

Diagramme de double projection – Guide

Pierre Legendre

Département de sciences biologiques, Université de Montréal,
C.P. 6128, succursale Centre-ville, Montréal, Québec H3C 3J7, Canada
E-mail: Pierre.Legendre@umontreal.ca

Mars 2000

Le progiciel R (version 4) permet de construire des diagrammes de double projection à partir, par exemple, des fichiers de sortie du programme CANOCO.

Diagramme de double projection

Un diagramme de double projection (“*biplot*” en anglais) est un diagramme d’ordination dans lequel on représente à la fois:

- les objets et les variables de la matrice de données \mathbf{Y} en ordination classique, soit en analyse en composantes principales (ACP) et en analyse factorielle des correspondances (AFC);
- les objets, les variables de la matrice réponse \mathbf{Y} et les variables de la matrice explicative \mathbf{X} en ordination canonique, soit en analyse canonique de redondance (ACR) et en analyse canonique des correspondances (ACC).

Deux types principaux de diagrammes sont utilisés en ACP et ACR:

- le diagramme préservant les distances entre objets (“*distance biplot*” en anglais) dans lequel les vecteurs propres sont normés à 1;
- le diagramme préservant les corrélations entre les variables (“*correlation biplot*” en anglais) dans lequel les vecteurs propres sont normés à $\sqrt{\lambda_j}$.

La théorie de ces normalisation est exposée à la section 9.1.4 de Legendre & Legendre (1998). En ordination canonique (ACR), les variables explicatives de la matrice \mathbf{X} sont représentées dans le diagramme en calculant d’abord leur corrélation linéaire avec chaque vecteur k donnant la position des sites dans l’espace canonique (“*fitted site scores*”), puis en multipliant ces corrélations par $\sqrt{\lambda_k} / (\text{variance totale de la matrice réponse } \mathbf{Y})$. Les variables explicatives binaires peuvent être représentées dans le diagramme par le centroïde des objets possédant la caractéristique en question. Pour plus de détails, voir la section 11.1.1 de Legendre & Legendre (1998).

Il y a également deux types principaux de diagrammes utilisés en AFC et en ACC:

- La mise à l'échelle (*scaling* en anglais) de type 1 est approprié si on s'intéresse d'abord à l'ordination des sites (objets, lignes du tableau de données). Dans l'espace factoriel total, les distances entre les sites sont les distances du khi-carré entre les lignes du tableau d'origine. Dans ce diagramme de double projection, les sites sont au barycentre (centre de masse) des espèces (colonnes du tableau de données).
- La mise à l'échelle de type 2 est appropriée si on s'intéresse d'abord aux relations entre les espèces. Dans l'espace factoriel total, les distances entre les espèces sont les distances du khi-carré entre les colonnes du tableau d'origine. Dans ce diagramme de double projection, les espèces sont au barycentre (centre de masse) des sites (lignes du tableau de données).

La théorie de ces normalisation est exposée à la section 9.4.1 de Legendre & Legendre (1998). En ordination canonique (ACC), les variables explicatives de la matrice \mathbf{X} sont représentées dans le diagramme en calculant leur corrélation linéaire **pondérée** avec chaque vecteur k donnant la position des sites dans l'espace canonique ("*fitted site scores*"). Les pondérateurs sont les sommes des lignes du tableau de fréquences (matrice \mathbf{Y}). Comme en ACR, les variables explicatives binaires peuvent être représentées dans le diagramme par le centroïde des objets possédant la caractéristique en question.

Exemple

À titre d'exemple, considérons les résultats présentés au tableau 11.5 de Legendre & Legendre (1998). Il s'agit d'une ACC avec mise à l'échelle de type 2. L'analyse porte sur 10 sites, 9 espèces et 4 variables environnementales. La première variable environnementale est quantitative ("Depth") alors que les trois autres sont qualitatives ("Coral", "Sand", "Other substrate"). Les blocs pertinents du fichier de sortie CANOCO.SOL (présent dans le dossier "Biplot data" sous le nom "CANOCO.SOL/CCA, SC=2") se présentent comme suit:

```
Coral reef fish: Numerical Ecology, Table 11.3
CCA Canonical axes: 3 Covariables: 0 Scaling: 2
```

```
No transformation
Spec: Species scores
```

N	NAME	AX1	AX2	AX3	AX4	WEIGHT	N2
	EIG	.3661	.1869	.0788	.0823		
1	Sp.1	.1104	-.2824	-.2030	.0019	60.00	6.82
2	Sp.2	.1414	-.3035	.3954	.1413	50.00	6.72
3	Sp.3	-1.0155	-.0958	-.1983	.1048	40.00	2.87
4	Sp.4	-1.0362	-.1096	.2210	-.2236	30.00	2.83
5	Sp.5	1.0537	-.5372	-.4381	-.2235	20.00	3.33
6	Sp.6	.9986	-.5740	.6799	.3900	10.00	3.33
7	Sp.7	.2552	.1782	-.2041	-.4334	45.00	7.01
8	Sp.8	.1466	.8574	-.0152	-.0528	35.00	7.52
9	Sp.9	.4137	.7079	.2157	.6903	25.00	7.18

Coral reef fish: Numerical Ecology, Table 11.3

CCA Canonical axes: 3 Covariables: 0 Scaling: 2

No transformation

Samp: Sample scores

N	NAME	AX1	AX2	AX3	AX4	WEIGHT	N2
	EIG	.3661	.1869	.0788	.0823		
1	Site 1	.7106	3.0817	.2197	1.2453	11.00	3.27
2	Site 2	.5848	3.0067	-.9474	-2.6997	12.00	2.32
3	Site 3	.7627	3.1526	2.1392	3.1163	6.00	2.57
4	Site 4	1.1123	-1.0715	-1.8753	-.6664	32.00	4.23
5	Site 5	-.9791	.0603	-.6963	.6126	54.00	5.21
6	Site 6	1.0432	-.4594	-.6398	-.2872	38.00	5.27
7	Site 7	-.9545	.0847	.1325	.4214	52.00	5.93
8	Site 8	.9473	.1084	.5261	.0056	38.00	6.39
9	Site 9	-1.1481	-.4904	.4784	-1.1702	47.00	5.03
10	Site 10	1.0329	-1.0351	2.7469	1.2808	25.00	4.03

Coral reef fish: Numerical Ecology, Table 11.3

CCA Canonical axes: 3 Covariables: 0 Scaling: 2

No transformation

BipE: Biplot scores of environmental variables

N	NAME	AX1	AX2	AX3	AX4
	R(SPEC,ENV)	.9985	.9401	.8832	.0000
1	Depth	-.1864	-.6403	.7452	.0000
2	Coral	-.9938	-.0977	-.0522	.0000
3	Sand	.2131	.9761	.0426	.0000
4	Other	.8809	-.4724	.0279	.0000

Coral reef fish: Numerical Ecology, Table 11.3

CCA Canonical axes: 3 Covariables: 0 Scaling: 2

No transformation

CenE: Centroids of environmental variables (mean.gt.0) in ordination diagram

N	NAME	AX1	AX2	AX3	AX4
	R(SPEC,ENV)	.9985	.9401	.8832	.0000
1	Depth	-.0670	-.2303	.2681	0.0000
2	Coral	-1.0227	-.1006	-.0538	0.0000
3	Sand	.6693	3.0653	.1339	0.0000
4	Other	1.0305	-.5527	.0327	0.0000

Coral reef fish: Numerical Ecology, Table 11.3
 CCA Canonical axes: 3 Covariables: 0 Scaling: 2

No transformation

SamE: Sample scores which are linear combinations of environmental variables

N	NAME	AX1	AX2	AX3	AX4	WEIGHT	% FIT
	EIG	.3661	.1869	.0788	.0823		
1	Site 1	.6921	3.0805	-.3287	.0000	11.00	97.01
2	Site 2	.6646	3.0621	.2302	.0000	12.00	92.96
3	Site 3	.6370	3.0438	.7892	.0000	6.00	92.59
4	Site 4	1.1089	-.5004	-1.5561	.0000	32.00	89.87
5	Site 5	-.9700	-.0655	-1.1206	.0000	54.00	90.99
6	Site 6	1.0537	-.5372	-.4381	.0000	38.00	84.08
7	Site 7	-1.0252	-.1023	-.0026	.0000	52.00	97.15
8	Site 8	.9986	-.5740	.6799	.0000	38.00	71.64
9	Site 9	-1.0803	-.1390	1.1154	.0000	47.00	94.33
10	Site 10	.9434	-.6107	1.7979	.0000	25.00	91.84

Notons au passage que tous les signes correspondant à l'axe 1 sont inversés par rapport à ceux de l'exemple publié au tableau 11.5 de Legendre & Legendre (1998). Cela ne changera en rien l'interprétation du diagramme de double projection.

La première étape consiste à sélectionner les lignes qui feront partie du fichier soumis au progiciel R. On pourra utiliser un éditeur ASCII ou encore un programme tel EXCEL ou StatView. Pour cet exemple, nous conserverons:

- Les "Species scores" (premier tableau).
- Nous avons le choix d'utiliser, dans le diagramme, soit les "Sample scores", soit les "Sample scores which are linear combinations of environmental variables". Nous utiliserons ici les "Sample scores" de façon à obtenir le même diagramme qu'à la Fig. 11.5 de Legendre & Legendre (1998), à des inversions d'axes près. Le tableau des "Sample scores which are linear combinations of environmental variables" sera donc éliminé pour cet exemple.
- Dans le tableau des "Biplot scores of environmental variables", seule la première ligne sera conservée puisque "Depth" est la seule variable quantitative.
- Dans le tableau "Centroids of environmental variables", la ligne "Depth" sera éliminée puisqu'il ne s'agit pas d'une variable binaire.
- Seules les valeurs d'ordination sur les trois premiers axes, qui sont les axes canoniques, seront conservées.
- On s'assurera que les noms des sites, des espèces et des variables environnementales ne contiennent pas d'espaces (blancs) si on utilise des espaces comme séparateurs des colonnes. Voir la note sous le prochain tableau.

Le fichier simplifié, aussi présent dans le dossier “Biplot data” sous le nom “Fichier d'entrée pour R/1”, se présente maintenant comme suit:

Sp.1	.1104	-.2824	-.2030
Sp.2	.1414	-.3035	.3954
Sp.3	-1.0155	-.0958	-.1983
Sp.4	-1.0362	-.1096	.2210
Sp.5	1.0537	-.5372	-.4381
Sp.6	.9986	-.5740	.6799
Sp.7	.2552	.1782	-.2041
Sp.8	.1466	.8574	-.0152
Sp.9	.4137	.7079	.2157
Site.1	.7106	3.0817	.2197
Site.2	.5848	3.0067	-.9474
Site.3	.7627	3.1526	2.1392
Site.4	1.1123	-1.0715	-1.8753
Site.5	-.9791	.0603	-.6963
Site.6	1.0432	-.4594	-.6398
Site.7	-.9545	.0847	.1325
Site.8	.9473	.1084	.5261
Site.9	-1.1481	-.4904	.4784
Site.10	1.0329	-1.0351	2.7469
Depth	-.1864	-.6403	.7452
Coral	-1.0227	-.1006	-.0538
Sand	.6693	3.0653	.1339
Other	1.0305	-.5527	.0327

Dans le fichier qui sera soumis au Progiciel R v.4, les colonnes peuvent être séparées soit par les tabulateurs, soit par un ou plusieurs espaces. On doit cependant s'assurer d'utiliser un type uniforme de séparateur dans tout le fichier; en d'autres termes, il faut éviter d'employer des espaces *et* des tabulateurs comme séparateurs dans le même fichier.

Avant de soumettre ce fichier à R v.4, on peut lui ajouter différents éléments qui faciliteront le travail par la suite:

- Ajouter des identificateurs (noms) pour les colonnes. Ces noms ne doivent pas contenir d'espaces (blancs). On peut donner, ou non, un identificateur à la première colonne qui contient les noms des éléments-lignes. Si on le fait, il faudra le spécifier au moment de la lecture du fichier par R v.4.

- Ajouter une colonne spécifiant quel symbole sera utilisé pour représenter les objets (sites) dans le graphique. Voici les codes des 14 symboles disponibles:

0	pas de symbole
1	●
2	●
3	○
4	■
5	■
6	□
7	▲
8	▲
9	△
10	▼
11	▼
12	∨
13	+
14	×

La valeur -999 (ou toute autre valeur sélectionnée par l'utilisateur dans les **Préférences** du menu **Édition**) désigne l'absence d'information; cette valeur se traduira dans le graphique par l'absence de symbole. On obtient le même résultat en employant le code 0. Toute autre valeur produira le symbole 1 ou le symbole 14; voir le *User_Guide* de R v.4. Dans cet exemple, les sites seront représentés par des cercles noirs (symbole 1).

- Dans la même colonne, spécifier le type de symbole qui sera utilisé pour représenter le centroïde des variables environnementales binaires dans le graphique. Dans cet exemple, les variables environnementales binaires seront représentées par des triangles (code 9).
- Ajouter une colonne spécifiant quel type de flèche sera utilisé pour représenter les espèces dans le graphique. Voici les codes des 3 types de flèches disponibles:

0	pas de flèche
1	_____ flèche sans pointe
2	_____▶ pointe échançrée
3	_____▶ pointe régulière

Toute autre valeur produit une flèche de type 2, sauf le code défini comme représentant l'absence d'information, dans les **Préférences** du menu **Édition**, qui ne produit aucune flèche. Dans cet exemple, les espèces seront représentées par une flèche sans pointe (de type 1). Le code pour l'absence d'information (ce qui produit une absence de flèche) est -999.

- Dans la même colonne, spécifier quel type de flèche sera utilisé pour représenter les variables environnementales quantitatives dans le graphique. Dans cet exemple, la variable environnementale quantitative (“Depth”) sera représentée par une flèche échançrée (code 2).
- On peut spécifier la couleur des symboles ou des flèches en ajoutant une autre colonne au fichier de données. Les codes de couleurs sont les suivants:

1	Rouge
2	Vert
3	Bleu
4	Cyan
5	Magenta
6	Jaune
7	Noir

Les flèches sont noires si on ne spécifie pas de couleur particulière, comme dans le présent exemple. Les symboles auront la couleur spécifiée dans la liste des symboles ci-dessus. On peut aussi faire en sorte que tout le diagramme soit dessiné en noir: voir **Utilisation de la couleur** dans les **Diagrammes** au paragraphe *Production du diagramme* ci-dessous.

- On peut enfin ajouter une colonne spécifiant un critère de groupement. Si, par exemple, les sites du présent exemple appartenaient à trois groupes, cette colonne contiendrait, pour chaque site, la valeur 1, 2 ou 3. Cette variable pourrait être employée pour tracer des ellipses autour des membres des différents groupes ainsi spécifiés.

Le fichier prêt à être importé (voir “Fichier d'entrée pour R/2” dans le dossier “Biplot data”) dans R v.4 est le suivant:

Identif.	Axe_1	Axe_2	Axe_3	Symboles	Flèches
Sp.1	.1104	-.2824	-.2030	0	1
Sp.2	.1414	-.3035	.3954	0	1
Sp.3	-1.0155	-.0958	-.1983	0	1
Sp.4	-1.0362	-.1096	.2210	0	1
Sp.5	1.0537	-.5372	-.4381	0	1
Sp.6	.9986	-.5740	.6799	0	1
Sp.7	.2552	.1782	-.2041	0	1
Sp.8	.1466	.8574	-.0152	0	1
Sp.9	.4137	.7079	.2157	0	1
Site.1	.7106	3.0817	.2197	1	-999
Site.2	.5848	3.0067	-.9474	1	-999
Site.3	.7627	3.1526	2.1392	1	-999
Site.4	1.1123	-1.0715	-1.8753	1	-999
Site.5	-.9791	.0603	-.6963	1	-999
Site.6	1.0432	-.4594	-.6398	1	-999
Site.7	-.9545	.0847	.1325	1	-999
Site.8	.9473	.1084	.5261	1	-999
Site.9	-1.1481	-.4904	.4784	1	-999
Site.10	1.0329	-1.0351	2.7469	1	-999
Depth	-.1864	-.6403	.7452	0	2
Coral	-1.0227	-.1006	-.0538	9	-999
Sand	.6693	3.0653	.1339	9	-999
Other	1.0305	-.5527	.0327	9	-999

Production du diagramme

Régler les **Préférences** dans le menu **Édition** du Progiciel R v.4. Pour les diagrammes, deux points retiendront l'attention:

- Noter la valeur de l'absence d'information ("Valeurs manquantes"), ou changer celle-ci pour la valeur utilisée dans le tableau de données ci-dessus.
- **Utilisation de la couleur** dans les **Diagrammes**: si cette case est cochée, les axes du diagramme seront en bleu et les symboles pourront porter différentes couleurs, tel qu'indiqué ci-dessus. Sinon, les diagrammes produits seront en noir. Cela est préférable pour des diagrammes qui doivent être incorporés à une thèse ou à une publication.

Nous sommes maintenant prêts à soumettre le fichier au Progiciel R v.4. Lorsqu'on demande la lecture du fichier (*via* la commande **Importer** du menu **Fichier** ou encore en faisant glisser le fichier sur l'icône de R v.4), il faut spécifier les éléments suivants dans la boîte de dialogue:

- Il s'agit d'une matrice rectangulaire.
- Les données sont séparées par des espaces dans le présent exemple. Si les données sont séparées par des tabulateurs, il faut le spécifier.
- Le fichier contient des noms d'objets en première colonne.
- Le fichier contient des noms de variables en première ligne. Il faut ignorer la toute première donnée du fichier.

Le fichier importé se présente comme suit:

23 éléments/lignes (23 inclus), 5 variables/colonnes (5 inclus)					
Nombre d'objets: n/d (n/d)					
	Axe_1	Axe_2	Axe_3	Symboles	Flèches
Sp.1	0.1104	-0.2824	-0.2030	0.0000	1.0000
Sp.2	0.1414	-0.3035	0.3954	0.0000	1.0000
Sp.3	-1.0155	-0.0958	-0.1983	0.0000	1.0000
Sp.4	-1.0362	-0.1096	0.2210	0.0000	1.0000
Sp.5	1.0537	-0.5372	-0.4381	0.0000	1.0000
Sp.6	0.9986	-0.5740	0.6799	0.0000	1.0000
Sp.7	0.2552	0.1782	-0.2041	0.0000	1.0000
Sp.8	0.1466	0.8574	-0.0152	0.0000	1.0000
Sp.9	0.4137	0.7079	0.2157	0.0000	1.0000
Site.1	0.7106	3.0817	0.2197	1.0000	-999
Site.2	0.5848	3.0067	-0.9474	1.0000	-999
Site.3	0.7627	3.1526	2.1392	1.0000	-999
Site.4	1.1123	-1.0715	-1.8753	1.0000	-999
Site.5	-0.9791	0.0603	-0.6963	1.0000	-999
Site.6	1.0432	-0.4594	-0.6398	1.0000	-999
Site.7	-0.9545	0.0847	0.1325	1.0000	-999
Site.8	0.9473	0.1084	0.5261	1.0000	-999
Site.9	-1.1481	-0.4904	0.4784	1.0000	-999
Site.10	1.0329	-1.0351	2.7469	1.0000	-999
Depth	-0.1864	-0.6403	0.7452	0.0000	2.0000
Coral	-1.0227	-0.1006	-0.0538	9.0000	-999
Sand	0.6693	3.0653	0.1339	9.0000	-999
Other	1.0305	-0.5527	0.0327	9.0000	-999

Pour produire le diagramme, il faut demander un **Nouveau** fichier dans le menu **Fichier** de R v.4. Dans la boîte de dialogue, on spécifie qu'on désire produire un **Diagramme** de type **Dispersion X-Y**. Une fenêtre munie d'un grand X apparaît.

En cliquant deux fois dans cette fenêtre, une boîte de dialogue apparaît. Celle-ci permet de spécifier quelle colonne du fichier sera utilisée pour chaque élément du graphique. On peut aussi faire apparaître cette boîte de dialogue en choisissant **Caractéristiques** dans le menu **Édition**. Attention: s'il y a plusieurs fichiers ouverts, ils apparaissent tous, avec leurs colonnes, dans ce menu; il faut choisir les variables du fichier à tracer et non d'un autre. Pour l'exemple, on peut spécifier ce qui suit:

- Choisir la colonne "Axe_1" comme variable X (abscisse).
- Choisir la colonne "Axe_2" comme variable Y (ordonnée).
- Si désiré, on pourrait représenter l'axe canonique 3 par des bulles de différentes tailles (bulles noires en positif, bulles vides en négatif) en spécifiant choisissant "Axe_3" comme "Z bulle". Cela n'est cependant pas habituel dans un diagramme de double projection.
- Choisir la variable "Symboles" comme "Z symbole".
- Choisir la variable "Flèches" comme variable "Z flèches".
- Cliquer la case "Utiliser les noms d'objets de X comme labels" si on désire voir apparaître des identificateurs dans le graphique.
- Cliquer la case "Axes proportionnels" si on désire que l'échelle demeure la même sur les deux axes du graphique.

Le graphique apparaît dans la fenêtre lorsqu'on clique **OK**. On peut agrandir la fenêtre à volonté à l'aide de la case dans le coin inférieur droit de la fenêtre; le graphique est redessiné à la nouvelle taille, mais le texte est récrit à la même taille.

Dans cette première version du graphique, les points extrêmes (en haut, en bas, à gauche et à droite) sont situés sur les axes. Cliquer deux fois sur le graphique pour remédier à cette situation: la boîte de dialogue réapparaît. Cela permet maintenant de fixer les bornes des axes en cliquant les différentes cases de l'encadré intitulé **Fixer les limites** et en spécifiant des valeurs. On choisira les limites de façon à obtenir des valeurs simples sur les axes et à laisser une marge entre les points extrêmes et la bordure du graphique. Le graphique peut être modifié de deux autres façons:

- On peut inverser la position des objets et des variables sur l'un ou l'autre des axes du graphique lorsque cela peut en faciliter la lecture et la compréhension. Pour ce faire, il suffit de changer tous les signes de la colonne correspondante dans le fichier de données (colonne "Axe_1" ou "Axe_2").

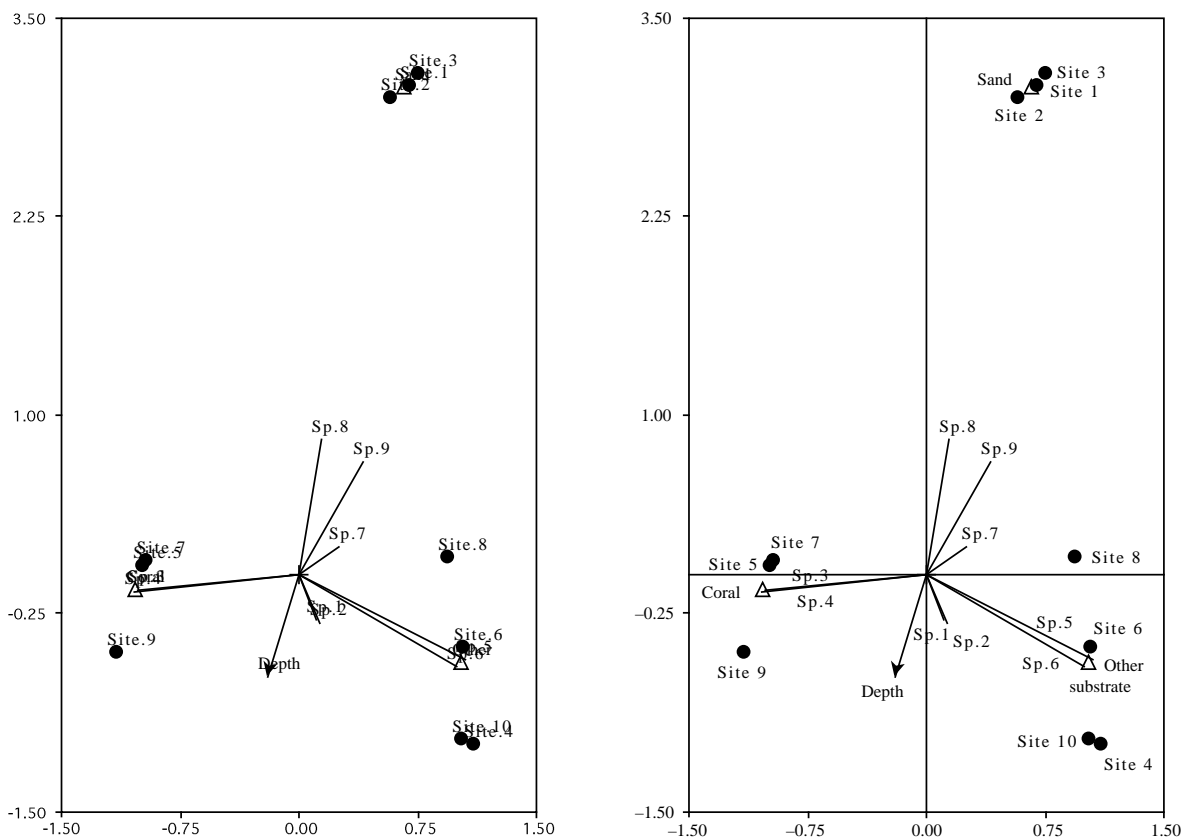


Figure 1 Diagramme de double projection. À gauche, le diagramme produit par le Progiciel R v.4. À droite, le diagramme retouché à l'aide d'un programme de dessin.

- Si les flèches représentant les espèces (ou les variables environnementales) sont trop courtes par rapport à la dispersion des sites, on peut en augmenter la longueur pour faciliter l'interprétation des relations entre ces variables en multipliant les deux coordonnées des espèces (ou des variables environnementales) par une constante, e.g. 2, 5 ou 10.

Le diagramme de double projection produit par R v.4 est représenté dans la portion de gauche de la figure 1. On peut sauver ce graphique en format PICT à l'aide de la commande **Enregistrer sous...** du menu **Fichier**. Ce fichier peut être importé dans un programme de dessin pour finition. On pourra alors déplacer les noms d'objets afin d'éviter les superpositions. On pourra également, si on le désire, modifier la police de caractère utilisée dans le graphique ou procéder à toute autre altération. Une version retouchée du même graphique est présentée dans la portion de droite de la figure.

Référence

Legendre, P. & L. Legendre. 1998. *Numerical ecology. 2nd English edition*. Elsevier Science BV, Amsterdam. xv + 853 pp.