $\mathbf{R}$:

$$\mathbf{R} = Diag(r_1/R...r_j/R...r_V/R)$$

où R est le nombre total de relevés effectués lors de toutes les campagnes. L'AFC ainsi modifiée correspond alors à l'analyse du triplet $(\mathbf{D}_S^{-1}\mathbf{\Gamma}\mathbf{R}^{-1} - \mathbf{1}_{SV}, \mathbf{R}, \mathbf{D}_S)$.

Dans la section 5.3, nous avions indiqué que la matrice principale du triplet de l'AFC était $\mathbf{D}_S^{-1}\mathbf{\Gamma}\mathbf{R}^{-1}$. Ici, nous centrons cette matrice en lui soustrayant le terme $\mathbf{1}_{SV}$. Les deux analyses sont en fait équivalentes. Dans la première, le vecteur $\mathbf{1}_S$ est un vecteur propre associé à la valeur propre 1, et il faut l'éliminer volontairement de l'analyse dans la mesure où il n'a aucun intérêt. Dans la seconde, ce même vecteur propre est associé à la valeur propre 0, et donc s'élimine tout seul. Pour le reste, les résultats sont identiques entre les deux analyses.

Si l'on établit un parallèle entre la problématique de l'analyse des peuplements ichtyologiques et celle de l'analyse de la sélection de l'habitat, on remarque que l'AFC décentrée est identique à l'analyse factorielle des rapports de sélection : si l'on assimile les campagnes de relevés aux types d'habitat, les relevés aux pixels de la carte de la zone d'étude, et les espèces aux individus suivis par radio-pistage, on peut noter les identités suivantes :

$$\begin{aligned} \mathbf{W} &= \mathbf{D}_S^{-1}\mathbf{\Gamma}\mathbf{R}^{-1} \\ \mathbf{A} &= \mathbf{R} \\ \mathbf{D} &= \mathbf{D}_S \end{aligned}$$

A propos de cette analyse, D ´ *et al.* (1995) indiquent : "*Contrairement à l'AFC, pour laquelle l'élément de référence dépend donc du contenu du tableau étudié, l'AFCD définit le point de référence a priori en prenant en compte la situation expérimentale des objets étudiés. (...) L'AFCD doit être considérée comme un cas particulier de la famille des AFC sur modèles qui n'a pas, à notre connaissance, été introduite en écologie.*" En effet, le simple fait de modifier la matrice de pondération des colonnes se traduit par un changement de point de référence dans l'espace multidimensionnel défini par les types d'habitat. Ce changement d'origine de l'espace, comme nous le montrons dans la section suivante, fait de cette analyse un cas particulier de l'analyse OMI.

7.2.2.3 Relations avec l'Analyse OMI

En effet, l'analyse factorielle des rapports de sélection peut également être exprimée sous la forme d'une analyse OMI particulière. Reformulons le problème de la façon suivante : soit **N** la matrice $Q \times S$ donnant le nombre de localisations des individus (en colonnes) dans chacun des Q pixels (en lignes) de la carte de la zone d'étude. Par ailleurs, soit **Z** la matrice $Q \times V$ décrivant l'appartenance (0 ou 1) d'un pixel à un type d'habitat. On peut alors montrer les identités suivantes :

$$\begin{aligned} \mathbf{D} &= Diag(\mathbf{N}^t\mathbf{1}_Q) \\ \mathbf{A} &= \frac{1}{Q}\mathbf{Z}^t\mathbf{Z} \\ \mathbf{W} &= \mathbf{D}^{-1}\mathbf{N}^t\mathbf{Z}\mathbf{A}^{-1} \end{aligned}$$

L'analyse factorielle des rapports de sélection est donc l'analyse du triplet :

$$(\mathbf{D}^{-1}\mathbf{N}^t\mathbf{Z}\mathbf{A}^{-1}, \mathbf{A}, \mathbf{D})$$

Notons que l'analyse factorielle des rapports de sélection est strictement équivalente à l'analyse du triplet :

$$(\mathbf{D}^{-1}\mathbf{N}^t\mathbf{Z}, \mathbf{A}^{-1}, \mathbf{D})$$

En effet, dans les deux cas, la matrice diagonalisée est la matrice :

$$\mathbf{D}^{-1}\mathbf{N}^t\mathbf{Z}\mathbf{A}^{-1}\mathbf{Z}^t\mathbf{N}$$

Par ailleurs, l'analyse OMI des deux tableaux **N** et **Z** est l'analyse du triplet $(\mathbf{L}, \mathbf{Q}, \mathbf{D}_S)$, avec :

$$\begin{aligned} \mathbf{D}_S &= \frac{1}{N}\mathbf{D} \\ \mathbf{L} &= \frac{1}{N}\mathbf{D}\mathbf{N}^t\mathbf{Z} \end{aligned}$$

On définit ici comme poids des colonnes de **V** la matrice **Q** :

$$\mathbf{Q} = \mathbf{A}^{-1}$$

On peut alors reformuler cette analyse OMI comme l'analyse du triplet :

$$(\frac{1}{N}\mathbf{D}^{-1}\mathbf{N}^t\mathbf{Z}, \mathbf{A}^{-1}, \frac{1}{N}\mathbf{D})$$

On peut choisir d'ignorer la constante $1/N$ (le nombre total de localisations reste fixe quel que soit la distribution de ces localisations sur la zone). Alors, il apparaît que l'analyse factorielle des rapports de sélection n'est qu'un cas particulier de l'analyse OMI, avec une métrique du

χ^2 (la matrice $\mathbf{A}^{-1}$) associée aux variables environnementales. C'est également à cette métrique que l'on doit les relations de parenté entre cette méthode et l'AFC.

7.3 C

L'étude de la distribution de plusieurs catégories de points dans l'espace écologique peut soulever une très grande diversité de questions. Dans ce chapitre, nous avons illustré deux grands types de questions fréquemment posées dans ce type d'étude, à savoir la discrimination des niches et la sélection de l'habitat.

En ce qui concerne la discrimination, nous avons encore une fois souligné la place centrale de l'analyse discriminante. Que la séparation des espèces se fasse par des analyses dans l'espace géographique ou écologique, c'est toujours l'analyse discriminante qui est au cœur du problème. Lorsque la discrimination se fait dans l'espace géographique, la question se pose de savoir comment mesurer la position des individus, et plusieurs solutions sont décrites dans le paragraphe 5.4. Lorsqu'on travaille dans l'espace écologique, la grande difficulté est de choisir les variables environnementales descriptives de cet espace. En effet, comme le notent C *et al.* (1987), il est impératif que le nombre de variables environnementales soit très inférieur au nombre d'occurrences. En outre, il est indispensable que les variables environnementales ne soient pas trop corrélées entre elles (§ 5.4.1). Il ne faut donc pas ignorer les étapes préliminaires d'exploration des données.

Le cas de l'étude de la sélection de l'habitat est édifiant. En effet, nous avons décrit trois grandes méthodes pouvant être utilisées. Nous avons également pu montrer que ces trois analyses n'étaient que des variations mineures d'une seule et même méthode. L'analyse K-select, développée dans le cadre de cette thèse, est adaptée à l'étude de la marginalité lorsque l'utilisation et la disponibilité des habitats sont définies séparément pour chaque catégorie de points. Si la disponibilité des habitats est identique entre les catégories, alors l'analyse K-select se réduit à l'analyse OMI. Lorsque l'environnement est décrit par une unique variable qualitative, et que l'on choisit de pondérer chaque type d'habitat par l'inverse de sa proportion disponible, l'analyse OMI devient une analyse factorielle des rapports de sélection. Cette dernière méthode montre la très grande cohérence entre les indices de sélectivité développés par M *et al.* (1972) et le test de N *et al.* (1974), deux des mesures les plus fréquemment utilisées actuellement dans la littérature traitant de la sélection de l'habitat. En conséquence, l'analyse K-select est une analyse très générale qui englobe un très grand nombre de méthodes qui possèdent des propriétés optimales pour l'étude de la sélection de l'habitat.

Ces liens très solides qui unissent les analyses montrent la très grande cohérence du modèle statistique utilisé pour les construire. Le schéma de dualité permet de développer une quantité considérable d'analyses. La difficulté vient ensuite de la terminologie employée. En effet, on rencontre souvent des méthodes qui portent des noms différents en fonction des champs d'application dans lesquels elles sont utilisées (e.g. l'analyse factorielle des rapports de sélection et l'AFC décentrée, l'ACC et l'AFCVI, etc.). Par ailleurs, des variations mineures autour

d'une même méthode peuvent entraîner des changements de noms des analyses (e.g. l'ACC et l'analyse discriminante). Il est donc particulièrement difficile de trouver son chemin dans le labyrinthe généré par le schéma de dualité. Ce modèle général repose sur la définition de trois matrices. Il suffit de changer un seul paramètre à une analyse, de modifier légèrement une matrice, d'utiliser un triplet particulier sur une problématique différente, et l'analyse change de nom. Bien sûr, le biologiste peut comprendre le principe sur lequel repose une méthode particulière. Mais chaque jeu de données soulève des contraintes ou une problématique spécifique qui impliquent de construire une analyse particulière, et peut-être de modifier une méthode existante pour l'adapter à la question posée. Il est donc nécessaire de garder en tête ces relations étroites entre les différentes méthodes. C'est là le rôle du biométricien, et c'est ce que nous montrons dans la partie suivante.

Une application de la démarche

Chapitre 8

La distribution des mouflons dans le massif des Bauges

L'objectif de l'analyse des données écologiques est de construire un modèle d'un système complexe. D'après L et S (2004), un système complexe est un système qui change de nature lorsqu'un des éléments qui le constituent est retiré. Celui-ci *émerge* des interactions entre ses éléments constitutifs. Le but de l'analyse est donc la mise en évidence de cette complexité, ou plus précisément, de la nature de ces interactions. Mais les jeux de données diffèrent entre eux non seulement par la variabilité des systèmes qu'ils représentent, mais aussi par celle des méthodes mises en œuvre pour les collecter. Enfin, la diversité des problématiques biologiques ajoutent une couche supplémentaire de complexité à l'analyse.

Les données doivent donc guider l'analyse, et des méthodes spécifiques doivent être adaptées à chaque jeu de données si l'outillage statistique existant est insuffisant pour prendre en compte toutes les contraintes. Les méthodes que nous avons développées dans le cadre de cette thèse pour répondre à des problématiques particulières peuvent être trouvées en Annexe, avec les applications de ces méthodes aux problèmes qui les ont générées (l'analyse K-select, l'analyse factorielle des rapports de sélection, l'analyse discriminante sur vecteurs propres du graphe de voisinage).

Dans ce chapitre, nous illustrons en détail la pratique de l'analyse de données, grâce à l'étude de la distribution du mouflon (*Ovis ammon*) dans le massif montagneux des Bauges (Alpes Françaises). Ces données correspondent à la distribution d'un unique semis de points sur une zone auquel nous associons un certain nombre de cartes de variables environnementales. Il s'agit donc de données collectées selon un protocole de type I (§ 2.3). *A priori*, nous devrions donc plutôt utiliser les méthodes décrites dans les chapitres 4 et 6 pour analyser ces données. Pourtant nous montrons ici que la structure particulière des données implique le développement de nouvelles méthodes, et l'utilisation d'outils élaborés pour d'autres types de protocoles.

8.1 P ´

8.1.1 Cadre de l'étude : le programme herbivorie

La plupart des études à long terme (> 10 ans) menées sur les ongulés ont pour objectif de déterminer les caractéristiques du fonctionnement des populations d'une seule espèce. Le grand nombre de publications scientifiques générées par ces études font de certaines zones des lieux de référence pour la biologie d'une espèce. Ainsi, les réserves de Chizé (Deux-Sèvres, P - *et al.* 2003, 2005) ou de Trois-Fontaines (Haute Marne, G *et al.* 2003) sont des zones de référence pour l'étude des populations de chevreuil (*Capreolus capreolus*). De même pour le sanglier dans la forêt d'Arc-en-Barrois (Haute-Marne, G *et al.* 1988, C *et al.* 2002b) ou le mouflon dans le massif du Caroux (Hérault, G *et al.* 2005). La liste est longue, et en comparant les résultats obtenus sur une même espèce entre plusieurs zones, on peut se faire une idée des grands patrons de la biologie de l'espèce.

Or, chacune de ces zones est particulière, et l'espèce étudiée n'est qu'un élément de l'écosystème. D'une zone à l'autre, les comportements des espèces diffèrent ; cela tient à la complexité des écosystèmes, dont nous avons brièvement discuté en introduction de ce chapitre. Il y a encore peu de connaissances sur la façon dont s'organisent les populations de plusieurs espèces sur une même zone. Or, l'étude des interactions interspécifiques est une question qui préoccupe de plus en plus les biologistes, en particulier en ce qui concerne les herbivores. Cette question a de sérieuses implications dans le domaine de la gestion de la faune sauvage. Par exemple, l'impact que peut avoir une espèce sur une autre peut limiter les chances de succès des opérations de réintroduction d'une espèce dans une zone où elle a disparu. De même, la compétition entre espèces sauvages et espèces domestiques est du plus grand intérêt pour limiter les conflits entre l'homme et la faune sauvage.

Mais les herbivores ne sont pas les seuls à être au centre des préoccupations des biologistes. En effet, le loup (*Canis lupus*) est actuellement en train de recoloniser le massif alpin, ce qui cause de vives polémiques parmi les utilisateurs de cette zone (éleveurs, chasseurs, associations de protection de la nature, etc.). Nombreux sont ceux qui souhaiteraient connaître l'impact réel de cette espèce sur le comportement, la dynamique des populations, ou la sélection de l'habitat par les herbivores, qu'ils soient sauvages ou domestiques. D'où l'idée de travailler non plus sur une seule espèce, mais sur un système composé de plusieurs espèces.

Cette idée a amené des spécialistes appartenant à plusieurs organismes (Centre national pour la recherche scientifique, Université de Savoie, Université de Lyon, Office national de la chasse et de la faune sauvage, Institut national de la recherche agronomique, Office national des forêts, CEMAGREF) à monter un programme d'étude nommé "herbivorie" dans le massif des Bauges (Savoie), en collaboration avec des organismes régionaux (Parc naturel des Bauges) et les populations locales (chasseurs, éleveurs, etc.). Ce programme, qui vient à peine d'être mis en place, a pour objectif l'étude des populations d'herbivores dans le massif des Bauges. Le choix de ce massif tenait surtout à ce que des études y sont menées depuis longtemps sur le chamois (*Rupicapra rupicapra*). Mais l'un des arguments principaux justifiant ce choix est que le loup n'a pas encore colonisé la zone. Ainsi, les données collectées aujourd'hui permettront la mise

en évidence de l'impact du loup sur les communautés d'herbivores de demain. Et les espèces d'herbivores rencontrées sur cette zone sont nombreuses, qu'elles soient sauvages (chamois, mouflons, chevreuils, cerfs), ou domestiques (vaches, chèvres, moutons).

La mise en place de ce programme a généré un certain nombre d'études sur des sujets aussi variés que la biodiversité, la compétition entre les ongulés sauvages et domestiques, la dynamique des maladies affectant la faune, la structuration spatiale au niveau génétique des peuplements d'ongulés, l'impact des ongulés domestiques et sauvages sur la forêt et les pâturages du massif, ou la structuration des paysages dans les Bauges.

8.1.2 L'étude de l'utilisation de l'espace par le mouflon

Parmi les nombreuses études menées dans le massif des Bauges, l'étude des interactions entre les ongulés est centrale. Cet intérêt s'est concrétisé par la mise en place d'une thèse actuellement menée par Gaëlle D (Laboratoire de Biométrie, Université Lyon 1), dont l'un des thèmes centraux est l'étude des interactions entre deux espèces d'herbivores rencontrées sur le massif : le chamois et le mouflon.

La cohabitation entre les ongulés sauvages utilisant les mêmes ressources est susceptible d'entraîner une compétition entre les deux espèces, qui peut s'exprimer de différentes façons (partage du temps et de l'espace, compétition par exploitation, etc.). Comprendre comment s'exprime cette cohabitation passe par l'étude de la distribution spatiale des deux espèces, l'étude des interactions entre ces distributions, la mise en évidence des caractéristiques de l'environnement qui les affectent et l'étude du régime alimentaire des deux espèces.

Afin de mener à bien un programme aussi vaste, Gaëlle D a entrepris un grand nombre d'études visant à permettre l'acquisition de ces connaissances. Dans le cadre de ces études, nous avons tous deux établi une collaboration pour permettre l'analyse de la distribution des chamois et des mouflons, et la mise en évidence des interactions entre ces deux distributions. Nous avons utilisé les résultats des recensements effectués par Jean-Michel J (ONCFS, directeur adjoint de la réserve nationale de chasse et de faune sauvage des Bauges) sur le massif chaque année au mois de juin, de 1980 à 2004. A l'heure de la rédaction de ce mémoire, ce travail est toujours en cours.

En effet, la question n'est pas simple. Chacune des deux distributions possède ses propres caractéristiques ; les processus qui les ont générées diffèrent. L'organisation sociale n'est pas la même pour les deux espèces, ce qui complique l'analyse. La zone d'étude possède une forme complexe, qu'il est difficile de gérer avec la plupart des méthodes présentées dans les parties précédentes. Les données sont structurées dans le temps, et cette dimension temporelle (24 années de données) ne doit pas être ignorée. Enfin, les connaissances biologiques sur le mouflon sont inexistantes dans ce massif, le programme venant juste d'être mis en place. Il est important de construire une démarche d'analyse qui prenne en compte toutes ces contraintes.

L'objectif de ce chapitre est de détailler comment se déroule en pratique une collaboration entre biométricien et biologiste. Il n'est pas d'apporter des réponses à cette problématique com-

plexe. Nous nous concentrons donc sur une petite fraction des analyses effectuées dans le cadre de cette collaboration. En effet, comme nous l'avons indiqué dans les parties précédentes (cf. chapitre 5), avant de se poser la question des interactions entre deux espèces, il est nécessaire de caractériser les propriétés des distributions de ces deux espèces séparément. Nous détaillons donc ici uniquement l'analyse de la distribution spatiale des mouflons, et l'étude de la sélection de l'habitat par cette espèce.

8.1.3 Les recensements du mouflon

La population actuelle du massif des Bauges est issue d'un lâcher de 16 individus effectué en 1954 et 1955 (figure 28). La population s'est bien développée depuis, et le mouflon est aujourd'hui présent dans tout le massif. Le recensement de juin 2004 a permis de dénombrer 427 mouflons, et ce nombre est certainement sous-estimé dans la mesure où ces animaux ne peuvent pas être observés dans les zones forestières.

Ces recensements sont effectués chaque année au mois de juin depuis 1980. Vingt-quatre circuits sont placés sur la zone d'étude de façon à ce que tous les mouflons présents en zones ouvertes soient détectés. Des observateurs (techniciens de l'ONCFS, de l'ONF, du parc, bénévoles) parcourent ces circuits et notent la position et la composition des groupes de mouflons détectés (nombre de mâles, de femelles et d'agneaux). Lors de ces opérations, les groupes de chamois sont également recensés et leur composition est notée. Les circuits sont présents sur toute la zone.

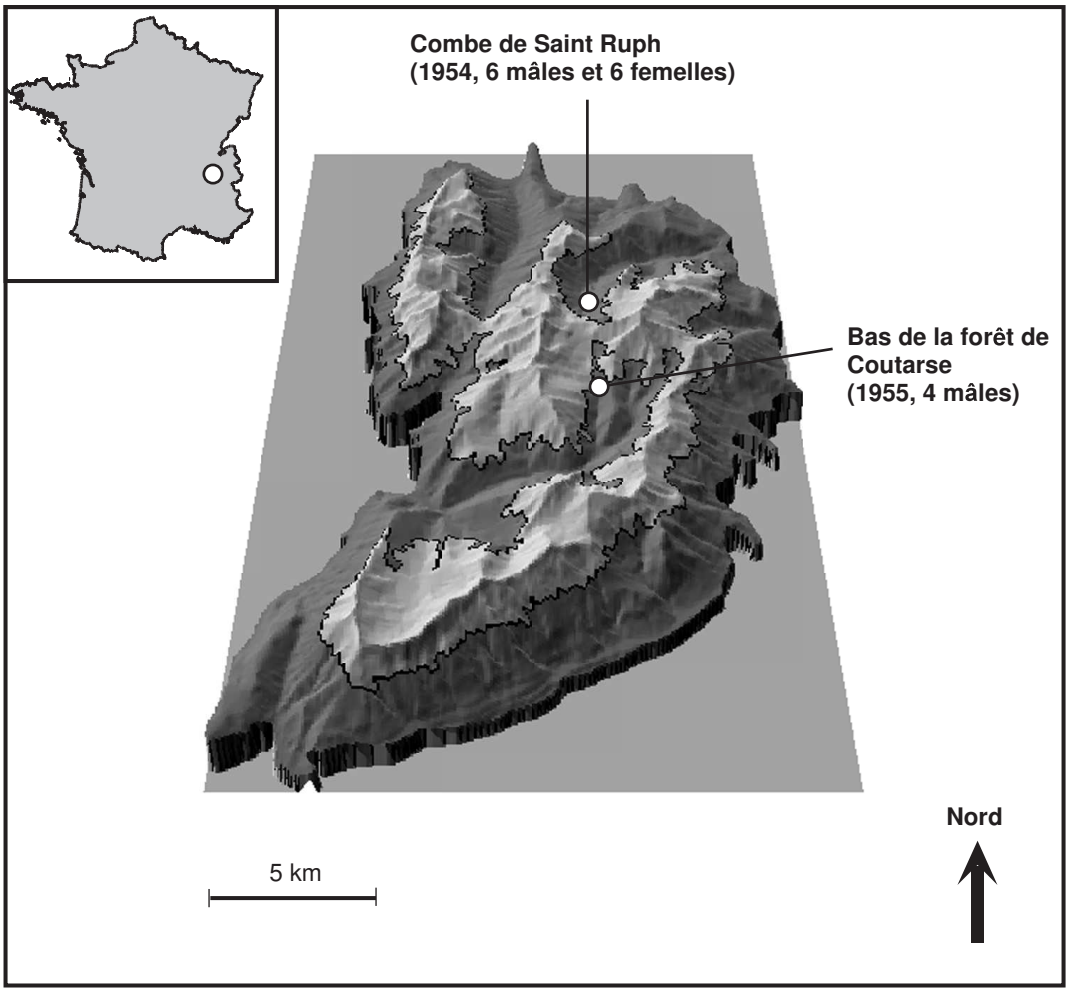
8.2 M ´

8.2.1 Etablissement des données

Les recensements de mouflons effectués chaque année depuis 24 ans ont permis de recueillir une quantité conséquente de données. L'objectif général de cette étude est de déterminer les facteurs qui affectent la distribution des mouflons dans le massif : quelles sont les caractéristiques de l'environnement recherchées par ces animaux ? Quel est leur nombre sur la zone d'étude ? Quel est leur mode d'organisation sociale et socio-spatiale ?

Etablir les données est la première étape de l'analyse. L'objectif est ici de construire la ou les bases de données qui seront analysées, donc d'éliminer ou de renseigner les données manquantes ou douteuses, de redéfinir les limites de la zone d'étude en fonction des connaissances du biologiste et/ou d'explorations préliminaires, de préciser ou de modifier la problématique en fonction des données, etc.

Cette étape de "nettoyage" (*data management*) peut être très longue, mais elle est toujours indispensable. Quelle que soit la base de données analysée, il y a toujours des informations manquantes pour lesquelles le biométricien doit prendre une décision en collaboration avec le biologiste. Nous nous limitons ici à un bref aperçu des nombreuses corrections que nous avons

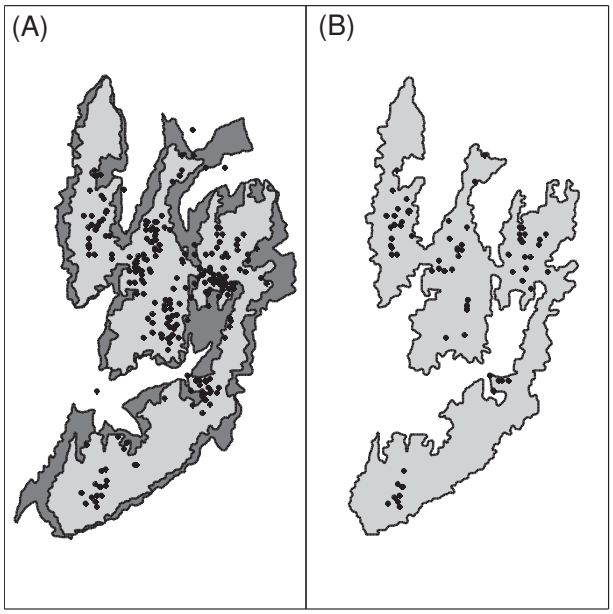


F . *28 – Localisation et représentation en perspective du massif des Bauges. Les zones foncées correspondent aux zones forestières et les zones claires aux zones ouvertes. La localisation des points de lâchers des mouflons est indiquée sur la carte.*

apportées aux données initiales.

Les recensements ont permis de localiser 7252 mouflons répartis en 285 groupes, de 1980 à 2004. Or, le protocole de relevé des données n'a été standardisé qu'à partir de 1994. Auparavant, le nombre de circuits était variable d'une année sur l'autre, et ce n'était pas toute la zone qui était échantillonnée. Par ailleurs, les chamois n'étaient qu'occasionnellement recensés lors de ces dénombrements. Or, le but *in fine* de cette étude étant de permettre l'analyse des interactions entre chamois et mouflons, il a fallu se concentrer sur les recensements effectués uniquement à partir de 1994, et supprimer de la base de données les localisations collectées avant.

Ce choix implique la perte d'un grand nombre d'occurrences de l'espèce. En effet, seulement 2611 mouflons répartis en 112 groupes ont été détectés entre 1994 et 2004. Entre la quantité et la qualité des données, nous avons choisi la qualité.



F . *29 – (A) Distribution des 285 groupes de mouflons détectés lors des recensements effectués au mois de juin, dans le massif des Bauges de 1980 à 2004, la zone gris foncé correspond à la réserve de faune sauvage des Bauges, et la zone gris clair à la zone ouverte. (B) En ne conservant que les mouflons détectés dans les zones ouvertes après 1994, il reste 107 groupes de mouflons détectés.*

Par ailleurs, sur ces 112 groupes détectés en 10 ans, 5 seulement étaient localisés en zone forestière. Or, la forêt occupe une surface importante dans ce massif (figure 29). Cette faible probabilité de détecter un groupe de mouflons dans ce type de milieu peut être le résultat d'un évitement de la forêt par le mouflon (ce qui est possible), et/ou d'une plus faible visibilité (ce qui est certain). Nous ne conservons donc que les zones ouvertes pour les analyses ultérieures (soit une surface de 6430 hectares). La base de données est donc finalement constituée de 2557 mouflons répartis en 107 groupes (figure 29).

Ces étapes de nettoyage sont les principales opérations préliminaires que nous avons effectuées, mais ne sont pas les seules. Ainsi, nous avons dû corriger de mauvaises assignations des groupes aux différents circuits parcourus, répondre à des interrogations sur des groupes dont la composition semblait "bizarre", déduire approximativement les coordonnées pour 15 groupes qui n'avaient pas été localisés. Ces petits problèmes sont autant de questions qui se posent habituellement lors des premières étapes de l'analyse des données, et permettent déjà de se familiariser avec le jeu de données. C'est grâce à la persévérance du biologiste qui ne rechigne pas à rechercher les fiches de recensement remplies par les observateurs pour trouver des solutions à ces petits problèmes que la base de données peut enfin être établie.

8.2.2 Analyse de la structure sociale

Que la base de données soit établie ne signifie pas pour autant que l'on sache *a priori* quelle méthode utiliser pour répondre à la question du biologiste. En effet, il faut tout d'abord examiner les structures présentées par les données.

En premier lieu, il apparaît que les individus sont répartis dans des groupes. L'examen des données révèle la présence de deux types de groupes, comme dans les autres zones sur lesquelles le mouflon est étudiée (L P *et al.* 1995, B *et al.* 1993) :

- les groupes constitués uniquement de mâles (18 groupes détectés en 11 ans).
- les groupes mixtes, que nous appellerons *groupes familiaux*, qui sont constitués de femelles, d'agneaux et de mâles (89 groupes détectés en 11 ans).

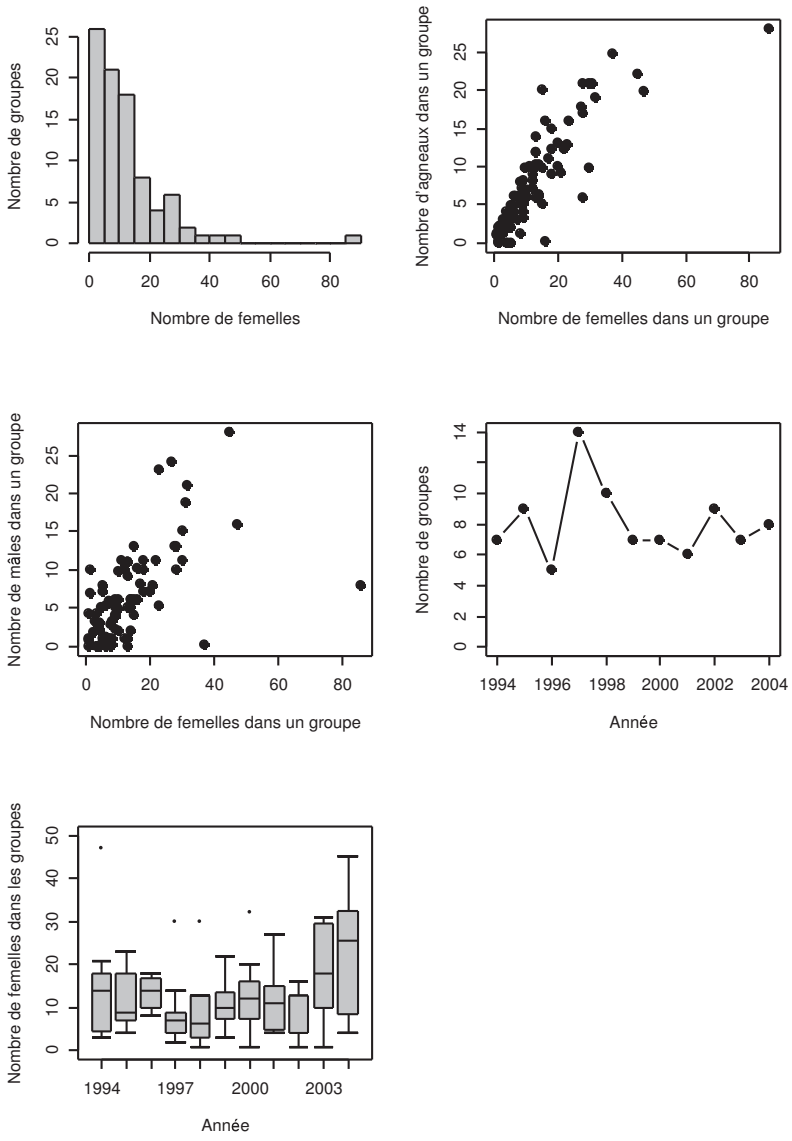
Les individus ne peuvent donc être considérés comme indépendants dans les analyses de la distribution spatiale.

Les groupes de mâles sont les plus simples à décrire. Leur taille est la seule caractéristique que l'on puisse étudier. Ces groupes sont en général constitués de 2 à 10 individus, bien que dans deux cas en 10 ans, des groupes plus importants ont été observés (18 et 20 animaux).

Les groupes familiaux sont plus complexes (figure 30). En effet, un examen des données, toutes années confondues, montre qu'il y a une corrélation importante entre le nombre de femelles et le nombre d'agneaux (r = 0.84), de même qu'entre le nombre de femelles et le nombre de mâles (r = 0.58). La discussion avec le biologiste permet d'expliquer ces relations : chaque année, les femelles mettent bas des agneaux (d'où la corrélation entre le nombre de femelles et d'agneaux). Ces agneaux grandissent, et deviennent détectables l'année suivante en tant que mâles ou femelles. En général, à l'âge d'un an, les femelles restent dans le groupe et une partie des mâles en partent. Toutefois, certains jeunes mâles restent dans le groupe pour encore quelque temps (D communication personnelle). Ainsi, le nombre de femelles dans un groupe familial est un déterminant essentiel de la composition de ce groupe (figure 1). Ce point est également relevé dans la littérature (L P *et al.* 1995, B *et al.* 1993). En moyenne, ces groupes sont constitués de 13.2 femelles (S.E. = 1.33), 7.9 agneaux (S.E. = 0.68) et 6.21 mâles (S.E. = 0.62).

T . *1 – Nombre de groupes familiaux de mouflons détectés de 1994 à 2004 en zone ouverte dans le massif des Bauges (N). Le nombre moyen de femelles par groupe (F) ainsi que l'erreur type (S.E.) sur ce nombre sont également donnés.*

Année	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
N	7	9	5	14	10	7	7	6	9	7	8
F	15.57	11.89	13.40	8.21	8.70	11.00	13.14	12.17	9.11	26.29	22.75
S.E.	5.82	2.31	1.94	1.88	2.74	2.33	3.84	3.42	2.02	10.80	5.22



F . *30 – Quelques graphiques permettant l'exploration de la structure sociale des groupes familiaux de mouflons détectés dans les Bauges de 1994 à 2004.*

Mais les données sont aussi structurées dans le temps, une structuration qu'il est indispensable de prendre en compte dans les analyses de l'utilisation de l'espace (R *et al.* 1981, S 1994, L˜ 2003). Il faut alors examiner la stabilité de la structure sociale dans le temps. En premier lieu, la stabilité du nombre de groupes détectés sur la zone est au centre de notre attention. La présence de variations temporelles du nombre de groupes pourrait avoir des implications considérables pour la suite de l'analyse. En effet, ces variations temporelles pourraient être générées par des causes qu'il faudrait prendre en compte dans la suite des analyses, telles que des biais d'observation (variations des performances des observateurs) ou des biais

de l'observé (variations de la taille de la population ou de la taille des groupes en réponse à un changement de la prédation).

En moyenne, 2 groupes de mâles (SE = 0.4) et 8.1 groupes familiaux (SE = 0.73) sont détectés chaque année. Sous l'hypothèse d'une répartition aléatoire des groupes entre les années, le nombre de groupes devrait être distribué selon une loi de Poisson. Or, une propriété bien connue de cette loi est que sa variance est égale à sa moyenne. En conséquence, le rapport variance / moyenne peut nous donner un indice de la variabilité interannuelle du nombre de groupes détectés (§ 4.2.4). Ce rapport vaut 0.88 pour les groupes de mâles et 0.73 pour les groupes familiaux. Dans les deux cas, la moyenne du nombre de groupes détectés est bien supérieure à sa variance. On observe donc une grande stabilité du nombre de groupes détectés par an (figure 30).

Puis, nous nous concentrons sur la stabilité de la composition des groupes. La taille des groupes de mâles ne semble pas varier en fonction de l'année (hypothèse testée par une analyse de la variance, F = 0.93, P = 0.53, P-value calculée grâce à 500 randomisations, i.e. par assignation aléatoire des groupes dans les années), mais il est vrai que le nombre total de groupes de mâles est très faible et que ce test est de fait très peu puissant.

Nous avons montré que le nombre de femelles dans un groupe familial est un déterminant important de la composition de ce groupe. Or, le nombre de femelles par groupe ne varie pas non plus en fonction de l'année (F = 1.26, P = 0.26), pas plus que le nombre total d'individus dans ce groupe (F = 0.99, P = 0.44). Il semble donc y avoir une très grande stabilité dans le temps de la situation démographique sur la zone.

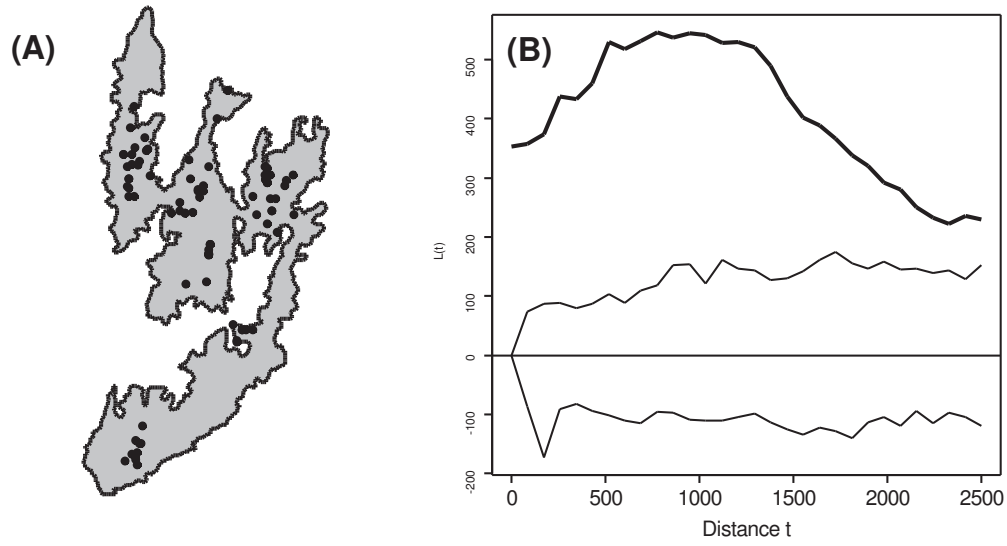
Nous ne poursuivons pas plus avant les analyses de la structure sociale. Nous avons assez d'éléments à présent pour conclure que le nombre de groupes observés lors des recensements reste le même d'une année sur l'autre, et que la composition de ces groupes est également stable. Concernant le nombre de groupes et le nombre d'individus dans ces groupes, c'est le même processus qui génère les données d'année en année. C'est à travers la dimension spatiale que nous devons à présent caractériser ce processus.

8.2.3 Modélisation spatiale

8.2.3.1 Etude des causes de l'agrégation

Etant donné le très faible nombre de groupes de mâles détectés, nous ne nous concentrons ici que sur la distribution spatiale des groupes familiaux. La carte ponctuelle de la distribution des groupes, toutes années confondues, révèle des agrégations de points, impression confirmée par l'estimation de la fonction $L(t)$ de Ripley (figure 31, cf § 4.3.2).

Le problème est donc maintenant de déterminer les causes de cette agrégation. Lorsqu'on considère les distributions des groupes mixtes *année par année*, cette agrégation semble disparaître (figure 32). Et lorsqu'on effectue un test de l'hypothèse d'une distribution aléatoire des groupes (CSR ; test de randomisation basé sur la différence maximale entre fonction G et



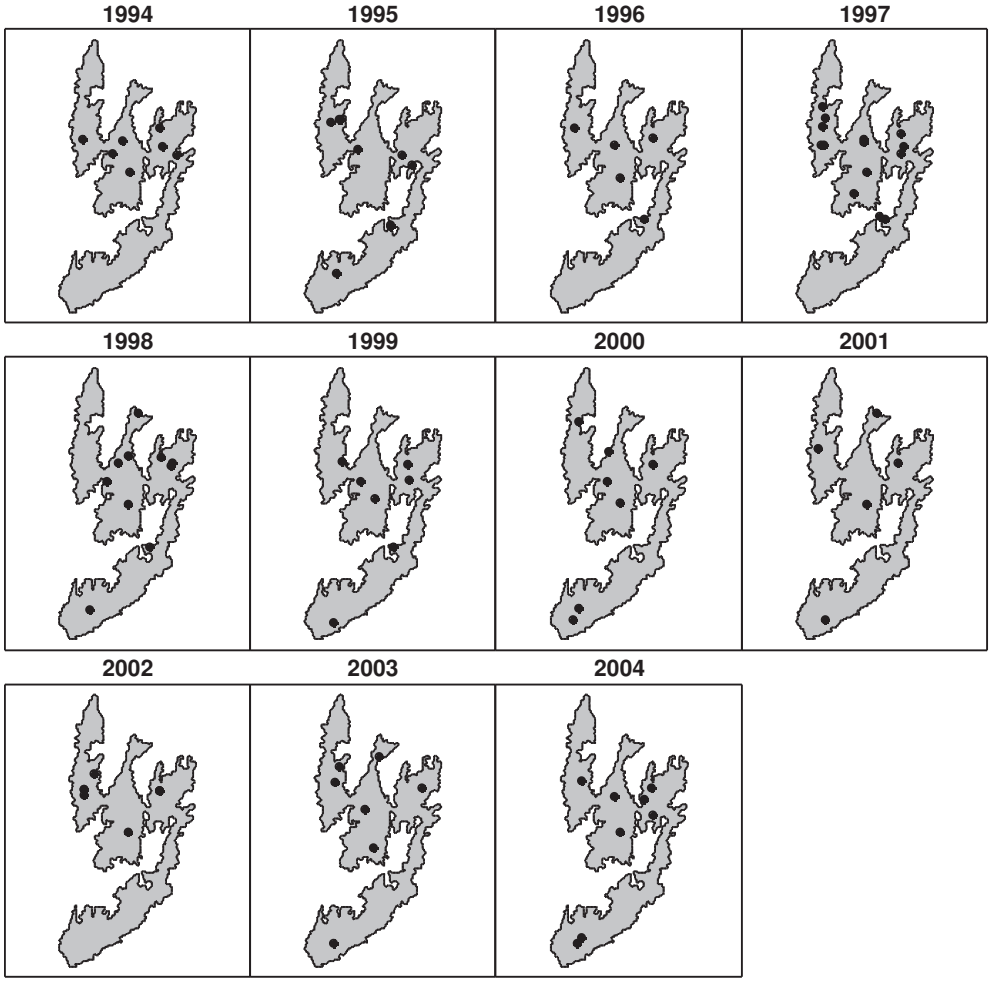
F . *31 – (A) Distribution des groupes familiaux détectés dans le massif des Bauges de 1994 à 2004. (B) Fonction L(t) de Ripley calculée sur le semis des groupes familiaux de mouflons dans le massif des Bauges. Les lignes fines correspondent aux limites inférieures et supérieures de cette fonction sous l'hypothèse de la CSR.*

fonction *F*, § 4.2.4), on ne peut mettre en évidence d'écart à la CSR pour aucune des années (cf. tableau 2). Ainsi, la CSR semble être le meilleur modèle pour la distribution des groupes une année donnée, mais lorsque toutes les années sont poolées, on note une très forte agrégation.

T . *2 – Résultats des tests de la CSR effectués année par année. La statistique utilisée est la différence maximale entre la fonction F et la fonction G. Ce critère est ensuite comparé aux valeurs obtenues après 200 réalisations de la CSR.*

	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Obs	0.49	0.25	0.43	0.28	0.40	0.21	0.23	0.44	0.24	0.32	0.24
Pvalue	0.27	0.85	0.51	0.54	0.30	0.97	0.95	0.47	0.89	0.73	0.89

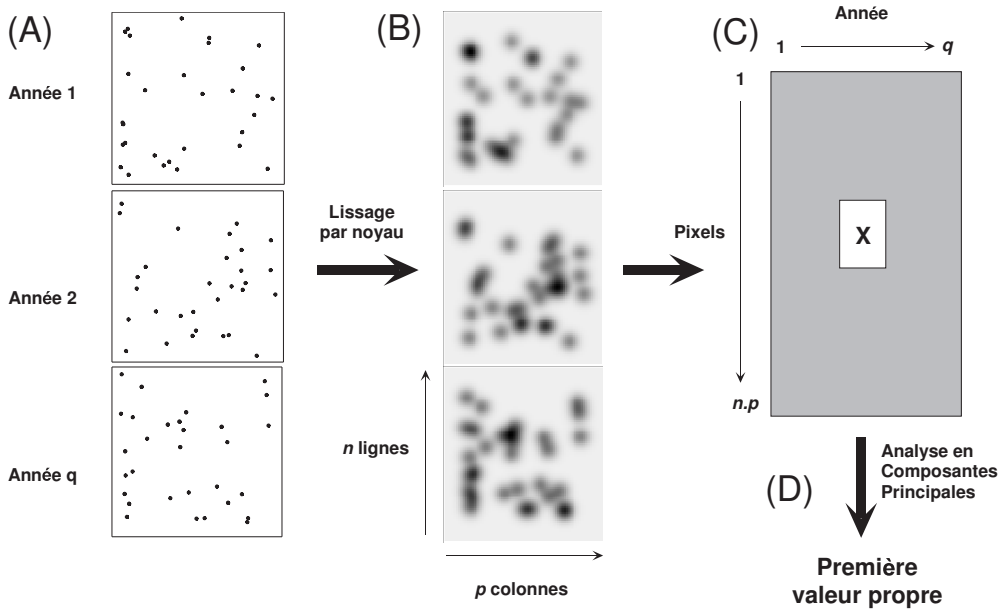
Cette agrégation pourrait être expliquée par une corrélation interannuelle de la position des groupes. Nous pouvons construire un test pour tester explicitement cette hypothèse, dont le principe est illustré figure 33. La distribution des localisations est séparée par année, et un lissage des semis de points est effectué en utilisant la méthode du noyau ; la même grille de pixels est utilisée pour toutes les années. Une très petite valeur est choisie pour ce paramètre (h = 200 m) car s'il existe une corrélation spatiale dans la position des localisations à petite échelle, cette corrélation est nécessairement retrouvée à grande échelle. On peut utiliser les estimations de la méthode du noyau pour construire un tableau **X** contenant la valeur de ces lissages pour chaque



F . *32 – Distribution des 83 groupes familiaux de mouflons détectés lors des recensements effectués au mois de juin, dans le massif des Bauges de 1994 à 2004, année par année.*

pixel (lignes) et pour chaque année (colonnes). Une analyse en composantes principales normée est ensuite effectuée sur ce tableau. Cette analyse a pour propriété de trouver un axe qui maximise la corrélation multiple avec les colonnes de **X**. Le maximum atteint est la première valeur propre de l'analyse. Dans la mesure où la CSR est un bon modèle pour la distribution des localisations, on peut se servir de ce modèle pour simuler l'indépendance entre les années, estimer à nouveau un lissage par la méthode du noyau sur les simulations, et recalculer une première valeur propre sous cette hypothèse. Le critère observé est ensuite comparé à la distribution des valeurs propres obtenues sous cette hypothèse. On peut à partir de cette première valeur propre calculer le pourcentage de la variabilité inter-annuelle expliquée par le premier axe de l'analyse, ce qui constitue une mesure de la corrélation entre les années.

Appliqué sur les données "mouflons", ce test révèle une très forte corrélation inter-annuelle dans la position des points (26% de variabilité expliquée sur le premier axe de l'analyse, $P < 0.001$, test basé sur 1000 répétitions). Notons qu'un paramètre de lissage égal à 500 m



F . *33 – Principe du test de la corrélation inter-annuelle de la position des localisations.*

permet d'arriver à 41% de la variabilité expliquée sur le premier axe. Un lissage par la méthode du noyau nous permet enfin d'identifier les zones dans lesquelles les groupes de mouflons sont observés chaque année. Un certain nombre de grandes zones peuvent être identifiées, au sein desquelles les groupes de mouflons sont détectés chaque année (figure 34).

L'interprétation biologique est délicate, étant donné le peu de connaissances disponibles sur cette espèce dans ce massif. Mais encore une fois, les connaissances de la biologie de l'espèce viennent à notre secours. Le mouflon est un animal grégaire (L P *et al.* 1995, B *et al.* 1993). On ne sait rien de la stabilité dans le temps des groupes de mouflons. Les groupes sont ils toujours constitués par les mêmes animaux ? Peuvent-ils se scinder en sous-groupes ? Un individu peut-il changer de groupe au cours du temps ? Dans le massif du Caroux, des observations de terrain semblent indiquer que le mouflon a un comportement très flexible, c'est-à-dire que les groupes s'assemblent et de désagrègent très facilement (C 1993, G , communication personnelle), mais aucune étude n'a pu être menée sur le sujet. En revanche on sait que cet animal est sédentaire. Des suivis d'individus par colliers GPS sur cette zone montrent que le mouflon possède un domaine vital de quelques centaines d'hectares (D , communication personnelle), ce qui correspond à la taille des zones identifiées par l'analyse. La littérature reporte également des tailles de domaine vital similaires pour cette espèce (D *et al.* 1992, 1993). Ainsi, un individu localisé dans une zone du massif sera localisé dans la même zone l'année suivante s'il est toujours en vie et s'il est détecté par les observateurs. Le biologiste émet alors l'hypothèse suivante : les mêmes groupes seraient détectés dans les mêmes zones d'année en année.

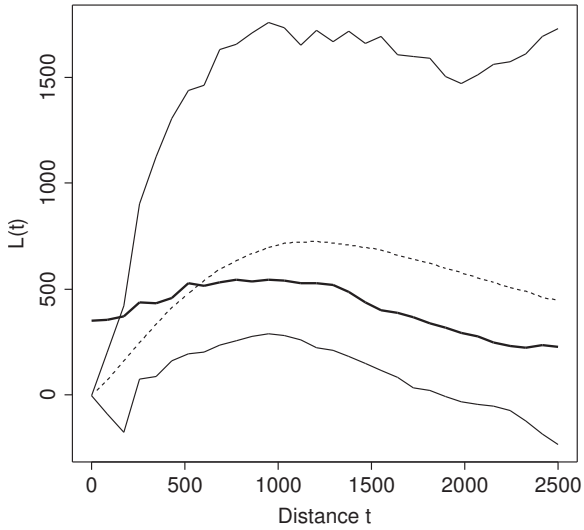
F . *34 – Le lissage par la méthode du noyau de la distribution des localisations des groupes de mouflons détectés de 1994 à 2004, avec un paramètre de lissage égal à 200 m. Sept grandes zones sont identifiées (voir le texte).*

8.2.3.2 Modélisation du processus

Le modèle proposé pour la distribution des groupes familiaux de mouflons pourrait être décrit par un processus de Neyman-Scott (§ 4.3.3.3). En effet, chaque “parent” correspondrait à un groupe, et chaque “descendant”, à une observation du groupe une année donnée. Il serait en théorie possible d’ajuster un tel processus à la distribution de localisations observée toutes les années confondues, en utilisant l’équation 4.6. Ceci nous permettrait alors de déterminer exactement combien de groupes sont présents sur la zone, et avec quelle fréquence ils sont observés (nombre de descendants).

Or, bien que techniquement possible (grâce à la fonction `pcp()` de la bibliothèque **splancs** du logiciel R), l’ajustement d’un tel processus est difficile ici. En effet, comme nous l’avons montré dans la section 4.3.3.3, cet ajustement est une opération compliquée qui implique un certain nombre de choix *ad hoc* de paramètres. Le choix de ces paramètres a été assez peu étudié, mais quelques recommandations existent lorsque la zone d’étude a une forme carrée ou rectangulaire. En revanche, cette opération devient divinatoire lorsque la forme de la zone est aussi complexe que celle du massif des Bauges, la variabilité de la fonction $K(t)$ de Ripley étant beaucoup trop grande.

Mais avec les informations déjà obtenues lors de l’exploration des données, un processus de Neyman-Scott peut être ajusté, avec comme paramètres 7 parents (le nombre de zones mises en évidence), 11 descendants par parent (le nombre d’années de recensements) et une distribution normale bivariée des descendants autour des parents d’écart-type 500 m (i.e. un domaine vital



F . *35 – La fonction $L(t)$ de Ripley estimée à partir des localisations de mouflons récoltées dans les Bauges, de 1994 à 2004. Les enveloppes de confiances sont obtenues après 100 réalisations d'un processus de Neyman-Scott dont les paramètres sont 7 parents, 11 descendants par parent et une distribution normale bivariée des descendants autour des parents d'écart-type 300 m. la ligne pointillée indique l'espérance d'un tel processus.*

à 95% de $\pi \times (500 \times 1.96)^2 \sim 300$ ha). Un tel processus conduit à une enveloppe de confiance de la fonction $L(t)$ de Ripley qui ne diffère pas significativement de ce qui est observé (figure 35). C'est seulement aux petites échelles (< 150 m) que l'on observe une agrégation plus forte qu'attendue sous l'hypothèse de ce modèle. Deux explications peuvent être données à ce mauvais ajustement :

– les observateurs, lorsqu'ils localisent les groupes de mouflons, sont munis d'une carte quadrillée par une grille dont les quadrats font 50 mètres de côté. Lorsqu'un groupe est détecté, il est placé dans un de ces quadrats. Il y a donc ainsi, à petite échelle, une certaine discrétisation des données qui conduit à une "agrégation" artificielle des groupes ;

– lors de l'étape d'établissement des données (§ 8.2.1), nous avons du déduire approximativement les coordonnées pour 15 groupes qui n'avaient pas été localisés par les observateurs. Ces groupes ont été placés à la moyenne des coordonnées des autres groupes localisés sur les mêmes circuits. Cette opération permet de ne pas perdre ces groupes pour l'analyse, mais se traduit nécessairement par un "empilement" de points aux mêmes coordonnées, donc à nouveau par une agrégation.

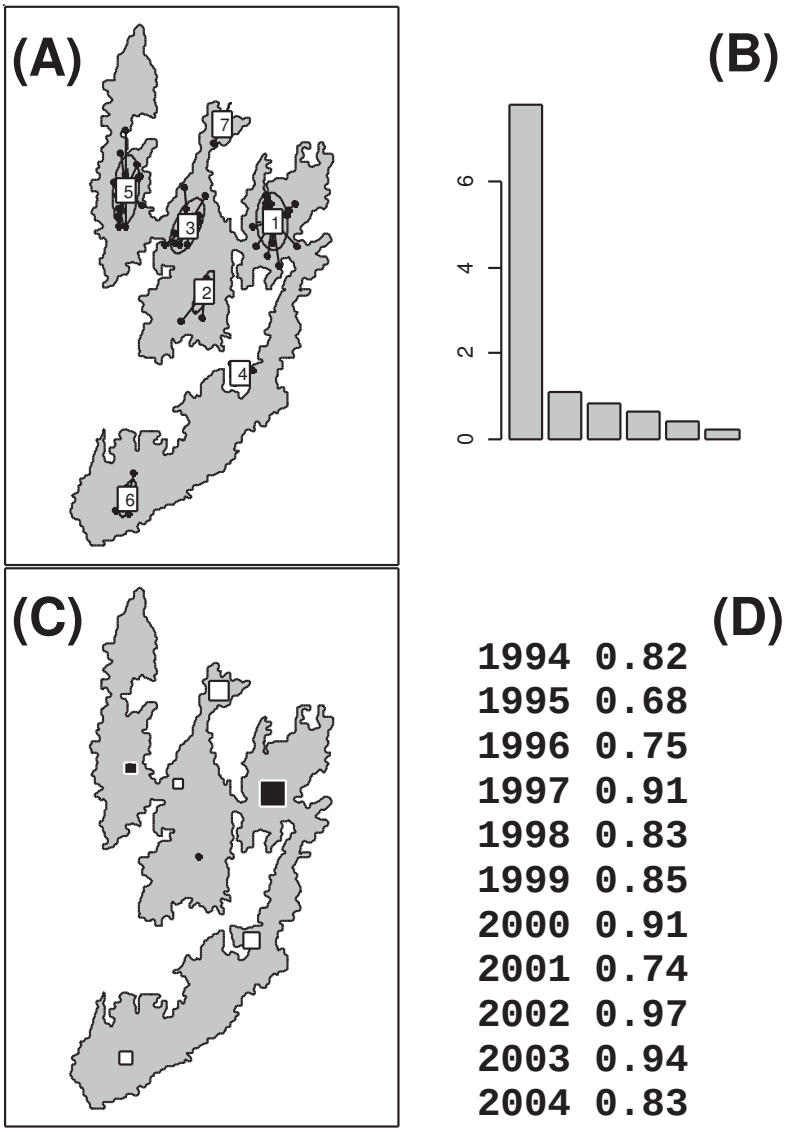
Ce modèle reste malgré tout une hypothèse très difficile à vérifier sans étude confirmatoire. En effet, on ne dispose d'aucun moyen d'identifier les animaux dans les groupes. Comme nous l'avons signalé, on ne dispose même pas d'informations sur la dynamique de ces groupes, et

il n'y a aucun moyen de savoir si un groupe observé une année donnée a encore une existence l'année suivante. On pourrait essayer d'utiliser la composition des groupes afin de déterminer si ce sont les mêmes qui sont détectés d'une année sur l'autre. Cependant, d'une année sur l'autre, la composition d'un groupe change (les agneaux deviennent mâles ou femelles, certains individus sont tués à la chasse, le taux de natalité peut changer, etc.). Il est alors très difficile de savoir si un groupe de 40 femelles une année est le même que le groupe de 30 femelles détecté dans la même zone l'année suivante.

Malgré cette limite, nous pouvons étudier l'effectif *en individus* au sein de chacune des 7 grandes zones que nous avons mises en évidence précédemment. Les femelles étant le cœur des groupes familiaux, nous ne nous concentrerons ici que sur cette catégorie d'animaux. Nous étudions alors, pour chaque année et au sein de chaque zone, le total du nombre de femelles détectées (en sommant ces nombres pour tous les groupes de chaque zone, lorsque plusieurs groupes sont détectés). Pour ce, nous définissons 7 zones, d'après la figure 34, et nous assignons chacun des 83 groupes de mouflons détectés à la zone la plus proche (figure 36). Puis nous construisons un tableau **Y** contenant le nombre total de femelles détectées dans chaque zone (lignes) pour chaque année (colonnes). Une ACP normée de ce tableau nous permet de savoir si le nombre d'individus dans chacune de ces zones est stable d'année en année, c'est-à-dire si c'est toujours dans les mêmes zones que l'on observe la plus forte densité en individus, quelle que soit l'année. Les résultats de cette ACP sont présentés figure 36.

L'ACP ne met en évidence qu'un seul axe, qui est très fortement corrélé à toutes les années (aucun coefficient de corrélation n'est inférieur à 0.68). Cet axe est un axe de taille : les fortes valeurs sur cet axe correspondent à de fortes densité en individus, et inversement pour les faibles valeurs. Dans la mesure où nous avions montré précédemment que le nombre de femelles dans les groupes restait stable d'une année sur l'autre, ce résultat permet de compléter le modèle, en affirmant que la densité en femelles dans chacune de ces zones est relativement stable.

L'hypothèse proposée par le biologiste semble donc se confirmer. La première année, un petit nombre de groupe familiaux est présent sur la zone, distribués de façon aléatoire. Ces emplacements sont en revanche stables dans le temps et les mêmes groupes sont observés d'une année sur l'autre dans une zone donnée. Or, la population de mouflons provient d'une introduction. La distribution actuelle des mouflons sur la zone est le résultat de la colonisation du massif. Afin d'émettre des hypothèses sur cette dynamique de colonisation, Gaëlle D a fait des recherches dans les journaux locaux et dans les rapports d'activité scientifiques écrits depuis 1954. D'après ses recherches, elle arrive au modèle de colonisation suivant : 16 individus ont été lâchés en deux points en 1954 et 1955. Ces individus se sont immédiatement établis dans la zone la plus proche – la zone 1 (sur laquelle la densité en individus est la plus forte). En 1965, des groupes de mouflons sont observés dans les zones 1 et 2. La colonisation des autres zones est incertaine. En 1967, des mouflons sont observés dans les régions 3 et 4. C'est en 1971 que les premiers mouflons sont observés dans la région 6, bien qu'il semblent s'y être installés beaucoup plus tard, dans les années 1980, en même temps que la région 5. Nous ne disposons d'aucune information concernant la région 7.



F . *36 – (A) les 7 zones définies pour l'étude de l'effectif des femelles ; (B) valeurs propres de l'ACP du tableau contenant le nombre de femelles dans chaque zone (ligne) et pour chaque année (colonne) ; (C) cartographie des scores des zones sur le premier axe de l'ACP dans le massif des Bauges ; (D) corrélations entre le nombre de femelles observées dans les zones et les scores des zones sur le premier axe de l'ACP, année par année.*

8.2.4 Analyse de la sélection de l'habitat

8.2.4.1 Mise en forme et exploration préliminaire

Nous avons ici une illustration très claire du fait que la densité en individus sur une zone ne reflète pas nécessairement la qualité de l'habitat (§ 2.2.2). La zone 1 est la plus dense, peut être parce qu'elle est la plus anciennement colonisée. La zone 6 est colonisée depuis une époque plus récente, ce qui explique peut-être sa faible densité. La population du mouflon dans le massif des

Bauges a une histoire qui commence en 1954, et comme de nombreux auteurs l'ont signalés (pour une revue, cf. M *et al.* 1992), on ne peut ignorer l'histoire dans les études de sélection de l'habitat.

Mais que la distribution des groupes dans l'espace géographique ne diffère pas d'une distribution aléatoire ne signifie pas qu'il en est de même pour la distribution des groupes dans l'espace écologique. En effet, les caractéristiques de l'habitat recherché par les mouflons lors du processus de colonisation sont peut-être elles-mêmes distribuées de façon aléatoire sur le massif. Il nous faut donc comparer la composition des zones colonisées par le mouflon à celle du reste de la zone d'étude.

Gaëlle D a pu réunir un certain nombre de cartes de variables susceptibles d'avoir influencé ces choix. Les variables environnementales en question sont :

- l'altitude
- la pente
- l'ensoleillement
- l'hydrographie (carte des crêtes et des fonds de vallons)
- la distribution des sentiers touristiques sur la zone
- le type de végétation (11 types définis)

La carte des chemins a tout d'abord été convertie en carte des distances aux chemins. De façon similaire, la carte des types de végétation a été transformée en 11 cartes de distances aux types de milieux. En effet, un individu peut utiliser une zone à cause de la proximité de certaines structures de l'habitat. Par exemple, un mouflon préfère peut-être rester à proximité des zones de forêts (qui lui assurent la protection) tout en utilisant les zones ouvertes (qui lui assurent la nourriture).

En outre, les variables de distances, x_i, ont été transformées par l'équation $-\exp(-0.00003x_i^2)$), afin de prendre en compte le fait que pour le mouflon, être plus proche de 100 mètres d'un type de milieu donné n'a pas la même signification selon qu'il se trouve à 50 mètres ou à 2000 mètres de ce type de milieu. La pente a été ensuite recodée en trois classes : <20%, entre 20 et 40% et >40% (classes que nous avons appelées respectivement "pente faible", "pente moyenne" et "pente forte"). En effet il est probable que le mouflon recherche un optimum de pente (e.g. même s'il apprécie les pentes fortes, les falaises à pic ne peuvent pas être utilisées). Cet optimum sera mieux mis en évidence avec une variable en classes qu'avec une variable continue. L'ensemble des cartes étudiées est présenté figure 37.

Une ACP normée peut dans un premier temps nous aider à déterminer les structures des variables d'habitat sur la zone d'étude. Un seul axe est mis en évidence par l'analyse, qui explique 18.5% de la variabilité présente sur la zone. La structure mise en évidence par cet axe est induite par l'altitude (figure 38). Les types de végétations sont en effet différents à forte altitude (présence d'éboulis, de prairies à laîches sempervirente et à sesleries, pente relativement forte) et à faible altitude (proche du milieu fermé, des prairies à brachipodiums et des fourrages, pente relativement faible). Cet étagement altitudinal de la végétation est bien connu en montagne.

8.2.4.2 La sélection de l'habitat par le mouflon

Considéré *a priori*, le jeu de données correspond à un protocole de type I. Nous avons en effet un seul type de localisations, les groupes familiaux de mouflons. Mais il faut être lucide. Effectuer brutalement une analyse conçue pour ce type de protocoles serait une mauvaise stratégie, car elle ignorerait totalement les structures mises en évidence par l'étape de modélisation spatiale.

Il faut donc procéder différemment. Nous avons mis en évidence 7 zones sur le massif dans lesquelles la probabilité de détecter un mouflon lors des recensements est plus importante que dans le reste du massif. On peut, grâce aux outils de la bibliothèque **adehabitat**, isoler les localisations de ces 7 zones (figure 36A). Nous avons émis l'hypothèse que ce sont les mêmes individus qui sont localisés dans ces zones d'une année sur l'autre. En d'autre termes, du point de vue de l'analyse, ces zones peuvent être considérées comme des domaines vitaux de groupes de mouflons. Nous pourrions effectuer une analyse OMI afin de déterminer ce qu'il peut y avoir de commun dans leur composition.

Il faut tout d'abord prendre en compte les structures spatiales mises en évidence précédemment, afin de déterminer si l'analyse OMI permet de faire ressortir une quantité significative de la marginalité présente dans le jeu de données sur un petit nombre d'axes. En d'autres termes, peut-on considérer qu'il y a une similarité de la composition de l'habitat dans les différentes zones ?

En fait, il est possible de tester si la quantité de marginalité expliquée par le premier axe de l'analyse OMI est plus importante qu'attendue sous l'hypothèse d'une utilisation aléatoire de l'habitat. Pour ce il faut construire un test adapté à la question posée. Nous avons précédemment montré que, d'un point de vue strictement spatial, la distribution des points ne diffère pas significativement de celle qu'aurait généré un processus de Neyman-Scott. Nous avons identifié sept zones. Nous pouvons alors construire un test de randomisation de la façon suivante. Pour chaque zone i ($i = 1, ..., 7$), le barycentre des n_i localisations dans l'espace géographique peut être calculé. Par ailleurs, on peut calculer les variances des coordonnées des localisations de groupes par rapport à ce barycentre. A chaque étape du processus de randomisation et pour chaque zone, un nouveau barycentre est placé aléatoirement sur la zone d'étude, et un nouveau jeu de localisations est généré en tirant au sort n localisations d'une loi normale bivariée centrée sur ce point. La matrice de variance-covariance de cette distribution est supposée diagonale (i.e. la corrélation entre les coordonnées X et Y est supposée nulle ; hypothèse raisonnable si l'on considère la figure 36A). Cette matrice est construite à partir des variances calculées précédemment. Une analyse OMI est alors effectuée sur ce nouveau jeu de localisations. Le première valeur propre de l'analyse constitue une mesure de la quantité de marginalité expliquée sur le premier axe de l'analyse, et fournit en conséquence un excellent critère pour mesurer la similarité de l'écart entre la composition moyenne sur la zone d'étude et la composition de chaque zone.

Ce test n'est pas significatif (42% de marginalité expliquée sur le premier axe de l'analyse, P = 0.59). Or, nous devons faire attention. En effet, chaque groupe pourrait avoir eu sa propre

stratégie de sélection du domaine vital lors de la colonisation du massif, stratégie différente d'un groupe à l'autre. Dans ce cas, il y aurait une sélection de l'habitat significative pour chacun des groupes, mais cela expliquerait que rien ne soit apparu au niveau de l'analyse OMI. On peut tester cette hypothèse grâce à une extension du test précédent. L'algorithme de randomisation reste le même, mais un test est effectué par zone, et le critère est la marginalité associée à chacune des zones. Ce test permet de comparer la marginalité observée pour chacune des zones à celle attendue sous l'hypothèse d'une répartition aléatoire des zones *dans l'espace écologique*.

T *. 3 – Test de l'hypothèse d'une marginalité nulle pour chacune des 7 zones occupées par le mouflon dans le massif des Bauges.*

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5	Zone 6	Zone 7
Marginalité	7.01	5.86	3.61	15.05	2.82	6.41	6.29
P-value	0.13	0.49	0.66	0.13	0.77	0.57	0.80

La marginalité n'est significative pour aucune des zones (tableau 3). En d'autres termes, tout se passe comme si les groupes s'étaient réellement distribués de façon totalement aléatoire à la fois dans l'espace géographique et dans l'espace écologique. Bien que de tels résultats soient décevant pour le biologiste, ils n'en demeurent pas moins des résultats. Aucune des variables étudiées ne permet d'expliquer la distribution sur la zone. Plusieurs hypothèses peuvent être formulées pour les expliquer. Il est possible que la sélection des zones par le mouflon se fasse sur d'autres variables environnementales que celles sur lesquelles nous avons travaillé. Ainsi, la présence d'autres espèces compétitrices, comme le chamois, a peut-être une influence. Il est aussi possible qu'à cette échelle l'environnement ne soit pas déterminant de la position des groupes. Le mouflon est un animal qui s'adapte à une grande variabilité d'environnements, et qui adopte un comportement d'utilisation de l'espace particulier dans chaque contexte.

8.2.5 Discussion

Ainsi, le seul modèle que l'on puisse proposer au biologiste est celui que nous avons décrit à la fin du § 8.2.3. La distribution actuelle des groupes de mouflons sur la zone est le résultat d'une colonisation de proche en proche du massif, qui s'est vraisemblablement effectuée très rapidement, en une décennie. En France la majorité des populations de mouflons sont des populations issues d'un lâcher. Ce type de processus de colonisation avait déjà été décrit dans d'autres zones (notamment le Caroux, P 1969, G et C *en préparation*). Mais aucune des variables d'habitat étudiées n'a permis de déterminer ce qui a influencé l'établissement des individus colonisateurs. Ce modèle est incontestablement utile, car il formalise le savoir du biologiste. Cependant, nous devons garder en tête que nous ne nous sommes concentrés que sur une très courte période de l'année (mois de juin), et que rien ne permet de généraliser ces résultats au reste de l'année.

Si ce modèle apporte quelques réponses, il permet surtout de poser de nouvelles questions, voire de préciser les anciennes. En premier lieu, comment caractériser un groupe de mouflons ? Dans ces analyses, nous avons considéré que les groupes étaient constitués des mêmes individus d'une année sur l'autre, mais les informations manquent à ce sujet (C 1993). Les seules raisons qui nous ont poussé à faire cette supposition est que le mouflon est un animal sédentaire et que les femelles, le cœur des groupes familiaux, utilisent en général le domaine vital que leur mère (D , com. pers.). Mais il y a des zones sur lesquelles on peut trouver plusieurs groupes une même année (e.g. la zone 1 en 1994, figure 32). Le biologiste souligne alors la difficulté sur le terrain de définir ce qu'est un groupe. Là où certains observateurs ne verrons qu'un grand groupe, d'autres en noterons deux petits. Nous avons pu définir des zones utilisées par le mouflon grâce à la méthode du noyau. La taille de ces zones correspond à la taille du domaine vital généralement mise en évidence chez cette espèce. Mais peut-être que plusieurs groupes utilisent le même domaine vital (i.e. absence de territorialité), peut-être que les mouflons peuvent passer d'un groupe à l'autre. La question de la définition du groupe reste ouverte à la fin de cette étude, malgré son caractère central pour la compréhension de la biologie du mouflon sur cette zone.

Un autre grand problème soulevé par l'analyse est la mise en évidence des causes du processus de colonisation. Ce type de processus est fréquemment reporté dans la littérature (P 1967, G et C *en préparation*). Pourquoi n'observe-t-on pas sur le massif un gradient du nombre de groupes depuis le point de lâcher ? comment s'effectue la colonisation ? Quels sont les facteurs (évitement du parasitisme ou de la consanguinité, etc.) qui poussent un ou plusieurs individus à quitter un groupe pour s'établir dans une autre zone ? Et combien y-a-t-il d'individus fondateurs pour un nouveau groupe ? Quels sont les facteurs qui influencent l'établissement de ces individus disperseurs dans une zone (variables d'habitat non étudiées ici, distance au groupe de départ) ? Pourquoi la colonisation du massif a-t-elle été aussi rapide (une décennie) ? Nous ne poursuivrons pas plus avant l'analyse de ces questions ici.

8.3 C

Dans ce chapitre, nous avons illustré la démarche de modélisation de la distribution spatiale d'une espèce sur une zone, en utilisant des données "typiques" dans ce type d'étude. A travers les analyses effectuées dans ce chapitre, on comprend alors l'écart qui peut exister entre l'application automatique d'une méthode "standard" (e.g. l'utilisation "à la chaîne" des fonctions de sélection des ressources) et la démarche de *construction d'une analyse*, c'est-à-dire la démarche de modélisation de données écologiques. Le problème posé par les données analysées ici est particulier car la problématique, l'espèce, son histoire et les données sont elles-mêmes particulières. Ces nombreuses contraintes obligent le biométricien à "se laisser faire par les données", c'est-à-dire à laisser l'analyse se construire progressivement autour des données.

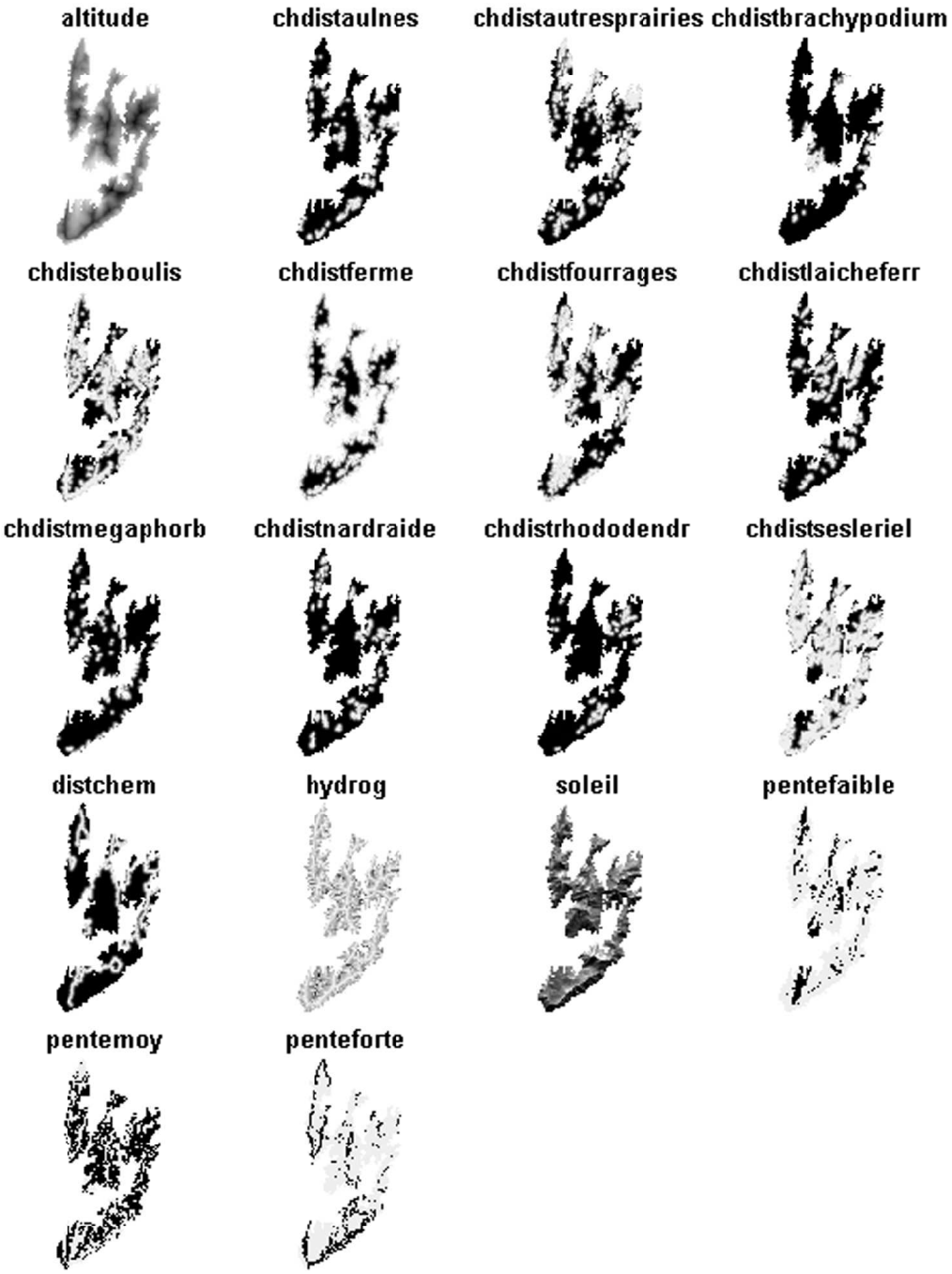
La question des outils à employer est donc secondaire. Nous avons vu qu'un protocole de type I n'implique pas nécessairement l'utilisation de méthodes développées initialement pour ce type de protocole. Notons que les outils que nous avons utilisés ne sont pas très compliqués. Il a

parfois fallu “inventer” une nouvelle méthode pour tester une hypothèse spécifique aux données (e.g. test de la corrélation interannuelle de la position des groupes, test de la significativité de la première valeur propre de l’analyse OMI). Ceci est une bonne illustration du point de vue de T *et al.* (1993), pour qui la Biométrie est avant tout un *métier* qui implique l’utilisation de concepts et de symboles mathématiques, outils nécessitant un emploi intensif d’ordinateurs.

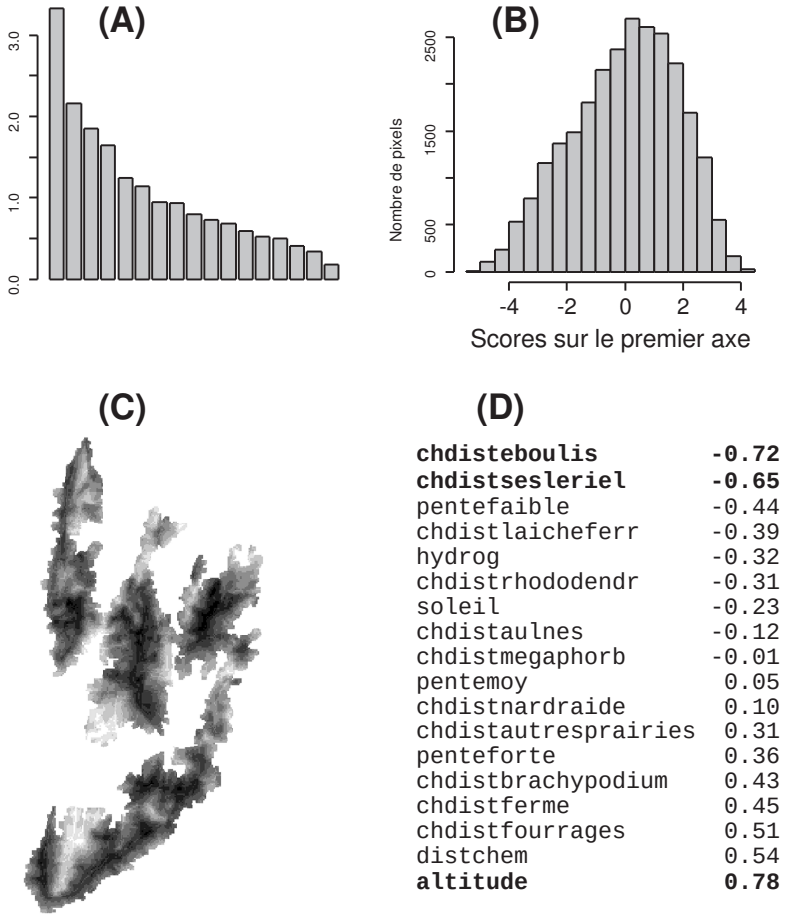
Un problème dont nous n’avons pas discuté est la question de l’inférence. Dans le cas présent, on étudie une seule population de mouflons dans une situation précise, avec son histoire, son habitat, etc. Nous avons ici une parfaite illustration du point de vue de L (1986), lorsqu’il affirme que la complexité est une donnée de l’Ecologie, et qu’elle n’est pas a démontrer ici. L’inférence consiste à déterminer comment les résultats obtenus sur cette zone peuvent être extrapolés à d’autres zones.

La variabilité entre les zones du comportement du mouflon n’est plus à démontrer. En fonction de la situation dans laquelle il se trouve, il va réagir différemment. Ainsi, le mouflon des Bauges n’a probablement pas le même comportement que celui du Caroux (G , communication personnelle). Bien sûr, *nous n’affirmons pas ici qu’il est impossible de trouver des patrons communs dans l’utilisation de l’espace par les animaux entre différentes zones*. Ce serait affirmer l’absence de théorie en Ecologie. Les individus d’une espèce donnée – par exemple le mouflon – ont des besoins identiques qu’ils doivent combler dans des milieux différents (même anatomie, donc régime alimentaire similaire, etc.). Dans un milieu donné, ces individus s’adapteront, mais ils ne s’adapteront pas n’importe comment, car il existe des contraintes spécifiques à l’espèce. La mise en évidence de ces contraintes ne pourra se faire que par la comparaison des résultats obtenus sur différentes zones. Il est en effet impossible de dissocier la part du fonctionnement d’une population liée à l’adaptation à un milieu particulier et celle qui est intrinsèque à l’espèce en n’étudiant qu’une seule zone. En déterminant ce qu’il peut y avoir de commun entre différentes zones où une espèce est présente, le biologiste peut éliminer les contraintes particulières liées aux milieux pour en déduire le dénominateur commun, c’est-à-dire la part du fonctionnement des populations que l’on retrouve de façon systématique dans toutes les populations. C’est sur cette base commune que pourront être définies des politiques de gestion de la faune.

Nous défendons ici l’idée que le biométricien qui doit analyser les données d’une zone ne peut faire de supposition *a priori* sur la part d’adaptation liée à l’espèce et celle qui est liée au milieu, à moins qu’il ne dispose des données relevées dans différentes zones. Dans le cas où il n’étudie qu’une population, il sera dans l’obligation d’accepter la complexité de l’objet qu’il étudie, et en décrire le fonctionnement. C’est le biologiste qui pourra après l’analyse, par comparaison avec d’autres études, en tirer des conclusions sur les caractéristiques de l’espèce.



F . *37 – Les 18 cartes de variables d'habitat utilisées pour l'étude de la sélection de l'habitat. Une variable dont le nom commence par "chdist" correspond à une variable mesurant la distance à un milieu.*



F . *38 – Résultats de l'ACP effectués sur les cartes des variables environnementales du massif des Bauges. (A) Diagramme des valeurs propres ; (B) histogramme des scores des pixels sur le premier axe ; (C) cartographie des scores des pixels sur le premier axe dans l'espace géographique ; (D) score des variables sur le premier axe de l'ACP (i.e. coefficients de corrélation).*

Chapitre 9

Discussion

Dans ce mémoire, un certain nombre de méthodes sont présentées, permettant l'analyse de la distribution d'un ou plusieurs semis de points dans les espaces écologique et géographique. Seuls les outils dont je me suis servi au cours de ces trois années de thèse y sont développés. Ils ont été organisés de façon à illustrer une démarche possible pour l'étude de l'utilisation de l'espace par la faune sauvage. Cette thèse a permis un certain nombre d'innovations méthodologiques, telles que l'analyse K-select (§ 7.2.1.2), l'analyse factorielle des rapports de sélection (§ 7.2.2) ou l'analyse discriminante sur vecteurs propres du graphes de voisinage (§ 5.4.2). Certaines améliorations mineures ont été apportées à des méthodes déjà existantes, telle que la généralisation de l'ENFA présentée § 6.3.2. La bibliothèque de fonctions **adehabitat** pour le logiciel R a été développée pour permettre l'application de la démarche proposée ici. Enfin, cette démarche a pu être appliquée dans le cadre de nombreuses collaborations, concrétisées par la rédaction de plusieurs articles (voir en Annexe).

Dans ce chapitre, nous discutons des principaux aspects théoriques liés à ce travail. L'importance de l'approche systémique lors de l'analyse d'un système complexe est tout d'abord mise en exergue. L'apport essentiel de l'évolution récente de l'informatique, et en particulier l'apport de la bibliothèque **adehabitat** est également discutée. Nous décrivons enfin le rôle essentiel de la consultation dans le travail du biométricien, de même que l'importance des biométriciens pour l'étude menée afin d'améliorer la gestion de la faune sauvage.

9.1 I ’ ´

Nous avons montré que la méthode qui permettrait, *de façon automatique*, de différencier les aspects environnementaux des contraintes spatiales dans les études de sélection de l'habitat est le Saint Graal des biologistes. Recherchée par beaucoup, elle tient plus de la légende que de la réalité. Cette constatation relève de la simple logique : une cause ne peut avoir qu'un seul effet, mais un effet peut avoir plusieurs causes. Un semis de points, récolté dans le cadre d'une étude observationnelle, est étudié dans l'espace géographique. Ce semis de points est l'effet. De nombreuses causes, c'est-à-dire des contraintes d'ordre spatial ou écologique, ont généré ce semis. il est impossible de dissocier les causes par la simple observation de l'effet. Aucune mé-

thode mathématique ne permettra jamais de différencier les contraintes spatiales et écologiques, à moins d'avoir précisément orienté le protocole de collecte des données *a priori* de façon à maîtriser l'un de ces deux aspects. Dans ce dernier cas, nous ne sommes plus dans le cadre d'une étude observationnelle, mais dans le cadre d'une étude expérimentale.

Or, les études écologiques sont par essence observationnelles. En effet, l'objet d'étude est la population biologique, c'est-à-dire un système complexe (L et S 2004). L'un des buts de l'Ecologie est précisément la mise en évidence de cette complexité. Il est donc impossible de prédire *a priori*, c'est-à-dire avant même la collecte des données, tous les facteurs qui peuvent avoir une influence sur le système, puisque c'est précisément l'objectif de l'analyse que de les mettre en évidence. Chaque système est particulier et les contraintes auxquelles il est soumis sont différentes des autres systèmes. Le caractère observationnel des études écologiques implique que d'autres arguments seront nécessaires pour permettre cette dissociation entre contraintes spatiales et écologiques, à savoir des arguments biologiques.

L'un des relecteurs chargés d'expertiser l'article introduisant l'analyse discriminante sur vecteurs propres de voisinage (C *et al.* 2006, cf. Annexe 4) – aujourd'hui sous presse dans *Candollea* mais précédemment soumis aux *Comptes Rendus Biologie* – nous a fait la remarque suivante : "*la discrimination spatiale reste un sujet d'actualité mais les méthodes factorielles semblent vieillottes dans ce contexte*". Cette remarque est caractéristique d'une opinion répandue, qui prétend que l'analyse se résume au choix de la meilleure méthode statistique. Or, il n'existe pas de "bonnes" ni de "mauvaises" méthodes. Il n'y a que de bonnes et de mauvaises démarches. L'utilisation automatique de méthodes "à la mode" sous-entend que l'analyse est une simple étape technique de l'étude. Comme de nombreux auteurs, nous défendons que l'analyse est centrale à la compréhension des données, et qu'en cela, la Biométrie est un *métier* (T *et al.* 1993). Les méthodes statistiques constituent alors des outils qui doivent être adaptés au besoin, et non des recettes de cuisine à appliquer à la lettre et de façon systématique. Le rôle de la Biométrie est de faciliter la compréhension du système étudié dans sa globalité.

9.2 L'

Cette approche systémique de l'analyse de la sélection de l'habitat a été grandement facilitée par le développement récent de l'informatique, et notamment du logiciel R, outil biométrique par excellence. En cela, la bibliothèque de fonctions **adehabitat** facilite l'application de la démarche défendue dans ce mémoire.

En effet, avant l'apparition du logiciel R, utiliser une méthode statistique particulière pour étudier la sélection de l'habitat pouvait prendre un temps considérable. L'étude de la sélection de l'habitat nécessitait l'utilisation d'un grand nombre de logiciels – logiciels de statistiques, systèmes d'information géographique, logiciels de gestion de bases de données, etc. Appliquer une méthode particulière impliquait de jongler continuellement avec ces logiciels, et se traduisait souvent par la mise en œuvre de tâches très répétitives. Il valait mieux, pour celui qui était responsable de l'analyse de données, savoir *a priori* quelle méthode était susceptible de lui renvoyer les meilleurs résultats concernant les structures qu'il étudiait. Etant donné le temps limité

consacré aux analyses par rapport à la durée totale de l'étude (préparation du protocole, collecte des données, etc.), la tentation était grande "de faire comme les autres", c'est-à-dire d'appliquer des méthodes "qui ont fait leur preuve". Beaucoup y cédaient, contribuant à l'expansion de la vision "technique" de l'analyse de données.

Selon nous, la bibliothèque **adehabitat** change la donne. En effet, cette bibliothèque fournit un très grand nombre d'outils de base qui peuvent être facilement combinés avec le reste de l'environnement R pour permettre la construction d'une grande variabilité d'analyses, en un temps relativement court. Cette bibliothèque facilite donc la mise en œuvre de la démarche défendue dans ce mémoire. L'apport d'**adehabitat** à l'étude de la sélection de l'habitat est parfaitement résumé par C (1992) : "*Pour faire des images il faut du logiciel. On peut acheter du logiciel commercial. Mais que penserait-on d'un expérimentateur qui achèterait un stock de phrases toutes faites pour s'exprimer ? (...). Bien que ce ne soit pas normal d'un point de vue scientifique classique, les logiciels servent actuellement de support de diffusion d'idées. Documentation des logiciels et apprentissage méthodologique sont maintenant liés.*"

9.3 L

Le biométricien est fréquemment appelé en consultation. Le biologiste lui présente alors brièvement ses données, et lui demande alors quelle est la meilleure méthode à appliquer pour les analyser. Le biométricien est dans l'embarras. La rapidité avec laquelle les données ont été présentées ne lui permet pas de répondre à la question. En effet, dans sa pratique quotidienne de l'analyse, il peut s'écouler un laps de temps considérable entre le moment où le biométricien "met les mains dans le cambouis" et celui où il sait ce qu'il va faire. Par exemple, 396 messages électroniques et une vingtaine de réunions ont permis l'aboutissement des analyses menées en collaboration avec les Herbiers de Genève.

Il est alors impossible, lors de consultations, d'orienter le biologiste vers une méthode particulière sans une connaissance approfondie de ses données et de sa problématique, connaissance qui ne s'acquiert pas en une journée, et encore moins en quelques heures. Comme le soulignait L (1984), "*il ne serait pas très difficile à un écologue d'apprendre un peu de statistique et d'informatique, mais on peut se demander si cela en vaut toujours la peine, au-delà d'assurer la communication. (...) Il est clair qu'il est indécent de demander aux écologues d'apprendre un peu de mathématique pour répondre à ces questions*". Le biométricien ne doit pas – il ne peut pas – indiquer au biologiste une démarche qui lui permettra d'analyser ses données.

En revanche, il peut analyser lui-même les données, dans le cadre d'une collaboration. La situation est inconfortable, car le biologiste attend souvent des solutions à ses problèmes, et espère pouvoir les mettre en œuvre *lui-même*. Or, on ne peut être à la fois bon biométricien et bon biologiste. Le biométricien est nécessairement un mauvais biologiste (C 1992), mais l'inverse est aussi vrai.

9.4 P : L ´ GPS

Le récent développement de la technologie GPS (*Global Positionning System*) permettant le suivi automatique d'animaux sur une zone se traduit par l'apparition d'un nouveau type de données. Développer des méthodes permettant l'analyse de ce type de données était à l'origine l'un des objectifs de ma thèse. Cependant, par manque de temps, je n'ai pas pu traiter ce sujet. En revanche, j'ai eu l'occasion d'effectuer une étude bibliographique sur la question, ainsi que de collaborer avec des biologistes concernés par ce problème. La question "Comment traiter les données GPS ?" est devenue le leitmotiv des nombreuses réunions auxquelles j'ai pu assister au cours de ces trois années de thèse. Nous en expliquons les raisons dans ce paragraphe, et soulignons encore une fois l'importance de la démarche biométrique dans ce type d'analyse.

La principale différence entre la technologie GPS et le radio-pistage "classique" est que le récepteur du collier peut être programmé à l'avance pour collecter *de façon automatique* les localisations de l'animal qui le porte, à n'importe quel moment de la journée. La triangulation permettant cette localisation est effectuée grâce à des satellites, ce qui permet une économie en temps et en personnel par rapport au radio-pistage classique. On comprend alors le vif succès que rencontre cette technologie parmi les biologistes.

Mais cette facilité de collecter les localisations modifie le type de données récoltées. Puisque le relevé est automatique, rien n'interdit plus de collecter les localisations à des intervalles de temps très courts (e.g. toutes les 10 minutes) sur de très longues périodes (e.g. 6 mois). Et les données ne sont alors plus du type "*semis de points*", mais du type "*semis de trajectoires*". En effet, le plus souvent dans ce type d'études, la position d'un animal au temps t dépend de sa position au temps $t-1$. On parle ici d'autocorrélation séquentielle dans la position des localisations (*serial autocorrelation*).

Les méthodes d'étude de la sélection de l'habitat permettant de prendre en compte l'autocorrélation séquentielle sont très rares dans la littérature. Faute de grives, on mange des merles. La plupart des auteurs ignorent alors cette contrainte des données lors de l'analyse, et utilisent en général des outils qui supposent l'indépendance entre les localisations (O et W 1999).

Jodie M (2004) a conduit dans le cadre de son D.E.A. une étude sur l'impact de cette autocorrélation séquentielle sur la sélection de l'habitat par une ourse (*Ursus arctos*) introduite dans les Pyrénées, étude à laquelle j'ai collaboré en aidant à la programmation des analyses sous R. Elle utilise deux types de simulations de l'utilisation de l'habitat par l'ourse pour tester l'existence d'une sélection de l'habitat. Le premier type de test repose sur la comparaison de la marginalité et de la tolérance observées (§ 6.1) avec les valeurs simulées sous l'hypothèse d'une distribution aléatoire des points dans le domaine vital de l'animal, c'est-à-dire sans prise en compte des contraintes de déplacement. Le second type de test est similaire, mais l'utilisation aléatoire de l'habitat est simulée en construisant des *trajectoires* aléatoires possédant les mêmes propriétés que la trajectoire observée (mêmes angles entre les déplacements successifs, mêmes longueurs des déplacements), c'est-à-dire en prenant en compte spécifiquement les contraintes de déplacement dans les tests. Jodie M montre que dans le premier cas, la sélection de l'habitat apparaît très significative alors que dans le second, elle ne l'est pas. Ce travail souligne

l'importance de l'analyse des sources de l'autocorrélation séquentielle dans ce type d'étude. Un article sur ce sujet est actuellement en préparation.

Pourtant, même lorsque la dépendance entre localisations successives est prise en compte dans ce type d'étude, peu s'intéresse à la nature de cette dépendance. La plupart de ces études n'utilisent que des processus markoviens de premier ordre (R *et al.* 2005), c'est-à-dire des outils qui supposent que la position d'un point dépend de la position du point précédent. Mais il peut aussi y avoir une autocorrélation séquentielle entre les trajets successifs : l'orientation et la longueur du trajet effectué du temps t au temps $t + 1$ dépendent alors de l'orientation et de la longueur du trajet effectué entre le temps $t - 1$ et t. En outre, de même qu'il existe plusieurs types de processus de points, il existe plusieurs modèles de processus pour les trajectoires (R et K 1984, M C et C 1989, S et S 2001, N 2005). Enfin, de même que la distribution de points sur une zone n'est pas seulement causée par les variables environnementales, les trajectoires ne sont pas seulement induites par l'environnement. L'environnement a une influence, mais l'état interne de l'animal aussi (satiété / faim, repos, recherche de partenaires sexuels, etc.).

Les problèmes posés par l'autocorrélation séquentielle sont donc de même nature que ceux posés par l'autocorrélation spatiale. Il ne s'agit pas d'un défaut des données, contrairement à ce que pensent certains auteurs (S et S 1985), mais d'une de ses qualités (D S *et al.* 1999). La présence d'une autocorrélation séquentielle des localisations révèle la présence de structures dans ces déplacements. Ces structures doivent être interprétées biologiquement.

Indiquons cependant que la bibliothèque de fonctions **adehabitat** est au cœur de cette problématique. Lorsque cette bibliothèque a été rendue disponible sur CRAN en septembre 2004, plusieurs chercheurs étrangers ont pris contact avec moi afin de monter un groupe de travail international appelé "*Animove*" autour de la question de l'analyse de données GPS. Un site internet a été construit et est actuellement maintenu par Paolo C , avec pour principal objectif d'héberger un forum de discussion autour de l'analyse des déplacements d'animaux, sur lequel **adehabitat** occupe une place de choix (cf. URL : http ://www.faunalia.com/animov/). J'ai co-écrit avec Paolo C et Carlotta C un court didacticiel disponible sur ce site internet (Annexe 14), qui présente quelques outils de base de la bibliothèque de fonctions pouvant être utilisés lors des premières étapes de l'analyse des déplacements des animaux (URL : http ://www.faunalia.com/ animov/ howto.php). Le groupe, créé en Octobre 2004 compte à l'heure de la rédaction 73 membres appartenant à 10 pays.

Mais cette problématique a également donné lieu à la création d'un groupe "habitat" à l'échelon national, actuellement animé par Mathieu B (doctorant au laboratoire de Biométrie), et qui regroupe des biométriciens et des biologistes appartenant à différents organismes (CNRS, Université Lyon 1, Office national de la chasse et de la faune sauvage, Institut national de la recherche agronomique, CEMAGREF), tous concernés par la problématique de l'analyse des données GPS. La bibliothèque **adehabitat** occupe ici aussi une place importante, puisque le but de ce groupe de travail est de permettre la communication entre biométriciens et biologistes à travers l'analyse concrète de jeux de données, donc l'utilisation et le développement de fonctions dans la bibliothèque, ceci afin d'illustrer la démarche d'analyse de ce type de données. Les

interactions entre les 31 membres du groupe devraient permettre, à moyen terme, de se faire une idée des moyens à mettre en œuvre pour analyser les trajectoires des animaux dans les espaces géographiques et écologiques.

9.5 A

Cette thèse, financée par l'Office national de la chasse et de la faune sauvage (ONCFS), avait initialement pour objectif de développer des outils statistiques utilisables par les biologistes de cet organisme pour analyser la sélection de l'habitat par la faune sauvage, quel que soit le type de données initialement disponible. Or le principal résultat de cette thèse est que l'analyse de la sélection de l'habitat n'est pas une question d'outils, mais de démarche. Cette démarche s'apprend, mais l'analyse de données reste un métier. Les apports *directs* de cette thèse en matière de gestion de la faune sauvage sont en conséquence assez faibles. Il m'aurait été impossible de prévoir ce résultat à mon entrée en thèse, alors que je ne voyais le biométricien que comme un ingénieur appliquant des techniques statistiques.

En revanche, de façon plus indirecte, cette thèse met en lumière l'importance du rôle des biométriciens pour aider à l'analyse des données dans le domaine de la gestion. Les organismes chargés de la gestion de la faune sauvage collectent une quantité importante de données dans un but de recherche appliquée, et l'analyse peut les aider considérablement à construire un modèle des systèmes étudiés. Le chapitre 8 présente un modèle de colonisation d'un massif montagneux après un lâcher d'animaux. On comprend l'importance de ce type de modèles pour l'ONCFS, étant donné le nombre important d'opérations de réintroductions menées par cet organisme (D - 1990, K 1990, C *et al.* 2005b).

Soulignons que le biométricien restera toujours dépendant du biologiste pour l'interprétation des analyses ; il n'est pas et ne peut pas être autonome. Ainsi, le modèle de la colonisation des Bauges par le mouflon n'aurait pu être construit sans les données recueillies par Jean-Michel J , ni les connaissances biologiques et le travail considérable de Gaëlle D . L'analyse a pour rôle d'aider à formaliser le savoir du biologiste sous la forme de modèles conceptuels, et à préciser les questions qui se posent à travers les données qu'il collecte.

Ainsi, un biométricien pourrait apporter beaucoup à un organisme de gestion de la faune sauvage. En outre, les nouvelles technologies qui facilitent la collecte des données (e.g. les colliers GPS, ou les bases de données cartographiques de plus en plus complètes) ou leur analyse (e.g. le logiciel R, les systèmes d'information géographiques) sont aujourd'hui plus abordables, ce qui se traduit par une augmentation considérable du nombre d'études portant sur l'analyse de l'utilisation de l'espace par la faune sauvage. De plus en plus de données sont récoltées, et la nécessité de professionnels chargés de leur analyse se fait de plus en plus sentir.

Conclusion

"*Notre siècle est celui de la précision. Appuyée sur la statistique. La statistique est une science étonnante. Elle donne des certitudes chiffrées. Elle a prouvé que dans huit cas sur dix, les boulangers sont des hommes qui fabriquent du pain. Ce qui confirme un pressentiment qu'on avait déjà de cette affaire, mais sans preuve scientifique et par pure intuition. Et voilà ce qu'il y a de beau avec la statistique : ce qu'on savait bêtement avant elle, on le sait ensuite scientifiquement*."
Alexandre V (1989).

L'opinion d'Alexandre V est souvent exprimée par de nombreux biologistes et en particulier par le personnel de terrain, qui est au centre du système étudié et qui en a la vision la plus directe. L'un d'eux m'a dit un jour "*J'ai l'impression qu'avec la statistique, on passe son temps à vouloir faire rentrer de force les données dans des tiroirs, et on perd toute l'information intéressante*". Et il est vrai que c'est ce qui se passe lorsque l'analyse consiste simplement en l'application d'une méthode choisie *a priori*, avant même l'examen des caractéristiques des données. Avec cette stratégie, les données qui reflètent le système étudié sont "formatées" afin de les faire rentrer dans le cadre de l'analyse. Or, l'information considérée comme intéressante par le biologiste doit être prise en compte – mieux, valorisée – dans cette analyse. Ce ne sont donc pas les données qui doivent s'adapter à l'analyse, mais au contraire les analyses qui doivent s'adapter aux données. C'est là le rôle du biométricien.

Le biométricien formalise le savoir du biologiste par des modèles, c'est-à-dire des représentations simplifiées de la réalité qui permettent la communication entre les biologistes. L'Ecologie est une science qui étudie des objets complexes. Comme nous l'avons indiqué à de nombreuses reprises dans ce mémoire, la complexité des objets étudiés n'est pas à démontrer en Ecologie. Elle est établie. Cette complexité est exprimée par les interactions entre les différents éléments du système. On ne peut alors utiliser des techniques qui supposent *a priori* l'indépendance entre les éléments du système, ce qui ne peut que conduire à une approche trop réductionniste de l'analyse des données. Ce serait comme de vouloir comprendre le fonctionnement d'une montre en supposant que ses rouages sont indépendants. Or, la nature et l'intensité de ces interactions sont souvent des inconnues. C'est là que l'analyse peut aider le biologiste à trouver des réponses.

Il faut donc se garder d'une vision trop technique de l'analyse des données écologiques. Nous avions en introduction de cette thèse opposé le maçon et l'architecte, afin de souligner la différence entre technicien et concepteur. Nous pouvons à présent approfondir l'analogie entre

biométricien et architecte : pour construire une maison (un modèle), il faut un terrain stable (une bonne connaissance de la biologie de la ou les espèces, et de la zone d'étude), des matériaux tels que briques ou tuiles (les données), des outils tels que truelles ou grues (les méthodes statistiques), mais surtout un client qui sait ce qu'il veut (une problématique bien définie). Le dernier modèle de perceuse à percussion ne sera d'aucun secours si l'objectif est de construire une réplique de la tour Eiffel en pisé dans une zone marécageuse. En d'autres termes, ce ne sont pas les outils qui font l'analyse, mais la cohérence de tous les éléments de l'étude. Ces éléments doivent être examinés *en détail* par le biométricien, qui modifie au besoin l'un ou l'autre pour parvenir à une plus grande cohérence. Celle-ci est établie par le dialogue entre le biologiste et le biométricien, dialogue qui peut être très long.

Faire de la Biométrie est avant tout faire partie du monde scientifique. Ce *métier* est intégré dans un édifice plus vaste, interdisciplinaire. Refuser ce statut à la Biométrie ne peut que mener à une vision technique de la statistique. Et l'utilisation des outils statistiques avec une telle vision n'a d'autres objectifs que de prouver ce que le biologiste sait déjà. Ce serait alors donner raison à Alexandre V concernant le rôle de la statistique, c'est-à-dire qu'elle n'est là que pour donner un faux label de scientificité à la vision du monde du biologiste, lequel n'en a peut-être pas besoin. A cette étrange position, nous préfèrerons l'opinion de T , qui soulignait lors d'une récente conférence que "*la modélisation est un guide de la démarche scientifique*", idée que nous avons cherché à illustrer dans ce mémoire.

Références

R´ ´

A , N., R , P., et K , R. (1993). Compositional analysis of habitat use from animal radio-tracking data. *Ecology*, 74 : 1313–1325.

A , J. et R , J. (1986). Comparison of some statistical techniques for analysis of resource selection. *Journal of Wildlife Management*, 50 : 157–165.

A , J. et R , J. (1992). Further comparison of some statistical techniques for analysis of resource selection. *Journal of Wildlife Management*, 56 : 1–9.

A , S., M , B., M D , L., et G , G. (1996). Assessing habitat selection when availability changes. *Ecology*, 77 : 215–227.

A , P. (2000). *Le traitement des variables régionalisées en écologie. Apports de la géomatique et de la géostatistique*. Thèse de doctorat, Université Claude Bernard Lyon 1.

A , N., M , M., et B , S. (1996). An autologistic model for the spatial distribution of wildlife. *Journal of Applied Ecology*, 33 : 339–347.

A , M. (1999). A silent clash of paradigms : some inconsistencies in community ecology. *Oikos*, 86 : 170–178.

A , M. (2002). Spatial prediction of species distribution : an interface between ecological theory and statistical modelling. *Ecological modelling*, 157 : 101–118.

B , A. et T , R. (2000). Practical maximum pseudolikelihood for spatial point patterns. *Australian and New Zealand Journal of Statistics*, 42 : 283–322.

B , A. et T , R. (2005). spatstat : An R package for analysing spatial point patterns. *Journal of Statistical Software*, 12 : 1–42.

B , T. et G , A. (1995). *Interactive spatial data analysis*. Longman Group Limited.

B , S. et W , A. (2002). Generalized additive modelling and zero inflated count data. *Ecological modelling*, 157 : 179–188.

B , M. (2004). *Le lynx, l'ENFA et le SIG : histoire de la sélection de l'habitat chez le lynx*. Mémoire de dea, Université Claude Bernard Lyon 1.

B , L. et H , L. (2002). Potential changes in the distributions of latitudinally restricted Australian butterfly species response to climate change. *Global Change Biology*,

8 : 954–971.

B , S. (1998). Le domaine vital des mammifères terrestres. *Revue d'écologie - La Terre et la Vie*, 53 : 309–334.

B , J. et A , D. (1995). Multiple significance tests : the Bonferroni method. *British Medical Journal*, 310 : 170–170.

B , R., D , M., et M , M. (1993). Does age influence between-rams companionship in Mouflon (Ovis gmelini). *Revue d'écologie - La Terre et la Vie*, 48 : 57–64.

B , D. et L , L. (1994). Environmental control and spatial structure in ecological communities : an example using oribatid mites (Acari, Oribatei). *Environmental and Ecological Statistics*, 1 : 37–61.

B , D., L , P., et D , P. (1992). Partialling out the spatial component of ecological variation. *Ecology*, 73 : 1045–1055.

B , M. et M D , L. (1999). Relating populations to habitats using resource selection functions. *Trends in Ecology and Evolution*, 14 : 268–272.

B , M., V , P., N , S., et S , F. (2002). Evaluating resource selection functions. *Ecological modelling*, 157 : 281–300.

B , J., E , L., R , L., N , C., C , C., L , P., et M , E. (2005). Is body mass of rock ptarmigan greater in the pyrenees than in the Alps ? *Wildlife Biology*, sous presse.

B , L., T , W., A , M., et H , A. (2004). Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography*, 27 : 437–448.

B , B. et B , G. (1984). Habitat selection by fox and gray squirrels : a multivariate analysis. *Journal of Wildlife Management*, 48 : 616–621.

B , K. et A , D. (1998). *Model selection and inference.* Springer, Berlin.

B , J. (1991). BIOCLIM - a bioclimate analysis and prediction system. In M , C. et A , M., editors, *Nature conservation : cost effective biological surveys and data analysis*, pages 64–68. CSIRO, Melbourne.

B , S. et M , J. (2003). Defining availability and selecting a currency of use : key steps in modeling resource selection. In H , S., editor, *Resource selection methods and applications. Proceedings of the First International Conference on Resource Selection*, pages 1–11, Laramie, Wyoming. Omnipress, Madison.

B , C., S , R., et K , P. (1981). Clarification of a technique for analysis of utilization-availability data. *Journal of Wildlife Management*, 48 : 1050–1053.

C , C. (2002). Problèmes des études de la sélection de l'habitat reposant sur des données de radio-pistage : cas du sanglier en milieu méditerranéen. Rapport technique, Office National de la Chasse et de la faune sauvage.

C , C. (Soumis). The package adehabitat for the R software : a tool for the analysis of space and habitat use by animals. *Ecological modelling*.

C , C., B , M., et L , P. (2003). Separation of ecological niches of galliform mountain birds in the Northern Alps (Vanoise National Park). *Game and Wildlife Science*, 20 : 259–285.

C , C., C , C., et S , R. (Soumis). The Kernel method for the estimation of the geographical distribution of species occurrences. *Applied Vegetaion Science*.

C , C. et D , A. (Soumis). Factorial analysis of selection ratios from animal radio-tracking data. *Ecology*.

C , C., D , A., et M , D. (2005a). K-select analysis : a new method to analyse habitat selection in radio-tracking studies. *Ecological modelling*, 186 : 143–153.

C , C., M , D., F , P., et F , C. (2004). Efficiency of spreading maize in the garrigues to reduce wild boar damage to Mediterranean vineyards. *European Journal of Wildlife Research*, 50 : 112–120.

C , C., M , D., G , J., M , L., et P , R. (2002a). Elephant damage to trees of wooded savanna in Zakouma National Park. *Journal of Tropical Ecology*, 18 : 599–614.

C , C., M , D., I , N., et G , J. (2005b). Reintroduction of roe deer into a Mediterranean habitat : female mortality and dispersion. *Wildlife Biology*, 11 : 153–161.

C , C., M , D., V , J., et B , S. (2002b). Summer and hunting season home ranges of wild boar (Sus scrofa) in two habitats in France. *Game and Wildlife Science*, 19 : 281–301.

C , C., S , R., C , D., et C , C. (2006). The discriminant analysis of the spatial distribution of vegetal species occurrences : I. theoretical aspects. *Candollea*, sous presse.

C , G., G , A., et W , J. (1993). DOMAIN : a flexible modelling procedure for mapping potential distributions of plants and animals. *Biodiversity and Conservation*, 2 : 667–680.

C , S., H , C., et M , J. (1999). Home range and habitat selection of bog turtles in southwestern Virginia. *Journal of Wildlife Management*, 63 : 853–860.

C , E., A , M., et B , B. (2002). Regional vegetation mapping in Australia : a case study in the practical use of statistical modelling. *Biodiversity and Conservation*, 11 : 2239–2274.

C , J. (1998). *Programming with data. A guide to the S language*. Springer, New-York.

C , R., H , J., et L , J. (2005). Potential distribution modelling, niche characterisation and conservation status assessment using GIS tools : a case study of Iberian Copris species. *Biological conservation*, 122 : 327–338.

C , S. (1996). A comparison of confidence interval methods for habitat use-availability studies. *Journal of Wildlife Management*, 60 : 653–658.

C , S. (1998). Statistical tests in publications of The Wildlife Society. *Wildlife Society*

Bulletin, 26 : 947–953.

C , D. (1992). *Echanges interdisciplinaires en analyse de données écologiques*. Mémoire d'habilitation à diriger des recherches, Université Claude Bernard Lyon 1.

C , D., D , A., et T , J. (2004). The ade4 package - I : One table methods. *R News*, 4 : 5–10.

C , D. et G , C. (1998). Analyses canoniques et listes d'occurrences d'espèces. Documentation thématique du logiciel ade-4.

C , D., L , J., et Y , N. (1987). Propriétés de l'analyse canonique des correspondances. Une utilisation en hydrobiologie. *Revue de Statistique Appliquée*, 35 : 55–72.

C , J. (1983). The estimation and analysis of preference and its relationship to foraging models. *Ecology*, 64 : 1297–1304.

C , J., D , J., et S , K. (1993). A multivariate model of female black bear habitat use for a geographic information system. *Journal of Wildlife Management*, 57 : 519–526.

C , W. et D , S. (1988). Locally weighted regression : an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83 : 596–610.

C , A. et O , J. (1981). *Spatial processes. Models and applications*. Pion limited, London.

C , D. et C , D. (1995). Landscape-level habitat use by brown-headed cowbirds in Vermont. *Journal of Wildlife Management*, 59 : 631–637.

C , R. et S , C. (1999). Using spatial pattern analysis to distinguish causes of mortality : an example from kelp in north eastern New Zealand. *Journal of Ecology*, 87 : 963–972.

C , L. (1998). Measuring the degree of sexual segregation in group-living animals. *Journal of Animal Ecology*, 67 : 217–226.

C , N. (1991). *Statistics for spatial data*. Wiley series in probability and mathematical statistics, New York.

C , J. (1993). Problèmes posés par la flexibilité du comportement social du Mouflon de Corse (Ovis ammon musimon) pour le dénombrement, par "approche et affût combinés". *Gibier Faune Sauvage*, 10 : 77–80.

C , J. et C , D. (1992). Statut ancien et actuel du Chevreuil dans le département du Gard, perspectives. *Bulletin Mensuel de l'Office National de la Chasse*, 164 : 26–38.

D , M., D , P., F , M., L , L., M , D., et R , D. (2002). Conceptual and mathematical relationships among methods for spatial analysis. *Ecography*, 25 : 558–577.

D , M. R. T. (1999). *Spatial pattern analysis in plant ecology*. Cambridge Studies in Ecology. Cambrige University Press.

C , T. (2005). *Composantes de la distribution spatiale d'un prédateur : effets respectifs de l'habitat, des ressources alimentaires et des interactions comportementales. Analyse de processus ponctuels non homogènes*. PhD thesis, Université Claude Bernard

Lyon 1.

D S , S., B , R., et B , R. (1999). Eliminating autocorrelation reduces biological relevance of home range estimates. *Journal of Animal Ecology*, 68 : 221–234.

D , P. (1983). *Statistical analysis of spatial point patterns*. Academic Press, London.

D , M., B , A., et R , D. (2001). Tree mortality in an unmanaged mountain pine (Pino mugo var. uncinata) stand in the Swiss National Park impacted by root rot fungi. *Forest Ecology and Management*, 145 : 79–89.

D ´ , S., C , D., et O , J. (1995). L'analyse des correspondances décentrée : application aux peuplements ichtyologiques du haut-Rhône. *Bulletin Français de la Pêche et de la Pisciculture*, 336 : 29–40.

D , S., C , D., et G -C , C. (2000). Niche separation in community analysis : a new method. *Ecology*, 81 : 2914–2927.

D , N. (2004). *Eléments d'écologie de la population d'éléphants du parc national de Zakouma (Tchad)*. Thèse de doctorat, Ecole Nationale du Génie Rural, des Eaux et Forêts.

D , C. (1990). Non-parametric estimates of interaction from radio-tracking data. *Journal of Theoretical Biology*, 143 : 431–443.

D , M., R , D., et O , C. (1987). Use of geographic information systems to develop habitat suitability models. *Wildlife Society Bulletin*, 15 : 574–579.

D , S. (1999). *Utilisation des listes d'occurrences spécifiques spatialisées en Ecologie et en Biogéographie*. Mémoire de dea, Université Claude Bernard Lyon 1.

D , S. (2003). *Elements d'interface entre analyses multivariées, systèmes d'information géographique et observations écologiques*. Thèse de doctorat, Université Claude Bernard Lyon 1.

D , M., M , M., G´ , J., K , K., et H , G. (1992). Etude des domaines saisonniers de femelles de Mouflon Corse (Ovis ammon) dans le massif du Caroux-Espinouse (Hérault). *Bulletin Mensuel de l'Office National de la Chasse*, 172 : 29–34.

D , M., Q , P., B , E., et M , M. (1993). Seasonal range use by European mouflon rams in medium altitude mountains. *Acta theriologica*, 38 : 185–198.

D , D. (1990). Réintroduction du cerf de Corse (Cervus elaphus corsicanus) en Corse : Problématique et état actuel de l'opération. *Revue d'écologie - La Terre et la Vie*, Suppl. 5 : 135–144.

E , C. (1927). *Animal ecology*. Sidgewick and Jackson, London.

E , R., G , A., et R , L. (2004). An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, 41 : 263–274.

E , W., M D , T., et S , R. (1998). Habitat selection using GIS data : a case study. *Journal of Agricultural, Biological, and Environmental Statistics*, 3 : 296–310.

E , Y. (1987). The duality diagram : a means of better practical applications. In L - , P. et L , L., editors, *Development in numerical ecology*, Series G, pages 139–156. Springer Verlag, Berlin.

ESRI (1996). Using ArcView GIS. Technical report, Environmental Systems Research Institute, Inc.

F , J., L , D., N , H., S , J., et S , J. (2001). Climate and animal distribution : a climatic analysis of the Australian marsupial Trichosurus caninus. *Journal of Biogeography*, 28 : 293–304.

G , K. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58 : 453–467.

G , J., D , P., D , D., V L , G., P , N., M , D., et R , G. (2003). Effects of Hurricane Lothar on the population dynamics of european roe deer. *Journal of Wildlife Management*, 67 : 767–773.

G , J., V , J., et K , F. (1988). Quelques caractéristiques de la dynamique des populations de Sangliers (Sus scrofa scrofa) en milieu chassé. *Gibier Faune Sauvage*, 4 : 31–47.

G , D., C , F., E ` , M., et G , J. (2004). Descriptive biogeography of Tomicus (Coleoptera : Scolytidae) species in Spain. *Journal of Biogeography*, 31 : 2011–2024.

G , O. et R , K. (1999). Gestalt Psychologie et cognition sans langage. Actualité d'une figure historique. *Intellectica*, 28 : 229–250.

G , M. et C , J. (En préparation). Evolution de la distribution du Mouflon méditerranéen (Ovis gmelini musimon x Ovis sp.) dans le Massif du Caroux-Espinouse (Hérault-France).

G , M., C , J., L , A., G , J., V , C., et M , D. (2005). Monitoring the abundance of mouflon in South France. *European Journal of Wildlife Research*, 51 : 69–76.

G , W. et W , C. (2004). A local nearest-neighbor convex-hull construction of home ranges and utilization distributions. *Ecography*, 27 : 489–505.

G , J., D , C., et B , S. (1999). Comparing the performances of Diggle's tests of spatial randomness for small samples with and without edge-effect correction : application to ecological data. *Biometrics*, 55 : 156–164.

G -C , C. (1999). *Analyse de la biodiversité à partir d'une liste d'occurrences d'espèces : nouvelles méthodes d'ordination appliquées à l'étude de l'endémisme dans les Ghâts occidentaux*. Thèse de doctorat, Université Claude Bernard Lyon 1.

G -C , C., D , S., et P , J. (2003). Large-scale biodiversity pattern analyses of the endemic tree flora of the western Ghats (India) using canonical correlation analysis of point data. *Ecography*, 26 : 429–444.

G , R. (1968). Trend-surface analysis of ecological data. *Journal of Ecology*, 56 : 845–869.

G , F. et P , R. (2000). Spatial structure analysis of heterogeneous point patterns : examples of application to forests stands. Fiche thématique du logiciel ade-4.

G , F. et P´ , R. (1999). On explicit formulas of edge correction for Ripley's K function. *Journal of Vegetation Science*, 10 : 433–438.

G , J. (2003). Unified Biplot geometry. In F , A. et M , A., editors, *Developments in applied statistics*, pages 3–23. Metodoloski zveski, Ljubljana.

G , R. (1971). A multivariate statistical approach to the hutchinsonian niche : bivalve molluscs of central Canada. *Ecology*, 52 : 544–556.

G , M. (1984). *Theory and applications of correspondence analysis*. Academic Press, London, UK.

G , J. (1917). Field tests o theories concerning distributional control. *American Naturalist*, 51 : 115–128.

G , A., E , T., et H , T. (2002). Generalized linear and generalized additive models in studies of species distributions : setting the scene. *Ecological modelling*, 157 : 89–100.

G , A. et H , U. (2003). Predicting reptile distributions at the mesoscale : relation to climate and topology. *Journal of Biogeography*, 30 : 1233–1243.

G , A., W , S., et W , A. (1999). GLM versus CCA spatial modeling of plant species distribution. *Plant Ecology*, 143 : 107–122.

G , A. et Z , N. (2000). Predictive habitat distribution models in ecology. *Ecological modelling*, 135 : 147–186.

H , P. (1995). Spatial pattern analysis in ecology based on Ripley's K function : introduction and methods of edge correction. *Journal of Vegetation Science*, 6 : 575–582.

H , P., P , F., C , S., et I , L. (1996). Spatial patterns in a two tiered semi-arid shrubland in southeastern Spain. *Journal of Vegetation Science*, 7 : 527–534.

H , P., P , F., C , S., et I , L. (1997). Spatial pattern in Anthyllis cytisoides shrubland on abandoned land in southeastern Spain. *Journal of Vegetation Science*, 8 : 627–634.

H , A. (1993). Habitat selection by mountain beavers recolonizing Oregon coast range clearcuts. *Journal of Wildlife Management*, 57 : 847–853.

H , L., K , P., et M , M. (1997). The habitat concept and a plea for standard terminology. *Wildlife Society Bulletin*, 25 : 173–182.

H , D. (1997). *Matrix algebra from a statistician's perspective*. Springer, New York.

H , T. et T , R. (1990). *Generalized additive models*. Chapman and Hall, London.

H , W. (1971). Contingency-table analysis of rain forest vegetation. In P , G., P , E., et W , W., editors, *Statistical Ecology. III Many species populations ecosystems and systems analysis*, pages 271–314. Pennsylvania State University Press.

H , J. (1995). *Säugetiere der Schweiz - Mammifères de la Suisse - Mammiferi della Svizzera.* Birkhäuser Verlag, Basel.

H , F., Z , J., et Z , H. (2003). Autologistic regression model for the distribution of vegetation. *Journal of Agricultural, Biological, and Environmental Statistics*, 8 : 205–222.

H , E. (2002). The outer border and central border for species-environmental relationships estimated by non-parametric generalized additive models. *Ecological modelling*, 157 : 131–139.

H , D. (1985). Analysing selection experiments with log-linear models. *Ecology*, 66 : 1744–1748.

H , E. et V , M. (1991). Macrohabitat use by black bears in a southeastern wetland. *Journal of Wildlife Management*, 55 : 442–448.

H , C. et H , B. (2002). Distance-based parametric bootstrap tests for clustering of species ranges. Technical report, Zürich, Switzerland.

H , L., M , B., H , D., E , R., et S , H. (2003). A resource selection probability function for the northern spotted owl in Plum Creek's Central Cascade habitat conservation plan, Washington state. In H , S., editor, *Resource selection methods and applications. Proceedings of the First International Conference on Resource Selection*, pages 1–9, Laramie, Wyoming. Omnipress, Madison, Wisconsin & Omnipress, Madison, Wisconsin, and Western EcoSystems Technologies, Cheyenne, Wyoming.

H , M. (1973). Reciprocal averaging : an eigenvector method of ordination. *Journal of Ecology*, 61 : 237–249.

H , M. (1991). Patterns of species distribution in Britain elucidated by canonical correspondence analysis. *Journal of Biogeography*, 18 : 247–255.

H , M. et S , A. (1976). Principal component analysis of taxonomic data with multi-state discrete characters. *Taxon*, 25 : 249–255.

H , A. (2004). Biomapper 3 User's Manual. Technical report, Laboratory for Conservation Biology, University of Lausanne.

H , A. et A , R. (2003a). Environmental-envelope based habitat-suitability models. In M , B., editor, *1st Conference on Resource Selection by Animals*, pages 67–76, Laramie, Wyoming. Omnipress, Madison.

H , A. et A , R. (2003b). Modeling habitat suitability for complex species distributions by environmental-distance geometric mean. *Environmental Management*, 32 : 614–623.

H , A. et G , A. (2002). Which is the optimal sampling strategy for habitat suitability modelling. *Ecological modelling*, 157 : 331–341.

H , A., H , J., C , D., et P , N. (2002). Ecological-niche factor analysis : How to compute habitat suitability maps without absence data ? *Ecology*, 83 : 2027–2036.

H , A., P , B., O , P., C , Y., G , C., et A , R. (2004). Ecologi-

cal requirements of reintroduced species and the implications for release policy : the case of the bearded vulture. *Journal of Applied Ecology*, 41 : 1103–1116.

H , N. et H , T. (1990). Habitat evaluation : do use/Availability data reflect carrying capacity. *Journal of Wildlife Management*, 54 : 515–522.

H , K. et L , F. (2001). Vienna and R : Love, Marriage and the Future. In D , R., editor, *Festschrift. 50 Jahre Österreichische Statistische Gesellschaft*, pages 61–70. Österreichische Statistische Gesellschaft, Vienna, Autriche.

H , G. (1957). Concluding remarks. In *Cold Spring Harbour Symposium*, volume 22, pages 415–427. Quantitative Biology.

J , C. et G , A. (2001). Modelling the distribution of bats in relation to landscape structure in a temperate mountain environment. *Journal of Applied Ecology*, 38 : 1169–1181.

J , F. (1971). Ordinations of habitat relationships among breeding birds. *The Wilson Bulletin*, 83 : 215–236.

J , R., L , R., et C , T. (2000). Habitat utilization and home range of the redwing francolin, Francolinus levaillantii, in highland grasslands, Mpumalanga province, South Africa. *African Journal of Ecology*, 38 : 329–338.

J , D. (1980). The comparison of usage and availability measurements for evaluating resource preference. *Ecology*, 61 : 65–71.

J , D. (1981a). How to measure habitat - a statistical perspective. In C , D., editor, *The use of multivariate statistics in studies of wildlife habitat*, pages 53–57. USDA Forest Service.

J , D. (1981b). The use and misuse of statistics in wildlife habitat studies. In C , D., editor, *The use of multivariate statistics in studies of wildlife habitat*, pages 11–19. USDA Forest Service.

J ¨ , J. et S , J. (1998). Distribution and habitat selection of wintering birds in urban environments. *Landscape and Urban Planning*, 39 : 253–263.

J , A., J , J., L , A., M , D., C , C., D , D., et L , J. (2004). Sélection de l'habitat par le chamois (Rupicapra rupicapra rupicapra L.) dans la Réserve nationale de chasse et de faune sauvage des Bauges (Savoie / Haute-Savoie). *Rapport scientifique de l'Office national de la chasse et de la faune sauvage.*, Sous presse.

K , J. (1981). Rationale and techniques for sampling avian habitats : introduction. In C , D., editor, *The use of multivariate statistics in studies of wildlife habitat*, pages 26–28. USDA Forest Service.

K , W. (2001). Spatial point pattern analysis of aerial survey data to assess clustering in wildlife distributions. *JAG*, 3 : 139–145.

K , F. (1990). La réintroduction du cerf (Cervus elaphus). *Revue d'écologie - La Terre et la Vie*, Suppl. 5 : 131–134.

K , S. et D , D. (1997). Distribution of black-tailed jackrabbit habitat determined by GIS in southwestern Idaho. *Journal of Wildlife Management*, 61 : 75–85.

K , S. et R , J. (1998). Limitations to mapping habitat use areas in changing landscapes using the Mahalanobis distance statistic. *Journal of Agricultural, Biological, and Environmental Statistics*, 3 : 311–322.

K , W. (2003). The role of habitat selection behavior in population dynamics : source-sink systems and ecological traps. *Oikos*, 103 : 457–468.

L , J. et D , B. (2004). Spatial point pattern analysis of available and exploited resources. *Ecography*, 27 : 94–102.

L , J., D , B., et R , P. (2003). Linking landscape patterns of resource distribution with models of aggregation in oviposing stream insects. *Journal of Animal Ecology*, 72 : 969–978.

L P , Y., B , L., G´ , J.-F., et M , M. (1995). Inter-individual associations and social structure of a mouflon population (Ovis orientalis musimon). *Behavioural processes*, 34 : 67–80.

L , J., C , D., P , R., et Y , N. (1988a). L'analyse des relations espèces-milieu par l'analyse canonique des correspondances. I - Variables de milieu quantitatives. *Acta Oecologica*, 9 : 53–71.

L , J., C , D., R -C , M., et Y , N. (1988b). L'analyse des relations espèces-milieu par l'analyse canonique des correspondances. II - Variables de milieu quantitatives. *Acta Oecologica*, 9 : 137–151.

L , J., S , R., B , G., et B , A. (1991). Principal component and correspondence analyses with respect to instrumental variables : an overview of their role in studies of structure-activity and species-environment relationships. In D , J. et K , W., editors, *Applied Multivariate Analysis in SAR and Environmental Studies*, pages 85–114. Kluwer Academic Publishers.

L , J. (1984). Sur les relations Biométrie-Ecologie. *Bull. Ecol.*, 15 : 117–119.

L , J. (1986). Contribution à l'étude de la complexité dans les systèmes biologiques. In *XIII colloque international d'économétrie appliquée*, pages 1–19, Sophia-Antipolis.

L , J. et S , A. (2004). *Philosophie de l'interdisciplinarité. Correspondance (1999-2004) sur la recherche scientifique, la modélisation et les objets complexes*. Transphilosophiques, Paris.

L , P. (1993). Spatial autocorrelation : trouble or new paradigm ? *Ecology*, 74 : 1659–1673.

L , P., D , M., F , M., G , J., H , M., et M , D. (2002). The consequences of spatial structure for the design and analysis of ecological field survey. *Ecography*, 25 : 601–615.

L , P. et F , M. (1989). Spatial pattern and ecological analysis. *Vegetatio*, 80 : 107–138.

L , P. et L , L. (1998). *Numerical Ecology, 2nd English edition*. Elsevier Science, Amsterdam.

L , A. (1998). GIS modeling of submerged macrophyte distribution using Generalized Additive models. *Plant Ecology*, 139 : 113–124.

L , A., L , J., et O , J. (2002a). Assessing New Zealand fern diversity from spatial predictions of species assemblages. *Biodiversity and Conservation*, 11 : 2217–2238.

L , A., O , J., et A , M. (2002b). Regression models for spatial prediction : their role for biodiversity and conservation. *Biodiversity and Conservation*, 11 : 2085–2096.

L , A., O , J., et L , J. (2002c). GRASP : generalized regression analysis and spatial prediction. *Ecological modelling*, 157 : 189–207.

L , A. (1995). The niche concept revisited : mechanistic models and community context. *Ecology*, 76 : 1371–1382.

L , J. (1999). Resource selection functions : taking space seriously ? *Trends in Ecology & Evolution*, 14 : 399–400.

L , S. (1992). The problem of pattern and scale in Ecology. *Ecology*, 73 : 1943–1967.

L˜ , A. (2003). Are certain habitats better every year ? A review and a case study on birds of prey. *Ecography*, 26 : 545–552.

L , J., S , J., et B , J. (1986). Bobcat habitat use and home range size in relation to prey density. *Journal of Wildlife Management*, 50 : 110–117.

L , E. et K , J. (1988). Comparison of pellet-group and radio-triangulation methods for assessing deer habitat use. *Journal of Wildlife Management*, 52 : 524–527.

L , A., G , W., et S , C. (2003). Black bear resource selection in the northeast Cascades, Washington. *Biological conservation*, 113 : 55–62.

M , D. (1996). *Occupation et utilisation de la garrigue et du vignoble méditerranéens par le sanglier (Sus scrofa L.)*. Thèse de doctorat, Université de droit, d'économie et des sciences d'Aix-Marseille III.

M , D. (2001). *De la recherche à la gestion des populations des grands mammifères terrestres*. Mémoire d'habilitation à diriger des recherches, Université Paul Valéry - Montpellier III.

M , D., B , B., et G , J. (1989). La masse corporelle : un bioindicateur possible pour le suivi des populations de chevreuil (Capreolus capreolus L.). *Gibier Faune Sauvage*, 6 : 57–68.

M , D., C , C., I , N., et G , J. (2002). Home range size and reproduction of female roe deer re-introduced into a Mediterranean habitat. *Zeitschrift für Jagdwissenschaft*, 48 : 194–200.

M , D., C , C., J , T., G , J., et M , L. (2001). The Kilometric

Index as a monitoring tool for populations of large terrestrial animals : a feasibility test in Zakouma National Park, Chad. *African Journal of Ecology*, 39 : 306–309.

M , D., M , N., et G , J. (1999). Variation saisonnière du domaine vital et sélectivité de l'habitat par le chevreuil en milieu méditerranéen : cas d'une femelle et d'un mâle adultes. *Revue d'écologie - La Terre et la Vie*, 54 : 71–87.

M , S., D , J., et O , S. (1999). Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions : a case study with a Himalayan river bird. *Ecological modelling*, 120 : 337–347.

M , B. (1994). *Multivariate Statistical Methods. A primer. Second edition*. Chapman & Hall, London.

M , B., M D , L., et T , D. (1993). *Resource selection by animals. Statistical design and analysis for field studies*. Chapman & Hall, London.

M , B., M D , L., T , D., M D , T., et E , W. (2002). *Resource selection by animals. Statistical design and analysis for field studies. Second Edition.* Kluwer Academic Publisher, London.

M , B., M , P., et C , L. (1972). Analysis of a selective predation experiment. *American Naturalist*, 106 : 719–736.

M , J. (2004). *Importance des structures spatiales et des contraintes comportementales dans les études de sélection de l'habitat*. Rapport de dea, Université Claude Bernard Lyon 1.

M C , S., R , M., K , R., et B , W. (1998). Evaluation of resource selection methods with different definitions of availability. *Journal of Wildlife Management*, 62 : 793–801.

M C , P. et N , J. (1989). *Generalized linear models. Second Edition.* Chapman & Hall, London.

M C , C. et C , M. (1989). Analysing discrete movement data as a correlated random walk. *Ecology*, 70 : 383–388.

M L , B. et H , F. (2001). Habitat selected by grizzly bears in a multiple use landscape. *Journal of Wildlife Management*, 65 : 92–99.

M , D. (2003). *Modélisation de la distribution spatiale du chamois de Chartreuse.* Licence professionnelle traitement de l'information géographique, Institut Universitaire de Technologie de Perpignan.

M , D., L , B., H , G., et G , P. (2000). Habitat selection models for eastern wild turkeys in central Mississippi. *Journal of Wildlife Management*, 64 : 765–776.

M , S. (1992). Tests of spatial and temporal interaction among animals. *Ecological applications*, 2 : 178–188.

M , C. (1947). Table of equivalent populations of North American small mammals. *American Midland Naturalist*, 37 : 223–249.

M , J., K , P., et F , T. (1987). Habitat use and seasonal range size of white-tailed deer in northcentral Minnesota. *Journal of Wildlife Management*, 51 : 644–648.

M , N. (1998). *Des outils biométriques appliques aux suivis des populations animales : l'exemple des cervidés. Vers un indice de consommation de la flore lignifiée.* Thèse de doctorat, Université Claude Bernard Lyon 1.

M , N., C , S., G , J., B , P., et B , Y. (2001). The browsing index : new tool uses browsing pressure to monitor deer populations. *Wildlife Society Bulletin*, 29 : 1243–1252.

M , M. (2001). A proposed research emphasis to overcome the limits of wildlife-habitat relationships studies. *Journal of Wildlife Management*, 65 : 613–623.

M , M., M , B., et M , R. (1992). *Wildlife-Habitat relationships. Concepts and applications*. The University of Wisconsin Press.

M´ , A., C , D., et S , R. (1993). Opérateurs de voisinage et analyse des données spatio-temporelles. In L , J. et A , B., editors, *Biométrie et environnement*, pages 45–72. Masson, Paris.

M , A. et I , R. (1998). Functional responses in habitat use : availability influences relative use in trade-off situations. *Ecology*, 79 : 1435–1441.

N , V. (2005). Using animal movements paths to measure response to spatial scale. *Oecologia*, Sous presse.

N , C., B , C., et P , J. (1974). A technique for analysis of utilization-availability data. *Journal of Wildlife Management*, 38 : 541–545.

N , S., B , M., et S , G. (2004). Grizzly bears and forestry. I Selection of clearcuts by grizzly bears in west-central Alberta, Canada. *Forest Ecology and Management*, 199 : 51–65.

N , M. et R , J. (1996). Microhabitat analysis using radiotelemetry locations and polytomous logistic regression. *Journal of Wildlife Management*, 60 : 639–653.

N -M , I. (1973). Data transformation in ecological ordination. I. Some advantages of noncentering. *Journal of Ecology*, 61 : 329–341.

O , J. et M , P. (2002). Continuum theory revisited : what shape are species responses along ecological gradients ? *Ecological modelling*, 157 : 119–129.

O , S. (2004). *Des outils pour l'intégration des contraintes spatiales, temporelles et évolutives en analyse des données écologiques*. Thèse de doctorat, Université Claude Bernard Lyon I.

O , P. et S -S , S. (2002). Should data be partitioned spatially before building large-scale distribution models ? *Ecological modelling*, 157 : 249–259.

O , D. (1997). Analysis of habitat selection studies with multiple patches within cover types. *Journal of Wildlife Management*, 61 : 1016–1022.

O , D. (1998). Analysis of the influence of spatial pattern in habitat selection studies. *Journal*

of Agricultural, Biological, and Environmental Statistics, 3 : 254–267.

O , D. et W , G. (1999). Autocorrelation of location estimates and the analysis of radiotracking data. *Journal of Wildlife Management*, 63 : 1039–1044.

P , M. (1993). Putting things in even better order : the advantages of canonical correspondence analysis. *Ecology*, 74 : 2215–2230.

P , G. (1993). Habitat-overlap of four wild ungulates in a Hugarian contiguous lowland forest. In T , I., editor, *Proceedings of the XXI IUGB Congress*, pages 348–356, Halifax N.S., Canada. International Union of Game Biologists.

P , P. (2003). *Habitat and corridor selection of an expanding red deer (Cervus elaphus) population*. Thèse de doctorat, Université de Lausanne.

P , J. et F , S. (2001). The practical value of modelling relative abundance of species for regional conservation planning : a case study. *Biological conservation*, 98 : 33–43.

P , G., T , K., D G , E., F , C., et L , R. (1998). Compositional analysis and GIS for study of habitat selection by goshawks in southeast Alaska. *Journal of Agricultural, Biological, and Environmental Statistics*, 3 : 280–295.

P , N. (1984). *Contribution à l'écologiedu genre Cepaea (Gastropoda) : approche descriptive et expérimentale de l'habitat et de la niche écologique*. Thèse de doctorat, Faculté des Sciences de l'Université de Lausanne.

P , J., L , A., R , M., D , J., M , M., J , A., et C - P , S. (2002). Illustrations and guidelines for selecting statistical methods for quantifying spatial pattern in ecological data. *Ecography*, 25 : 578–600.

P , D. et H , D. (1991). Home range and habitat use of coyotes in a farm region of Vermont. *Journal of Wildlife Management*, 55 : 433–441.

P , P. (1979). Likelihood measures of niche breadth and overlap. *Ecology*, 60 : 703–710.

P , N., G , J., D , P., M , D., V L , G., et D , D. (2003). Age and density modify the effects of habitat quality on survival and movements of roe deer. *Ecology*, 84 : 3307–3316.

P , N., G , J., Y , N., D , P., M , D., D , D., V L , G., et T , C. (2005). The response of fawn survival to changes in habitat quality varies according to cohort quality and spatial scale. *Journal of Animal Ecology*, 74 : 972–981.

P , P. (1967). Le mouflon de Corse (Ovis ammon musimon Schreber, 1782) ; Position systématique, écologie et éthologie comparées. *Mammalia*, 31 : 1–262.

P , D. et P , J. (1984). Moose habitat use and selection patterns in North-central Idaho. *Journal of Wildlife Management*, 48 : 1335–1343.

P , W., C , G., F , P., et C , R. (2001). Evaluation of Museum collection data for use in biodiversity assessment. *Conservation Biology*, 15 : 648–657.

P , W. et C , K. (1987). Effects of environmental pattern on habitat preference analysis. *Journal of Wildlife Management*, 51 : 681–685.

R , N., L , A., et J , P. (2002). Modeling spatial distribution of amphibian populations : a GIS approach based on habitat matrix permeability. *Biodiversity and Conservation*, 11 : 2143–2165.

R , R., L , J., et B , V. (2000). Spatial relationship of resident and migratory birds and canopy openings in diseased ponderosa pine forests. *Environmental Modelling & Software*, 15 : 189–197.

R , B., H , V., H , A., et V , P. (2003). Modelling habitat-suitability using museum collections : an example with three sympathic Apodemus species from the Alps. *Journal of Biogeography*, 30 : 581–590.

R , E., P , F., et D , M. (2000). Defining key habitats for low density populations of Eurasian badgers in Mediterraean environments. *Biological conservation*, 95 : 269–277.

R , J., M A , C., L , D., et P , H. (2005). A spatially explicit habitat selection model incorporating home range behavior. *Ecology*, 86 : 1199–1205.

R , J., O , R., et A , B. (1981). Bird community use of riparian habitats : the importance of temporal scale in interpreting discriminant analysis. In C , D., editor, *The use of multivariate statistics in studies of wildlife habitat*, pages 186–196. USDA Forest Service.

R , B. (1977). Modelling spatial patterns. *Journal of the Royal Statistical Society*, B39 : 172–212.

R , M., C , N., et V , M. (2001). A PCA-based modelling technique for predicting environmental suitability for organisms from presence records. *Diversity and Distributions*, 7 : 15–27.

R , H. (1981). Wildlife science : gaining reliable knowledge. *Journal of Wildlife Management*, 45 : 193–313.

R , R. et K , P. (1984). The search for resources by cabbage butterflies (Pieris Rapae) : Ecological consequences and adaptive significance of markovian movements in a patchy environment. *Ecology*, 65 : 147–165.

R , D. et M K , K. (1999). Estimation of habitat selection for central-place foraging animals. *Journal of Wildlife Management*, 63 : 1028–1038.

R , J. (1981). Why measure bird habitat ? In C , D., editor, *The use of multivariate statistics in studies of wildlife habitat*, pages 29–32. USDA Forest Service.

R , G. et R , M. (1967). A propos de quelques méthodes de classification en phytosociologie. *Revue de Statistique Appliquée*, 15 : 59–72.

R , M., W , M., J , B., et K , J. (2000). Elk distribution and modeling in relation to roads. *Journal of Wildlife Management*, 64 : 672–684.

R , B., B , A., T , R., et D , P. (2003). Rasp : a package for spatial statistics. *DSC 2003 Working Papers*, pages 1–8.

R , B. et D , P. (1993). Splancs : spatial point pattern analysis code in S-plus. *Computers and Geoscience*, 19 : 627–655.

S , S., P , N., et N , C. (2003). Winter habitat selection by two sympatric forest grouse in western switzerland : implication for conservation. *Biological conservation*, 112 : 373–382.

S , M. (2000). Corrélation entre facteurs écogéographiques et capacité de soutien chez le Chevreuil (capreolus capreolus) dans le Valais. Technical report, Université de Lausanne.

S , F. et S , L. (2001). Multifractal random walk in copepod behavior. *Physica A*, 301 : 375–396.

S , R. L. (1994). Annual variation in habitat selection : patterns concealed by pooled data. *Journal of Wildlife Management*, 58 : 367–374.

S , D., G , B., et P , R. (1998). KernelHR : a program for estimating animal home ranges. *Wildlife Society Bulletin*, 26 : 95–100.

S , D., M , J., K , B., B , G., R , K., et G , R. (1999). Effects of sample size on kernel home range estimates. *Journal of Wildlife Management*, 63 : 739–747.

S , C. (1983). Grizzly bear food habits, movements, and habitat selection in the mission mountains, Montana. *Journal of Wildlife Management*, 47 : 1026–1035.

S , B. (1986). *Density estimation for statistics and data analysis*. Chapman and Hall, London.

S , F. (2000). Potential plant distribution mapping based on climatic similarity. *Taxon*, 49 : 503–515.

S , R. et R , F. (1981). *Biometry. The principles and practice of statistics in Biological research. Second Edition*. W.H. Freeman and company, New York.

S , R., B , B., C , C., et C , C. (2006a). Biogeography of the forests of the Paraguay-paranà basin. In P , T., editor, *Neotropical Savannas and Dry Forests*, page Sous presse.

S , R., C , C., et B , B. (2004). The geographical zonation in the Neotropics of tree species characteristic of the Paraguay-Paraná Basin. *Journal of Biogeography*, 31 : 1489 – 1501.

S , R., C , C., et B , B. (2006b). The discriminant analysis of the spatial distribution of vegetal species occurrences : II. The spatial distribution of major tree communities in Paraguay. *Candollea*, 60.

S , R., P , R., C , A., et R , L. (1995). Origin, affinities and diversity hot spots of the Paraguayan dendrofloras. *Candollea*, 50 : 515–537.

S , D. et P , A. (2000). Recent applications of point process methods in forestry statistics. *Statistical Science*, 15 : 61–68.

S (1914). The elimination of spurious correlation due to position in Time or Space.

Biometrika, 10 : 179–180.

S , R. et S , N. (1985). Testing for independence of observations in animal movements. *Ecology*, 66 : 1176–1184.

T B , C. (1985). Correspondence analysis of incidence and abundance data : properties in terms of a unimodal reponse model. *Biometrics*, 41 : 859–873.

T B , C. (1986). Canonical correspondence analysis : a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67 : 1167–1179.

T B , C. (1987). The analysis of vegetation-environment relationships by canonical correspondence analysis. *Vegetatio*, 69 : 69–77.

T B , C. et L , C. (1994). Biplots in reduced-rank regression. *Biometrical Journal*, 36 : 983–1003.

T B , C. et P , I. (1988). A theory of gradient analysis. *Advances in Ecological research*, 18 : 271–317.

T , J. et C , D. (1992). A method for reciprocal scaling of species tolerance and sample diversity. *Ecology*, 73 : 670–680.

T , J., C , D., et C , S. (1995). Multivariate analysis of spatial patterns : a unified approach to local and global structures. *Environmental and Ecological Statistics*, 2 : 1–14.

T , D. et T , E. (1990). Study designs and tests for comparing resource use and availability. *Journal of Wildlife Management*, 54 : 322–330.

T , R., D , C., et M , J. (1993). *Modélisation de phénomènes biologiques*. Masson, Paris.

T , J., A , R., et L , J. (1996). Habitat use and ecological correlates of home range size in a small cervid : the roe deer. *Journal of Animal Ecology*, 65 : 715–724.

U , G. et F , B. (1985). *Spatial data analysis by example. Vol. 1 : Point pattern and quantitative data*, volume 1. John Wiley & Sons, Chichester.

V , P., S , F., et C , S. (2002). Modelling bird abundance from forest inventory data in the boreal mixedwood forests of Canada. In S , J., H , P., et M , M., editors, *Predicting Species Occurrences : Issues of Accuracy and Scale*, Island Press, pages 559–572. Washington D.C., USA.

V , A. (1989). *Chronique des grands micmacs*. Pocket, Paris.

V , J., G , J., et B , E. (1991). Kilometric index as biological indicator for monitoring forest roe deer populations. *Acta theriologica*, 36 : 315–328.

W , J. et M , R. (1997). Grizzly bear habitat selection in the swan mountains, Montana. *Journal of Wildlife Management*, 61 : 1032–1039.

W , Z., P , S., L , S., et L , Z. (2003). Spatial pattern of Cryptocarya chinensis life stages in lower subtropical forest, China. *Botanical Bulletin of Academia Sinica*, 44 : 159–166.

W , D. (1985). Canonical trend surface analysis : a method for describing geographic patterns. *Systematic Zoology*, 34 : 259–279.

W , N. (1993). *La méthode du noyau comme outil d'investigation de l'utilisation de l'espace dans une population de campagnols roussâtres*. Mémoire de dea, Université Claude Bernard Lyon 1.

W , G. et G , R. (1990). *Analysis of wildlife radio-tracking data*. Academic press, London, UK.

W , R. (1981). Applied aspects of choosing variables in studies of bird habitats. In C , D., editor, *The use of multivariate statistics in studies of wildlife habitat*, pages 28–41. USDA Forest Service.

W , T. et M , K. (2004). Rings, circles, and null-models for point pattern analysis in ecology. *Oikos*, 104 : 209–229.

W , B. (1989). Kernel methods for estimating the utilization distribution in home-range studies. *Ecology*, 70 : 164–168.

W , B. (1995a). A convex hull-based estimator of home-range size. *Biometrics*, 51 : 1206–1215.

W , B. (1995b). Using Monte Carlo simulation to evaluate kernel-based home range estimators. *Journal of Wildlife Management*, 59 : 794–800.

Z , A., L , A., et O , J. (2002). Predicting species spatial distributions using presence-only data : a case study of native New Zealand ferns. *Ecological modelling*, 2002 : 261–280.