REVUE DE STATISTIQUE APPLIQUÉE

GEORGES ROUX
MAURICE ROUX

À propos de quelques méthodes de classification en phytosociologie

Revue de statistique appliquée, tome 15, n° 2 (1967), p. 59-72
<http://www.numdam.org/item?id=RSA_1967__15_2_59_0>

A PROPOS DE QUELQUES MÉTHODES DE CLASSIFICATION EN PHYTOSOCIOLOGIE

Georges ROUX
Biologie végétale, Orsay

Maurice ROUX
Laboratoire de Calcul, ISUP, Paris

INTRODUCTION

L'étude des groupements végétaux (ou phytosociologie) d'une région commence par l'établissement de "relevés" faits en des points les plus nombreux possibles de cette région. Ces relevés portent chacun sur une surface de l'ordre de quelques mètres carrés (Cf. Guinochet, 1955, pour la délimitation des surfaces de relevés) et consistent à noter, soit uniquement les noms des espèces présentes dans la surface considérée, soit les espèces présentes accompagnées d'un coefficient dit d'abondance-dominance (Cf. Braun-Blanquet) représentant la proportion de cette espèce dans la surface étudiée. Ces relevés sont alors rassemblés en un tableau à double entrée, que nous appellerons matrice des données, où l'usage veut que l'on réserve une colonne par relevé et une ligne par espèce.

Désignons par t_{ij} la valeur portée à l'intersection de la ligne i et de la colonne j de ce tableau. Si l'on a simplement noté les présences, $t_{ij} = 1$ ou 0 suivant que l'espèce i est présente ou absente dans le relevé j. Si l'on a noté les abondances, t_{ij} représente alors l'abondance de l'espèce i dans le relevé j. (Cf. Braun-Blanquet, 1932).

Le travail de classification porte sur les colonnes de cette matrice des données, c'est-à-dire sur les relevés. Pour faire ce travail de nombreux auteurs ont proposé divers "indices de similarité", ou coefficients de ressemblance, entre colonnes. Notre but était de calculer ces différents indices pour un même ensemble de 55 relevés, comportant 174 espèces, faits en 1965 au Plateau d'Emparis (Hautes-Alpes) à l'instigation de M. Guinochet (Biologie Végétale, SPCN, Orsay). Puis pour chaque indice nous voulions essayer de construire une figure plane représentant au mieux cet ensemble de relevés, compte-tenu des proximités ainsi calculées, espérant ainsi faire apparaître visuellement les divers groupes de la classification à établir. Enfin, comparer les résultats entre eux d'une part, avec la réalité "botanique" d'autre part.

Dans une première partie nous présenterons les résultats afférents aux indices calculés à partir des présences. Puis, nous parlerons d'un indice calculé à l'aide des coefficients d'abondance. Enfin nous réserverons une place à part à l'analyse factorielle.

Avant d'exposer ces résultats, disons un mot de nos méthodes. Nous avons établi des programmes-machine permettant de calculer les divers indices dont nous avions besoin ; pour que les résultats, fort volumineux $\left(\frac{55 \times 54}{2} = 1\,485 \text{ proximités}\right)$, soient utilisables, nous avons

fait en sorte qu'outre la matrice des similarités, tableau dans lequel on lit la proximité entre les relevés i et j à l'intersection de la ligne i et de la colonne j, la machine inscrive aussi les coefficients calculés par ordre de grandeur croissante, en rappelant au-dessous de chaque valeur les deux relevés dont cette valeur représente la proximité. C'est ainsi que nous obtenons "l'ordonnance" de l'ensemble des relevés, c'est-à-dire la relation d'ordre sur les paires de relevés (Cf. Benzécri, 1966).

Rappelons que d'après les travaux de Shepard (1962) on peut reconstruire (à une similitude près) la figure formée par un ensemble de points dont on connaît l'ordonnance, pourvu que ceux-ci soient assez nombreux.

D'autre part, B. Cordier s'est très obligeamment chargée de l'analyse factorielle de 53 de nos relevés, deux autres relevés ayant été adjoints par la suite pour former les 55 relevés dont nous parlions ci-dessus.

1 - INDICES UTILISANT LES COEFFICIENTS DE PRESENCE-ABSENCE

On peut ranger en deux catégories les indices utilisant des matrices de données ne comportant que des zéros ou des 1 :

1/ Les indices dans le calcul desquels on considère uniquement comme coïncidences entre deux relevés K et L les lignes où il y a un 1 dans chacune des deux colonnes K et L ; nous appellerons n_{KL} le nombre de ces coïncidences (c'est-à-dire le nombre d'espèces communes aux deux relevés) ;

2/ Les indices pour lesquels les coïncidences entre K et L sont à la fois les lignes où il y a un 1 dans les colonnes K et L, et les lignes où il y a un zéro dans chacune de ces deux colonnes (c'est-à-dire qu'une absence commune aux 2 relevés considérés est comptée au même titre qu'une présence commune), nous noterons n_{kl} le nombre d'absences communes aux deux relevés K et L (Cf. Benzécri, 1966, pour cette notation).

Nous poserons aussi :

$m_{kl} = n_{kl} + n_{KL}$ (m = "matched")

et u_{kl} = le nombre de lignes pour lesquelles les colonnes K et L sont différentes (u = "unmatched").

1.1 - Indices utilisant n_{KL}.

Ce sont les plus nombreux. Parmi ceux-ci, on trouve :

Jaccard (1908) $\quad a_{kl} = \dfrac{n_{KL}}{n_{KL} + u_{kl}}$

Dice (1945), Sorens en (1948) $\quad b_{kl} = \dfrac{2n_{KL}}{2n_{KL} + u_{kl}}$

Kulczinsky (1927) (1) $\quad c_{kl} = \dfrac{n_{KL}}{u_{kl}}$

Kulczinsky (1927) (2) $\quad d_{kl} = \dfrac{n_{KL}}{2}\left[\dfrac{1}{n_K} + \dfrac{1}{n_L}\right]$

Ochiai (1957) $$e_{kl} = \frac{n_{KL}}{\sqrt{n_K n_L}}$$

Sokal et Sneath $$f_{kl} = \frac{n_{KL}}{n_{KL} + 2u_{kl}}$$

Nous notons n_K le nombre total d'espèces présentes dans le relevé K.

Nous allons montrer d'abord que les indices a, b, c et f donnent exactement la même ordonnance.

Supposons que l'on ait $a_{ij} \leqslant a_{kl}$ (1), c'est-à-dire que les relevés i et j sont plus proches entre eux, au sens de Jaccard, que les relevés k et l.

Alors (1) $\Longleftrightarrow$ (équivalente à) $\frac{n_{IJ}}{n_{IJ} + u_{ij}} \leqslant \frac{n_{KL}}{n_{KL} + u_{kl}}$

$$\Longleftrightarrow \quad \frac{1}{1 + \frac{u_{ij}}{n_{IJ}}} \leqslant \frac{1}{1 + \frac{u_{kl}}{n_{KL}}} \quad \text{si} \quad n_{IJ} \neq 0 \quad \text{et} \quad n_{KL} \neq 0$$

$$\Longleftrightarrow \quad \frac{u_{ij}}{n_{IJ}} \geqslant \frac{u_{kl}}{n_{KL}} \quad (2)$$

$$\Longleftrightarrow \quad \frac{2}{2 + \frac{u_{ij}}{n_{IJ}}} \leqslant \frac{2}{2 + \frac{u_{kl}}{n_{KL}}}$$

$$\Longleftrightarrow \quad \frac{2n_{IJ}}{2n_{IJ} + u_{ij}} \leqslant \frac{2n_{KL}}{2n_{KL} + u_{kl}}$$

$$\Longleftrightarrow \quad b_{ij} < b_{kl}$$

c'est-à-dire que les relevés i et j sont plus proches entre eux, au sens de Dice, que les relevés k et l.

De même (2)

$$\Longleftrightarrow \quad \frac{n_{IJ}}{u_{ij}} < \frac{n_{KL}}{u_{kl}}$$

$$\Longleftrightarrow \quad c_{ij} < c_{kl}$$

même conclusion que précédemment mais au sens de Kulczinsky (1)

enfin (2) $$\Longleftrightarrow \quad \frac{1}{1 + \frac{2u_{ij}}{n_{IJ}}} < \frac{1}{1 + \frac{2u_{kl}}{n_{KL}}}$$

$$\Longleftrightarrow \quad \frac{n_{IJ}}{n_{IJ} + 2u_{ij}} < \frac{n_{KL}}{n_{KL} + 2u_{kl}}$$

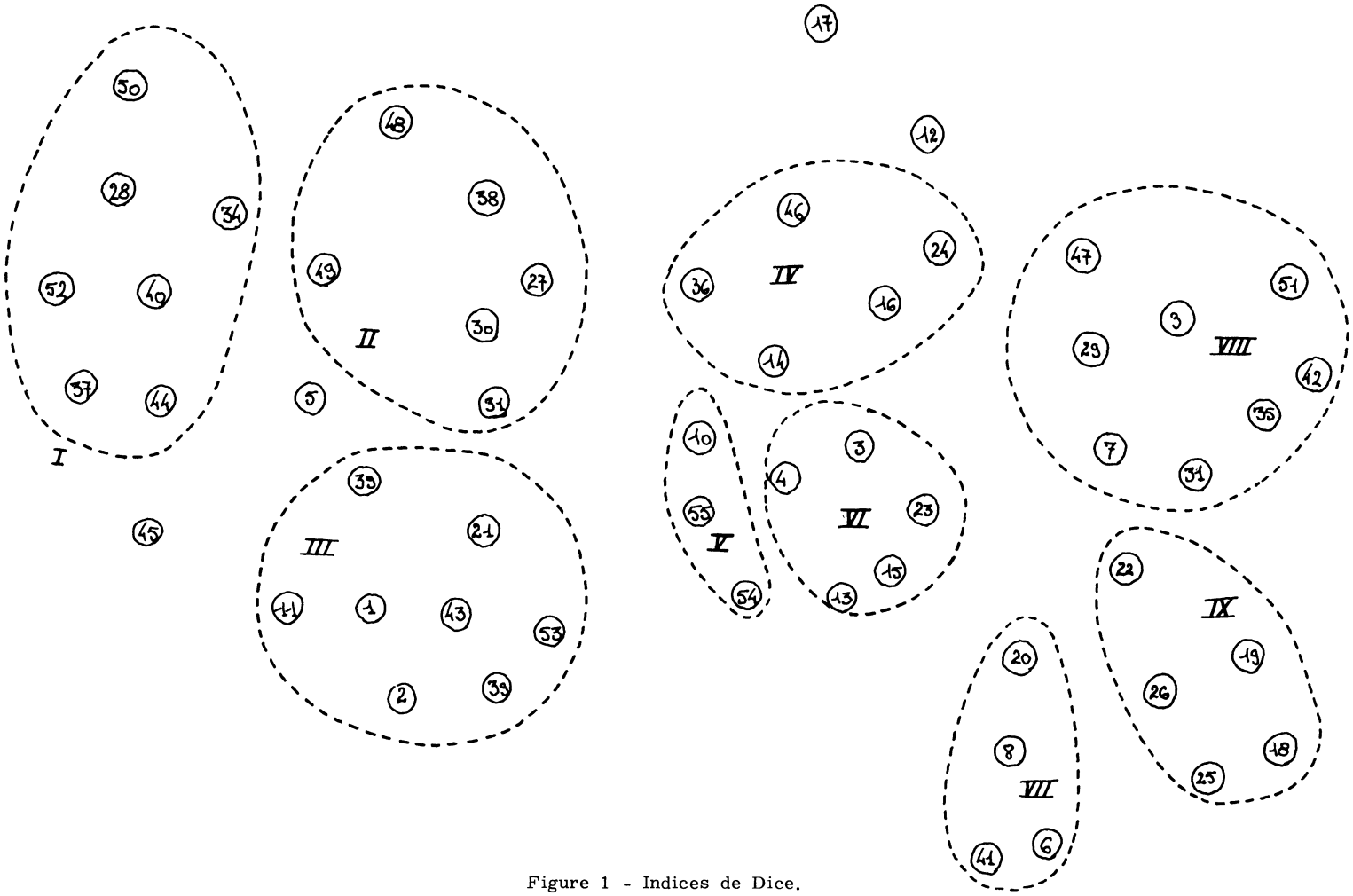

Figure 1 - Indices de Dice.

même conclusion que ci-dessus, mais au sens de Sokal et Sneath. Pour cette raison nous n'avons calculé qu'un seul de ces quatre indices : celui de Dice (1945), noté b ici. Ce calcul nous a permis de construire la figure 1 ci-contre, dans laquelle les groupes II, VII et IX se dessinaient dès le premier abord. Les groupes I, II et VIII ont été établis par l'examen direct des relevés correspondants et des espèces y contenues. Enfin les séparations entre les groupes IV, V et VI n'ont pu être obtenues que par le procédé qui consiste à "visualiser" la matrice de corrélation en déplaçant les lignes (et corrélativement les colonnes) de façon à obtenir sur la diagonale principale les plus fortes similarités (figure 2). Les 9 groupes ainsi obtenus correspondent bien à des associations végétales classiques. Restent les relevés 45, 12 et 17 qui n'ont pu être placés dans aucun groupe : manquent-ils d'"homogénéité"

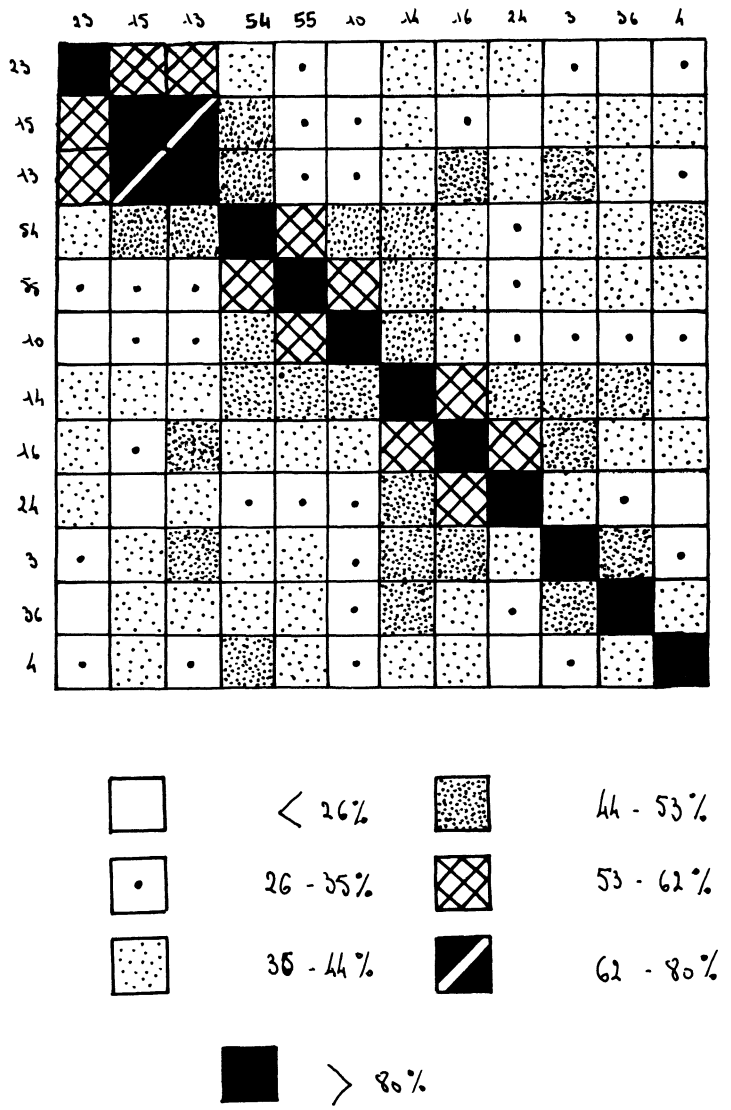

Figure 8

Analyse différentielle de Czekanowski

(la surface qui les délimite serait alors mauvaise) ou sont-ils, au contraire, les témoins de l'existence d'associations intermédiaires stables ? Leur rareté ne nous semble pas permettre de conclure.

Nous avons fait, en outre les quelques remarques suivantes : $2n_{KL} + u_{kl}$, dénominateur des indices b, s'écrit aussi $n_K + n_L$. Donc b_{kl} se calcule en divisant n_{KL}, le nombre d'espèces communes à K et L, par la moyenne arithmétique de n_K et n_L, nombre d'espèces présentes dans K et L, respectivement. D'autre part e_{kl}, indice d'Ochiai, se calcule en divisant n_{KL} par la moyenne géométrique de n_K et n_L, tandis que d , indice (2) de Kulczinsky, est obtenu en divisant n par la moyenne harmonique de n_K et n_L. La moyenne harmonique étant toujours inférieure à la moyenne géométrique, elle-même toujours inférieure à la moyenne arithmétique, les indices correspondants seront dans l'ordre inverse :

$$b_{kl} < e_{kl} < d_{kl}$$

Comme ces 3 moyennes sont peu différentes entre elles, pour des nombres assez peu différents, comme c'est notre cas (min $n_J = 9$, max $n_J = 35$) les indices b, e et d doivent donner des résultats à peu près concordants. C'est bien ce que nous avons constaté, puisque nous avons calculé ces trois indices. Les plus grandes discordances ont lieu pour des paires de relevés où les présences de l'un sont de beaucoup inférieures aux présences de l'autre, comme on pouvait s'y attendre car c'est dans un tel cas que les moyennes géométriqueset harmonique sont sensiblement inférieures à la moyenne arithmétique.

Examinons, par exemple, la similarité entre les relevés 44 et 49. Au sens de Dice on a $b_{44.49} = 0,4138$ ce qui place cette paire au 242e rang au sens de Kulczinsky (2) $d_{44.49} = 0,4833$ ce qui place cette paire au 149e rang (nous comptons les rangs de l'ordonnance "à l'envers" en donnant le rang 1 à la paire formée des deux relevés les plus proches). Il y a donc un décalage de près d'une centaine de rangs entre les deux manières de calculer pour ce couple particulier de relevés. Cela ne serait pas grave s'il s'agissait d'un couple de relevés peu ressemblants, car leur considération n'entrerait pas en ligne de compte pour l'élaboration de la figure représentative de ces proximités. Mais ce n'est pas le cas ! La construction de cette figure nécessite l'examen de l'ordonna ce depuis la fin jusqu'au 200e rang environ. Il y a plus ! Regardons les relevés 31 et 49.

Nous avons $b_{31.49} = 0,4255$ ce qui place cette paire au 219e rang,
et $d_{31.49} = 0,4352$ ce qui place cette paire au 216e rang.

Quand on passe d'une méthode de calcul à l'autre il y a donc <u>interversion</u> des couples 44.49 et 31.49 : c'est-à-dire que pour Dice 31.49 est un couple plus "grand" (relevés plus similaires) que 44.49 alors que pour Kulczinsky (2) c'est 44.49 qui est plus grand que 31.49.

Or sur la figure 1, qui serre de près la réalité, avons-nous dit, 31 et 49 appartiennent tous deux au groupe II alors que 44 est un relevé du groupe I, dans ce cas particulier Dice semblerait donc donner de meilleurs résultats que Kulczinsky (2).

Mais nous avons découvert un autre exemple dans lequel deux paires se trouvent interverties lorsque l'on passe d'un indice à l'autre :

On a :

$$b_{7.51} = 0,3830 \quad \text{et} \quad d_{7.51} = 0,4052$$

tandis que

$$b_{7.10} = 0,3922 \quad \text{et} \quad d_{7.10} = 0,3997$$

On voit que 7.10 est une paire plus grande que 7.51 au sens de Dice alors qu'au sens de Kulczinsky (2) c'est 7.51 que est plus grand que 7.10. Et dans ce cas c'est Kulczinsky (2) qui serait plus dans le vrai puisque 7 et 51 appartiennent au même groupe VIII, tandis que 10 est un relevé du groupe V.

Il est donc difficile de dire lequel est le meilleur (pour le but poursuivi) de ces deux indices, à plus forte raison lorsque l'on envisage l'indice d'Ochiai qui est intermédiaire entre les deux précédents. Toutefois nous penchons pour Dice, qui, s'il est peu net pour le groupe VIII, nous a permis de le découvrir quand même, alors que Kulczinsky (2) "mélange" les groupes I et II comme le montre la figure 3 élaborée à partir de cet indice.

1.2 - Indices utilisant m_{kl}.

Parmi ceux-ci se rangent, entre autres,

Sokal et Michener : $g_{kl} = \dfrac{m_{kl}}{n}$

Sokal et Sneath : $h_{kl} = \dfrac{2m_{kl}}{m_{kl} + n}$

(n est le nombre total d'espèces répertoriées dans le tableau des données). L'indice de Rogers et Tanimoto (1960) dont la formule est $\dfrac{m_{kl}}{m_{kl} + u_{kl}}$ est identique à celui de Sokal et Michener (SM, en abrégé) puisque $m_{kl} + u_{kl} = n$. De plus SM et Sokal et Sneath donnent la même ordonnance. Si $g_{ij} \leqslant g_{kl}$ celà veut dire que l'on a :

$$\frac{m_{ij}}{n} \leqslant \frac{m_{kl}}{n} \Longleftrightarrow \frac{n}{m_{ij}} \geqslant \frac{n}{m_{kl}} \qquad (3)$$

Or :

$$h_{ij} = \frac{2}{1 + \dfrac{n}{m_{ij}}} \quad \text{et} \quad h_{kl} = \frac{2}{1 + \dfrac{n}{m_{kl}}}$$

donc $(3) \Longleftrightarrow h_{ij} \leqslant h_{kl}$.

Parmi ces trois indices nous n'avons donc calculé que SM. Ces calculs ont révélé au premier coup d'œil des faits assez surprenants, par comparaison avec ceux donnés par Dice, ou par Jaccard qui sont les mêmes, rappelons-le, du point de vue de l'ordonnance.

On avait par exemple :

pour S.M. $g_{18.52} = 0,8851$ ce qui place cette paire au 68ème rang de l'ordonnance et pour Dice $b_{18.52} = 0,0$ ce qui place cette paire au dernier rang (1485ème) de l'ordonnance, (nous comptons toujours les rangs à partir de la fin, c'est-à-dire que $b_{18.52}$ se trouve au premier rang où

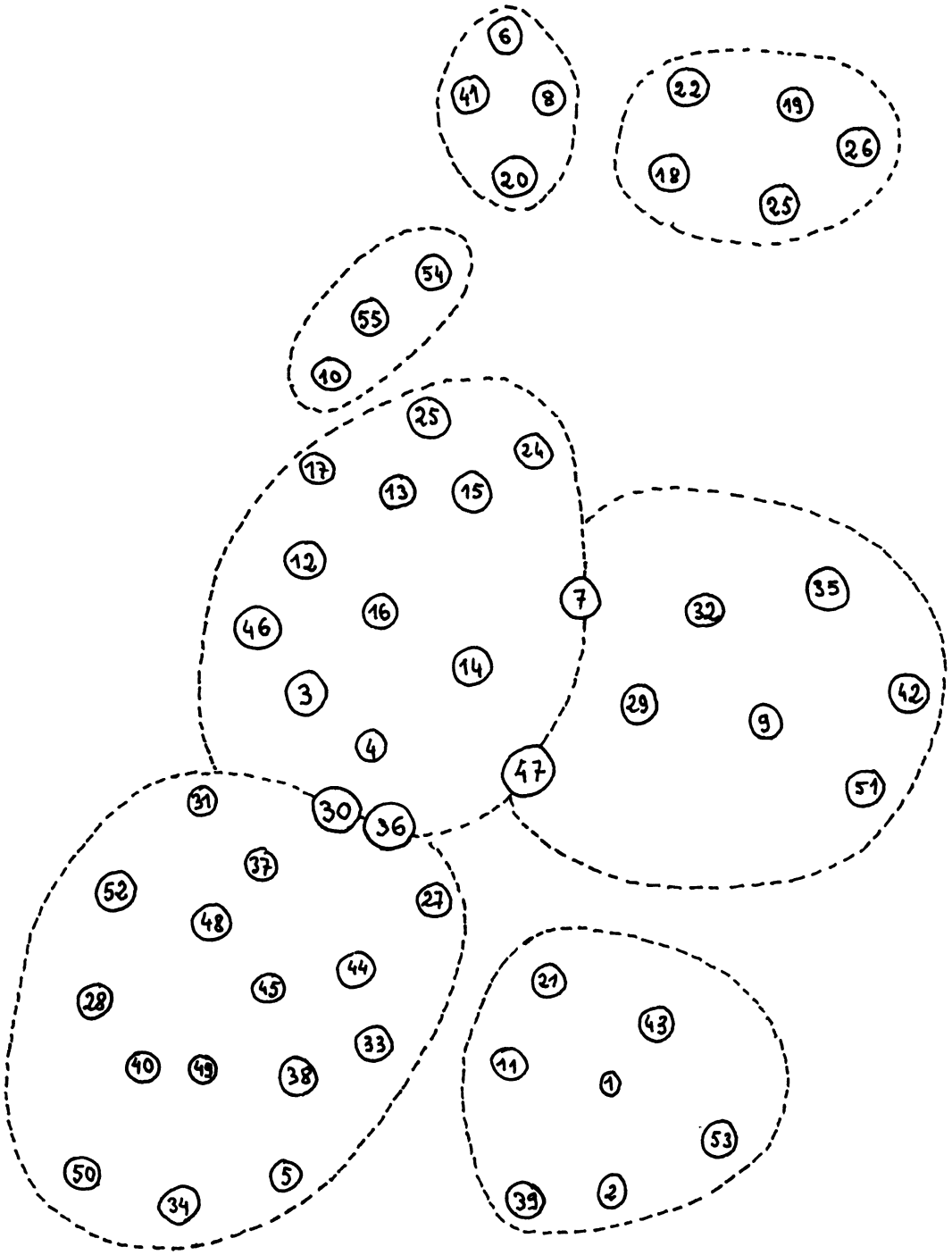

Figure 3 - Indices de Kulczinsky (2).

sont donc les proximités nulles qui signifient distances les plus grandes, entre relevés). De même pour S.M. $g_{41.44} = 0,8736$ ce qui place cette paire au 115ème rang et pour Dice $b_{41.44} = 0,0833$ ce qui place cette paire au 1316ème rang !

Cela s'explique facilement :

$$b_{ij} = \frac{2n_{IJ}}{2n_{IJ} + u_{ij}} = 1 - \frac{u_{ij}}{2n_{IJ} + u_{ij}}$$

$$g_{ij} = \frac{n_{IJ} + z_{ij}}{n_{IJ} + u_{ij} + z_{ij}} = 1 - \frac{u_{ij}}{n}$$

(z_{ij} désigne le nombre d'espèces absentes simultanément dans i et j).

Supposons que les relevés i et j comportent peu d'espèces et aient peu d'espèces communes, alors le nombre $2n_{IJ}+u_{ij}$, égal à $n_I + u_J$ somme des présences des deux relevés, sera faible puisque n_I et n_J sont faibles, donc $\frac{u_{ij}}{2n_{IJ} + u_{ij}}$ sera voisin de 1 et b_{ij} voisin de zéro ce qui est normal puisque nous avons supposé que i et j ont peu d'espèces communes. Par contre u_{ij} étant faible g_{ij} sera voisin de 1. Cela est dû au fait que pour SM nos deux relevés se ressemblent par leurs nombreuses absences communes ; mais cela contredit l'expérience. M. Guinochet, consulté, a bien voulu nous le confirmer.

2 - INDICE UTILISANT LES COEFFICIENTS D'ABONDANCE-DOMINANCE

Le seul indice de cette catégorie que nous ayons calculé est encore dû, selon Dagnelie, à Kulczinsky. Nous le noterons Kulczinsky (3). La formule qui le donne est la suivante :

$$p_{kl} = \frac{\sum_{i=1}^{n} F_{ikl}}{\sum_{i=1}^{n} t_{ik} + t_{il}}$$

où t_{ij} est l'abondance de l'espèce i dans le relevé j et $F_{ikl} = \min(t_{ik}, t_{il})$.

Ce calcul nous a donné des résultats que l'on peut considérer comme bons quoiqu'il semble dessiner des séparations plus floues entre certains groupes, en particulier les groupes I, II et III (figure 4).

Nous ferons remarquer, avec Dagnelie, que cet indice donne les mêmes résultats "en sens inverse" que le paramètre introduit par Odum (1950) sous le nom de "percentage difference", dont la formule est la suivante :

$$s_{kl} = \frac{\sum_{i=1}^{n} |t_{ik} - t_{il}|}{\sum_{i=1}^{n} (t_{ik} + t_{il})}$$

(Le numérateur est la somme des valeurs absolues des différences des abondances des relevés k et 1).

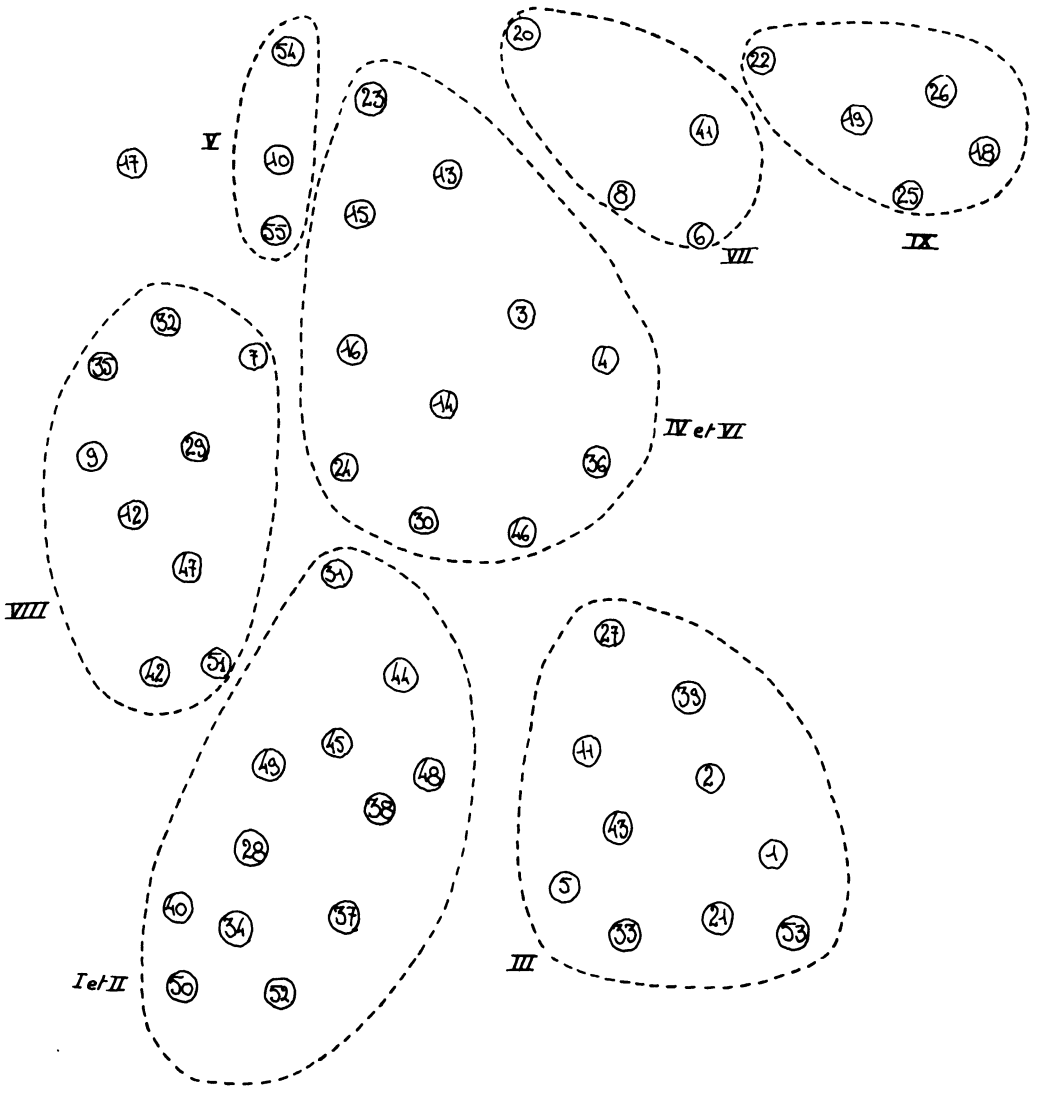

Figure 4 - Indices de Kulczinsky (3).

On a, en effet, $|t_{ik} - t_{il}| = t_{ik} + t_{il} - 2F_{ikl}$

donc : $$\sum_{i=1}^{n} |t_{ik} - t_{il}| = \sum_{i=1}^{n} (t_{ik} + t_{il}) - \sum_{i=1}^{n} F_{ikl}$$

et $$s_{kl} = 1 - p_{kl} \ .$$

Le paramètre d'Odum est donc un coefficient de dissemblance alors que celui de Kulczinsky (3) est un coefficient de ressemblance. Mais, d'après cette dernière formule leurs résultats seront exactement inversés, du point de vue de l'ordonnance.

3 - L'ANALYSE FACTORIELLE

Nous avons fait deux analyses factorielles de nos données, l'une à partir des coefficients présence-absence, l'autre à partir des coefficients d'abondance, suivant la thèse de B. Cordier : "Sur l'analyse factorielle des correspondances". Cette méthode est très voisine de "l'analyse en composantes principales" ("The principal axes method" dans la littérature de langue anglaise) qui consiste essentiellement en une extraction des valeurs propres les plus grandes, en valeur absolue, et des vecteurs propres correspondants, de la matrice de corrélation.

Dans cette théorie on considère, en effet, les colonnes (respectivement : les lignes) de la matrice $n \times p$, des données comme un ensemble de réalisations d'une variable aléatoire multivariate à n dimensions (resp. : à p dimensions). On cherche à rendre compte "au mieux" du phénomène à étudier, à l'aide du plus petit nombre possible de variables aléatoires (univariates) non corrélées, par une rotation des axes de référence. Ce qui distingue cette méthode des travaux antérieurs c'est l'utilisation d'une nouvelle forme quadratique pour le calcul des corrélations qui permet d'établir une identification entre les deux ensembles entrant dans la composition des données, ensembles formés pour nous des espèces, d'une part, et des relevés de l'autre.

Les résultats, contrairement à ce que nous attendions, sont meilleurs avec les coefficients de présence-absence comme on peut le voir sur les figures 5 et 6. Cela semble confirmer l'hypothèse de M. Guinochet suivant laquelle la présence d'une espèce dans un relevé constitue un caractère "génétique" pour ce relevé tandis que l'abondance est d'ordre "morphologique".

Le gros avantage que nous voyons dans la méthode d'analyse factorielle est qu'elle donne directement la figure représentative de l'ensemble à classer et ce avec une totale objectivité évidemment. L'un des défauts communs aux modes de constructions par tatonnement à partir de l'ordonnance est en effet que l'intervention humaine est toujours plus ou moins influencée par ce que l'on croit savoir par ailleurs des relevés à classer.

CONCLUSION

En ce qui concerne le traitement des données en présence-absence, nos préférences vont vers la famille formée des indices de Jaccard, Dice, Kulczinsky (1), etc. Les petites démonstrations que nous avons faites à ce propos nous montrent d'ailleurs que c'est finalement le rapport $\frac{n_{JK}}{\mu_{jk}}$, c'est-à-dire concordances "divisées" par discordances qui donne l'ordonnance. (C'est, de plus, ce qui est le plus facile à calculer !).

En ce qui concerne le traitement des données en abondance-dominance, quoique les résultats obtenus soient satisfaisants avec l'indice de Kulczinsky (3), nous faisons une réserve quant à l'objectivité dans l'établissement de la matrice des données, l'évaluation de l'abondance d'une espèce étant tout de même fonction du "coup d'œil" de l'expérimentateur. Et, à propos d'objectivité, nous ne saurions trop louer la méthode d'analyse factorielle qui présente à ce point de vue toutes les garanties que l'on est en droit d'exiger !

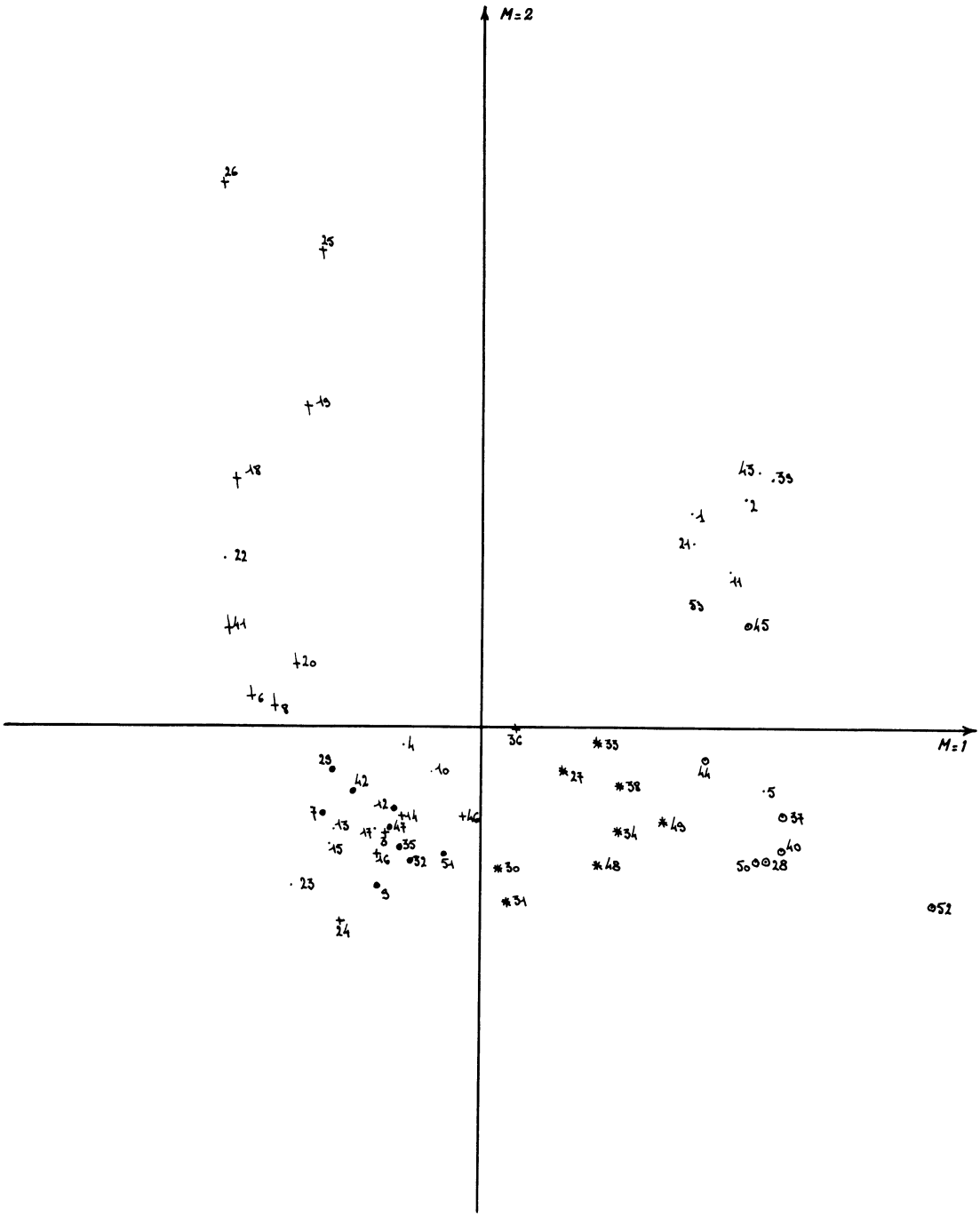

Figure 5 - Résultats de l'analyse factorielle à partir des coefficients de présence-absence.

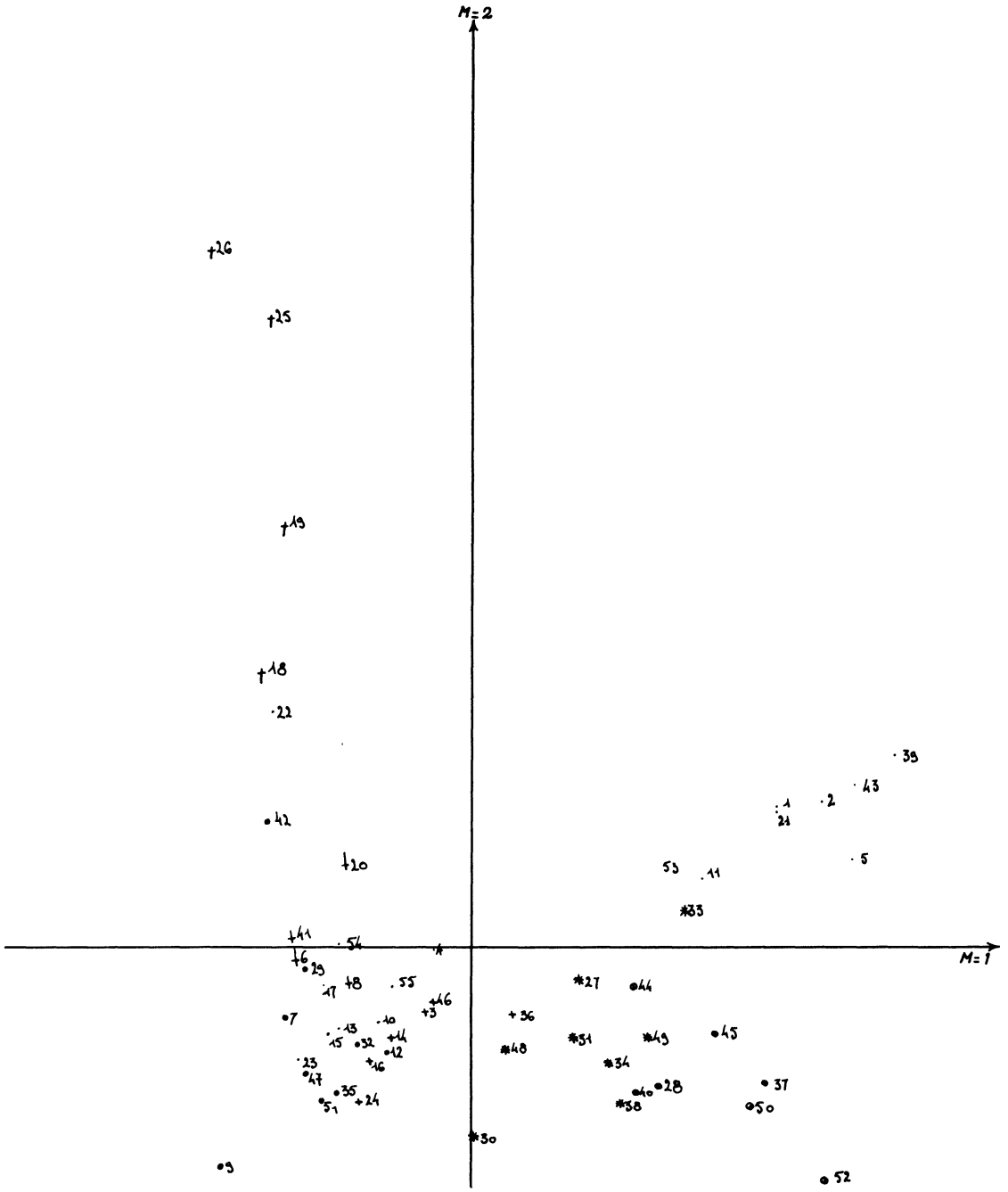

Figure 6 - Résultats de l'analyse factorielle à partir des coefficients d'abondance.

BIBLIOGRAPHIE SOMMAIRE

J.P. BENZECRI - Sur l'analyse factorielle des proximités. ISUP. Paris. 1964.

J.P. BENZECRI - Leçons sur l'analyse factorielle et la reconnaissance des formes. Cours de 3ème cycle ISUP. Paris. 1966.

J. BRAUN-BLANQUET, G.D. FULLER, H.S. CONRAD - Plant sociology - The study of plant communities (Authorized English Translation of Planzensoziologie) New-York and London. 1932.

B. CORDIER - Sur l'analyse factorielle des correspondances. Thèse. Rennes. 1965.

P. DAGNELIE - Contribution à l'étude des communautés végétales par l'analyse factorielle. Bull. Serv. Carte Phytogéog. (B) T5 pp. 7-71, 93-195 ; 1960.

M. GUINOCHET - Logique et dynamique du peuplement végétal pp. 61-72. Masson. Paris. 1955.

A. OCHIAI - Zoogéographic studies on the soleoid fishes found in Japan and its neighbouring regions. Bull. Jap. Soc. Sci. Fish. T22, pp. 526-530. 1957.

E.P. ODUM - Bird populations of the Highlands (North Carolina). Plateau in relation to plant succession and avian invasion. Ecology 31 pp. 587-605. 1950.

R.N. SHEPARD - The analysis of proximities : scaling with an unknown distance function. I and II ; Psychometrika. 1962.

R.R. SOKAL et P.M.A. SNEATH - Principles of numerical taxonomy. - W.H. Freeman and Co : San Francisco and London. 1963.