Quelques problémes concernant l'application de la taxonomie numérique en zoologie systématique par

LIVIU DRAGOMIRESCU, MIHAI SERBAN, PETRU BANARESCU

Extrait de "Travaux du Muséum d'Histoire naturelle Grigore Antipa"
Vol. XXV
Bucharest, 1984

QUELQUES PROBLÉMES CONCERNANT L'APPLICATION DE LA TAXONOMIE NUMÉRIQUE EN ZOOLOGIE SYSTÉMATIQUE

LIVIU DRAGOMIRESCU, MIHAI SERBAN, PETRU BANARESCU

The heuristic tool of the numerical taxonomy and some alternatives of the algorithms choice are presented. Two methods are applied in five families of Harpacticoida (Crustacea, Copepoda) and in the genus *Onychostoma* (Pisces, Cyprinidae). The second studied case is an unconventional extension of the cluster analysis in zoogeography.

INTRODUCTION

Il y a plus de deux siècles que Adanson (1763) a formulé pour la première fois les principes de la taxonomie numérique. 200 ans plus tard exactement Sokal et Sneath (1963) publient leur travail fondamanental "Principles of Numerical Taxonomy", qui ouvre une ère nouvelle dans le domaine des classifications. Comme suite d'un accès de plus en plus large aux ordinateurs, on assiste ajourd'hui à une permanente diversification des modèles mathématiques de l'analyse des groupements (cluster analysis), ainsi que des aplications des classifications automatiques, qui connaissent un développement tout-à-fait explosif dans les domaines les plus divers. C'est ainsi qu'on est arrivé à définir ce qu'on appell l'analyse des données (data analysis), constituée par la taxonomie numérique et par l'analyse fatorielle (factor analysis) et dont l'utilité a été succinctement formulée par Benzecri (1973): "Représenter les données avec un minimum de perte d'information … et un maximum d'explication".

En suivant, avec certaines modifications, l'opinion Anderberg (1973), les étapes qui composent la démarche de la taxonomie numérique sont les suivantes:

1. Choix des objets d'étude
2. Choix des prédicats (des caractères dans le cas de la zoologie) qui servent à la description des objets (especès ou taxons de divers rangs).
3. Établissement des unités à classifier: objets (analyse Q) ou caractères (analyse R).
4. Choix des règles de codification pour chaque caractère et élaboration de la matrice objet x prédicat (taxon x caractères).
5. Choix des l'algorithme de classification
6. Calcul du graphe arborescent ayant comme terminaisons les objets ou les caractères (le dendrogramme).
7. Interprétation des résultats.

Dans le présent travail nous allons nous référer d'abord au choix de l'algorithm de classification et, en ce qui concerne la 2-ème application, à un problème d'interprétation aussi.

Dans la littérature que nous avons eue jusqu'à présent à notre disposition il y a, exprimées implicitement ou explicitement, plusieurs conceptions:

1. Watanabe (1969) est d'avis qu'il existe un algorithme *naturel* de classification, unique, bassé sur un groupe detrois règles que l'auteur appelle

stratégies. Cet algorithme s'applique en *divisant.* L'ensemble des unités à classifier jusqu'aux éléments. L'ensemble est doté d'une *cohésion,* une fonction positive d'ensemble, supraadditive. On aura donc une fonction d'ensemble.

$$C:P(X)\rightarrow R_+$$

dénommée cohésion et qui doit être supraaditive, c'est-à-dire

$$C(AUB) \geq C(A)+C(B), \quad (\forall)\ A,B\in P(X)$$

Les stratégies sont les mêmes quel que soit le problème réel. Nous croyons pourtant que la fonction de cohésion peut dépendre de la nature du problème à résoudre car WATANABE (1969) ne construit qu'une seule variante, basée sur l'entropie. Il insiste sur l'idée de l'unicité subjacente au *naturel,* en l'exemplifiant par la tableau de Mendeleev, qui est unique parce qu'il est le résultat d'une classification naturelle.

2. Au pôle opposé à l'idée d'unicité se place la conception d'Anderberg (1973), qui considère que: "il ne faut pas chercher d'une manière étroite une bonne classification", car chaque algorithme dévoile une certain aspect des donnés. Il exemplifie ce point de vue par la gamme très large de classifications (groupements) qui peuvent être faites avec un jeu de cartes, chaque groupement intéressant certains joueurs. Dans cette optique, la taxonomie numérique est considérée comme une *instrument de découverte* ou un générateur d'hypothèses.

3. Hanson (1958), cité par Anderberg (1973), montre que la démarche de la classification automatiquen'est ni inductive, ni déducitve, mais *rétroductive,* c'est-à-dire analogue à la manière dont les physiciens émettent des hypothéses à partir des données; il exemplifie la pensée rétroductive par la façon dans laquelle Kepler est arrivé à proposer une orbite elliptique pour Mars alors que cette orbite était considérée comme circulaire.

4. Une position intermédiaire et celle conformént à laquelle il y a plusieurs classes de problèmes chacune exigeant un algorithm à part. Dans ce sens, le traité d'"Ecologie numérique" de Legendre et Legendre (1979) présente un guide systématique semblable à un schéma logique, pour l'utilisation adéquate des divers algorithmes dans la résolution des divers problèmes écologiques.

5. Une opinion similaire est exprimée par Jambu (1973-1975), qui affirme que: "le choix de la formule de distance appartient au seul utilisateur selon l'idée qu'il se fait de la proximité entre les variables". Il s'ensuit que chaque problème réel nécessite des instruments adéquats de modelage. Nous rappelons que c'est justement dans ce cas qu'il s'agit des algorithmes les plus usités, c'est-à-dire d'algorithmes qui démarrent d'une distance ou, inversement, d'une similarité définie entre les objets, dans le but d'associer objet par objet pour former des groupements en suivant des procédés divers d'agglomération.

6. Une méthode dfe compromis entre la position idéale de Watanabe et celle pragmatique d'Anderberg est appliquée par Benzecri

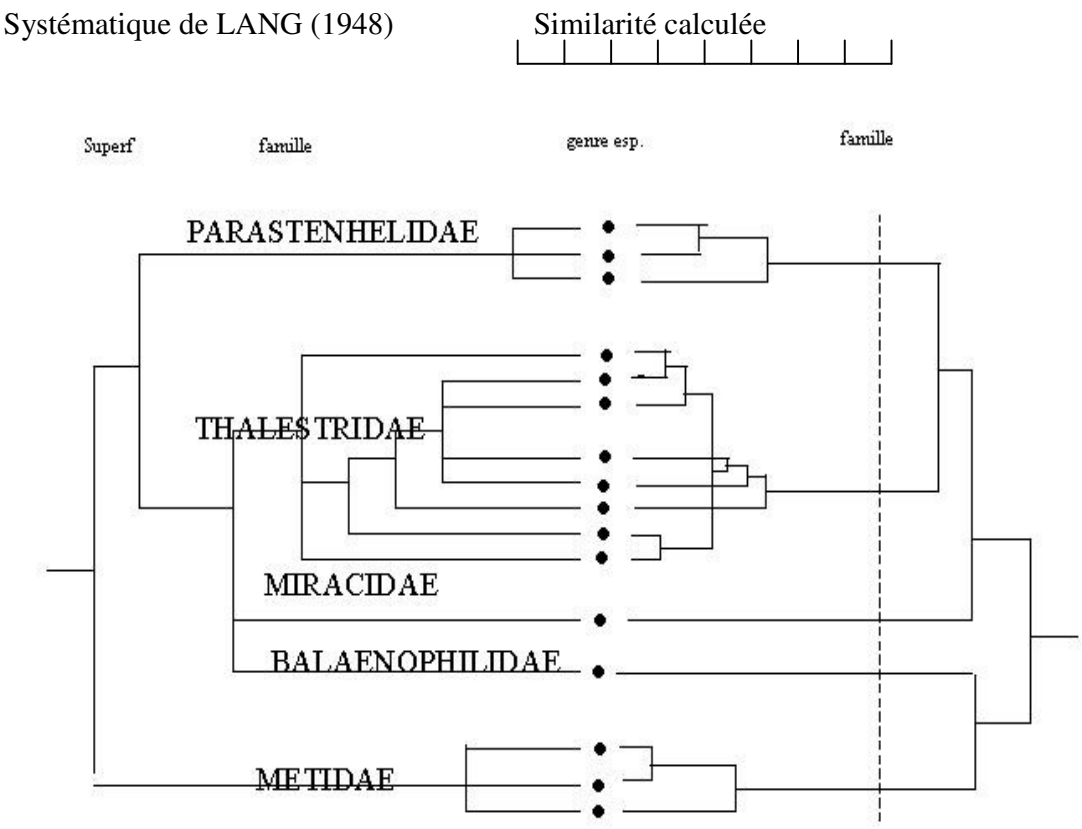

fig. 1. Comparaison entre l'arbre phylogénétique établi d'après la systématique de LANG (1948) et le dendogramme calculé par la méthode de la liaison moyenne. Il faut remarquer la concordance entre la répresentation des familles.

(1973) qui, après avoir expérimenté plusieurs algorithmes sur divers matériaux, arrive à la conclusion qu'il ne faut retenir que ceux qui offrent la plus grande satisfaction pour n'importer quel de ces matériaux. Il menttione:

a) "la maximisation de la variance interclasse, la métrique de base étant celle du χ^2".
b) "l'agrégation suivant la distance moyenne, à utiliser surtout pour un tableau en (0,1) avec l'indice de Jaccard", et
c) "l'algorithme d'échange".

7. Enfin, Sokal et Sneath (1963) considèrent la taxonomie numérique comme une discipline strictement empirique, conformément au 5-ème des principes qu'ils ont formulés.

DEUX APPLICATIONS EN ZOOLOGIE SYSTÉMATIQUE

1)*Reconstruction d'une classification de Karl LANG pour un groupe de crustacés*

La taxonomie numérique est utilisée d'habitude pour classifier des groupes systématiques peu étudiés, pour inclure des espèces nouvelles dans des classifications anciennes ou pour résoudre des situations problématiques. Dans ce travai nous essayons de suivre une autre voie, celle de la reconstruction à l'aide de la taxonomie numérique de l'arbre phylogénétique d'un groupe élaboré par les méthodes classiques. Il s'agit donc d'étudier les algorithmes qui, par leurs résultats, dendrographiques, sont "équivalents" aux procédés éuristiques de la taxonomie classique.

Notre étude à porté sur la groupe de crustacés Harpacticoides systématisés par Lang (1948), pour lequel {erban, Alb, Neagu et Racovi]\ (1979) ont utilisé le procédée éuristique – algorithmique de Hayashi et coll. (1965). Nous avons utilisé d'une part la méthode de la liaison moyenne, et d'autre part la méthode du plus proche voisin, que nous avons appliquées à l'aide de l'ordinateur. Pour les deux nous avons testé tant le coefficient de simple concordance que le coefficient Jaccard.

Les résultats obtenus montrent que la méthode la plus adéquate est celle liaison moyenne appliquée au coefficient de simple concordance. Elle aboutit à la même division en familles de l'ordre pris en considération (fig.1), les différences de détail étant dues au caractère non – binaire de l'arbre de Lang et au fait que, pour le début, nous n'avons utilisé qu'une partie des données taxonomiques sur lsequelles est fondé cet arbre.

Afin d'approfondir le problème, nous avons à présant envue d'utiliser en totalité l'information taxonomique disponible et en même temps , de développer la méthode de la liaison moyenne pour aboutir à dyes arbres non – binaires aussi.

Le principe de cette méthode – l'agglomération des groupements avec "les centres moyens" les plus proches – se montre compatible avec la supposition d'une origine commune des taxons apparentés, ayant comme ancêtre un taxon hypothétique dont les "propriétés" sont celles du centre moyen. Nous réalisons ainsi un modelage de l'idée de phylogénie, dérivé de la pensée géométrique qui dirige les méthodes employées.

Beaucoup d'autres expériments numériques de taxonomie appliqués aux problèmes de systématique seront sans doute nécessaires afin de valider l'utilité de l'idée du taxon hypothétique dans les études de phylogénie. Si nous avons cru pourtant nécessaire de présenter ici cette idée, c'est parce qu'elle a été conçue d'une manière indépendante par les deux premiers d'entre nous, tandis que les deux derniers ont prouvé sa fertilité non seulement en systématique, mais en zoogéographie aussi, comme nous allons le voir dans la seconde application.

2) *Déduction de l'"histoire naturelle" du genre de poissons dulçaquicoles Onychostoma de l'Asie Orientale*

En partant d'un dendogramme obtenu par la méthode " de la liaison moyenne" pour 17 espèces et 2 sous – espèces de poissons cyprinides – graphe basé sur l'idée d'un ancêtre hypothétique commun - , ou, en d'autres termes,

Fig. 2.

en équivalant ce dendrogramme avec un arbre phylogénétique du genre – sans doute un groupe homogène - , on a pu reconstituer, par déduction, l'histoire de la colonisation de sept bassins fluviaux de la Chine et du nord du Vietnam, qui constituent des aires géographiques isolées (fig.2). Pour l'étude ichtyologique du genre *Onychostoma* nous nous rapportons aux travaux de Banarescu (1971 a et b).

L'équivalence en question, qui devient possible si l'on admet le concept du taxon hypothétique, a été suggérée et, en même temps, peu être soutenue dans notre cas par le fait que le dendrogramme associe 16 des 19 taxons en 8 paires d'espèces cantonnées dans des aires voisines, c'est-à-dire des espèces vicariantes. On peut donc en déduire, dans le genre Onychostoma, la spéciation s'est déroulée exclusivement par évolution divergente déterminée par l'isolement géographique .

L'interprétation que nous venons de formuler s'appuie d'une part sur la position isolée que le taxon 5 (*O. macracantha*) occupe dans le dendogramme, position qui est concordante avec la qualité d'espèce la plus primitive du genre, et d'autre part sur trois règles fondamentales de la zoogéographie:

1. Les espèces vicariantes localisées dans des aires géographiques avoisinantes sont des taxons phylogénétiquement apparantés.
2. Les espèces occupant la même aire (sympatriques) sont des taxons divergents.
3. Les régions tropicales constituent les principaux centres de genèse des espèces.

Il en résulte dans notre cas une direction générale de migration de sud vers le nord et une seconde direction, de l'ouest vers l'est dans la colonisation des deux îles: Taiwan et Hainan.

En tenent compte de ces trois régles et en comparant le dendrogramme avec la distribution géographique des espèces, on peut déduire:

a. Une concordance entre la partie extrême sud du taxon 5 et sa position isolée dans le dendogramme, ce qui confirme la qualité supposée d'espèce la plus primitive.
b. D'une manière complémentaire, le fait que le groupement C (corespondant au sous – genre *Scaphestes*), qui conformément au dendrogramme comprend les taxons les plis différenciés, occupe les aires géographique les plus éloignées par rapport au centre hypothétique de genèse du genre, est en concordance avec le principe de l'augmentation de la divergence entre les espèces évoluées et celles primitives déterminée par l'amplification du gradient des facteurs du milieu parallèlement à l'éloignement du centre génétique.

Sans faire appel à des détails, nous pouvons donc proposer une image de la colonisation des sept bassins fluviaux, qui résulte de la superposition du dendrogramme sur une cart schématique obtenue par transformation continue de la carte réelle (fig.2). Il s'ensuit que, par cette voie originale de la taxonomie numérique, on arrive finalement à résoudre un problème de la systématique classique, que Banarescu (1973) avait consideré difficile, en indiquant toutefois la méthode numérique comme solution possible.

DISCUSSION ET CONCLUSIONS

1. Dans le première application, le coefficient de Jaccard s'est montré moins convenable que celui de simple concordance, à cause de l'homogénéité du groupe. Nous le considérons approprié pour les groupes plus grands et surtout pour les groupes hétérogènes, dans lesquels il est évident que l'absence simultanée d'un caractèr ne peut pas constituer un argument en faveur de la similarité. On ne peut pas affirmer, par exemple, que l'hippopotame est plus semblable au dauphin qu'à l'éléphant parceque les deux premières espèces sont dépourvues de trompe.

En employant un terminologie inspirée de celle médicale, on peut dire que "le coefficient de simple concordance" est plus convenable aux taxonomies différentielles – par analogie avec le diagnostic différentiel - , tandis que "le coefficient de Jaccart" convient aux taxonomies de triage pour les espèces des grands groupes diversifiés – par analogie avec le diagnostic de triage.

2. C'est toujours dans le cadre de la premièr application qu'on doit mentionner le fait que l'idée d'utiliser une méthode bassée sur un principe d'agglomération, dans le but de "simuler la logique du probleme, est analogue aux modèles déterministes (par example, aux équations différentielles). Dans la littérature que nous avons consultée nous n'avons trouvé que l'expression d'une philosophie analogue aux modèles probabilistes, dans le sens que la méthode de classification est d'autant meilleure que le dendrogramme généré est plus proche de celui présumé, sans aucune référence à la liaison entre le mode de construction et la logique du problème.

Le jugement nécessaire pour le choix de la méthode taxonomique appropriée à un certain genre de problèmes peut être illustré, par exemple, en recommandant, "la méthode du plus proche voisin" dans l'étude d'une chaîne de copies d'un texte afin de déduire la cronologie de cette chaîne.

3. Par l'entremise de la taxonomie numérique, le point de vue du spécialiste est formulé en deux structures, identiques au jugement informatique:

a. une structure de données, exprimée par la matrice taxons – caractères, et
b. une structure de programme, exprimée par l'algorithme de calcul.

4. En ce qui concerne la structure de données, un poids égal accordé à tous les caractères constitue, à notre avis, une forme analogue au principe de l'information maxima de la théorie de l'information, principe qui impose une distribution uniforme en absence d'une information préalable.

5. La révolution informatique offre de très larges possibilités d'application de la taxonomie numérique à la systématique. C'est ainsi que, par l'extension de ce qu'on appelle expériment numérique (simulation sur calculatrices numériques programmables), on pourra arriver, dans tous les domaines d'application, à valider par consensus certaines algorithmes de classification, cequi se trouve en accord avec le 5-ème principe de Sokale et Sneath (1963), que nous avons mentionné.

6. La taxonomie numérique n'objective pas la classification dans on sens absolu, mais la rend contrôlable étape par étape. Elle peut contribuer ainsi à l'objectivation du subjectivisme du spécialiste. Dans le contexte de la tendance centrale de la civilisation contemporaine, marquée à notre avis par la dominance de l'algorithme, le rôle que joue le

subjectivisme du spécialiste ne diminue donc pas, mais augmente d'une manière tout à fait paradoxale, en recevant une valeure supérieure.

UNELE PROBLEME PRIVIND APLICAREA TAXONOMIEI NUMERICE IN ZOOLOGIA SISTEMATICA

REZUMAT

Lucrarea este alcatuita din trei sectiuni:

1. O prezentare succinta a demersului taxonomiei numerice si expunerea principalelor "filosofii" de alegere a algoritmilor, intalnite in literatura;
2. Doua aplicatii ale taxonomiei numerice in zoologia sistematica – reconstructia clasificarii unui grup de Crustacee studiat de Lang in 1948 si deducerea "istoriei naturale" a genului de pesti dulcicoli Onychostoma din Asia rasariteana – si
3. Unele concluzii asupra "artei" de a aplica taxonomia numerica.

"Experimentele numerice" realizate au condus la rezultate, care concorda satisfacator cu clasificarile efectuate euristic de specialisti. Dintre algoritmii experimentati, cel mai adecvat la sistematica gruparilor omogene, s-a dovedit metoda "legaturii medii" aplicata coeficientului "de simpla concordanta".

A doua aplicatie extinde utilizarea taxonomiei numerice din sistematica in zoogeografie.

BIBLIOGRAFIE

ADANSON (M.) 1763- Familles des plantes, 1:515, Paris

ANDERBERG (M. R.), 1973 – Cluster Analysis for Applications: 1-390 New York, San Francisco, London

BANARESCU (P.), 1971 – A review of the species of the subgenus Onychostoma s. str. With Description of a new species (Pisces Cyprinidae). Rev: Roum. Biol. Zool., 16, 4: 241-248.

BANARESCU (P.), 1971 a – Revision of the Onychostoma – subgenus Scaphester (Pisces, Cyprinidae). Rev. Roum. Biol. Zool., 16, 6:1 – 357

BANARESCU (P.), 1973 – Principiile si metodele zoologiei sistematice : 1-219, Bucuresti

BENZECRI (J. P.), 1973 – L'analyse des données. I. La taxonomie numérique; 1-200, Paris, Bruxelles, Montréal

JAMBU (M.), 1974 – Sur les indices des distances envue de la construction d'une clasification hiérarchique. Consommation, 2:73-87

LANG (K.), 1948 – Monographie der Harpacticiden, 1-1682:Lund

LEGENDRE (L.), LEGENDRE (P.), 1979 – Écologie numeriqué: 1-200. Paris, New York. Barcelona. Milano.

SOKAL (R.), SNEATH (A.) – Principles of numerical taxonomy : 1-359. San Francisco, London

SERBAN (M.), ALB (Maria), NEAGU (Livia), RACOVITA (G.), 1979 – Application des méthodes numériques à la systématique des Harpacticoides. I. Analyse des formules d'armature des pattes natatoires entant que critère taxonomique. Trav. Inst. Speól. "Emile Racovitza", 18:33-52

WATANABE (S.), 1969 – Knowing and Guessing: 1-400 New York

Liviu Dragomirescu:
Institutul "Victor Babes"
Splaiul Independentei 99-101
76.201, Bucuresti, Romania

Mihai Serban:
Institutul de speologie "Emil Racovita"
Strada Clinicilor 5
3400, Cluj-Napoca, Romania

Petru Banarescu:
Institutul de Stiinte Biologice
Splaiul Independentei 296
77.703, Bucuresti, Romania

Tehoredactare:
Toma Camelia Laura

Data
17.11.2002