Statistique et analyse des données

R. J. Pankhurst

Les méthodes de l'identification

Statistique et analyse des données, tome 2, n° 3 (1977), p. 43-53.

<http://www.numdam.org/item?id=SAD_1977__2_3_43_0>

Statistique et Analyse des données
3 - 1977

les méthodes de l'identification

P A N K H U R S T R. J.

Botany Department, British Museum (Natural History), Cromwell Road,
LONDON SW7 5BD, GRANDE BRETAGNE

Résumé : Les méthodes modernes de l'identification.

La méthode traditionnelle de l'identification des exemplaires de plantes et d'animaux est l'utilisation de la clé diagnostique. Des développements récents ont rendu possible la construction des clés par les ordinateurs. En même temps, d'autres moyens sont devenus beaucoup plus praticables : par exemple, (1) la construction directe des clés sur des cartes perforées, (2) l'identification par comparaison sur les ordinateurs, (3) l'utilisation de l'ordinateur dans le mode " conversationnel ". Les avantages et désavantages de ces méthodes sont discutés.

Abstract : Modern identification methods.

The traditional method for the identification of specimens of animals and plants is the diagnostic key. Recent developments have made it possible to construct such keys with the help of computers. At the same time, other types of method have become much more practicable : for example (1) the direct construction of keys on punched cards, (2) identification on a computer by calculation of comparison (similarity), (3) the use of computers in conversational, on-line identification programs. The advantages and disadvantages of theses methods are discussed.

1. Introduction.

Je vais présenter ici un aperçu général des méthodes modernes d'identification. Par " moderne ", j'entends les méthodes qui utilisent un ordinateur, soit directement, soit indirectement. En dépit de ce moyen, les caractères de l'objet qu'il faut identifier doivent être réglés par un observateur humain. A l'heure actuelle, il n'est pas question de faire faire ces observations par une machine de quelque sorte que ce soit, sauf sous certaines conditions très limitées. Aussi, la décision finale concernant l'identité d'un animal ou d'une plante reste entre les mains des scientifiques et n'est pas prise par une machine.

Il existe des méthodes de deux types généraux : " monothétique " et " polytnétique ". Les méthodes monothétiques sont celles qui fonctionnent par l'élimination de taxa par degrés, en utilisant leurs caractères l'un après l'autre. Les clés diagnostiques et les clés sur cartes perforées sont de style monothétique. Les méthodes polythétiques sont celles pour lesquelles une comparaison basée sur tous les caractères est faite à la fois ; il n'est pas besoin d'insister sur le fait que chaque caractère utilisé doit être correct.

Comme exemple de ces méthodes, nous avons :

- i les programmes de comparaison à l'aide de coefficients de ressemblance ou à l'aide de mesures de distance,

- ii les méthodes statistiques utilisant des arguments probabilistes, le maximum de vraisemblance et le théorème de Bayes. Le point de départ pour toutes ces méthodes est la matrice de données taxonomiques qui n'est qu'une table rectangulaire de tous les taxa et de tous les caractères, où sont inscrites les valeurs ou états des caractères, représentant les définitions des taxa suivant une classification donnée.

2. Clés diagnostiques.

La clé diagnostique est la méthode traditionnelle pour l'identification biologique. Le botaniste français LAMARCK (1778), dans sa "Flore française" fut le premier à publier quelque chose qui peut-être reconnu nettement comme une clé. Il écrit : " *Je doits observer ici que la manière de procéder dans une analyse ne peut être arbitraire, et qu'encore qu'il paraisse indifférent au premier coup d'oeil d'employer telle division plutôt que telle autre, la marche qui fera trouver le nom de la plante, doit cependant être combinée d'après certaines règles que je réduis à deux. La première est que l'on parvienne au but par la voie la plus sûre. La seconde est que cette voie soit en même temps la plus courte possible.*"

Cet avis est encore valable deux siècles plus tard. La logique de la construction d'une clé est plus ou moins pareille, que l'on se serve ou non d'un ordinateur. On doit admettre, quand même, que les principaux textes de la taxonomie sont plutôt vagues à ce sujet. Depuis 1970, une série de programmes pour ordinateur a été décrite (DALLWITZ (1974), HALL (1973), MORSE (1974), PANKHURST (1971), PAYNE (1975)). Un exemple du type de résultats obtenu par le programme de PANKHURST est montré dans la figure I.

KEY TO JURINEA

1 STERILE ROSETTES PRESENT. 2
2 OUTER INVOLUCRAL BRACTS PATENT, OR RECURVED. 3
3 CAULINE LEAVES AURICULATE, PAPPUS SHORTER THAN ACHENE. 11.J.POLYCLONOS
3 CAULINE LEAVES WITHOUT AURICLES, PAPPUS 1-2 TIMES ACHENE. 4
4 CAULINE LEAVES AMPLEXICAUL. 12.J.LEDEBOURII
4 CAULINE LEAVES NOT AMPLEXICAUL. 5
5 ACHENE 1-2 MM. 14.J.GLYCACANTHA
5 ACHENE 2-5 MM. 10.J.MOLLIS
2 OUTER INVOLUCRAL BRACTS ERECT. 6
6 UPPER SURFACE OF LEAVES WHITE, OR GREY, BASAL LEAVES ARACHNOID-TOMENTOSE ABOVE. 7
7 BASAL LEAVES ENTIRE, LEAF MARGINS REVOLUTE, CAPITULA HEMISPHERICAL, CORONA OF ACHENE INCONSPICUOUS, PAPPUS 1-2 TIMES ACHENE. 7.J.KIRGHISORUM
7 BASAL LEAVES PINNATIFID, LEAF MARGINS PLANE, CAPITULA OBCONICAL, CORONA OF ACHENE CONSPICUOUS, PAPPUS 2-4 TIMES ACHENE. 4.J.PINNATA
6 UPPER SURFACE OF LEAVES GREEN, BASAL LEAVES SUBGLABROUS ABOVE, OR SETOSE ABOVE. 8
8 BASAL LEAVES SETOSE ABOVE, CAPITULA OBCONICAL, CORONA OF ACHENE CONSPICUOUS, PAPPUS 2-4 TIMES ACHENE. 3.J.TZAR-FERDINANDI
8 BASAL LEAVES SUBGLABROUS ABOVE, CAPITULA SUBGLOBOSE, OR HEMISPHERICAL, CORONA OF ACHENE ABSENT, OR INCONSPICUOUS, PAPPUS 1-2 TIMES ACHENE. 9
9 CAPITULA SUBGLOBOSE, CORONA OF ACHENE ABSENT. 17.J.FONTQUERI
9 CAPITULA HEMISPHERICAL, CORONA OF ACHENE INCONSPICUOUS. 13.J.CONSANGUINEA

Fig. 1 : Clé diagnostique construite oar ordinateur (partie)

Il est bien clair que l'ordinateur ne sert qu'à construire la cle et que son utilisation pour identifier quelque chose est faite de façon traditionnelle.

Une brève explication de la manière de procéder dans ces programmes est donnée ici. Le problème est d'éviter la recherche de toutes les clés possibles, puisqu'il en existe un nombre généralement astronomique. La procédure commence avec tous les taxa et tous les caractères. On doit choisir entre les caractères, par le moyen d'une fonction calculée sur les valeurs. Cette fonction doit reconnaître plusieurs aspects des caractères. incluant :

1) la distribution des valeurs, avec une préférence pour les caractères qui divisent les taxa en deux groupes plus ou moins égaux,

2) une préférence pour les divisions en deux et non en trois ou plus de trois

3) une préférence pour les caractères constants et bien connus (sans trop de valeurs manquantes) et,

4) pour les caractères préférés par l'utilisateur du programme.

Il y a d'autres facilités possibles avec de tels programmes : on peut choisir entre les deux styles de clé, parallèle ou jumelée et on peut pondérer (préférence numérique) les taxa ou les caractères choisis. Parfois, on a une matrice de données qui n'est pas complète, c'est-à-dire qu'il existe des paires de taxa qui ne sont pas bien distinguées. Si l'on accepte cette situation, on peut aboutir à une clé " partielle " dont les divisions ne se terminent parfois pas par le nom d'un seul taxon, mais par plusieurs. En chaque division d'une clé, on a un caractère principal et il est souvent possible de trouver d'autres caractères secondaires, qui sont moins utiles mais qui correspondent dans la distribution de leurs valeurs aux valeurs du caractère principal. un programme peut chercher ces caractères auxiliaires à chaque niveau. Dans la nature elle-même, il existe fréquemment une variabilité des caractères et un programme doit la prendre en compte. Par exemple, on peut trouver des caractères qui distinguent deux taxa, bien qu'ils soient variables, mais avec des valeurs différentes des deux.

3. Clés sur cartes perforées.

La forme de clé sur cartes perforées qui est probablement la mieux connue est celle qui est celle qui est perforée au bord, et qui est souvent utilisée pour des index bibliographiques. Chaque carte représente un taxon et les positions au bord représentent les valeurs des caractères. Si une valeur peut être observée pour un taxon, on troue le bord de la carte afin que, lors du tri des cartes avec une aiguille, en choisissant les caractères, les cartes qui représentent les taxa qui sont en accord soient séparées de celles qui ne le sont pas. Le nombre de caractères est limité par la longueur du bord de la carte (de l'ordre d'une centaine à peu près) mais le nombre de taxa n'a pas de limite. Des cartes perforées de ce type doivent être fabriquées à la main et il n'existe pas de moyens pour en faire des copies automatiquement. L'avantage de cette forme de clé est qu'on peut choisir n'importe quel caractère et dans n'importe quel ordre ; celle-ci donc, est une méthode à accès multiple.

L'autre forme de clé est basée sur la carte perforée classique.

Cette fois, chaque carte représente une seule valeur de caractères et la présence ou l'absence d'un trou sur les 10 lignes et les 80 colonnes indique le taxon qui possède cette valeur , ou qui ne la possède pas. Le numéro du taxon peut être facilement retrouvé par les numéros qui sont imprimés sur la carte ; par exemple, la colonne 12 et la ligne 9 représentent le taxon N° 129. Il y a assez de place pour 800 taxa, qui n'est pas une limite sévère et il n'y a pas de limite pour le nombre de caractères. Pour trouver l'identité d'un exemplaire, on prend entre les doigts les cartes correspondant aux valeurs observées et on les aligne ou superpose très exactement. Les trous indiquant les taxa qui sont en accord peuvent être vus facilement contre la lumière, ou contre un fond noir. L'identification se termine quand il ne reste qu'un seul trou. Les clés sur cartes perforées IBM peuvent être facilement fabriquées par le moyen d'un programme (PANKHURST & AITCHISON, 1975) et coûtent relativement peu cher, et sont simples à corriger.

La figure 2 ci-dessous montre une carte choisie d'une clé pour les levures (BARNETT & PANKHURST, 1974). Bien que les deux types de clé sur cartes perforées soient à accès multiple, ce sont encore des méthodes monothétiques, parce que l'identification se fait quand même par élimination progressive. Si l'on se trompe dans la reconnaissance des caractères en utilisant les deux types de clés précédents, on risque de se perdre complètement ou de tomber sur une identité fausse sans le savoir.

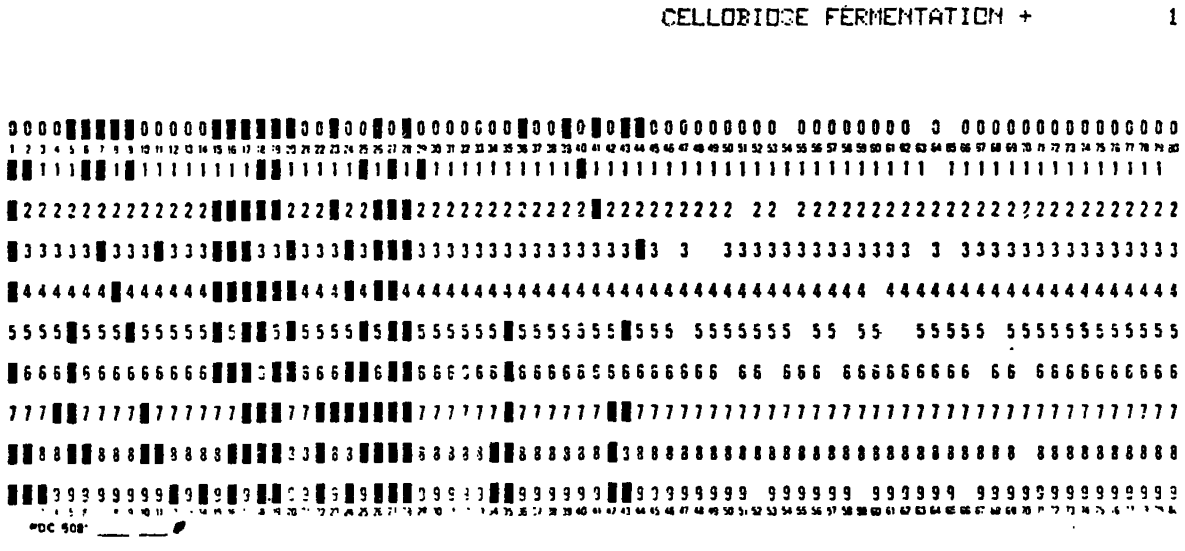

Fig. 2 : Carte perforée tirée d'une clé pour les levures.

4. Identification par comparaison.

Le principe de cette méthode est assez simple. On commence par une description, plus ou moins complète, de l'objet inconnu et on calcule une mesure de ressemblance ou de distance quelconque entre l'objet et un ensemble de taxa. Les taxa avec la ressemblance la plus grande, ou à la distance la plus proche, fournissent l'identification la plus raisonnable, ou la plus probable. La mesure de ressemblance peut être un coefficient de "similarité", ou de corrélation, ou une probabilité calculée par le théorème de BAYES ou par le maximum de vraisemblance. A l'exception des méthodes de comparaison directe sur matrice de caractères, il est nécessaire de faire les calculs sur un ordinateur, de préférence avec un système de traitement par lot ; un exemple des résultats se trouve à la figure 3.

```
/BAMISIA CF. FENESTRIFERA MALE
 SPECIAL CHARACTERS ARE -
      UNCUS/DIVISION
      GNATHUS/PRESENCE

  SEQ     SIM.   COUNT            SPECIES

    1     91.6    53   *++        FENESTRIFERA
    2     78.0    25              INOPIATA
    3     71.9    52    ++        FURVA
    4     70.5    49   *++        LETONSA
    5     66.7    48    ++        IDALIALIS

 ENSEMBLES GROUP     4

 SPECIAL TAXA COMPARED
      1    91.6 *                 FENESTRIFERA

REPORT ON TAXON     15
 CHARACTER STATE
      5       1
       DISTINGUISHED TAXA-     1    2    3    4    5    6    7    8    9
                              10   11   12   13   16   17   18   21   26
                              27   28   29   30   31   32   35

      27      1
       DISTINGUISHED TAXA-     2   11   23   25   27   31   32   35
```

Fig. 3 : Résultats d'un programme de comparaison.

Une proportion des taxa, avec la ressemblance la plus élevée, a été imprimée avec le nom du taxon et la ressemblance exprimée comme un pourcentage. Dans ce programme, on peut choisir des caractères spéciaux, ce qui veut dire que ce sont,de façon plus ou moins subjective, les caractères importants vus sur l'objet.

On attend que la plupart des caractères spéciaux soient en accord avec la détermination finale, ce qui aide à choisir entre les taxa de la liste. Le programme imprime aussi le nombre total de caractères qui ont contribué au calcul de la ressemblance, parce que, si ce total est plus bas que la moyenne, la ressemblance risque d'être fausse.

Si les taxa sont les espèces d'un genre, par exemple: le programme marque avec un astérisque les taxa sur la liste, qui appartiennent au sous-genre qui ressemble le plus à l'objet ; tous ces moyens aident dans le choix de l'identité finale.

Le programme ci-dessus n'emploie pas de probabilités ; les méthodes statistiques d'identification ont beaucoup d'avantages théoriques mais elles sont peu praticables en biologie, parce qu'on n'a pas de connaissance sur les probabilités (pour les fréquences de taxa, et les fréquences des valeurs de caractères). Assez souvent aussi, ce ne sont pas de vraies probabilités, mais des fonctions variables de l'environnement.

Les méthodes de comparaison ont l'avantage qu'elles permettent des erreurs dans les descriptions des objets ; malgré ces erreurs, on peut parvenir tout de même à l'identité précise. L'inconvénient est qu'elles exigent une description plus ou moins complète de chaque objet, ce qui demande plus de travail que pour la clé diagnostique, où l'on peut trouver l'identité sans être obligé de se référer à tous les caractères.

5. Identification par un ordinateur à système direct.

Certains programmes pour ordinateur ont été préparés pour l'identification en mode conversationnel. En général, ces programmes utilisent la méthode d'élimination progressive de style à accès multiple. Ils offrent les meilleures possibilités du point de vue de la souplesse et de l'efficacité. L'ordinateur doit disposer d'un système en temps partagé, ou doit être à la disposition d'une seule personne à la fois. Cette façon de travailler exige une technologie complexe et coûte relativement cher ; elle ne peut être mise en service en dehors d'un bureau ou d'un laboratoire. Un tel système est maintenant disponible au BRITISH MUSEUM, au département d'Histoire Naturelle.

Il est possible, quand même, que des ordinateurs de poche deviennent suffisamment puissants dans les cinq prochaines années pour pouvoir faire ces identifications.

Un système direct vous offre de nouvelles possibilités :

1) Une idée très utile est celle de la"limite de variabilité". On peut demander à l'ordinateur d'accepter les identifications jusqu'à une limite d'erreur, donnée comme nombre de caractères. Si vous choisissez une limite de deux, l'ordinateur acceptera, comme identifications valables, tous les taxa qui sont exactement en accord avec l'exemplaire et ceux qui diffèrent par un ou deux caractères. Une limite de zéro est équivalente à celle de la clé diagnostique et une limite égale au nombre total de caractères équivalente à la méthode de comparaison.

2) Chaque fois que la machine attend l'enregistrement d'un caractère, on peut lui demander quel est le meilleur, soit pour donner le maximum d'information, soit pour confirmer une identité souhaitée, afin qu'on puisse atteindre le résultat aussi rapidement que possible.

3) L'ordinateur peut être utilisé comme système d'information pour répondre aux questions telles que " Quelles sont les différences entre les taxa A et B ? " " Quelles sont les différences entre cet objet et le taxon C ? " . Un exemple complet d'une telle conversation avec l'ordinateur se trouve dans PANKHURST (1976).

6. Méthodes voisines.

On a déjà remarqué que tous les programmes d'identification sont basés sur la matrice de données taxonomiques. Cette matrice peut être utile à d'autres fins. Par exemple, un programme pour calculer une matrice de coefficients de ressemblance permet d'appliquer les méthodes de la taxonomie numérique. Il est relativement simple de construire et d'imprimer des descriptions taxonomiques directement de la matrice (figure 4, PANKHURST (1977)), et cela représente une étape dans la production automatique des textes taxonomiques.

Une question relative à toutes les méthodes d'identification est la suivante : comment trouver qu'un ensemble de caractères qui est suffisant contient des caractères utiles, mais en un nombre pas trop grand? Des algorithmes existent pour cela (WILLCOX & LAPAGE (1972)) mais, pour être complets, ils exigent de l'ordinateur , soit beaucoup de place en mémoire, soit beaucoup de temps pour calculer. Il y a aussi des méthodes approximatives, mais elles risquent de se tromper sur le vrai minimum. Dans les deux cas, il est difficile de trouver un ensemble minimum qui contienne seulement des caractères utiles. Un problème analogue est celui qui consiste à trouver un ensemble minimum de caractères pour distinguer un taxon particulier de tous les autres taxa,ce qu'on appelle "une description diagnostique". Il y a également des méthodes exactes et approximatives, étroitement liées à celles dont je viens de parler. En exemple de la méthode approximative, on peut se reporter aux descriptions de levures dans BARNETT & PANKHURST (1974).

Pour conclure, la figure 5 fournit une comparaison entre les méthodes d'identification. La complexité d'utilisation augmente de gauche à droite, à l'exception de la dernière colonne qui garde une certaine souplesse.

Programmes disponibles

Une série de programmes en FORTRAN IV, avec leurs descriptions est disponible auprès de l'auteur.

TAXACUM X GREAT BRITAIN

TARAXACUM

Plant small, delicate. Leaves not many, 0-100mm, medium width, lanceolate, medium, green, unmarked, lobed, neutral texture. Leaf bases persistent. Leaf lateral lobes 3-6, recurved, narrow, triangular, acuminate. Upper margin of leaf lateral lobes entire, or denticulate. Distal margin of leaf lateral lobes sigmoid. Leaf terminal lobe medium length, triangular, or sagittate, acute, not mucronate, entire. Leaf interlobes denticulate, medium width, not parallel-sided. Midrib purple. Petiole purple, short to quarter leaf.

Scape 0-100mm, slender, green, arachnoid below. Exterior bracts length 5-7mm, 0-2mm, erect, not stiff, lanceolate, bordered. Upper(inner) surface of exterior bract pale green, glaucous. Lower(outer) surface of exterior bract green. Exterior bracts smooth, or corniculate, contrasting.

Capitulum 21-30mm, yellow, flat. Ligule flat, striped. Styles discoloured, exserted. Pollen present. Achene 3.1-3.5mm to 3.6-4mm, spirulose above, dark red. Achene cone 0.7-0.9mm, cylindrical. Achene rostrum 9-11mm.

Fig. 4. Description taxonomique construite par ordinateur.

Bibliographie.

LAMARCK J.B.P. (1975) "Flore Françoise", 1ère édition, 3 vol., Paris, Imprimerie Royale.

DALLWITZ M.J. (1974) A flexible computer program for generating diagnostic keys. Syst. Zool. 23 (1), 50 - 57.

HALL A.V. (1973) The use of a computer-based system of aids for classification, Contr. Bolus Herb. G, Univ. of Cape town, 110 pp.

MORSE L.E. (1974) Computer programs for specimen identification, key construction and description printing. Publ. Mich. St. Univ., Biol 5 (1), 128 pp.

PANKHURST R.J. (1971) Botanical keys generated by computer, Watsonia 8, 357 - 368.

PAYNE R.W. (1975) Genkey : a program for constructing diagnostic keys, dans "Biological Identification with computer", ed. R.J. PANKHURST, pp 65 - 72, Academic Press, London 8 New-york.

PANKHURST R.J., AITCHISON R.R. (1975) A computer program to construct pólyclaves, dans "Biological identification with computers", ed. J.R. PANKHURST, pp 73 - 78, Academic Press, London 1 New-York.

BARNETT J.A., PANKHURST R.J. (1974)"A new key to the yeasts", North Holland, Amsterdam, 273pp.

PANKHURST R.J. (1976) "On-line identification program" British Museum (Natural History), 18 pp. (demander à l'auteur).

PANKHURST R.J. (1977) The printing of taxonomic descriptions by computer, Taxon.

WILLCOX W.R. LAPAGE S.P. (1972) Automatic construction of diagnostic tables, Comp. J. 15 (3), 263 - 267.

Méthode / Critère	Clé diagnos-	Cartes perforées	Comparaison par		Système Conversa-tionnel
			Ressemblance	probabilité	
Construite par ordinateur	oui ou non	oui ou non			
Utilisation exigée par l'ordinateur	non	non	oui ou non	oui	oui
Utilisable en stage pratique	oui	oui	peut-être	non	non
Nombre de caractères nécessaires dans chaque exemplaire	peu	peu	beaucoup	beaucoup	peu à beaucoup
Marche bien avec des exemples incomplets ou endommagés	non	oui	oui	oui	oui
Effet de quelques erreurs dans l'observation des caractères	souvent en erreur	souvent en erreur	normalement correct	normalement correct	normalement correct
Peut détecter un taxon non inclus	souvent non	souvent non	oui	oui	oui
Mesure numérique de l'identité	non	non	oui	oui	oui

Fig. 5 : Les Méthodes Comparées.