Bull. Ecol. Res. Comm. (Stockholm) 17: 433–442 (1973).

Measures of Rank Distances Followed by Repeated Clustering and Tests of Rank Correlations in the Study of Biological and Chemical Data from Tropical Forest Soils (Ivory Coast)

A. M. Aubry, R. Van den Driessche, D. Bauzon, A. Perraud & Y. Dommergues

Investigations on soil microbial activity or microbial populations in the field are usually based on large numbers of microbiological, chemical and physical data. The aim of this paper is to present two methods which may speed up and facilitate their interpretation.

Methods

Four programs written in Fortran have been used on the 1108 and 370 computers (Aubry, 1972).

Capacity

The program package allows for a maximum number of 37 samples, 100 variables and no more than 17 samples per cluster.

Data Sheet

The raw data matrix consists of m rows and v columns. The row entries are the sample numbers, the column entries are the biological or chemical variables. Data are punched in the F5.0 format (with 9999 as greatest value instead of 99999) or in a smaller Fn.0 format. Missing data are punched -1 right justified. As a single card can only accomodate a limited number of variables, several cards may have to be used in sequence for each sample.

The programs manipulate simultaneously ordinal and interval variables. Only interval variables appear in the example given in Table 2, which has been punched into 2 x 32 cards.

The Distance Program

A distribution-free measure of distance between samples using ranks has been adopted (Kendall & Stuart, 1966; Van den Driessche & Garcia Gomez, 1973). The data matrix of the x_{hi} is replaced by a rank matrix of r_{hi}, where ordering has taken place a column at a time, in other words independently for each variable. No replacement occurs (the coded form -1 is used) for missing data. A weighing coefficient I_i which takes into account the number of missing data, the number e_i of sets of tied ranks and their size t_{qi} is used:

$$I_i = [\, m_i^3 - m_i - \sum_{q=1}^{e_i} (t_{qi}^3 - t_{qi}) \,]/12 \qquad (1)$$

where m_i is the actual number of samples for variable i (m_i = m - missing data of i). I_i equals zero and variable i is then deleted when all the data of variable i are equal.

The distance index between samples h and k varies between zero and two and is computed as follows:

$$D_{hk} = \frac{1}{v} \sum_{i=1}^{v} \frac{(r_{hi} - r_{ki})^2}{I_i} \qquad (2)$$

If for a given i one or both ranks r_{hi} r_{ki} are missing, due to missing data, no difference occurs and the variable i vanishes for the distance in question.

The $\binom{m}{2}$ distances are computed. Both punched and printed output is provided: punched cards with all the distances in the F9.4 format and a printed matrix in the A1 format, where class symbols 0, 1, 2,... 9, A, B,... N, P, ... Z take the place of the distances less than .34, the higher distances being deleted.

The Clustering Program

Input is the lower half of the distance matrix in the F9.4 format output from the distance program. A clustering procedure is followed using two criteria:

1. mean distance between two clusters > mean distance within both clusters;
2. distances between unclustered samples and the clusters > mean distance within each cluster.

Mean distances between and within clusters are printed with two decimals in matrix form. Content of each cluster appears as row and column entries.

The Program for Averaging the Data within the Clusters

Data are averaged among the samples of each cluster. Punched output as well as printed output is provided in the F9.4 format.

Repeated Clustering

The distance program is then used a second time with this punched output and repeated clustering follows. The whole procedure is repeated once again and the final clusters appear in the matrix form of Table 4.

The Correlation Program

A rank measure of correlation between variables has been used (Spearman, 1904). The same ranking procedure is followed as for the distance approach.

Two weighting coefficients I_i and I_j are needed for the variables i and j:

$$I_i = \frac{m_{ij}^3 - m_{ij}}{12} - \sum_{q=1}^{e_i} \frac{t_{qi}^3 - t_{qi}}{12} \qquad (3)$$

$$I_j = \frac{m_{ij}^3 - m_{ij}}{12} - \sum_{q=1}^{e_j} \frac{t_{qj}^3 - t_{qj}}{12} \qquad (4)$$

m_{ij} stands for the sample size expressed in number of samples, where both variables i and j occur. Symbols e_j and t_{qj} are homologous to e_i and t_{qi} already encountered.

The correlation coefficient:

$$R_{ij} = [\, I_i + I_j - \sum_{h=1}^{m} (r_{hi} - r_{hj})^2]/2 \sqrt{I_i I_j} \qquad (5)$$

varies between -1 and 1.

Tabular values (Beyer, 1966) are available for the correlation test at the .01 and .05 probability levels, for sample sizes comprised between 6 and 37, and are part of the program.

The same printed output appears on the left and right half of each page. Output is divided into non-significant, positive and negative correlations. The actual sample size appears underneath the two variables and the probability level is indicated. Variables appear in plain language.

Examples

The above methods are exemplified by the results of an investigation on forest soils from the Ivory Coast.

This work was based on the analysis of 32 soil samples (0-10 cm horizon) collected in November 1969 and March 1970 from 16 stations (Tables 1 and 2). For each soil sample 20 variables were used:

1. Chemical variables sensu stricto (plus clay content)
   - C/N
   - PH (pH H_2O 1:2.5)
   - S/ (total exchangeable cations, meq/100 g soil)
   - T (cation exchange capacity, meq/100 g soil)
   - TCC (total cation content, meq/100 g soil)
   - TFE (total Fe, expressed as Fe_2O_3 percentage)
   - TP (total P, mg P/g soil)
   - CLA (clay content, percentage)
2. Humic variables
   - FA (fulvic acids expressed as mg carbon/g soil)
   - HA (humic acids expressed as mg carbon/g soil)
   - GHA (gray humic acids, percentage of HA)
   - HU (humine expressed as mg carbon/g soil)
3. Biological variables
   - CO2 (potential CO_2 evolution, expressed as mg CO_2 evolved/g soil/7 days)
   - DES (dehydrogenase activity expressed as μl H transferred/g soil/24 hr)
   - SAC (saccharase, expressed as μmole reducing sugars/g soil/24 hr)
   - AMY (amylase, expressed as μmole reducing sugars/g soil/96 hr)
   - GLU (beta-glucosidase, same expression as amylase)
   - URE (urease, expressed as mg NH_3-N/g soil/3 hr)
   - ASP (asparaginase, expressed as mg NH_3-N/g soil/21 hr)
   - PHO (phosphatase, expressed as mg phenol/g soil/3 hr)

1. Interpretation through the Clustering Method

Table 3 and Fig. 1 show the clustering of the 32 samples. On step 3, six clusters appear, which have been called A, B, C, D, E, F. A contains 14 samples of northern and central forests. D contains 8 samples from southern forests. The remaining 10 samples are grouped into

the 4 clusters B, C, E, F. Moreover, the distance between clusters A and D is the highest observed: .43 (Table 4 and Fig. 1) and it may be observed that A and D constitute the two main clusters. Clusters C and F are near to cluster D as distances DC and DF are small (.09 and .13). Clusters B and E are at the same distance from A and D (.16, .20, .20, .19): they cover intermediate sites.

In addition, clustering suggests (1) that the time of sampling (November or March) has no influence, (2) that the geographic location is of importance but the assumed partition into three zones (northern, central, southern) is not satisfactory, (3) that, within the two main clusters A and D, the bedrock factor is the determining factor.

2. Interpretation through the Rank Correlation Method

Despite several shortcomings, correlation is of considerable use in ecology. The existence of a positive or negative interdependency between two variables does not prove causality; it can, however, support stimulating hypotheses for further studies, especially for experiments on models.

In most soils, it may be postulated that biological and humic characteristics are governed by two groups of causal factors: (1) chemical (sensu stricto) and physical soil-factors, (2) vegetation-factors. If that is so, the presence or absence of interdependency between the different variables may give rise to hypotheses on the role of the assumed factors.

Thus Table 5 induces the hypothesis that biological activity (saccharase excepted) is predominantly governed by the exchangeable cation status (S/ and T), total cation content (TCC), total Fe content (TFE), clay content (CLA). Humine (HU) and gray humic acids (GHA) seem also to be governed by one of these assumed factors, presumably clay; this hypothesis agrees with the known fact that gray humic acids are adsorbed on clay minerals and protected by these colloids from bacterial and enzymatic degradation. On the other hand it can be hypothesized that saccharase (SAC), fulvic (FA) and humic acids (HA) are essentially vegetation-dependent variables.

Another conclusion drawn from Table 5 is that the comparison of the 0-10 cm horizons of the different Ivory Coast forest soils can be easily performed through but a few determinations such as dehydrogenase (DES), saccharase (SAC), humic acids (HA), humine (HU), cation exchange capacity (T), clay content (CLA).

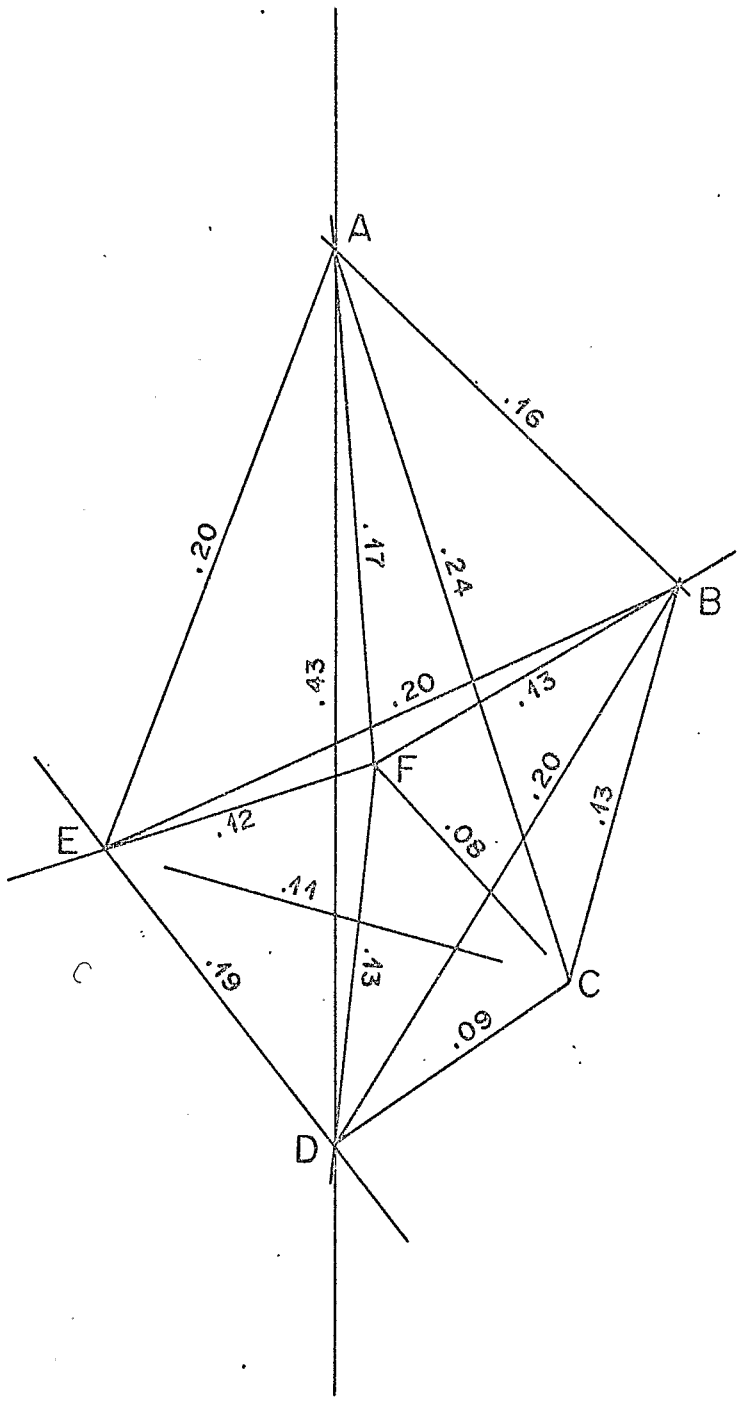

Fig. 1 Mean distances between arbitrary locations of clusters, step 3

Table 1 Sample numbers and sites (Ivory Coast forest soils)

Sample No	Sites	Rocks	Sample ref.	Month
1	Southern forest	Schists	A11	Nov.
2	Southern forest	Schists	A21	Nov.
3	Southern forest	Granites	B11	Nov.
4	Southern forest	Granites	B21	Nov.
5	Southern forest	Tertiary sands	C11	Nov.
6	Southern forest	Tertiary sands	C21	Nov.
7	Central forest	Schists	E11	Nov.
8	Central forest	Schists	D71	Nov.
9	Central forest	Granites	D31	Nov.
10	Central forest	Granites	F11	Nov.
11	Central forest	Amphibolites	G11	Nov.
12	Central forest	Amphibolites	G21	Nov.
13	Northern forest	Schists	N41	Nov.
14	Northern forest	Schists	N21	Nov.
15	Northern forest	Granites	K21	Nov.
16	Northern forest	Granites	M11	Nov.
17	Southern forest	Schists	A11	March
18	Southern forest	Schists	A21	March
19	Southern forest	Granites	B11	March
20	Southern forest	Granites	B21	March
21	Southern forest	Tertiary sands	C11	March
22	Southern forest	Tertiary sands	C21	March
23	Central forest	Schists	E11	March
24	Central forest	Schists	D71	March
25	Central forest	Granites	D31	March
26	Central forest	Granites	F11	March
27	Central forest	Amphibolites	G11	March
28	Central forest	Amphibolites	G21	March
29	Northern forest	Schists	N41	March
30	Northern forest	Schists	N21	March
31	Northern forest	Granites	K21	March
32	Northern forest	Granites	M11	March

Table 2 Ivory Coast forest soils: chemical and biological data (units in the text)

Sample No	Variables																			
	C/N	PH	CO_2	DES	SAC	AMY	GLU	URE	ASP	PHO	S/	T	TCC	TFE	TP	FA	HA	GHA	HU	CLA
1	12.8	4.9	2.45	10.8	80.7	34.1	24.6	.16	.59	.94	2.84	8.62	9.69	2.00	.30	2.93	1.46	48	8.04	11.8
2	13.3	4.5	1.78	3.27	56.6	24.5	15.2	.14	.35	.63	1.24	6.93	2.08	1.45	.23	2.32	1.22	47	5.01	8.8
3	15.0	4.5	1.30	6.13	29.7	29.7	17.6	.16	.27	.81	3.47	7.41	5.74	.78	.21	1.65	.70	42	6.24	8.8
4	13.1	4.9	1.79	6.90	26.2	24.5	21.5	.28	.33	.72	2.97	7.44	7.77	2.75	.52	2.63	1.21	45	8.81	19.5
5	16.3	3.9	1.08	5.71	15.2	20.7	11.9	.06	.22	.72	1.01	11.1	3.17	1.80	.41	3.01	3.16	50	7.90	9.1
6	12.2	4.4	1.11	4.30	15.1	15.1	10.9	.11	.23	.37	1.19	5.04	2.73	1.50		1.48	1.25	50	6.47	8.8
7	12.2	4.6	2.38	14.9	30.3	26.7	29.8	.29	.58	.92	5.57	12.6	11.0	4.75	.65	4.77	5.40	49	18.4	21.9
8	9.3	6.1	3.08	34.8	52.3	46.5	44.6	.52	.73	1.21	10.4	10.9	19.8	4.45	.78	3.01	1.78	57	16.2	21.3
9	10.0	5.4	2.06	17.9	24.7	29.7	30.0	.24	.49	1.05	6.04	8.36	12.3	1.00	.41	2.85	2.18	55	9.12	16.6
10	11.4	6.2	3.90	31.0	47.8	48.3	89.3	.49	.72	1.19	15.6	14.4	19.7	3.00	.54	3.01	1.85	58	17.3	23.2
11	9.2	7.4	5.22	21.9	56.5	62.7	31.7	3.44	1.30	1.65	66.4	45.0	159	9.25		2.71	1.06	58	43.7	41.9
12	9.4	7.5	5.16	15.9	58.0	60.3	24.3	3.07	1.13	1.64	73.6	49.6	145	10.5		3.37	1.29	52	46.6	41.4
13	9.9	5.8	2.96	28.9	47.8	48.3	72.1	.27	.61	1.12	11.3	13.2	28.9	5.25		3.12	3.02	65	20.2	26.4
14	10.7	5.8	5.34	16.5	41.0	35.5	36.3	.20	.70	1.03	7.90	10.3	25.1	3.50		3.48	3.30	61	15.7	23.8
15	14.1	6.3	1.11	40.1	74.2	37.0	43.1	.12	.37	.41	7.19	9.09	14.0	2.75	.40	.92	1.92	61	8.84	15.8
16	9.3	7.6	2.93	70.9	57.8	63.9	90.1	2.22	.77	1.13	28.4	20.0	40.3	4.25	.75	3.12	2.91	68	24.0	28.3
17	14.8	4.9	1.59	11.2	64.6	28.0	28.3	.17	.33	.92	3.02	8.67	9.69	2.00	.30	3.03	1.93	43	9.26	11.8
18	13.7	4.5	.76	5.06	36.7	15.2	15.0	.07	.24	.48	1.28	6.92	2.08	1.45	.23	2.16	1.61	47	5.76	8.8
19	13.4	4.7	.64	5.18	22.5	11.5	10.5	.09	.20	.60	2.84	6.95	5.74	.78	.21	1.56	.92	46	6.03	8.8
20	13.2	4.8	1.12	5.61	37.0	20.8	30.5	.08	.34	.79	4.32	9.00	7.77	2.75	.52	2.66	1.88	45	9.19	19.5
21	15.8	3.9	.90	1.02	17.2	26.2	29.3	.02	.25	.67	.94	10.2	3.17	1.80	.41	3.24	2.56	52	7.64	9.1
22	12.2	4.7	.68	1.52	17.6	17.0	14.3	.06	.20	.43	1.39	4.91	2.73	1.50		1.62	1.26	50	6.74	8.8
23	13.4	4.6	.89	9.19	27.8	43.0	37.1	.15	.30	.89	5.88	14.4	11.0	4.75	.65	6.28	5.39	52	20.3	21.9
24	10.5	5.9	2.00	21.2	59.4	33.4	41.1	.29	.75	1.03	10.1	11.5	19.8	4.45	.78	2.78	1.81	55	15.6	21.3
25	10.0	5.3	1.40	6.38	28.0	22.7	19.4	.26	.51	.84	5.86	9.10	12.3	1.00	.41	2.32	2.37	55	8.06	16.6
26	10.6	6.0	1.87	28.7	40.4	40.3	67.3	.34	.60	.99	13.7	13.9	19.7	3.00	.54	2.43	2.30	58	15.6	23.2
27	10.2	7.2	2.99	10.9	61.1	55.3	35.9	1.99	1.00	1.56	65.5	44.8	159	9.25		3.12	.91	61	44.0	41.9
28	10.1	7.2	3.43	8.34	56.2	50.2	18.5	2.56	.85	1.59	68.8	47.9	145	10.5		2.32	.70	56	42.0	41.4
29	12.3	5.9	1.75	22.7	52.0	50.2	51.1	.29	.62	1.11	10.5	13.7	28.9	5.25		3.29	3.04	64	21.2	26.4
30	11.1	5.6	2.72	16.5	44.8	44.9	48.2	.20	.69	.97	9.48	12.2	25.1	3.50		3.50	3.37	59	13.8	23.8
31	12.4	6.4	1.00	33.4	60.9	33.2	36.1	.09	.72	.59	6.37	7.81	14.0	2.75	.40	1.52	1.25	63	7.74	15.8
32	10.4	7.6	2.21	69.8	56.5	34.0	81.9	1.36	.83	1.09	30.6	23.0	40.3	4.25	.75	3.59	2.75	66	27.9	28.3

Table 3 Sample names and results of repeated clustering. Mean distances after step 3 appear in Table 4

Sites	Rocks	Sample ref.	Month	Sample No	Clusters Step 1	Step 2	Step 3
Central forest	Granites	F11	Nov.	10	]	]	]
Central forest	Granites	F11	March	26	]	]	]
Central forest	Schists	D71	Nov.	8	]	]	]
Central forest	Schists	D71	March	24	]	]	]
Northern forest	Schists	N21	Nov.	14	]	]	]
Northern forest	Schists	N21	March	30	]	]	]
Northern forest	Schists	N41	Nov.	13	]	]	] A
Northern forest	Schists	N41	March	29	]	]	]
Northern forest	Granites	M11	Nov.	16	]		]
Northern forest	Granites	M11	March	32	]		]
Central forest	Amphibolites	G11	Nov.	11	]	]	]
Central forest	Amphibolites	G11	March	27	]	]	]
Central forest	Amphibolites	G21	Nov.	12		]	]
Central forest	Amphibolites	G21	March	28		]	]
Central forest	Schists	E11	Nov.	7			] B
Central forest	Schists	E11	March	23			]
Southern forest	Schists	A11	Nov.	1		]	]
Southern forest	Schists	A11	March	17		]	] C
Southern forest	Granites	B21	Nov.	4		]	]
Southern forest	Granites	B21	March	20		]	]
Southern forest	Tertiary sands	C21	Nov.	6	]	]	]
Southern forest	Tertiary sands	C21	March	22	]	]	]
Southern forest	Schists	A21	March	18	]	]	]
Southern forest	Granites	B11	March	19	]	]	] D
Southern forest	Schists	A21	Nov.	2	]	]	]
Southern forest	Granites	B11	Nov.	3	]	]	]
Southern forest	Tertiary sands	C11	Nov.	5	]		]
Southern forest	Tertiary sands	C11	March	21	]		]
Northern forest	Granites	K21	Nov.	15	]		E
Northern forest	Granites	K21	March	31	]		E
Central forest	Granites	D31	Nov.	9	]		F
Central forest	Granites	D31	March	25	]		F

Table 4 Mean distances within clusters on the diagonal (underlined) and mean distances between clusters, step 3. Content of clusters in Table 3

Clusters	A	B	C	D	E	F
A	.06	.16	.24	.43	.20	.17
B	.16	.05	.13	.20	.20	.13
C	.24	.13	.06	.09	.11	.08
D	.43	.20	.09	.07	.19	.13
E	.20	.20	.11	.19	.01	.12
F	.17	.13	.08	.13	.12	.01

Table 5 Results of rank correlation test: N stands for a negative, P for a positive correlation; blank space for absence of correlation at the .01 probability level

	C/N	pH	CO_2	DES	SAC	AMY	GLU	URE	ASP	PHO	S/	T	TCC	TFE	TP	FA	HA	GHA	HU	CLA
C/N	⊙	N	N	N		N		N	N	N	N	N	N	N	N			N	N	N
pH	N	⊙	P	P	P	P	P	P	P	P	P	P	P	P				P	P	P
CO_2	N	P	⊙	P	P	P	P	P	P	P	P	P	P	P					P	P
DES	N	P	P	⊙	P	P	P	P	P	P	P	P	P	P	P			P	P	P
SAC		P	P	P	⊙	P			P		P		P	P						
AMY	N	P	P	P	P	⊙	P	P	P	P	P	P	P	P		P		P	P	P
GLU		P	P	P		P	⊙	P	P	P	P	P	P	P	P	P	P	P	P	P
URE	N	P	P	P		P	P	⊙	P	P	P	P	P	P	P			P	P	P
ASP	N	P	P	P	P	P	P	P	⊙	P	P	P	P	P	P			P	P	P
PHO	N	P	P	P		P	P	P	P	⊙	P	P	P	P	P	P		P	P	P
S/	N	P	P	P	P	P	P	P	P	P	⊙	P	P	P	P			P	P	P
T	N	P	P	P		P	P	P	P	P	P	⊙	P	P	P	P		P	P	P
TCC	N	P	P	P	P	P	P	P	P	P	P	P	⊙	P	P			P	P	P
TFE	N	P	P	P	P	P	P	P	P	P	P	P	P	⊙	P	P		P	P	P
TP	N			P			P	P	P	P	P	P	P	P	⊙	P			P	P
FA						P	P			P		P		P	P	⊙	P		P	P
HA							P									P	⊙			
GHA	N	P		P		P	P	P	P	P	P	P	P	P				⊙	P	P
HU	N	P	P	P		P	P	P	P	P	P	P	P	P	P	P		P	⊙	P
CLA	N	P	P	P		P	P	P	P	P	P	P	P	P	P	P		P	P	⊙

References

Aubry A.M. 1972. Programmes pour l'obtention des distances et des corrélations non-paramétriques en pédologie. O.R.S.T.O.M., Bondy, mimeogr., 54 pp.

Beyer W.H. (ed.) 1966. Handbook of tables for probability and statistics, Chemical Rubber, Cleveland, 517 pp.

Kendall M.G. & Stuart A. 1966. The advanced theory of statistics, vol. 3. Design and analysis, and time series, Griffin, London, 552 pp.

Spearman C. 1904. The proof and measurement of association between two things. Am. J. Psych., 15: 72–101.

Van den Driessche R. & Garcia Gomez A. 1973. Distances non-paramétriques entre profils. Rev. Ecol. Biol. Sol 9: 617–628.

Discussion

J.S. Waid: I wondered whether you, in your last example, have found any seasonal correlations.
Y. Dommergues: For the characteristics which were investigated, no seasonal variations could be detected, neither in the first example (presented in this paper), nor in the second one (survey of the Lamto IBP station soils).
V. Sundman: Did you make any efforts to correlate data on soils, climate and vegetation. You started considering all three types of data, but showed only correlations between data on soils?
Y. Dommergues: In the case of the example presented here, only soil characteristics have been investigated. But both clustering and rank correlation programs can be used for interpreting not only data on soils but also, and simultaneously, data on climate and vegetation.
J. Remacle: How do you explain that there is no correlation between enzyme activities and clay content in the second example?
Y. Dommergues: In the second example (related to a survey of the Lamto IBP station, Ivory Coast, which is to be published in another paper), the absence of correlation may be attributed to the low and very irregular clay content of the soil samples.