

1049

COLLECTION PROGRAMMATION
Publiée sous la direction de L. NOLIN

*Publications de l'Institut de Programmation
de la Faculté des Sciences de Paris*

Les bases de la classification automatique

par

I.C. LERMAN

GAUTHIER-VILLARS ÉDITEUR
55, quai des Grands-Augustins — PARIS 6^e

1970

BIBLIOTHEQUE DU CERIST

DANS LA MEME COLLECTION

Parus :

ARSAC J., LENTIN A., NIVAT M. et NOLIN L.— Algol. Théorie et pratique.

GROSS M. et LENTIN A.— Notions sur les grammaires formelles.

FAURE R.— Eléments de la recherche opérationnelle. Avec la collaboration de *N.S. Guillot*
Le Garff et de *M. Bloch*.

NOLIN L.— Formalisation des notions de machine et de programme.

Sous presse :

FAURE R. et HEURGON E.— Structures ordonnées et algèbres de Boole.

© GAUTHIER-VILLARS, 1970

Toute reproduction, même partielle, de cet ouvrage est interdite.
La copie ou reproduction, par quelque procédé que ce soit : photographie, microfilm, bande magnétique, disque ou autre, constitue une contrefaçon passible des peines prévues par la loi du 11 mars 1957 sur la protection des droits d'auteur.

TABLE DES MATIÈRES

	Pages
INTRODUCTION & RESUME	IX
Chapitre 0 — SUR QUELQUES QUESTIONS LIEES AU TREILLIS DES PARTITIONS	1
I — Treillis des partitions d'un ensemble fini	1
1 — Définitions et propriétés	1
2 — Dénombrement de chaînes de $\mathcal{Q}$	3
II — Partitions d'un entier n	4
1 — Généralités	4
2 — Représentations	5
III — Type d'une partition et cardinal du graphe de la relation d'équivalence correspondante	5
1 — Dénombrement de partitions de type donné	5
2 — Type d'une partition et cardinal du graphe correspondant	6
IV — Espaces ultramétriques	10
1 — Propriétés	10
2 — Chaîne du treillis des partitions d'un ensemble fini et espaces ultramétriques associés	11
Chapitre I — MESURE DE SIMILARITE ET PREORDONNANCE	13
I — Introduction	13
1 — Caractères et attributs	13
2 — Signification d'un attribut pour une classification	14
3 — Choix des attributs	14
II — Définition d'une mesure de similarité	15
1 — Généralités	15
2 — Définitions et propriétés	16

	Pages
III — Préordonnance associée à une mesure de similarité	19
1 — Définitions et propriétés générales	19
2 — Ecart entre deux préordonnances associées à deux similarités . .	19
IV — Proposition de quelques préordonnances particulières sur E	24
1 — La préordonnance la “plus fine”	24
2 — La préordonnance la “moins fine”	25
3 — Préordonnances associées à des indices centrés et réduits	26
Chapitre II — CRITERES	27
I — Introduction	27
II — Critères définis à partir d’une matrice d’incidence des données	28
1 — Généralités	28
2 — Définition d’un critère	28
III — Critères définis à partir d’une relation binaire symétrique sur E. . . .	31
1 — Origine de la relation	31
2 — Représentation d’une partition	31
3 — Définition d’un critère	31
4 — Aspect limite du critère défini	32
IV — Critères définis à partir de la donnée d’une famille finie de partitions .	33
1 — Origine de la donnée	33
2 — Représentation d’une partition	34
3 — Définition d’un critère	34
4 — Objets équivalents	37
V — Critères définis à partir de la donnée d’un indice de similarité	38
A — <i>Recherche d’une partition</i>	38
1 — Introduction	38
2 — Critères définis pour la recherche d’une partition	38
B — <i>Recherche d’une chaîne de partitions</i>	39
1 — Représentation d’une chaîne de partitions	39
2 — Définition d’un critère	40
VI — Critères définis à partir d’une préordonnance sur E.	41
A — <i>Recherche d’une partition</i>	41
1 — Représentation d’une partition	41
2 — Définition d’un critère	41
3 — Comparaison des deux critères	43
4 — Distribution asymptotique de $ \omega \cap R \times S $	49
5 — Autres critères	51

	Pages
B — Recherche d'une chaîne de partitions	54
1 — Représentation d'une chaîne de partitions	54
2 — Généralisation des critères précédents	54
3 — Rang lexicographique et nombre d'inversions	57
VII — Conclusion	63
Chapitre III — ALGORITHMES	65
I — Introduction	65
II — Algorithmes des transferts	65
1 — Définition	66
2 — Calcul des partitions centrales	66
3 — Application de l'algorithme à d'autres critères	67
III — Algorithme "lexicographique"	67
1 — Définition d'une préordonnance ultramétrique ω_u liée à ω . . .	68
2 — Algorithme permettant d'obtenir la préordonnance ultramétrique ω_u définie par H	68
3 — Hiérarchie de classifications associée à une ordonnance	71
Chapitre IV — ASPECTS STATISTIQUES	75
I — Introduction	75
II — Deux problèmes de convergence relatifs à la méthode des partitions centrales	75
A — Premier problème	75
1 — Introduction	75
2 — Proposition	75
B — Deuxième problème	77
1 — Introduction	77
2 — Proposition 1	78
3 — Proposition 2	78
III — H-classificabilité	80
1 — Introduction	80
2 — Ecartement entre la structure de ω et celle d'une préordonnance ultramétrique	81
3 — Test de l'hypothèse N de la non existence d'une h-classification "naturelle"	83
Annexe I, Programmes	89

	Pages
Annexe II, Exemples réels	105
1 — Un exemple de géographie économique	105
2 — Un exemple de psychosociologie	107
3 — Un exemple sur l'évolution des espèces	110
BIBLIOGRAPHIE	115

INTRODUCTION ET RÉSUMÉ

La plus vaste classification que l'homme ait édiflée est celle du règne animal et du règne végétal. Bien que contestée dans ses ramifications, cette classification a été l'indispensable langage commun à tous les naturalistes et a constitué la matière brute dont l'élaboration suscite encore maintenant de nombreuses recherches fructueuses.

L. Blaringhem introduisant *Les Mathématiques de l'hérédité* de G. Malécot rappelle que "Francis Galton vers 1880, après Ad. Quetelet (1835), avant Alphonse Bertillon dont les *Instructions signalétiques* (1893) sont en usage familier, réalisa les portraits composites. Tous les Indiens, dit-il, ont, à première vue, des traits communs qui les distinguent des autres races ; les Anglais de même ne sont pas confondus avec les Espagnols ; et surtout, les membres d'une même famille ont dans leurs traits, dans leurs gestes, dans leur marche une allure qui malgré l'âge et le sexe donnent un air de famille", forment une classe "naturelle" dirons-nous ?

Utilisé couramment par les taxinomistes, le concept de classe "naturelle" reste mal défini. Danser (1950) s'aperçoit de la difficulté de donner un statut scientifique à ce terme. Comme déjà le faisait remarquer Vicq d'Azyr, (1792), la définition d'une classe par une propriété caractéristique, (classe monothétique), est trop rigide pour les Sciences de la nature et de l'homme. Il faut attendre Beckner (1959) pour une formulation intuitive, (qui laisse donc une part d'arbitraire), d'une classe "naturelle". Une telle classe se réfère à un ensemble fini d'attributs tel que chaque élément de la classe possède une proportion importante, mais non spécifiée, de ces

attributs et réciproquement chaque attribut est possédé par une proportion importante des éléments de la classe, sans qu'il soit nécessaire qu'un des attributs soit possédé par tous les éléments de la classe. On appelle une telle classe polythétique (par opposition à monothétique).

C'est relativement à une visée classificatoire (Gilmour, 1937) que le spécialiste établit les attributs de description de la population qu'il étudie ; pour suivre la démarche d'Adanson (1757) : "Je me contenterai de rapprocher les objets suivant le plus grand nombre de degrés de leurs rapports et de leurs ressemblances. . . Les objets ainsi réunis formeront plusieurs petites familles que je réunirai encore ensemble, afin d'en faire un tout dont les parties soient unies et liées intimement".

Respectant l'intuition du spécialiste, il importe par conséquent pour une bonne classification (partition) de la population étudiée, que, relativement aux attributs de description, le degré de ressemblance de deux objets soit grand s'ils appartiennent à la même classe et petit sinon. De plus une hiérarchie de classifications emboîtées (chaîne ordonnée de partitions) respecte d'une manière satisfaisante les ressemblances entre objets, si la ressemblance de deux objets d'une même classe est d'autant plus grande que le niveau de la hiérarchie où ils se trouvent réunis pour la première fois est plus élevé (que la partition qui les réunit est plus fine).

Un des intérêts pratiques les plus immédiats de la taxinomie est de reconnaître un nouvel individu appartenant à une population déjà étudiée en lui attribuant son "adresse" dans la classification ; par

exemple savoir à quelle espèce appartient un poisson qu'on vient de découvrir. Il s'agit là d'un problème de *classement* qu'il ne faut nullement confondre avec celui de la *classification* où l'on souhaite dans ce cas, comme dit P. Dagnélie, "définir ou faire des classes".

Un objet plus élevé de la taxinomie est la connaissance d'un système (par exemple celui des vivants pour le règne animal). En cela le spécialiste peut être aidé par la détermination statistique, parmi les propriétés qui servent à décrire une population, de celles qui expliquent le mieux la classification obtenue selon le critère présenté ci-dessus.

Une telle synthèse de l'information apparaît comme un préalable très précieux pour l'étude de populations dans des domaines nouvellement explorés tels que la sociologie végétale, l'archéologie, la psychologie, la sociologie, l'économie, etc.

L'importance du temps nécessaire au spécialiste pour se familiariser avec l'objet de sa recherche montre l'intérêt d'une synthèse *automatique*, ayant recours à des algorithmes de calcul. Dans ce cadre, P. Dagnélie trouve surprenant le désintérêt des statisticiens pour les problèmes de la classification. Mais en général l'importance des données est telle que la masse des calculs à mettre en œuvre ne peut être conçue sans l'existence des calculatrices électroniques ; or il n'y a guère plus de dix ans, en France tout au moins, que le développement des ordinateurs permet d'entreprendre des calculs importants.

Se faisant l'écho de plusieurs chercheurs, l'ouvrage de R. Sokal et P.H.A. Sneath (1963) *The principles of numerical taxonomy* marquait un pas décisif dans l'introduction des méthodes automatiques. Le point de départ de chacune des méthodes proposées était le choix d'un indice ou mesure de similarité, fonction numérique sur $E \times E$, associant à chaque couple d'objets, (x, y) , x et y appartenant à E , un nombre positif, $S(x, y)$, censé mesurer la ressemblance des deux objets et établi à partir des vecteurs logiques

$$\alpha(x) = (x_1, x_2, \dots, x_i, \dots, x_T)$$

et

$$\alpha(y) = (y_1, y_2, \dots, y_i, \dots, y_T)$$

où $x_i = 1$ si l'objet x possède l'attribut α_i et 0 sinon, (de même pour y), $\{\alpha_1, \alpha_2, \dots, \alpha_T\}$ étant l'ensemble des attributs.

Au choix de la similarité succédait le choix d'un algorithme établissant une classification ou une hiérarchie de classifications susceptible de respecter les ressemblances entre objets à travers les nombres $S(x, y)$.

Il restait que seule l'intuition du spécialiste guidait le choix de la mesure de similarité et que les algorithmes proposés étaient des procédures d'approximation plus ou moins justifiées et souvent ambiguës ; néanmoins les différentes méthodes donnaient entre elles et celles des méthodes non automatiques des résultats très voisins.

Dés lors un souci de rigueur mathématique dans l'élaboration des méthodes s'imposait.

Dans un tout autre contexte, C.T. Zahn (1964) étudiait le problème de trouver une relation d'équivalence Π qui "approxime au mieux" une relation symétrique R sur un ensemble fini E , dans le sens suivant : "Minimiser le cardinal de la différence symétrique des graphes de R et de Π dans $E \times E$ ".

Ce problème illustre exactement la nature mathématique de celui de la classification automatique : la donnée de base est une structure de type σ sur l'ensemble E des objets qui est moins "riche" que celle, τ , que l'on désire (partition ou chaîne de partitions). La démarche générale consiste à injecter la classe Θ , des structures de type τ , dans celle, Σ , des structures de type σ . Il s'agit alors de déterminer celui ou ceux des éléments de la classe Θ qui sont au mieux compatibles avec la donnée de base. Pour cela on définit un *critère* ; un critère est, très généralement, une relation de préordre (qui peut être partielle) sur l'ensemble Θ . Pour cette relation τ est préférable à τ' ($\tau' < \tau$) si, en un sens qu'il y a lieu de préciser, τ est plus proche de la structure définie par la donnée de base que ne l'est τ' . Le critère est le plus souvent défini à partir de la donnée d'une métrique ou pseudo-métrique sur Σ , mais nous verrons des cas où il n'en est pas ainsi.

Chacun des caractères de description, $a_1, a_2, \dots, a_h, \dots, a_p$, présente un nombre fini de modalités et

détermine une partition de E . Celle relative au caractère a_h , notée S_h , réunit deux objets s'ils sont identiques du point de vue de a_h . S. Régnier (1965) juge d'autant plus préférable une partition X que

$$\frac{1}{p} \sum_{h=1}^p d(X, S_h) \quad (1)$$

est le plus petit ; $d(X, S_h)$ désignant le cardinal de la différence symétrique des graphes de X et de S_h dans $E \times E$. Les "partitions centrales" sont celles qui réalisent le minimum de (1).

De plus, S. Régnier définit un algorithme dit des "Transferts" qui en un nombre de pas acceptable permet d'obtenir une partition qui réalise un optimum local de (1).

Cette méthode était un net progrès par rapport aux techniques utilisées. La démarche revenait par ailleurs à adopter comme mesure de similarité de deux objets x et y la proportion des caractères qui présentent la même modalité chez x et y , S_{xy} ; à plonger le cube des relations sur E , $\{0, 1\}^{E \times E}$, dans $R^{n \times n}$ et à chercher les points "partition" de ce cube les plus voisins du point de $R^{n \times n}$ ($S_{xy} : (x, y) \in E \times E$).

Il restait que l'imperfection des données était telle qu'on ne pouvait pas considérer en général les différentes partitions S_h comme également discriminantes ou, ce qui revient au même, le choix de la similarité qui suppose également significatives de la ressemblance deux modalités quelconques, ne pouvait toujours être satisfaisant. De plus la méthode ne se généralisait pas au cas de la recherche d'une chaîne de partitions.

S'inspirant des travaux de R.N. Shepard (1962), J.P. Benzecri et W.F. de la Véga retenaient comme donnée, non plus les distances ou similarités $S(x, y)$, mais le système d'inégalités entre distances appelé ordonnance (1965), ordre total sur l'ensemble F des paires d'objets distincts de E . Conformément à cette appellation nous avons préféré distinguer ce cas de celui plus général et le plus fréquent, où la donnée est une *Préordonnance*, préordre total sur F .

Bien qu'associée au choix d'une distance ou d'une similarité, on pouvait espérer qu'une telle donnée fût plus stable. Sur une question de B. Jaulin, (1966),

nous entreprîmes d'étudier l'influence du choix de la similarité sur la préordonnance associée. Moyennant une définition axiomatique, assez générale pour inclure la totalité des indices de similarités proposés, nous montrâmes que si le nombre d'attributs possédés par un même objet de E est invariable la préordonnance associée est la même quel que soit l'indice choisi ; plus généralement si ce nombre d'attributs a une variance faible, la préordonnance associée, lorsqu'on remplace un indice par un autre, varie peu au sens d'une métrique naturelle sur l'ensemble des préordres totaux sur F . Nous terminâmes en proposant des préordonnances qui nous semblaient naturelles, et qui ne dépendaient plus d'une seule similarité mais d'un couple ou d'un triplet. Cette étude fait l'objet du chapitre 1.

La donnée d'une partition P sur E définit sur F , un préordre à deux classes R et S ; où R (resp. S) est l'ensemble des paires réunies (resp. séparées) par la partition. Tout élément de R (paire de deux objets proches) précède tout élément de S (paire de deux objets éloignés).

Cherchant à adapter l'algorithme des "Transferts" au cas où la donnée est une ordonnance ω , J.P. Benzecri proposait de chercher la partition P qui "maximise le nombre d'inégalités spécifiées par P et compatibles avec ω ". En d'autres termes, il s'agissait de maximiser $|\omega \cap R \times S|$, où ω désigne le graphe, dans $F \times F$, de la préordonnance et où (R, S) est associé à la partition P .

W.F. de la Véga préféra maximiser le cardinal du complémentaire dans $F \times F$ de la différence symétrique entre ω et $R \times S$, et envisagea le problème de la comparaison des deux critères en 1965. Ce problème, que nous reprîmes en 1967, nous a permis, avec des méthodes combinatoires, de montrer définitivement l'intérêt du critère de la différence symétrique.

Introduisant l'ensemble $\mathcal{R}$, des partitions pour lesquelles $|R| = r$ et Ω désignant l'ensemble de toutes les ordonnances sur E ; la distribution de $|\omega \cap R \times S|$, sur $\Omega \times \mathcal{R}$, muni d'une mesure uniforme est, pourvu que le cardinal de E ne soit pas trop petit, très approximativement normale. Sous ce point de vue, le critère de la Véga paraît être obtenu en centrant la

variable $|\omega \cap R \times S|$; nous proposons donc, comme nouveau critère, de réduire la variable centrée. Une telle statistique permettra en outre de tester la signification d'une classification obtenue par un algorithme.

Devant notre ignorance des exigences que doit satisfaire une classification, on est conduit parfois à limiter la confiance que l'on peut accorder à tout critère qui préordonne totalement l'ensemble des partitions. Conformément à la définition d'un ordre partiel sur l'ensemble des partitions, posée par W.F. de la Véga — définition qui utilise la notion de dominance statistique — J.P. Benzécri étudie un critère "qui ne donne pas toujours entre deux classifications une préférence ou une équivalence mais sur qui, lorsqu'il en donne, on est comme forcé de s'accorder". Ce critère dit d'ordre latéral est sûr mais partiel, son intérêt pratique reste à établir.

Nous avons bien établi deux critères d'ordre partiels qui peuvent être énoncés à propos de la recherche d'une classification mais qui sont plus naturels pour la recherche d'une hiérarchie de classifications, (chaîne de partitions), ce qui semble être l'idéal de ce que désirent les taxinomistes.

Bien que la nature profonde des techniques utilisées dans ce cas soit la même que pour la recherche d'une partition, ce problème ne se ramène pas mathématiquement au précédent. Introduisant la notion de préordonnance ultramétrique, dont la donnée est équivalente à celle d'une chaîne de partitions, nous avons généralisé les différents critères.

Nous avons établi dans ce cadre un algorithme qui est très lié à la nature de la donnée de base, préordonnance sur E . Cet algorithme que nous appelons "lexicographique" s'apparente à celui de P.H.A. Sneath (1957) ; son caractère très naturel nous a fait soupçonner une propriété d'optimalité, que nous avons cherché à expliciter. Ceci nous a conduit à définir le premier des deux critères d'ordre partiel, mentionnés ci-dessus, que nous avons désigné par "critère d'ordre lexicographique". Quant au second critère dit "du nombre d'inversions", il est plus classique, s'apparente au critère de la différence symétrique et tient son origine dans les travaux de M.G.

Kendall. Nous avons tenu à juger le premier critère du point de vue du second.

Le chapitre 2 est consacré à la définition et à l'étude des critères selon que la donnée de base est :

- a) le tableau des données $A \times E$ (où A est l'ensemble des attributs et E celui des objets) ;
- b) une relation binaire symétrique sur E ;
- c) une famille finie de partitions sur E ;
- d) une distance ou pseudo-distance, (mesure de similarité), sur E ;
- e) une préordonnance sur E .

Au chapitre 3 nous définissons d'une part l'algorithme des "transferts" ; d'autre part l'algorithme "lexicographique" et nous en étudions les propriétés.

La population E d'objets est en général un échantillon d'un ensemble E' plus vaste, qu'il est souvent impossible d'observer dans sa totalité. Par ailleurs une description exhaustive des éléments de E est encore souvent impossible ; les caractères ou attributs, retenus pour la description, constituent un échantillon d'un ensemble plus vaste de caractères ou attributs. Considérant cette situation, S. Régnier a étudié (1966) deux problèmes de convergence statistique qui permettaient de se rendre compte du comportement asymptotique des partitions centrales ; c'est ce que nous analysons dans la première partie du chapitre 4. Puis, dans le cadre où la donnée est une préordonnance ω , nous envisageons le problème de juger si, au vu des données, (ici de ω), il n'est pas "artificiel" d'imposer sur E une structure de chaîne de partitions, censée respecter les ressemblances entre objets. D'où la proposition d'une statistique susceptible de mesurer l'aptitude de E à être organisé en une hiérarchie de classifications, à être h -classifié dirons-nous ; et, la définition d'un test éprouvant l'hypothèse de la non existence d'une h -classification "naturelle" ; hypothèse que nous caractérisons ici d'une manière qui nous semble satisfaisante.

Enfin, pour terminer cette présentation, indiquons que le chapitre 0 introduit le treillis des partitions d'un ensemble fini et celui d'un entier, fixant ainsi le vocabulaire utilisé par la suite. Nous traitons dans ce chapitre de certaines questions de nature

combinatoire auxquelles nous nous sommes trouvés confrontés, notamment lors de la comparaison des critères. Ce chapitre reste essentiellement de référence.

Toutes les méthodes que nous considérons dans ce texte supposent donnée une information relative à la ressemblance des objets. Cette information est généralement établie sur les apparences sensibles, les caractères externes des objets, sans tenir compte des liaisons possibles entre ces caractères, c'est-à-dire de la structure sous-jacente qui parfois fait intervenir des informations extérieures aux objets (par exemple historiques, stratigraphiques etc.). Les classifications proposées sont donc phénotypiques. Relativement aux Sciences de la nature, on oppose à ce type de classifications celui des classifications phylogéniques, (par filiation), établies à partir de la parenté présumée des espèces. Bien que leur thèse puisse prêter à discussion, R. Sokal et P. Sneath insistent dans leur ouvrage en assurant que le seul moyen d'établir une parenté est de comparer avec le plus de soin possible les manifestations externes

[cf. pp. 55, 94, 216]. Il arrive même, disent-ils, que soit retenue une classification qui contredise ce que l'on suppose de la parenté des espèces. On croit que les crocodiles et les oiseaux ont des ancêtres communs dont ni lézard ni tortues ne descendent et pourtant on classe les crocodiles avec les autres reptiles, [cf. Sokal et Sneath p. 226]. Notre dernier mot rejoindra le premier ; la condensation de l'information, que fournissent nos classifications, permettra l'élaboration d'hypothèses et par conséquent de recherches futures.

Qu'il me soit permis de terminer en exprimant mes plus vifs remerciements à Messieurs B. Jaulin et S. Régner pour les nombreuses conversations que j'ai pu avoir avec eux et dont chacune enrichissait mes associations d'idées. Tous mes remerciements à Monsieur M. Eytan qui a contribué à la rédaction du chapitre 0 et nous a fourni une caractérisation d'une matrice de distances ultramétrique.

Je dois au Professeur J.P. Benzécri des réactions et commentaires qui m'ont été d'un grand profit.