JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

JEAN-PAUL VALOIS

Approche graphique en analyse des données

Journal de la société française de statistique, tome 141, n° 4 (2000), p. 5-40

<http://www.numdam.org/item?id=JSFS_2000__141_4_5_0>

APPROCHE GRAPHIQUE EN ANALYSE DES DONNÉES

Jean-Paul VALOIS [1]

RÉSUMÉ

Les méthodes graphiques font historiquement partie intégrante de l'analyse des données; elles peuvent jouer un rôle pratique important, notamment en aidant à acquérir les intuitions préalables à l'adoption d'un modèle ou en facilitant le contrôle de son opportunité. Pour le statisticien, les clefs d'un graphique sont le nombre et le type des variables prises en compte, et l'objectif recherché : description globale de la population, repérage des unités statistiques, ou comparaison à un modèle. Plusieurs systèmes de représentation sont possibles, en coordonnées orthonormées, parallèles, ou radiales, pouvant parfois se combiner. La typologie qui en découle montre que certaines représentations sont sous-utilisées : courbes de densité en 2D ou représentations en axes parallèles. Jusqu'alors approchée par voie intuitive (décennie 1970-1980), la représentation graphique peut être analysée aujourd'hui en tenant compte d'apports venant des sciences cognitives, de la neurophysiologie, et des études d'ergonomie. Ces apports permettent de mieux utiliser la syntaxe visuelle des graphiques pour acquérir des informations nouvelles à partir d'un lot de données. Un exemple industriel illustre la discussion.

ABSTRACT

Historically, the graphic methods are an integral part of data analysis; they can significantly foster those intuitions that can tell us which model is best suitable. For the statistician, the keys of a graph are the number and type of variables taken into account and the target : overall description of the population, pointing of statistical units, or comparison to a model. There are several systems of representation : orthonormal, parallel or radial coordinates, possibly combined. Looking at the resulting typology, we can see that some representations, such as the 2D density curves or the representations in parallel axes, are sometimes underused. An intuitive process in the 1970's-1980's, the graphical analysis benefits nowadays from the knowledge of the cognitive sciences, the neurophysiology and the ergonomics studies which contribute to a better application of the visual syntax of graphs in order to extract as much information as possible from a set of data. The discussion is illustrated by means of an industrial example.

1. Total Fina Elf, 64018 Pau Cedex E-mail : jean-paul.valois@totalfinaelf.com

1. INTRODUCTION

Le nom de J.W. TUKEY est associé à l'idée de méthodes exploratoires. Son ouvrage de 1977 visait à réhabiliter les « simple descriptions », effectuées sans hypothèses préalables, et dans lesquelles on ne cherche pas à inférer d'hypothèse distributionnelle.

L'idée d'approche exploratoire, longtemps associée aux travaux de TUKEY, est aujourd'hui reconnue et tend à englober l'ensemble des outils d'Analyse des Données (LEBART *et al.*, 1995).

TUKEY *(ib.)* ne soulevait pas seulement la question statistique d'une classe d'outils fonctionnant sur des populations comportant des écarts importants à la multinormalité ; il recherchait une sorte d'excellence des méthodes graphiques : « we demand impact from our pictures », *(ib.)* ; ses travaux ultérieurs (1990, 1993) confirment qu'il s'agissait là d'un point central de sa réflexion, alors partagée par d'autres auteurs (BERTIN, 1977, TUFTE, 1983).

Ces interrogations ou pratiques nouvelles marquent un tournant dans l'histoire des graphiques. Celle-ci est souvent présentée isolément. Le présent article ne détaille pas l'historique complet des méthodes graphiques ; le lecteur pourra trouver par exemple dans BENIGER et ROBIN (1978), une recension historique des méthodes graphiques et une bibliographie spécifique. Nous nous attachons en paragraphe 2 à souligner certains points particuliers, à savoir les liens des techniques graphiques avec l'histoire technique et culturelle générale, et à suggérer le rôle des approches graphiques dans le développement même de certaines méthodes statistiques. Ces deux considérations nous paraissent propres à interpeller le statisticien sur l'utilisation qu'il fait – ou non – des graphiques, davantage qu'un examen historique traditionnel.

Les graphiques n'ont pas seulement une importance historique, ils ont un rôle à jouer dans la démarche du statisticien pour comprendre les données et leurs relations (paragraphe 3).

Les logiciels grand public induisent cependant à choisir les graphiques surtout en fonction d'effets esthétiques, de leur « look », au détriment d'une réflexion sur la nature des données et sur l'objectif de la représentation. A l'inverse, cette facilité de produire des effets visuels variés, alors que le rapport aux données reste inchangé, rend aujourd'hui manifeste la distinction (BENETT et FLACH, 1992, CLEVELAND, 1993b, VALOIS 1993) entre la sémantique, question que le graphique cherche à résoudre, et la syntaxe, moyens visuels mis en œuvre.

La sémantique est abordée dans le paragraphe 4. Un inventaire détaillé des différents graphiques disponibles a été proposé par différents auteurs (par ex. CHAMBERS *et al.*, 1985, SPENCE et LEWANDOWSKI, 1990, WAINER et THISSEN, 1981) et n'est pas repris dans cet article. La contribution apportée ici propose un cadre pour classer les différents graphiques. La démarche statistique permet un éclairage important de ce thème et conduit à valoriser le contenu du graphique et les rapports qu'il entretient avec les données. La

typologie proposée suggère que de nouveaux graphiques peuvent prendre place dans la panoplie des outils disponibles.

Quelle que soit l'étape de son travail (exploratoire, confirmatoire, ou communication de résultats), l'utilisateur des méthodes graphiques ne peut se désintéresser des questions posées dans les années 80 sur l'impact d'un graphique. Abordée alors par des choix empiriques ou intuitifs, cette question est éclairée par les recherches actuelles. Nous nous attachons ici à quelques apports récents des neurosciences et des sciences cognitives (paragraphe 5).

Le paragraphe 6 présente pour illustrer le propos quelques figures (2 à 7) correspondant à un exemple industriel.

Ces différentes approches convergent (paragraphe 7) pour souligner l'intérêt des graphiques et en particulier de certaines représentations moins fréquemment utilisées.

2. QUELQUES REPÈRES HISTORIQUES

2.1 La mise au point des graphiques standard

L'histoire des graphiques est très ancienne. Certaines formes graphiques étaient utilisées dès le Moyen-Age, voire dès la haute antiquité : un système de coordonnées rectangulaires était par exemple pratiqué par les égyptiens (BENIGER et ROBIN, 1978, COLLINS, 1993). Pour le développement de la technique graphique occidentale, un moment notable se situe à la Renaissance italienne, lorsque la formulation des lois de la perspective (ALBERTI, 1435) permet de représenter rigoureusement la réalité sur un plan. Ces lois sont l'occasion d'une première conceptualisation du lien entre représentation graphique et mathématiques : « Au peintre est nécessaire la mathématique de son art...L'œil est le prince des mathématiques »(Léonard de VINCI).

Mais après que KEPLER (1611) eut fait admettre l'analogie entre le globe oculaire et les systèmes optiques, une méfiance latente contre ces dispositifs s'étend à la vision; elle est liée à une conception antique du rayon visuel (HAVELANGE, 1998). Le principe de précaution heuristique formulé par DESCARTES (1637) (« A cause que nos sens nous trompent quelque fois »), traduit un doute sur les perceptions qui envahit rapidement l'époque classique (MALEBRANCHE, 1674 : « Nos yeux nous trompent généralement »). Les développements du calcul des probabilités se font alors sans recourir aux méthodes graphiques (FERMAT, PASCAL, HUYGHENS, BERNOULLI, MOIVRE, BAYES) et la mathématique est appelée pour remédier aux erreurs de perception (méthode des moindres carrés, GAUSS, LAPLACE, LEGENDRE). Il faut attendre le XVIII$^{\text{ème}}$ siècle (à la suite de LOCKE, 1693) pour que soit réhabilitée la conception aristotélicienne de confiance raisonnée dans les sens, ce qui remet à l'honneur la description naturaliste (Encyclopédie, 1751-1772).

Ces considérations culturelles expliquent nous semble-t-il pourquoi le XVIII$^{\text{ème}}$ siècle marque un temps de stagnation ou de recul dans l'utilisation des

méthodes graphiques; ce fait est noté mais non expliqué par différents auteurs (BENIGER et ROBIN, *ib.*, TILLING, 1975). La revalorisation des sens, et donc des ouvrages descriptifs, que HAVELANGE remarque à la fin du XVIII^{ème} fournit le contexte culturel dans lequel paraissent les œuvres de PRIESTLEY (1765) et surtout de PLAYFAIR (1786). WAINER et THISSEN (1991) soulignent le sens esthétique de PLAYFAIR, mais insistent sur le fait qu'il est le premier à avoir donné une signification symbolique à la représentation spatiale : une dimension peut être utilisée pour signifier une grandeur ni spatiale ni temporelle. Il devient dès lors nécessaire d'expliciter les axes, que DESCARTES lui-même ne figure pas (Figure 1).

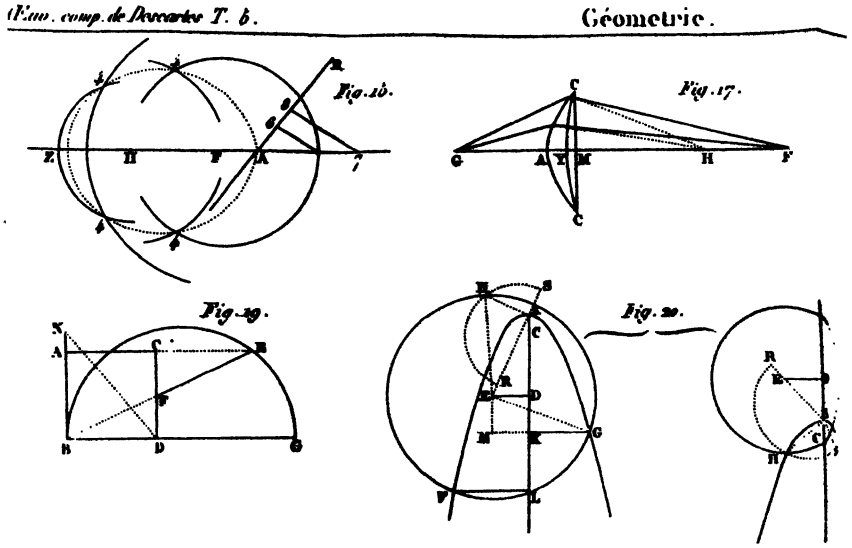

FIG 1. — Extrait d'une planche de graphiques de la « Géométrie » de Descartes (1637).

Même si d'heureux précurseurs peuvent être rencontrés parfois loin dans le passé, la fin du XVIII^{ème} et le début du XIX^{ème} siècle apparaissent comme un moment fécond où sont établies en moins d'un siècle nombre de formes graphiques aujourd'hui classiques : 1752, cartographie en contours, 1760-65, ajustement, interpolation, erreur de mesure, 1779, variations périodiques, 1786, diagramme en barres, 1794, papier muni de coordonnées, 1833, histogramme, 1843, courbes d'isovaleurs, 1880, stéréogrammes par PEROZZO (voir détails dans BENIGER et ROBIN, *ib.*, ou pour la cartographie dans PALSKY, 1996).

BENIGER et ROBIN voient dans le XIX^{ème} siècle la prise en compte graphique de distributions continues, avec les statistiques démographiques. L'utilisation abstraite des repères spatiaux s'étend avec l'échelle logarithmique (LALANNE, 1846) : le repère doit s'adapter aux grandeurs à représenter.

2.2 La fin du XIX^{ème} et le début du XX^{ème} siècle

La fin du XIX^{ème} et la première moitié du XX^{ème} siècle voient se développer une utilisation standardisée des graphiques (WAINER et THISSEN, *ib.*). MAYR (1874) et MAREY (1879) présentent un panorama raisonné des graphiques, MAREY insiste sur la possibilité d'« embrasser d'un coup d'œil une quantité énorme de documents » (*ib.*). Chez les statisticiens, la description réhabilitée de l'environnement trouve son écho au XIX^{ème} siècle. En parcourant cette « ère de la biométrie », BENZECRI (1982) note que la notion de corrélation au sens moderne de cette expression – liaison entre deux variables aléatoires – est due à PEARSON (1896) ; avant lui GALTON, puis BRAVAIS utilisaient chacun des deux axes pour représenter la même grandeur, les écarts seuls étant corrélés. C'est PEARSON qui introduit l'idée que les valeurs vraies de deux variables – et non seulement les écarts de mesure d'une même variable – soient corrélées. Ceci le conduit à proposer un graphique dont les deux axes représentent chacun une variable différente, la formalisation des relations lui permettant comme on le sait de poser les fondements méthodologiques à la fois de la régression et des méthodes factorielles. Mais cette idée que des variables puissent entretenir des relations était implicite dans l'œuvre graphique de PLAYFAIR, qui disposait des variables différentes sur chaque axe un siècle plus tôt (WAINER et THISSEN, *ib.*).

On peut voir dans le XX^{ème} siècle un développement de l'approche multidimensionnelle (BENIGER et ROBIN, *ib.*). FISCHER puis JORESKOG élargissent ces possibilités en décrivant les données dans un espace euclidien multidimensionnel. Cet effort vient dans le prolongement des intuitions de PLAYFAIR. Il culmine dans le « multidimensionnal scaling » (KRUSKAL et WISH, 1978) ou dans l'oeuvre de BENZECRI (1979), qui placent la projection spatiale euclidienne au coeur des aides interprétatives, attitude que l'on retrouve dans les méthodes PLS (WOLD, 1985).

2.3 Le renouvellement de l'approche graphique dans la décennie 80

Dans la décennie 1970-80 émergent deux tendances différentes qui ont en commun l'idée de tirer un meilleur parti des graphiques. D'une part, de nouvelles formes graphiques sont proposées (faces de CHERNOFF (1973), courbes d'ANDREWS (1972), graphique des quantiles de GNANADESIKAN, 1977,...). D'autre part, une réflexion méthodologique se fait jour, plusieurs auteurs cherchant à répondre à la question suivante : quelles règles suivre pour qu'un graphique soit efficace, c'est-à-dire pour qu'il restitue une information technique qui nous frappe ? BERTIN (1977) pose que le graphisme a des lois propres et commence à les formuler. Cet effort est développé alors que la connaissance des processus neurophysiologiques de vision est balbutiante et guère connue des non-spécialistes (prix Nobel attribué en 1981 à HUBERT et WIESEL pour les travaux sur l'aire corticale V1, détails dans IMBERT, 1983).

Après que l'intérêt des méthodes graphiques eut décru chez les statisticiens au milieu du XX^{ème} siècle (BENIGER et ROBIN, ib), l'apparition en une décennie d'ouvrages d'origine variée donnant une impulsion nouvelle aux

graphiques et à leur utilisation ne manque pas de poser question. L'impulsion est antérieure à l'utilisation de l'ordinateur, aussi bien pour TUKEY (1977) que pour BERTIN (communication orale). Suggérons un possible mouvement culturel de revalorisation des perceptions visuelles sous l'influence des médias qui se répandent à partir des années 60 (diffusion des illustrations de presse et télévision, voir par exemple GAUTHIER, 1996).

Le terme d'analyse graphique des données est proposé par WAINER et THISSEN (1981) pour marquer l'apparition conjointe d'un ensemble d'outils et de règles d'utilisation dont le corpus paraît désormais constituer une technique à part entière.

Cette dénomination rend justice à la place nouvelle des graphiques pour analyser les données. Mais sa formulation suggère que les données peuvent être analysées soit par voie numérique soit par voie graphique et prolonge implicitement le hiatus apparu après DESCARTES entre méthodes numériques et méthodes graphiques. Ancrée comme on l'a vu sur un mouvement culturel ancien, la méfiance à l'égard des graphiques reste persistante chez les statisticiens. Elle a ralenti la diffusion des méthodes graphiques nouvelles ainsi que l'audience des pistes de réflexion sur la conception des graphiques et leur rôle. En particulier, le repère dit cartésien reste aujourd'hui dominant.

Quoique très rapide, la mise en situation de l'utilisation des graphiques dans le contexte culturel général offre l'intérêt de proposer une explication au fait que des formes graphiques parfois très anciennes n'ont été utilisées massivement que bien après leur création. COLLINS (1993) suggère que l'histoire des graphiques a pour trait dominant l'extension à des domaines de plus en plus abstraits de formes visuelles créées précédemment pour décrire le monde physique. Mais en soulignant que les techniques graphiques, et surtout leur utilisation, sont liées à un contexte culturel, nous sommes conscients de fournir une vue à la fois enrichissante et frustrante : dans ce cadre en effet, on ne peut prétendre que sensibiliser le lecteur à un point de vue et admettre que d'autres points de vue complémentaires sont évidemment possibles et, pour ce qui concerne le présent article, s'en tenir à des points de repères ponctuels pour ne pas donner à cette partie une place disproportionnée. Notre but est de suggérer aux lecteurs une réflexion remettant en cause la réticence encore répandue à l'égard des graphiques.

3. LE RÔLE DES GRAPHIQUES

3.1 Validation des modèles

L'ouvrage de CLEVELAND (1993a) apporte une interpellation très forte pour le statisticien et peut être vu comme un livre charnière (GUNTER, 1994). CLEVELAND reprend divers lots de données classiquement traités dans la littérature statistique, y compris par R. FISHER. Des représentations graphiques appropriées l'amènent à penser que le modèle utilisé antérieurement était inadapté, ou que telle particularité ou erreur dans les données n'avait pas été prise en compte. Ces indications le conduisent à une modélisation

différente qui produit des conclusions nouvelles. CLEVELAND conclut à la nécessité de vérifier graphiquement si le modèle choisi est approprié : « se fier uniquement à des méthodes numériques est une stratégie dangereuse » (ib), on peut aboutir à des conclusions erronées, même lorsque les tests inférentiels paraissent justifier le modèle adopté. Il faut donc tout autant se méfier des modèles que des perceptions, un modèle choisi... à l'aveuglette peut conduire à des résultats inappropriés, même s'ils sont formellement corrects.

Le statisticien en a l'expérience, quand au cœur d'une démarche de modélisation linéaire, il produit le graphique des résidus, méthode la plus rapide pour vérifier le caractère aléatoire de leur répartition, donc la validité du modèle. La démarche formalisée par CLEVELAND (*ib.*) a donc des précédents dans les habitudes des statisticiens.

3.2 Recherche de nouvelles informations

L'apport des graphiques ne se limite pas à vérifier la validité d'un modèle. La préoccupation commune des graphiciens de la décennie 70-80 est de tirer plus d'information des graphiques qu'on ne le fait habituellement. L'objectif est de dégager un élément nouveau des données (CLEVELAND, 1993a, WOODS, 1991) et de le faire apparaître clairement (« what hits you », TUKEY, 1990).

Différents logiciels proposent aujourd'hui des facilités interactives pour faire coexister selon les besoins des types de graphiques variés et pour animer ceux-ci (en permettant de repérer immédiatement une sous-population sur l'ensemble des graphiques représentés) ; ces possibilités nouvelles étendent l'efficacité des graphiques dans l'exploration des données (CHAU, 1995, THEUS, 1995, VALOIS, 1998b).

3.3 Limites de l'approche graphique

Malgré ces apports, il ne s'agit ni de porter les graphiques au pinacle ni de dénigrer le rôle fécond des modèles, mais de laisser à chacun sa place dans la démarche pratique. Les graphiques ont un rôle à jouer dans l'acquisition d'indispensables intuitions (LEGUEN, 1995). TUKEY (1990) a souligné le paradoxe suivant : les graphiques utilisent des données quantitatives, mais nous en retirons surtout des impressions semi-quantitatives. Quand on accepte cette limite, les graphiques bien utilisés jouent le rôle de « partenaires » (TUKEY, ib). Dans une étude pratique peuvent prendre place en alternance des phases d'analyse graphique et de modélisation (VALOIS, 1998a).

Utiliser les graphiques comme partenaires implique de produire, après essais parfois nombreux (TUKEY, ib), la représentation adaptée à la question posée donc, en définitive, demande un codage approprié des données initiales. Cette démarche nécessite de clarifier les liens entre les données, les questions posées et les différents types de graphiques.

4. UNE TYPOLOGIE DES GRAPHIQUES

L'idée de proposer une typologie des graphique n'est pas nouvelle : MAYR (1874), MAREY (1879), plus récemment BERTIN (1977), CHAMBERS *et al.* (1983), pour se limiter à quelques exemples. Les inventaires des graphiques disponibles sont souvent organisés en fonction de la dimension du graphique. Or le passage aujourd'hui aisé d'une représentation 2D à une représentation en fausse perspective (effets pseudo-3D) rend inadéquat ce point de vue. Il apparaît nécessaire également de prendre en compte la présence fréquente de variables catégorielles, qui déterminent des sous-populations (CLEVELAND, 1993b). Enfin, le type de question à résoudre est rarement évoqué.

Nous proposons donc ici de mettre l'accent sur trois questions principales : le nombre et le type des variables, la question à résoudre, et le système de coordonnées employé pour la représentation. Sur ces questions, la démarche statistique et la réflexion mathématique de base ont un éclairage important à fournir.

4.1 Nombre et type des variables, question à résoudre

Le nombre (C) de variables catégorielles et le nombre (Q) de variables quantitatives peuvent former un tableau à double entrée (voir Annexe 1 et figure 8).

Dans chaque case (C_i, Q_j) de ce tableau, on peut rechercher quels sont les caractères globaux de la population, ou comment les détails s'organisent par rapport à ces tendances. Ces deux attitudes distinguées par J. BERTIN (1977) recouvrent respectivement – dans le domaine de l'analyse des données – les catégories d'approche globale et de repérage proposées auparavant par RICHAUDEAU et GAUQUELIN (1966) ; elles correspondent à deux modalités des processus attentionnels (« divided /focused attention », dans CORBETTA *et al.*, 1990). Le statisticien est familier d'une telle démarche ; par exemple, en analyse factorielle, un point de vue global peut être fourni par les variables et le repérage par l'examen des individus. Mais en fait, aussi bien pour les variables que les individus, on peut préférer une représentation de synthèse qui privilégie les faits dominants (figures 2 et 5A) au détriment des détails, ou au contraire une représentation qui donne accès au détail de l'information au détriment d'une bonne vue d'ensemble (figure 3, ou table plus ou moins aménagée).

Dans le domaine statistique, il convient d'ajouter la comparaison à un modèle, pratiquée depuis longtemps (droite de Henry). Un exemple plus élaboré d'une telle démarche peut être trouvé dans WORSLEY (1987).

Il est souvent nécessaire d'adopter des représentations différentes selon le point de vue retenu (VALOIS, 1986). Pour chaque couple (C_i, Q_j) caractérisant la nature des données, il y a donc trois situations possibles concernant l'objectif du graphique : description globale (« G »), repérage des unités statistiques (« U »), ou comparaison à un modèle (« M »).

4.2 Système de coordonnées

4.2.1 Présentation

Le système dit cartésien place les points de données (x_k, y_k) dans l'espace, à l'intersection de parallèles aux axes passant par $x_k y_k$. Les quantités x et y sont quantitatives, et les positions le long des axes x et y ne sont pas permutables.

Moins traditionnel, un autre mode de représentation repère les positions $x_k y_k$ sur les axes eux-mêmes et joint ces positions par un segment de droite. En représentation plane multivariée, une unité statistique est alors représentée par un segment polygonal, fermé si les demi-axes sont en disposition concentrique (coordonnées radiales), ouvert si les axes de coordonnées sont parallèles (INSELBERG, 1985),

4.2.2 Coordonnées radiales

Aux coordonnées radiales se rattachent, outre le classique graphique en « camembert », les glyphes (ANDERSON, 1960), et les polygones (SIEGEL et al 1971). Système centré, les coordonnées radiales peuvent être adjointes localement à la représentation d'une unité statistique dans un système orthonormé (CHERNOFF, 1973), par exemple pour illustrer les descripteurs d'une ville sur une carte de géographie, ou en cartographie géologique (VALOIS, 1998b). L'impression visuelle fournie par le polygone dépend simultanément de toutes les variables. Les ergonomes soulignent l'existence d'une propriété émergente (par exemple la symétrie) pour ce type de configuration qui est dite « configurale » (POMERANZ *et al.* 1977, CARSWELL et WICKENS, 1990).

4.2.3 Coordonnées parallèles

Dans les coordonnées parallèles (INSELBERG, 1985), on répète le long d'une dimension D (habituellement horizontale) une succession d'axes correspondant chacun à une variable. Chaque dimension est représentée sans interférence avec les autres. Il s'agit donc d'une représentation « séparable » au sens de CARSWELL et WICKENS, 1987. L'orthogonalité entre la dimension des axes et celle de leur répétition crée une ressemblance avec le repère cartésien, de sorte que la distinction entre ces deux systèmes est rarement faite, par exemple pour les boîtes de dispersion de TUKEY (1977). En système de coordonnées parallèles, les positions selon la dimension D sont totalement permutables. On peut changer l'ordre d'apparition des boîtes de TUKEY selon l'axe horizontal ; cette permutation ne change pas la signification du graphique (sa sémantique), mais ce changement de syntaxe visuelle (ce que l'œil perçoit) peut avoir des effets importants sur la clarté de lecture, comme montré par BERTIN (1977) qui a remarqué l'intérêt des permutations.

Toute utilisation extensive du repère cartésien, produite en utilisant l'un des axes pour représenter une variable catégorielle ou une suite arbitraire, rejoint la disposition étudiée par INSELBERG (*ib.*). Sous un vocable différent, CHERNOFF (1973) mentionne ce mode de représentation ; il propose que les valeurs graphiques de représentation des variables soient normées (0,1) sur les extrema des données. Si chacun des deux axes reçoit une variable catégorielle

ou nominale, le graphique représente une matrice permutable dans ses deux dimensions (BERTIN, 1977).

4.2.4 Coordonnées quelconques

Une autre utilisation non euclidienne de l'espace de représentation tient compte des variables pour déformer les éléments d'un motif prédéfini, comme le propose CHERNOFF (1973) pour des visages stylisés. Nous désignons cette pratique comme système de coordonnées quelconques; nous ne la reprenons pas ici car d'usage plus délicat, tant en programmation qu'en utilisation : l'ensemble fourni peut s'avérer très lisible, mais il faut veiller à mettre les variations visuelles en accord avec le sens implicite des données pour l'utilisateur (sourire pour une progression jugée bénéfique et non l'inverse), donc mettre le système de signes en accord avec le système de sens, selon un principe de compatibilité (KOSSLYN, 1994).

4.2.5 Graphiques multiples

Il est bien sûr possible de répéter n fois un graphique pour diverses variables ou sous-populations. L'intérêt pratique d'une telle répétition en « treillis » est clair (BECKER *et al.* 1994, THEUS, 1995), mais le même motif étant répété plusieurs fois, cette opération ne crée pas un nouveau cas dans la typologie et n'est donc pas explicitée ici.

4.3 Apport de la présentation typologique : exemples

Une présentation visuelle condensée de cette typologie (Annexe 1) conduit à remarquer immédiatement que certaines représentations sont très courantes, alors que d'autres sont négligées ou sous-employées, telle la densité de points en 2 dimensions, ou l'exploitation des possibilités offertes par les coordonnées parallèles.

La représentation d'une seule variable est généralement abordée par une représentation globale en densité (histogramme ou courbe de fréquence cumulée). Quand on dispose de deux variables quantitatives, c'est un report cartésien des unités statistiques (U) que l'on effectue communément. Il est cependant possible de calculer une densité de points en 2D et de la visualiser par un tracé d'isovaleurs. Pour un objectif d'approche globale des données représentées, ce type de graphique apparaît très efficace (Figure 5A, voir aussi VALOIS, 1993, WILKINSON, 1994; pour une utilisation concrète, voir de LA ROCHE *et al.*, 1980). Différents algorithmes de calcul de la densité ou de tracé des isovaleurs sont imaginables : nous renvoyons le lecteur vers plusieurs articles abordant ce sujet (ATKINSON, 1997, BEBBINGTON, 1978, GOLDBERG et IGLEWITZ, 1992, ROUSSEEUW et RUTS, 1996, ZANI *et al.*, 1998).

Limitée à quelques valeurs clefs (quantiles), les boîtes de dispersion proposées par TUKEY (1977) reposent sur une représentation schématisée de la densité et fournissent une approche globale (G) de la population. La normalisation proposée par CHERNOFF (1973) réduit les effets de variance; elle permet de

plus d'étendre les possibilités de ce type de graphique à une comparaison de sous-populations (en normalisant en fonction de la population de référence).

Le système de coordonnées parallèles permet également une représentation des unités statistiques, ou au moins d'un certain nombre d'entre elles (voir détails en Annexe 2 et utilisation en figure 6). Par ailleurs, l'adjonction en marge du graphique d'un repère correspondant à une population de référence (par exemple normale) permet une comparaison visuelle de la population à ce modèle.

Ainsi mis en œuvre, le système de coordonnées parallèles couvre la représentation multivariée, en approche globale d'une population (boîtes de dispersion), en repérage des unités statistiques, ou en facilitant la comparaison à un modèle. Elle peut être utilisée pour une population isolée ou pour comparer des sous-populations (existence d'une variable catégorielle). Cette représentation n'est pas optimale dans tous ces cas, la lisibilité contraignant à ne représenter pratiquement qu'un petit nombre d'unités statistiques. Ces possibilités, et ces limites, illustrent les différents points abordés dans cette démarche typologique.

5. L'IMPACT D'UN GRAPHIQUE

5.1 Les stimuli visuels et leur efficacité

BERTIN (1977) postule que le graphique a des lois propres, distinctes de celles des données sous-jacentes. Les données sont représentées par des signes, décomposables en différents stimuli, dont il évalue empiriquement les propriétés et l'efficacité visuelle : certains signes peuvent traduire des quantités (position, taille, densité optique), d'autres sont utilisables pour coder des catégories (trame, couleur, orientation, forme). BERTIN estime que les différences d'orientation se remarquent mieux que celles de couleurs et que les différences d'intensité prévalent sur les teintes. CLEVELAND (1993b) propose un classement légèrement différent de l'efficacité visuelle des stimuli, mais ne contredit pas ce principe d'analyse que l'on retrouve dans WILKINSON (1999). En outre, BERTIN *(ib.)* accorde une importance majeure aux regroupements obtenus par permutation des lignes ou colonnes de signes ; il note l'importance de s'en tenir aux données porteuses d'information, selon un principe de sobriété que, de son côté, TUFTE (1983) érige en principe absolu. Bien qu'elles servent de base encore aujourd'hui pour différents travaux (WILKINSON, 1999), nous ne reprenons pas le détail des conclusions de BERTIN, disponibles en français (1977), pour développer des apports plus récents et moins connus.

Depuis la première partie du XX^{ème} siècle (EELLS, 1926), des expériences ont été entreprises sur des échantillons de personnes pour évaluer statistiquement l'efficacité de divers graphiques ou éléments de graphiques. Les travaux sont aujourd'hui nombreux (CLEVELAND *et al.*, 1982, LEWANDOWSKI et SPENCE, 1989, SPENCE et LEWANDOWSKI, 1990, LOHSE, 1991,

SIEGRIST, 1996, SIMKIN et HASTIE, 1987). La base intuitive ou empirique sur laquelle reposent les travaux antérieurs est critiquée par ces différents auteurs.

De telles expériences peuvent aider à connaître le comportement de l'utilisateur vis à vis du graphique. Si, comme proposé par PINKER (1990), on modélise le cheminement visuel de l'utilisateur (trajet visuel et temps consacré aux différents items), il semble possible de prédire le temps global mis par l'utilisateur pour capter l'information. Un programme proposé par LOHSE (1993) donne des résultats encourageants, mais suppose un graphique parfaitement standardisé; son intérêt est de valider le modèle. Mais il n'y a pas d'accord unanime sur la force « attractive » des différents stimuli visuels ni sur leur organisation dans le graphique; le positionnement de celui-ci dans la page (titre, légende...) doit aussi être pris en compte. En outre les résultats empiriques gagnent à être confrontés à une base neurophysiologique qui peut en éclairer les points forts et les limites.

KOSSLYN (1994) a montré que les résultats acquis par la neurophysiologie pouvaient donner des indications pratiques pour mieux construire les graphiques. Nous utilisons également ci-après des contributions neurophysiologiques plus récentes, ainsi que des résultats produits par les sciences cognitives (CARSWELL et WICKENS, 1990, WOODS, 1991, DANEK et KOUBEK, 1995, CORBETTA *et al.*, 1991...). Ces travaux donnent aujourd'hui des éléments techniques de base pour discuter ce sujet. Il apparaît que ces avancées ne remettent pas totalement en cause les intuitions initiales, auxquelles elles donnent un début de justification.

Ajoutons néanmoins que le domaine des sciences cognitives est très vaste. Le graphique est une frontière entre le monde des données, qu'il représente en les symbolisant, et le monde de la connaissance qui cherche à mémoriser et interpréter les données; il met en jeu des processus cognitifs variés. Le présent article vise à souligner brièvement le rôle de certains de ceux-ci et à les mettre en relation. Cette synthèse est certainement partielle; le lecteur pourra l'approfondir en fonction de son propre parcours ou du contexte de tâche dans lequel il utilise les graphiques. Le point fort que nous voulons souligner ici est de montrer des liens possibles entre le point de vue pragmatique et empirique qui a prévalu dans les années 80 et certains apports plus récents des sciences cognitives. Une telle intuition se retrouve dans les travaux de Mac EACHREN (1995).

5.2 La perception visuelle et son rôle

S'appuyant à la fois sur des expériences de comportement et sur des observations de neurophysiologie, DEHEANE *et al.* (1999) proposent un modèle pour la compréhension des nombres. Les ordres de grandeur et les comparaisons sont traitées dans le cerveau par des aires corticales voisines de celles traitant la perception visuelle, alors que le traitement arithmétique exact est au contraire en lien avec les aires gérant le langage. Ce modèle est cohérent avec la remarque de TUKEY (1990) sur les indications semi-quantitatives que nous tirons des graphiques, et sur la constatation qu'une valeur exacte

est mieux perçue grâce à une table (CHAU, 1995), quitte à aménager celle-ci pour la rendre plus lisible (EHRENBERG, 1977).

Une convergence peut être notée entre les stimuli visuels explicités par BERTIN, 1977 et les attributs perçus en vision préattentive (JULESZ, 1991, TREISMAN, 1997, MALIK et PERONA, 1990). Les figures 4 et 5B sont conçues dans l'esprit des travaux de JULESZ (*ib.*), c'est-à-dire en utilisant la capacité de la vision préattentive de reconnaître immédiatement différents signes s'ils diffèrent suffisamment par leur couleur, taille ou orientation. On désigne classiquement par vision préattentive ce qui est perçu immédiatement, avant l'intervention d'une lecture consciente. L'usage du mot « inconscient » trouve son origine dans de telles constatations (HELMOLTZ, 1896). DEHEANE (1998) montre que certaines opérations arithmétiques simples sont prises en compte dans les étapes préattentives, des parties du cerveau pouvant hériter de propriétés présentes semble-t-il chez certaines espèces animales.

L'intérêt de ce stade précoce de la vision n'est pas seulement sa rapidité : les informations acquises à ce stade influencent et conditionnent la poursuite du processus perceptif (KOSSLYN, 1991). A la conception aujourd'hui ancienne d'une perception faisant intervenir un enchaînement continu de processus a succédé un modèle de traitement coopératif faisant intervenir simultanément différents centres fonctionnels répartis dans le cortex cérébral ; dans ce schéma, les informations acquises dans les stades précoces du traitement de la perception influencent les processus plus tardifs. Ainsi le décodage des intensités lumineuses (en « noir et blanc ») est achevé avant celui des couleurs, ce qui n'est pas sans évoquer la hiérarchie intensité/couleur proposée par BERTIN. Divers apprentissages culturels semblent intervenir très tôt dans la perception (l'opposition classique de couleurs comme bleu/rouge résulte d'un codage culturel dont l'origine historique est documentée, PASTOUREAU, 2000).

5.3 Hiérarchiser l'information

Cette modélisation de la perception visuelle permet de rendre compte des constatations empiriques de BERTIN. Le slogan proposé par le maître français « Voir, c'est percevoir immédiatement » indique bien l'importance de cette vision préattentive dans ses constatations. MARTIN (1989) avait postulé que l'élément important d'un texte doit être vu en premier ; cette remarque peut être étendue aux graphiques (VALOIS, 1993).

Les ressources cognitives sont limitées (MILLER, 1956, KOSSLYN, 1994), et il est avantageux de répartir les informations en un nombre réduit d'unités perceptives (3 ou 4 maximum). Cette hiérarchisation de la représentation en sous-ensembles (plans visuels de CHAPPE, 1993, « perceptual units », KOSSLYN, *ib.*) utilise mieux la vision préattentive et libère les ressources cognitives pour traiter des tâches de plus haut niveau (MITCHELL et BIERS, 1982) ; elle rend le graphique plus efficace au double sens de quantité d'information perçue (TUKEY, 1990, CLEVELAND, 1993b) et de niveau de réponse obtenue (BERTIN, 1977). Les avantages du regroupement en unités perceptives, remarqués de longue date (WERTHEIMER, 1945), rendent compte de l'intérêt majeur accordé aux permutations par BERTIN (1977).

Ce regroupement en unités perceptives est traditionnellement conçu comme un agencement spatial. Mais on peut étendre cette notion à un regroupement des signes en fonction d'autres caractéristiques, par exemple leur densité optique (VALOIS, 1998b), voir figure 6.

La sobriété réclamée par TUFTE (1983) a été critiquée car elle ne fait pas toujours preuve d'une efficacité supérieure (BUCKER et CLEVELAND, 1993) ; le graphique a besoin d'éléments auxiliaires pour être compris (titre ou légende, graduations). Le rapport à la question posée doit être clair (il conditionne la perception, KOSSLYN, 1994) et la présentation doit être globalement subordonnée à la réponse apportée par le graphique à la question posée (*ib.*). Un graphique doit être aménagé pour répondre à une question précise et bien définie. Les éléments qui chargent inutilement les ressources perceptives sont donc à éviter. La sobriété peut être resituée ici comme aidant à percevoir la hiérarchisation du graphique en unités perceptives.

5.4 Les composants du graphique

L'adjonction d'éléments autres que les données (effets décoratifs pour attirer l'attention) est sévèrement critiquée par les différents auteurs, en tête desquels BERTIN, 1977 et TUFTE, 1983. Le graphique échappe alors au domaine de l'analyse technique pour répondre à d'autres critères (mise en page), et même en ce domaine, des limites sont rappelées (CHAPPE, 1993), mais pas toujours respectées (exemples dans VALOIS, 1998b).

KOSSLYN (1994) indique que les différents stimuli, tels qu'étudiés par BERTIN, sont traités par des faisceaux de neurones différents. Les capacités d'un canal étant limitées, il est avantageux d' utiliser conjointement plusieurs faisceaux de neurones (par exemple on discrimine mieux des symboles qui sont à la fois de forme et de couleur différentes). Pour la même raison, la discrimination entre des symboles est plus efficace si on utilise des seuils bien séparés (30° pour les orientations).

Des particularités de détail de la perception (CLEVELAND, 1985, KOSSLYN, 1994) sont à prendre en compte pour optimiser la représentation graphique et en ajuster la syntaxe visuelle. En particulier (TVERSKY *et al.*, 1989), l'information perçue tend à subir l'effet attracteur d'un modèle perceptif sous-jacent (bissectrice dans le repère orthonormé, même si elle n'est pas figurée). Les différences par rapport à ce modèle tendent à être perçues comme moins importantes qu'elles ne le sont en réalité. C'est la raison pour laquelle il est intéressant par exemple de reporter les résidus d'une régression selon une ligne horizontale, alors que structurellement (sémantique des données) l'information est la même que sur le graphique d'origine, seule la syntaxe visuelle ayant changé.

5.5 Lien avec la tâche

Les expériences sur l'efficacité visuelle des graphiques sont intéressantes, mais elles sont réalisées en considérant le graphique sans égard à son contexte d'utilisation. Ce contexte implique une tâche à résoudre à l'aide du graphique ; à des

contextes différents peuvent correspondre des processus attentionnels variés, ce qui a des conséquences sur la perception (POSNER, 1990, CORBETTA *et al.*, 1991, SPITZER *et al.*, 1988). Le choix de la représentation pour résoudre un problème doit donc être mis en rapport avec le type de tâche (BENETT *et al.*, 1992, GOETTL *et al.*, 1991, MITCHEL et BIERS, 1992, SIMKIN et HASTIE, 1987), ce qui rejoint la distinction (approche globale/repérage) proposée en paragraphe 4.

Les expériences menées par les ergonomes prennent en compte globalement ces aspects. La représentation intégrale fournie par les coordonnées radiales est bien adaptée pour les situations de décision nécessitant la prise en compte conjointe de nombreuses variables. On vise alors à regrouper les informations graphiques quand la tâche nécessite la prise en compte conjointe d'informations nombreuses (principe de compatibilité, CARSWELL et WICKENS, 1990). Mais les expériences montrent que les avantages de la représentation intégrale persistent également dans un contexte d'attention sélective (recherche d'une information particulière).

6. ILLUSTRATION SUR UN EXEMPLE INDUSTRIEL

L'exemple présenté est issu de l'ingénierie pétrolière (COSTE et VALOIS, 2000) ; les données sont les débits mensuels d'huile et d'eau pour 179 puits d'un gisement pétrolier en production. La question posée par l'utilisateur est de cerner rapidement les zones fournissant les plus forts débits et de tenir compte des influences géologiques et hydrodynamiques pour pronostiquer les zones gardant un potentiel de production d'hydrocarbures dans le futur proche.

Le traitement des données fait appel à une extraction d'informations à partir des courbes de débit pour refléter le niveau des productions et leur évolution. Cette opération réduit les données initiales tridimensionnelles (puits, quantité produite, temps) à un tableau bidimensionnel (puits, indicateurs). Une redondance existe entre les indicateurs extraits des courbes et une analyse factorielle est d'abord entreprise pour condenser l'information.

6.1 Résultats d'Analyse Factorielle (Figure 2)

La présentation des corrélations des variables aux axes factoriels met en œuvre une permutation dans chaque colonne : les variables sont rangées par ordre de valeurs décroissantes (+1 en haut à -1 en bas, 0 pour le trait horizontal) . En outre, un décalage horizontal est réglé sur la valeur absolue (position extrême gauche pour les corrélations absolues > 0.8, décalage de 2 caractères vers la droite pour les corrélations absolues entre 0.8 et 0.6, etc ; les variables dont la corrélation est entre 0.2 et – 0.2 ne sont par reportées, Valois, 1986).

Cette présentation est immédiatement lisible. Le facteur 1 oppose ainsi DAC (dates tardives d'arrivées d'eau) à BSWmax (% d'eau maximum le plus élevé), et caractérise les puits dont le débit d'eau est faible et/ou tardif. Le facteur 2 isole les puits dont le débit d'huile est le plus élevé (Q_{omax}), ils ont aussi un

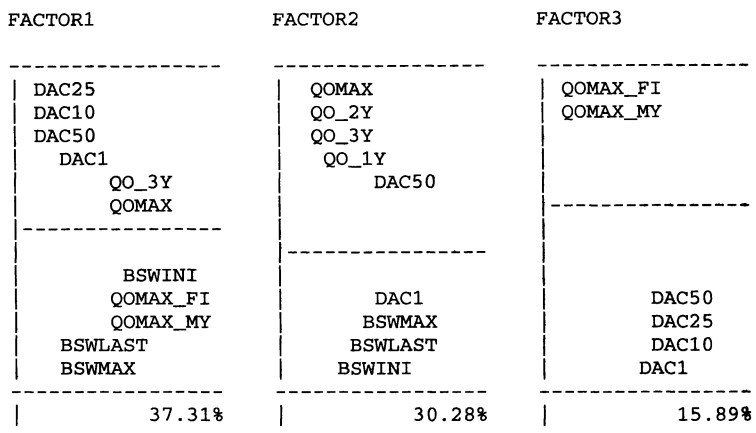

FIG 2. — Résultats d'analyse factorielle ACP, corrélation des variables aux axes factoriels

pourcentage d'eau faible (BSW initial, maximum et final). Le facteur 3 marque les puits dont la décroissance du débit d'huile est la plus rapide (Q_{omax}-f_i = rapport Q_{omax}/Q_o final).

Cette représentation ne se substitue pas aux traditionnels plans factoriels. Mais, dans le cas de facteurs nombreux, elle permet de choisir très rapidement le graphique le plus approprié en fonction des variables d'intérêt ; par exemple ici le plan (F1, F2) permet la représentation des variables BSW relatives au pourcentage d'eau dans la production.

Cette présentation semi-graphique exploite verticalement l'intérêt des permutations, en outre elle utilise la dimension horizontale pour représenter une grandeur. On porte ici l'identité des variables dans la représentation elle-même, et non sur les marges comme pratiqué habituellement pour les graphiques mono-variés. Cette représentation est adaptée à une perception globale des informations majeures, la représentation traditionnelle reste nécessaire pour obtenir les détails.

6.2 Résultats de Classification Hiérarchique (Figure 3)

Une classification hiérarchique a été effectuée sur les coordonnées factorielles des individus (les puits).

A gauche de l'arbre de classification est portée une représentation symbolisée des variables utilisées (ici les 3 facteurs de la figure 2), la partie encadrée au centre de l'arbre étant agrandie dans la sous-figure inférieure. La représentation des variables recourt à un centrage réduction, un trait médian représentant la moyenne. De part et d'autre de ce trait, sont portées 1 à 4 étoiles proportionnellement à la valeur pour chaque individu. La graduation est ici réglée pour fournir 4 étoiles pour les individus atteignant 2 écarts-types. Au delà de cette valeur, le nombre d'écarts-types est substitué à la dernière étoile (4 pour un individu situé à 4 écart-types, @ au-delà de 10 écarts-types). Cette représentation permet une bonne lisibilité des tendances d'ensemble (approche

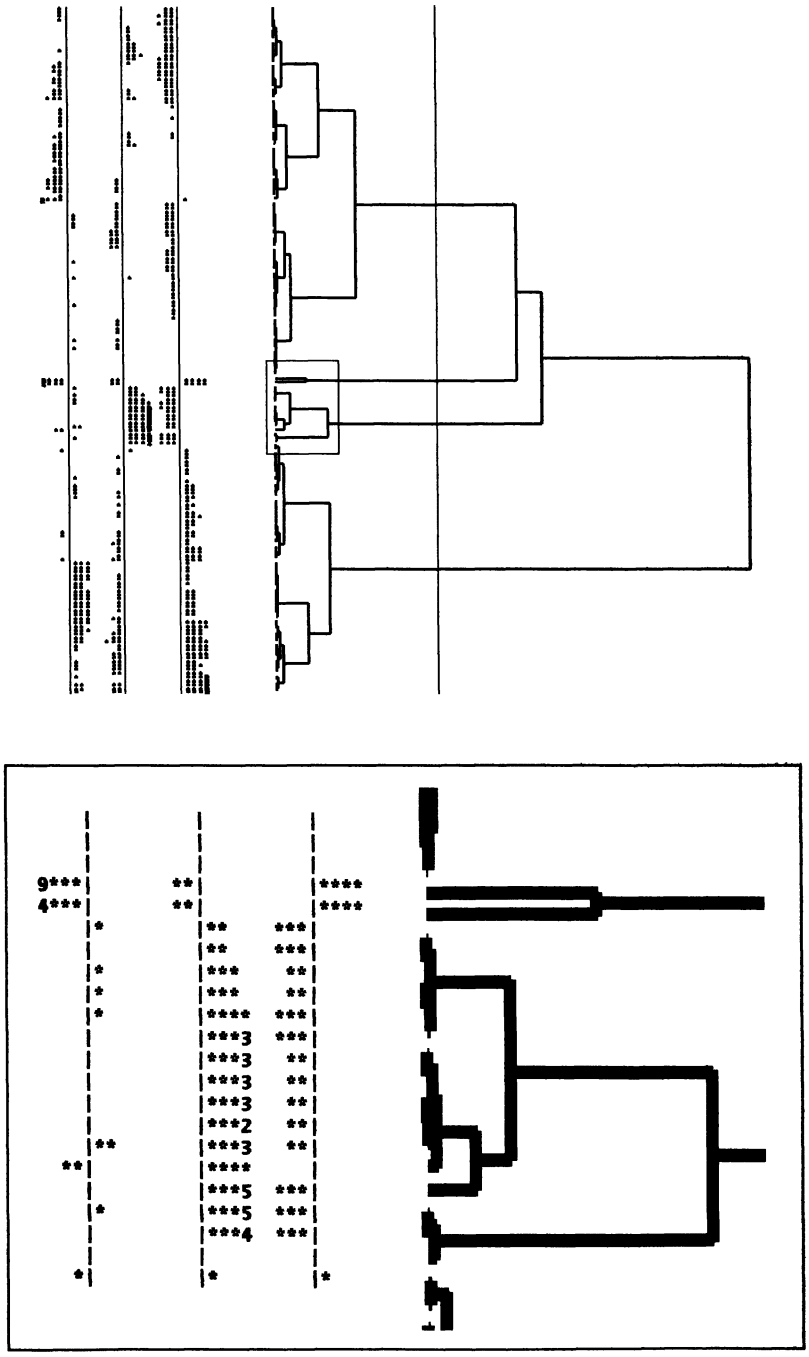

FIG 3. — Arbre de classification hiérarchique (méthode de Ward.)

globale), aussi bien pour la répartition des individus que pour le comportement des variables qui sous-tend la classification ; elle facilite aussi le repérage du niveau des valeurs pour des individus atypiques.

(Une adaptation en couleurs de cette figure peut consister à générer automatiquement une couleur par classe pour les colonnes situées à gauche, ce qui accroît la lisibilité des limites de classe.)

Cette représentation donne un accès simple et immédiat à l'interprétation de la classification. On adopte ici une partition en 4 classes (trait vertical porté sur l'arbre). La classe 1 a des valeurs négatives pour le facteur 1 (arrivées d'eau rapide, % maximal d'eau élevé). La classe 2 se distingue par le facteur 2 en négatif (faibles débits d'huile liés à une arrivée d'eau précoce). La classe 3 est isolée par le facteur 3 (décroissance rapide du débit d'huile). La classe 4 représente les puits fournissant peu d'eau et les meilleurs débits d'huile, c'est le « cœur » du gisement.

Les figures 2 et 3 montrent que l'approche graphique peut être utilisée en conjonction avec les méthodes numériques (ici ACP ou CAH), et n'est pas réservée aux stades les plus amont de la reconnaissance exploratoire. L'utilisation d'outils statistiques multivariés permet d'étendre les possibilités de représentations graphiques effectuées en mode séparable (en figure 2 et en partie gauche de la figure 3, chaque facteur est représenté séparément, il pourrait constituer une figure isolée).

6.3 Graphique des Quantiles et première cartographie (Figure 4)

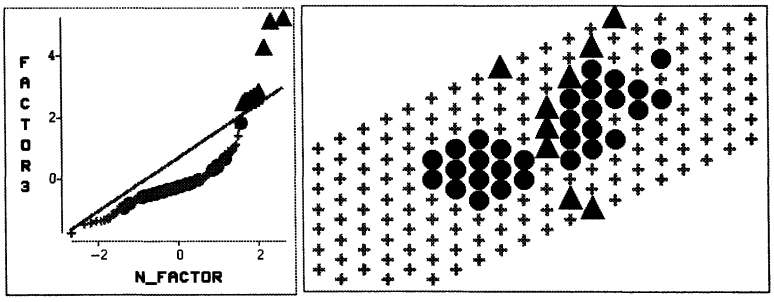

FIG 4. — Graphique des quantiles du facteur 3 et cartographie.

Le graphique des quantiles (figure 4, à gauche) compare la distribution des quantiles d'une variable (ici le facteur 3) à ceux d'une normale. Il fournit un moyen plus précis que le traditionnel histogramme pour délimiter la partie de la population la plus responsable de l'écart à la normalité.

Sur cet exemple, les puits pétroliers ainsi repérés (triangles noirs) ont une organisation géographique bien délimitée. Isoler précisément cette partie de population met en évidence une structure en arc de cercle qui recoupe le gisement. Elle est interprétée comme une faille (les puits ont un débit qui décroît rapidement, cf. figure 2). Sur la carte (partie droite de la figure 4),

les ronds noirs figurent le cœur du gisement – puits à bon débit d'huile –, les autres puits sont figurés par une croix grise.

Cette disposition graphique compare la population à un modèle (ici le modèle normal).

6.4 Densité de points (Figures 5A et 5B)

Une information plus détaillée peut être tirée de la représentation des individus sur le plan factoriel (F1, F2), traitée ici en courbes de densité de points (Figure 5A).

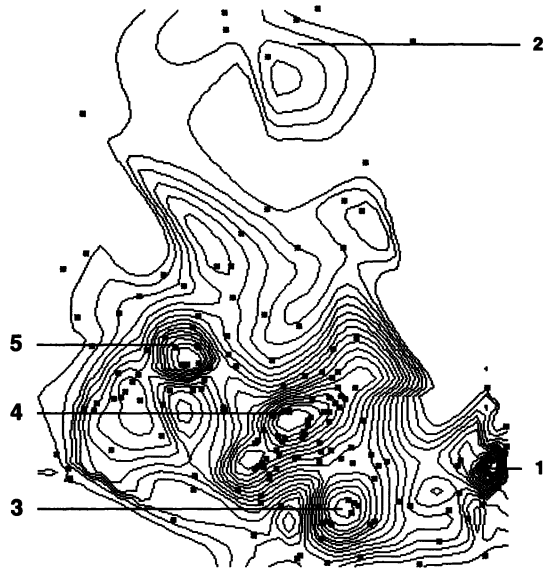

FIG 5A. — Traitement en isodensité de points du plan factoriel (F1,F2).

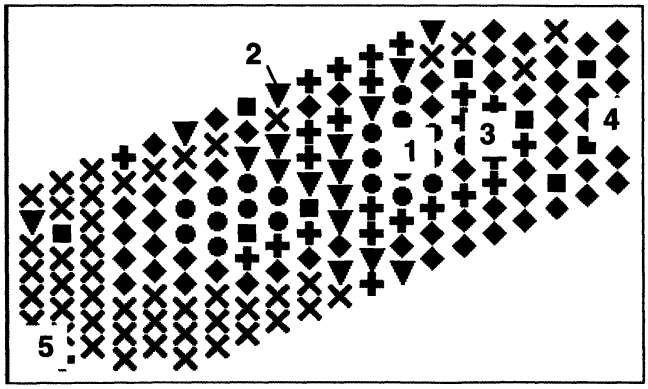

FIG 5B. — Répartition cartographique des groupes déduits de la Figure 5A.

On effectue un pavage du graphique (ici le plan factoriel); la fréquence observée dans chaque pavé est utilisée pour un tracé d'isovaleurs. Pour effectuer ce tracé, les pavés sont classés par fréquence décroissante, et l'on somme les effectifs à partir de la fréquence maximale (les ex-aequo étant classés par proximité au barycentre). Un programme d'interpolation reprend les fréquences ainsi cumulées avant le tracé des isovaleurs.

Cette procédure empirique permet de donner une bonne idée de la densité des points. Prévue pour les graphiques comportant plusieurs milliers de points, dans lesquels les superpositions limitent la lisibilité, elle peut cependant comme ici renseigner utilement sur la répartition des points dans le graphique orthonormé même quand les points sont moins nombreux (ici 179). Ce recours à une visualisation de densité, habituel pour la représentation monovariée (histogramme), n'est pas répandu pour les graphiques 2D. C'est pourtant un moyen pour effectuer une approche globale de la population. La puissance de cette représentation est liée en 2D à la sensibilité visuelle aux directions (BERTIN, 1977).

Cinq regroupements de points peuvent être ici distingués. Ces groupes sont figurés par différents symboles dans la cartographie (Figure 5B). Outre la zone centrale (1 = ronds noirs) et la zone faillée (2 = triangles, cf. figure 4), les groupes de points (3,4,5) trouvés sur le plan factoriel se montrent cohérents cartographiquement : ils correspondent à différentes zones périphériques (ici distinguées par des signes plus, croix ou losanges). Le traitement en isodensités du plan factoriel permet donc de décrire avec plus de détails la zone périphérique du gisement.

6.5 Boîtes de dispersion (Figure 6)

Selon le modèle d'INSELBERG (1985), on considère chaque variable comme portée par un axe indépendant. Une normalisation des valeurs est ici effectuée sur chacun des axes, entre le minimum et le maximum de la population de référence (ensemble des 179 puits), ce qui permet de visualiser tout de suite les caractères de la sous-population étudiée. Chaque axe doit donc, si opportun, porter des graduations propres (non reportées ici pour raisons de lisibilité).

La figure montre les caractères du cœur du gisement (zone dite '1' en figure 5B). Par rapport à l'ensemble des 179 puits, on y observe de faibles pourcentages d'eau en début d'exploitation (BSWini), une arrivée d'eau lente (date DAC10 élevée), une faiblesse du pourcentage d'eau maximum (BSWmax), et bien sûr des débits d'huile élevés (maximum instantané Qomax, et cumulé sur 10 ans NP10A).

Les différents puits repérés en figure 4 comme atypiques, sont figurés par ailleurs en portant leur enveloppe (zone noire) dans le système de coordonnées parallèles. Ils ont en commun avec la zone 1 un débit d'eau initial faible (BSWini). Mais l'eau arrive très rapidement (DAC10 plus faible), et le pourcentage d'eau maximal atteint est beaucoup plus élevé, rejoignant les maxima connus sur l'ensemble de la population. Les débits d'huile y restent modérés, par suite d'un déclin fort après le débit maximum initial (variable Q_{omax}/Q_{ofin}).

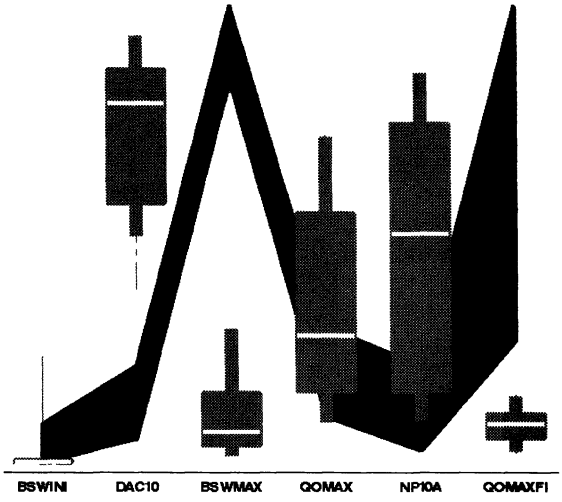

FIG 6. — Boîtes de dispersion, traitées en coordonnées parallèles.

La figure souligne donc le comportement antagoniste des puits du cœur du gisement (boîtes de dispersion grises) et de ceux de la zone faillée (enveloppe noire), excepté pour la variable BSWini.

La représentation en coordonnées parallèles, en autorisant une permutation des variables (on les a rangées ici en fonction des résultats d'ACP) et un examen de sous-populations, après normalisation par rapport à une population de référence, se montre un puissant outil descriptif (VALOIS, 2000).

Une superposition de deux types de représentations (boîtes de dispersion et enveloppe) est possible en respectant les règles proposées par KOSSLYN (1994) : séparation suffisante en angles et en densité optique pour permettre de bien différencier les deux unités perceptuelles correspondant à chaque type d'information.

6.6 Cartographie en représentation intégrale (Figure 7)

La représentation intégrale vise à fournir simultanément le maximum d'informations; les effets de symétrie (ici déformation des polygones) rendent immédiatement perceptibles des corrélations entre variables ou facilitent des regroupements visuels d'individus (ici les puits pétroliers dans une cartographie). Ce type de représentation a un intérêt majeur quand la prise de décision nécessite l'examen conjoint de variables nombreuses. Les tests expérimentaux tendent à montrer que ce type de représentation maintient néanmoins ses avantages si l'on recherche une information isolée.

Les variables représentées sont ici : débit d'huile maximum (en position trigonométrique 0°), le débit d'huile au bout de 3 ans (à 72°), le déclin du débit d'huile (Q_{omax}/Q_{ofinal}, à 144°), le pourcentage d'eau initial à l'ouverture du puits (à 216°), et le pourcentage d'eau maximal atteint (à 288°). Le cœur du gisement se distingue par des polygones allongés vers l'Est et le Nord-Est

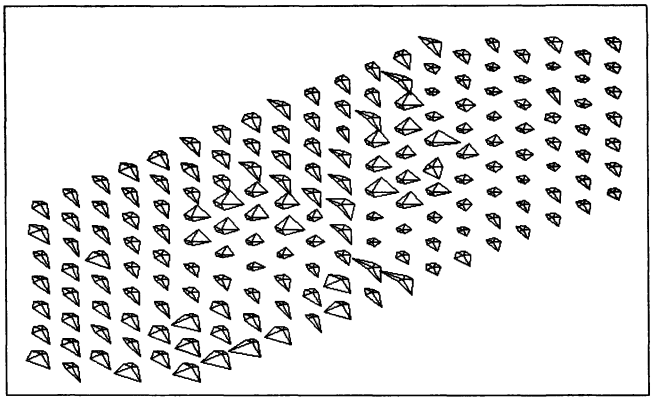

FIG 7. — Cartographie en représentation intégrale.

au centre de la zone (bons débits d'huile) ; la zone faillée se signale par des polygones très irréguliers, allongés vers le Nord-Ouest : débits modestes, mais s'effondrant de façon importante. Les arrivées d'eau commencent par le bord méridional de la carte, mais s'étendent à toute la zone occidentale, et à la zone faillée. Le compartiment oriental du gisement connaît de faibles débits, d'huile et d'eau, en relation probable avec des caractères géologiques différents (porosité du réservoir plus faible). La figure permet donc de récapituler l'ensemble des observations effectuées en figures 2 à 6.

Ce type de figure se prête donc à une représentation simultanée des variables, il facilite la perception des classes d'individus, tout en autorisant une lecture nuancée des variations des variables. Il pourrait être utilisé pour d'autres propos que l'illustration cartographique ici présentée, par exemple pour la représentation des facteurs 'secondaires' sur le plan factoriel principal (F1,F2).

7. CONCLUSION – DISCUSSION

Un lien ancien entre méthodes numériques et représentation graphique s'est distendu, pour des raisons liées à une histoire culturelle longue, dont nous avons essayé d'éclairer quelques jalons historiques. Les possibilités importantes offertes aujourd'hui par l'élaboration informatisée des représentations, la réflexion et les développements effectués depuis les années 80 dans ces domaines, enfin la confirmation de la place importante qu'occupent les représentations visuelles dans l'appréhension des nombres, doivent contribuer à redonner aujourd'hui aux représentations graphiques une place majeure dans la pratique de l'analyse des données.

Le rôle des graphiques est de faire passer une information statistique. Ils mettent le statisticien en dialogue direct avec les données. Les capacités de

synthèse spontanée de la perception permettent de décomposer le graphique en différents ensembles ou de l'aborder selon plusieurs niveaux de détail ; elles font du graphique un allié pour le statisticien dont la tâche principale est de dégager les faits principaux d'un ensemble de données.

Les représentations graphiques ne peuvent donc être limitées à un rôle illustratif des résultats obtenus par ailleurs par les méthodes numériques ; elles peuvent aussi prendre une place importante tout au long de la démarche du praticien, pour suggérer pistes de recherche et intuitions, permettre de pressentir quelle est la méthode de modélisation ou de traitement qui sera la plus appropriée et quelles seront les précautions d'emploi à prendre. Les graphiques peuvent jouer ce rôle de concert avec les méthodes numériques tout au long du processus qui va de la réception (validation) des données à la présentation des résultats.

Une discussion typologique paraît indispensable pour proposer de nouvelles représentations ayant un sens mathématique et non un simple privilège de mode ou de look. Le rôle de la démarche statistique nous paraît essentiel pour clarifier la typologie des graphiques. La typologie présentée ne prétend pas être exhaustive ni englober tous les graphiques possibles ou existants ; d'autres typologies plus détaillées sont possibles (par ex. WILKINSON, 1999). Le rôle d'une typologie est simplement de fournir un cadre de réflexion permettant d'englober une majorité de cas courants et de susciter de nouveaux cas.

La typologie présentée fait apparaître que certaines représentations sont sous-utilisées : densité des points en 2D par exemple. Une meilleure compréhension des coordonnées parallèles peut en élargir l'usage et permettre une description féconde et simple d'emploi de données comportant des sous-populations.

Les apports des études ergonomiques tendent elles aussi à confirmer l'intérêt pratique de visualisations globales de l'information, comme les coordonnées radiales ou parallèles, qui offrent toutes deux la représentation conjointe de différentes dimensions.

Les conclusions tirées de l'aperçu historique, de la grille typologique proposée et des études ergonomiques convergent donc pour proposer d'analyser les données avec une panoplie élargie de graphiques, sans se limiter au système d'axes que la tradition a curieusement attribué à DESCARTES.

RÉFÉRENCES

ANDERSON E. (1960), A semigraphical method for the analysis of complex problems, *Technometrics*, 2, 3, 387-391.

ANDREWS D.F. (1972), Plots of high dimensional data, *Biometrics*, 28, 125-36.

ATKINSON A.C. (1997), Fast very robust methods for the detection of multiple outliers. *J. Am. Statist. Ass.*, 89, 1329-1339.

BEBBINGTON A.C. (1978), A Method of Bivariate Trimming for Robust Estimation of the Correlation Coefficient. *Appl. Statist.*, 27, n.3, 221-226.

BECKER R.A., CLEVELAND W.S. (1993), Discussion of "Graphic comparisons of Several Linked Aspects" by John Tukey, *J. Am. Statist. Ass.*, 2, 1, 41-48.

BECKER R.A., CLEVELAND W.S., SHYU M.J. (1994), Treillis display, a framework for visualizing 2D and 3D data, ATT Bell Laboratories, *Research Report*, N°8.

BENETT K.B., FLACH J.M. (1992), Graphical displays : implications for divided attention, focused attention and problem solving, Human Factors, 34 (5), 513-533.

BENIGER J.R., ROBYN D.L. (1978). Quantitative Graphics in Statistics : A brief history, *The American Statistician*, 32, 1,1-11.

BENZECRI J .P. (1979), L'analyse des données, t.I, La taxinomie, t. II, L'analyse des correspondances, 3^{e} éd., Dunod.

BENZECRI J.P. (1982), Histoire et préhistoire de l'analyse des données, Dunod, 159 pp.

BERTIN J. (1977), La graphique et le traitement graphique de l'information, Flammarion, 271 pp.

CARSWELL C.M., WICKENS C.D. (1987), Information integration and the object display, Ergonomics, 30, 511-527.

CARSWELL C.M., WICKENS C.D. (1990), The perceptual interaction of graphical attributes, configurality stimulus homogeneity and object integration, Perception and Psychophysics, 47, 157-168.

CHAU P.Y.K. (1995), An empirical study evaluating the usefulness of dynamic graphical display in decision support, *Journal of Information Science*, 21, 3, 201-208.

CHAMBERS J., CLEVELAND, W.S., KLEINER B., TUKEY P. (1983), Graphical methods for data analysis, The Wadesworth statistics, Probability Series, Duxburg Press, 396 pp.

CHAPPE (1993) L'infographie de presse, Centre de Formation aux Techniques du Journalisme, 125 pp.

CHERNOFF H. (1973), The use of faces to represent points in k-dimensionnal space graphically, *J. Am. Statist. Ass.*, 68, 342, 361-368.

CLEVELAND W.S., HARRIS C.S., Mc GILL R. (1982), Experiments on quantitative judgements of graphs and maps, Bell System Technical Journal.

CLEVELAND W.S. (1993a), Visualising Data, Hobart Press, 369 pp.

CLEVELAND W.S. (1993b), A model for studying display methods of statistical graphics, Journal of computational and graphical statistics, 2, 4, 323-343.

COLLINS B.M. (1993), Data visualisation - has it all been seen before? in R.A. Earnshaw and D. Watson (Eds.), Animation and Scientific Visualisation : Tools and applications, New-York, Academic Press, 3-28.

CORBETTA M, MIEZIN F.M., DOBMEYER S., SHULMAN G.L., PETERSEN S.E. (1991), Selective and divided attention during visual discrimination of shape, color and speed : functional anatomy by positrons emission tomography, *J. Neurosci.*, 11, 2383-2402.

COSTE J.-F., VALOIS J.-P. (2000), An Innovative Approach for the Analysis of Production History in Mature Fields : A Key Stage for Field Re-engineering. Ann.Fall Meeting, Soc. of Petroleum Eng., Dallas, 1-4 oct. 2000, ref SPE 62880.

DANEK A.M., KOUBEK R.J. (1995), Mapping perceptual and cognitive processing for the effective use of graphical displays in shop floor scheduling tasks, The Intern. Journal of Human factors in manufacturing, 5, 4, 401-415.

DEHAENE S., DEHAENE-LAMBERTZ G., COHEN L. (1998), Abstract representations of numbers in the animal and human brain. Trends Neurosci., 1998, 21, 355-361.

DEHEANE S., SPELKE E., PINEL P., STANESCU R., TSIVKIN S. (1999), Sources of mathematical thinking : behavioral and brain-imaging evidence, Science, 284, 970-4.

DESCARTES R. (1637), Discours de la méthode, 4e partie, in Œuvres publiées par Victor Cousin, F.G. Levrault, Paris, 1824, T.5.

EHRENBERG A.S.C. (1977), Rudiments of numeracy, J. Royal Statist. Soc., Series A, 140, 277-297.

EELLS W.C. (1926), The relative merits of circles and bars for representing components parts, JASA, 22, 473-482.

GAUTHIER A. (1996), Du visible au visuel, anthropologie du regard, Presses Universitaires de France, Sociologie d'aujourd'hui, 197pp.

GNANADESIKAN R. (1977), Methods for Statistical Data Analysis of multivariate observations, Wiley, New-York.

GOETTL B.P., WICKENS C.D. et KRAMER A.F. (1991), Integrated displays and the perception of graphical data, Ergonomics, 34, 8, 1047-1063.

GOLDBERG K.M., IGLEWICZ (1992), Bivariate extensions of the boxplot, *Technometrics*, 34, 307-320.

GUNTER B. (1994), Visualizing data by W.S. Cleveland, Summit, *Technometrics*, 36, 3, 314-315.

HAVELANGE C. (1998), De l'œil et du monde, une histoire du regard au seuil de la modernité, Fayard, Paris, 427pp.

HELMOLTZ H. von (1896), Handbuch des Physiologischen Optik, Dritter Abschnitt, Zweite Auflage, Voss, Hambourg, Pub. Ang. Helmoltz's treatise on Physilogical Optiks, éd. J.P.C. Southall, 1924, The Optical Society of America, rééd. Dover, New-York, 1962.

IMBERT M. (1983), La neurobiologie de l'image, La Rercherche, 14, 144, 615-623.

INSELBERG A. (1985), The plane with parallel co-ordinates, Special issue on computational geometry, *The visual Computer*, I, 69-97.

JULESZ B. (1991), Early vision and focal attention, Reviews of Modern physics, 63, 735-772.

KEPLER J. (1611), La dioptrique, cité par HAVELANGE, 1998.

KOSSLYN S.M. (1991), A cognitive neuroscience of visual cognition, further developments, in Mental images in human cognition, Logie R.H. et Denis M. Ed, Elsevier, 351-381.

KOSSLYN S.M. (1994), Elements of graph Design, Freeman and Co., New-York, 309 pp.

KRUSKAL J.B., WISH M. (1978), Multidimensionnal scaling, Sage, Berverly Hills.

LALANNE L. (1846), Mémoire sur les Tables Graphiques et sur la Géométrie Anamorphique appliquées à diverses questions qui se rattachent à l'art de l'ingénieur, Annales des Ponts et Chaussées, 2e série, 11, 1-69.

LA ROCHE H. de, STUSSI J.M., CHAURIS L. (1980), Les granites à deux micas hercyniens francais, essais de cartographie géochimique appuyés sur une banque de données, implications pétrologiques et métallogéniques, in 26e Congrés géologique international, Livert-guide d'excursion, pub. Sciences de la Terre, XXIV, 1980, 1, 5-121.

LEBART L., MORINEAU A., PIRON M. (1995) Analyse exploratoire des données, Dunod.

LE GUEN M. (1995), Statistique, Imagerie et Sciences Cognitives, Bull. Méthodologie Sociologique, 9, 90-100.

LEONARD de VINCI, Textes choisis, pensées, théories, préceptes, fables et facéties, traduits dans leur ensemble pour la première fois d'après les manuscrits origi-

naux et mis en ordre, avec introduction, par Peladian, Mercure de France, 1929 ; et Traité de la peinture, édition André Chastel, Paris, Berger-Levrault 1987.

LEWANDOWSKI S., SPENCE I. (1989), Discriminating strata in scatterplots, *Journ. Of Am. Stat. Ass.*, 84, 407, 682-688

LOHSE J. (1993), A cognitive model for the understanding graphical perception, Human Computer Interaction, 8, 353-388.

LOCKE J. (1693), Essai philosophique concernant l'entendement humain, où l'on montre quelle est l'étendue de connoissances certaines et la manière dont nous y parvenons. Traduit de l'anglais par M. Coste, 4e éd. fr., Amsterdam, aux dépens de la compagnie, 1758, 4 vol. (1ère éd. anglaise 1690). Cité par Havelange 1998.

Mac EACHREN (1995), How Maps Work. New-York, The Guilford Press.

MALEBRANCHE N. (1674), De la recherche de la vérité. Où l'on traite de la nature de l'esprit de l'homme et de l'usage qu'il doit en faire pour éviter l'erreur dans les sciences. Sixième édition revue et augmentée de plusieurs éclaiscissements. Paris, Gallimard, 1979 (date de la 6e éd. 1 712) cité par Havelange, 1998.

MALIK J. et PERONA P. (1990), Preattentive texture discrimination with early vision mechanisms - J. Opt. Soc. Am. A., 7, 5, 923-932.

MAREY E.J. (1879), La méthode graphique dans les sciences expérimentales, et principalement en physiologie et en médecine, Masson.

MARTIN M. (1989), The semiology of documents, in IEEE Transactions on professional communications, 32, 3, 171-177.

MAYR G. von (1874) , Gutachen ber die Anwendungen der graphischen und geographischen methode in der Statistik. Munich, Gotteswinter und Mssl.

MILLER G.A. (1956), The magical number seven, plus or minus two, some limits on our capacity for processing information, The psychological review, 63, 2, 81-97.

MITCHELL J.A., BIERS D.W. (1992), Decision statistic mapping and number of information dimensions on decision making with graphical displays, in : Proceedings of the human factor society, 36th. annual meeting, 1503-1507.

PALSKY G. (1996), Des chiffres et des cartes, naissance et développement de la cartographie quantitative française au XIXe siècle, Paris, Comité des travaux historiques et scientifiques.

PASTOUREAU M. (2000), Bleu, histoire d'une couleur, Seuil, 216 pp.

PEROZZO L. (1880), Della Rappresentazione Graphica di una Collectivita di Individui nella Successione del Tempo, Annali di Statistica, 12, 1-16.

PINKER S. (1990), A theory of graph comprehension. In R. Freedle (Ed.), Artificial Intelligence and the Future of Testing, Hillsdale, NJ : Lawrence Erlbaum Associates, 73-126.

PLAYFAIR W. (1786), The commercial and political atlas, London, Carry.

POMERANZ J.R., SAGER L.C., STOEVER R.J. (1977), Perception of wholes and of their components parts : some configural superiority effects, Journal of experimental psychology, Human perception and preformance, 3, 422-435.

POSNER M.I. et PETERSEN S.E. (1990), The attention system of the human brain, Annu. Rev. Neurosci., 13, 25-42.

PRIESTLEY J. (1765), A chart of Biography, London.

RICHAUDEAU F., GUAQUELIN M. et F. (1966), La lecture rapide, éd. Retz, Paris, 320 pp.

ROUSSEEUW P.J., RUTS I. (1997), Bivariate Location Depth, *Appl. Statist.* (JRSS C), 45, 153-168.

SIEGEL J.H., GOLDWYN R.M., FRIEDMAN H.P. (1971), Pattern and Process in the evolution of human septic shock, Surgery, 70, 232-245.

SIEGRIST M. (1996), The use or misuse of three dimensional graphs to represent lower-dimensional data, Behaviour et Information Technology, 15, 2, 96-100.

SIMKIN D. et HASTIE R. (1987), An Information-Processing Analysis of Graph Perception, J. Am. Statist. Ass., 82, 398, 454-465.

SPENCE I., LEWANDOWSKI S. (1990), Graphical Perception, in Modern Methods of Data Analysis, Fox J. Long J.S. Ed, Beverly Hills, Sage, chap. 1, 13-57.

TILLING L. (1975), Early Experimental Graphs, British Journal for the History of Science, 8, 193-213.

THEUS M. (1995), Treillis displays vs. Interactive graphics, Computational statistics, 10, 113-127.

TREISMAN A. (1997), Features and Objects in Visual processing.

TUFTE E. (1983), The Visual Display of Quantitative Information, Graphics Press, Cheshire, Connecticut.

TUKEY J.W. (1977), Exploratory data analysis, Addison Wesley Publishing Company.

TUKEY J.W. (1990), Data-Based Graphics : Visual Display in the decades to come, statistical science, 5, 3, 327-339.

TUKEY J.W. (1993), Graphic comparison of several linked aspects : alternatives and suggested principles, *Journal of computational and graphical statistics*, 1993, 2, 1, 1-33.

TVERSKY B. et SCHIANO D.J. (1989), Perceptual and conceptual factors in distortions in memory for graphs and maps, J. Exp. Psy. Gen. 118, 4, 387-398.

VALOIS J.-P. (1986), Mise en oeuvre interactive des choix algorithmiques : application à l'analyse factorielle des données géochimiques, in Data Analysis and Informatics, IV, Diday and al éd., Elsevier (North Holland), 625-641.

VALOIS J.P. (1991), Le leucogranite peralumineux de Mortagne (Vendée, France) : analyse statistique et cartographie géochimique appliquées à la recherche de gisements d'uranium, Bull. Centres Rech. Elf Exploration Prod. Elf Aquitaine, 15, 2, 507-521.

VALOIS J.-P. (1993), Perception visuelle des documents écrits, implications en ergonomie logicielle, Revue des questions scientifiques, 164, 2, 151-180.

VALOIS J.-P. (1998a), Analyse exploratoire et régression, application à l'étude de l'enfoncement d'une plateforme pétrolière, *Journées SFdS*, Rennes.

VALOIS J.-P. (1998b), Infographie, outil de data mining, Compte-rendus du Club SAS 1997, 8/10 oct. 1997, Paris, pub. SAS Institute S.A., Grégy-sur-Yerres.

VALOIS J.-P. (1999), Une typologie des graphiques statistiques, Journées SfdS, Grenoble.

VALOIS J.-P. (2000), Les boîtes de dispersion en coordonnées parallèles, 6^{e} Journées Agro-Industrie et méthodes statistiques, Pau, Janv. 2000, organisées par la Soc. Fr. de Statistique.

WAINER H., THISSEN D. (1981), Graphical data analysis, *Ann. Rev. Psychol.*, 32, 191-241.

WERTHEIMER M. (1945), Productive thinking, New-York, Harper, réée. 1982, Phoenix éd. Chicago, Univ. of Chicago Press.

WILKINSON L. (1994), Less is more : two and three dimensional graphics for data display, in : Symposium on data visualization, Behaviour research methods, intruments and computers, 26, 2, 172-176

WILKINSON L. (1999), The Grammar of Graphics, Statistics and Computing, *Springer*, 408pp.

WOLD H., (1985), Partial Least Squares, in Encyclopedia of Statistical Sciences, Vol. 6, Kotz S. et Johnson N.L. ed., John Wiley & Sons, Now-York, 581-91.

WOODS D.D. (1991), The cognitive engineering of problem representations. In G.R.S. Weir and J.L. Alty ed, Human computer interaction and complex systems, 169-188, London Academic.

WORSLEY K.J., (1987), Un exemple d'identification d'un modèle log-linéaire grâce à une analyse des correspondances, *Revue de Statistique Appliquée*, XXXV, (3), 13-20.

ZANI S., RIANI M., CORBELLINI A. (1998), Robust bivariate boxplots and multiple outlier detection. *Computational Statistics and Data Analysis*, 28, 257-270.

ANNEXE 1 : SCHÉMATISATION DE LA TYPOLOGIE PROPOSÉE

• Présentation générale

La typologie proposée, schématisée dans la figure 8, distingue trois systèmes de coordonnées : cartésiennes, radiales ou parallèles. On assimile à ce dernier système décrit par INSELBERG (1985) tous les cas où une variable catégorielle est portée le long d'une direction, autorisant alors des permutations (BERTIN 1977) qui n'ont pas de sens sur un axe cartésien. La possibilité de permutations est symbolisée dans les sous-figures par un trait tireté.

Le système radial représente les variables à partir d'un point repère, par la longueur de segments répartis selon différentes directions. Le système radial peut être combiné au système cartésien, par exemple en utilisant chaque localisation géographique comme centre d'un polyèdre (figure 6), propriété qui n'est pas explicitée dans cette figure de synthèse. On assimile ici au système radial la représentation de rayons concentriques de longueur égale, les grandeurs étant restituées par les angles successifs.

Le second principe de cette typologie porte sur le type et le nombre de variables : existence d'une (ou plusieurs) variable catégorielle permettant de distinguer des sous-populations, et nombre de variables quantitatives.

À chaque situation définie par un système de coordonnées et une configuration (type, nombre de variables) correspond une rangée horizontale du schéma. Les rangées sont partagées en trois colonnes correspondant à trois situations : représentation globale de la population, représentation destinée au repérage de chaque unité statistique, ou comparaison de celles-ci à un modèle.

Cette typologie ne prétend pas couvrir tous les graphiques ; le lecteur trouvera des graphiques qui ne « rentrent » pas bien dans ce cadre ; l'intérêt d'une typologie est de montrer l'existence de cases inexplorées ou sous-utilisées. En outre les contraintes de place ou de typographie ne permettent pas d'explorer de façon exhaustive tous les cas possibles ; le lecteur pourra compléter le cadre proposé. Pour des raisons de lisibilité, chaque sous-figure correspond à un exemple schématisé, les sous-figures étant sans lien de contenu les unes par rapport aux autres.

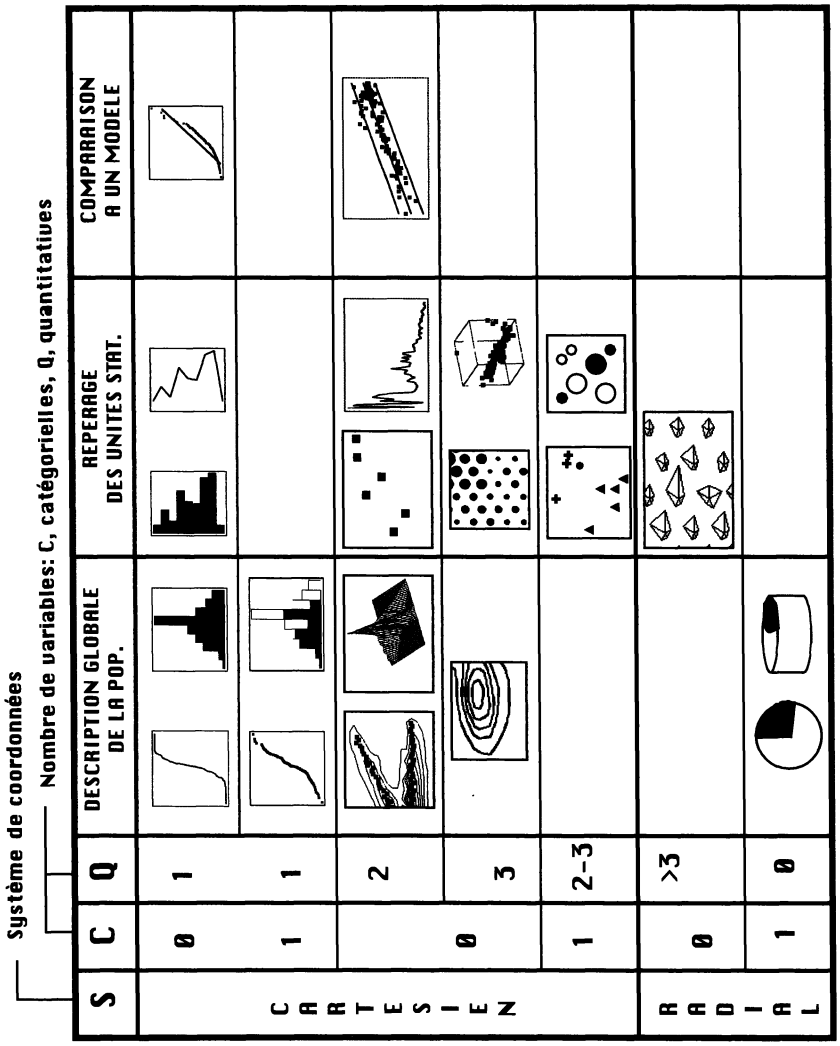

FIG 8. — Présentation schématique de la typologie.

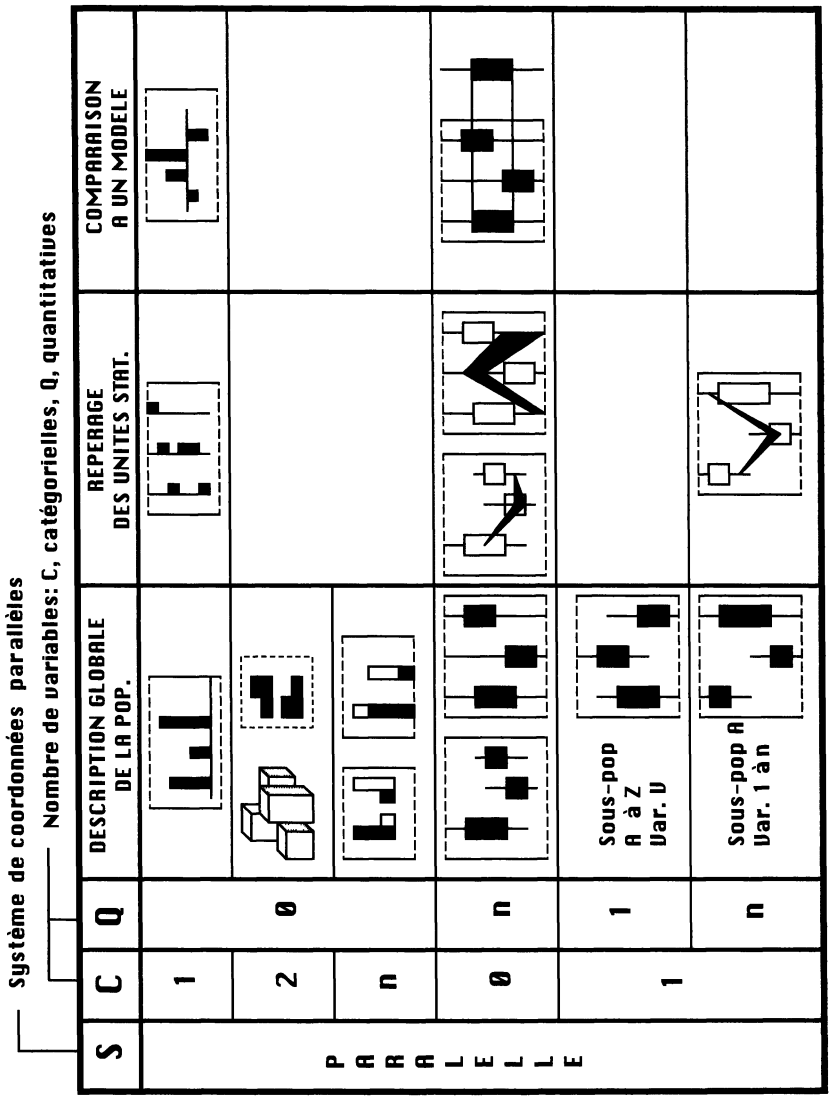

FIG 8 (suite). — Présentation schématique de la typologie.

• Description globale de la population, en coordonnées cartésiennes ou radiales

Si l'on recherche une description globale de la population, la première rangée (coordonnées cartésiennes, une seule population, une variable quantitative) regroupe la courbe de fréquence cumulée, et l'histogramme (dans lequel la variable quantitative est codée en classes, donc d'une certaine façon

transformée en variable catégorielle). Si une variable catégorielle permet de distinguer deux sous-populations (seconde rangée), on peut former le graphique des quantiles, ou porter deux couleurs dans l'histogramme.

Cette représentation globale de la population, qui fait appel à la densité de la variable, est parfaitement courante pour une variable quantitative, mais est sous-employée quand on dispose de deux variables quantitatives (troisième rangée) : une représentation en iso-valeurs de la densité des points dans un graphique 2D est pourtant possible (Valois, 1993), que cette représentation se fasse dans le cadre orthonormé ou avec effet de perspective en 3D (troisième rangée).

La quatrième rangée (3 variables quantitatives) propose les courbes d'isovaleurs, d'usage courant en cartographie.

Si l'on dispose uniquement d'une variable catégorielle, sans variable quantitative (dernière rangée), le mode de représentation couramment pratiqué est le « camembert », qu'il soit sous forme de cercle ou avec un effet de perspective en 3D : on voit ici, comme dans le cas des isovaleurs, que la typologie proposée évite les confusions qui pourraient naître si l'on raisonnait sur la dimension du graphique. Les études d'ergonomie montrent dans le cas du « camembert » qu'il n'y a guère de différence pour l'efficacité de la perception entre la représentation 2D ou pseudo-3D.

• Repérage des unités statistiques en coordonnées cartésiennes ou radiales

Le repérage des unités statistiques pour lesquelles on dispose seulement d'une variable quantitative peut être obtenu en les représentant individuellement dans leur ordre d'apparition par des barres horizontales proportionnelles à la valeur quantitative représentée (première rangée) ; une ligne brisée peut être également utilisée.

Si l'on dispose de deux variables quantitatives, on obtient la classique représentation orthonormée. Une représentation par une ligne joignant les points peut être préférée si la succession des données est organisée (série chronologique par exemple).

Dans le cas de trois variables quantitatives, on peut porter à chaque point du graphique orthonormé des boules dont la surface (ou le rayon) est proportionnel à la troisième grandeur. Cette représentation est pratiquée en cartographie, où les deux premières variables sont les coordonnées dans l'espace. Les ordinateurs rendent aujourd'hui l'accès aisé à une présentation des points en fausse perspective donnant l'apparence d'un cube (quatrième rangée).

La cinquième rangée traite le cas d'une variable catégorielle et de 2 ou 3 variables quantitatives. Les graphiques orthonormés admettent la prise en compte de plusieurs populations (en jouant sur le type de symboles, leur orientation, ou leur couleur) ; ce type de représentation est très répandu et il n'est pas besoin d'insister ici. Les symboles peuvent être de taille constante (2 variables quantitatives) ou proportionnelle à une troisième variable.

L'utilisation de coordonnées radiales permet de prendre en compte un nombre plus important de variables quantitatives. Les polygones peuvent être disposés linéairement dans la page selon l'ordre des individus, ou être affectés aux coordonnées xy des points, par exemple en fonction de leur position géographique. L'épaisseur des traits, ou leur couleur, permet de représenter l'appartenance à des sous-populations donc la prise en compte d'une variable catégorielle (non représentée ici pour raisons typographiques).

• Comparaison à un modèle

La comparaison à un modèle est possible graphiquement. En mode monovarié, on retrouve la droite de Henry ; en mode bivarié, l'usage de graphiques pour juger de la qualité d'une régression avec sa zone de confiance est trop classique pour qu'il soit nécessaire d'insister.

• Coordonnées parallèles, variables catégorielles

La représentation d'une variable catégorielle s'effectue en coordonnées parallèles par une série de barres reflétant le pourcentage de chaque catégorie, ce qui relève donc d'une description globale de la population.

Le croisement de deux variables catégorielles fait intervenir une matrice représentant ces variables, les fréquences étant portées soit en relief (diagrammes en barre en 3 dimensions) soit en colorant en à plat une proportion de la cellule (matrice permutable de Bertin, 1977).

En troisième rangée, l'utilisation de couleurs dans les barres permet de combiner plusieurs catégories (exemple classique : les proportions de A, de B... dans des années successives). Un autre mode de représentation, visuellement peu efficace, consiste à placer différentes barres côte à côte ; ces graphiques sont très utilisés notamment en gestion (troisième rangée).

Le repérage des unités statistiques peut être effectué aisément en axes parallèles (un axe par variable).

On assimile ici à une comparaison à un modèle la représentation en coordonnées réduites (centrage sur la moyenne, BERTIN, 1977), l'effet visuel obtenu dépend en effet de l'écart à une équirépartition.

• Coordonnées parallèles, variables quantitatives

Les boîtes de dispersion proposées par Tukey (1977) forment une représentation globale de la population en axes parallèles (on peut permuter l'ordre horizontal des boîtes sans altérer le sens).

La typologie proposée suggère d'autres utilisations de ces boîtes. Si l'on veut comparer des sous-populations, on peut représenter chaque sous-population en normant le graphique sur les minima et maxima de la population globale (rangée : 0,n, sous-figure droite en première colonne) : la répartition des boites de la sous-population indique alors immédiatement si celle-ci, pour une variable donnée, couvre l'ensemble de la distribution globale, ou au

contraire est décalée vers les minima ou maxima. En présence d'une variable quantitative, on peut représenter côte à côte les sous-populations; si l'on dispose de plusieurs variables quantitatives, on peut représenter côte à côte les variables en produisant un graphique par sous-population.

L'intégration des résultats d'INSELBERG (1985) permet de représenter des unités statistiques ou des groupes cohérents (pour raisons de lisibilité) d'unités statistiques sur les coordonnées parallèles (détails en annexe 2), on range cette possibilité dans la colonne repérage. Moyennant précautions de lisibilité, cette représentation peut être superposée aux boîtes de Tukey (Figure 6) .

Enfin, la comparaison (graphique) à un modèle peut être obtenue en portant en marge la distribution d'une variable aléatoire de référence, par exemple de distribution normale.

ANNEXE 2 : PRINCIPES DE REPRÉSENTATION EN COORDONNÉES PARALLÈLES

En coordonnées parallèles, une unité statistique peut être représentée en joignant par des segments de droite ses coordonnées sur les différents axes. Schématiquement, à un point du repère cartésien (limité pratiquement à deux axes, en haut à gauche) correspond donc un segment (ou une suite de segments) en coordonnées parallèles (où le nombre d'axes, donc de variables, n'est plus limité). Inversement, un ensemble d'unités statistiques alignées selon une droite du repère cartésien va se traduire en coordonnées parallèles par un point unique, intersection des segments représentatifs de chaque individu (en haut à droite). L'organisation des segments dans le repère parallèle répond à une logique étudiée par INSELBERG (1985). Notons m la pente et b l'intercept entre deux variables, en nous limitant ici au cas simple de la relation linéaire $y = mx + b$ (sous-figure inférieure).

Les segments se coupent dans le couloir entre les deux axes si la pente m est négative; le point de rencontre est équidistant des axes si $m = -1$, décalé vers l'un ou l'autre des deux axes si m est différent de -1. Les segments sont parallèles si $m = 1$ et, dans ce cas, horizontaux si $b = 0$ (pour des axes pareillement gradués). Si m est positif et différent de 1, ils convergent vers l'extérieur des deux axes, à droite si m est entre 0 et 1, à gauche si m est supérieur à 1.

Une heureuse surprise attend le statisticien dans ce repère. Quand le coefficient de corrélation r^2 entre les deux variables tend vers 1, la zone d'intersection des segments tend vers un point unique. Sinon, la zone d'intersection des segments va être d'autant plus large que le coefficient de corrélation va se dégrader, la figure illustre le cas où m est négatif. Les directions des segments vont sembler anarchiques pour les plus mauvaises corrélations. Outre que la lecture d'un point unique pour une bonne corrélation est sympathique, rappelons que l'avantage de la représentation en coordonnées parallèle est d'offrir la vision simultanée des relations entre un grand nombre de variables (prises deux à deux, dans la pratique une vingtaine ou une trentaine de variables peuvent être admises dans le format A4).

Les axes peuvent être habillés avec les boîtes de dispersion proposées par Tukey (voir figure 7). Comme l'intersection des segments peut donner lieu à des figures embrouillées si les unités statistiques sont nombreuses et les corrélations faibles, l'expérience conduit à utiliser cette représentation pour illustrer la position dans le repère multivarié d'un groupe particulier d'individus, soit peu nombreux, soit suffisamment corrélés; on peut ainsi montrer les différences que présente une sous-population particulière par rapport à la population d'ensemble, ou la position multivariée des individus aberrants que l'on ne souhaite pas intégrer aux analyses numériques et pour lesquels on dispose ainsi d'une arme descriptive puissante et concise.

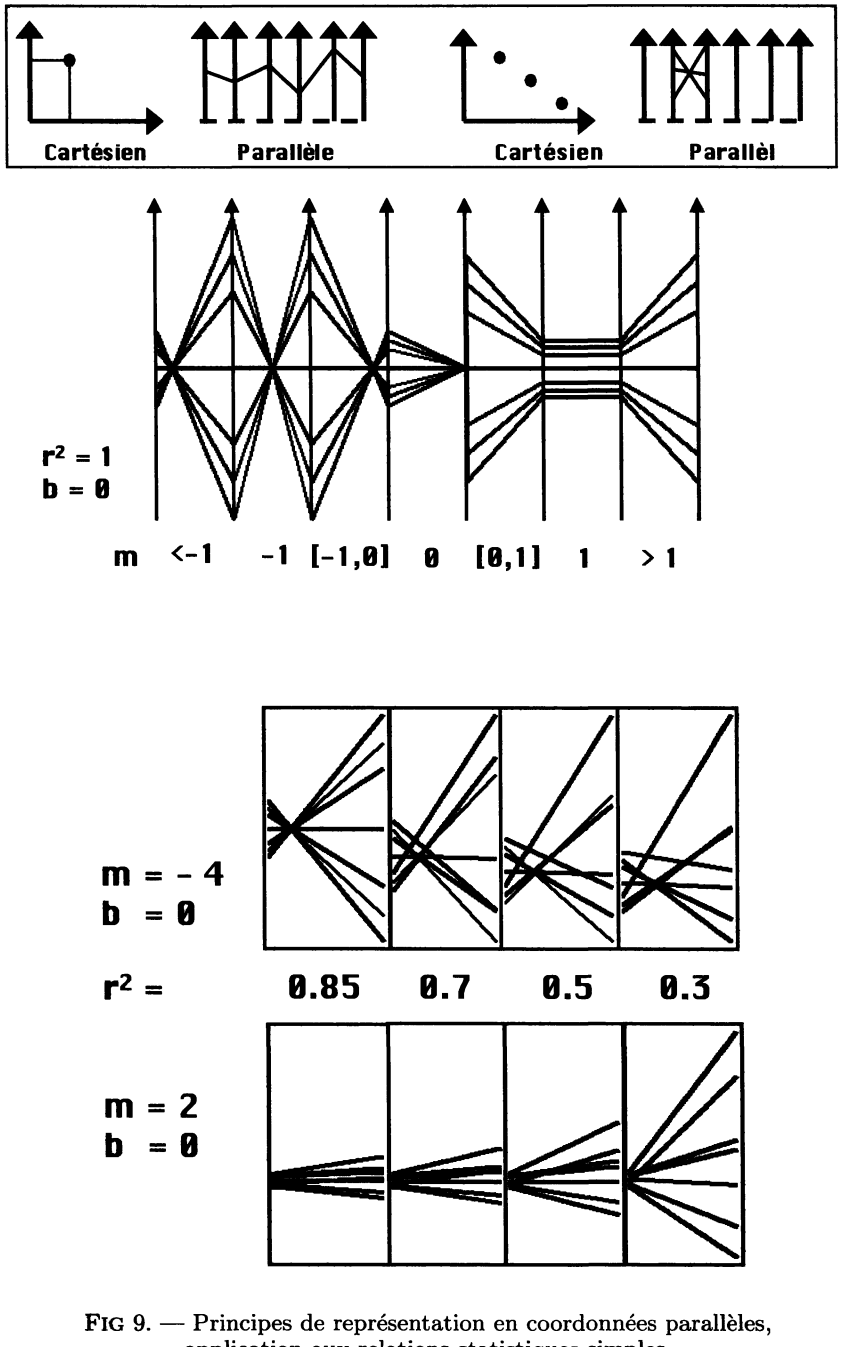

FIG 9. — Principes de représentation en coordonnées parallèles, application aux relations statistiques simples.