LEGENDRE, P. & D. BORCARD [2003], Quelles sont les échelles spatiales importantes dans un écosystème? In: J.-J. DROESBEKE, M. LEJEUNE & G. SAPORTA (éds), *Analyse statistique de données spatiales.* Paris, Editions TECHNIP. (À paraître.)

CHAPITRE 19

QUELLES SONT LES ÉCHELLES SPATIALES IMPORTANTES DANS UN ÉCOSYSTÈME ?

par

Pierre LEGENDRE[1] et Daniel BORCARD

Département de sciences biologiques, Université de Montréal,
C.P. 6128, succursale Centre-ville, Montréal, Québec H3C 3J7, Canada

Courriel : Pierre.Legendre@umontreal.ca, Daniel.Borcard@umontreal.ca
Site WWWeb : http://www.fas.umontreal.ca/biol/legendre/

19.1 Introduction

L'écologie a pour objet l'étude des relations entre les êtres vivants et le milieu organique ou inorganique dans lequel ils vivent. Les écosystèmes sont des systèmes stochastiques : toutes les variables qu'on y mesure possèdent une variabilité naturelle. Cette variabilité se manifeste dans les données d'inventaire recueillies à l'aide de plans d'échantillonnage qui nous conduisent à échantillonner en différents sites (variabilité spatiale) ou à différents moments (variabilité temporelle), ou encore dans les résultats des expériences de terrain dans lesquelles la source principale de variabilité est générée par l'expérimentateur.

Nos recherches ont pour but de comprendre et de modéliser la structure spatiale des communautés plurispécifiques d'êtres vivants. Nous utilisons les tableaux d'abondances d'espèces (sites x espèces) car ceux-ci représentent la (multi)variable-réponse la plus pertinente pour étudier la réponse des écosystèmes à la variabilité naturelle ou celle qui est induite par l'action humaine.

Nous comprenons maintenant que les structures spatiales des communautés (gradients, agrégats, etc.) peuvent être générées par deux mécanismes principaux, agissant séparément ou de concert (Legendre et Legendre [1998]) :

• Dans le modèle d'autocorrélation, la valeur y_j observée au point (site) j d'une aire géographique est donnée par la moyenne μ_y du processus sur toute la surface, à laquelle s'ajoute la somme des influences provenant des points (sites) i situés près de j, ainsi qu'une erreur indépendante ε_j :

$$y_j = \mu_y + \sum f(y_i - \mu_y) + \varepsilon_j \qquad (1)$$

Les valeurs y_i sont les valeurs de **y** aux sites i, voisins du point j qui nous intéresse, qui sont situés à l'intérieur de la zone d'influence du processus générant l'autocorrélation spatiale. L'influence des points voisins de j peut s'exprimer, par exemple, comme une fonction de la distance entre i et j. Le terme total d'erreur est $[\sum f(y_i - \mu_y) + \varepsilon_j]$; il contient la composante « autocorrélation » de la variation. Le modèle (1) est supposé stationnaire.

[1] De septembre à décembre 2002 : Professeur invité, Chaire de Statistique Appliquée, Conservatoire National des Arts et Métiers (CNAM), 292 rue Saint Martin, F-75141 Paris cedex 03.

• Le modèle de dépendance spatiale suppose plutôt que la structure spatiale observée dans **y** résulte de l'influence d'une ou plusieurs variables explicatives qui possèdent elles-mêmes une structure spatiale :

$$y_j = \mu_y + f\,(\text{variables explicatives}) + \varepsilon_j \qquad (2)$$

y_i est la valeur de la variable-réponse au site j alors que ε_j est un terme d'erreur dont la valeur est indépendante d'un site à l'autre. f peut être n'importe quel type de fonction reliant les prédicteurs à la variable réponse **y**. Les prédicteurs agissent habituellement à plus grande échelle spatiale que la variable **y** ; si c'est le cas, lorsque la fonction f ou les variables responsables de la structure spatiale ne sont pas connues, on peut modéliser la structure spatiale de différentes façons, par exemple en lui ajustant une surface polynomiale, et la retirer de **y**. Les résidus $\mathbf{y}_{res}$ pourront alors être modélisés indépendamment de la tendance représentée par l'équation (2).

Cet article présente un survol des travaux que nous avons réalisés depuis une douzaine d'années en vue de décomposer la variation spatiale d'une variable **y** ou d'une multivariable **Y** en différentes fractions qui pourront être attribuées, selon nos hypothèses, à différentes causes ou processus générateurs.

19.2 Surfaces polynomiales multivariables

Dans un premier temps, nous avons utilisé la méthode des surfaces polynomiales pour ce type de modélisation (Borcard et al. [1992] ; Borcard et Legendre [1994] ; Legendre et Borcard [1994]). À la suite de Student [1914], nous exprimons les relations spatiales entre les sites d'échantillonnage sous la forme d'un polynôme des coordonnées géographiques des sites. Le polynôme est, le plus souvent, limité au degré 3. Les termes successifs du polynôme sont construits après avoir centré les coordonnées géographiques X et Y des sites, ce qui réduit la colinéarité entre les termes du premier et du second degré du polynôme.

Le polynôme peut être utilisé pour modéliser la réponse d'une seule variable par régression multiple. On élimine habituellement les termes du polynôme qui ne contribuent pas significativement à l'explication de la variable-réponse **y** ; cela permet d'obtenir un modèle parcimonieux. L'examen de la surface-réponse permet souvent de tirer des conclusions quant au processus générateur de la structure spatiale modélisée par la surface-réponse ; on trouve des exemples de modélisation spatiale d'une espèce à la fois dans Legendre et al. [1997].

La même méthode peut être employée pour étudier la réponse spatiale d'un tableau de données multivariables, comme l'a proposé Legendre [1990] ; la méthode est décrite en détail à la section 13.4 de Legendre et Legendre [1998]. La régression multiple est remplacée par l'analyse canonique de redondance (ACR : Rao [1964, 1973]) ou l'analyse canonique des correspondances (ACC : ter Braak [1986, 1987]). Les tableaux de structure des communautés écologiques (abondances d'espèces, **Y**), en particulier, peuvent être soumis à cette forme d'analyse. Puisqu'il s'agit souvent de « matrices creuses » de fréquences contenant de très nombreux zéros, des métriques appropriées à ces tableaux doivent être préservées lors de leur analyse. L'ACC préserve la métrique du χ^2 entre les lignes ou les colonnes du tableau. D'autres métriques, comme la distance de corde ou de Hellinger, peuvent être préservées entre les lignes (sites) si les données sont d'abord soumises à l'une des transformations décrites par Legendre et Gallagher [2001] ; les données ainsi transformées peuvent être analysées à l'aide des méthodes linéaires telles que l'ACP ou l'ACR. On peut enfin calculer toute autre matrice de distance appropriée aux données, puis revenir à une matrice rectangulaire par la méthode des coordonnées principales (Gower [1966]). Ce nouveau tableau de données peut faire l'objet d'une modélisation par analyse canonique de redondance (Legendre et Anderson [1999]) à l'aide du polynôme spatial des coordonnées des sites.

19.3 Partition de la variation

Nous avons étendu cette modélisation spatiale à l'analyse des relations entre les tableaux d'espèces **Y** et les tableaux de variables environnementales **X** en présence d'un tableau de variables spatiales **W**. Considérons pour le moment que le tableau **W** contient le polynôme des coordonnées spatiales des sites d'échantillonnage.

Par analyse canonique partielle, nous décomposons la variation du tableau d'abondances d'espèces en quatre composantes : [a] la variation non spatiale expliquée par les variables environnementales du modèle, [b] la variation structurée dans l'espace qui est explicable par ces mêmes variables environnementales, [c] la variation structurée spatialement qui n'est pas expliquée par les variables environnementales du modèle et [d] la variation résiduelle, non expliquée par les variables explicatives incluses dans le modèle (Figure 19.1). Il importe de rappeler ici que les fractions de variation isolées par ce procédé ne sont pas orthogonales entre elles et que la fraction [b] ne correspond nullement à une interaction au sens de l'analyse de variance. Des centaines d'applications de cette méthode ont été publiées, à ce jour, dans la littérature écologique.

Ce type d'analyse permet de se débarrasser de la structure spatiale, si celle-ci est considérée comme une source de fausses corrélations, et de dégager la fraction [a] non spatialisée pour l'interprétation. Si, par ailleurs, les composantes spatiale et non spatiale des variables explicatives **X** sont intéressantes pour expliquer la variation spatiale de **Y**, on s'intéressera à représenter et interpréter la composante [a+b] de la variation. Enfin, la composante [c] représente une fraction de la variation spatialisée de **Y** dont l'explication ne se trouve pas dans les variables **X** utilisées dans l'analyse. Une carte de la composante [c] permet de visualiser le patron de variation représenté par cette fraction, de réfléchir à son origine et de formuler de nouvelles hypothèses quant aux processus générateurs de cette structure.

Des tests de signification, réalisés par la méthode des permutations, permettent de déterminer si les quantités [a], [a+b], [b+c], [c] et [a+b+c], qui sont des statistiques de type R^2, sont significatives. On ne peut pas tester la signification de la fraction [b] prise de façon isolée puisque celle-ci est obtenue par soustraction ; elle ne correspond pas à un paramètre estimé au cours des analyses canoniques (Borcard et al. [1992] ; Legendre et Legendre [1998] ; Méot et al. [1998]).

On peut estimer les quatre fractions de variation à l'aide de trois analyses canoniques simples. Pour obtenir les vecteurs permettant de positionner les sites sur les axes canoniques des fractions [a] et [c], il faut réaliser deux analyses canoniques partielles supplémentaires. Cela permet de tracer des diagrammes d'ordination contenant les sites, les espèces, ainsi que les variables environnementales ou spatiales. Puisqu'on connaît la position géographique des sites, les composantes canoniques peuvent également être représentées sur des cartes, ce qui donne au chercheur une perception intuitive des facteurs responsables de la dynamique de la communauté. On trouvera des exemples de telles cartes dans Borcard et Legendre [1994], Legendre et Legendre [1998, section 13.5] et Borcard et al. [en préparation].

19.4 Modélisation spatiale multi-échelle : méthode CPMV

Le polynôme spatial utilisé dans les paragraphes précédents ne permet de modéliser que des phénomènes à échelle large[2]. Il faudrait trop de termes au polynôme, et donc trop de paramètres, pour arriver à modéliser les variations à échelle spatiale fine. Après plusieurs années de recherche, nous avons trouvé une méthode permettant de réaliser une décomposition spectrale de l'espace (Borcard et Legendre [2002]). Cette méthode fournit des variables correspondant à toutes les échelles spatiales perceptibles par le plan d'échantillonnage. On n'arrive évidemment pas à

[2] Les écologistes parlent d'*échelle large* (concernant une grande surface) ou *fine* (touchant une petite surface). Pour désigner l'échelle d'une carte, les géographes utilisent les expressions *grande échelle* (par ex. 1/25000) et *petite échelle* (par ex. 1/1000000) dans un sens opposé à l'acception écologique.

modéliser des phénomènes plus grands que la zone d'étude ou plus petits que l'intervalle entre les observations.

La méthode est décrite dans la Figure 19.2. Dans une première étape, nous fabriquons des variables spatiales correspondant au plan d'échantillonnage. À partir des coordonnées géographiques des sites, on calcule une matrice de distances géographiques entre ceux-ci. On tronque cette matrice à une certaine distance, choisie par l'utilisateur, en ne conservant que les distances entre sites proches les uns des autres. Les autres distances sont remplacées par une grande valeur ; n'importe quelle grande valeur fera l'affaire, pourvu qu'elle soit égale ou supérieure à 4 fois la valeur à laquelle on a tronqué les distances, comme nous l'avons constaté à l'usage. Une analyse en coordonnées principales (ACoP) de la matrice de distances tronquée, que nous appelons la « matrice de voisinage », produit plusieurs valeurs propres positives, au moins une valeur propre nulle et un certain nombre de valeurs propres négatives. On conserve les coordonnées principales (vecteurs propres de l'ACoP normés à $\lambda_i^{0,5}$) correspondant aux valeurs propres positives. Ceux-ci forment la matrice des « coordonnées principales d'une matrice de voisinage » (CPMV). Pour un échantillonnage à pas régulier le long d'un transect, si on a tronqué après la distance séparant deux sites immédiatement voisins, le nombre de variables CPMV est égal aux deux tiers du nombre de sites *n* le long du transect (valeur arrondie vers le haut).

L'intérêt de cette méthode apparaît lorsqu'on trace les variables CPMV. La Figure 19.3 montre les CPMV résultant d'un transect composé de 100 sites échantillonnés à pas régulier. La distance de tronquage est de 1 pas dans cet exemple. On s'aperçoit que les 67 variables CPMV constituent une décomposition spectrale des relations spatiales entre les sites d'échantillonnage. La méthode est directement utilisable pour des sites situés sur une ligne (un transect) ou une surface (une carte), qu'ils soient équidistants ou non. Pour des sites qui n'ont pas été échantillonnés à pas régulier ou qui ne se trouvent pas sur un transect, les variables CPMV ne se présentent pas comme des sinusoïdes, mais on peut encore reconnaître si elles appartiennent aux échelles large ou fine.

La seconde étape consiste à utiliser la matrice des variables CPMV comme matrice explicative de la régression multiple (s'il n'y a qu'une seule variable réponse *y*) ou de l'analyse canonique (si la matrice **Y** est multivariable), à la place du polynôme spatial des sections précédentes. Suivant le principe de parcimonie, on réalise une sélection des variables CPMV qui contribuent de façon significative à l'explication de **Y**. Plusieurs types de représentations peuvent s'ensuivre : on peut tracer des diagrammes de double ou triple projection (*biplot* ou *triplot* en anglais) des sites, des espèces composant la matrice **Y** et des variables CPMV retenues lors de l'analyse ; on peut diviser les variables CPMV en sous-modèles spatiaux orthogonaux entre eux, correspondant, par exemple, aux échelles large, moyenne et fine, puis calculer les valeurs ajustées à ces sous-modèles ; on peut enfin réaliser des analyses canoniques portant seulement sur les variables d'un sous-modèle spatial et tracer les diagrammes de double ou de triple projection correspondants.

L'article de Borcard et Legendre [2002] contient plusieurs séries de simulations visant à établir les propriétés statistiques ainsi que la robustesse de la méthode.

1. Nous avons d'abord vérifié si les tests statistiques associés à la méthode avaient une erreur de type I correcte. Pour 100 points équidistants le long d'un transect, nous avons simulé des valeurs d'une variable réponse ayant différents types de distribution aléatoire : uniforme, normale, exponentielle et exponentielle au cube. La matrice des variables CPMV contenait 67 colonnes lorsque nous tronquions les distances à la valeur 1, ou 55 colonnes lorsque nous tronquions les distances à la valeur 4. Nous avons calculé le coefficient de détermination (R^2) de la régression multiple (sans sélection préalable des variables explicatives) et réalisé un test par permutation (999 permutations). Après 5000 répétitions indépendantes de chaque simulation, nous avons calculé le taux de rejet de l'hypothèse nulle, de même que son intervalle de confiance. Les résultats montrent que le test a toujours une erreur de type I correcte, quelle que soit la distribution des valeurs aléatoires **y** soumises à l'analyse. Le pourcentage de variance expliqué par la régression fut, en moyenne, égal à son espérance pour des variables-réponse aléatoires. Ces résultats montrent que,

pour les données simulées tout au moins, le test de signification réalisé au cours de la méthode CPMV ne trouvera des structures spatiales significatives dans des données aléatoires qu'au niveau de signification α du test.

2. Une seconde série de simulations fut réalisée pour estimer la puissance du test du coefficient de détermination. Pour ce faire, nous avons généré des structures spatiales de différents types. (a) Nous avons d'abord testé la capacité de la méthode à détecter des structures spatiales dues à l'autocorrélation de données par ailleurs aléatoires (équation 1). Le test conservait une bonne puissance tant que le seuil de tronquage était plus petit que la largeur de la structure spatiale générée par l'autocorrélation dans les données aléatoires. (b) Nous avons ensuite généré des structures spatiales déterministes (équation 2) semblables à celles qui peuvent être induites par des variables environnementales : des structures spatiales de formes gaussiennes de différentes tailles, avec ou sans bruit, ainsi que des sinus de différentes périodes avec bruit. Les structures bruitées ont été générées avec différents rapports signal/bruit. Pour les structures de forme gaussienne et les sinusoïdes, le test détectait aisément les structures générées tant que la courbe gaussienne était plus large que le seuil de tronquage et que le rapport signal/bruit était élevé. (c) Nous avons enfin généré des gradients spatiaux linéaires qui représentent une autre forme courante de structure déterministe (équation 2) rencontrée dans les données écologiques. La méthode CPMV détectait toujours les gradients dans les données, quel que soit le rapport signal/bruit. Elle utilisait pour ce faire la moitié des variables CPMV (une sur deux en alternance), ne laissant que l'autre moitié des variables CPMV pour détecter d'autres structures éventuellement présentes dans **y**. Nous en concluons qu'il vaut mieux extraire la tendance linéaire des données, avant l'analyse, afin de conserver la puissance des variables CPMV pour détecter des structures plus intéressantes.

3. Des données à la structure complexe, ressemblant à certaines données écologiques réelles, ont été générées comme suit : sur un transect (100 points), nous avons additionné un gradient linéaire, une grosse bosse, 4 bosses plus petites, une sinusoïde dont la période était 1/17 de la longueur du transect, une variable aléatoire fortement autocorrélée et enfin un bruit blanc N(0,4) représentant 50 % de la variance des données (Figure 19.4). Nous avons fait l'analyse CPMV comme nous l'aurions faite pour un jeu de données réel : nous avons éliminé la tendance par régression, puis nous avons analysé les résidus par régression multiple sur les 67 variables CPMV. La sélection des variables a permis de retenir 8 variables CPMV qui représentaient ensemble 43,3 % de la variance de la série. La corrélation entre la somme des composantes spatiales introduites dans la série (sauf le gradient) et la structure ajustée aux 8 variables CPMV était 0,775. La corrélation entre le bruit blanc injecté dans la série et les résidus du modèle spatial obtenu après extraction de la tendance était 0,796. Le détail des calculs est illustré dans les figures 9 à 11 de Borcard et Legendre [2002]. Cet essai montre que la méthode CPMV est capable de retrouver dans les données des structures complexes appartement à différentes échelles spatiales.

À titre de complément, la Figure 19.5 présente une analyse CPMV sur une surface géographique. Un échantillonnage à pas régulier (maille de la grille : 1 km) a été réalisé à 63 sites de l'étang de Thau (Hérault, sud de la France ; longueur de l'étang : 19 km) (Amanieu et al. [1989]). Quatre des 45 variables CPMV sont présentées sous forme de cartes dans la Figure 19.5 (a-d). Les variables CPMV sont les éléments à partir desquels le modèle spatial sera construit. L'une des variables mesurée dans l'étang, la chlorophylle *a* (Chl *a*, Figure 19.5e), est modélisée à titre d'exemple ; cette variable est un bon indicateur de la quantité de phytoplancton présente dans l'eau. La variable Chl *a* fut régressée sur les 45 variables CPMV. Après élimination des CPMV qui ne contribuaient pas significativement au modèle, 12 variables CPMV furent conservées. Elles expliquent ensemble 78 % de la variance des données de Chl *a*. Ces variables furent divisées en trois groupes correspondant aux échelles large (CPMV 1, 3, 5 et 8), intermédiaire (CPMV 13, 14, 17, 19 et 20) et fine (CPMV 24, 28 et 36). Les valeurs de Chl *a* ajustées à ces sous-modèles ont été calculées ; elles sont représentées sous forme de cartes à la Figure 19.5 (f-h). Les structures spatiales sont moins évidentes dans les cartes des modèles d'échelle intermédiaire et fine à cause de la nature discontinue de l'échantillonnage. D'autres exemples d'application de l'analyse CPMV sont présentés dans l'article de Borcard et al. [en préparation].

19.5 Scalogramme

Lorsque l'échantillonnage a été réalisé à pas régulier le long d'un transect, les variables CPMV forment une série de sinusoïdes de période décroissante, comme celles de la Figure 19.2. On calcule une régression multiple de la variable-réponse **y** sur l'ensemble des variables CPMV. Les coefficients de régression indiquent la contribution des CPMV à l'explication de **y**. À chaque coefficient de régression est associée une statistique *t* (ou *F*) qui peut être testée pour sa signification statistique, soit à l'aide d'un test paramétrique, soit par la méthode des permutations. Notons au passage que, puisque les variables CPMV sont orthogonales les unes aux autres (linéairement indépendantes), les coefficients de régression simples et partiels sont les mêmes en régression simple (une variable CPMV à la fois) et en régression multiple ; les tests de signification diffèrent cependant.

Pour une matrice-réponse **Y**, on peut réaliser une analyse canonique (ACR ou ACC) partielle pour chaque variable CPMV, en plaçant toutes les autres variables CPMV dans la matrice des covariables. On obtiendra une statistique *F* partielle pour chaque variable CPMV ainsi qu'une probabilité permutationnelle associée.

Un scalogramme est un diagramme ayant les différentes variables CPMV en abscisse et une statistique indiquant la réponse de **y** ou **Y** en ordonnée. L'abscisse représente donc les différentes échelles spatiales, les plus larges à gauche et les plus fines à droite. Pour une variable-réponse **y**, la statistique pourra être le coefficient de régression partielle, la valeur absolue de la statistique *t* ou encore son carré qui est une statistique *F*. Pour un tableau-réponse **Y**, la statistique sera la valeur propre canonique divisée par la variation totale de **Y** ou encore la statistique *F*. Chaque point du scalogramme peut être représenté par un symbole correspondant au degré de signification de la statistique, selon le résultat du test statistique. La Figure 19.6 présente un scalogramme calculé pour la richesse spécifique des fougères dans un transect formé de 260 quadrats de végétation de 5 m x 5 m, échantillonné en Amazonie péruvienne (Tuomisto et Poulsen [2000]). La structure spatiale de ce transect a été analysée dans Borcard et al. [en préparation].

19.6 Modélisation multi-échelle de périmètres

Brind'Amour et al. [article soumis] ont étudié la répartition spatiale multi-échelle des poissons autour du lac Drouin, un lac de 31 ha, d'une profondeur maximale de 22 m, situé dans la région de Lanaudière au Québec (46°09' N, 73°55' W ; Figure 19.8a). Huit espèces de poissons ont été dénombrées par recensement visuel en apnée à 90 sites situés le long de la rive du lac. Les recensements ont été réalisés en juin et en août 2001. Les 90 sites ont été visités trois fois de suite chaque mois. Plusieurs variables environnementales ont également été mesurées à chaque site.

À l'occasion de cette étude, une variante de l'analyse CPMV a été développée pour étudier des transects fermés, comme ceux qu'on peut étudier à intérieur ou à extérieur du périmètre d'un lac ou d'une île, ou encore autour d'un bosquet. Le principe de cette analyse (Figure 19.7) est simplement de connecter les deux extrémités de la boucle, dans la matrice de distances, avant de calculer les coordonnées principales. Seules les distances entre sites voisins (Figure 19.7a, traits gras) sont transcrites dans la matrice de distance (Figure 19.7b) ; toutes les autres distances (traits fins entre sites non-adjacents) sont tronquées et représentées par une valeur égale à 4 fois la valeur maximale (Max = 1 dans cet exemple artificiel). Les coordonnées principales successives produisent des contrastes d'échelle, de plus en plus fins, qui se comparent tout à fait à ce qu'on obtient pour un transect rectiligne, à cette différence près qu'elles sont produites par paires ayant la même valeur propre ; on peut le vérifier en « circulant » autour de la boucle. CPMV 1 et 2 contrastent tous deux les sites en deux blocs ; CPMV 3 et 4 produisent quatre blocs (en deux paires opposées) orientés selon des angles différents ; CPMV 5 et 6 fragmentent le pourtour en 6 blocs, avec des regroupements dessinant des triangles. Les variables CPMV 1 et 2 ont des valeurs propres égales ; il en va de même de 3 et 4, puis de 5 et 6.

Dans l'article de Brind'Amour et al. [soumis], l'analyse a porté sur la structure de la communauté de poissons (abondance de 7 des 8 espèces), l'abondance totale des poissons toutes espèces confondues, ainsi que leur biomasse totale. Nous n'examinerons ici, à titre d'illustration de la méthode, que les résultats obtenus pour la communauté de 7 espèces de poissons au mois de juin ; la huitième espèce, qui représentait moins de 1% des observations, a été exclue des analyses. L'objectif était d'évaluer quelles étaient les variables environnementales responsables des patrons spatiaux observés à différentes échelles dans la communauté de poissons. L'analyse en coordonnées principales de la matrice de voisinage entre les 90 sites a produit 60 variables CPMV. Les abondances d'espèces furent transformées en utilisant la transformation de Hellinger proposée par Legendre et Gallagher [2001] ; les données transformées formèrent la matrice **Y** d'une analyse canonique de redondance (ACR) visant à établir, par sélection ascendante, quelles étaient les variables CPMV (matrice **X**) qui expliquaient la variation des espèces de façon significative. 15 variables CPMV ont été retenues par cette analyse. Celles-ci furent regroupées en 4 échelles : l'échelle très large (portée de plus de 2 km) correspondait à la CPMV 1 ; l'échelle large (500-1000 m) était formée des CPMV 3 à 7 ; l'échelle moyenne (200-450 m) était formée des CPMV 10, 11, 13, 18, 19 et 26; l'échelle fine (< 100 m) était formée des CPMV 36, 44 et 58. Pour chaque échelle, une nouvelle ACR fut calculée entre la matrice **Y** décrite ci-dessus et une matrice **X** ne contenant que les variables CPMV contribuant à l'échelle en question. Les valeurs propres canoniques furent testées (test par permutation), ce qui permit d'établir quels axes représentaient une réponse significative des poissons aux variables CPMV représentant l'échelle en question. Les espèces contribuant le plus fortement à un axe canonique significatif furent notées, dans chacune des figures (Figure 19.8 b-e), avec le signe de leur contribution au vecteur propre. Les espèces ayant un signe positif se retrouvent en plus grande abondance dans les sites représentés par des bulles pleines alors que les espèces munies d'un signe négatif se retrouvent surtout dans les bulles vides.

Dans la suite de l'analyse, chaque modèle composite correspondant à une échelle fut régressé sur les variables environnementales en vue de trouver les variables qui permettent d'expliquer la variation des espèces à cette échelle. Ainsi, 51 % de la variation de la composition en espèces à échelle très large est explicable par la présence de débris de bois au fond de la section du lac, la présence de forêt sur la berge et de macrophytes dans la section littorale du lac, ainsi que par la variable quantitative *fetch* qui mesure l'exposition au vent. 26 % de la variation à échelle large est explicable par la pente du littoral et de la rive ainsi que la présence de sable dans la zone littorale. La variation à échelle moyenne est expliquée à 20% par la présence de macrophytes dans la section littorale du lac ainsi que par la variable *fetch*. La variation à échelle fine n'est pas significativement expliquée par les variables environnementales relevées au cours de cette étude.

19.7 Conclusion

L'analyse CPMV représente un instrument puissant pour l'analyse de la répartition spatiale des variables écologiques, qu'il s'agisse de variables analysées une à la fois (**y**), comme la richesse spécifique, ou de toute la structure d'une communauté écologique qui forme un tableau multivariable (tableau **Y**). Cette forme d'analyse est applicable à des données à espacement régulier ou irrégulier. Elle permet d'obtenir une partition de la variation entre les variables explicatives environnementales et spatiales, des diagrammes mettant en évidence les relations entre les sites, les espèces et les variables environnementales. Pour les données récoltées à pas régulier le long d'un transect, l'analyse permet d'obtenir un scalogramme qui montre quelles sont les échelles spatiales importantes pour la variable ou la multivariable à l'étude. Un programme (SPACEMAKER), écrit par D.B., permet de fabriquer les variables CPMV correspondant à des surfaces ou des transects échantillonnés de façon régulière ou non. Ce programme est disponible gratuitement sur le site <http://www.fas.umontreal.ca/biol/legendre/>.

De nombreux aspects de l'analyse CPMV restent à explorer, notamment son comportement en présence de plans d'échantillonnage irréguliers, uni- ou bidimensionnels. Dans ces contextes, l'analyse peut déjà être appliquée pour quantifier la part de variation attribuable à la structuration spatiale des données, mais l'interprétation des variables spatiales en termes d'échelles précises est

plus délicate. Des analyses par simulations ainsi que des applications à des données réelles permettront d'optimiser ces aspects de la méthode.

Remerciements

Nous remercions Hanna Tuomisto, université de Turku (Finlande), qui nous fourni les données de richesse spécifique des fougères du transect Nauta, en Amazonie péruvienne. Ces données nous ont servi à calculer le scalogramme présenté à la Figure 19.6.

Références bibliographiques

AMANIEU, M., P. LEGENDRE, M. TROUSSELLIER & G.-F. FRISONI [1989], Le programme Écothau : théorie écologique et base de la modélisation, *Oceanologica Acta*, **12**, pp. 189-199.

BORCARD, D., P. LEGENDRE & P. DRAPEAu [1992], Partialling out the spatial component of ecological variation, *Ecology*, **73**, pp. 1045-1055.

BORCARD, D. & P. LEGENDRE [1994], Environmental control and spatial structure in ecological communities: an example using Oribatid mites (Acari, Oribatei), *Environmental and Ecological Statistics*, **1**, pp. 37-53.

BORCARD, D. & P. LEGENDRE [2002], All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices, *Ecological Modelling*, **153**, pp. 51-68.

BORCARD, D., P. LEGENDRE, C. AVOIS-JACQUET & H. TUOMISTO, Dissecting spatial structures of ecologial data at all scales, (en préparation.)

BRIND'AMOUR, A., D. BOISCLAIR, P. LEGENDRE & D. BORCARD, Multiscale spatial distribution of a littoral fish community in relation to environmental variables, (article soumis).

GOWER, J.C. [1966], Some distance properties of latent root and vector methods used in multivariate analysis, *Biometrika*, **53**, pp. 325-338.

LEGENDRE, P. [1990], Quantitative methods and biogeographic analysis, pp. 9-34 *in*: D.J. GARBARY & R.G. SOUTH (eds.), *Evolutionary biogeography of the marine algae of the North Atlantic*. NATO ASI Series, Vol. G 22. Berlin, Springer-Verlag.

LEGENDRE, P. [1993], Spatial autocorrelation: trouble or new paradigm? *Ecology*, **74**, pp. 1659-1673.

LEGENDRE, P. & M.J. ANDERSON [1999], Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments, *Ecological Monographs*, **69**, pp. 1-24.

LEGENDRE, P. & D. BORCARD [1994], Rejoinder, *Environmental and Ecological Statistics*, **1**, pp. 57-61.

LEGENDRE, P. & E. GALLAGHER [2001], Ecologically meaningful transformations for ordination of species data, *Oecologia*, **129**, pp. 271-280.

LEGENDRE, P. & L. LEGENDRE [1998], *Numerical ecology, 2nd English edition.* Amsterdam, Elsevier Science BV.

LEGENDRE, P., S.F. THRUSH, V.J. CUMMINGS, P.K. DAYTON, J. GRANT, J.E. HEWITT, A.H. HINES, B.H. MCARDLE, R.D. PRIDMORE, D.C. SCHNEIDER, S.J. TURNER, R.B. WHITLATCH & M.R. WILKINSON [1997], Spatial structure of bivalves in a sandflat: scale and generating processes, *Journal of Experimental Marine Biology and Ecology*, **216**, pp. 99-128.

MÉOT, A., P. LEGENDRE & D. BORCARD [1998], Partialling out the spatial component of ecological variation: questions and propositions in the linear modeling framework, *Environmental and Ecological Statistics*, **5**, pp. 1-27.

RAO, C. R. [1964], The use and interpretation of principal component analysis in applied research, *Sankhyaá, Ser. A*, **26**, pp. 329-358.

RAO, C. R. [1973], *Linear statistical inference and its applications. 2nd edition.* New York, Wiley.

STUDENT [W.S. GOSSET] [1914], The elimination of spurious correlation due to position in time or space, *Biometrika*, **10**, pp. 179-180.

TER BRAAK, C.J.F. [1986], Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis, *Ecology*, **67**, pp. 1167-1179.

TER BRAAK, C.J.F. [1987], The analysis of vegetation-environment relationships by canonical correspondence analysis, *Vegetatio*, **69**, pp. 69-77.

TUOMISTO, H. & A.D. POULSEN [2000], Pteridophyte diversity and species composition in four Amazonian rain forests, *Journal of Vegetation Science*, **11**, pp. 383-396.

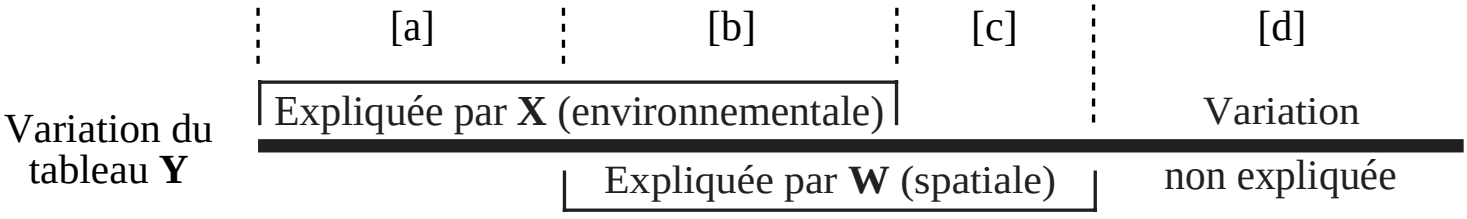

Figure 19.1 Partitionnement de la variation d'un tableau-réponse **Y** entre une matrice de variables explicatives environnementales (**X**) et un tableau de variables spatiales (**W**). Le trait horizontal représente la variation de **Y**. Figure adaptée de Borcard et al. [1992] et Legendre [1993].

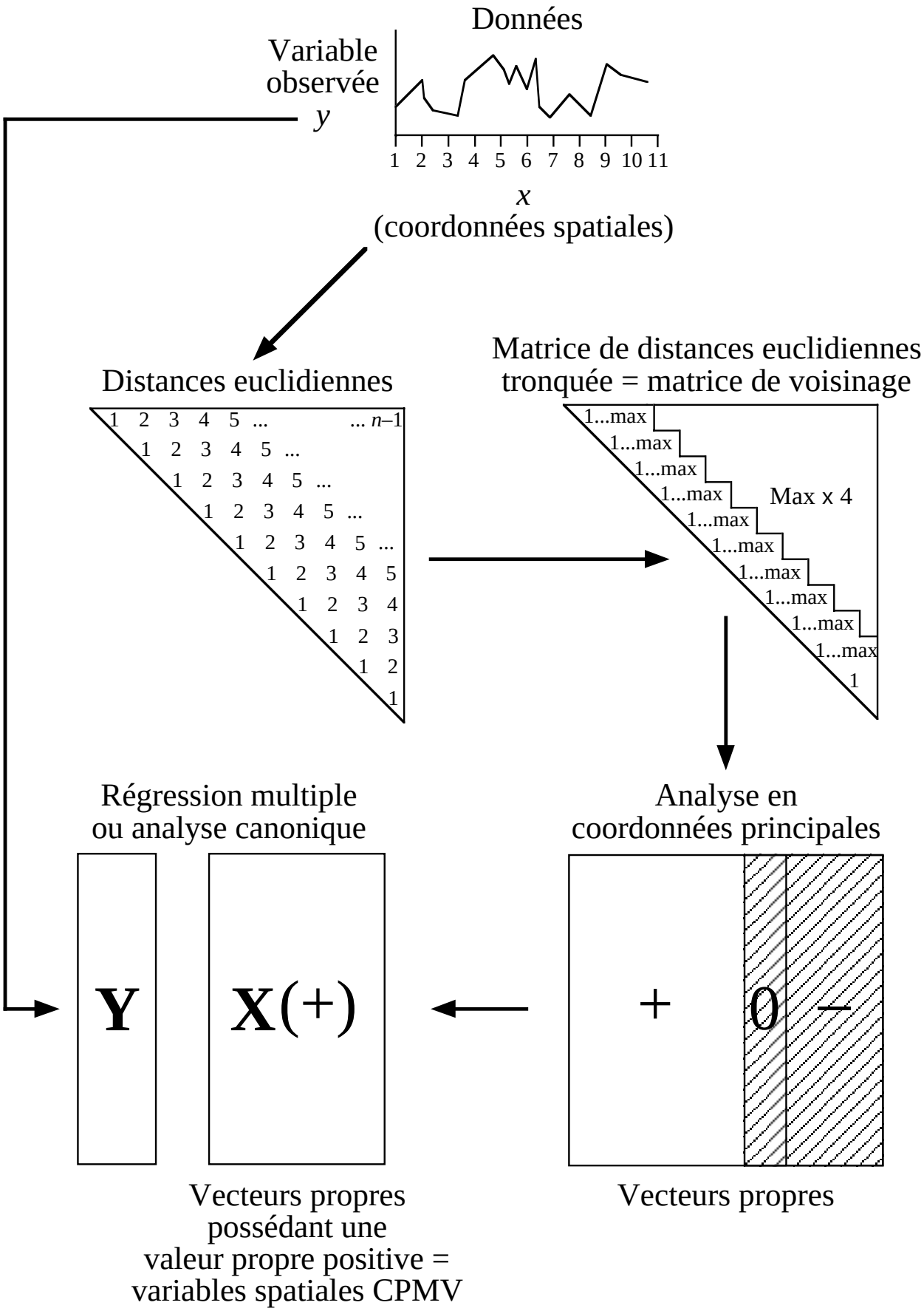

Figure 19.2 Principe de l'analyse spatiale CPMV, basée sur les coordonnées principales d'une matrice de voisinage (matrice tronquée de distances euclidiennes entre les sites).

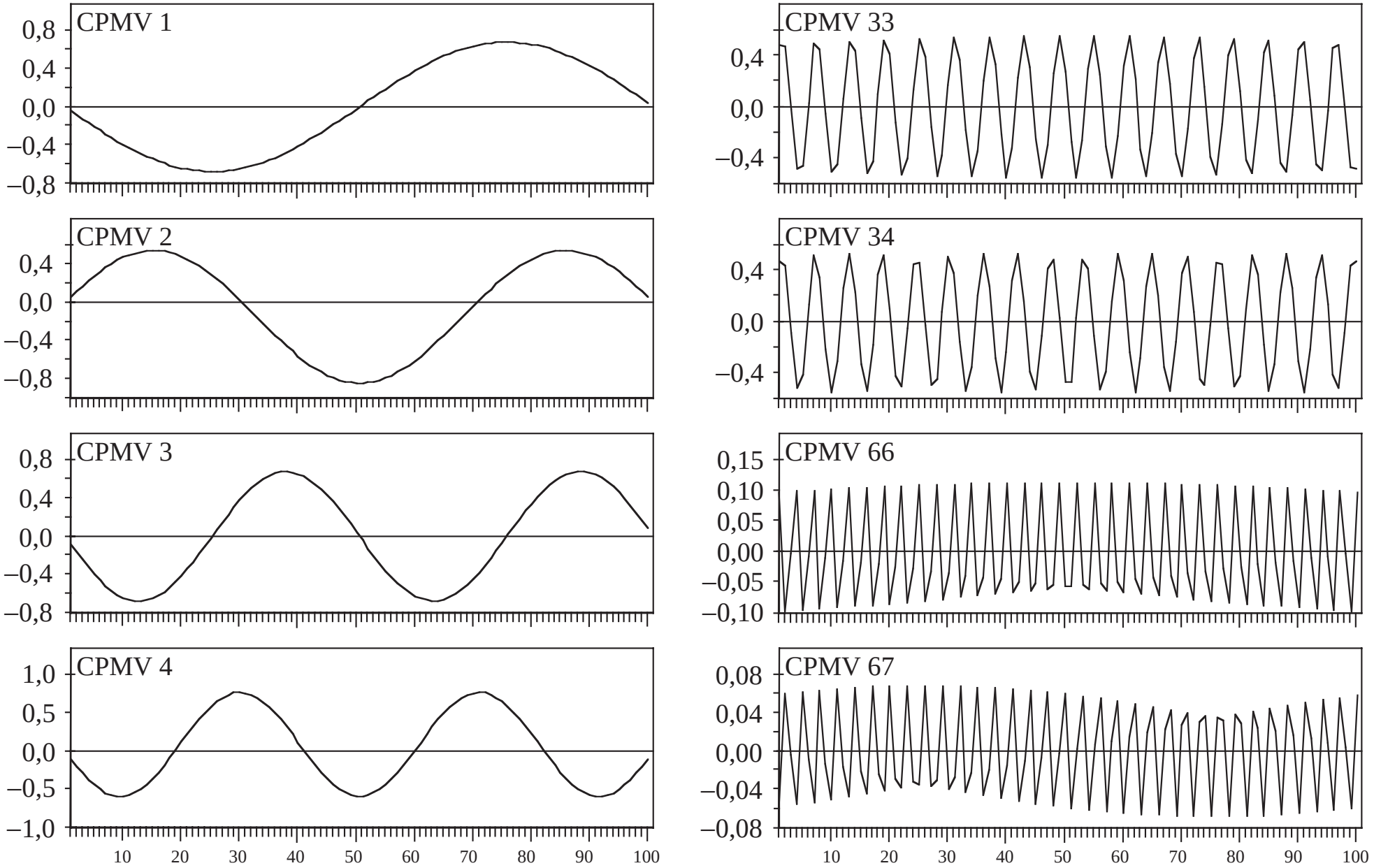

Figure 19.3 Huit des 67 variables CPMV obtenus pour 100 sites équidistants le long d'un transect. La matrice de distances euclidiennes a été tronquée à 1 pas (Max = 1, Figure 19.2). L'abscisse représente la position des points le long du transect.

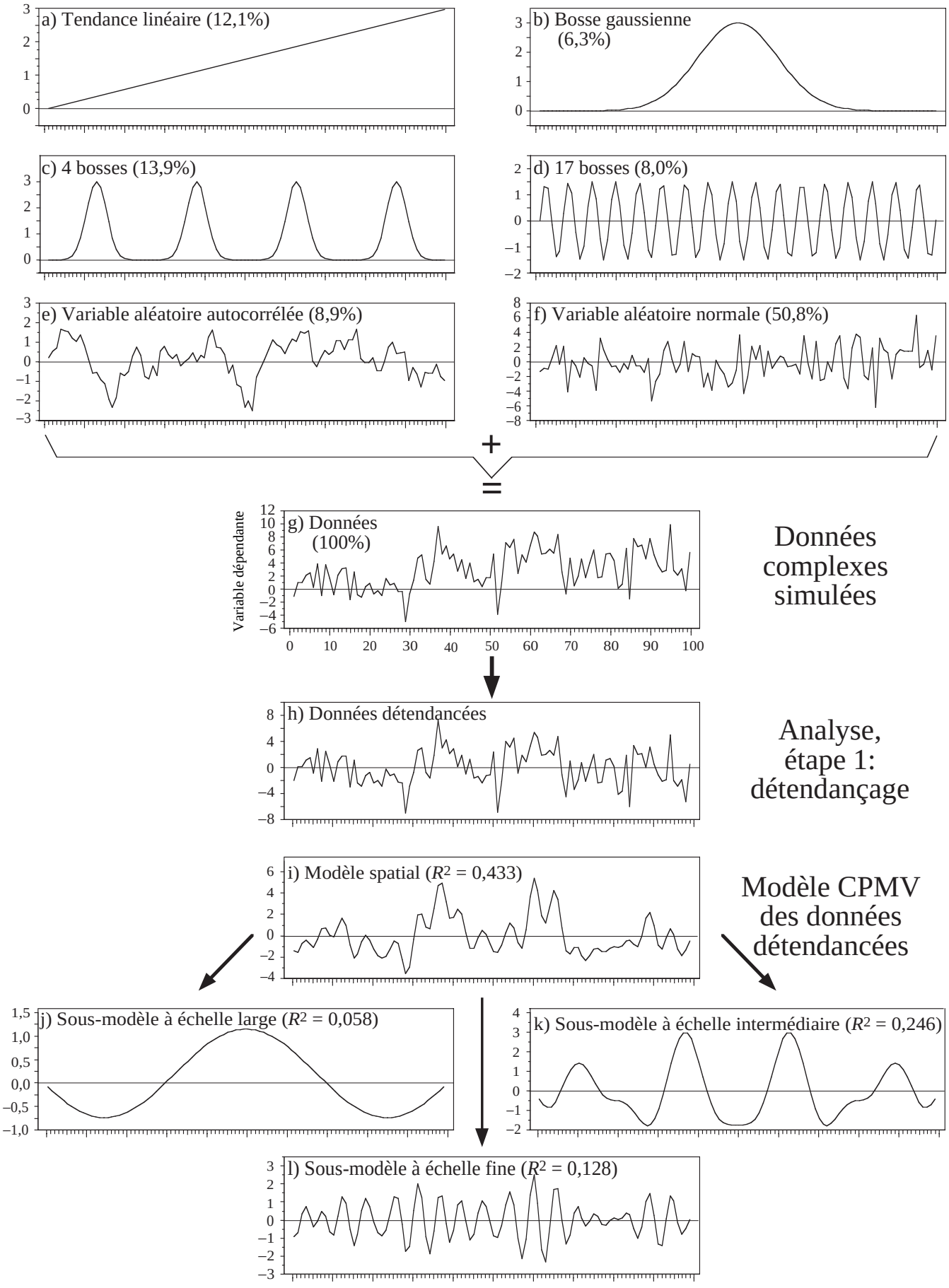

Figure 19.4 Construction de données complexes artificielles, et modèles obtenus par analyse CPMV. Les coefficients R^2 des modèles sont calculés par rapport aux données détendancées. Les sous-modèles à trois échelles spatiales sont orthogonaux entre eux et additifs.

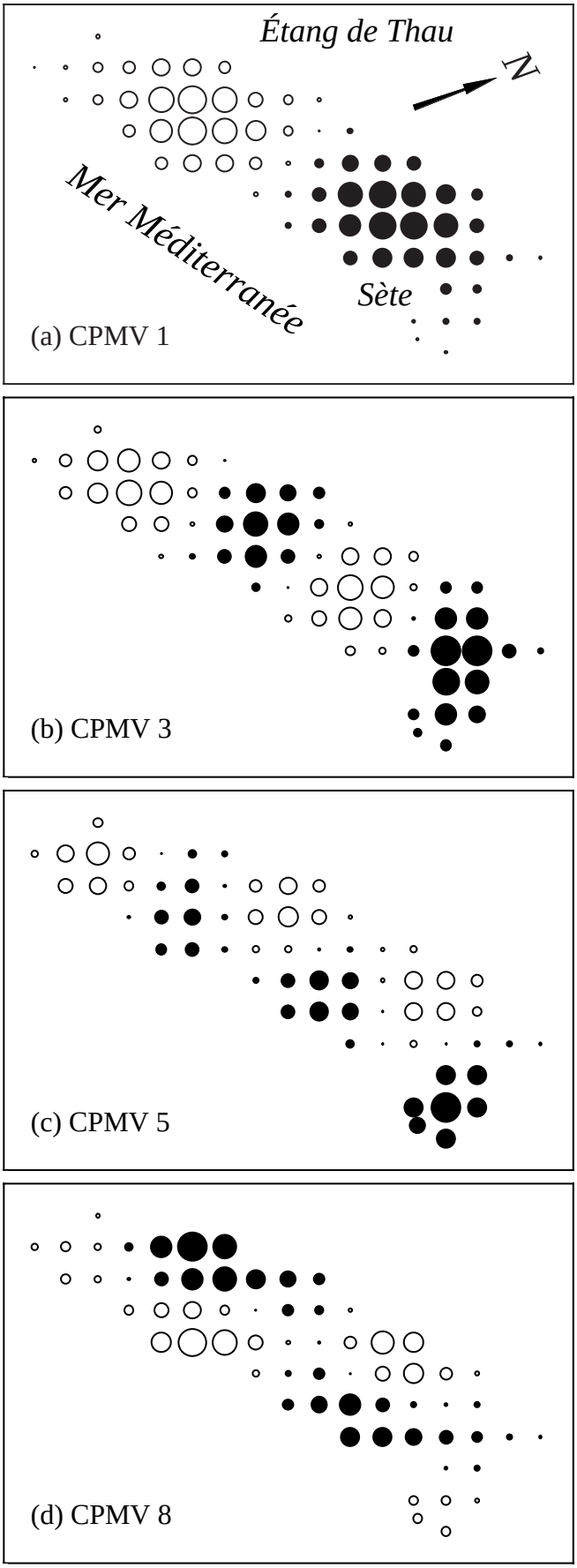

Figure 19.5 Analyse CPMV de 63 sites d'échantillonnage de l'étang de Thau. (a-d) Carte des variables CPMV 1, 3, 5 et 8 qui formeront le modèle spatial de Chl *a* à échelle large. Les bulles pleines correspondent aux valeurs positives de la variable ; les bulles vides représentent les valeurs négatives. (e) Carte des valeurs observées de la variable Chl *a* (centrées sur 0). (f) Modèle spatial de Chl *a* à échelle large. (g) Modèle spatial de Chl *a* à échelle intermédiaire. (h) Modèle spatial de Chl *a* à échelle fine. (f-h) Les valeurs ajustées sont centrées sur 0.

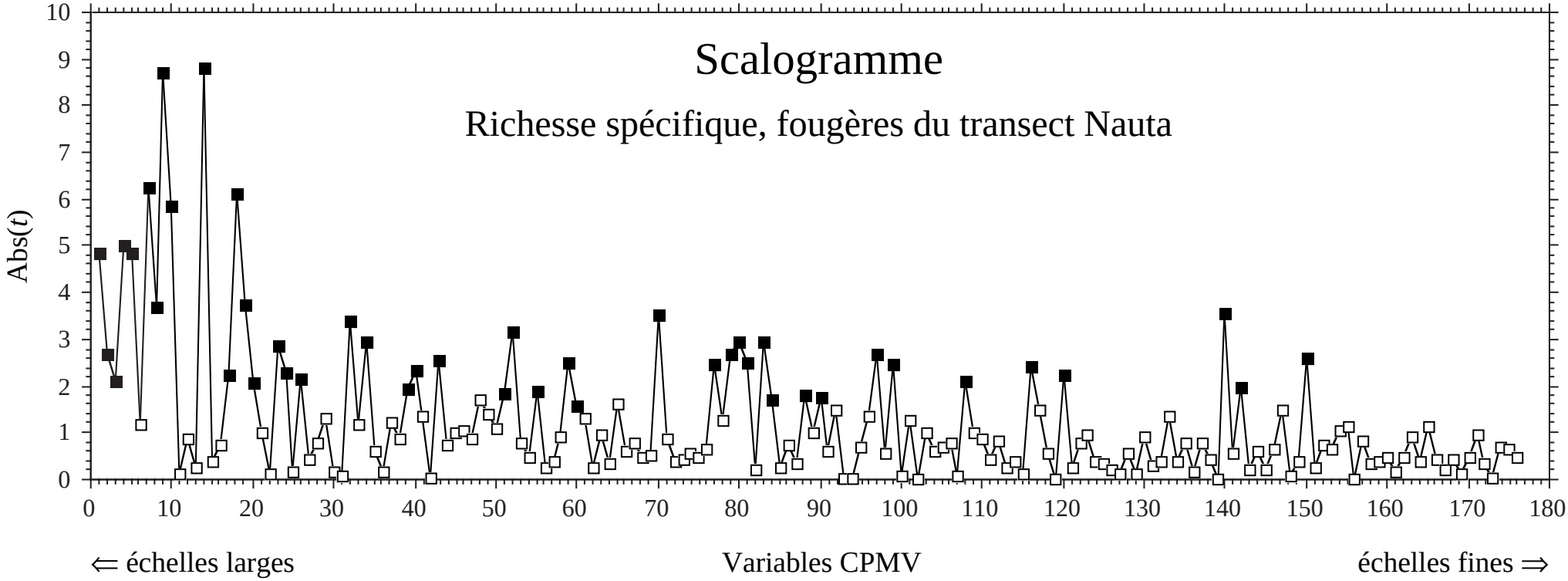

Figure 19.6 Scalogramme de la richesse spécifique des fougères dans le transect Nauta en Amazonie péruvienne. Les statistiques *t* significatives au seuil $\alpha = 0{,}05$ sont représentées par des carrés noir (tests permutationnels, 999 permutations). Parmi les 176 variables CPMV, 44 sont significatives.

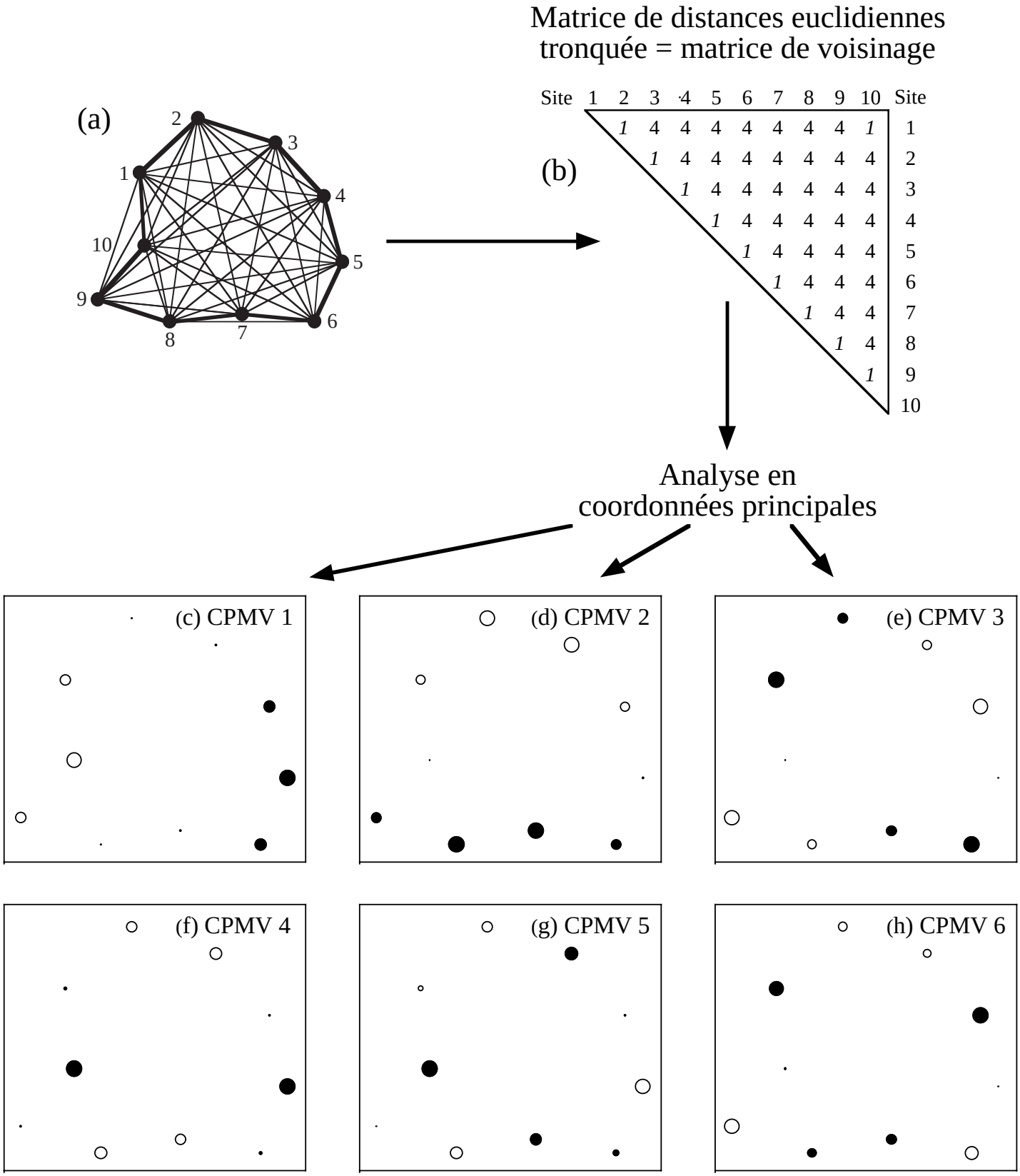

Figure 19.7 Modélisation CPMV autour d'une structure fictive en boucle. (a) Position des sites sur la carte. (b) Matrice de voisinage. Les distances entre sites voisins (traits gras en a) sont transcrites dans la matrice en b ; ces distances sont toutes égales à 1 dans l'exemple fictif. Notez la distance 1 entre les sites 1 et 10. Les distances entre sites non-adjacents (traits fins en a) sont remplacées par 4 fois la valeur maximale (max = 1 dans l'exemple, donc 4 x max = 4). (e-h) Les variables CPMV successives sont représentées par des bulles sur la carte des sites, comme à la Figure 19.5.

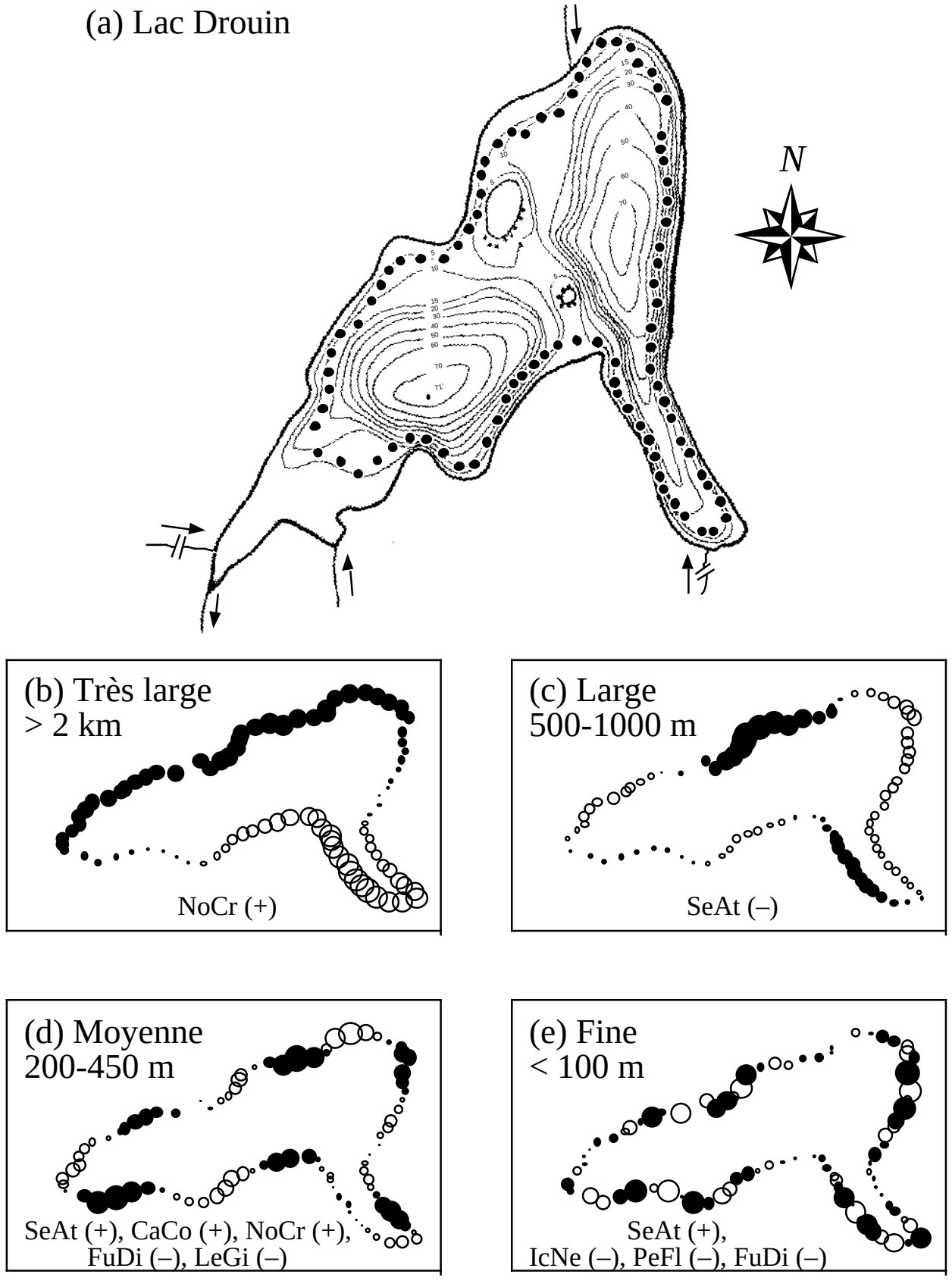

Figure 19.8 (a) Carte bathymétrique du lac Drouin ; les 90 sites d'échantillonnage sont représentés par des losanges. (b-e) Cartes des variables CPMV regroupées en quatre échelles, spécifiées dans les figures. Pour ces cartes, la projection cartésienne des coordonnées des sites a déformé le lac, l'étirant dans la direction est-ouest. Les valeurs des échelles (fonction linéaire de plusieurs CPMV) sont représentées par des bulles, comme à la Figure 19.5. Les espèces de poissons qui contribuent fortement à une échelle sont indiquées, sous la figure, avec le signe de leur contribution au vecteur propre ; les espèces sont ordonnées de la plus forte contribution positive à la plus forte contribution négative. Code des espèces : CaCo, *Catostomus commersoni* (meunier noir) ; FuDi, *Fundulus diaphanus* (fondule barré) ; IcNe, *Ictalurus nebulosus* (barbotte brune) ; LeGi, *Lepomis gibbosus* (crapet-soleil) ; NoCr, *Notemigonus crysoleucas* (chatte de l'est) ; PeFl, *Perca flavescens* (perchaude) ; SeAt, *Semotilus atromaculatus* (mulet à cornes).