

Régression linéaire

Résumé

La fiche décrit deux méthodes de régression linéaire (module LinearReg) dans le cadre d'un problème de prédiction d'une variable biologique par des variables d'environnement, problème posé dans un article récent de P. Baran & Coll. (1993 *Bull. Fr. Pêche Piscic* : 331, 321-340). On accorde une certaine importance à l'examen des variables initiales et à la définition de l'objectif visé. On aborde la régression multiple classique, les difficultés qu'elle soulève, et la solution proposée par la régression PLS, ou régression partiellement aux moindres carrés. Inventée en chimométrie, dont elle est un standard méthodologique la régression PLS gagne à être connue en écologie. L'algorithme utilisé est décrit par Ter Braak & Juggins (1993, *Hydrobiologia* : 269/270: 485-502, p. 487).

Plan

1 — Le problème : Habitat et abondance de la truite commune	2
2 — Liaisons entre variables à prédire.....	3
2.1 — Changement de variable préliminaire	3
2.2 — L'automodélisation par ACP normée	7
2.3 — Régression et projection : approche élémentaire	10
3. — Liaisons entre variables explicatives	12
4 — MLR : la régression linéaire multiple.....	16
5 — Sélection de variables en régression linéaire	19
6 — Régression PLS.....	24
Références	26

D. Chessel et J. Thioulouse

1 — Le problème : Habitat et abondance de la truite commune

Un article récent de P. Baran & Coll.¹ pose avec une précision incontestable la notion de variables instrumentales du point de vue de l'expérimentateur. Le résumé est explicite :

Les relations entre les caractéristiques de l'habitat et les biomasses et densités de truites communes (*Salmo trutta* L.) ont été recherchées dans 33 stations de la rivière Neste d'Aure et trois de ses affluents: la Neste du Louron, la Neste du Rioumajou et le ruisseau d'Espiaube dans le département des Hautes-Pyrénées. L'étude a été conduite sur un cycle annuel.

Dans un premier temps, la validité du modèle d'Indice de Qualité d'Habitat (HQI) (BINNS et EISERMAN, 1979), basé sur 10 variables de l'habitat, a été testée. Les biomasses théoriques prévues par le modèle ne sont pas linéairement corrélées aux biomasses observées par pêche électrique. Le meilleur ajustement linéaire est obtenu grâce à des transformations par les logarithmes. Toutefois, la pente de la droite de régression est significativement différente de 1 ($t = 2.53$ ($p < 0.01$)). Le modèle de l'Indice de Qualité d'Habitat ne constitue pas, dans le cas de la vallée d'Aure, un outil satisfaisant de prévision des biomasses de truites.

Dans un deuxième temps, l'influence de chaque variable de l'habitat a été testée individuellement. Les biomasses observées sont significativement corrélées à l'altitude (entre 1350 et 600 m), aux surfaces d'abris, à la température mensuelle maximale (pour une gamme allant de 10 à 16 °C), à la conductivité électrique, à la vitesse moyenne au fond, à la profondeur moyenne et au rapport largeur/profondeur. Les densités sont significativement corrélées aux mêmes variables, à l'exception de la profondeur moyenne; il faut également ajouter des corrélations significatives avec la pente de la ligne d'eau et la largeur de la rivière. L'étude par classe d'âge montre que l'abondance de la cohorte 0+ est liée à l'altitude, la température et la conductivité. La largeur moyenne constitue la seule caractéristique de l'habitat physique corrélée avec les biomasses et densités de 0+. L'étude par saison indique seulement une corrélation négative entre les densités et biomasses échantillonnées en hiver et la profondeur moyenne. En ce qui concerne la cohorte 1+, on observe des corrélations avec les mêmes variables altitude, température et conductivité auxquelles il faut ajouter la variable abris. Les densités de truites de taille supérieure à la taille légale de capture (180 mm) sont positivement corrélées à la surface d'abris, la profondeur moyenne, la température et la conductivité, et négativement avec l'altitude.

Dans une troisième étape, à partir de régressions multiples progressives, il a été possible d'établir un modèle statistique à 5 variables qui explique 86% de la variation de biomasse totale de truites. Ce type d'outil peut constituer un élément de gestion pour les populations de truites de la Vallée de la Neste d'Aure.

Les auteurs nous permettent de reproduire exactement le tableau de données publiées (op. cit. p. 327) dans le tableau 1. Il s'agit clairement d'une question de modèle prédictif de l'abondance des individus d'une espèce par les paramètres environnementaux. Les auteurs citent un article de 1988 qui propose *70 modèles permettant d'estimer l'abondance des salmonidés à partir des variables de l'habitat* (op. cit. p. 322). C'est donc une question qui intéresse les écologues pratiquant la statistique.

Il s'agit de variables **instrumentales** parce qu'on trouve deux ensembles de variables formés d'une part des variables explicatives ou prédictives, d'autre par des variables à prédire. Lorsqu'il n'y a qu'une variable à prédire et plusieurs explicatives (régression multiple) la situation est simple. Elle se complique ici, et le résumé cité le montre bien,

en ce sens qu'on peut multiplier les modèles indépendants pour chacune des variables à expliquer ou qu'au contraire on peut chercher des **modèles communs** à plusieurs variables.

Stations	Altitude	Temp.	Cond.	Pente	Larg.	Prof.	V. Fond	V. Surf	Abris	DensInv	Module	Debit E.
N1	1017	12	169	1.8	5.6	0.23	0.29	0.55	16	139	482	53
N2	1010	12.2	162	2.02	8.4	0.2	0.3	0.41	7	104	1281	23.5
N3	970	12.6	155	4	7	0.3	0.32	0.48	19	120	1350	25
N4	830	13	165	0.5	5	0.3	0.26	0.73	22	727	2575	26
N5	800	11	93	2.6	14.5	0.32	0.52	1.45	5	321	8205	36
E1	1100	10.5	183	3	3.3	0.24	0.4	1.01	30	102	400	36.4
E2	810	12.5	156	1.7	2.4	0.12	0.33	0.63	14	145	100	29
SG2	840	12.5	180	1.1	1.1	0.13	0.2	0.48	28	291	55	31
R1	1364	10.3	41	2.5	10.9	0.21	0.29	0.78	9	122	2180	37.7
R2	1120	11	85	7.5	10	0.24	0.21	0.37	32	160	620	28.5
R3	1070	11.1	95	5	6.6	0.29	0.26	0.46	14	160	720	28.5
R4	906	11.5	134	10.5	8	0.33	0.13	0.32	42	228	550	22
L1	1250	10	46	12	5.1	0.43	0.16	0.28	51	474	200	56
L2	1200	10.5	50	7.4	7.4	0.24	0.13	0.31	18	860	360	83
L3	1185	10.7	55	5.3	6.7	0.28	0.25	0.35	20	700	430	75
L4	1110	11	60	0.9	9.3	0.15	0.28	0.5	6	287	380	79
L5	986	13	95	2.2	6.1	0.25	0.25	0.3	22	694	1200	42
L6	980	13	95	0.8	8.5	0.18	0.27	0.4	6	694	1200	42
L7	965	10.5	78	1.8	9.6	0.27	0.39	1.5	2	450	3600	33
L8	900	13.5	108	1.2	10.6	0.34	0.4	0.6	14	373	4400	27.9
L9	895	13.5	109	0.1	7.3	0.2	0.17	0.3	8	373	770	40.3
L10	895	13.5	109	1.1	6.7	0.21	0.21	0.4	18	373	770	40.3
L11	860	13.5	110	1.7	6.9	0.17	0.24	0.48	7	400	1170	41.6
L12	860	13.5	110	0.5	5.6	0.43	0.09	0.39	60	400	1170	41.6
L13	847	13.5	110	1.2	10.1	0.17	0.28	0.48	4	350	1290	42.3
L14	847	13.5	110	0.8	6.2	0.45	0.11	0.36	45	350	1290	42.3
L15	820	13.7	115	1.7	6.8	0.3	0.19	0.44	35	350	1450	42.8
L16	730	14	118	0.4	7.7	0.25	0.19	0.54	22	182	1700	43.1
L17	710	13.5	96	1	12.3	0.32	0.33	0.38	11	350	6100	32.7
NB1	685	14.5	128	1.2	9.8	0.32	0.37	0.47	41	509	2000	13.3
NB2	637	15.5	132	0.5	10	0.37	0.28	0.4	19	480	3400	20
NB3	630	14.5	127	0.9	19	0.33	0.33	0.77	10	299	6500	52.7
RU1	1250	10	145	3.7	2.3	0.2	0.33	0.43	19	500	250	20

Tableau 1 : Données de P. Baran & Coll. (1993). Première partie : Variables mésologiques.

Quand des variables explicatives sont destinées à modéliser plusieurs variables à expliquer (en particulier quand celles-ci sont liées entre elles) les variables explicatives sont appelées instrumentales. On consultera l'article cité pour la définition de ces variables instrumentales (1-Altitude, 2-Température, 3-conductivité, 4-Pente, 5- Largeur, 6-Profondeur, 7-Vitesse au fond, 8-Vitesse en surface, 9-Abris, 10-densité d'invertébrés, 11-module et 12-Débit d'étiage). Les variables à prédire sont au nombre de 7 (1-Biomasse totale, 2-densité totale, 3-densité pêchable, 4- densité des 0+, 5- biomasse des 0+, 6-densité des 1+, 7-biomasse des 1+).

2 — Liaisons entre variables à prédire

2.1 — Changement de variable préliminaire

Toutes les variables à expliquer ont une distribution de fréquence dissymétrique et non gaussienne, du même type que la première (Figure 1). On sait que de telles distributions conviennent mal à la mesure de la corrélation linéaire et qu'un changement de variable qui normalise les données a souvent pour fonction de linéariser les relations. Comme indiqué dans l'article cité toutes les mesures biologiques sont transformées par $y = \text{Log}(x + 1)$.

Stations	BiomTot	DensTot	DensCapt	Dens0+	Biom0+	Dens1+	Biom1+
N1	816	0.25	0.02	0.03	13.1	0.06	103
N2	444	0.12	0.01	0.02	19.2	0.08	266
N3	444	0.07	0.02	0	5.1	0.03	97
N4	1690	0.37	0.06	0.01	4.4	0.16	404
N5	890	0.14	0.06	0.02	14.9	0.05	163
E1	505	0.13	0.03	0.03	12.8	0.05	130
E2	2046	0.48	0.08	0.07	26.1	0.15	307
SG2	3242	1.02	0.13	0.26	84	0.3	573
R1	183	0.06	0.01	0.01	2.1	0.01	12
R2	781	0.17	0.06	0	9.5	0.1	205
R3	591	0.15	0.03	0.01	16.1	0.06	154
R4	901	0.16	0.06	0.02	11.7	0.06	160
L1	907	0.23	0.04	0.02	12.1	0.1	217
L2	1110	0.17	0.07	0.02	49.2	0.05	95
L3	1014	0.14	0.06	0.02	48.8	0.04	76
L4	408	0.13	0.02	0.02	11.1	0.06	110
L5	720	0.13	0.04	0.04	45.3	0.04	155
L6	368	0.12	0.05	0.02	37.6	0.03	75
L7	290	0.05	0.02	0.01	9.5	0.02	39
L8	984	0.12	0.07	0.02	31.3	0.02	102
L9	984	0.23	0.05	0.18	32	0.08	199
L10	1187	0.23	0.07	0.07	23	0.06	206
L11	558	0.13	0.02	0.02	21.7	0.02	64
L12	2504	1.26	0.66	0.02	19.3	0.06	164
L13	585	0.14	0.02	0.04	32.8	0.04	118
L14	3125	0.45	0.17	0.06	46.2	0.05	182
L15	2944	0.24	0.14	0.06	39.3	0.1	387
L16	2050	0.39	0.12	0.07	60.9	0.11	339
L17	1345	0.22	0.09	0.07	52.6	0.05	885
NB1	2216	0.33	0.18	0.02	27.7	0.26	1437
NB2	1737	0.24	0.08	0.01	10.8	0.19	1144
NB3	1526	0.22	0.15	0.01	30.9	0.17	1005
RU1	713	0.23	0.02	0.09	68	0.05	135

Tableau 1 : Données de P. Baran & Coll. (1993). Deuxième partie : Variables biologiques.

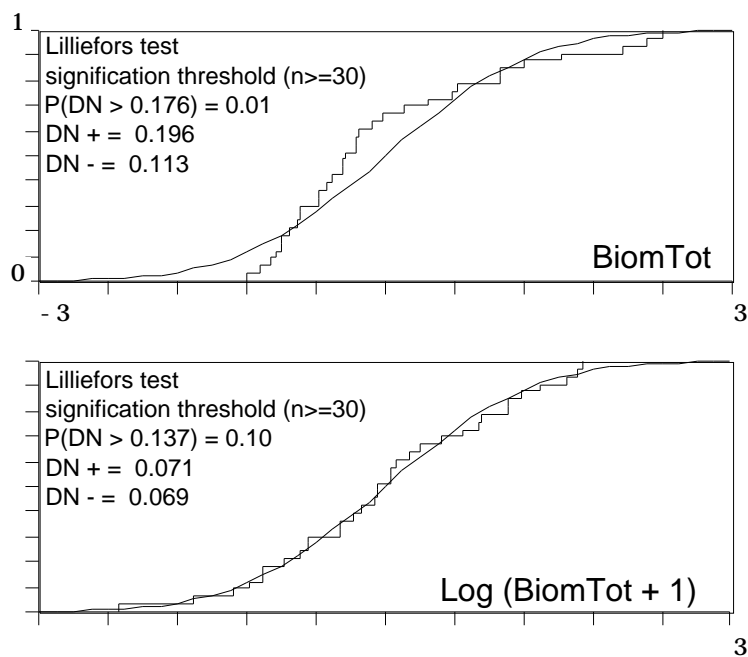


Figure 1 : Fonction de répartition théorique (courbe gaussienne) et empirique (fonction en escalier). Test de normalité de Lilliefors². Rejet de la normalité au seuil de 1% pour la variable brute (en haut) et normalité acceptable pour la variable transformée (en bas).

Tout ce qui suit porte sur les variables transformées. On remarquera que la transformation rend les distributions gaussiennes pour les mesures de biomasse mais pas pour les mesures de densité. La station L12 qui présente une densité de 1.26 truites au m², soit 25 fois la plus petite valeur, peut être considérée comme un point étranger (outlier). Les mesures de densité et de biomasse n'ont pas le même statut. La présence de ces très forte valeurs relatives rend délicate la mesure des relations inter-variables. Par exemple, la figure 2 montre l'instabilité de la relation estimée si on enlève les stations SG2 et L9 qui présentent des très fortes valeurs de Dens0+. Si les variables sont dépendantes certainement, la liaison n'est clairement linéaire ni avant ni après le changement de variable.

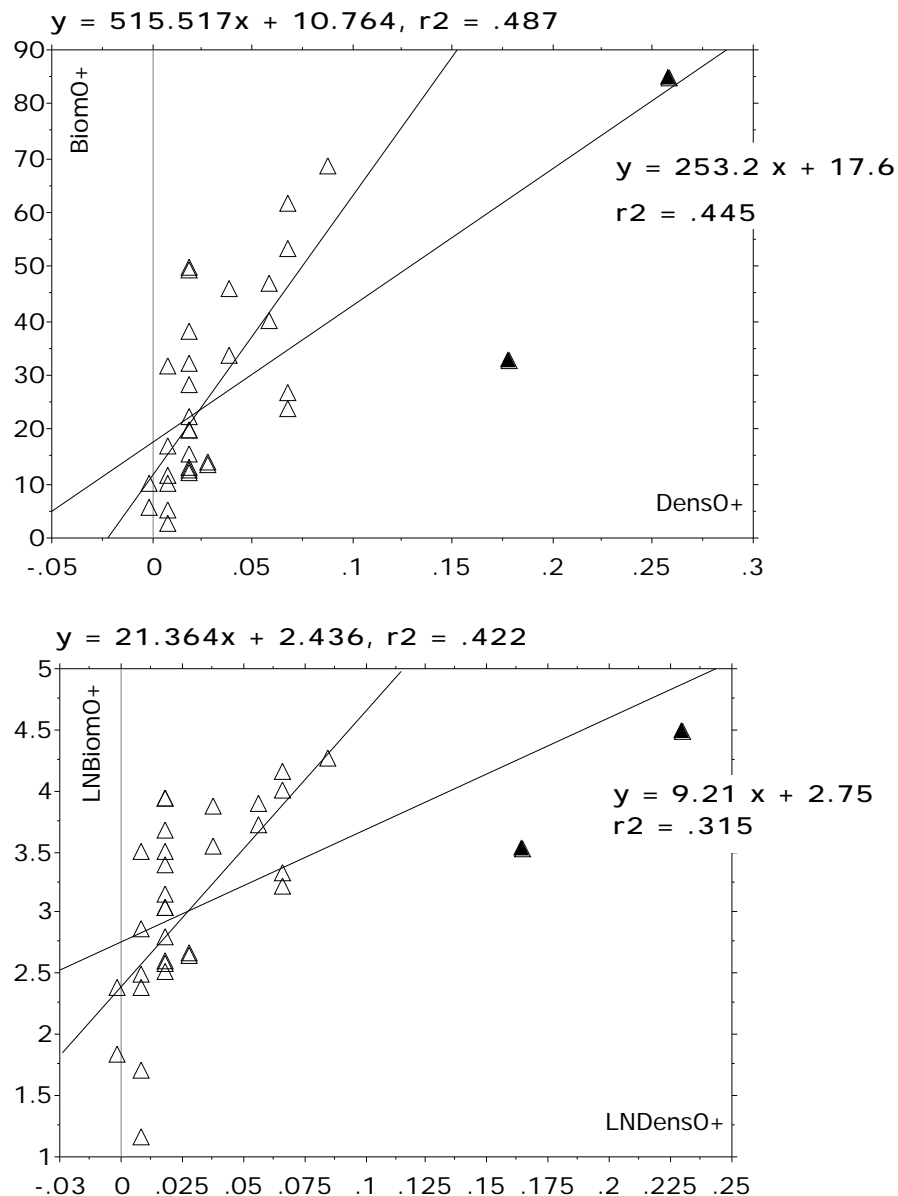


Figure 2 : Influence de deux station sur la corrélation densité-biomasse (individus 0+). (Logiciel StatViews™ SE+Graphics). Outre le rôle des points extrêmes, on notera la relation non linéaire.

Le passage en Log (x + 1) normalise la variable biomasse mais non la variable densité. La liaison n'est pas linéarisée (figure 3). Mais c'est par la présence de valeurs extrêmes, que le passage en Log n'a pas suffisamment transformées, qui pose problème. Si les liaisons sont non linéaires, les méthodes linéaires sont-elles non appropriées ? C'est une idée répandue qui confond principe de fonctionnement et résultats obtenus.

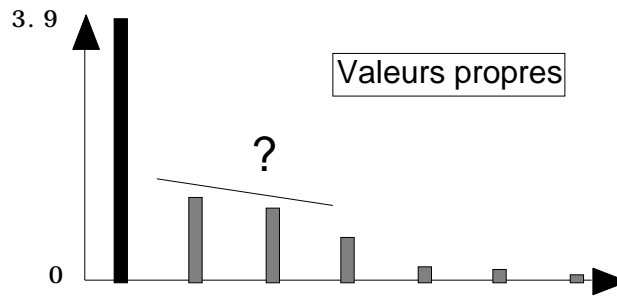


Figure 5 : Valeurs propres de l'ACP normée des variables biologiques.

2.2 — L'automodélisation par ACP normée

Il est clair *a priori* que les variables biologiques sont corrélées. La matrice des corrélations est dans la figure 4. Une des propriétés fondamentales des analyses linéaires est la formule de reconstitution des données (Théorème d'Eckart-Young³). Il s'agit de définir une variable artificielle qui est capable de prédire au mieux le plus grand nombre de variables observées.

L'ACP normée du tableau des variables biologiques (TLog, 33 lignes-stations, 7 colonnes-variables) donne une première valeur propre (figure 5) de 3.82 (55% d'inertie). Cela signifie qu'il existe une variable y (première composante principale) dont la somme des carrés des corrélations avec les variables de départ vaut 3.82, soit une corrélation moyenne de 0.74.

Cela est somme toute étonnant. Il existe une variable artificielle qui présente une corrélation avec chacune des variables qui dépasse la quasi totalité de toutes les corrélations bivariées. Ceci s'exprime dans la figure 6.

Cette variable prend en compte toutes les mesures d'abondance. Parmi celles-ci, la biomasse totale est la plus représentative. Si on se pose la question "que doit-on prédire avec les variables environnementales ?" la réponse, implicite dans l'article cité, et explicite après cette première approche, est double. D'une part il convient d'expliquer d'une part une variable abondance globale, d'autre part une structure descriptive des composantes de cette abondance.

Il y a une bonne partie de la variabilité du tableau considérée qui relève de la **redondance** pure (c'est le premier facteur ci-dessus) : si il y a une population abondante, en gros il y a plutôt plus de truites à la maille, plutôt plus de truites 0+ ou 1+ et les captures représentent des biomasses plutôt plus élevée.

La figure 7 donne la carte de cette abondance totale. On y voit directement l'un des problème majeur de l'écologie statistique. Dans le tableau, les stations pourraient passer pour une série d'échantillons indépendants (hypothèse sur laquelle repose toute l'analyse statistique inférentielle) alors que sur la carte l'autocorrélation spatiale invalide définitivement cette assertion. Ignorons, pour le moment la question.

Ensuite, clairement, il y a autour de cette redondance des variantes non réductibles à un effet aléatoire. Il y a certes redondance encore entre densité et biomasse de chaque catégorie mais des effets d'opposition entre différents types de communauté, ce qu'exprime la carte 2-3 des variables de l'ACP (figure 8).

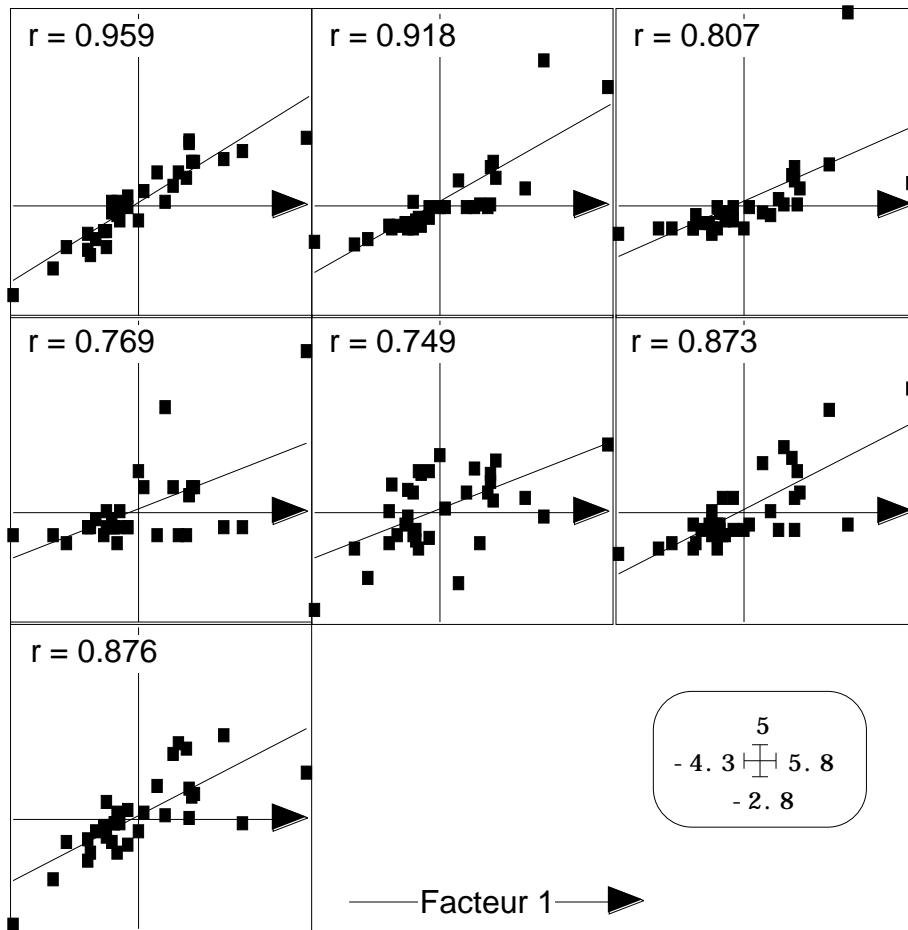


Figure 6 : Graphe canonique de l'ACP normée du tableau TLog.

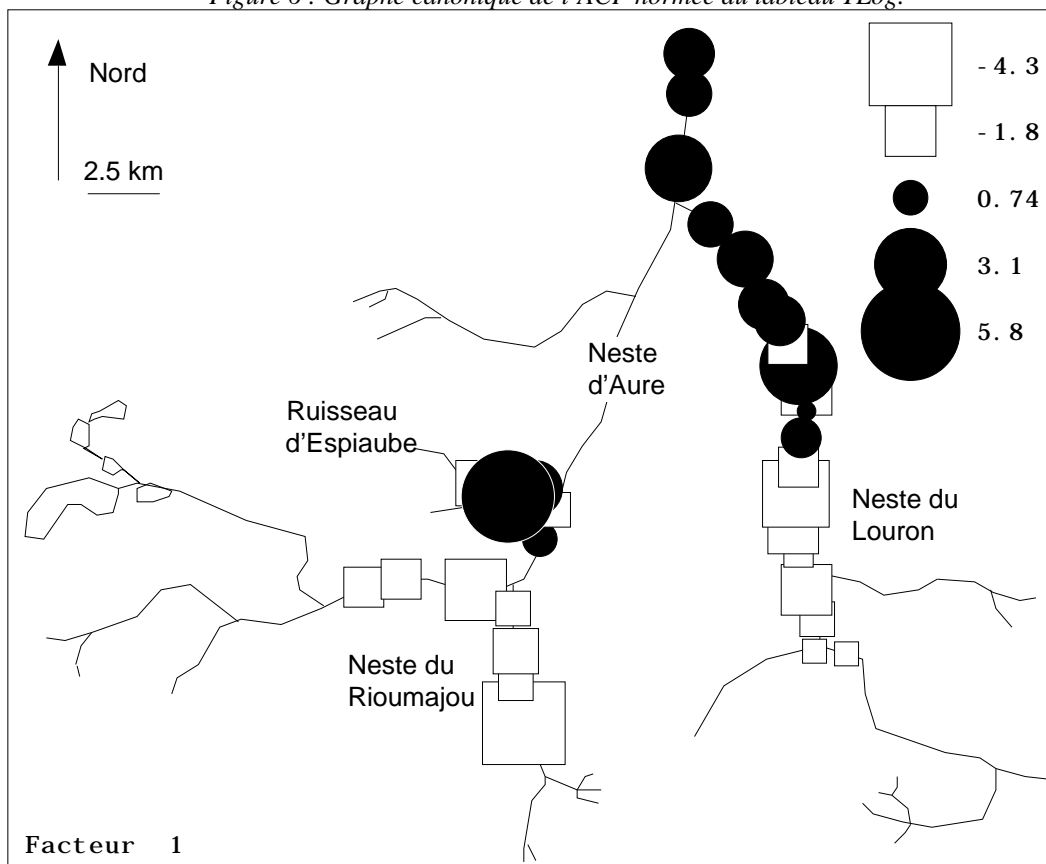


Figure 7 : Cartographie des coordonnées factorielles des lignes de l'ACP de TLog.

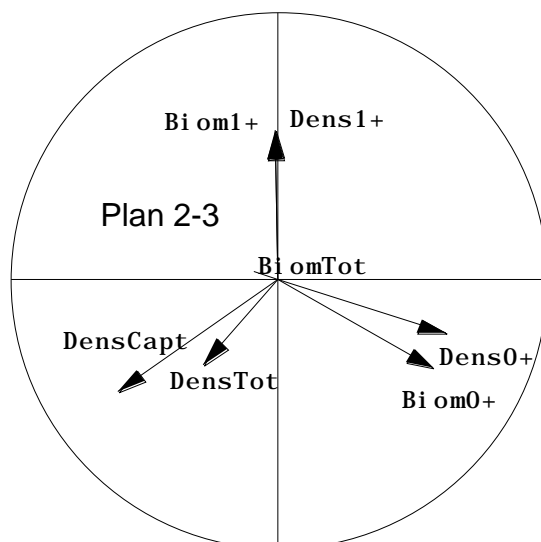


Figure 8: Plan 2-3 des colonnes de l'ACP de TLog.

Le vecteur BiomTot est perpendiculaire au plan de projection dans la figure 8. Reste trois directions représentées deux fois. Nous décomposerons donc le problème de prédiction en deux sous-problèmes. D'une part, nous considérerons la prévision de la biomasse totale, variable fondamentale, quasiment identique au facteur 1 de l'ACP, qu'on appelle **effet taille** en analyse des données, particulièrement en morphométrie⁴. D'autre part, et indépendamment, nous considérerons le tableau des résidus de la régression de chacune des autres variables sur la première, qui comportent surtout des indications sur l'**effet forme** sous-jacent. Ce faisant, nous employons une technique la plus transparente possible et proche de la structure de l'article étudié. La variable Biomasse totale (après passage en $\log(x + 1)$) est isolée dans un tableau Taille tandis que les résidus de prédiction indépendants de V1 par définition (ce qui reste des variables 2 à 7 de TLog quand on a enlevé la liaison avec V1) forme le tableau Forme (33-6).

Utiliser le modules FilesUtil. Extraire les 6 dernières colonnes de TLog dans Provi.:

Row-Col Selection			
Input file		TLog	33 7
Selection of rows (default = all)			
Selection of columns (default = all)		2a7	
Output file		Provi	

Extraire la première colonne dans Taille. Ajouter une colonne de 1 à Taille :

Add column 1n			
Input file		Taille	33 1
Output file		1n+Taille	

Utiliser le modules Projectors :

0.0340	-0.0431	-0.0058	-0.3866	-0.0106	-0.4202
0.0251	-0.0067	0.0010	0.3040	0.0408	1.1299
-0.0205	0.0032	-0.0188	-0.8934	-0.0066	0.1276
0.0047	-0.0601	-0.0452	-1.7428	0.0405	0.2122
-0.0726	-0.0112	-0.0179	-0.3137	-0.0247	-0.0514
0.0127	0.0031	0.0073	-0.1470	0.0057	0.2894
0.0502	-0.0559	0.0073	-0.2338	0.0216	-0.2525
0.2847	-0.0457	0.1582	0.6585	0.1194	-0.0899
0.1169	0.0606	0.0152	-1.0889	0.0212	-1.0098
-0.0249	-0.0013	-0.0341	-0.6575	0.0288	0.3070
0.0041	-0.0088	-0.0166	-0.0181	0.0067	0.3008
-0.0572	-0.0122	-0.0182	-0.5451	-0.0159	-0.0821
0.0003	-0.0317	-0.0184	-0.5177	0.0208	0.2143
-0.0833	-0.0187	-0.0239	0.7157	-0.0366	-0.8075
-0.0942	-0.0212	-0.0215	0.7570	-0.0413	-0.9377
0.0481	0.0096	0.0033	-0.1625	0.0266	0.3365
-0.0462	-0.0142	0.0073	0.8705	-0.0229	0.1102
0.0563	0.0464	0.0061	1.0537	0.0034	0.0605
0.0312	0.0355	0.0027	-0.1188	0.0064	-0.3440
-0.1070	-0.0095	-0.0206	0.3404	-0.0591	-0.6168
-0.0133	-0.0284	0.1251	0.3618	-0.0019	0.0468
-0.0444	-0.0238	0.0221	-0.0588	-0.0307	-0.1061
-0.0039	-0.0142	-0.0052	0.2964	-0.0286	-0.5109
0.4399	0.3586	-0.0461	-0.6328	-0.0708	-1.0785
-0.0029	-0.0178	0.0129	0.6688	-0.0118	0.0466
-0.0407	-0.0081	-0.0136	0.0903	-0.0922	-1.1963
-0.1872	-0.0296	-0.0120	-0.0352	-0.0424	-0.3852
-0.0129	-0.0197	0.0072	0.5911	-0.0139	-0.1556
-0.0733	-0.0148	0.0187	0.6767	-0.0469	1.2231
-0.0700	0.0266	-0.0427	-0.2199	0.1086	1.2086
-0.0996	-0.0434	-0.0460	-0.9761	0.0645	1.2241
-0.0943	0.0292	-0.0424	0.0890	0.0546	1.2240
0.0402	-0.0328	0.0545	1.2748	-0.0128	-0.0173

Tableau 2 : *Forme*, Résidus des prédictions des 6 dernières colonnes de TLog par la première.

File Reg.mod contains predicted variables
It has 33 rows and 6 columns

File Reg.coe contains weights
coefficients of linear combination of explanatory variables
It has 2 rows (explanatory v.) and 6 columns (dependant v.)

File Reg.res contains observed values - predicted values
It has 33 rows and 6 columns

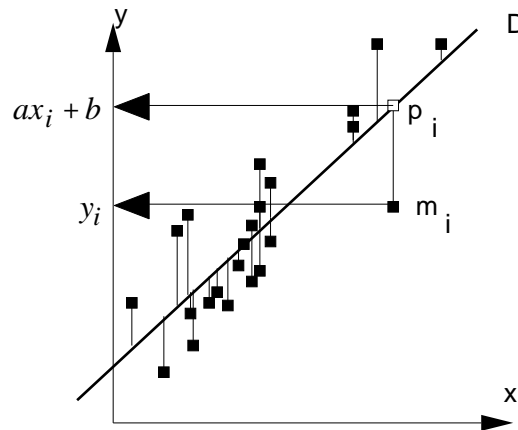
Renommer Reg.Res en Forme et éditer le résultat (tableau 2). Rencontrée pour la première fois, l'opération demandent quelques explications générales. On y trouve les objets essentiels manipulés dans le module Projectors dont l'esprit est celui de l'ouvrage de Takeuchi & Coll. (1982)⁵.

2.3 — Régression et projection : approche élémentaire

L'essentiel tient, pour commencer, dans la remarque suivante. On voit, en général la régression linéaire simple comme la recherche d'une droite D d'équation $y = ax + b$ qui rend minimum la quantité :

$$\sum_{i=1}^n \psi_i d^2(m_i, p_i)$$

où ψ_i est le poids de la ligne i , m_i le point observé de coordonnées (x_i, y_i) et p_i le point prévu de coordonnées $(x_i, ax_i + b)$:



Ceci revient à chercher les paramètres a et b qui minimisent :

$$E(a, b) = \sum_{i=1}^n \psi_i (y_i - ax_i - b)^2.$$

C'est le point de vue des n points de \mathbf{R}^2 . Dans le point de vue 2 points de \mathbf{R}^n , on considère au contraire les vecteurs $\mathbf{x} = (x_1, x_2, \dots, x_n)$ et $\mathbf{y} = (y_1, y_2, \dots, y_n)$. Il faut y ajouter le vecteur $\mathbf{1}_n = (1, 1, \dots, 1)$. Le modèle recherché est un vecteur $\hat{\mathbf{y}} = (ax_1 + b, ax_2 + b, \dots, ax_n + b)$ qui s'écrit plus simplement $\hat{\mathbf{y}} = a\mathbf{x} + b\mathbf{1}_n$. Faire la régression, c'est trouver la combinaison linéaire des vecteurs \mathbf{x} et $\mathbf{1}_n$ qui minimise la même quantité vue alors comme :

$$E(a, b) = \sum_{i=1}^n \psi_i (y_i - \hat{y}_i)^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \|\mathbf{y} - (a\mathbf{x} + b\mathbf{1}_n)\|^2$$

La solution est alors le projeté orthogonal, au sens de la métrique des poids :

$$\mathbf{D} = \text{Diag}(\psi_1, \psi_2, \dots, \psi_n)$$

du vecteur \mathbf{y} sur le sous-espace engendré par les vecteurs \mathbf{x} et $\mathbf{1}_n$. La régression multiple d'une variable \mathbf{y} sur p variables explicatives $\mathbf{x}_1, \dots, \mathbf{x}_p$ c'est la projection du vecteur \mathbf{y} sur le sous-espace engendré par les vecteurs $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p$ et $\mathbf{1}_n$. Le vecteur des écarts ou des résidus est le vecteur $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}$ perpendiculaire aux vecteurs définissant l'espace de projection. Faire une régression c'est projeter un vecteur sur le sous-espace engendré par d'autres vecteurs. On peut approfondir cette idée essentielle dans le chapitre IV (pages 77-112) de Rouanet et Le Roux (1993)⁶

Le module Projectors permet de définir des sous-espaces, de projeter des paquets de vecteurs et d'étudier de diverses manières l'effet de cette projection. C'est pourquoi nous sommes partis de la variable BiomTot (en Log, unique colonne de Taille). Le fichier 1n+Taille contient deux colonnes (la première est le vecteur $\mathbf{1}_{33}$ et la seconde la variable x). L'option XSubSpace définit le sous-espace associé par une base orthonormée (deux vecteurs colonnes à 33 composantes normée et orthogonale) par la procédure de Gram-Schmidt. Par défaut, la pondération est uniforme (1/33 pour chaque ligne). Ces deux colonnes forme le fichier Auxi.@ob (@ permet de repérer les sous-

espaces, ob pour *orthonormal basis*). Y est associé le fichier Auxi.@pl (pl pour poids des lignes, qui contient la pondération pour laquelle les colonnes du @ob correspondant sont orthonormés).

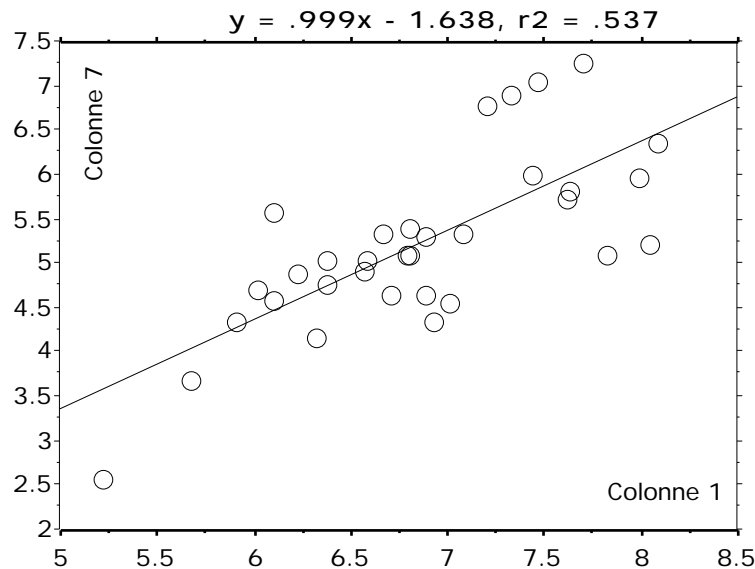
Le module @modelling projette un tableau quelconque (ici Provi) sur le sous-espace engendré par les colonnes de Auxi.@ob. On obtient les vecteurs projetés (modèles) en colonnes dans Reg.Mod (Mod pour modèle), les écarts dans Reg.Res (Res pour résidus) et les coefficients de régressions dans Reg.coe. Ce dernier fichier existe si et seulement si le tableau de départ (1n+Taille) est formé de colonnes linéairement indépendantes (intuitivement non totalement redondantes, comme, par exemple si une colonne est x_1 , une colonne est x_2 et une troisième est x_1+x_2).

Input file: Reg.coe
Row: 2 Col: 6

1	-0.9255	-0.4479	-0.1474	-0.6218	-0.2915	-1.6375
2	0.1662	0.0762	0.0273	0.5450	0.0537	0.9995

permet d'écrire que le modèle utilisé pour la variable 6 est :

$$\text{Log (Biom1+1)} = 0.9995 \text{ Log (BiomTot+1)} - 1.6375$$



On peut obtenir le tableau des résidus, appelé Forme, avec un programme classique de régression (ci-dessus StatViews™). Moins habituel pourra être l'usage de poids arbitraire, la définition de sous-espaces à partir de tous types de variables et les usages qui en seront faits.

Nous pourrons, sur cet exemple, aborder la question de la prédiction d'une variable (Taille) et celui de la prédiction d'un tableau (Forme). **Cette fiche porte désormais sur le premier problème.**

3. — Liaisons entre variables explicatives

On sait que la principale difficulté de la régression linéaire dérive des relations pouvant exister entre variables prédictives. Si ces relations sont parfaites, on a évidemment confusion parfaite d'effets, et si elles ne le sont pas, la redondance est une perturbation dans l'interprétation des résultats.

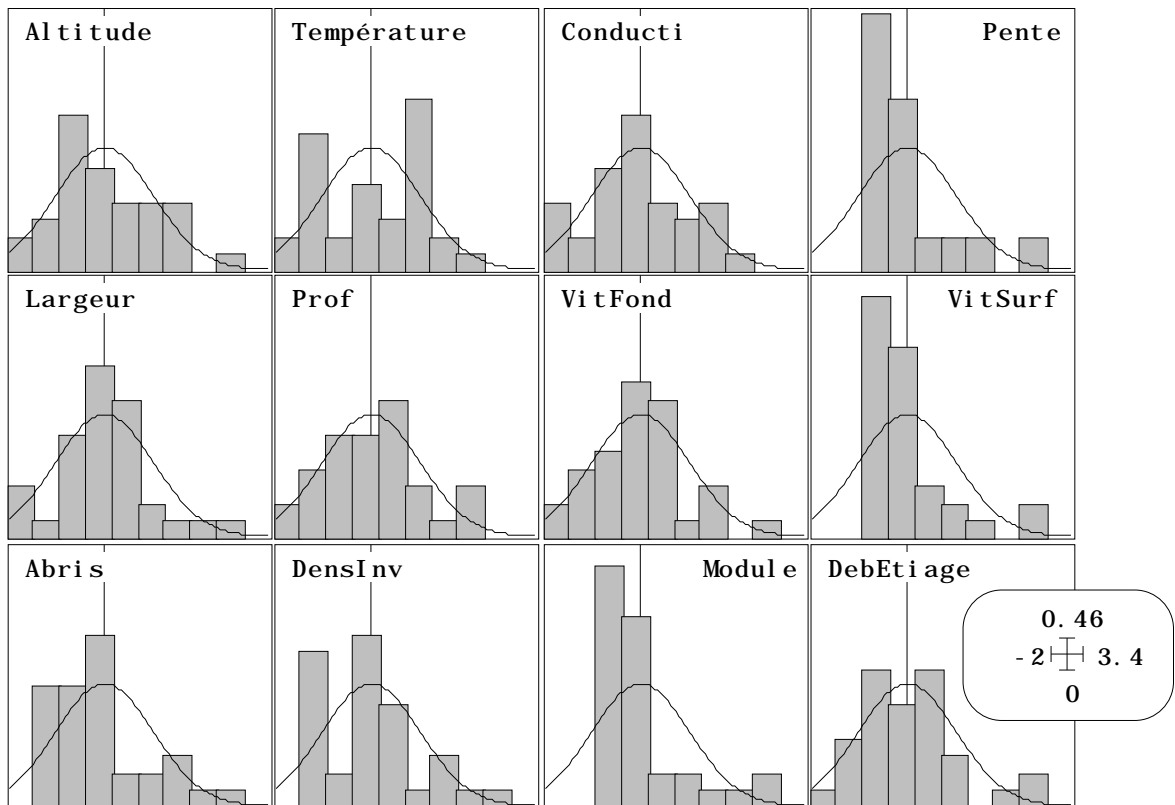
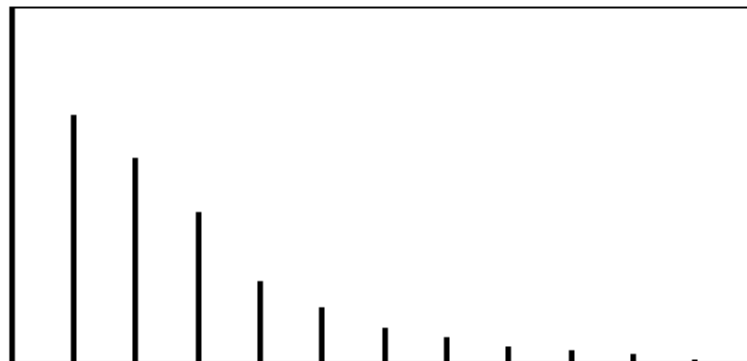


Figure 9 : Histogrammes des 12 variables explicatives normalisées.

Les histogrammes (figure 9) des variables mésologiques ne présentent aucune particularités pathologiques et on peut ne pas transformer les variables. L'ACP donne une matrice de corrélation (figure 10) avec peu de valeurs très importantes. On notera simplement la très bonne liaison linéaire entre altitude et température à laquelle on s'attend. La structure est relativement complexe et on peut garder au moins 4 axes :



Dans le plan 1-2 des variables (figure 11, a), on reconnaît en partie le gradient amont-aval. Les valeurs propres sont assez voisines, contrairement au cas précédent, il n'y a pas de direction privilégiée, mais un premier axe légèrement dominant et des sous-espaces de référence rendant compte de relations complexes. La configuration des variables dépend fortement du nombre de variables enregistrées⁷. Par exemple, si on enlève la profondeur de fond on obtient la figure 11 (b). Le gradient amont-aval (figure 10 c) reste en place (les coordonnées sont très voisines, figure 10d). Les cartes des variables sont voisines sans être identiques. Ceci indique qu'il n'y a pas de redondance (mesures de la même chose) mais de corrélation. La carte du facteur 1 n'est, évidemment pas sans rappeler la figure 7.

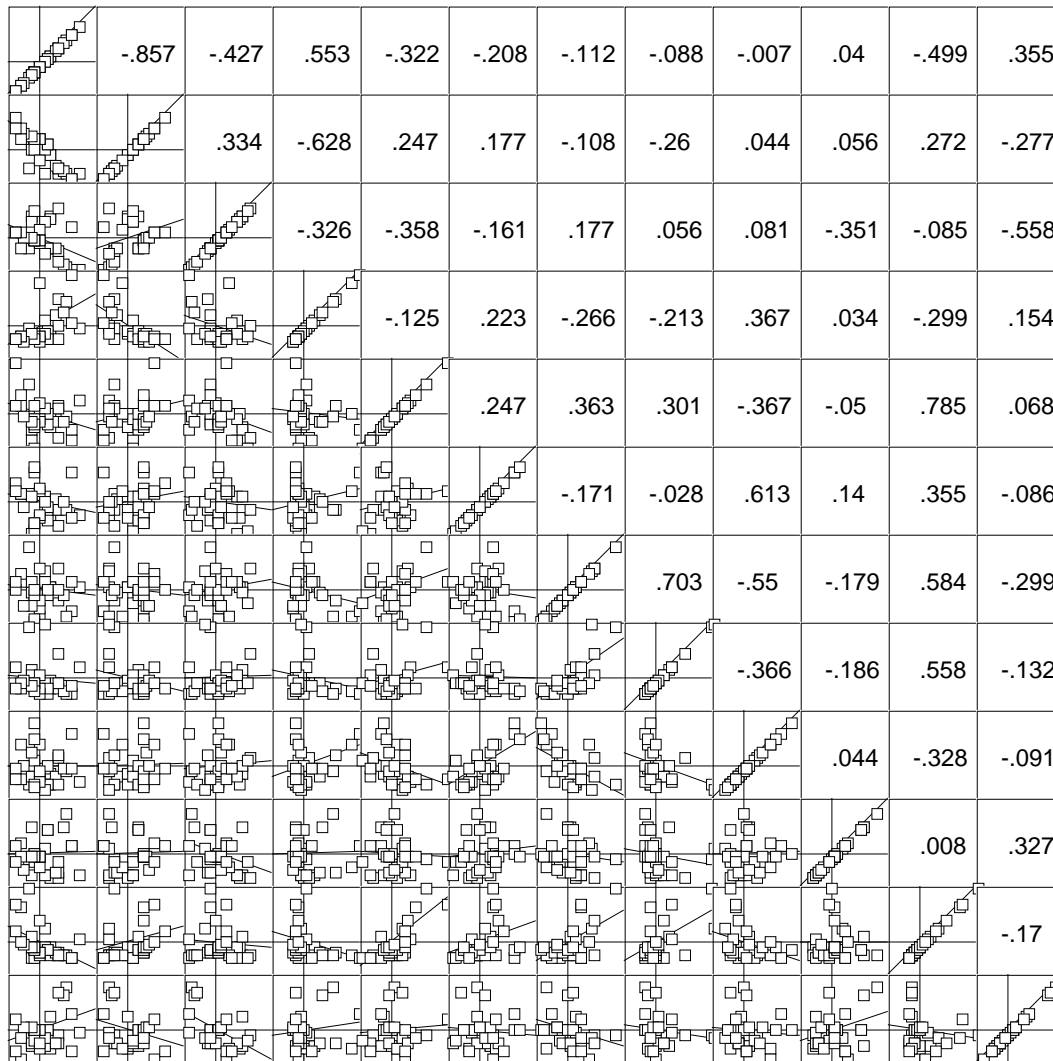


Figure 10 : Matrice de corrélation du fichier Mil.

On retiendra de cette partie préliminaire la transformation en Log impérative sur les variables biologiques, leur redondance et la séparation des effets taille (fichier Taille, 33-1) et forme (fichier Forme, 33-6), la corrélation complexe entre les variables du milieu gardées brutes (fichier Mil, 33-12), l'autocorrélation spatiale présente dans toute les données (cartographies) et la ressemblance des cartes de synthèse (forme et coordonnée de l'axe 1 de Mil).

On peut alors s'intéresser à l'examen des relations entre variables mésologiques et biologiques.

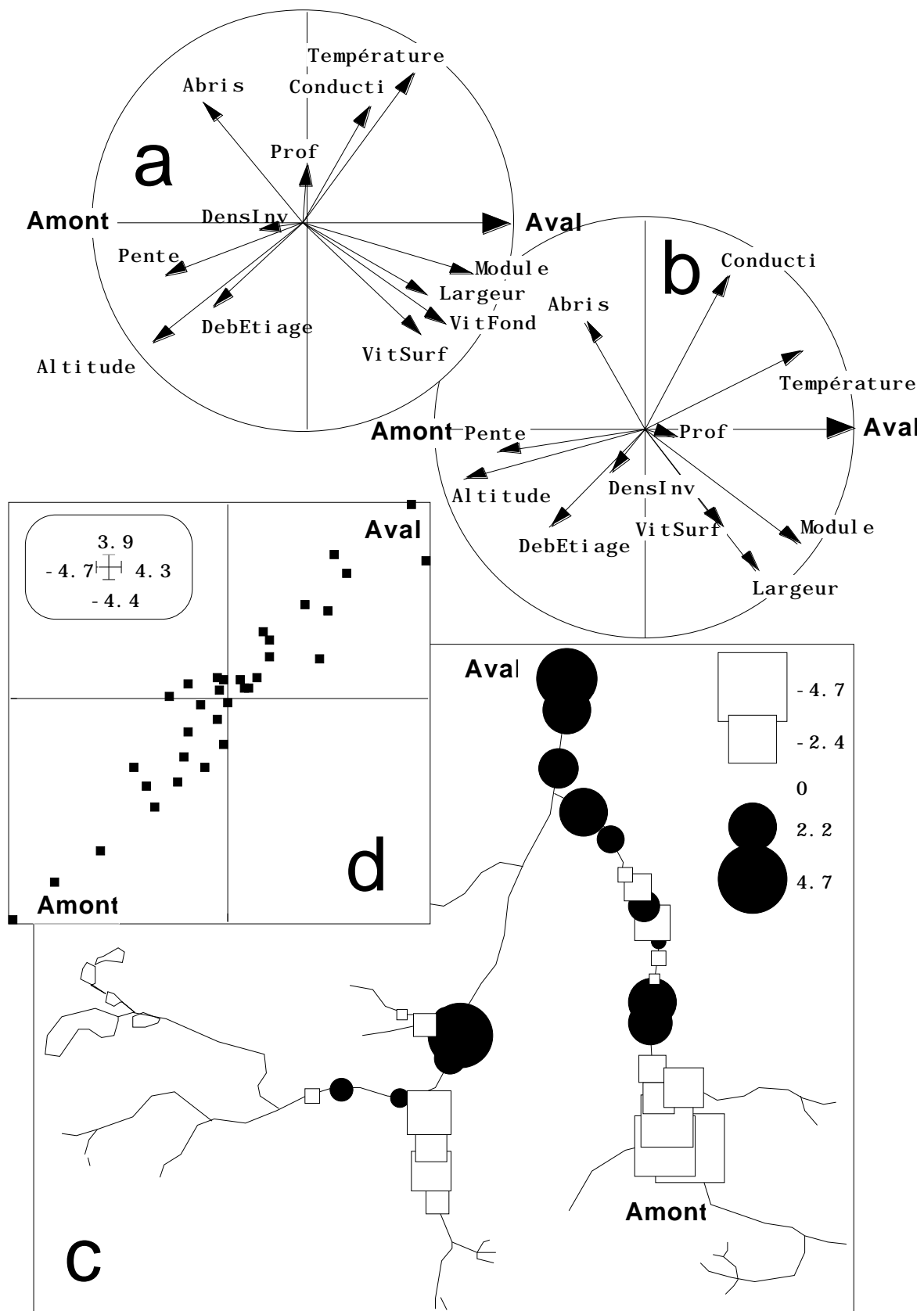


Figure 11 : Plan 1-2 de l'ACP (a - Mil. b - Mill après retrait de la variable 7, Vitesse au fond). c - Cartographie du facteur 1 de l'ACP de Mil. d - Coordonnées des lignes (axe 1) (en abscisse Mil, en ordonnée Mill).

4 — MLR : la régression linéaire multiple

Le tableau Mil contient 12 variables prédictrices sur 33 stations. Le tableau TLog contient 7 variables à prédire sur ces 33 stations. Que se passe-t'il quand on cherche brutalement un modèle linéaire de prédiction des unes par les autres ? Utiliser le module LinearReg et initialiser par l'option Initialize.

```
-----
New TEXT file TLog/Mil.reg contains the parameters:
----> Explanatory variables: Mil [33][12]
----> Dependant variable file: TLog [33][7]
----> Row weighting file: Uniform weighting
-----
```

On obtient (début) :

```
Multiple Linear Regression
-----
Explanatory variable file: Mil
It has 33 rows and 12 columns
Var. | Mean      | Variance |
  1 | 9.418e+02 | 1.808e+02 |
  2 | 1.240e+01 | 1.474e+00 |
...
 10 | 3.657e+02 | 1.977e+02 |
 11 | 1.762e+03 | 1.941e+03 |
 12 | 3.902e+01 | 1.602e+01 |
-----
Dependent variable file: TLog
It has 33 rows and 7 columns
R2 = Squared multiple correlation coefficient
Var. | Mean      | Variance | R2      |
  1 | 6.860e+00 | 7.046e-01 | 8.386e-01 |
  2 | 2.148e-01 | 1.615e-01 | 8.026e-01 |
  3 | 7.466e-02 | 8.738e-02 | 7.687e-01 |
  4 | 3.954e-02 | 4.667e-02 | 6.093e-01 |
  5 | 3.117e+00 | 7.661e-01 | 5.157e-01 |
  6 | 7.713e-02 | 5.895e-02 | 5.117e-01 |
  7 | 5.219e+00 | 9.606e-01 | 7.721e-01 |
-----
```

On voit que le module permet la régression linéaire multiple pour chacune des variables de TLog sur l'ensemble des variables de Mil. Entre 84% (variable 1) et 51% (variables 5 et 6) de variance est expliquée. Cette remarque est vraie pour toutes les options. On ne s'intéresse pour l'instant qu'à la seule variable biomasse totale extraite de TLog dans Taille. Réinitialiser et utiliser la même option Modelling.

```
Multiple Linear Regression
-----
```


Explanatory variable file: Mil
 It has 33 rows and 12 columns

Var.	Mean	Variance
1	9.418e+02	1.808e+02
...		
12	3.902e+01	1.602e+01

 Dependent variable file: Taille
 It has 33 rows and 1 columns
 R2 = Squared multiple correlation coefficient

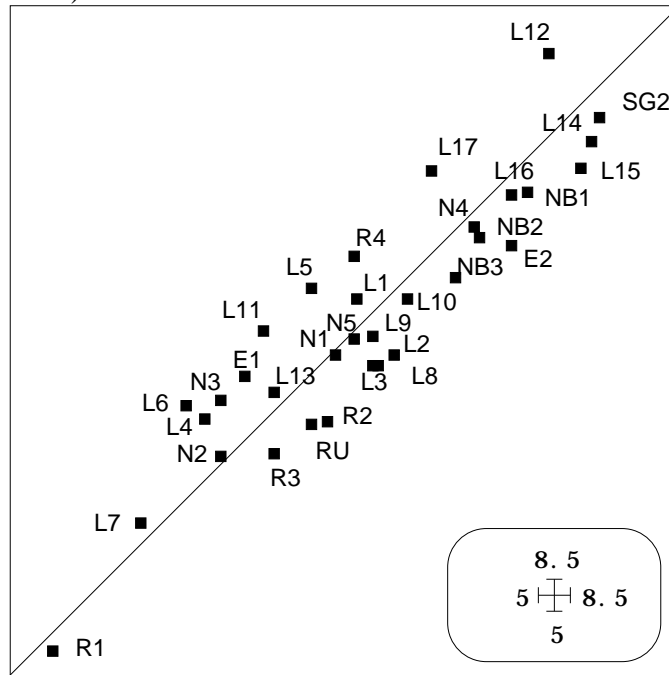
Var.	Mean	Variance	R2
1	6.860e+00	7.046e-01	8.386e-01

 File Taille/Mil.MLRmod has 33 rows and 1 columns
 It contains linear models
 from separate multiple linear regression of each dependant variable
 upon the set of explanatory variables

File :Taille/Mil.MLRmod

Col.	Mini	Maxi
1	5.116e+00	8.254e+00

84% de variance expliquée donne un bon modèle (en abscisse, les données, en ordonnée les prévisions) :



 File Taille/Mil.MLRres has 33 rows and 1 columns
 It contains (data - model) matrix
 File :Taille/Mil.MLRres

Col.	Mini	Maxi
1	-4.942e-01	3.760e-01

 File Taille/Mil.MLRw1 has 12 rows and 1 columns
 It contains regression coefficients
 Rows : explanatory variables / Columns : dependant variables
 Models for normalized (mean = 0 / variance =1) variables

File :Taille/Mil.MLRw1

Col.	Mini	Maxi
1	-8.010e-01	6.069e-01

Noter que le modèle est écrit pour les variables normalisées. Chercher alors les coefficients de corrélation entre la variable à prédire et les variables explicatives (MatAlg) :

Ouvrir le fichier Auxi qui contient les corrélations et le fichier Taille/Mil.MLRw1 qui contient les coefficients de régression :

Coefficients		Corrélation		
[1]	-0.8010	[1]	-0.6096	Altitude
[2]	-0.2816	[2]	0.5448	Température
[3]	0.0114	[3]	0.3082	Conducti
[4]	-0.1083	[4]	-0.1908	Pente
[5]	-0.4596	[5]	-0.1768	Largeur
[6]	-0.2039	[6]	0.3625	Prof
[7]	-0.0781	[7]	-0.3584	VitFond
[8]	-0.3371	[8]	-0.2817	VitSurf
[9]	0.6069	[9]	0.5524	Abris
[10]	0.0236	[10]	0.1527	DensInv
[11]	0.6014	[11]	0.0481	Module
[12]	0.2516	[12]	-0.0733	DebEtage

On voit immédiatement que la corrélation de l'abondance avec l'altitude est de -0.6 et avec la température de +0.5, ce qui est normal puisqu'on est dans la gamme de température 10°-16° en dessous de l'optimum (la biomasse augmente avec la température et donc diminue avec l'altitude, op. cit. p. 333). Le modèle utilisé commence cependant par :

$$\text{Biomasse} = -0.80 * \text{altitude} - 0.2 * \text{température} + \dots (\text{variables réduites}).$$

Une variable corrélée positivement (0.60) avec l'abondance (11 : module) donne un coefficient presque nul (0.05) alors que la variable 3 (conductivité) non corrélée (0.01) donne un coefficient non négligeable (0.30). Cette observation, tant de fois refaite, que les coefficients sont de peu de signification est caractéristique d'un déséquilibre entre nombre de points (33) et nombre d'explicatives (12). On a le choix entre augmenter le nombre de stations, diminuer le nombre de variables ou changer de méthode.

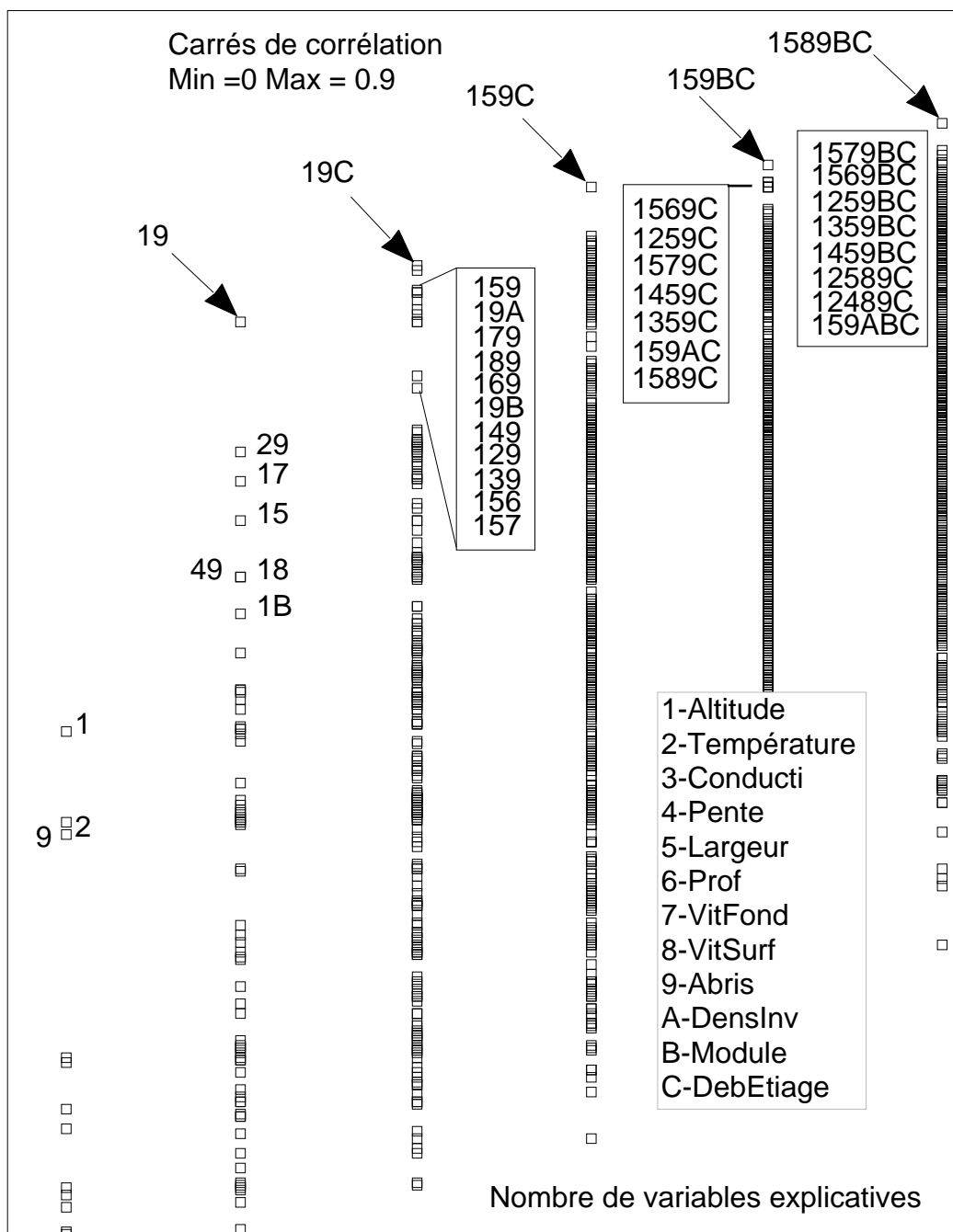


Figure 12 : 2509 carrés de corrélation des régressions multiples de *Taille* sur *Mil*.

5 — Sélection de variables en régression linéaire

La totalité des logiciels de statistique comporte des modules de régression multiple, c'est pourquoi ADE est assez pauvre sur la question. Le module RegMul et son option MultCorCoef permet de calculer tous les carrés de corrélation de toutes les régressions multiples utilisant au plus un nombre donné de variables explicatives. Il vaut mieux être raisonnable, car avec 12 variables on compte déjà 2509 modèles de régression à 1, 2, ..., 6 variables explicatives.

Utiliser MultiCorCoeff avec les paramètres :

MultiCorCoeff			
Explanatory variables		Mil	33 12
Dependent variables		Taille	33 1
Maximal number of dependent		6	
Option: row weighting			
Output file name		Taille/Mil	

On obtient :

Multiple correlation on all possible combination of variables

 Multiple correlation coefficient (R2): Taille/Mil.yr2

It has 2509 rows and 1 columns

On one row, R2 of each dependent variable for a given explanatory variable set

Label of the combinations in file Taille/Mil.lab with 2509 rows

Number of explanatory variables used in column 1 of file Taille/Mil.xr2 with 2509 rows

Rank of the projection subspace in column 2 of file Taille/Mil.xr2 with 2509 rows

L'utilisation de Curves permet d'obtenir la figure 12 :

Lines			
X file (default = 1, 2, 3, ..., n)		Taille/Mil.xr2	2509 2
X file column number (default = 1)			
Y file (no default)		Taille/Mil.yr2	2509 1
Cumulated data (1=yes, 2=no)			
Variable label file (or #)			
Draw curves (1=yes, 2=no)		2	
Draw points (1=yes, 2=no)		1	
Row label file (or #)		Taille/Mil.lab	
Number of curves by window		1	

La plus grande partie des 2509 étiquettes a été supprimées et Claris Draw™ demande 7 Mo de mémoire pour manipuler la figure. On pourrait avoir des ambitions plus limitées, sans le résultat est sans ambiguïté. On sait qu'ajouter des variables augmente certainement le pourcentage de variance expliquée, mais pas forcément l'intérêt expérimental du modèle. Du premier coup d'œil, on voit que la régression Abondance = f(Altitude, Abris) est optimale. L'altitude représente en fait la température et la discussion des auteurs (op. cit.) est parfaitement convaincante. Pour obtenir le modèle exporter les données dans un logiciel de statistiques classiques.

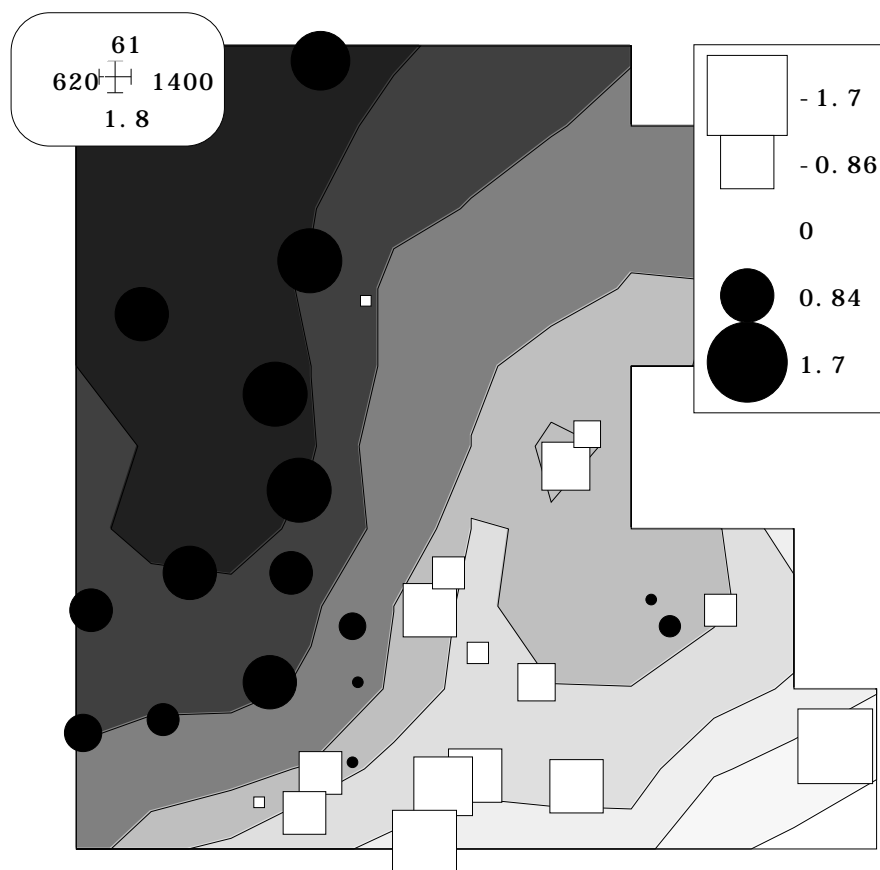
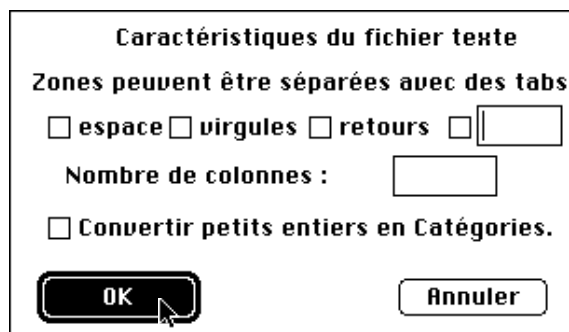


Figure 13 (première partie) : En abscisse, l'altitude en m. En ordonnée, les abris en 1/100. Représentation des données d'abondance centrée par valeurs (module Scatters de ADE-4, option Values) et par courbes de niveaux (module Levels).

Par exemple, juxtaposer les fichiers Taille et Mil dans un fichier A :



Editer le fichier binaire A avec le logiciel StatViews™:



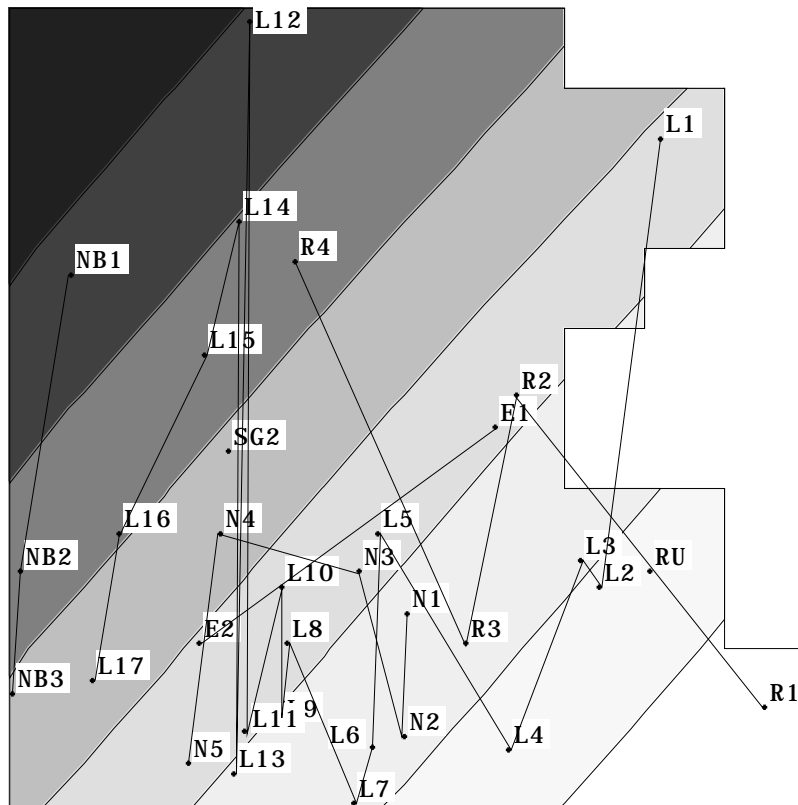


Figure 13 : (deuxième partie. En abscisse, l'altitude en m. En ordonnée, les abris en 1/100. Représentation du modèle linéaire par courbes de niveaux, étiquetage des stations et tracé des rivières. Il n'y a lieu d'introduire un terme d'interaction dans le modèle.

Sélectionner les variables et l'option régression multiple, activer la régression, copier la colonne du modèle dans un nouveau document, sauvegarder en texte et repasser le résultat en binaire :

Colonne 1	Colonne 2	Colo...	Colo...	Col...	Colo...	Colo...	Colo...	Colo...	Colonne 10
Y ₁	X ₁								X ₂
6.7056	1017	12.0	169	1.8	5.6	.23	.29	.55	16

Vue de A-t
Régression Multiple Y₁:Colonne 1 2 Variables X

Variable :	Coefficient :	Err. Std. :	Coeff. Std. :	(Valeur)-t :	Probabilité :
CONSTANTE	8.53				
Colonne 2	-.002	4.074E-4	-.606	5.793	.0001
Colonne 10	.027	.005	.548	5.242	.0001

Résiduel : Colonne 15 Résidu Std. : Colonne 16 Ajusté : Colonne 17 Prédite : Colonne 18

Sans-titre	
Colonne 1	
1	6.561
2	6.335
3	6.753
4	7.165
5	6.777
6	6.743

Enregistrer fichier sous : A1.txt HDdc Ejecter

Enregistrer Annuler Bureau

Format : Normal Texte

Ceci montre qu'il faut un instant pour exporter un fichier binaire d'ADE-4 dans un logiciel quelconque qui importe des fichiers textes et utiliser les résultats dans un nouveau fichier binaire. On construit ainsi les figures 13 et 14 qui explore plus à fond le modèle.

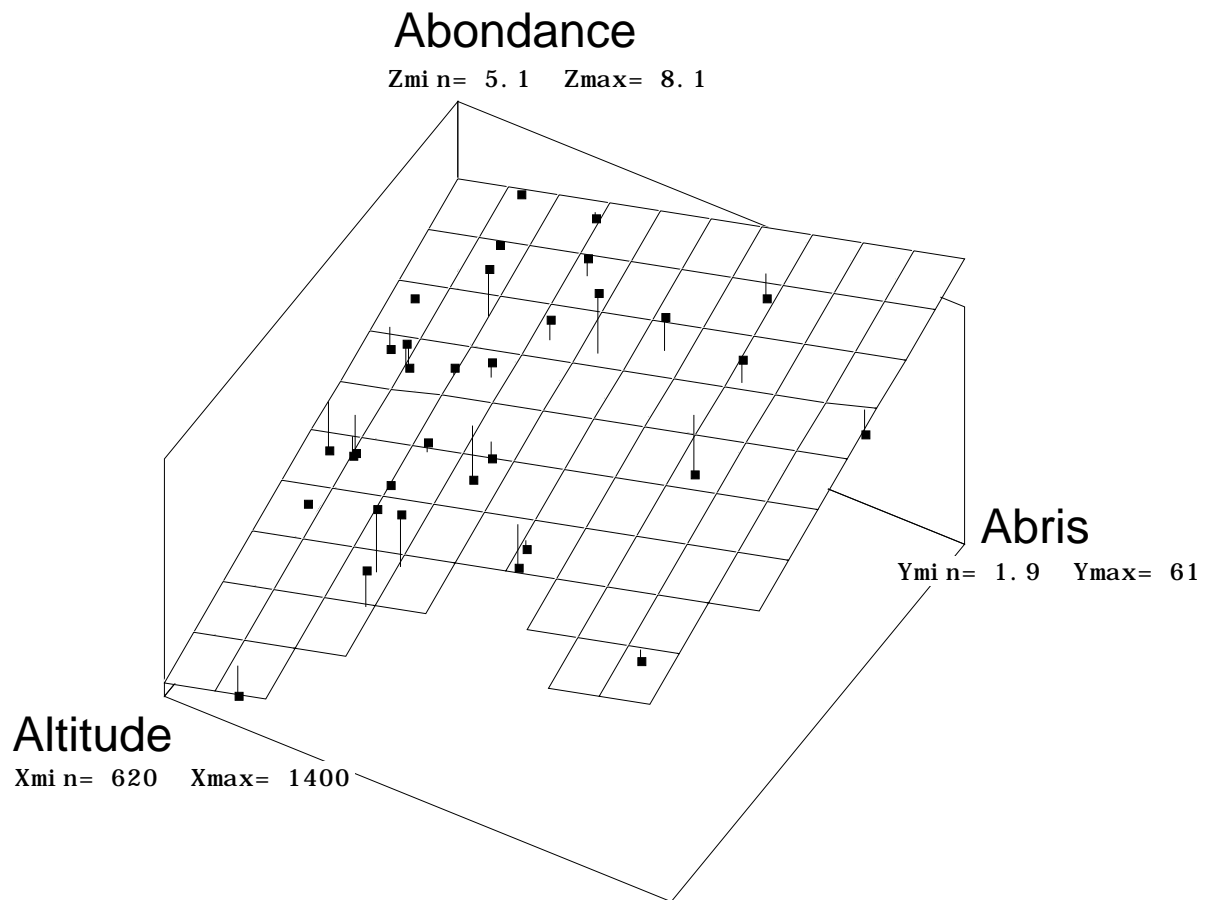


Figure 14 : Données en 3D et modèle linéaire ($z = 8.53 - 2.3610^{-3} x + 2.698 \cdot 10^{-2} y$). Module autoprogrammable en QuickBasic™ XYZ-Gen de ADE 3.7.

Dans RegMul, on obtient le même résultat en sélectionnant les colonnes 2 et 9 :

```
New TEXT file Taille/mil2-9.reg contains the parameters:
----> Explanatory variables: Mil2-9 [33][2]
----> Dependant variable file: Taille [33][1]
----> Row weighting file: Uniform weighting
-----
Multiple Linear Regression
-----
Explanatory variable file: Mil2-9
It has 33 rows and 2 columns
Var. | Mean | Variance |
  1 | 1.240e+01 | 1.474e+00 |
  2 | 2.048e+01 | 1.432e+01 |
Dependent variable file: Taille
It has 33 rows and 1 columns
Var. | Mean | Variance | R2 |
  1 | 6.860e+00 | 7.046e-01 | 5.766e-01 |
-----
File Taille/mil2-9.MLRmod has 33 rows and 1 columns
It contains linear models from separate multiple linear regression of
each dependant variable upon the set of explanatory variables
File :Taille/mil2-9.MLRmod
|Col.| Mini | Maxi |
```

1	5.905e+00	8.164e+00
---	-----------	-----------

File Taille/mil2-9.MLRres has 33 rows and 1 columns
 It contains (data - model) matrix
 File :Taille/mil2-9.MLRres

Col.	Mini	Maxi
1	-8.230e-01	1.003e+00

File Taille/mil2-9.MLRw1 has 2 rows and 1 columns
 It contains regression coefficients
 Rows : explanatory variables / Columns : dependant variables
 Models for normalized (mean = 0 / variance =1) variables
 File :Taille/mil2-9.MLRw1

Col.	Mini	Maxi
1	5.215e-01	5.295e-01

Le modèle s'écrit $\text{abondance} = 0.52 * \text{altitude} + 0.53 * \text{Abris}$ pour deux corrélations (p. 17) de 0.54 et 0.55. La cohérence est parfaite et souligne la bizarrerie du précédent modèle. On est passé de 84% de variance expliquée à 58% ; la perte est donc loin d'être négligeable. Il serait agréable d'augmenter la précision du modèle sans tomber sur une combinaison linéaire folklorique. C'est l'objet du dernier paragraphe.

6 — Régression PLS

Pour exécuter la régression partiellement aux moindres carrés, on utilise simplement après initialisation :

PLS -> Randomization Test	
Input file	<input type="text" value="Taille/Mil.reg"/>
Number of permutations	<input type="text" value="1000"/>

On obtient :



```

PLS1 - Permutation test
-----
Explanatory variable file: Mil
It has 33 rows and 12 columns
-----
Dependent variable file: Taille
It has 33 rows and 1 columns
-----
----- Vari number:      1 -----
-----
| Step | Nrepet | X>Xobs | Frequence |
|-----|-----|-----|-----|
| 1 | 1000 | 0 | 0.000e+00 |
| 2 | 1000 | 187 | 1.870e-01 |
| 3 | 1000 | 475 | 4.750e-01 |
| 4 | 1000 | 711 | 7.110e-01 |
| 5 | 1000 | 305 | 3.050e-01 |
| 6 | 1000 | 558 | 5.580e-01 |
| 7 | 1000 | 953 | 9.530e-01 |
| 8 | 1000 | 980 | 9.800e-01 |

```

Cette méthode est itérative. A chaque pas on cherche une combinaison linéaire explicative, on fait la prévision linéaire, on enlève la prévision de la variable à expliquer

et la combinaison prédictive des variables explicatives. On obtient une nouvelle variable à prédire, résidu du tour précédent et de nouvelles variables prédictrices indépendantes de la combinaison déjà utilisée. La question est donc le nombre d'itération à utiliser. Par tests de permutation réinitialisés à chaque tour, on compte la fréquence des permutations aléatoires qui donnerait un pourcentage d'explication aussi bon. Au premier tour, on n'en trouve aucune, signe que la prédiction a un sens. Au second tour on en trouve 19%, ce qui implique qu'il n'est possible de tenir compte que d'une seule itération. D'où le dialogue :

PLS -> Modelling	
Input file	 Taille/Mil.reg
Number of components (no default)	 1

On obtient :

```

PLS1 - Modelling
-----
Explanatory variable file: Mil
It has 33 rows and 12 columns
-----
Dependent variable file: Taille
It has 33 rows and 1 columns
-----
|-----| Col: 1 |-----| | |
|Step| Variance | Explained | Ratio | Exp. Sum |
| 1 | 1.000e+00 | 6.398e-01 | 6.398e-01 | 6.398e-01 |

```

La première itération donne 64% de variance expliquée soit mieux que la meilleure sélection de deux variables (58%) mais moins que la meilleure combinaison (84%). C'est pourquoi on dit partiellement aux moindres carrés, car on optimise non la corrélation modèle-donnée (ce qui minimise la somme des carrés des écarts) mais la covariance (ce qui maximise partiellement la variance, comme dans l'ACP du tableau des explicatives, et partiellement la corrélation, comme dans la régression classique). Ce faisant, on a simplement pris comme prédicteur les coordonnées de l'analyse de covariance entre le tableau Mil et le tableau Taille, qui ne comporte qu'une seule variable qui est à elle-même sa première composante principale.

```

File Taille/Mil.PLSw1 has 12 rows and 1 columns
It contains coefficients
File :Taille/Mil.PLSw1
|Col.| Mini | Maxi |
|----|-----|-----|
| 1 | -2.589e-01 | 2.346e-01 |
|----|-----|-----|
-----
File Taille/Mil.PLSmod has 33 rows and 1 columns
It contains components models
File :Taille/Mil.PLSmod
|Col.| Mini | Maxi |
|----|-----|-----|
| 1 | 5.626e+00 | 8.052e+00 |
|----|-----|-----|
-----
File Taille/Mil.PLSres has 33 rows and 1 columns
It contains residuals
File :Taille/Mil.PLSres
|Col.| Mini | Maxi |
|----|-----|-----|
| 1 | 6.064e+00 | 7.915e+00 |
|----|-----|-----|

```

	Coeff. MLR		Coeff. PLS		Corrélation	
[1]	-0.8010	[1]	-0.2589	[1]	-0.6096	Altitude
[2]	-0.2816	[2]	0.2314	[2]	0.5448	Température
[3]	0.0114	[3]	0.1309	[3]	0.3082	Conducti
[4]	-0.1083	[4]	-0.0810	[4]	-0.1908	Pente
[5]	-0.4596	[5]	-0.0751	[5]	-0.1768	Largeur
[6]	-0.2039	[6]	0.1540	[6]	0.3625	Prof
[7]	-0.0781	[7]	-0.1522	[7]	-0.3584	VitFond
[8]	-0.3371	[8]	-0.1197	[8]	-0.2817	VitSurf
[9]	0.6069	[9]	0.2346	[9]	0.5524	Abris
[10]	0.0236	[10]	0.0649	[10]	0.1527	DensInv
[11]	0.6014	[11]	0.0204	[11]	0.0481	Module
[12]	0.2516	[12]	-0.0311	[12]	-0.0733	DebEtiage

On retrouve automatiquement le couple Altitude-Température et les Abris comme meilleures variables prédictives. La cohérence coefficients-corrélations est respectée, le rôle des variables Conductivité, Profondeur et Vitesses rentre en ligne de compte. La régression PLS l'emporte sans conteste sur la régression MLR en interprétabilité au prix d'une perte certaine en précision. Ce qui reste a un sens. La sélection des variables s'est faite sans effort, avec des nuances.

Cette première approche très intuitive sera poursuivie, car la régression PLS (algorithme dans ⁸, s'étend à plusieurs variables à prédire (synthèse complète dans Lindgren (1994) ⁹), comme la régression linéaire s'étend dans l'ACP sur variables instrumentales. Il est logique d'attendre la même différence dans le cas multivarié des variables à prédire.

Références

- ¹ Baran, P., Delacoste, M., Lascaux, J.M. & Belaud, A. (1993) Relations entre les caractéristiques de l'habitat et les populations de truites communes (*Salmo trutta* L.) de la vallée de la Neste d'Aure. *Bull. Fr. Pêche Piscic* : 331, 321-340.
- ² Lilliefors, H.W. (1967) On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association* : 64, 387-389.
- ³ Eckart, C. & Young, G. (1936) The approximation of one matrix by another of lower rank. *Psychometrika* : 1, 211-218.
- ⁴ Yoccoz, N. (1988) *Le rôle du modèle euclidien d'analyse des données en biologie évolutive*. Thèse de doctorat, Université Lyon 1. 1-254.
- ⁵ Takeuchi, K., Yanai, H. & Mukherjee, B.N. (1982) *The foundations of multivariate analysis. A unified approach by means of projection onto linear subspaces*. John Wiley and Sons, New York. 1-458.
- ⁶ Rouanet, H. & Le Roux, B. (1993) *Analyse des données multidimensionnelles*. Dunod, Paris. 1-310.
- ⁷ Ramsey, F.L. (1986) A fable of PCA. *The American Statistician* : 40, 4, 323-324.
- ⁸ Ter_Braak, C.J.T. & Juggins, S. (1993) Weighted averaging partial least squares regression (WA-PLS): an improved method for reconstructing environmental variables from species assemblages. *Hydrobiologia* : 269/270: 485-502.
- ⁹ Lindgren, F. (1994) *Third generation PLS. Some elements and applications*. Research Group for Chemometrics. Department of Organic Chemistry. Umeå University. S-901 87 Umeå, Sweden. ISBN 91-7174-911-X, 1-57 & 5 papers.