L'analyse exploratoire des données : une approche interactive nouvelle de l'information statistique

Philippe Waniez[a]

Résumé

Peu normative, l'Analyse Exploratoire des Données (en anglais *Exploratory Data Analysis, EDA*) imaginée par le statisticien J.W. Tukey (de l'Université de Princeton et des Laboratoires AT&T Bell) insiste sur l'inadaptation fréquente des hypothèses sous-jacentes à la statistique classique, hypothèses souvent trop fortes au regard de la complexité des univers analysés. Elle cherche, de plus, à prendre mieux en compte les anomalies ou les cas extrêmes, trop souvent considérés comme aberrants, car s'ajustant mal aux "lois" statistiques.

Au lieu de rechercher à tout prix l'adéquation à un test statistique, et de prendre, de manière quasi rituelle, une décision de type probabiliste, l'analyse exploratoire s'intègre à un processus de recherche combinant les deux méthodes. L'approche exploratoire conduit à "radiographier les données", à chercher ce qui se passe dans les chiffres, sans a priori.

Faisant suite à la publication, par le GIP RECLUS, d'un ouvrage consacré à l'Analyse Exploratoire, cette communication tente de montrer l'originalité de l'outil central de l'analyse exploratoire multivariée : la toupie.

[a]ORSTOM & Maison de la Géographie 17, rue Abbé de l'Epée 34000 Montpellier

1 Les principes de l'exploration multivariée

En mettant au point leur logiciel PRIM-9, en 1972, à l'Université de Standford, J.W. Tukey, M.A. Fishkeller et J.H.Friedman ont mis en pratique les principes de l'exploration multivariée telle qu'ils la proposaient. En effet, PRIM est formé par les initiales des 4 opérations de base grâce auxquelles l'exploration d'un nuage de points multidimensionnel devient une réalité.

1.1 Quatre principes pour une méthode

- P pour Projection
  Dans le monde réel, les objets sont observés en perspective : un même objet apparaît d'autant plus petit qu'il est éloigné de l'observateur. De plus, la combinaison par le cerveau des images transmises par les deux yeux permet de rendre aux objets leur relief. Malheureusement, les nuages de points multidimensionnels auxquels font appel les statisticiens pour analyser leurs données n'ont pas d'existence matérielle. Il faut donc recourir, comme le font les différentes méthodes d'analyse factorielle, à la projection des points de l'espace multidimensionnel sur un plan .
- R pour Rotation
  La rotation permet de créer l'illusion de la troisième dimension. En regardant le nuage de points sous divers angles on cherche à identifier des organisations particulières. Cette reconnaissance des formes du nuage de points ouvre la voie de l'interprétation des données statistiques .
- I pour Isoler
  Isoler un ensemble de points pour mieux les observer revient à s'interroger sur l'existence de groupes présentant des caractéristiques particulières. L'isolement consiste, d'une part, à étudier le groupe pour lui-même, en définissant un sous-ensemble d'observations devant être analysé à part, et d'autre part, à examiner ce groupe par rapport aux autres observations "ou aux autres groupes", en les marquant par un signe ou une couleur particulière .
- M pour Masquer
  En masquant certaines parties du nuage de points, en fonction de critères qui n'ont pas contribué directement à sa construction, on cherche à discriminer les observations a priori. Ainsi, il est possible de faire des hypothèses sur le rôle joué par telle ou telle autre caractéristique .

1.2 Des formes significatives

Sans limiter l'Analyse Exploratoire à une simple observation de graphique, son principal apport réside néanmoins dans la recherche de formes récurrentes sur les graphiques bi- ou trivariés, formes qu'il faut s'efforcer de reconnaître. En les classant de la moins intéressante à la plus intéressante, on peut distinguer (figure n° 1) :

- les nuages de points en forme de disque (A) ou d'ellipse peu allongée correspondant à une distribution normale. Très importantes en statistique inférentielle, car elles correspondent à certaines conditions d'échantillonnage, les distributions normales sont les moins intéressantes en analyse exploratoire ;

- les alignements de points (B) sont d'un plus grand intérêt. En effet, ils expriment l'existence de tendances, de relations entre les variables ;

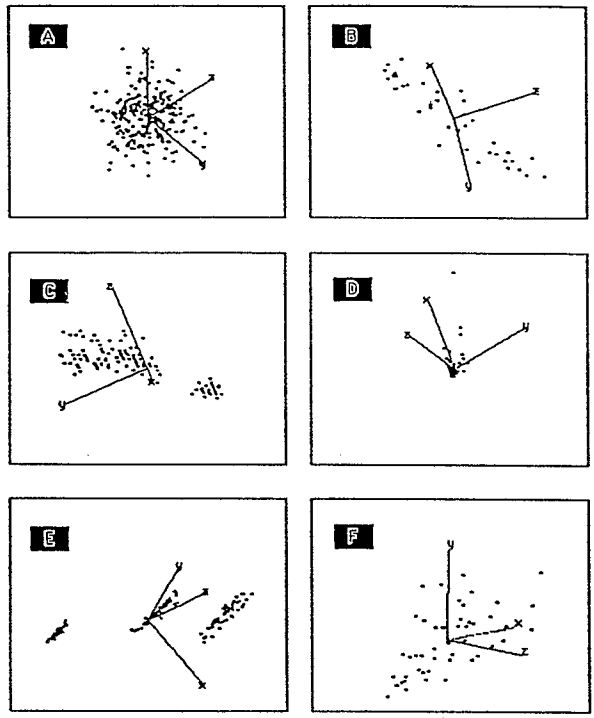

Figure 1: Exemples de formes "significatives" de nuages de points.

- les groupes de points séparés (C) nettement les uns des autres traduisent l'existence de populations différentes au sein du même tableau de données. Dans un tel cas de figure, il apparaît souvent préférable d'isoler chaque groupe d'individus pour les examiner séparément ;

- les surfaces minces traduisent l'existence de combinaisons de variables qui interagissent sur une autre variable. Cette configuration correspond à la régression multiple de la statistique "classique" ;

- les observations exceptionnelles (D), qui n'entrent pas dans les formes décrites ci-dessus, doivent toujours faire l'objet d'un examen particulier. Il peut s'agir d'erreurs de saisie, mais si cela n'est pas le cas, on doit s'interroger sur ces exceptions ;

- enfin, d'autres formes plus complexes apparaissent quelquefois. Elles prennent divers noms comme celui de "bâtons" (E), "d'aile d'oiseau" (F) ou du "lapin à oreille molle".

2 L'analyse d'un nuage de points tridimensionnel : la toupie

L'une des méthodes les plus intéressantes et originales de l'analyse exploratoire multivariée repose sur un graphique trivarié, que l'on peut faire tourner autour de ses trois axes, afin d'observer le nuage de points sous divers angles. Cette figure porte différents noms dans la littérature anglo-saxonne : *3D plot, Spin*, ou bien encore *Rotating plot.* En français, l'expression "Graphique Rotationnel" apparaît parfois, mais il semble à la fois plus français et plus imagé de parler de Toupie, dont la traduction anglaise est Spinning top. En effet, d'après le Dictionnaire alphabétique et analogique de la langue française Robert, une toupie est "un jouet d'enfant, formé d'une masse conique, sphéroïdale, etc., munie d'une pointe sur laquelle elle peut se maintenir en équilibre en tournant". De cette définition, on retient les idées de volume et de rotation qui apparaissent précisément comme les caractères les plus originaux de ce graphique. La métaphore peut s'étendre à la méthode d'analyse elle-même : d'une certaine manière, la toupie constitue un vaisseau d'exploration des galaxies. Chaque étoile représente une observation localisée dans l'espace multidimensionnel (ou multivarié) en fonction de ses valeurs sur les variables formant le système d'axes.

2.1 Construire une toupie

Sur les habituels graphiques bivariés, les nuages de points sont construits en localisant chaque observation en fonction de ses valeurs sur deux variables formant les axes orthogonaux d'un plan. En considérant une troisième variable, on introduit une troisième dimension représentée par un axe orthogonal aux deux autres : le nuage de points acquiert ainsi une épaisseur.

On peut représenter un tel nuage en perspective. Par exemple, chaque commune de Nouvelle-Calédonie forme un point (gros et rond) sur le graphique construit en fonction du pourcentage de ses habitants nés en Nouvelle-Calédonie (%NENC), de celui des agriculteurs par rapport à la population active (enfin, de celui des personnes âgées de 0 à 14 ans par rapport à la population totale (%0-14 ANS). En chaque plan formé par les variables prises deux à deux, on obtient une "boîte" qui renforce l'impression de volume (figure n° 2). Les communes peuvent être projetées sur chaque face, l'ensemble de ces projections formant à son tour un nuage de points (petits et carrés) bivarié.

On notera que, si la lecture d'un seul nuage de points bivarié est aisée, il apparaît plus difficile de retenir et de mettre en relation trois graphiques bivariés simultanément. Par ailleurs, la perspective adopte un angle de vue qui n'est pas toujours le meilleur pour examiner chaque groupe de points. Le rôle de la toupie est précisément de faciliter l'examen du volume sous tous les angles.

La construction d'une toupie correspond au premier principe de l'analyse exploratoire multivariée : P pour Projection. Au lieu de recourir à l'artifice de la perspective pour rendre compte du volume formé par le nuage de points tridimensionnel, on le projette sur un plan figuré par l'écran de l'ordinateur. Selon l'orientation de ce plan par rapport aux axes de référence, la projection du nuage de points révèle diverses configurations, diverses formes qu'il faut interpréter.

Pour faciliter l'observation, on place en général le système d'axes au centre du nuages, sur le point correspondant à la médiane de chacune des variables.

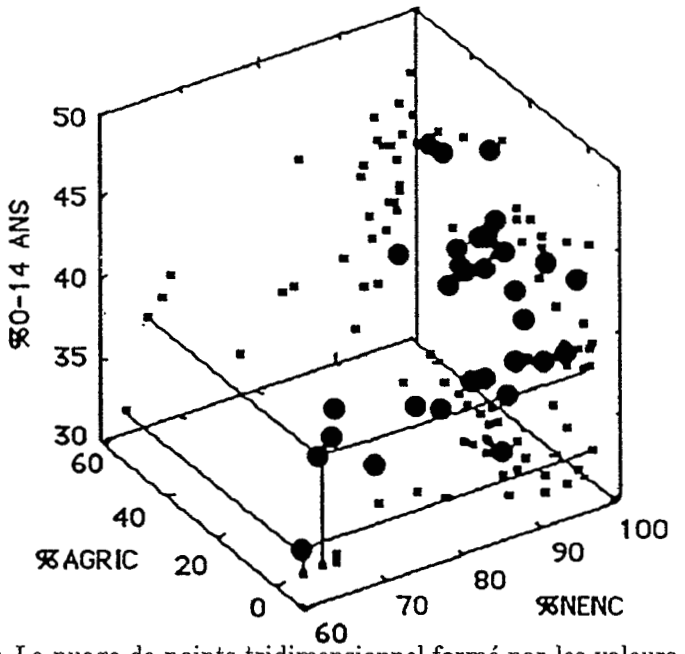

Figure 2: Le nuage de points tridimensionnel formé par les valeurs de communes sur les variables %NENC, %AGRIC et %0-14 ANS.

Lorsque le plan de projection est parallèle à deux axes, le troisième disparaît, ou plus précisément, il est confondu avec l'origine du système d'axes. Dans tous les autres cas, lorsque le plan de projection forme un angle compris entre 0 et 90°, tous les axes demeurent visibles.

Pour illustrer ce propos, nommons X le pourcentage d'agriculteurs dans la population active (%AGRIC), Y, celui des 0-14 ans (%0-14) dans la population totale et Z, la proportion de la population totale née en Nouvelle-Calédonie (%NENC).

Les parties A, B, et C de la figure n° 3 représentent le nuage de points projeté sur des plans parallèles, respectivement à XY, YZ et XZ : seuls les axes concernés sont visibles. Par contre, sur la partie D, les 3 axes (ou, plus exactement, leurs projections) sont visibles, leur longueur dépendant de l'inclinaison par rapport à l'un ou l'autre des axes.

La construction d'une toupie revient donc à choisir les variables relatives aux trois dimensions, à placer les axes sur le nuage de points afin de pouvoir le faire tourner ensuite autour de l'un des axes.

2.2 Un exemple élémentaire de rotation de la toupie

Pour illustrer la technique de rotation, une toupie a été construite à l'aide de trois indicateurs statistiques, issus du recensement de 1980, relevés sur les 27 Etats de la Fédération Brésilienne : le pourcentage de la population de race blanche dans la population totale (nommé BLANCS sur le graphique), le PIB par habitant (PIB) et le taux d'immigration (IMMIG). L'opération consiste à découvrir des regroupements d'Etats, en animant le graphique d'un mouvement de rotation,

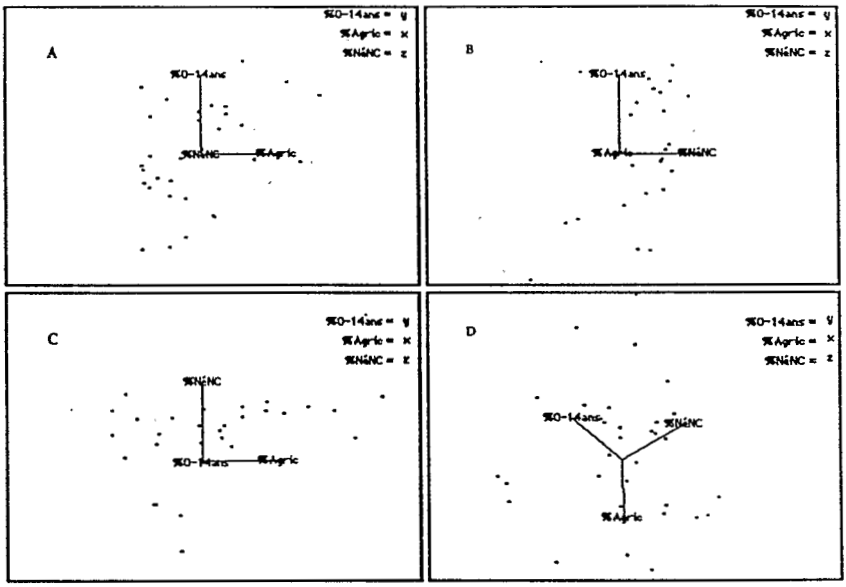

Figure 3: La projection d'un nuage de points tridimensionnel sur des plans d'orientations différentes.

puis en le figeant à chaque fois qu'un groupe semble faire son apparition.

En position initiale, le système d'axes est centré sur la moyenne arithmétique (figure n° 4.A) et figure le plan BLANCS *versus* PIB. Deux ensembles s'individualisent : dans la partie inférieure droite sont agglomérés tous les Etats qui présentent des PIB par habitant et des proportions de blancs inférieurs à la moyenne. En les marquant du symbole distinctif -, on procède à une première partition qui peut directement être exportée, par un simple copier/coller, dans le tableur de Cartographie-2D (les Etats membres du groupe - prennent la valeur 1, et les autres Etats, la valeur 0). La carte obtenue présente une nette dissymétrie entre les Etats des régions *Norte* et *Nordeste*, et le reste du pays.

Pour étudier l'influence de l'immigration sur la régionalisation du Brésil, on procède alors à une première série de rotations du nuage de points (figure n° 4.B) ; celle-ci se fait après effacement des axes. Un regroupement d'Etats se dessine marqué du symbole X. Après restitution du système d'axes, ce groupe figure dans la partie supérieure de l'axe IMMIG : les Etats concernés présentent donc un taux d'immigration supérieur à la moyenne. Après un nouveau copier/coller dans son tableur, Cartographie-2 trace une carte sur laquelle s'individualisent les Etats de la région *Centro-Oeste* (Mato Grosso do Sul, Mato Grosso, Rondonia, Distrito Federal) et d'une partie de la région *Norte* (Para, Amapa et Roraima). Il s'agit des principaux espaces de "frontière" qui, par leur potentiel réel ou supposé, attirent à la fois les firmes industrielles et agro-alimentaires, nationales ou multinationales, et les laissés pour compte du *Nordeste* ou du *Sudeste*.

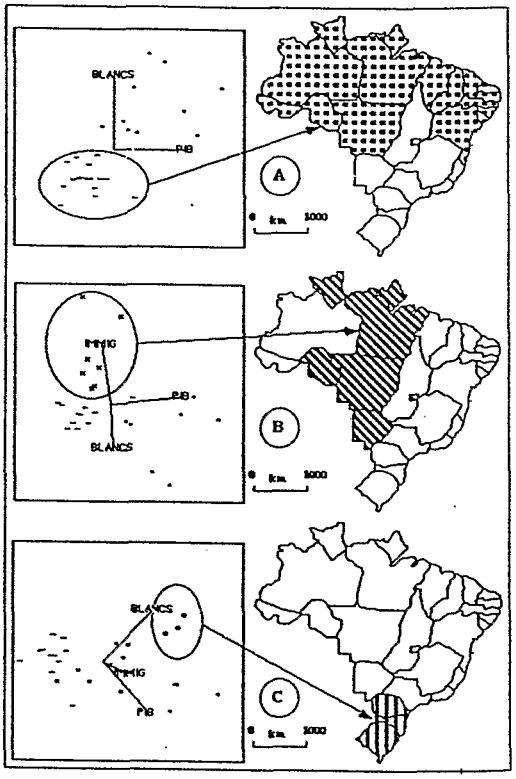

Figure 4: Utilisation de la toupie et cartes associées.

A : position initiale,
B : après une première série de rotations,
C : après une seconde série de rotations.

Enfin, une autre série de rotations de la toupie (figure n° 4.C) permet d'isoler un groupe formé par les Etats de la région Sul (Parana, Santa Catarina, Rio Grande do Sul) caractérisés par une sur-représentation de la race blanche dans la population totale. De fait, il s'agit des espaces d'immigration européenne (Italie et Allemagne, principalement) de la fin du siècle dernier et du début du XXème siècle.

2.3 Toupie et analyse des données

Il y a naturellement une très grande parenté entre la recherche de groupes dans le nuage de points tridimensionnel de la toupie, et une approche plus systématique par la classification automatique, précédée ou non d'une décomposition en facteurs (par l'analyse en composantes principales ou l'analyse factorielle des correspon-

dances). L'analyse exploratoire n'exclut aucunement le recours à l'une de ces techniques ; la toupie peut même contribuer à une meilleure interprétation des résultats.

Par exemple, une analyse factorielle des correspondances a été réalisée sur les résulats de l'élection présidentielle qui s'est déroulée au Brésil en Novembre 1989 (21 candidats, 4500 municipios). Cette AFC conduit à interpréter trois facteurs exprimant chacun une tendance politique (figure n° 5 : populisme (Brizola), libéralisme/autoritarisme (Collor/Maluf), socialisme/libéralisme+autoritarisme (Lulla/Collor + Maluf).

ITER	VAL PROPRE	POURCENT	CUMUL
2	1	0.33669	47.690	47.690
3	1	0.16995	24.072	71.763
4	1	0.07842	11.108	82.871
5	2	0.06271	8.883	91.754
6	1	0.02632	3.728	95.482
7	5	0.02430	3.441	98.923
8	2	0.00760	1.077	100.000

J1	QLT	POID	INR	1#F	COR	CTR	2#F	COR	CTR	3#F	COR	CTR
1 BRI1	1000	165	391	*1288*	*994*	*815*	79	4	6	-57	2	7
2 COL1	981	305	123	-288	289	75	*-353*	*436*	*223*	*-220*	*170*	*188*
3 COV1	796	115	102	-285	131	28	*640*	*658*	*278*	68	7	7
4 LUL1	956	172	89	-188	96	18	-27	2	1	*554*	*838*	*672*
5 ULY1	995	47	104	-148	14	3	*-677*	*296*	*128*	56	2	2
6 MAL1	909	89	138	-468	198	58	*804*	*587*	*337*	*-305*	*84*	*105*
7 AFI1	102	48	35	-145	41	3	53	6	1	-148	43	14
					1000			1000			1000	

BRI1=Leonel Brizola, COL1=Fernando Collor, COV1=Mario Covas
LUL1=Lula da Silva, ULY1=Ulysses Guimarães, MAL1=Paulo Maluf AFI1=Afif Domingos

Figure 5: Eléments d'interprétation de l'AFC sur l'élection présidentielle de 1989 au Brésil.

En complément de l'AFC, une classification ascendante hiérarchique faite à partir des coordonnées sur les trois premiers facteurs a permis de grouper les municipios en 6 classes. Leur cartographie donne une idée de la régionalisation de l'expression électorale. Mais pour en interpréter la signification, la toupie s'avère un excellent auxiliaire (figure n° 6). En effet, sur le même écran d'ordinateur, on peut examiner la forme du nuage de points (ici, un trépied), la localisation des classes dans ce volume (ici la classe n° 6) et sur les histogrammes de chacun des facteurs. On pourrait aussi tracer les histogrammes des variables d'origine.

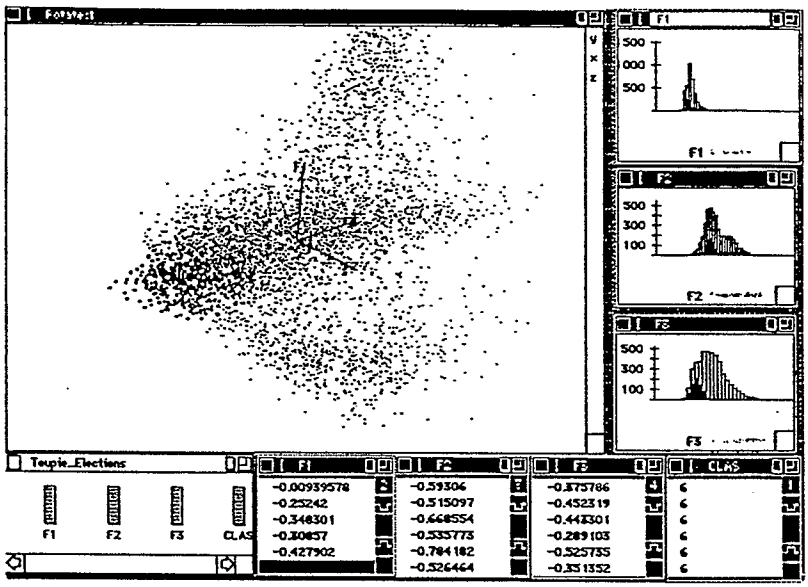

Figure 6: Toupie réalisée à partir de l'AFC sur l'élection présidentielle de 1989 au Brésil.

3 Conclusion : l'analyse exploratoire inséparable de l'informatique

L'approche exploratoire pour l'analyse des données statistiques se compose de deux volets complémentaires formés, d'une part, de techniques spécifiques et d'autre part, d'un environnement informatique particulier. Les techniques spécifiques ont été développées par J.W. Tukey et ses élèves. Elles mettent l'accent sur la visibilité de l'information statistique par des graphiques originaux comme la toupie, et par des résumés statistiques résistants basés sur la médiane et les quartiles, plus que sur la moyenne et l'écart-type. A sa manière, l'approche exploratoire réhabilite les graphiques des "ingénieurs" d'autrefois, en faisant de l'écran de l'ordinateur le moyen d'une "hyper-vision" nécessaire aux décideurs d'aujourd'hui.

Mais l'approche exploratoire, qui ne se limite pas aux représentations graphiques, peut être étendue à toute la statistique "classique" ainsi qu'à l'Analyse des Données. Au lieu d'utiliser ces techniques comme des "boîtes noires" ou des "moulins à statistiques", comme cela est encore très souvent le cas, le point de vue exploratoire invite à regarder les données et les résultats, pas à pas, afin de vérifier en permanence si les analyses sont effectuées dans les conditions d'application optimales. Cette manière de procéder garantit l'analyste contre les problèmes qui ne peuvent manquer de surgir lorsqu'on traite une information avec une méthode

inadaptée.

Ces deux volets de l'approche exploratoire présentent un commun dénominateur : l'informatique. Ainsi, le bon usage d'un logiciel d'analyse statistique, d'un statisticiel, semble désormais aussi important que la connaissance des procédés de calcul. On peut même avancer sans risque que l'ordinateur s'avère être un excellent moyen d'approche expérimentale de la statistique auprès des non-mathématiciens.

Tous les exemples présentés ici ont été réalisés sur Macintosh avec *DataDesk*, logiciel en parfaite adéquation avec les techniques et l'approche exploratoires. Odesta Corporation qui le fabrique, a mis de son côté un atout décisif en consultant P.F. Velleman dont l'expérience en matière d'analyse exploratoire fait autorité aux Etats-Unis. La principale originalité de ce logiciel réside sans doute, dans les liens dynamiques généralisés entre les fenêtres complétés par les menus *hyperview*. De ce fait, l'interactivité autorise une véritable navigation dans les structures numériques pour en découvrir toutes les facettes, même les moins accessibles. On ne peut manquer d'être impressionné par un tel chef-d'-oeuvre de génie logiciel. D'autres logiciels d'analyse exploratoire pour Macintosh sont disponibles sur le marché : *SYSTAT, JMP, MacSpin.* Enfin, pour ceux qui penseraient encore que l'analyse exploratoire n'est pas une voie d'avenir, signalons que le leader en matière de logiciel d'analyse statistique, SAS Institute, vient de publier un nouveau module dénommé *SAS/INSIGHT*, spécialement destiné au système d'exploitation UNIX. De belles explorations en perspective...

Références bibliographiques

Tukey J.W. - 1977 -*Exploratory Data Analysis.* Reading, MA., Addison-Wesley Publishing Company, 688p.

Il s'agit du traité fondamental de l'analyse exploratoire, par son propre inventeur. On y trouve l'exposé des méthodes de représentations graphiques, de résumés numériques robustes, de lissages, etc. Il n'y est fait aucune mention des moyens informatiques utiles.

Hartwig F., Dearing B. - 1982 - *Exploratory Data Analysis.* Sage University Paper, Series : Quantitative Applications in the Social Sciences, n° 16. Beverly Hills, CA, Sage Publications, 83p.

Ce petit livre d'initiation explique comment utiliser les principales techniques de l'analyse exploratoire dans le domaine de la science politique. Un glossaire facilite la familiarisation du lecteur à la terminologie propre à l'EDA.

Jambu M. - 1989 - *Exploration Statistique et informatique des données.* Paris, Dunod, Col. Dunod informatique, 506p.

Issu d'un enseignement d'analyse des données à l'Institut National des Télécommunications, cet ouvrage demande, pour être intégralement compris, un niveau supérieur en mathématiques. Cependant, l'exposé de certaines méthodes

exploratoires ne nécessite pas une telle spécialisation. Les aspects informatiques n'y sont que peu étudiés. Il s'agit, par ailleurs, d'un excellent traité d'Analyse des Données, par l'un des grands spécialistes français de la question.

Velleman P.F. - 1981 - *Applications, Basics, and Computing of Exploratory Data Analysis.* Boston, MA, Duxbury Press.

L'analyse exploratoire exposée par l'un des créateurs de DataDesk.

Bertrand R., Valiquette C. - 1986 -*Pratique de l'analyse statistique des données.* Québec, Presses de l'Université de Québec, 379p.

Cet excellent ouvrage d'initiation à l'analyse des données statistiques en sciences humaines est l'un des rares à reprendre, en langue française, les exposés fondamentaux de Tukey. Les auteurs y expliquent, dans un langage accessible aux non-mathématiciens, les différences fondamentales entre analyse exploratoire et analyse confirmatoire. L'informatique n'y est malheureusement qu'évoquée, surtout pour insister sur les risques qu'elle est supposée faire courir à ses utilisateurs.

Waniez P., - 1991 - *Analyse exploratoire des données.* Montpellier, GIP RECLUS, Col. Reclus Modes d'Emploi, n° 17, 159p.

Adresses utiles

SYSTAT est une marque déposée de SYSTAT Inc.. Importateur en France : Statilogie, 40 rue du Colonel Pierre Avia, 75015 Paris.

DataDesk est une marque déposée de Data Description Inc.. Hyperview est une marque déposée de Data Description. Ce logiciel est diffusé aux Etats-Unis par Odesta Corporation. Importateur en France : Alpha Systèmes Diffusion, Miniparc ZIRST Grenoble Meylan, 43 chemin du Vieux Chêne, 38240 Meylan.

JMP et **SAS/INSIGHT** sont des marques déposées de SAS Institute Inc.. Importateur en France : SAS Institute, 50 avenue Daumesnil, 75012 Paris.

MacSpin est une marque déposée de D2 Software. Importateur en France : Bruno Rives & Associés, 6 avenue Franklin Roosevelt, 75008 Paris.