AUTO-MODELISATION EN ANALYSE DES DONNEES

Daniel Chessel, Jean Thioulouse

AUTO-MODELISATION EN ANALYSE DES DONNEES

Daniel Chessel et Jean Thioulouse

Pour les utilisateurs occasionnels, qui sont nombreux, l'analyse des données est la partie de la statistique descriptive qui produit des cartes factorielles, selon le principe suggéré dans la figure 1. Le choix du plan de projection est défini par un critère d'optimalité.

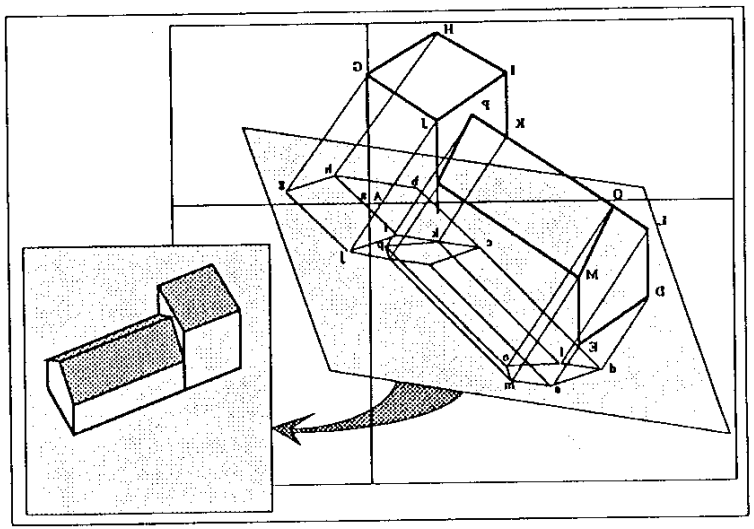

Figure 1 : Schéma de principe d'une représentation euclidienne . Les objets sont des ensembles de points (A, ..., P) d'un espace euclidien (ici de dimension 3) projetés (a, ..., p) sur un plan. La représentation cartésienne des points dans une base orthonormée du plan est dite image euclidienne. L'image (a, ..., p) est aussi un objet. Substituer l'image-objet à l'objet de départ est à la base de la reconstitution des données ou auto-modélisation.

L'idée de regarder "*Sur le plan*", qui semble aujourd'hui si naturelle, n'a pas, loin s'en faut, présidé à la naissance de cette méthodologie. On a d'abord cherché[1] un plan qui s'ajuste au nuage multidimensionnel en minimisant une somme de carrés d'écarts, ce qui a fondé la stratégie de l'analyse factorielle. Substituer aux valeurs observées des valeurs prédites de structure plus simple a été le premier objectif de la statistique multidimensionnelle, laquelle s'impose donc comme une entreprise de *simulation* au sens de la théorie des modèles[2] . Maîtrisé comme outil exploratoire d'une réalité le simulateur devient un *modèle instrument* [3] et l'analyse des données est une technique de modélisation. Son identité propre est associée à la possibilité de laisser émerger les propriétés du modèle des données elles-mêmes[4] : l'exposé propose quelques illustrations de ce point de vue. On y défendra l'idée que la statistique exploratoire n'est en rien étrangère à la modélisation mathématique mais plutôt aide au diagnostic utile en des circonstances variées.

1 Pearson K., On lines and planes of closeest fit to systems of points in space. Phil. Mag., 2, 559-572, 1901

2 Legay J.M., la méthode des modèles, état actuel de la méthode expérimentale, Informatique et Biosphère, Paris, 11-69, 1973. "Un simulateur est un modèle pour lequel l'intersection modèle-objet ne concerne que le domaine des performances".

3 Dès qu'elle est acceptée cette notion première (Legay J.M., Méthodes et modèles dans l'étude des systèmes complexes, Colloque National du ministère de la Recherche et de la Technologie, 17-18 avril 1986, 10 p.) rend caduques les querelles entre "écoles" de statistique.

4 Benzecri J.P., Statistical analysis as a tool to make patterns emerge from data. In Watanab S. Ed. Methodologies of pattern recognition, Academic Press, New York, 35-60, 1969.

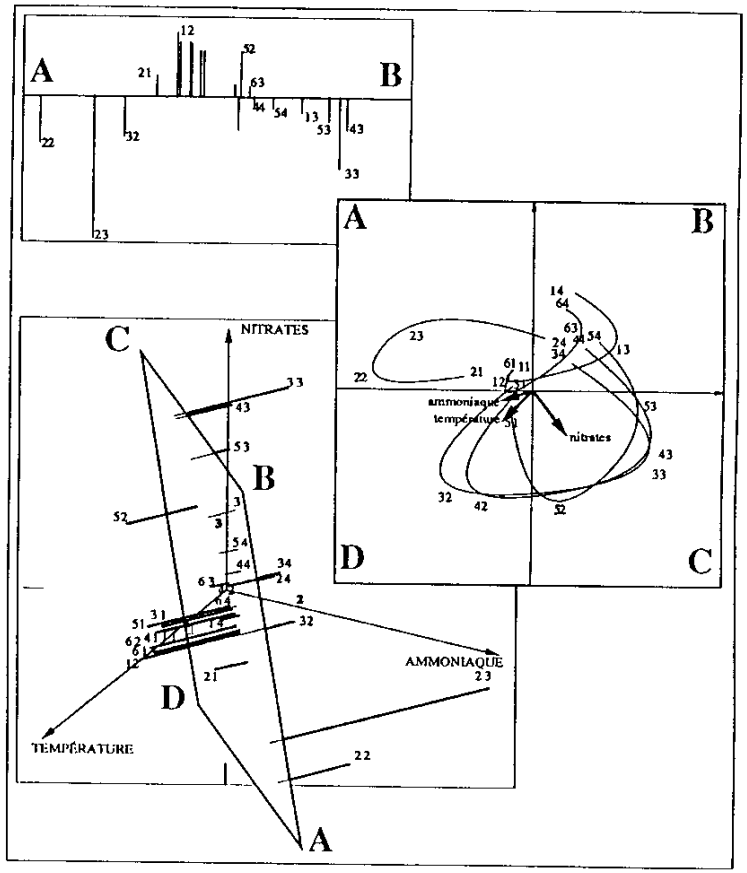

Figure 2 : Projection (opération géométrique) et reconstitution (opération numérique) sont de même nature. 6 stations et 4 dates donnent 24 points de mesure (11,12, ..., 64) où sont enregistrées 3 variables. On obtient 24 points de $\mathbb{R}^3$. ABCD est le plan d'inertie maximale. Vu "de face" il donne la carte factorielle. Vu "de profil" il donne les résidus de l'approximation . Les données sont des points de $\mathbb{R}^3$, les modèles également, les erreurs sont des vecteurs liés (traits gras des projections). Maximiser la variabilité sur le plan ou minimiser l'erreur autour du plan sont équivalents.

Données, Modèles et erreurs

La figure 2 rappelle le principe technique de l'auto-modélisation. Si p variables sont mesurées sur n individus on considère le nuage de n points M_i $(1 \leq i \leq n)$ dans l'espace $\mathbb{R}^p$. Chaque point est projeté sur un sous-espace de dimension réduite ce qui génère une décomposition :

$$x_{ij} = m_{ij} + r_{ij} \quad \text{(donnée = modèle + erreur)}$$

où x_{ij} est la mesure de la variable j sur l'individu i, m_{ij} est la composante correspondante du point projeté et r_{ij} la différence entre ces deux valeurs.

Les méthodes classiques diffèrent entre elles par le choix des poids des points, celui de la métrique euclidienne utilisée dans $\mathbb{R}^p$ et enfin (on pourrait dire surtout) celui de manipulations préliminaires du tableau traité (centrage, normalisation, transformation de variables, ...).

Une diagonalisation génère simplement une base orthonormée (axes principaux), le passage dans cette base donnant les images euclidiennes (cartes factorielles), le retour, après projection, dans la base canonique donnant les modèles et les erreurs. Cette approche géométrique[5], qui se dit sans hypothèse explicite, confère curieusement le même statut au modèle et à l'erreur commise en ce sens que si $m_k(i,j)$ est la $j^{ème}$ composante dans la base canonique de la projection sur le $k^{ième}$ axe principal du $i^{ème}$ point on peut toujours écrire :

$$x_{ij} = \sum_{k=1,r_0} m_k(i,j) + \sum_{k=r_0,r_1} m_k(i,j) + \sum_{k=r_1,p} m_k(i,j)$$

équation du type

$$\text{donnée} = \text{modèle} + ? + \text{résidu}$$

Le ? pouvant être intégré aussi bien au modèle qu'au résidu. Une telle plasticité, à laquelle est associée l'analyse d'inertie, est simultanément outil d'exploration et source de difficultés.

5 Extraordinairement claire chez les fondateurs par exemple Bartlett M.S., The vector representation of a sample *Proc. Cambridge Phil. Soc.* 70, 327-340, 1934, elle est née avec la statistique inférentielle : Fisher R.A., Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, X, 507-521, 1915, et a d'abord concerné les p points de $\mathbb{R}^n$ tandis que Pearson (op. cit.) considèrent les n points de $\mathbb{R}^p$. Le schéma de dualité de l'école dite française a unifié l'ensemble.

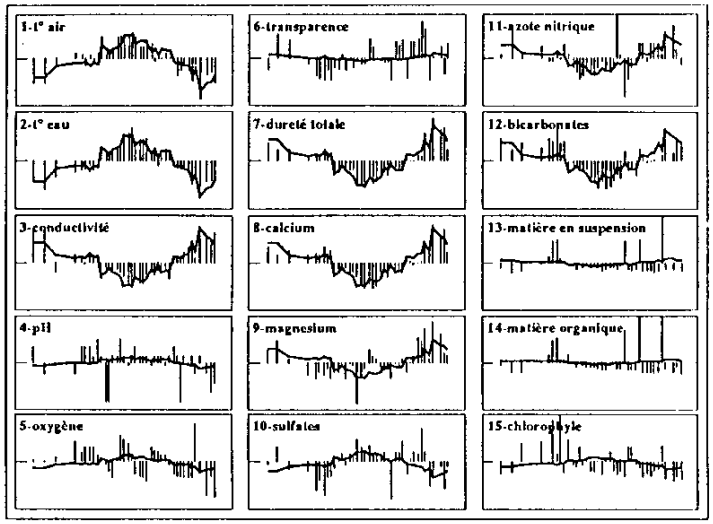

Figure 3 : Auto-modélisation d'une chronique multidimensionnelle par l'analyse en composantes principales normée. Lecture directe des observations (diagrammes en bâtons) et du meilleur modèle de rang 1 (en trait gras), courbe unique qui s'ajuste au mieux au plus grand nombre de variables à un facteur d'échelle près. On exprime ici en langage graphique les aides à l'interprétation numériques. Echelle commune à toutes les fenêtres, en abscisse [0,400] (jours), en ordonnée [-3,+3] (valeurs normalisées). Mise en œuvre de Y. Auda, Rôle des méthodes graphiques en analyse des données : application au dépouillement des enquêtes écologiques. Thèse de 3° cycle, Université Lyon 1, 127 p., 1983. Logiciels GRAPHMU et MACMUL de J. Thioulouse, Laboratoire de Biométrie, Université Lyon 1.

La figure 3 montre comment on peut utiliser cette propriété en analyse en composante principale normée[6] . Si x_{ij} est la valeur centrée réduite de la variable j et si les points M_i ($1 \leq i \leq n$) forment une chronique le modèle utilisant le premier axe principal s'écrit :

$$x_{ij} = \alpha_i \beta_j + r_{ij}$$

qui autorise de placer sur chacune des courbes observées la courbe unique (α_1, ..., α_n) au coefficient d'échelle β_j près. La

6 Données et illustrations extraites de Carrel G. et coll., Approche graphique de l'analyse en composantes principales normée : utilisation en hydrobiologie, *Œcol. Gener.*, 7 (2), 189-203, 1986. On trouvera des illustrations de même logique en analyse des correspondances dans Persat H. et Chessel D. Typologie de distribution en classes de taille : intérêt dans l'étude des populations de poissons et d'invertébrés. *Œcol. Gener.*, 10 (2), 175-195 (1989).

somme des carrés des résidus est minimum. La figure 4 montre l'intégration progressive de plusieurs facteurs. La logique du calcul est linéaire, au sens algébrique, le "simulateur" ne l'est en aucun cas.

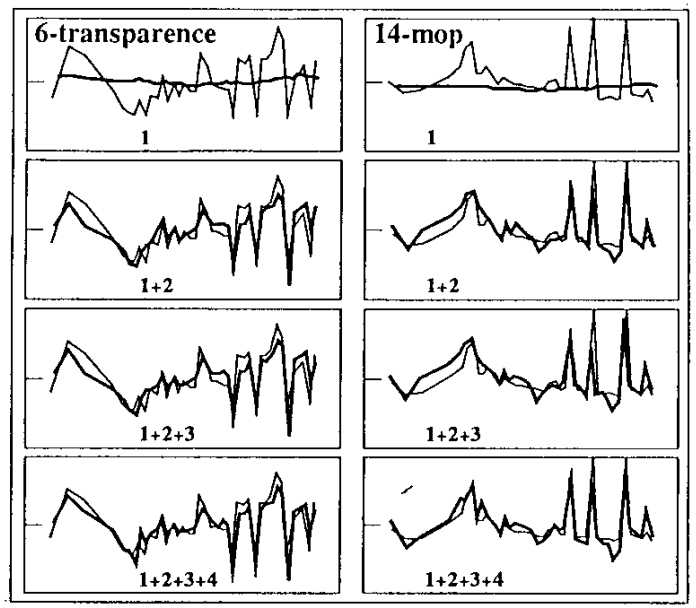

Figure 4 : Dépouillement graphique d'une analyse en composantes principales normée. Reconstitution sur les 4 premiers facteurs des variables 6 et 14 (matière organique particulaire). En trait fin les données, en gras les modèles. L'évolution actuelle des moyens graphiques modifie sensiblement la pratique de la statistique, en particulier celle des méthodes qui donnent "à voir" dans l'information numérique.

Modèles implicites et modèles triviaux

En pratique l'analyse des données n'est perçue comme entreprise d'auto-modélisation que par les spécialistes de la discipline, lesquels sont très minoritaires parmi les utilisateurs des programmes correspondants. Il s'en suit des confusions sévères dont il n'est pas inutile de parler dans un dialogue interdisciplinaire.

Les données numériques contiennent souvent une structure inhérente à l'objet : la température de l'eau est plus forte en juillet qu'en décembre ou le taux d'équipement des ménages est croissant, ... Faire une typologie de courbe de température ou de courbes d'évolution de biens d'équipement demande d'éliminer

ces *effets triviaux*. Pour un tableau à n lignes et p colonnes cet effet t peut s'écrire :

$$x_{ij} = t_{ij} + (x_{ij} - t_{ij})$$

avec :

valeur de t_{ij}	estimation	méthode associée
$t_{ij} = 0$	aucune	ACP non centrée
$t_{ij} = cte$	$\bar{x}$ moyenne générale	ACP à centrage unique
$t_{ij} = a_i$	$\bar{x}_i$ moyenne par ligne	ACP centrée par ligne ou mode Q
$t_{ij} = b_j$	$\bar{x}_j$ moyenne par colonne	ACP centrée par colonne ou mode R
$t_{ij} = a_i + b_j$	$\bar{x}_i + \bar{x}_j - \bar{x}$	ACP doublement centrée
$t_{ij} = a_i b_j$	$\bar{x}_i \bar{x}_j / \bar{x}$	AFC (pondération non uniforme)
$t_{ij} = a_i b_j$	par diagonalisation	ACP non centrée avec élimination du 1° facteur (pondération uniforme)

La combinaison d'une partie imposée (effet trivial) et d'une partie extraite par l'analyse (automodélisation) engendre une grande quantité disponible. Citons :

$$x_{ij} = \sum_{k=1,m} \alpha_{ik} \beta_{jk} + r_{ij} \quad [7]$$

$$x_{ij} = \mu + \rho_i + \gamma_i + Q_i \gamma_j + r_{ij} \quad [8]$$

On comprend alors que pour certains usagers ces méthodes montrent des évidences car la même technique peut être destinée à éliminer cette évidence (pour étudier les résidus) ou au contraire à la mettre en valeur (pour simplifier les données). Les logiciels publics sont pauvres en option, ces options sont implicites. Les programmes ouverts embarrassent par contre l'utilisateur. Le "génie" du programme d'analyse des correspondances (AFC) est de ne pas poser de questions et d'impliquer un double centrage

7 obtenu à partir d'une ACP non centrée, introduit par Whittle P., On principal components and least square methods of factor analysis *Skand. Aktuar.*, 35, 223-239, 1952.

8 obtenu à partir d'une ACP doublement centrée et le premier facteur, introduit par Mandel J. Non additivity in two-way analysis of variance. *J. Amer. Statist. Ass.*, 65, 878-888, 1961.

multiplicatif ... sans préavis. La figure 5 montre qu'on peut y perdre une part substantielle de l'information.

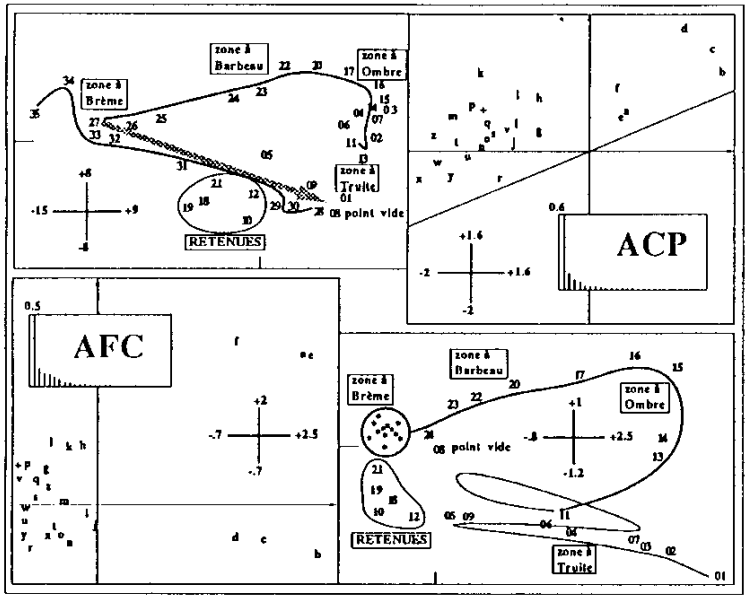

Figure 5 : Automodélisation implicite en œuvre dans un programme d'analyse multivariée. 35 stations réparties le long du cours du Doubs (01, ..., 35) et 27 espèces de Poissons (a,b, ..., +) forment un tableau faunistique.

a Chabot	*b Truite c.*	*c Vairon*	*d Loche f.*	*e Ombre c.*	*f Blageon*
g Hotu	*h Toxostome*	*i Vandoise*	*j Chevaine*	*k Barbeau*	*l Spirlin*
m Goujon	*n Brochet*	*o Perche*	*p Bouvière*	*q Perche s.*	*r Rotengle*
s Carpe	*t Tanche*	*u Brème*	*v Poisson-Chat*	*w Grémille*	
x Gardon	*y Brème b.*	*z Ablette*	*+ Anguille*		

L' ACP centrée de ce tableau (en haut) prend d'abord en compte l'évolution de la richesse et de l'abondance des peuplements, lesquelles augmentent d'amont en aval et diminuent brusquement sous l'effet de la pollution (entre 27 et 28). La rivière se restaure entre les stations 30 et 35. Ce point est éliminé automatiquement en AFC (en bas) par le double centrage. Il est donc possible d'enlever des données par la même technique une perturbation nuisible (par exemple une hétérogénéité de l'intensité d'échantillonnage ou de l'accessibilité des peuplements) ou une part essentielle des résultats. Données de J. Verneaux, Cours d'eau de Franche-Comté (Massif du Jura). Recherches écologiques sur le réseau hydrographique du Doubs. Essai de biotypologie. Thèse d'état, Besançon, 260 p., 1973.

Ces quelques remarques simples sur les effets triviaux invitent à considérer des méthodes factorielles en oeuvre *après une*

modélisation[9] . En effet on peut généraliser la démarche en considérant un modèle paramétrique m du tableau entier (modélisé par ligne, par colonne, ou globalement) et mettre en oeuvre une analyse du tableau :

$$y_{ij} = x_{ij} - m_{ij}$$

qui mettra ou non en évidence un sous-paramétrage du modèle si une part structurée (automodélisable) apparaît ou non dans les résidus de la modélisation. Cette idée est en oeuvre dans les travaux de B. Escofier [10] et si le modèle m est lui-même dérivé d'une approche euclidienne cette stratégie renvoie à la grande famille des analyses sur variables instrumentales[11] .

Dans la même logique il est donc possible :

a – de générer des simulations numériques jouant le rôle de modèle, les développements multitableaux autorisant maintenant un plus grand nombre de dimension du type

$$x_{ijk} = \sum_{l=1,r} \alpha_{il}\beta_{jl}\gamma_{kl} + r_{ijk}$$ [12]

b – d'étudier les structures de résidus autour d'un modèle naïf (structure triviale) ou sophistiqué (structure fonctionnelle),

c – de générer des simulateurs numériques et modéliser ensuite les résidus comme par exemple

$$x_{ij} = \beta_j(\sum_{k=1,r_j} \alpha_k y_{ik}) + r_{ij}$$

$$\text{et } r_{ij} = \sum_{l=1,m} \gamma_{il}\delta_{jl} + s_{ij}$$

dans l'analyse en composantes principales sur variables instrumentales conçue comme régression multiple simultanée[11]. C'est le cas dans les analyses inter et intra classes[13] , les analyses de nuages projetés en particulier des analyses des correspondances canonique, partielle et interne[14] .

Au delà des programmes classiques, devenus arbres cachant la forêt, l'analyse des données est donc un ensemble méthodologique diversifié dans ces pratiques et ses interactions avec la modélisation tendent à s'accroître pour plusieurs raisons. Nous en citerons trois.

Manipulation des variables qualitatives

On peut toujours passer d'un enregistrement quantitatif à une variable qualitative en groupant les observations.

On peut assurer la transformation inverse en codant numériquement les modalités. L'analyse des données obtient des codes optimaux en maximisant sous contraintes des formes quadratiques[15]. La figure 6 montre que les codes peuvent être "meilleurs" que les données elles-mêmes en ce sens que le passage d'une mesure quantitative à une mesure qualitative puis d'une mesure qualitative à un code numérique restitue une information plus signifiante expérimentalement de l'objet étudié. L'analyse des correspondances multiples[16] et ses dérivées reposent sur ce principe.

9 Caussinus H. et Falguerolles A. de, Tableaux carrés : modélisation et méthodes factorielles, *Rev. Statistique Appliquée*, 35, 3, 35-52, 1987.

10 Escofier B., Analyse factorielle en référence à un modèle. Application à l'analyse d'un tableau d'échanges. *Rev. Statistique Appliquée*, 32, 4, 25-36, 1984. Escofier B., Analyse des différences entre plusieurs tableaux de fréquence. *Les Cahiers de l'Analyse des Données*, VIII, 4, 491-499, 1983.

11 Cf. les synthèses de R. Sabatier (Approximation d'un tableau de données. Application à la reconstitution des paléoclimats. Thèse de 3° cycle, Montpellier, 184 p., 1983. Méthodes factorielles en analyse des données : approximations et prise en compte de variables concomitantes. Thèse d'Etat, Montpellier, 242 p., 1987.

12 Cf la synthèse de Kroonenberg P.M., Three-mode Principal Component analysis, DSWO Press, Leiden, Pays-Bas, 398 p., 1983.

13 Benzecri J.P., Analyse de l'inertie intra classe par l'analyse d'un tableau de correspondances. *Les Cahiers d'analyse des Données*, 8, 351-358, 1983. Foucart T., Sur les suites de tables de contingence indexées par le temps. *Statistique et analyse des données*, 2, 67-84, 1978.

14 Ter Braak C.J.F., Canonical correspondence analysis : a new eigen vector method for multivariate direct gradient analysis. *Ecology*, 67, 1167-1179, 1986. Cazes P., Chessel D. et Doledec S., L'analyse des correspondances internes : son usage en hydrobiologie. *Rev. Statistique Appliquée*, 36, 1, 39-54 Ter Braak (C.J.F.) Unimodal models to relate species to environments, Agricultural Mathematics group Box 100. NL-6700 Al Wageningen, The Netherlands, 1987, en particulier p. 83-89.

15 Par exemple les codes des lignes et des colonnes d'une table de contingence maximisant la corrélation sont obtenus par la première coordonnée en analyse des correspondances : Williams E.J., Use of scores for the analysis of association in contingency tables. *Biometrika*, 39, 274-289, 1952.

16 Synthèse bibliographique et théorique dans Tenenhaus M. et Young F.W., An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical data. *Psychometrika*, 50, 1, 91-119, 1985.

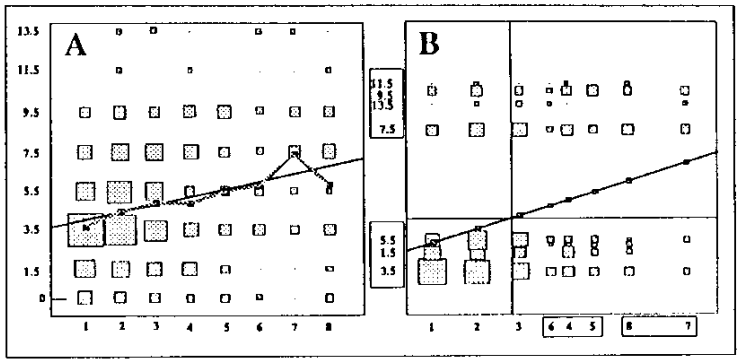

Figure 6 : Pour 350 chattes on connait l' âge (1 an à 8 ans et plus) et le nombre de chatons produits dans l'année 1 ou 2 (1.5), 3 ou 4 (3.5), ..., 13 ou 14 (13.5). La liaison ente les 2 variables est exprimée par le nuage bivarié élémentaire (A : surface des carrés proportionnelle au nombre d'observations) , la courbe de régression (rapport de corrélation de 0.089) et la droite de régression (carré de corrélation de 0.073). On y "voit" une liaison linéaire comme dans un cas d'école de statistique descriptive. L'AFC de la table de contingence qui ne tient aucun compte de l'origine quantitative des données recode les lignes et les colonnes du tableau en maximisant la corrélation observée et , pour ce faire, rend linéaire la liaison entre les codes lignes et colonnes : HIRSCHFELD (H.O.), A connection between correlation and contingency. *Proc. Camb. Ph. Soc.*, 31, 520-525. Les modalités associées à l'âge sont ordonnées sur 5 niveaux (1, 2, 3, 4 à 6, 7 et plus) et celles qui concernent la fécondité sont groupées en deux classes (1 à 6, 7 à 14). Le rapport de corrélation, le carré de corrélation et la première valeur propre valent 0.114. La liaison se fait d'abord par le nombre de portées. Le codage d'objets par des méthodes euclidiennes est un des moyens les plus sûrs de prendre en compte des liaisons non linéaires. Données de Legay (J.M.) et PONTIER (D.), Relation âge-fécondité dans les populations de chats domestiques, *Felis catus*. *Mammalia*, 49, 3, 395-402.

Modèles qualitatifs

Ce qui caractérise les structures des systèmes écologiques (dispersions dans l'espace, associations entre espèces, relations à l'environnement) est la diversité des possibles, égale à la diversité des formes des organismes vivants. Nous avons donc besoin de liberté dans l'émergence des modèles censés décrire ces structures, liberté au sens qu'une image inattendue, c'est-à-dire

n'appartenant à une classe préétablie, puisse prendre place[17] .

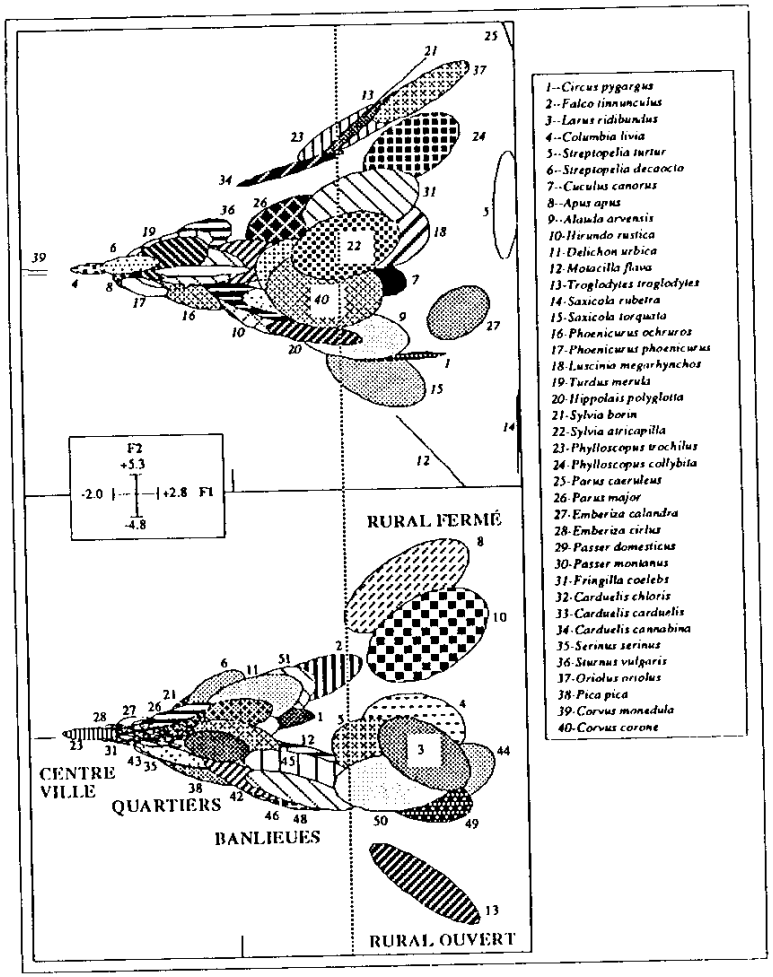

Figure 7 : exemple de modèle qualitatif d'une structure écologique (écotone). Données de F. Tatibouet et J. Broyer : Etude des peuplements d'oiseaux nicheurs de la zone urbaine de Lyon. Rapport final du Contrat 237-01-78-00314, Ministère de l'Environnement (Ecologie Urbaine), 106-156, 1980.

17 Cet élément est une des racines de la biométrie au sens de J.M. Legay (Pour une biométrie, *Statistique et analyse des données*, 1, 2, 5-11, 1976. Sur les relations biométrie-écologie, *Bull. Ecol.*, 15, 2, 117-119, Quelques réflexions à propos d'écologie : défense de l'indisciplinarité, *Œcol. Gener.*, 7, 4, 391-398) comme de l'analyse des données au sens J.P. Benzecri (L'analyse des données, T. II, L'analyse des correspondances, Dunod, Paris, 619 p., 1973).

La méthodologie euclidienne est un moyen d'expression de tels modèles discursifs; la figure 7 montre une exemple de cette démarche. 40 espèces d'Oiseaux dont la liste figure sur la droite de la figure sont présentes dans un ensemble de 51 relevés par point d'écoute. Ces points sont systématiquement répartis sur un transect débutant en milieu rural de la région lyonnaise, traversant des communes est de la communauté urbaine, atteignant le centre ville de Villeurbanne, pour se prolonger jusqu'àux zones rurales très ouvertes situées au sud de l'agglomération. Chacune des correspondances (couples à valeur non nulle) espèce-relevé est positionnée sur un plan factoriel par la somme des coordonnées factorielles (AFC) ligne et colonne correspondantes, la variabilité totale étant ramenée à l'unité. Cette pratique très peu utilisée est associée à l'AFC comme analyse canonique (Benzecri J.P., 1973, op. cit.). En haut de la figure chaque espèce est représentée par une ellipse d'inertie, image de son amplitude d'habitat. Chaque relevé est traité de même (en bas) ce qui exprime la diversité des taxons qui s'y trouvent. Les centres des ellipses forment tout simplement les cartes factorielles classiques. L'axe 1 (horizontal) est le gradient urbain-rural, l'axe 2 est le gradient ouvert-fermé, facteur écologique premier de la répartition des Oiseaux. La comparaison des deux cartes construites à partir d'une même logique exprime le modèle difficilement quantifiable d'*écotone* : il existe deux types de relevés (urbain/rural) et une frontière marquée entre groupes de relevés. Cette interface est le point moyen de dispersion d'un nombre élevé d'espèces et c'est dans la zone de contact qu'on trouve la diversité la plus grande.

Endo-modèles

La capacité de construction d'images numériques de référence avec un minimum d'a priori devient franchement étonnante quand elle précède le traitement des observations proprement dites mais ne prend en compte que l'objectif affirmé. On doit à L. Lebart[18] d'avoir introduit les graphes de voisinage en analyse des données. La diagonalisation d'un endomorphisme symétrique (exemple dans la figure 8) engendre des codes de référence (vecteurs propres) ne dépendant que de la relation proposée. Ces codes peuvent servir de variables explicatives dans une procédure de régression multiple . On pourrait les appeler des endo-modèles comme liés à un endomorphisme et issus d'un procédé endogène.

18 Lebart L., Analyse statistique de la contiguïté, *Pub. Inst. Stat.*, 18, 81-112, 1969. Correspondence analysis of graph structures, Bulletin technique du CESIA, 2, 1-2, CESIA, Paris, 1984.

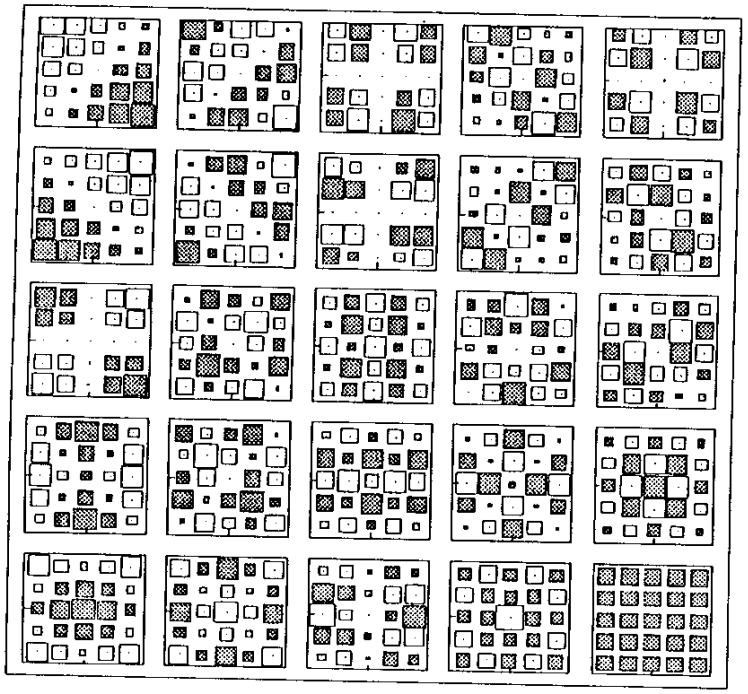

Figure 8 : codes numériques associés à un graphe de voisinage. 25 points forment un échiquier 5-5. Deux points sont voisins si et seulement si ils se suivent sur la même ligne ou la même colonne (relation de la tour). M est la matrice 25-25 associée (m(i,j) = 1 si i est voisin de j, 0 sinon). Dn est la matrice diagonale 25-25 de la pondération uniforme. Dn* est la matrice diagonale contenant la somme des poids des voisins. Dn*-MDn est la matrice d'un opérateur Dn-symétrique qui donne une base Dn-orthonormée de vecteurs propres. Les composantes de chacun d'entre eux sont représentées sur une carte par carrés (le 25ème est constant). Ils forment une liste finie de figures de référence dans la logique des polynômes orthogonaux ou des séries de Fourier.

La figure 9 explicite cette démarche d'endo-modélisation. On y a affaire à un cas typique d'examen préliminaire des données qui conduira à la recherche d'une classe de modèles du type MARMA, autorégressif (de pas 1 à autocorrélation négative, alternance), à moyenne mobile (croissance), composante exogène (climatologie) et multidimensionnel (235 courbes de production sont à modéliser simultanément).

En conclusion l'analyse des données peut *orienter* le choix d'un modèle par une première approche sans a priori, *modifier* ce choix en exhibant des structures résiduelles, le *valider* ou le

compléter par des observations d'une autre nature. Ses pratiques sont d'une plasticité étendue : les plus connues comportent des choix implicites qui ne sont pas toujours les meilleurs. Sa logique interne proprement algébrique permet de construire une grande diversité de variantes dont il faudra :

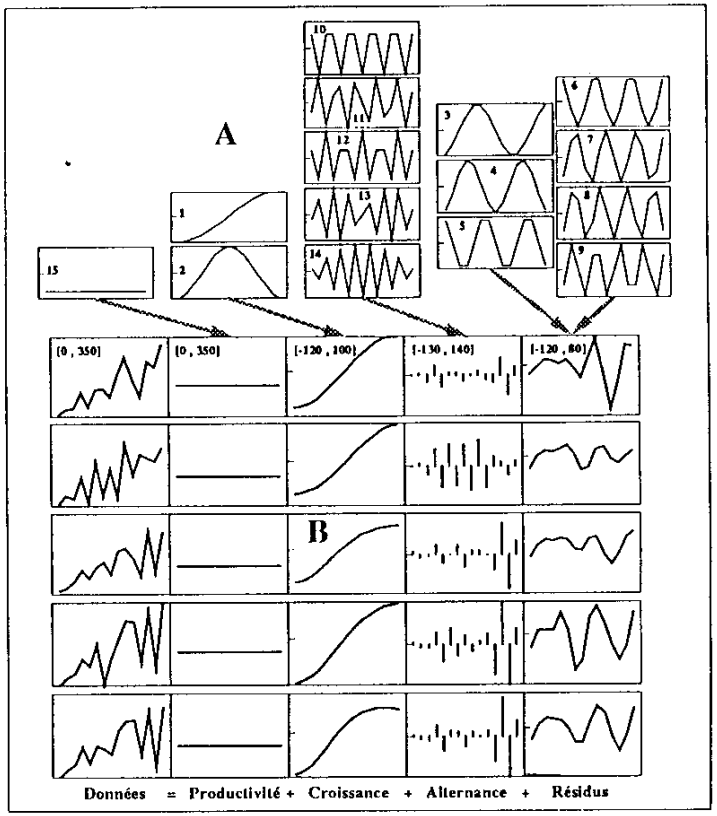

Figure 9 : Exemple d'endo-modélisation. 15 points sont alignés régulièrement sur un axe. Deux points successifs sont voisins. La diagonalisation de l'endomorphisme D_{15^-} symétrique associé à la matrice de voisinage donnent 15 courbes de référence établies avant tout usage de données (A : bornes de variation $[-\sqrt{2}, +\sqrt{2}]$). En B on a 5 courbes observées qui sont la production (kg) en fruits d'un mandarinier relevée durant 15 années consécutives (Données de la station INRA de Corse. D. Tisne-Agostini, Description par l'analyse en composantes principales de l'évolution de la productivité du Clémentinier en association avec 12 types de porte-greffe, Rapport de DEA, laboratoire de Biométrie, Université Lyon 1, 1988). Dans la colonne 2 on extrait la moyenne, puis on teste par analyse de variance sur la base des vecteurs définis en A la signification (évidente) de la croissance (colonne 3) et de l'alternance (colonne 4). La structure des résidus (colonne 5) qu'on attendait aléatoire invite à faire l'hypothèse d'une composante exogène). La décomposition oriente vers une modélisation paramétrique complexe.

bientôt faire un dictionnaire[19]. Cette logique coordonne et identifie chaque cas particulier mais la vocation de l'ensemble est au service de la méthode expérimentale : en ce sens elle est un point d'inter-disciplinarité et peut-être quand l'innovation porte simultannément sur l'outil et l'objet qu'il aide à comprendre un point d'*indisciplinarité* au sens de J.M. Legay.

Soulignons alors, remarque concrète et largement partagée, que la notion de modèles-outils rend cruciale la question des *logiciels*, en particulier dans le champ graphique. Là encore nous manquons singulièrement de liberté créatrice. La plupart des illustrations de cette communication sont des produits des programmes GraphMu (dérivé du travail déjà cité de Y. Auda) et MacMul , qu'on obtiendra sur simple demande auprès du second auteur[20] .

19 Yoccoz N., Le rôle du modèle euclidien d'analyse des données en biologie évolutive. Thèse de doctorat, Université Lyon 1, 254 p., 1988.

20 Thioulouse J., Statistical analysis and graphical display of multivariate data on the Macintosh. *Comput. Applic. Biosci.*, 5, 4, 287-292, 1989.