MATHÉMATIQUES ET SCIENCES HUMAINES

B. JAULIN

Mesure de la ressemblance en anarchéologie

Mathématiques et sciences humaines, tome 26 (1969), p. 33-43
<http://www.numdam.org/item?id=MSH_1969__26__33_0>

M.S.H. (7e année, nº 26, 1969, p. 33-43)

MESURE DE LA RESSEMBLANCE EN ANARCHÉOLOGIE[1]

par

B. JAULIN[2]

RÉSUMÉ

Le mot anarchéologie a été utilisé pour la première fois, semble-t-il, par J.-C. Gardin, dans un sens légèrement différent de celui que nous lui donnons. En effet, on se préoccupe simplement ici de certaines pratiques que l'on rencontre maintenant en archéologie : il s'agit de l'usage des algorithmes de calcul qui permettent, à partir d'une description des objets rencontrés dans les fouilles, de définir des classifications de ceux-ci. Le manque de justification dans l'emploi anarchique de ces procédures peut se manifester à différents moments. L'on s'intéresse plus particulièrement aux hypothèses relatives à la détermination des inégalités entre les « ressemblances» des objets.

Les algorithmes de calcul utilisés en archéologie, bien que se rapportant à l'analyse des objets rencontrés dans les fouilles, sont souvent élaborés de façon abstraite, ou, plutôt, indépendamment de la visée des archéologues, comme s'il pouvait y avoir des techniques de calcul universelles pour les problèmes qui se posent dans cette discipline. Aussi est-il parfois difficile à l'archéologue de donner un sens aux résultats obtenus de cette façon.

Il a été question lors de ce congrès des méthodes de calcul relatives au problème de la sériation des objets et à celui de leur classification. Les questions qui se posent en archéologie à propos de la notion de type, notamment celles analysées par M. Borillo pour la statuaire archaïque grecque, montrent clairement la nécessité d'étudier les propriétés des modèles à partir desquels on définit les algorithmes permettant d'obtenir des classifications.

Les principaux algorithmes de classification dont on dispose maintenant supposent définis sur l'ensemble fini des objets une structure, par exemple une distance ou encore, un pré-ordre sur l'ensemble des paires d'objets visant à représenter les comparaisons entre les « ressemblances». Ce sont pour la plupart des procédures d'approximation : on rend compte au « mieux» de la structure donnée par une classification ou une hiérarchie de partitions. Or, la notion d'approximation correspond à une technique et non pas à un principe du genre de ceux qui permettent, en statistique par exemple, d'accorder quelque crédit ou d'interpréter les résultats obtenus.

Aussi S. Régnier a étudié le problème de savoir si des hypothèses uniquement probabilistes sur les données permettraient d'apporter une réponse au problème de la classification dans les cas qui nous intéressent. Il postule simplement que les classes que l'on recherche sont caractérisées par des distributions différentes dans l'ensemble des descriptions et montre, dans le cas non paramétrique, c'est-à-dire celui rencontré lorsque l'on utilise des codes descriptifs constitués de traits, que le modèle statistique du maximum de vraisemblance conduit à un ensemble de classification défini par une partition du nombre des objets. Il semble donc nécessaire, dans ce cas, de prendre en compte une « structure» sur l'ensemble des objets si l'on veut construire par le calcul une classification.

Toutefois si l'on considère des exemples géométriques simples, ensemble fini de points répartis dans le plan pour lesquels on connaît les inégalités entre leur distance euclidienne, il n'est pas toujours

1. Ce texte correspond à celui d'un exposé fait à l'occasion d'un congrès organisé par le C.N.R.S. sur le thème : « Emploi des calculateurs en archéologie. Problèmes semiologiques et mathématiques», Marseille, 7-12 avril 1969.
2. École Pratique des Hautes Études. 6e section.

raisonnable de construire une partition à partir d'une telle donnée. D'où le problème abordé dans (2) par I. C. Lerman, de la construction d'un test permettant de juger de la « classificabilité » d'un ensemble dont l'ensemble des couples est muni d'un pré-ordre.

Mais est-on sûr que les codes descriptifs rencontrés en archéologie permettent de définir les inégalités entre les « ressemblances » des objets ? La notion de ressemblance est en effet, du point de vue de l'archéologue, liée à une visée : on peut être ressemblant du point de vue du style, sans l'être du point de vue de l'usage ou de l'époque, etc., et, bien souvent, l'interprétation donnée aux classifications construites par le calcul correspond à une de ces préoccupations. Ainsi, on peut craindre que le souci d'« exhaustivité » qui préside dans la plupart des cas à la construction des codes soit en opposition avec la pratique des algorithmes de classification, du fait de la prise en compte de traits ayant des statuts différents vis-à-vis de ces diverses visées, et, il semble nécessaire de savoir, lorsque l'on utilise de telles procédures, si les traits que l'on considère pour décrire les objets sont en accord avec l'interprétation ou l'usage que l'on fera des classifications obtenues.

Mais que signifie « être en accord » dans ce contexte. C'est ce que nous avons tenté de préciser dans le cadre d'un modèle théorique de description qui nous a conduit à utiliser, à la façon d'un exercice, certaines notions de topologie générale. Ajoutons donc ici quelques explications afin de justifier, du point de vue de l'archéologue, la prise en compte de ces notions.

Une des caractéristiques des codes utilisés en archéologie est que les propriétés ou traits qui les constituent sont nommés dans le langage, ou, à défaut, spécifiés par un dessin pris comme modèle. L'ensemble des formes (vases, haches, etc.) qu'une telle liste de propriétés entend « coder » a un caractère de « continu » (au sens intuitif du terme) alors que l'ensemble des descriptions possibles est fini si l'on utilise un nombre fini de traits. On admet donc que chaque propriété existant dans un code peut être « affinée », c'est-à-dire est la conjonction d'un nombre fini de propriétés. Ainsi la propriété « avoir une anse » peut donner lieu aux propriétés « avoir une anse concave » et « avoir une anse convexe », chacune de ces propriétés pouvant elle-même être précisée, etc. On est donc conduit à introduire la notion de code infini hiérarchisé. Pour ceci on utilise un alphabet fini Y, $Y = \{y_1, \ldots, y_T\}$ et l'ensemble L(Y) des mots que l'on peut construire avec cet alphabet afin de nommer les propriétés constituant le code de telle sorte que la propriété de nom f (f étant un mot constitué de lettres appartenant à Y) est la conjonction des propriétés de nom fy_i, y_i appartenant à Y, fy_i étant ce mot obtenu en faisant suivre le mot f de la lettre y_i. Ainsi si $Y = \{y_1, y_2\}$ est réduit à deux éléments, on pourra figurer un tel code par un arbre du type de celui indiqué sur la figure 1., la flèche entre le sommet de nom $y_1\, y_2$ et celui de nom $y_1\, y_2\, y_1$ signifiant simplement que la propriété de nom $y_1\, y_2\, y_1$ implique la propriété de nom $y_1\, y_2$.

On voit sur ce dessin apparaître la notion de niveau d'analyse, mesurée par la longueur des mots que l'on utilise pour désigner les propriétés. Ainsi la propriété de nom $y_1\, y_1\, y_2$ est associée au troisième niveau d'analyse. Cette notion de niveau d'analyse est d'usage fréquent en archéologie et un des problèmes qui se posent dans ses termes est de savoir « à quel niveau on doit arrêter la description ». La réponse à cette question dépend évidemment de l'usage que l'on fait des codes. Si cette question est liée à la pratique des algorithmes de classification qui utilisent comme donnée initiale les inégalités entre les ressemblances des objets, on verra, dans le cadre des modèles que nous décrirons plus bas dans lesquels on donne un sens au mot ressemblance, qu'il existe un niveau d'analyse fini tel que les informations fournies par le code constitué des propriétés associées à ce niveau d'analyse permettent de définir les inégalités entre « les ressemblances » d'un nombre fini d'objets.

Cette notion de niveau d'analyse est, sous des noms variés et selon les situations, très courante. Elle intervient notamment en analyse numérique dans l'étude des problèmes d'approximation. Elle est également liée à la « perception » des formes dans le plan, dans la mesure où elle pourrait représenter l'aptitude à regarder de plus ou moins près. Ainsi si l'on ne retient d'une forme du plan que sa trace sur une « grille », et si l'on fait varier le pas de la grille, on obtiendra un code infini hiérarchisé pour lequel un trait correspond au fait qu'une forme coupe ou ne coupe pas un carré défini par la grille. Comme il fallait s'y attendre, le niveau d'analyse correspondra, ici, à la grandeur du pas de la grille (cf. fig. 2).

Ces deux exemples de code infini hiérarchisé suggèrent d'introduire une notion de ressemblance entre les objets à partir de la seule hypothèse suivante, qui semble être couramment admise dans la pratique des codes descriptifs : deux objets sont d'autant plus ressemblants, que leur description (ou trace) sont les mêmes à des niveaux d'analyse d'autant plus fins. On examinera cette hypothèse qui conduit à l'introduction d'une topologie sur l'ensemble des objets dont la métrique est « bien approximée » si l'on prend, à un niveau d'analyse fini suffisamment grand, le coefficient de la « différence symétrique

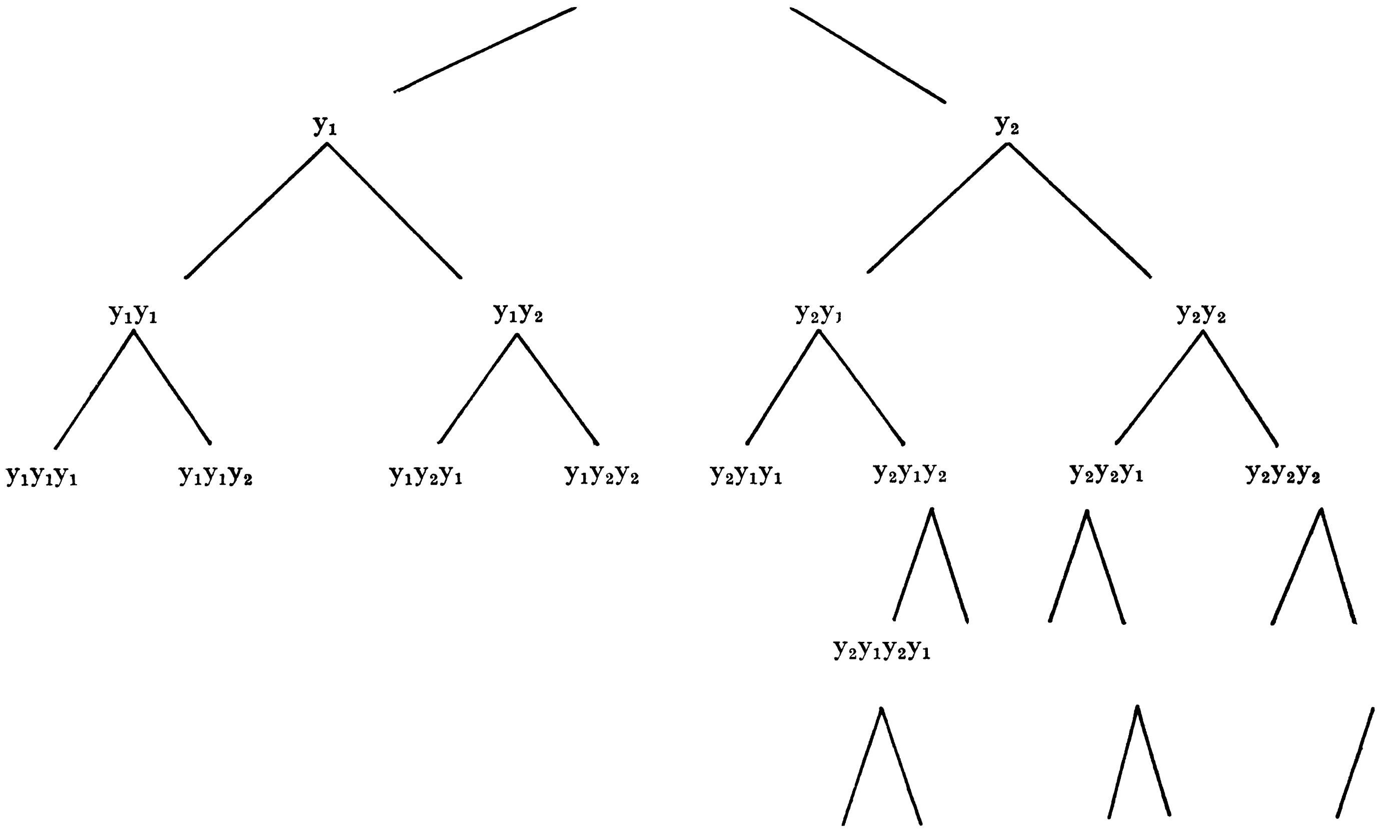

Fig. 1

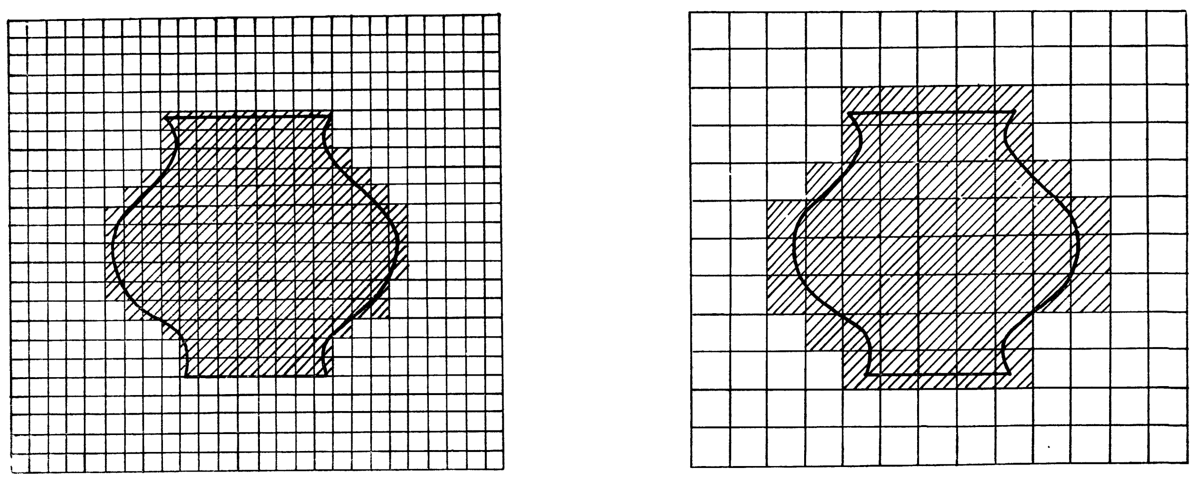

Fig. 2

cumulée» dans l'ensemble des représentations. Toutefois cette topologie est de dimension zéro et peut être définie en prenant comme seule hypothèse que tous les traits constituant le code sont tels que si x est un objet qui possède un trait, tout objet « suffisamment voisin» le possédera également et telle en outre que si x est un objet qui ne possède pas un trait, tout objet « suffisamment voisin» ne le possédera pas non plus. En fait, la topologie T_1 est la moins fine ou la moins riche pour laquelle tous les traits satisfont, vis-à-vis de la notion de voisinage qui lui est associée, ces deux propriétés. On voit apparaître ici la notion de voisinage et l'on remarque alors que cette caractérisation de la topologie T_1 ne fait intervenir que le statut des traits vis-à-vis de cette notion. La notion de ressemblance en archéologie étant subjective, on est conduit alors à introduire sur X d'autres topologies caractérisées seulement par les liens qu'entretiennent chacun des traits du code avec une notion de voisinage dont on postule l'existence sur l'ensemble des objets. Les topologies les moins fines satisfaisant à ces diverses hypothèses ne sont pas, en général, « approximables» par des topologies définissables par des écarts. Aussi étudierons-nous les cas où ces topologies coïncident avec celle correspondant à la première hypothèse mentionnée ci-dessus et l'on verra que ceci se produit « assez fréquemment».

1. — DESCRIPTION DE LA SITUATION ET EXEMPLES.

Soit X un ensemble infini et Y un ensemble fini. On utilise les éléments du monoïde libre L (Y) au-dessus de Y pour indexer une famille $(C_f)_{f \in L(Y)}$ de sous-ensemble de X (que l'on appellera parfois un code infini hiérarchisé) possédant les deux propriétés suivantes :

1) Si e est le mot vide de L (Y), $C_e = X$;

2) Pour tout $f \in L(Y)$, $C_f = \bigcup_{y_i \in Y} C_{fy_i}$

Exemple 1.

Cet exemple correspond à la situation que nous considérions dans l'introduction et qui peut se présenter en archéologie lorsque l'on utilise des propriétés que l'on « affine » pour décrire des objets. Les propriétés 1) et 2) mentionnées ci-dessus étant définies dans un modèle de la théorie des ensembles, nous devons rendre compte de cette situation dans les termes de cette théorie. Il s'agit là d'une traduction sans intérêt qui a cependant, pour notre propos, l'avantage de mettre en évidence le caractère « syntaxique » de la notion de propriété et celui « sémantique » de celle d'objet. On peut se satisfaire de ceci :

Soit U un univers, $Y = \{y_1, \ldots, y_T\}$ un ensemble fini et X un ensemble infini de U.

Soit $(T_f)_{f \in L(Y)}$ une famille de formules à une variable libre de la théorie des ensembles indexée (dans U) par des éléments de L(Y) telle que si $C_f = \{ a/a \in X \text{ et } T_f(a) \text{ est vrai dans } U \}$ on obtienne dans U une famille d'ensembles indexée par les éléments de L(Y) satisfaisant les propriétés (1) et (2), c'est-à-dire criblant l'ensemble X. (Ainsi dans la mesure où les objets considérés en archéologie sont des formes de l'espace et où les traits considérés dans la pratique sont exprimables dans la théorie des ensembles, on peut rendre compte dans cette théorie de la situation que nous avons en vue. On peut admettre ainsi qu'il existe un univers dans lequel on puisse trouver l'ensemble des haches comme élément de l'ensemble des parties de R^3 qui existe dans tout univers, et exprimer, dans cette théorie, le fait de posséder une arête elliptique... Toutefois, on ne voit absolument pas à quel élément d'un univers correspondrait la population des « spartiates » (à moins de les considérer comme « urs éléments »), ni comment exprimer dans la théorie des ensembles, le fait pour une statue, d'avoir une hanche marquée ou non. Aussi faut-il être prudent quant à l'adéquation de ce qui va suivre avec la situation rencontrée dans la pratique des codes réels. Les exemples qui suivent, n'ont pas, de ce point de vue là, de caractère ambigu ; les objets et les propriétés qu'ils mettent en cause sont bien du domaine de la théorie des ensembles).

Exemple 2.

L'origine de cet exemple est dans les recherches entreprises par les spécialistes des machines à calculer qui se préoccupent de « reconnaissance des formes ». (On dispose de « photocole », appareil qui ne retient d'une forme dans le plan, que sa trace sur une grille de pas variable.)

Soit $X = \mathcal{P}_f([0, 1])^2$ l'ensemble des parties fermées (pour la topologie usuelle du plan) du carré ayant pour côté le segment fermé [0, 1].

Soit $Y = \{ a, b, c, d \}$. Divisons ce carré en quatre autres (« égaux » entre eux et fermés) et nommons chacun d'entre eux par une lettre de Y. Recommençons ce découpage avec chacun des pavés obtenus... (cf. fig. 3).

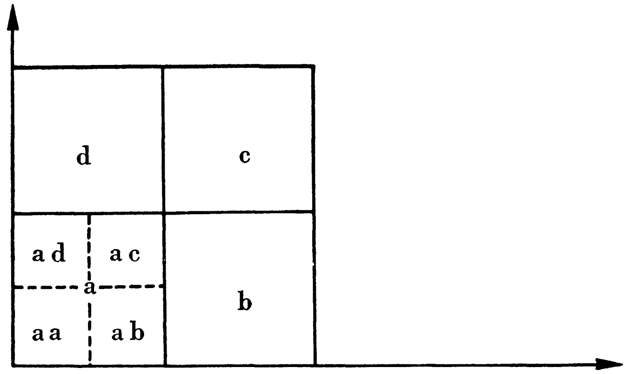

Fig. 3

Posons alors $C_f = \{ X/X \subset [0, 1]^2$ est une partie fermée qui coupe le carré de nom f $\}$. Cet exemple a une parenté avec le précédent. L'ensemble X que l'on considère est l'ensemble des « formes » de $[0, 1]^2$ et les propriétés qui servent à les décrire correspondent au fait qu'une forme coupe ou ne coupe pas un carré. L'on vérifie aisément que les propriétés (1) et (2) sont bien satisfaites pour cette famille $(C_f)_{f \in L(Y)}$.

Exemple 3.

Soit : $X = 2^N$ l'ensemble des parties de N, $2 = \{ 0, 1 \}$.
Soit : $Y = \{ a, b \}$.
On désigne par $f(i)$ la i^{eme} lettre du mot $f \in L(Y)$.
Soit alors : $C_f = \{ A \mid A \subset 2^N$, pour tout $i \leqslant l(f)$, $i \in A \Leftrightarrow f(i) = a \}$.
On rencontre là encore un code infini hiérarchisé.

Remarque. — Si l'on désigne par Y_n l'ensemble des mots de n lettres construit sur l'alphabet Y et $p_n : Y_{n+1} \rightarrow y_n$ la surjection qui, à un mot de $n + 1$ lettres associe le mot constitué des n premières lettres de ce mot — la famille $F = (Y_n, p_n)_{n \geqslant 0}$ est appelée un crible. Cette notion est utilisée, notamment en analyse numérique, pour l'étude de problèmes d'approximation dans des espaces métriques. (On associe à tout élément f de Y_n un fermé C_f de l'espace métrique, de telle sorte que les propriétés (1) et (2) ci-dessus, soient vérifiées.)

2. — PREMIÈRE HYPOTHÈSE SUR LA NOTION DE RESSEMBLANCE.

Les divers exemples précédents suggèrent d'introduire la terminologie suivante. Pour $i \in N$, (N désignant l'ensemble des entiers naturels), on appelle niveau d'analyse i, la famille $(C_f)_{f \in L(Y) \text{ et } l(f) = i}$, $l(f)$ étant la longueur du mot f. D'autre part, on entendra par description au niveau i, l'application : $f_i : X \rightarrow 2^{Y^i}$ où $f_i(x) = \{ h/l(h) = i, h \in L(Y), x \in C_h \}$. Y^i désigne l'ensemble des mots de i lettres, $i \in N$.

On dira que x et $x' \in X$ ont même description au niveau d'analyse i, $x \underset{i}{\equiv} x'$, si et seulement si $f_i(x) = f_i(x')$.

Remarquons ici que si $i < j$, $x \underset{j}{\equiv} x' \Rightarrow x \underset{i}{\equiv} x'$. Ainsi la notion intuitive de « finesse d'analyse » est mesurée par la longueur des mots f qui servent à nommer les propriétés considérées à un niveau d'analyse. Dans le cas de l'exemple 2, la longueur de ces mots définit la grandeur du « pas de la grille » et mesure, en un certain sens, la précision avec laquelle on regarde une forme. D'autre part, dans cet exemple, deux formes dans le carré ont même description au niveau i, si elles coupent les mêmes carrés, c'est-à-dire si elles ont même trace sur la grille dont le pas vaut $\frac{1}{2^i}$.

La première idée venant à l'esprit (notamment dans le cas de l'exemple 1) pour définir une notion de ressemblance sur les objets consiste à supposer que deux objets sont d'autant plus voisins ou semblables qu'ils ont des descriptions identiques à des niveaux d'analyse d'autant plus grand. Aussi dira-t-on que deux objets x et y sont « voisins d'ordre i », $i \in N$, si ils ont des descriptions identiques, $x \underset{i}{\equiv} y$, au niveau i d'analyse. Cette notion de voisinage est celle associée à la structure uniforme U sur X dont une base fondamentale d'entourage est constituée de la famille $(V_i)_{i \in N}$ où :

$$V_i = \{ (x, y) \,/\, x \in X, y \in X, f_i(x) = f_i(y) \}.$$

Si l'on désigne par U_i la structure uniforme sur X dont une base fondamentale d'entourage est constituée du seul sous-ensemble V_i de $X \times X$, on constate que U est borne supérieure des $(U_i)_{i \in N}$. Or la topologie T_i^1 associée à U_i peut être définie par l'écart :

$$e_1^i(x, y) = \frac{| f_i(x) (\oplus) f_i(y) |}{| Y |^i}$$

où : $| f_i(x) (\otimes) f_i(y) |$ désigne le cardinal de la différence symétrique dans Y^i des sous-ensembles $f_i(x)$ et $f_i(y)$.

Par suite, la topologie T_1 associée à la structure uniforme U est définie par l'écart :

$$e_1(x, y) = \sum_{i \in N} \frac{1}{2^i} e_1^i(x, y)$$

(T_1 est borne supérieure des T_i^1, $i \in N$ et s'identifie avec la topologie induite sur la diagonale de $\prod_{i \in N} X_i$, où X_i est X muni de la topologie T_i^1).

Cette hypothèse sur la notion de ressemblance des objets est-elle suffisante du point de vue de la pratique des codes en archéologie ? N'est-il pas nécessaire de considérer d'autres écarts non équivalents à e_1. Afin de discuter cette question, nous étudierons en premier lieu les propriétés de la topologie T_1, puis nous nous référerons à l'exemple (2) pour constater que, bien souvent, les traits utilisés dans les codes en archéologie ont un statut, vis-à-vis de la notion de voisinage associée à T_1, différent de celui qui leur est donné par cette topologie.

3. — PROPRIÉTÉS DE LA TOPOLOGIE T_1.

On dira que le code infini hiérarchisé $(C_f)_{f \in L(Y)}$ est discriminant ou fidèle si $\bigcap_{i \in N} V_i = \Delta$ où Δ désigne la diagonale de X x X. Puisque chacun des V_i, $i \in N$, s'écrit sous la forme $\bigcup_{\alpha \subset Y^i} f_i^{-1}(\alpha) \times f_i^{-1}(\alpha)$, dire que $(C_f)_{f \in L(Y)}$ est discriminant, c'est dire que pour tout couple d'objets x, y $\in$ X, il existe un niveau d'analyse i, $i \in N$ tel que $x \underset{i}{\not\equiv} y$, c'est-à-dire tel que x et y n'ont pas même description à ce niveau d'analyse. Dans ces conditions, la topologie T_1 est séparée, l'écart e_1 définit précédemment est donc une distance. (Remarquons que pour l'exemple (2), la famille $(C_f)_{f \in L(Y)}$ est discriminante, cela résulte du fait que si A et B sont deux fermés distincts de $[0, 1]^2$, il existe $x \in A$, $x \notin B$ (par exemple) tel que $d(x, B) = \min_{y \in B} d(x, y)$ soit strictement positif, d désignant la distance euclidienne habituelle.)

Si l'on considère alors la topologie T_i^1, $i \in N$, on constate que les ouverts pour cette topologie sont des réunions finies de sous-ensembles de la forme, $f_i^{-1}(\alpha)$, $\alpha \subset Y^i$, puisque U_i admet pour base fondamentale d'entourage le seul entourage $V_i = \bigcup_{\alpha \subset Y^i} f_i^{-1}(\alpha) \times f_i^{-1}(\alpha)$. Par suite les sous-ensembles C_f et $\complement C_f$ sont ouverts car ce sont des réunions de $f_i^{-1}(\alpha)$ et comme $f_i^{-1}(\alpha) = \left(\bigcap_{f \in \alpha} C_f \right) \cap \left(\bigcap_{f \notin \alpha} \complement C_f \right)$ la topologie T_i^1 est engendrée par les C_f et les complémentaires des C_f, $f \in Y^i$.

Ainsi la topologie T_1, borne supérieure de T_i^1 est engendrée par les C_f et les complémentaires de C_f, $f \in L(Y)$.

Cette topologie T_1 admet une base $\mathcal{C}_1$ dénombrable (puisque Y est fini) constituée d'ouverts fermés de la forme $f_i^{-1}(\alpha)$, car si $f_i^{-1}(\alpha) \cap f_j^{-1}(\beta) \neq 0$ alors, comme on le vérifiera aisément :

$$f_i^{-1}(\alpha) \cap f_j^{-1}(\beta) = f_j^{-1}(\beta) \text{ si } j > i.$$

T_1 est donc totalement discontinue et homomorphe à la topologie d'un sous-ensemble de l'ensemble triadique de Cantor $2^{\mathbf{N}}$, c'est-à-dire à la topologie d'un sous-ensemble de $2^{\mathbf{N}}$ muni de la topologie produit, $2 = \{0, 1\}$ étant muni de la topologie discrète.

En effet, si $D_1, D_2, \ldots$ sont les ouverts fermés de la base $\mathcal{C}_1$, et si l'on pose $f_i(x) = 1$ si $x \in D_i$, $f_i(x) = 0$ si $x \notin D_i$, $f_i : X \to \{0, 1\}$ est continue si $\{o, 1\}$ est muni de la topologie discrète, donc $f = \prod_{i \in \mathbf{N}} f_i : X \to 2^{\mathbf{N}}$ est continue, $2^{\mathbf{N}}$ étant muni de la topologie produit. En outre f est injective si la famille $(C_f)_{f \in L(Y)}$ est discriminante et $f^{-1} : f(X) \to X$ est continue car :

$$(f^{-1})^{-1}(D_i) = \{\vec{x} / \vec{x} \in 2^{\mathbf{N}}, pr_i(\vec{x}) = 1\} \cap f(X).$$

On remarquera que dans l'exemple (3), la topologie T_1 est exactement celle de cet espace.

Pour étudier la compacité de T_1 on utilise le lemme suivant faisant intervenir les notions intuitives d'arbre, de branche d'un arbre, de niveau d'un sommet dans l'arbre, etc. « Si A est un sous-ensemble quelconque de l'ensemble des sommets d'un arbre infini de type fini tel que pour toute branche B, $\overline{B} \cap A \neq \varnothing$ ($\overline{B}$ désignant l'ensemble des sommets de la branche B), alors il existe $n \in \mathbf{N}$, tel que pour toute branche B, $\overline{B} \cap A_n \neq \varnothing$, où $A_n \subset A$ désigne les éléments de A ayant un niveau inférieur à n [1]. »

Si l'on désigne alors par Z l'arbre (de type fini car Y est fini), dont les sommets correspondent aux éléments de la base $\mathcal{C}_1$, la relation considérée étant l'inclusion entre ces sous-ensemble de X, on remarque que pour que T_1 soit quasi compacte, il faut et il suffit que pour toute branche de Z, l'intersection des ouverts fermés associés aux sommets de cette branche soit non vide. La condition est, évidemment nécessaire (par définition de la compacité), elle est suffisante car à tout recouvrement ouvert de X correspond dans ces conditions un ensemble A de sommets de l'arbre Z qui coupe toute branche, donc en vertu du lemme précédent, il existe un sous-recouvrement fini. On dira dans ce cas là, que le code est quasi-compact.

On en déduit : si T_1 est compacte et sans point isolé, elle est homéomorphe à la topologie de l'ensemble triadique de Cantor $2^{\mathbf{N}}$. En effet, dire qu'elle est sans point isolé, c'est dire que pour tout niveau d'analyse i, aucune classe pour l'équivalence $\underset{i}{\equiv}$ n'est réduite à un seul élément de X. Par suite, toutes les branches de l'arbre Z sont infinies et correspondent à un et seul point de X (quasi-compacité et séparation de T_1) et il n'existe pas de branches de Z telle que de tout sommet ne partirait qu'une seule sous-branche. (Par suite, comme Y est fini, on peut construire (par regroupement à chaque niveau dans l'arbre Z), une base $\mathcal{C}_1$ pour la topologie T_1 constituée d'ouverts fermés tels que l'arbre Z associé vérifie les conditions de séparation et de quasi-compacité et en outre est telle que de chaque sommet partent deux sous-branches. Ainsi si l'on numérote convenablement les éléments de cette base, on vérifie aisément que l'application $f : X \to 2^{\mathbf{N}}$ définie précédemment est un homéomorphisme).

Remarquons que l'on peut obtenir très facilement des topologies T_1 non compactes. Ainsi, si $X \subset]0,1[$ est constitué des éléments rationnels qui ne sont pas de la forme $\frac{a}{2^n}$, on obtient la topologie

1. Ce lemme est une conséquence immédiate du fait qu'un produit d'espace compact est compact. En effet, si Z est un arbre de type fini et si A_i désigne l'ensemble des sommets au niveau i, alors $A = \prod_{i \in \mathbf{N}} A_i$ muni de la topologie produit est compact lorsque chacun des A_i est muni de la topologie discrète, puisque A_i est fini. Or l'ensemble X des branches de l'arbre Z s'identifie avec un sous ensemble $\hat{X}$ fermé dans A, donc compact (car si $\vec{x} \notin \hat{X}$, compte tenu du fait qu'une base de la topologie de A est constituée des sous-ensembles de la forme $\hat{w} = \{\vec{x} / \vec{x} \in \prod_{i \in \mathbf{N}} A_i \, \vec{x} > w\}$ où w désigne un élément de $\prod_{i=1}^{n} A_i$ et $\vec{x} > w$ signifiant que $pr_j \vec{x} = pr_j w$ pour $j = 1$ à n, il existe $w \in \prod_{i=1}^{n} A_i \, \vec{x} > w$ et w n'est pas le commencement d'une branche de l'arbre, donc $\hat{w} \subset \complement \hat{X}$). Vu la nature de cette base de A, expliciter la propriété de compacité de $\hat{X}$ conduit à écrire le lemme.

usuelle de l'ensemble Q des rationnels en découpant en deux « morceaux égaux » à chaque niveau. De façon générale, on peut montrer, à l'aide d'un procédé analogue à celui utilisé précédemment et du fait (supposé connu) que la topologie induite sur un sous-ensemble dénombrable dense de 2^N qui ne contient pas de suite (de 0 et de 1) stationnaire est celle de Q, que si X est dénombrable et si la topologie T_1 est séparée et sans point isolé, elle est homéomorphe à celle de Q.

On vérifiera, pour les exemples (2) et (3) que la topologie T_1 est compacte et sans point isolé. Elle est donc homéomorphe à celle de l'ensemble triadique de Cantor 2^N.

Ainsi l'hypothèse qui a servi à introduire l'écart e_1 conduit, dans les deux cas que nous venons de considérer, soit à la topologie de l'ensemble triadique de Cantor, soit à celle de la droite rationnelle Q. En fait, plus généralement, si l'on impose simplement au code d'être discriminant, la topologie que l'on obtient est celle d'un sous-ensemble de 2^N (elle serait donc de dimension 0). Or, ceci n'est qu'une conséquence simple du fait que la topologie T_1 est engendrée par les C_f et les complémentaires des C_f, $f \in L(Y)$. Aussi notre propos maintenant est de considérer d'autres topologies sur X que l'on peut associer naturellement à un code infini hiérarchisé et pour lesquelles les traits T_f considérés dans le code auront, vis-à-vis de la notion de voisinage, d'autres propriétés que celles que nous venons de mentionner.

Auparavant, considérons certaines propriétés utilisées en archéologie pour décrire les formes dans le plan et regardons, à titre de curiosité et peut-être aussi pour justifier la définition des modèles que nous introduirons dans le paragraphe suivant, leur statut vis-à-vis de la topologie T_1 définie pour l'exemple (2). Ainsi on se rend compte aisément que les propriétés suivantes « être connexe », « avoir un trou » au sens intuitif, c'est-à-dire présenter une cavité qui contiendra un carré à un niveau d'analyse suffisamment grand, « être une ligne polygonale convexe ou concave de moins de *n* sommets », « être un arc de cercle », etc., définissent des ensembles de « formes » qui, pour la topologie T_1, sont des fermés d'intérieur vide. En effet, les sous-ensembles de $\mathcal{P}_f\,[0, 1]^2$ dont les éléments satisfont la négation de ces propriétés, sont voisinages de chacun de leur point, car si $x \in \mathcal{P}_f\,[0, 1]^2$ ne satisfait pas l'une de ces propriétés, il existe un niveau d'analyse i tel que toute forme qui aura même description que x à ce niveau, n'aura pas cette propriété et, d'autre part, ces ensembles sont denses pour T_1, car les $f_i^{-1}(\alpha)$, $\alpha \subset Y^i$, $i \in N$, constituent une base pour cette topologie et que pour toute description α, il existe une forme ayant cette description qui ne possède pas l'une de ces propriétés. Ainsi, si l'on utilise ces propriétés dans un code descriptif pour représenter les fermés de $[0, 1]^2$, on sera conduit, si on leur fait jouer le rôle d'ouverts fermés, à une topologie T_1' sur $\mathcal{P}_f\,[0, 1]^2$ différente de celle obtenue dans le cadre de l'exemple 2.

Naturellement toutes les propriétés utilisées dans les codes descriptifs n'ont pas un statut, vis-à-vis de la topologie T_1, de fermé d'intérieur vide. Ainsi « être un fermé dont la frontière est une ligne continue », ou encore, ce qui correspond aux traits caractérisés par des dessins pris comme modèles, « être un fermé dont la frontière contient une ligne continue homothétique à une ligne donnée », « être un fermé d'intérieur vide », etc. Toutefois, on a pu constater lorsque les traits utilisés dans la pratique ont une définition suffisamment précise, que ceux-ci ne définissent pas des ouverts fermés pour la topologie T_1 de l'exemple (2). (Ainsi les ensembles de formes associées aux trois propriétés mentionnées ci-dessus ne sont pas fermés pour la topologie T_1, mais sont d'intérieur vide.) De façon générale, on peut dire que les traits considérés pour la description des formes du plan sont tels que leur négation a rarement le même statut du point de vue de la topologie T_1, que les propriétés elles-mêmes.

4. — AUTRES HYPOTHÈSES SUR LES RAPPORTS ENTRE UN CODE DESCRIPTIF ET UNE NOTION DE RESSEMBLANCE DÉFINIE SUR L'ENSEMBLE DES OBJETS.

On va considérer ici deux topologies T_2 et T_3 associées respectivement aux deux hypothèses suivantes :

a) Tout objet x « suffisamment voisin » d'un objet x′ qui possède la propriété C_f, $f \in L(Y)$, possède également la propriété C_f.

b) Tout objet x « suffisamment voisin » d'un objet x′ qui ne possède pas la propriété C_f, $f \in L(Y)$, ne possède pas non plus la propriété C_f.

Ainsi T_2 est la topologie sur X engendrée par les C_f, $f \in L(Y)$, et T_3 celle engendrée par les complémentaires des C_f.

Comme pour la topologie T_1, on constate que chacune de ces topologies est borne supérieure de topologies sur X : T_j, $j = 2, 3$, est borne supérieure des topologies T_j^i, $i \in N$, $j = 2, 3$, ou T_2^i (resp. T_3^i) est la topologie sur X engendrée par les C_f, $f \in L(Y)$ et $l(f) = i$ (resp. $\complement C_f$, $f \in L(Y)$ et $l(f) = i$).

Toutefois, en général, les topologies T_j^i, $j = 2, 3$, $i \in N$ ne sont pas définissables par un écart [1]. (On sait en effet que pour qu'il en soit ainsi, il est nécessaire et suffisant, Y étant fini, que T_j^i admette une base qui soit une partition de X) [1].

Les topologies T_2 et T_3 sont moins fines que la topologie T_1. Elles peuvent cependant lui être identiques. C'est notamment le cas lorsque pour tout x, tout $y \in X$, $|f_i(x)| = |f_i(y)| = k_i$ (ainsi dans l'exemple (3), on a $k_i = 1$ pour tout i). En effet, dans ces conditions, on a l'équivalence suivante pour tout $f \in L(Y)$, $\alpha \subset Y_i$:

C_f ouvert $\Leftrightarrow f_i^{-1}(\alpha)$ ouvert $\Leftrightarrow \complement C_f$ ouvert, et par suite $T_1^i = T_2^i = T_3^i$. (Il est trivial que si les $f_i^{-1}(\alpha)$ sont ouverts C_f et $\complement C_f$ sont ouverts puisque ces sous-ensembles sont réunions de $f_i^{-1}(\alpha)$. D'autre part si C_f est ouvert, alors $f_i^{-1}(\alpha)$ est ouvert, car $f_i^{-1}(\alpha) = \left(\bigcap_{f\in\alpha} C_f\right) \cap \left(\bigcap_{f\notin\alpha} \complement C_f\right)$ et $\bigcap_{f\in\alpha} C_f \subset \bigcap_{f\notin\alpha} \complement C_f$. Cette dernière inclusion provient du fait que s'il n'en était pas ainsi, il existerait $x \in X, |f_i(x)| > k_i = |\alpha|$ puisque alors $\complement\left(\bigcap_{f\notin\alpha} \complement C_f\right) \cap \left(\bigcap_{f\in\alpha} C_f\right)$ serait non vide et que $\complement\left(\bigcap_{f\notin\alpha} \complement C_f\right) = \bigcup_{f\notin\alpha} C_f$.

Ainsi, dans ces conditions, $f_i^{-1}(\alpha) = \bigcap_{f\in\alpha} C_f$ est ouvert si les C_f sont ouverts, Y étant fini. La démonstration de l'implication $\complement C_f$ ouvert pour tout $f \Rightarrow f_i^{-1}(\alpha)$ ouvert pour tout $\alpha \subset Y^i$ est à peu de chose près la même que la précédente.

Un autre cas où les trois topologies T_1, T_2, T_3 coïncident est lorsque T_2 et T_3 sont séparées et que T_1 est compacte puisque T_2 et T_3 sont moins fines que T_1. Or si l'on dit que x est i-inclus dans y, $x \underset{i}{\subset} y$, si $f_i(x) \subset f_i(y)$, on vérifie aisément que T_2 et T_3 sont simultanément séparées si et seulement si la condition suivante est satisfaite :

a) Pour tout x, tout $y \in X$, il existe un niveau d'analyse $i \in N$ tel que x ne soit pas i-inclus dans y, $x \underset{i}{\not\subset} y$.

(Intuitivement, cela signifie qu'il n'existe pas deux objets x, y de X tels que, quel que soit le niveau d'analyse i, les propriétés possédées par l'objet x soient également possédées par l'objet y. Ainsi les topologies T_2 et T_3, dans le cas de l'exemple (2) ne sont pas séparées, il suffit pour s'en rendre compte de prendre deux fermés dont l'un est inclus, au sens de $[0, 1]^2$, dans l'autre.)

Cette condition est suffisante car si pour tout i, $x \underset{i}{\subset} y$, alors comme les topologies T_2 et T_3 admettent respectivement comme base les intersections finies des C_f et des $\complement C_f$, si $x \in C_{g_1} \cap \ldots \cap C_{g_n}$ (resp. $y \in \complement C_{g_1} \cap \ldots \cap \complement C_{g_n}$) on a $y \in C_{g_1} \cap \ldots \cap C_{g_n}$ (resp. $x \in \complement C_{g_1} \cap \ldots \cap \complement C_{g_n}$) pour tout système fini de $g_i \in L(Y)$. La réciproque est évidente également.

Remarquons alors que si T_2 et T_3 sont séparées, il en va de même de T_1. Ainsi, si un code est quasi-compact et si la condition (*a*) est satisfaite, les trois topologies T_1, T_2, T_3 coïncident.

---

1. Dans la première version de ce texte, on considérait les fonctions :

$$e_2^i(x, y) = \frac{|Y|^i - |f_i(x) \cap f_i(y)|}{|Y|^i}$$

et

$$e_3^i(x, y) = \frac{|Y|^i - |\complement f_i(x) \cap \complement f_i(y)|}{|Y|^i}$$

ces deux fonctions satisfont l'inégalité triangulaire et sont symétriques, mais ne sont pas telles que $e_j^i(x, y) = 0 \Rightarrow x = y, j = 2, 3$. Toutefois elles sont associées aux topologies T_j, $j = 2, 3$, en ce sens que les ouverts pour ces topologies sont déterminés par les boules $B_j^i(a, r) = \{ y / e_j^i(x, y) \leqslant r \}$, mais ces fonctions ne sont pas continues pour ces topologies. (Je suis redevable à M. P. Soury et à M. Eytan de cette remarque.)

Notons que, en général, les trois topologies T_1, T_2, T_3 peuvent être distinctes les unes des autres. Ainsi dans le cas de l'exemple (2), la topologie T_3 est moins fine que la topologie naturelle associée à la distance ρ de Hausdorff définie sur $\mathcal{P}_f\,[0, 1]^2$ alors que ni T_2 ni T_1 ne lui sont comparables. (On rappelle que $\rho\,(A, B) = \max\,[\delta\,(A, B),\ \delta\,(B, A)]$ ou $\delta\,(A, B) = \sup_{x \in A} d\,(x, B)$ avec $d\,(x, B) = \inf_{y \in B} d\,(x, y)$, d dans cette dernière expression désignant la distance euclienne habituelle et que, la topologie associée à cette distance admet, puisque $[0, 1]^2$ est compacte, une base constituée des parties de la forme $\mathcal{P}_f\,[0, 1]^2 \cap \mathcal{P}\,(0)$ et $\complement\,\mathcal{P}\,(F) \cap \mathcal{P}_f\,[0, 1]^2$ où 0 et F désignent respectivement un ouvert et un fermé de $[0, 1]^2$.)

Remarquons, dans le cadre de cet exemple (2), que si l'on considère un sous-ensemble $A \subset \mathcal{P}_f\,[0, 1]^2$ de fermés de $[0, 1]^2$ (en supposant qu'il en existe un) satisfaisant les deux propriétés suivantes :

(1) A est compact pour la topologie de Hausdorff;

(2) Pour tout x, tout $y \in A$, x n'est pas inclus dans y (l'inclusion étant prise dans $[0, 1]^2$), alors les trois topologies T'_1, T'_2, T'_3 définies par la restriction $(C'_f)_{f \in L(Y)}$ du code $(C_f)_{f \in L(Y)}$ à A, $C'_f = \{\, x \,/\, x \in A$ et x est un fermé de $[0, 1]^2$ qui coupe le carré de nom f $\}$, coïncident et en outre, ces trois topologies sont identiques à la restriction T_A de la topologie de Hausdorff à A. En effet, la condition (2) implique que T'_2 et T'_3 sont séparées, ainsi que T'_1.

D'autre part, la condition (1) implique que le code $(C'_f)_{f \in L(Y)}$ est quasi-compact. En effet si f'_i désigne la restriction à A de l'application f_i, $i \in \mathbf{N}$, alors $f_i'^{(-1)}\,(\alpha)$ est un fermé pour la topologie de Hausdorff sur $\mathcal{P}_f\,[0, 1]^2$ car $f_i'^{(-1)}\,(\alpha) = f_i^{-1}\,(\alpha) \cap A$ et si $\alpha = \{\, g_1, \ldots, g_p \,\}$, $g_k \in L(Y)$ $1\,(g_k) = i$ et $B \subset [0, 1]^2$ est la réunion des carrés fermés de nom $g_1, \ldots, g_p$, alors $f_i^{-1}\,(\alpha) = \{\, x \,/\, x$ est un fermé de $[0, 1]^2$ (donc de B), $x \subset B \,\}$. Or la topologie de Hausdorff sur $f_i^{-1}\,(\alpha)$ est compacte puisque B est compact et cette topologie n'est pas autre chose que la restriction à $f_i^{-1}\,(\alpha)$ de la topologie de Hausdorff sur $\mathcal{P}_f\,[0, 1]^2$. Par suite, comme la topologie de Hausdorff sur $\mathcal{P}_f\,[0, 1]^2$ est compacte la condition de quasi compacité donnée à la page 39 est satisfaite pour le code $(C'_f)_{f \in L(Y)}$. Les topologies T'_1, T'_2, T'_3 coïncident donc. Or la topologie T'_3 coïncide avec la topologie T_A, restriction à A de la topologie de Hausdorff sur $\mathcal{P}_f\,[0, 1]^2$, car $\complement\, C'_f = \{\, x \,/\, x \in A$, x est un fermé qui est contenu dans l'ouvert complémentaire du carré fermé de mot f $\}$ est ouvert pour T_A et que T'_3 est séparée et T_A est compacte. Par suite $T'_1 = T'_2 = T'_3 = T_A$.

5. — REMARQUES FINALES.

Ainsi dans ces différents cas, les hypothèses qui ont servi à introduire les topologies T_2 et T_3 sont équivalentes à celle suivant laquelle « deux objets sont d'autant plus ressemblants que leur description sont les mêmes à des niveaux d'analyse d'autant plus fins ». Dans ces conditions il apparaît justifié de privilégier l'écart :

$$e_1\,(x, y) = \sum_{k \in \mathbf{N}} \frac{1}{2^k} \frac{|\, f_i\,(x)\,(\oplus)\,f_i\,(y) \,|}{|\, Y \,|^k}.$$

Si l'on désigne par $d'_k\,(x, y) = \sum_{p=1}^{k} \frac{1}{2^p}\, e_1^p\,(x, y)$, on se rend compte aisément qu'il existe $j \in \mathbf{N}$ tel que :

$$d'_j\,(x, y) < d'_j\,(z, t) \Leftrightarrow e_1\,(x, y) < e_1\,(z, t).$$

Mais il faut noter que ces inégalités sont relatives à l'écart e_1; il se peut qu'il existe des écarts e'_1 équivalents à e_1 (c'est-à-dire définissant la même structure uniforme sur X) qui ne déterminent pas le même système d'inégalités entre les couples d'objets. A fortiori ceci peut se produire si l'on impose seulement à l'écart e_1 de définir la même topologie que l'écart e_1. Ainsi si l'on désigne par :

$$h : X \to 2^{L(Y)} \text{ l'application } h = \prod_{f \in L(Y)} h_f$$

où :

$$h_f : X \to \{\, 0, 1 \,\}, \qquad h_f\,(x) = 1 \Leftrightarrow x \in C_f,$$

on remarque que la topologie T_1 sur X est la moins fine qui rende continue l'application h, $2^{L(Y)}$ étant muni de la topologie produit, $2 = \{0, 1\}$ étant muni de la topologie discrète [1]. Si l'on pose alors $e'_1(x, y) = d(h(x), h(y))$ où $d(\vec{u}, \vec{v})$, $\vec{u}, \vec{v} \in 2^{L(Y)}$ étant défini, si l'on désigne par r(f), $f \in L(Y)$, le rang de f dans l'ordre lexicographique sur L(Y) obtenu à partir d'un ordre total sur Y :

$$d(\vec{u}, \vec{v}) = \begin{cases} 0 \text{ si } \vec{u} = \vec{v} \\ \dfrac{1}{r(f) + 1} \quad \text{si } \vec{v} \neq \vec{u} \end{cases}$$

(où $f \in L(Y)$ est l'élément de plus petit rang tel que $pr_f \vec{u} \neq pr_f \vec{v}$).

On constate que e'_1 est un écart qui définit sur X la même structure uniforme que e_1 et, cependant, que les inégalités entre les couples d'éléments de X définies par e_1 et e'_1 ne sont pas les mêmes. Enfin, pour terminer, remarquons que les trois topologies T_1, T_2, T_3 que nous avons considérées sont les moins fines qui satisfont aux hypothèses que nous avons posées et que rien donc ne permet d'affirmer que ce soit l'une de celles-là qui rende compte de la notion de voisinage dont on postule l'existence sur l'ensemble des objets. Ainsi le point de vue que nous avons pris ne permet pas vraiment de justifier le rôle central que l'on fait jouer à l'écart e_1 dans la pratique des codes descriptifs.

BIBLIOGRAPHIE

[1] BORILLO Mario. — « Le problème de la vérification d'hypothèse en archéologie. Éléments de méthode. » Proceedings du congrès sur « les applications des calculateurs en archéologie. Problèmes sémiologiques et mathématiques ». Editions du C.N.R.S. (à paraître.)

[2] LERMAN César. — « h-classificabilité. » Proceedings du congrès sur « les applications des calculateurs en archéologie. Problèmes sémiologiques et mathématiques ». Éditions du C.N.R.S. (à paraître.)

[3] RÉGNIER Simon. — « Non fécondité du modèle statistique général de la classification automatique. » Proceedings du congrès sur « les applications des calculateurs en archéologie. Problèmes sémiologiques et mathématiques ». Éditions du C.N.R.S. (à paraître.)

1. Si l'on munit $\{0,1\}$ de la topologie où $\{0\}$ est fermé et $\{1\}$ est ouvert, on obtiendra dans ces conditions la topologie T_2. En faisant jouer à $\{0\}$ le rôle d'ouvert et à $\{1\}$ le rôle de fermé, on obtiendrait la topologie T_3.