Zeitschrift: Boissiera : mémoires de botanique systématique
Band: 28 (1978)

Artikel: Etude taxonomique d'un groupe complexe d'espèces des genres Phaseolus et Vigna (Papilionaceae) sur la base de données morphologiques et polliniques, traitées paar l'analyse informatique
Kapitel: Les méthodes informatiques
Autor: Maréchal, Robert / Stainier, Françoise / Mascherpa, Jean-Michel
DOI: https://doi.org/10.5169/seals-895590

4. Les méthodes informatiques

4.1. Introduction

"La taxonomie numérique est le groupement par des méthodes numériques d'unités taxonomiques en taxons, sur la base de l'étude de leurs caractères" (SNEATH & SOKAL, 1973: 4).

Cette définition implique plusieurs remarques. Il faut à partir des données, ou caractères observés, transformer l'information qu'elles contiennent en des quantités numériques. Puis, il est nécessaire de trouver un algorithme, une statistique, qui tire de ces données, le maximum de l'information recueillie, et la traduise en termes de proximité ou de phylogénie.

Suivant les principes de Michel ADANSON (1757), Sneath & Sokal ont posé les principes de la taxonomie numérique:

- plus l'information contenue dans les taxons d'une classification est grande, plus sont nombreux les caractères sur lesquels elle est basée, meilleure sera la classification,
- à priori, chaque caractère a un poids égal dans l'élaboration des taxons,
- la ressemblance entre deux entités est une fonction de leur propre similarité en chacun des caractères où elles sont comparées,
- on reconnaît des taxons différents à ce que les corrélations des caractères diffèrent dans les groupes d'organismes étudiés,

– des déductions phylogéniques peuvent être tirées des structures taxonomiques d'un groupe et de la corrélation des caractères, fournissant des suppositions sur l'évolution des êtres,

– la taxonomie est conçue et appliquée comme une science empirique,

– les classifications sont basées sur des ressemblances phénotypiques.

Ces sept principes fondamentaux de la taxonomie numérique fournissent l'algorithme de travail du systématicien, c'est-à-dire l'ordre dans lequel il doit entreprendre sa recherche: observation des unités taxonomiques et leur quantification, établissement des mesures de ressemblance sur lesquelles il établira les nouveaux taxons, et enfin, généralisations sur les taxons, choix des nouveaux caractères à observer, leur pondération "à posteriori" par leur pouvoir discriminant, recherche des liens phylogéniques.

La méthode de travail que nous avons utilisée suit les principes fondamentaux de la taxonomie numérique exposés plus haut. Elle peut se résumer selon l'organigramme présenté à la figure 34 (cf. aussi MASCHERPA, 1976).

Etude du problème: on peut engager un programme de taxonomie numérique dans deux directions principales. On peut envisager que les groupes naturels aient été soigneusement étudiés, mais qu'il subsiste des points délicats, soit que plusieurs espèces aient une position taxonomique ambiguë, soit que l'adjonction de nouveaux spécimens remette en jeu la classification existante. C'est souvent le cas où la description de caractères uniquement morphologiques n'est pas assez précise. Le spécialiste doit alors faire appel à des moyens d'investigation plus fins: cytologie, palynologie, microscopie électronique, génétique, biochimie, etc. Mais très souvent, les résultats sont trop nombreux pour qu'il soit possible d'en tirer "à la main" quelque chose de positif. Ce sera aussi la position à prendre lorsque plusieurs classifications auront été proposées par différents auteurs, et qu'aucune d'elles ne soit parfaitement satisfaisante. Alors, l'objectivité de la méthode et de l'appréciation des caractères permettra une tentative de clarification. A notre idée, ce sont les cas les plus intéressants d'utilisation de l'informatique. Le problème étant posé, le module 2 de l'organigramme est résolu, car ce qui importe ici n'est pas tant de choisir des caractères à observer, que d'essayer d'en obtenir le plus possible, selon le principe 1 de la taxonomie numérique. Le module primordial est le cinquième. C'est à ce stade que doit commencer à s'élaborer la hiérarchisation des caractères, puis la recherche de la nouvelle systématique.

La seconde méthode consiste à envisager la taxonomie comme moyen de classification d'individus nouveaux. Nous pensons ici aux résultats qui peuvent être obtenus lorsque les individus sont représentés en majorité par des variétés issues de sélections nouvelles. La systématique des nouvelles introductions est en effet un point très actuel de la taxonomie numérique, les stations de recherches appliquées, horticoles ou agronomiques, ayant entrepris des travaux de plus en plus nombreux. Souvent les variétés sélectionnées ne diffèrent que par des caractères masqués: présence de résistance aux différentes maladies, de nouvelles protéines ou de nouveaux acides aminés. Une systématique nouvelle est donc à envisager. Il est important pour les sélectionneurs de connaître toutes les caractéristiques génétiques et morphologiques des variétés et cultivars qu'ils comptent en introduction. La recherche de nouveaux hybrides reste en effet du domaine de la pure spéculation, si on ne connaît pas les distances taxonomiques qui séparent les parents, donc aussi la

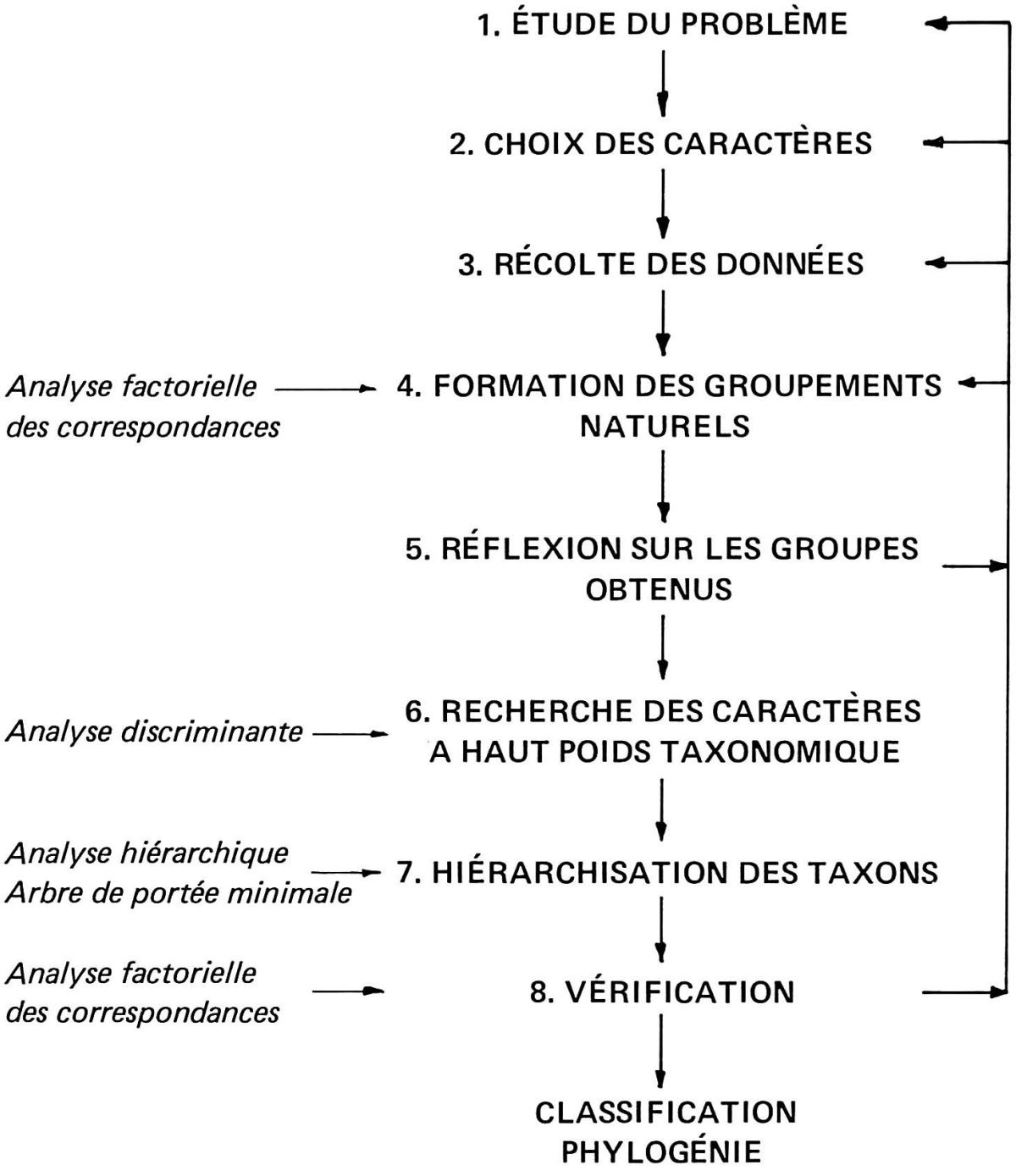

Fig. 34. – Organigramme méthodologique.

phylogénie, avant de commencer les programmes de sélection. On est ainsi amené à envisager le choix de caractères nouveaux et à préciser leur pouvoir discriminant. Les modules 2 et 6 sont alors les plus importants.

Une troisième application de la taxonomie numérique que nous voudrions citer ici, bien qu'étant en dehors de notre sujet, peut consister en un classement automatique d'individus inconnus. Celle-ci ne peut intervenir que lorsque les taxons ont fait l'objet de recherches approfondies, que plusieurs monographies ont précisé les points restés obscurs, les espèces mal classées ayant été correctement resituées et renommées, donc lorsque plus aucun doute ne subsiste sur la population envisagée. L'ordinateur peut alors servir à déterminer automatiquement de nouveaux individus. La création d'une banque de données est alors souhaitable, située si possible dans l'herbier qui contient la majorité des types du taxon envisagé.

La taxonomie numérique a pris ces dernières années une place importante dans la conception de la systématique. Elle a fait l'objet de nombreux travaux et ouvrages, dont la liste serait aussi fastidieuse qu'incomplète (EVERITT, 1974; SNEATH, 1957; SNEATH & SOKAL, 1973; SOKAL & MICHENER, 1958; MASCHERPA, 1976). D'autre part, l'essor de cette technique a engagé de nombreux mathématiciens à approfondir les théories mathématiques existantes, et surtout à en concevoir de nouvelles. Aussi, le taxonomiste se trouve-t-il souvent perdu devant le foisonnement des modèles et programmes informatiques qui lui sont proposés. Pour notre part, nous avons essayé de choisir celles qui s'adaptent le mieux à la démarche classique du systématicien devant sa table de travail, à la charge de chacun de retrouver dans le parc informatique dont il dispose, les programmes qui lui seront nécessaires. Pour chaque module de l'organigramme méthodologique, nous définirons donc la théorie mathématique qui est la plus logique et la plus performante, ainsi qu'un programme informatique.

4.2. Le choix des caractères

Nous définirons un caractère comme étant le "résultat d'une observation naturelle ou au moyen d'un appareil, susceptible de variation sur toute l'étendue des individus étudiés".

Chaque caractère peut donc être de nature très différente selon le type d'observation effectué. On les réunit généralement en trois groupes: 1) les caractères morphologiques – visibles ou microscopiques – cytologiques; 2) les caractères physiologiques et biochimiques; 3) les caractères écologiques. Toute classification, basée sur un choix arbitraire, tout au moins au départ, d'un certain nombre de caractères doit refléter le plus précisément possible la structure du génome, qui détermine le phénotype. Deux hypothèses rapportées par Sneath & Sokal, semblent être généralement vérifiées:

– l'hypothèse de la non-spécificité, selon laquelle il n'existe pas de groupes de gènes dont l'influence se limite exclusivement à une seule partie ou à une seule fonction de l'organisme,

– l'hypothèse selon laquelle tout caractère est dépendant de plusieurs gènes, et réciproquement, la plupart des gènes influencent plusieurs caractères.

Lorsque le nombre de caractères choisis est grand, ce qui est toujours le cas de la taxonomie numérique, il arrive souvent que certains d'entre eux soient logiquement ou partiellement corrélés. Il est alors très difficile de s'en apercevoir, et le taxonomiste devra utiliser des méthodes mathématiques permettant de passer outre cette difficulté. Heureusement, il en existe de nombreuses que nous exposerons dans le paragraphe suivant. On arrive ainsi logiquement au point 3 de l'organigramme, la récolte des données et leur transformation dans un support accessible à la machine, ou quantification. C'est le problème du codage ou non des caractères.

On appelle "état d'un caractère" les valeurs que peut prendre un caractère selon l'étendue de la population. Quand un caractère est mesuré, l'état de ce caractère est infini, la mesure pouvant prendre toutes les valeurs entre 0 et l'infini. On dit aussi que la variable est "continue". Mais, il peut aussi ne prendre que des valeurs entières précises, comme le nombre de bractées ou de pièces florales; on dit alors que cette variable est "discrète". Un caractère peut aussi être codé. Le codage des caractères, bien que permettant un travail ultérieur plus rapide de la machine, demande un effort supplémentaire de la part du taxonomiste, car la prise des données doit se faire en deux temps. En effet, pour être sûr d'avoir envisagé toutes les possibilités de codage, le taxonomiste doit parcourir une première fois toute l'étendue de la population, puis reprendre le travail pour coder les caractères. A notre avis, dans les travaux de révision, ce double passage au travers des échantillons est une bonne chose pour le taxonomiste, soit qu'il ne connaisse pas sa population de départ (!), et alors cela lui permettra de se familiariser avec son matériel, soit qu'il la connaisse, et dans ce cas le supplément de travail n'est pas insurmontable.

Rejoignant Adanson, si on considère que la systématique a pour but d'établir une classification d'individus, basée sur des caractères mesurés dont on ne connaît pas au départ l'importance relative, la pondération des caractères "à priori" nous semble relever du non-sens. Il paraît en effet plus judicieux, une fois la classification obtenue, de rechercher les caractères qui discriminent le mieux les différents groupes, ou les différents individus, qui mettent le mieux en évidence la structure taxonomique de la population. Nous serons ainsi amenés, par une analyse discriminante établie sur les groupements naturels issus de l'analyse numérique, à établir une hiérarchisation "à posteriori" des caractères observés, de manière à rendre plus précis les caractères distinctifs de chaque groupe. De plus, n'ayant pas d'à priori, le taxonomiste pourra voir surgir de l'analyse mathématique des combinaisons de caractères souvent peu visibles, ou insoupçonnées au début de l'étude.

4.3. Mesures de la similarité entre individus

Toute classification phénotypique a comme point de départ, sur la base des caractères observés, l'établissement d'une "mesure" de la ressemblance entre les

différentes unités taxonomiques. Ici, le choix est immense, tant les mathématiciens nous ont apporté d'indices différents (SNEATH & SOKAL, 1973). Le concept initial est de trouver une formulation reflétant au mieux l'information contenue dans les données originales. Notons ici que le terme "d'unités taxonomiques" est une généralisation pratique du terme "individu", car il s'applique aussi bien à un individu à proprement parler, qu'à une espèce, un genre, etc., selon le niveau de classification auquel on s'intéresse.

Les indices de similarité peuvent être rangés en quatre grandes catégories: les indices de distance, d'association, de corrélation et les indices probabilistes. Nous ne nous intéresserons ici qu'aux trois premiers.

4.3.1. Les indices de distance

Ils mesurent la distance qui sépare deux unités taxonomiques dans l'espace choisi des caractères, et donc reflètent plus précisément leur dissimilarité. Le cas le plus général est celui où l'espace des caractères est un espace euclidien, où tous les axes de coordonnées sont perpendiculaires. De nombreux indices de distance sont proposés dans la littérature: indice des différences moyennes de CAIN & HARRISON (1958) – qui est un indice simple, métrique, mais présentant le désavantage de sous-estimer les distances entre les individus – la distance euclidienne, le D^2 de MAHALANOBIS (1936). La distance euclidienne est l'indice de similarité le plus fréquemment employé en taxonomie numérique, défini pour la première fois par SOKAL en 1961. C'est cet indice que nous utiliserons pour la formation des dendrographes.

Quand le nombre de répétitions par individu est suffisant pour prendre en considération la variance entre ces répétitions, une mesure de la distance peut être fournie par le "coefficient of racial likeness" – CRL – de PEARSON (1926). Cet indice très intéressant ne peut se concevoir que si les caractères observés sont mesurables, continus et évidemment pas codés. Le D^2 de Mahalanobis a l'avantage de permettre le calcul des corrélations entre variables. Quand ces corrélations sont nulles, il est équivalent à une distance euclidienne mesurée sur des variables standardisées. Ce sera la mesure de distance de l'analyse discriminante.

4.3.2. Les indices d'association

Ils sont généralement utilisés lorsque les caractères observés sont codés, binaires ou à plusieurs états. Ils mesurent le nombre de fois où le même caractère a le même code pour les deux individus observés. Le plus ancien connu est l'indice de JACCARD (1908), bien souvent utilisé dans les travaux de systématique. Il ne tient pas compte des caractères absents pour les deux individus, et sera très utile lorsque le manque d'informations sur un caractère entraîne l'absence de plusieurs autres. Ainsi, l'absence de bractées florales entraînera forcément le manque d'information – involontaire – sur leur forme, leur pilosité,... Les coefficients de SOKAL & MICHENER (1958), de ROGERS & TANIMOTO (1960) ont été conçus pour des caractères binaires. Ils sont souvent utilisés en taxonomie numérique, d'autant

plus que Sneath a montré qu'on pouvait les ramener à une distance euclidienne mesurée sur des données non-standardisées. Leur généralisation à des caractères à plusieurs états ne pose pas de problème sérieux, et les classifications obtenues dans ce cas sont très voisines. Malheureusement, ils ne permettent pas le calcul des mesures de similarité après pondération, donc la hiérarchisation des taxons après pondération. Citons encore le coefficient général de similarité de GOWER (1971), qui présente l'avantage de pouvoir être utilisé avec toutes les combinaisons de variables, codées binaires, à plusieurs états, discrètes ou continues.

4.3.3. Le coefficient de corrélation

Le coefficient de corrélation de Pearson est le plus utile lorsque les données sont de type continu ou codé à plusieurs états. Il est identique au coefficient de la statistique de régression. Il peut normalement varier de +1 à −1, +1 donnant la corrélation parfaite. Remarquons qu'il est difficile de concevoir dans une population d'individus vivants, qu'un coefficient de corrélation puisse prendre la valeur −1, ce qui signifierait une antinomie sur tous les caractères des deux individus en présence. Les coefficients de corrélation ne sont pas métriques, mais on peut les transformer en distances euclidiennes, en prenant leur complément à 1, ou au moyen de la fonction arccosine. C'est la stratégie que nous emploierons pour la formation des dendrographes. Citons en outre, leur avantage de faire figurer la matrice de covariance des données dans le calcul des mesures de similarité.

Voyons maintenant les modules restant de l'organigramme méthodologique, ainsi que les théories mathématiques appliquées dans chaque cas.

4.4. Le stockage de l'information

Sauf pour les modules 1, 2 et 3, un programme correspond généralement à un module. Dans le choix des caractères, nous n'avons pas vérifié leur distribution au préalable, cette opération statistique de l'invariance d'un caractère étant vérifiée lors des différents programmes utilisés.

Une fois quantifiée, l'information est traitée par un langage spécial, INFOL – pour INFormation Oriented Language – proposé dès 1968 par l'Université du Northwestern (Evanston, Illinois) et nettement amélioré par l'Université de Genève depuis 1970 (CHENAIS & al., 1972). INFOL traite principalement des fichiers disques ou bandes magnétiques. De notre point de vue, l'utilisation de tels langages est essentielle en taxonomie, car elle permet de créer des banques de données utilisables pour les autres chercheurs du monde entier. Il est toujours plus facile de faire parvenir une bande magnétique, qu'une série de spécimens d'herbiers.

4.5. Recherche des groupements naturels, analyses des correspondances

Ce paragraphe correspond au module 4 de l'organigramme.

A ce point du travail, l'idée de base de la taxonomie numérique est d'engager le chercheur à oublier tout ce qu'il connaît de sa population, de manière à garder l'objectivité de la machine. Comprenons bien qu'il n'est pas question de mettre en doute les qualités scientifiques et l'objectivité des taxonomistes dits classiques. La taxonomie numérique part de l'axiome suivant: si la population observée possède une certaine structure, c'est-à-dire si les individus sont normalement groupés et possèdent donc entre eux certains liens, morphologiques, cytologiques, phylétiques ou autre, l'information contenue dans les données doit permettre, à elle seule, de retrouver cette structure. Alors, laissons faire la machine. Dans ce but, il existe une série d'algorithmes à notre disposition. Nous ne citerons que les principaux, comme les analyses hiérarchiques ascendantes ou descendantes, les analyses de partition, de recherche de densité, les analyses factorielles en composantes principales, l'analyse des variables canoniques, l'analyse factorielle des correspondances. Notre propos n'est pas ici de discuter la valeur respective de chacune d'entre elles, et nous renvoyons le lecteur à des ouvrages de compilation déjà parus (BENZÉCRI, 1973; BLACKITH & REYMENT, 1969; EVERITT, 1974; LANCE & WILLIAMS, 1966, 1967a et b; MASCHERPA, 1976; MCQUENN, 1967; RENÉ-CHAUME, 1975; SNEATH & SOKAL, 1973). Quant à nous, nous avons choisi l'analyse factorielle des correspondances de l'équipe de J. P. Benzécri. Nous pensons qu'elle va plus loin dans l'interprétation que les analyses factorielles classiques et les analyses des variables canoniques.

Historiquement, l'analyse factorielle en composantes principales est apparue la première. Elle permet d'étudier une série de données dont on ne peut pas connaître l'interdépendance à priori. En d'autres termes, lorsqu'on étudie une population, on recherche généralement plus d'information qu'il n'est nécessaire pour la décrire, pour la définir. Cette information peut être scindée en deux parties: l'information suffisante et nécessaire, qu'on ne connaît généralement pas, et une information supplémentaire, génératrice de parasites et qu'on aimerait bien supprimer.

Le principe de l'analyse factorielle en composantes principales consiste donc à trouver un jeu de variables nouvelles, y pouvant décrire l'information contenue dans le jeu de variables originales x, selon les conditions suivantes:

- chaque y est une combinaison linéaire des variables x:

$$y_i = a_{i1}x_1 + a_{i2}x_2 + \ldots\ldots + a_{ip}x_p$$

- la somme des carrés des a_{ij} est égale à 1.
- de tous les y_i possibles, y_1 possède la plus grande variance,
- de toutes les combinaisons y_i possibles qui ne sont pas corrélées avec y_1, y_2 est celle qui a la plus grande variance, etc.

De cette manière, le nouveau jeu de variables présente la propriété d'être composé de variables *"y" non corrélées et arrangées en variances décroissantes.* Donc seules les quelques premières variables y contiennent suffisamment d'information

pour expliquer la structure de la population. Cette méthode ne pose aucune hypothèse sur la distribution des variables, elle est générale et ne prétend tester aucune hypothèse originale. Ces premières variables, ou composantes principales, sont les axes principaux d'étirement du nuage des individus. L'avantage de cette méthode est donc de remplacer l'hyperespace des données originales, par un espace de dimensions réduites, non corrélées, donc orthogonales. Nous avons vu que les premiers axes principaux expliquaient la majorité de la variance des données. En taxonomie, on peut maintenant choisir, soit de procéder à une analyse hiérarchique dans cet espace simplifié, soit de représenter directement la population dans ce même espace, et d'étudier "de visu" la proximité respective des individus. On les représente alors dans une série de graphiques de coordonnées rectangulaires, chaque axe étant une des composantes principales conservée. Le problème qui se pose est d'essayer de trouver une correspondance entre ces axes et les variables originales qui ont servi à les établir. C'est la "reification", travail extrêmement ardu et complexe. L'analyse des correspondances permet, entre autres, de faciliter ce travail. C'est aussi une analyse factorielle, mais qui procède de la manière suivante: on peut représenter les individus d'une population dans l'espace des variables mesurées, mais inversément, on peut aussi représenter les variables dans l'espace des individus. L'information est exactement la même, et les deux espaces sont symétriques. Il suffira donc de trouver une relation permettant de passer de l'espace des variables à celui des individus, pour pouvoir projeter ces deux espaces sur le même plan du graphique. D'où le nom de "correspondance" entre les deux espaces. On arrive ainsi à visualiser en même temps les individus et les variables, et à reifier plus facilement les axes principaux d'inertie. Pour plus de détails sur les théories mathématiques que nous avons volontairement simplifiées, voir les textes de BENZÉCRI (1973) et de ESCOFIER-CORDIER (1969). Remarquons cependant que l'analyse des correspondances se différencie aussi de l'analyse factorielle en composantes principales par la pondération des lignes et des colonnes de la matrice des données, et par l'utilisation d'une distance probabiliste. Elle fait intervenir les corrélations entre "profils" des individus ou des variables, et non les corrélations linéaires classiques. Cette technique nouvelle nous a séduit par la diversité des enseignements qu'on peut en tirer, car elle ne se limite pas seulement à la description d'une structure, mais apporte aussi des renseignements sur la variabilité des caractères, leur corrélation, leur contribution.

Dans des domaines qui nous sont proches, citons quelques travaux qui ont utilisé l'analyse des correspondances. En taxonomie, ce sont par exemple les révisions des genres *Myosotis* (BLAISE, 1969; BLAISE & al., 1973), *Erodium* (GUITTONNEAU & al., 1971), *Panicum* (PERNÈS & al., 1975). En phytosociologie et phyto-écologie, ce sont les travaux de BARBE (1974), BÉGUIN & al. (1974), BRIANE & al. (1974), BRISSE & GRANDJOUAN (1971), LACOSTE & ROUX (1971, 1972), RAMEAU (1974), ROMANE (1972).

Souvent, l'analyse des correspondances est complétée par une analyse de groupement, hiérarchique ou non. Nous avons pour ce travail, fait intervenir l'analyse hiérarchique au niveau du module 7 de l'organigramme.

Une fois que la structure de la population est reconnue, que les groupements naturels sont précisés, nous passons logiquement au point 5 de l'organigramme: réflexion sur les groupes obtenus. En effet, l'ordinateur a travaillé, mais le taxonomiste reste le maître à jouer. Les résultats obtenus doivent maintenant être vérifiés, selon ce qui est déjà connu de la systématique de la population étudiée. Soit

les groupements obtenus et la taxonomie coïncident, et on peut passer directement au point suivant, soit, et c'est souvent le cas, il existe des différences entre les deux. Ces différences peuvent provenir d'une mauvaise étude du problème, d'un choix insatisfaisant des caractères – ceux recueillis ne permettant pas de déterminer la structure réelle de la population – ou d'une récolte incorrecte des données. Après réflexion et corrections éventuelles, on recommencera le travail informatique. Voir à ce sujet l'influence des données manquantes! Dans le cas le plus intéressant, lorsque le taxonomiste découvre une nouvelle structure, il sera logique de considérer les groupes ou les individus en litige d'une manière séparée. On définira les groupes par les individus certains, laissant les individus litigieux comme "spécimens à attribuer". D'une part, les individus à attribuer sont exclus des groupes de base de l'analyse discriminante (cf. paragraphe suivant). D'autre part, pour chacun d'eux, on calculera leur distance moyenne par rapport à tous les individus d'un même groupe. On connaîtra de plus la distance minimale et la distance maximale qui sépare ces individus de chacun des groupes, comme indiqué sur le tableau 14. Tous ces renseignements, liés à ceux obtenus par l'analyse discriminante, par l'analyse hiérarchique et par l'analyse des correspondances sur les données pondérées, permettront de juger objectivement de l'appartenance des individus à attribuer à l'un ou à l'autre des groupes naturels.

4.6. Recherche de la pondération des caractères, analyse discriminante multivariée

C'est une méthode simple de réduction de variables, portant sur la recherche des variables à haut poids informatif. Ce n'est pas à proprement parler une méthode d'ordination, puisqu'elle présuppose la connaissance des groupements naturels. Elle correspond au module 6 de l'organigramme.

L'analyse discriminante est une généralisation aux observations multivariées de l'analyse de variance simple à l'intérieur et entre les groupes, avec cette modification que toutes les variables sont considérées au même niveau. L'analyse se déroule en trois temps. A partir des groupements naturels définis antérieurement, on détermine un certain nombre d'individus, appelés "individus de base", sur lesquels on recherchera les critères discriminants, c'est-à-dire le "profil" caractéristique de chaque groupe. Ces individus de base seront choisis parmi ceux dont la position systématique ne fait plus aucun doute. Parmi ceux restants, on choisira un certain nombre d'"individus tests", qui permettront d'apprécier la finesse des caractères discriminants, à posteriori. Puis, une fois que le profil de chacun des groupes sera bien défini et testé, on cherchera l'allocation des individus à attribuer, dits ici "individus anonymes".

Considérons un ensemble de n individus, répartis en k groupes, sur lesquels on a mesuré p variables. Chaque groupe i peut être caractérisé par une valeur y définie comme suit:

$$y_i = a_1x_1 + a_2x_2 + ... + a_px_p$$

Cette fonction $y = f(a \cdot x)$ est appelée fonction discriminante, où x représente les variables et a les poids attachés à chaque variable. Le problème posé se résume à chercher les valeurs des poids a qui rendent maximales les différences entre les groupes, de manière à les séparer – à les discriminer – le mieux possible. Une fois que ces poids sont connus, on peut calculer la distance qui sépare chacun des groupes dans l'espace réduit des nouvelles variables, comme pour l'analyse des correspondances. La distance est mesurée ici au moyen du D^2 de Mahalanobis, comme dans le cas des correspondances, ce qui permettra de vérifier les résultats obtenus par chacune des deux analyses. Pour apprécier la finesse de la discrimination, le programme utilise la notion de "pourcentage de biens classés". Il calcule la fonction discriminante de chaque groupe, puis attribue chacun des individus de base au groupe dont il est le plus proche. Comme on connaissait l'attribution des individus de base à priori, on peut ainsi calculer le pourcentage des individus qui sont attribués à posteriori conformément à leur attribution originale. Il en sera de même pour les individus test, puis pour les individus anonymes. Pour plus de précisions sur la méthode, on se reportera au livre de BLACKITH & REYMENT (1971) et à celui de ROMEDER (1973).

Dans la logique systématique, on arrive ainsi "presque" au terme de l'analyse. En effet, après avoir étudié exactement le problème posé, on a choisi des caractères, puis récolté les données. Ensuite, on a formé les groupements naturels des taxons et recherché pour chacun d'eux, les variables qui les décrivaient le mieux. Les individus litigieux ont été traités. Avant d'entreprendre la partie purement de réflexion systématique et de proposer une nouvelle classification des individus de notre population, une dernière vérification reste à effectuer. Au moyen des variables de haut poids taxonomique, on peut maintenant, si le taxonomiste le désire encore, effectuer une analyse hiérarchique sur les taxons. Pour ce faire, on emploiera de nouveau l'analyse des correspondances. Nous dirons cependant quelques mots d'une autre analyse couramment employée en taxonomie numérique, l'analyse des groupements. Elle utilise les méthodes hiérarchiques, agglomératives ou divisives, les méthodes de partition, les recherches de densité, les méthodes d'agglutination...

4.7. Hiérarchisation, dendrographe

Afin de ne pas trop alourdir cet article par des considérations de méthodologie informatique, nous ne parlerons que des méthodes hiérarchiques, les plus couramment utilisées jusqu'à présent en taxonomie numérique. Dans les méthodes hiérarchiques, les classes sont rangées en groupes, le processus étant répété à différents niveaux, jusqu'à l'obtention d'un arbre de classification ou dendrogramme. On distinguera les arbres dont la variance intra-groupe est représentée, ce sont les dendrographes, de ceux où elle ne l'est pas, ou dendrogrammes. On peut diviser les méthodes hiérarchiques en agglomératives, ou ascendantes, et en divisives, ou descendantes. Dans les deux cas, le résultat final est un arbre de classification. Toutes commencent par l'établissement d'une matrice de similitude entre les taxons, au moyen de différentes mesures de distances.

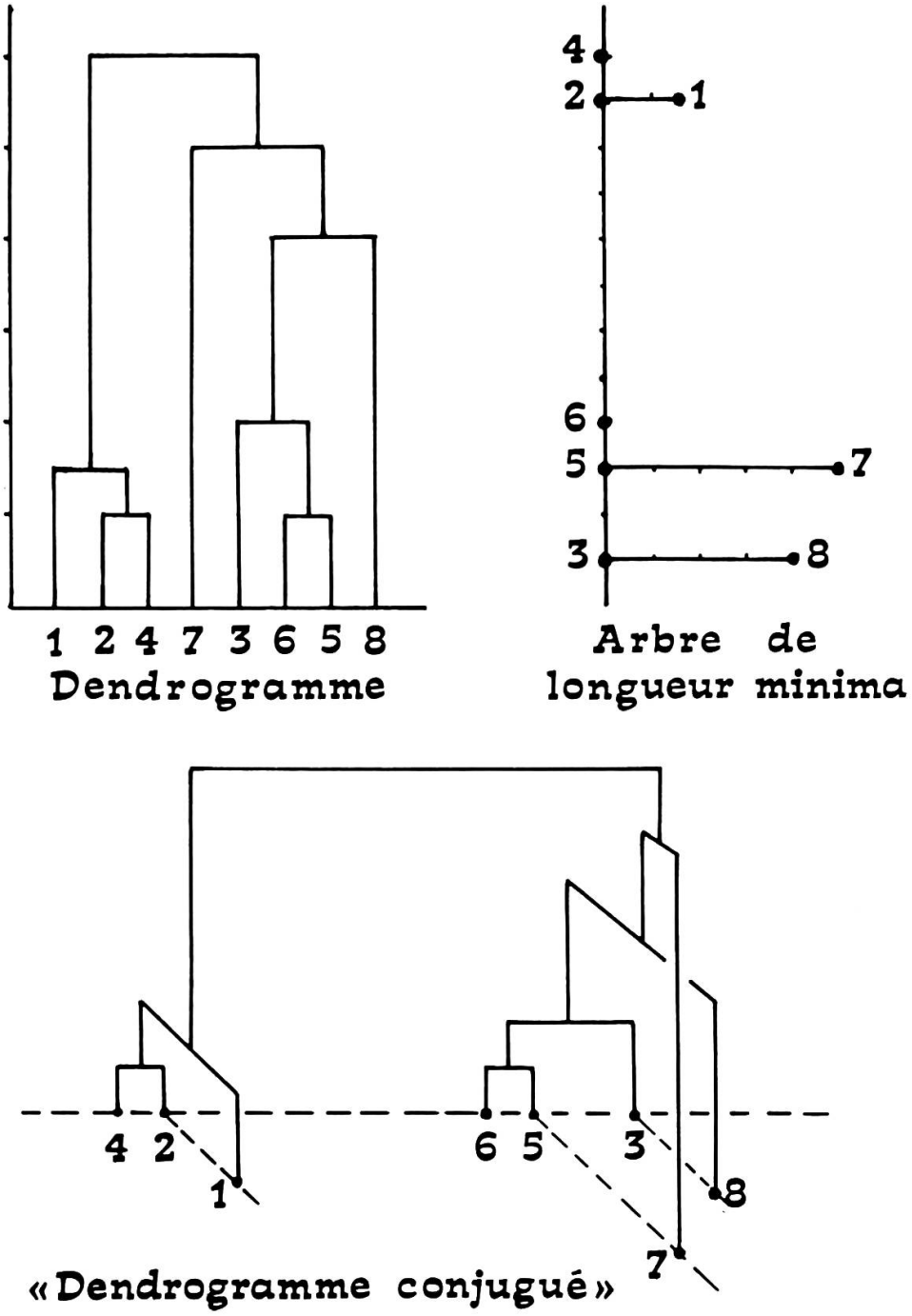

Fig. 35. – Fixation des nœuds de rotation d'un dendrogramme par l'utilisation conjuguée d'un arbre de longueur minima.

En taxonomie, le point le plus délicat dans l'utilisation des méthodes hiérarchiques reste l'interprétation du dendrographe. En effet, la théorie mathématique précise que les nœuds de l'arbre ne sont pas des points fixes, mais mobiles. Un arbre de classification peut tout à fait être comparé à un de ces "mobiles" qui servent de décoration. On peut donc considérer un arbre comme reflétant la structure de la population, dans son ensemble, mais pas dans le détail. Citant Lebart dans le livre de BENZÉCRI (1973), on peut dire que: "Quant à l'arbre global des classes, une multitude de combinaisons hypothétiques sont possibles."

Pour résoudre ce problème, c'est-à-dire pour fixer les pieds du dendrographe, BENZÉCRI & JAMBU (1976) viennent de proposer de lier une analyse hiérarchique selon le saut minimum et un arbre de longueur minimum (GOWER, 1971). Cette optique nous paraît fort intéressante pour la taxonomie, et pour illustrer ce point, nous donnerons la figure 35.

On voit sur cette figure que les individus 7 et 8 qui paraissaient fort éloignés sur le dendrographe, se révèlent plus proches lorsque les deux techniques sont employées conjointement. L'analyse hiérarchique présente aussi l'inconvénient de ne pas pouvoir remettre en cause une liaison entre deux individus, ou entre un individu et un groupe. Une fois qu'un individu est lié, il le reste jusqu'à la fin de l'agrégation, même si sa présence dans un groupe diminue la variance intergroupe. Dans ces conditions, nous croyons préférable de comparer les résultats d'une analyse hiérarchique avec ceux d'une autre méthode, pour nous l'analyse des correspondances. Nous avons trop vu de taxonomie numérique effectuée avec seulement l'analyse hiérarchique, pour nous méfier de ces résultats. Souvent, pour pallier ce défaut, les auteurs sont obligés de construire plusieurs dendrographes, en changeant une fois le nombre des individus, une fois le nombre des caractères, ou en prenant successivement des caractères de types différents. Mais dans ce cas, comment décider de la meilleure classification. Nous ne saurions donc trop conseiller au lecteur de ne pas tirer des conclusions immédiates des dendrographes proposés, mais de les comparer avec les diagrammes des correspondances.

4.8. Programmes utilisés et conclusions

Comme indiqué précédemment, les données ont été stockées sur disque magnétique au moyen du langage INFOL2 implémenté au Centre universitaire d'informatique de l'Université de Genève. Nous avons travaillé sur un ordinateur UNIVAC 1108 scientifique opérant en multiprogrammation, mémoire centrale 256K. Les terminaux utilisés sont des SUPERBEE. Le développement graphique des programmes emploient le Plotter Benson-France 1751. A part les programmes que nous avons développés nous-même, nous avons eu accès à la bibliothèque des programmes développés par le CUI.

Le programme de l'analyse des correspondances est dû à l'équipe de BENZÉCRI (1973, vol. II). Dans la version que nous possédons, le programme CORRESPO1 a été conçu par ROBERT & NICOLAU (1971, 1972). Le programme d'analyse hiérarchique DGRAPH5, utilise comme routine principale de formation des groupements naturels le programme DENDROGRAPH développé par MCCAMMON &

WENNINGER (1970), basé sur la méthode des groupements par paires non pondérées. Les auteurs ont déterminé aussi la variabilité à l'intérieur des groupes, qui permet de préciser l'homogénéité des groupes formés. La routine de dessin est due à J. M. Froidevaux & G. Gorin de l'Institut des sciences de la terre, Université de Genève, modifiée par nous-même. Toutes les mesures de similarité que nous avons employées fournissent les mêmes groupements naturels, à quelques exceptions près. En tenant compte des remarques que nous avions précédemment formulées quant aux analyses hiérarchiques, nous concluerons, avec BLACKITH & REYMENT (1971), que ce n'est pas tant les indices ou les algorithmes qui fournissent des classifications identiques, que la robustesse du matériel biologique, vivant, qui se laisse classer de la même manière. Pour l'analyse discriminante, nous avons employé le programme MAHAL3 de ROMEDER (1973), sans autres modifications que celles nécessaires à sa compatibilité aux compilateurs d'UNIVAC. La distribution statistique des caractères, les mesures de variances, tous les tests statistiques classiques ont été effectués à partir de SPSS – Statistical Package for the Social Sciences – de NIE & al. (1975).

Les conclusions taxonomiques que nous développons dans le dernier chapitre de cet article, procèdent donc d'une logique que nous espérons aussi rigoureuse qu'objective. Nous pensons ne pas avoir été tributaires d'une mathématique, et à chaque fois que nous n'avons pas été d'accord avec les résultats de la machine, nous avons essayé de savoir pourquoi et de donner une explication logique dans le texte. La méthodologie informatique est d'un grand secours en systématique, étant donnée la quantité d'information qu'elle permet d'intégrer. Mais surtout grâce à elle, nous avons pu mettre en évidence la stabilité de la matière biologique. La taxonomie numérique ne doit pas être une nouvelle méthode d'atomisation des classifications, mais au contraire doit tendre vers une simplification, doit présenter un effort unificateur de la systématique. Dans cette optique, nous pensons que la systématique doit devenir une science pluridisciplinaire, chaque vision d'un même phénomène apportant sa contribution à la vision de l'ensemble. Morphologistes, généticiens, agronomes, palynologistes, biochimistes, informaticiens peuvent ensemble aider à mieux comprendre le monde végétal.