Analyse de données : traitements visuels et mathématiques

Madeleine Brocard, Denise Pumain, Violette Rey

Analyse de données : traitements visuels et mathématiques

Madeleine Brocard, Mme Denise Pumain, Violette Rey

Résumé

Cet article compare quatre techniques d'analyse des données : analyse en composantes principales, analyse des correspondances, classification automatique et matrice ordonnable, appliquées aux départements de Roumanie, à la situation de la recherche scientifique dans les régions françaises, aux catégories socio-professionnelles de la population active dans les grandes villes françaises. La comparaison des résultats montre, sous certaines conditions, une assez grande stabilité des structures mises en évidence, mais aussi des divergences imputables aux caractères spécifiques de chaque procédure : une bonne connaissance de ceux-ci est donc indispensable. La qualité des résultats et la pertinence de leur interprétation dépendent davantage de cette mesure de la relativité que du choix de l'une plutôt que de l'autre.

Abstract

Data analysis : visual and mathematical methods. This article compares four analytical techniques, principal components analysis, analysis of correspondences, automatic classification, and the « orderable matrix », as applied to the departments of Romania, to the place of academic research in the regions of France, and the socio-professional categories of the working population in France's largest cities. The comparison of results shows, under conditions, that the structures considered are reasonably stable, but also that there are variations which can be ascribed to the particular characteristics of each procedure : a good knowledge of these is therefore indispensable. The quality of the results and the relevance of their interpretation depend rather more on this measurement of relativity than on the choice of one method rather than another.

L'Espace Géographique, n° 4, 1977, 247-260.
Doin, 8, place de l'Odéon, Paris-VI°

ANALYSE DE DONNÉES : TRAITEMENTS VISUELS ET MATHÉMATIQUES

Madeleine BROCARD, Denise PUMAIN et Violette REY *
Université de Rouen *Université de Paris I*

ANALYSE FACTORIELLE
ANALYSE SPATIALE
FRANCE (VILLES)
MATRICE ORDONNABLE
ROUMANIE
DONNÉES
(TRAITEMENT GRAPHIQUE DES)

RESUME. — Cet article compare quatre techniques d'analyse des données : analyse en composantes principales, analyse des correspondances, classification automatique et matrice ordonnable, appliquées aux départements de Roumanie, à la situation de la recherche scientifique dans les régions françaises, aux catégories socio-professionnelles de la population active dans les grandes villes françaises. La comparaison des résultats montre, sous certaines conditions, une assez grande stabilité des structures mises en évidence, mais aussi des divergences imputables aux caractères spécifiques de chaque procédure : une bonne connaissance de ceux-ci est donc indispensable. La qualité des résultats et la pertinence de leur interprétation dépendent davantage de cette mesure de la relativité que du choix de l'une plutôt que de l'autre.

FACTOR ANALYSIS
FRANCE (CITIES)
ORDERABLE MATRIX
ROMANIA
SPATIAL ANALYSIS
VISUAL DATA PROCESSING

ABSTRACT. — *Data analysis : visual and mathematical methods.* This article compares four analytical techniques, principal components analysis, analysis of correspondences, automatic classification, and the « orderable matrix », as applied to the departments of Romania, to the place of academic research in the regions of France, and the socio-professional categories of the working population in France's largest cities. The comparison of results shows, under conditions, that the structures considered are reasonably stable, but also that there are variations which can be ascribed to the particular characteristics of each procedure : a good knowledge of these is therefore indispensable. The quality of the results and the relevance of their interpretation depend rather more on this measurement of relativity than on the choice of one method rather than another.

Diverses techniques d'analyse des données ont fait l'objet de nombreuses utilisations en géographie. Il semble utile de tenter d'évaluer leurs avantages respectifs à partir d'une expérimentation sur des études de cas. L'objectif est de répondre à un certain nombre de questions souvent posées : Peut-on appliquer ces techniques indifféremment à tous les types de données avec les mêmes chances de succès ? Employées successivement sur un même ensemble de données, apportent-elles des enseignements redondants, complémentaires ou contradictoires ? Dans quelle mesure les résultats acquis à partir de données différentes par des techniques différentes sont-ils comparables ?

Quatre types de traitement des données font l'objet de cette comparaison : matrice visuelle, analyse factorielle des correspondances, analyse en composantes principales, classification ascendante hiérarchique. Ils ont porté sur trois tableaux de données différents, comportant tous une base spatiale : les catégories socio-professionnelles des villes françaises de plus de 50 000 habitants, la structure spatiale de la

* Nous remercions pour leur aide les assistants du Centre de Calcul de l'Université de Paris I: Mme Baran, MM. Cohen, Sastre et Vielajus.

Roumanie, les relations entre recherche et développement régional en France. Les résultats n'en sont pas parfaitement comparables car certaines variations n'ont pu être évitées : dans le mode de réduction des données, le nombre des traitements effectués, l'interprétation des résultats par des chercheurs différents (1). Mais chaque fois trois de ces traitements au moins ont été expérimentés sur chacun des trois objets. Et le propos était bien plus d'apprécier l'apport spécifique de chaque technique que la comparaison des résultats proprement dite.

Nous supposons déjà connus les différents types de traitement utilisés. Nous rappelons dans une première partie les caractéristiques propres à chacun, et leurs conséquences pour l'interprétation des résultats. Une deuxième partie montre, par les trois études de cas, les principaux résultats d'une confrontation plus détaillée.

I. PROCÉDURE COMPARÉE DE CHAQUE TECHNIQUE.

Toutes ces techniques se situent dans la même perspective d'analyse des données. Elles supposent que l'on définisse une portion du réel, composée d'un ensemble d'individus (unités géographiques, ou éléments de même nature) et de leurs attributs. Ces attributs (ou variables) peuvent faire l'objet d'une mesure. qualitative ou quantitative, et sont choisis de manière à décrire ou approcher au mieux un ou plusieurs phénomènes, caractéristiques et distinctifs des individus. Les techniques d'analyse multivariée et la matrice ordonnable se proposent de tirer du tableau de données ainsi constitué une structure descriptive, qui ressort de l'image globale des ressemblances et des différences entre les distributions des variables et les profils des individus : 1) elles font apparaître des combinaisons de variables, parfois décrites comme des dimensions fondamentales du phénomène étudié (ou relevant de dimensions latentes), et elles rassemblent les individus que ces combinaisons caractérisent; 2) elles ordonnent variables et individus à l'intérieur de ces groupes; 3) elles mettent en évidence et hiérarchisent les oppositions entre groupes de variables et d'individus.

Cette démarche descriptive, à finalité typologique, s'impose comme cadre *un modèle linéaire :* les liaisons mises en évidence entre variables sont de forme linéaire, les combinaisons de variables sont de forme additive. Les discontinuités, les effets multiplicatifs, les boucles de rétroaction, etc., ne sont prises en compte par aucun de ces types d'analyse. Au-delà de la simple communication des résultats, c'est alors une des tâches de l'interprétation que de réintroduire ces différents effets de pondération.

(1) Les résultats détaillés de chacun de ces traitements font l'objet d'articles parus ou à paraître. Cf. bibliographie *in fine*.

En dépit de ces fondements communs quant au cadre théorique d'analyse et au but poursuivi, les techniques étudiées diffèrent sensiblement par la méthode de traitement mise en œuvre. Ce qui les distingue, ce sont les critères de ressemblance qu'elles utilisent, ainsi que la procédure de recherche des corrélations. Nous étudierons ces différences à travers les étapes successives du traitement des données.

1. *La nature des données.*

Notre comparaison se situe en aval des opérations de collecte des données et de constitution de l'information à traiter. Toutefois, en dehors des considérations méthodologiques qui seront évoquées plus loin, le choix d'une méthode de traitement n'est pas totalement indépendant de la nature de ces données, et il convient de souligner quelques limites techniques :

a) la dimension du tableau étudié est limitée en pratique à 100×100 pour la matrice ordonnable; elle peut aussi prohiber l'emploi direct d'une classification automatique, dont le coût s'élève très vite avec le nombre d'individus à classer;

b) la matrice ordonnable et l'analyse en composantes principales (abrégée ici en ACP) conviennent au traitement de variables quantitatives quelconques, même mesurées dans des unités très disparates; en revanche, l'analyse factorielle des correspondances (abrégée en AFC) doit être utilisée avec précaution sur des tableaux autres que les tableaux de contingence ou de dépendance : il importe de s'assurer que l'hétérogénéité des unités de mesure n'entraîne pas de pondérations aberrantes entre les variables;

c) la matrice ordonnable et l'AFC permettent, sous certaines conditions de codage, de traiter des variables qualitatives et même un mélange de variables qualitatives et quantitatives; l'ACP ne permet en principe de traiter que des variables quantitatives, mais elle peut traiter des données ordinales (à partir d'une matrice de corrélation de rangs) et même des matrices de présence-absence.

2. *La transformation des données préalable au traitement.*

Une première transformation consiste à élaborer, à partir du tableau de données, une matrice qui permette de classer variables et individus. Elle vise principalement à égaliser le poids des variables, à réduire la diversité des unités, des ordres de grandeur et des dispersions de celles-ci. Certaines techniques atténuent ou éliminent totalement aussi les inégalités de taille entre individus. De cette première transformation des données dépend en partie ce qui sera utilisé comme critère de ressemblance (ou comme critère discriminant), pour regrouper (ou séparer), variables et individus.

a. *Le choix des classes et des paliers pour la matrice ordonnable.*

L'élaboration de la matrice ordonnable à partir du tableau de données consiste en une certaine standardisation des variables, qui élimine l'effet des différences d'unités de mesure, les inégalités des ordres de grandeur et qui atténue, en les supprimant parfois totalement, les effets de l'inégale dispersion des variables. Le domaine de variation de chacune des variables est en effet ramené à 11 classes (parfois 6), visualisées par 11 paliers gradués du blanc au noir, souvent centrés sur la moyenne de la série. Si les classes ainsi constituées sont de même étendue, les variables statistiquement les plus dispersées forment les lignes qui contiennent le plus de blanc et de noir et sont donc les plus discriminantes. Si les classes comportent des effectifs égaux (possibilité plus rarement utilisée, semble-t-il), toutes les variables ont la même importance visuelle et le classement s'opère d'après le rang des individus sur ces variables.

D'une façon générale, la matrice autorise une grande souplesse dans le choix des classes et des paliers. Elle permet de rester fidèle à l'hétérogénéité de l'information, par exemple en excluant les valeurs extrêmes avant le découpage en classes de la série, ou en utilisant pour ce découpage les seuils « naturels » de la distribution, tels qu'ils apparaissent dans l'histogramme des fréquences. Cela peut entraîner toutefois une plus grande difficulté au moment de l'interprétation des résultats, ou du moins une grande complexité, qui risque de n'être que le reflet de la « subjectivité » introduite. A la fois avantage et inconvénient de la matrice ordonnable sur les techniques plus automatisées, cette possibilité de mieux coller au réel, à la connaissance acquise, s'oppose à la simplification nécessaire, à la répétitivité et à la modélisation de l'expérience.

b. *La transformation des données dans l'analyse en composantes principales.*

Le plus souvent, l'ACP opère une standardisation (centrage et réduction) *des variables :* cela leur assure même ordre de grandeur et même dispersion, même « poids » dans l'analyse. L'analyse est donc sensible aux inégalités de taille ou de valeur entre *les individus :* elle isole souvent un individu extrême ayant de très fortes valeurs sur toutes les variables, et que l'on a parfois intérêt à extraire des analyses ultérieures; souvent aussi, lorsque les individus sont rangés à peu près dans le même ordre par toutes les variables, le premier facteur n'est que le reflet de cet ordre : il se produit ce qu'on appelle un « effet de taille ». L'importance de cet effet de taille est parfois telle qu'il peut obliger à une nouvelle transformation des données, égalisant les poids des individus (par exemple par des transformations en pourcentage), afin de faire apparaître d'autres relations entre variables. Pour la même raison, l'analyse est sensible aux « redondances », aux très fortes corrélations entre variables qui sont exprimées par les premiers facteurs et qui, si l'on n'y prend garde, risquent de masquer le reste de l'information.

c. *La métrique du χ^2 en analyse des correspondances.*

L'AFC se distingue des techniques précédentes en ce qu'elle atténue de façon symétrique les inégalités de poids des variables et des individus : l'utilisation de la métrique du χ^2 pour calculer les distances entre variables et entre individus revient à comparer des profils, aussi bien sur les lignes du tableau que sur les colonnes. Le poids (ordre de grandeur) des variables et des individus intervient dans le calcul de l'inertie et des contributions, mais il ne joue pas dans la proximité établie entre les variables et entre les individus. La position de ceux-ci sur les axes factoriels dépend donc surtout de *l'écart* entre le *profil* de la ligne (variable) ou de la colonne (individu), et le *profil* de la ligne ou de la colonne marginale (ligne ou colonne « total » du tableau des données). L'AFC est ainsi moins sensible que l'ACP aux inégalités de taille entre individus, mais n'échappe pas à ce genre de difficulté (cf. ci-après II, 2 b et c).

Elle est par contre sensible aux variables très dispersées statistiquement, aux individus dont le profil comporte des écarts importants au profil moyen.

d. *Le choix d'une distance en classification automatique.*

Ces techniques élaborent à partir du tableau de données une matrice de distances (ou d'indices de similarité) entre individus. Selon la distance choisie, les inégalités de poids des variables et des individus dans la classification seront prises en compte ou non, atténuées ou renforcées : par exemple, une distance euclidienne fait jouer les inégalités de taille entre individus et donne plus de poids aux variables s'exprimant par des valeurs numériques élevées, tandis qu'une distance du χ^2 égalise le poids des variables et des individus, le calcul de la distance s'effectuant d'après des différences de profils. Les résultats diffèrent sensiblement selon la distance utilisée; aussi convient-il d'ajuster le choix de celle-ci aux fondements retenus pour la typologie.

3. *Le traitement et la présentation des résultats.*

a. *Simultanéité et symétrie du traitement des variables et des individus.*

La matrice ordonnable et les analyses factorielles donnent des résultats qui portent sur les deux ensembles étudiés : l'image de la matrice se lit en

ligne aussi bien qu'en colonne; les analyses fournissent des listes de saturations des variables et de scores ou poids locaux des individus, ainsi que leur représentation sur les axes factoriels. Seules les classifications automatiques ne classent que l'un ou l'autre ensemble à la fois, certains programmes comportant toutefois le calcul de la contribution des variables à la formation des classes d'individus.

La procédure de traitement et la présentation des résultats sont les mêmes pour les variables et pour les individus dans le cas de l'analyse des correspondances et de la matrice ordonnable. L'analyse en composantes principales introduit une dissymétrie dans le traitement, qui s'effectue dans l'espace des variables seulement — sauf dans le cas des analyses R-mode, peu utilisées —, et dans la présentation des résultats (corrélation entre variables et facteurs d'une part et coordonnées des projections des individus sur les facteurs, d'autre part).

b. *Constitution et hiérarchie des systèmes matriciels et des axes factoriels.*

Il est possible d'assimiler, aux facteurs issus des analyses factorielles, ce que les utilisateurs de la matrice ordonnable appellent des « systèmes ». Il s'agit en effet de faisceaux de variables qui induisent sensiblement le même ordre — ou un ordre totalement inverse — entre les individus. Ces « systèmes » ou facteurs sont obtenus par des procédures différentes; ils ne sont pas hiérarchisés d'après les mêmes critères, et la combinaison des variables qu'ils rassemblent n'a pas tout à fait la même signification selon le type de traitement.

Dans la matrice ordonnable, la manipulation conduit à rapprocher les lignes au profil semblable, et à éloigner celles qui ont un profil opposé. Ces ressemblances ou oppositions de profil correspondent — plus ou moins étant donné le découpage des séries en paliers — à de fortes corrélations, positives ou négatives, entre les variables. Le classement d'ensemble privilégiée, aux extrémités de la matrice, les plus grandes quantités de noir et de blanc. Ainsi, lorsque la matrice est diagonalisée, peut-on considérer que les extrêmes de la diagonale représentent un premier « facteur » (faisceau de variables à fortes corrélations positives et négatives) et les oppositions intermédiaires un deuxième facteur, puis un troisième, etc. On hiérarchise donc différents groupes de variables en fonction de la plus ou moins grande opposition de profils des variables qui les composent. Il peut être toutefois difficile d'établir avec précision cette hiérarchie, et les « facteurs » ainsi hiérarchisés, contrairement à ceux des analyses factorielles, ne sont pas orthogonaux. Dans le cas où la matrice ne peut être totalement diagonalisée (lorsqu'un groupe de variables introduit un classement des individus totalement différent de celui du premier système), on obtient un second « système », dont les variables ne sont guère corrélées à celles du premier. C'est ce qui fonde l'analogie entre « système » et facteur d'analyse factorielle. Il est alors difficile de hiérarchiser les « systèmes » d'après la matrice, autrement que par le nombre des variables qui les constituent, et peut-être par l'importance des oppositions de noir et de blanc qui les caractérisent (degré de dispersion des variables).

Dans l'analyse en composantes principales, les facteurs correspondent aux groupes de variables qui ont entre elles les corrélations les plus élevées, positives ou négatives. La hiérarchie des facteurs se fonde sur le nombre et l'importance de ces corrélations.

La différence n'est qu'apparente entre la transition d'un groupe de variables à l'autre, offerte par l'image matricielle, et l'orthogonalité des facteurs de l'ACP. En effet, les variables peuvent présenter des corrélations notables avec plusieurs facteurs et donc se trouver à la charnière de plusieurs groupes, tout comme dans la matrice. Or, le plus souvent, l'interprétation est fondée sur les variables et pas seulement sur les facteurs.

Dans l'analyse des correspondances, l'identification des facteurs pour eux-mêmes a encore moins d'importance. Ils permettent de mettre en évidence des écarts entre le profil des variables et le profil moyen, en opposant sur les parties négatives et positives de l'axe des profils symétriques par rapport à ce profil moyen. La hiérarchie des facteurs correspond à une hiérarchisation de ces spécificités par rapport au profil moyen. Elle reflète essentiellement l'ordre de la dispersion statistique des variables, et non plus l'ordre des corrélations comme en composantes principales.

En résumé, dans le traitement des données, la matrice ordonnable ressemble à l'ACP parce qu'elle rapproche les variables d'après leurs corrélations, et qu'elle forme des facteurs d'après le nombre et l'importance des corrélations fortes. Elle a en commun avec l'AFC la symétrie du traitement des variables et des individus. L'AFC se distingue par la procédure de regroupement des éléments et de hiérarchisation des facteurs, qui mettent en première évidence les écarts aux profils moyens.

c. *Les conséquences de la forme de présentation des résultats.*

La présentation des résultats, sous forme d'images matricielles globales d'une part, de listes de chiffres et de graphiques de plans factoriels d'autre part, n'influence guère le chercheur, qui est conduit de toute manière à un va-et-vient entre le dépouillement des faits analytiques et l'examen de la structure d'ensemble. Les conséquences de cette diversité de présentation apparaissent plutôt lors de la communication des résultats, dans la forme du discours, et seront envisagées en conclusion.

II. RÉSULTATS COMPARÉS DE TROIS ÉTUDES DE CAS.

A travers les trois études citées et compte tenu des similitudes logiques des outils, nous pouvons établir la comparaison des résultats à trois niveaux successifs : niveau élémentaire des corrélations terme à terme entre variables, niveau des regroupements de variables, niveau des regroupements d'individus. Il en ressort une remarquable convergence des résultats d'ensemble.

1. *Le niveau des relations élémentaires entre variables.*

Dans la matrice graphique, la juxtaposition de deux lignes doit exprimer la plus grande ressemblance de profils entre deux variables, donc correspondre au coefficient de corrélation le plus élevé du tableau. Le tableau (fig. 1), établi pour comparer l'ordre des variables des catégories socio-professionnelles (CSP) des *villes françaises,* donné par la matrice avec leur coefficient de corrélation, montre qu'en général il en est bien ainsi (répartition diagonale sur le tableau) : les coupures qui isolent des *groupes* de variables sur la matrice se situent entre des CSP ayant des corrélations faibles entre elles (ex. : corrélation de 0,2 entre techniciens et cadres administratifs supérieurs).

Néanmoins, deux différences apparaissent : il y a des CSP plus corrélées avec celles d'un groupe voisin qu'avec celles du groupe auquel elles appartiennent (ex. : professions libérales avec le groupe e; armée, clergé et police avec le groupe d; employés de commerce avec le groupe d). Ainsi, à l'intérieur de chaque groupe graphique de variables, l'optimisation visuelle ne correspond pas toujours au regroupement qu'induiraient les plus fortes corrélations. Qu'en déduire ? Quel est le degré de fiabilité des rapprochements visuels ? Seront-ils répétés d'une manière identique par deux manipulateurs ? L'expression graphique de la ressemblance de deux variables est-elle moins valable que l'expression mathématique d'une corrélation ?

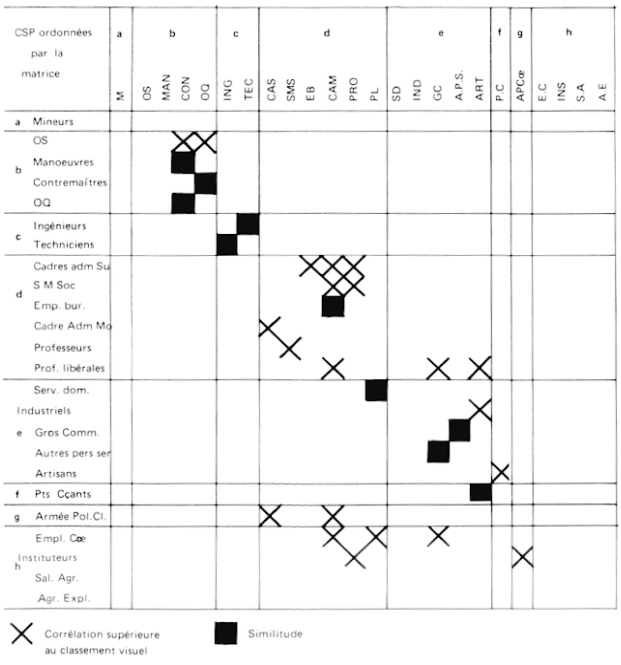


FIG. 1. - *Juxtaposition visuelle des variables et corrélations maximas : comparaison.*

h : non classés.

Deux éléments expliquent en partie ces différences : a) à l'intérieur de chaque groupe de variables, le chercheur, pendant l'interprétation, peut être amené à choisir pour les variables un ordre légèrement différent de l'ordre visuel strict, plus explicite pour son discours, sans que le contenu du groupe en soit modifié; de même, pour une variable de profil faiblement contrasté et difficile à classer, le chercheur peut opter pour un rangement dans un groupe en fonction seulement d'une portion de son profil; b) le découpage préalable de la distribution en 11 paliers introduit sans doute là une distorsion spécifique, non enregistrée par le coefficient mathématique.

Il reste qu'une manipulation visuelle très fine donne l'assurance d'une bonne mise en évidence des corrélations. C'est ce que montre la vérification faite sur la matrice Roumanie.

2. *Le niveau des regroupements de variables.*

La confrontation des regroupements de variables fournis par chaque méthode s'articule autour de deux points : celui de la hiérarchie entre les facteurs mathématiques par rapport à l'ordre de l'image; celui du degré de ressemblance entre le facteur mathématique et le groupe graphique de variables. Compte tenu de la convergence des corrélations constatée au niveau des variables élémentaires, il est normal que nous retrouvions des convergences approchées sur les regroupements de variables. Cependant, la comparaison ne peut plus s'effectuer d'une manière mécanique; elle passe par l'interprétation de contenu et nous oblige à entrer un peu dans chaque sujet d'étude.

a. *La recherche scientifique en France.*

Cet exemple a fait l'objet d'un double traitement, par l'analyse factorielle des correspondances et par la matrice ordonnable. L'objectif était de comprendre comment la recherche scientifique s'intégrait aux données socio-économiques, en analysant leur distribution dans l'espace. D'où le nombre d'indicateurs relatifs à la recherche (51 sur 123, largement redondants), les autres variables portant sur la population (quantité, densité, répartition, âge), l'enseignement

(étudiants, ingénieurs), l'industrie (structure), l'agriculture. D'après les deux traitements effectués, la répartition spatiale de la recherche est d'abord liée à celle de la population (poids de l'enseignement supérieur), puis de l'industrie, et il s'y ajoute pour la Région parisienne le rôle de la concentration administrative et politique (par interprétation, puisqu'aucune variable ne l'exprimait directement).

La fig. 2 permet de comparer les traitements effectués (l'AFC a été effectuée sur les valeurs brutes, après examen de la distribution des variables et comparaison de leurs ordres de grandeur).

L'axe 1 (82 %) oppose l'ensemble des indicateurs de recherche scientifique aux indicateurs agricoles; la plupart des données relatives à la population, l'enseignement, l'industrie sont groupées en milieu d'axe. L'axe 2 (6,4 %) est fortement caractérisé par les grands établissements industriels, l'axe 3 (4,5 %) par l'industrie lourde.

La manipulation visuelle aboutit à deux systèmes graphiques : pour obtenir un bon classement des données, on est en effet contraint de séparer l'ensemble des indicateurs en deux blocs, aboutissant à une typologie régionale différente. Le système 1 (93 indicateurs) reflète une partie de l'axe 1, le système 2 (30 indicateurs) regroupe le reste des indicateurs de l'axe 1, et les indicateurs discriminants de l'axe 2.

La matrice met toutefois en valeur un regroupement d'indicateurs qui n'apparaît pas dans l'analyse factorielle des correspondances : la recherche industrielle se rapproche en bloc de la plupart des indicateurs industriels, au lieu de s'assimiler aux indicateurs de recherche. Cela tient sans doute en partie à l'effet de taille introduit, dans l'AFC, par le poids excessif de Paris dans le domaine de la recherche. Or la Région parisienne a presque toujours été exclue du calcul de la moyenne lors de l'établissement des paliers de la matrice. Elle n'intervient donc dans la manipulation visuelle que comme une région forte, sans plus. Cela permet de mettre en valeur des regroupements secondaires : ceux-ci sont d'ailleurs très significatifs du point de vue de l'interprétation des résultats.

b. *La Roumanie.*

A partir d'une approche géographique globale de la Roumanie socialiste, l'objectif de l'analyse fut d'établir une *typologie* départementale et un découpage régional contemporain qui soient l'expression du modelage spatial apporté par le système socialiste. Trois questions étaient plus particulièrement posées dans le cadre de l'action des nouvelles structures collectives et étatisées : celle du rôle qu'il fallait encore attribuer aux campagnes et à l'agriculture dans une nation en pleine expansion industrielle, celle de l'influence exercée par l'industrialisation selon ses types, celle du poids des héritages légués par une longue appartenance à « l'Europe Centrale ».

Le traitement par matrice graphique a rangé les variables en 12 groupes (A à L), de netteté inégale, mais suffisamment ordonnés les uns par rapport aux autres pour qu'une seule image classée de l'ensemble des variables sur tous les individus soit possible (fig. 3 et 6); il s'agit là d'un cas de classement qui ne se produit pas toujours, comme l'a montré l'exemple précédent; ces 12 groupes se répartissent eux-mêmes en deux portions d'images (A à G et I à L).

L'apport de cette image pour l'interprétation fut très important; d'une part, les groupes de variables et leur organisation étaient fortement significatifs; d'autre part, la structure de l'image répondait à certaines de nos questions : les faits ruraux demeuraient primordiaux comme éléments de structuration de l'espace roumain (groupes A à G, système 1); les faits de production et les faits de structure et d'évolution restaient dissociés, particulièrement pour les activités urbaines et industrielles (I et J, système 2, B, système 1); l'impossibilité de classer la variable « minorités nationales », à elle seule groupe charnière (H) entre le système rural et le système urbain et industriel, confirmait le poids des héritages historiques.

Dans les analyses factorielles, le grand nombre de variables (75) par rapport aux individus (40 départements) et leur forte hétérogénéité statistique (valeurs absolues, pourcentages, indices, etc.) a donné dans toutes les analyses un grand nombre de facteurs sans forte hiérarchie entre eux, le premier ne regroupant jamais plus de 35 % de la variance totale. Dans le contexte méthodologique de ce travail, nous nous sommes limités à la confrontation de la matrice graphique avec les trois premiers facteurs de chaque analyse.

L'ACP, dont le traitement est le plus adapté à la nature hétérogène des données, a été faite deux fois : avec Bucarest-Ville (département introduisant un très fort effet-taille), et sans Bucarest. On obtient une fois 25, 21 et 10 %, et une autre fois 22, 20 et 10 % de la variance pour chacun des trois facteurs; mais ce qui était F1 dans l'une devient F2 dans l'autre et vice versa, tandis que F3 reste de contenu identique. Les variables à score élevé dans chacun des facteurs de l'analyse sans Bucarest ont été localisées sur l'image graphique (fig. 3) à la place qu'elles ont acquise par le traitement visuel : la concordance des résultats entre les deux types d'analyse est exceptionnelle. F1 prend la totalité des variables des groupes G et F, et la plupart de celles de E et A; il décrit dans le système rural les oppositions agricoles fortes (G $\neq$ A). F2 prend la totalité des variables des groupes I et J et une partie de celles de L, K, B, E; il décrit le système urbain et industriel avec son opposition interne (I $\neq$ L). Quant à F3, il a ses variables réparties en des groupes charnières de l'image, BCD, H, KL. D'ailleurs, si on les classe indépendamment des autres, ils ont une *forte opposition* (2) interne qui en

(2) Cf. notre article dans *l'Espace géographique,* 1973, n° 1, p. 45.

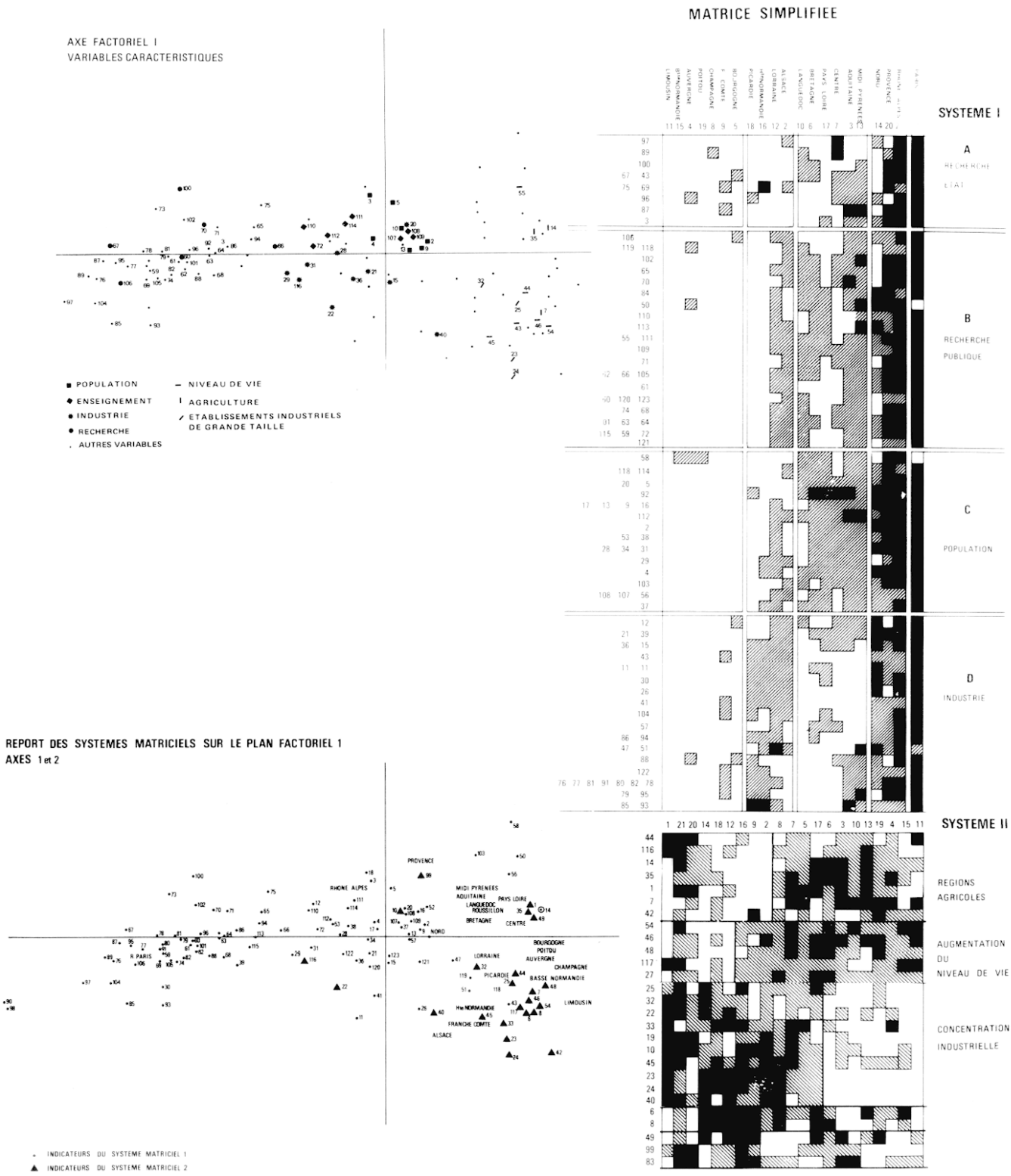

AXE FACTORIEL I
VARIABLES CARACTERISTIQUES
POPULATION
ENSEIGNEMENT
INDUSTRIE
RECHERCHE
AUTRES VARIABLES
NIVEAU DE VIE
AGRICULTURE
ETABLISSEMENTS INDUSTRIELS DE GRANDE TAILLE
MATRICE SIMPLIFIEE
SYSTEME I
A
RECHERCHE ETAT
B
RECHERCHE PUBLIQUE
C
POPULATION
D
INDUSTRIE
SYSTEME II
REGIONS AGRICOLES
AUGMENTATION DU NIVEAU DE VIE
CONCENTRATION INDUSTRIELLE
REPORT DES SYSTEMES MATRICIELS SUR LE PLAN FACTORIEL 1
AXES 1 et 2
PROVENCE
RHONE ALPES
MIDI PYRENEES
AQUITAINE
LANGUEDOC ROUSSILLON
PAYS LOIRE
BRETAGNE
CENTRE
NORD
BOURGOGNE
POITOU
AUVERGNE
LORRAINE
CHAMPAGNE
PICARDIE
BASSE NORMANDIE
LIMOUSIN
Hte NORMANDIE
FRANCHE COMTE
ALSACE
R PARIS
INDICATEURS DU SYSTEME MATRICIEL 1
INDICATEURS DU SYSTEME MATRICIEL 2

FIG. 2

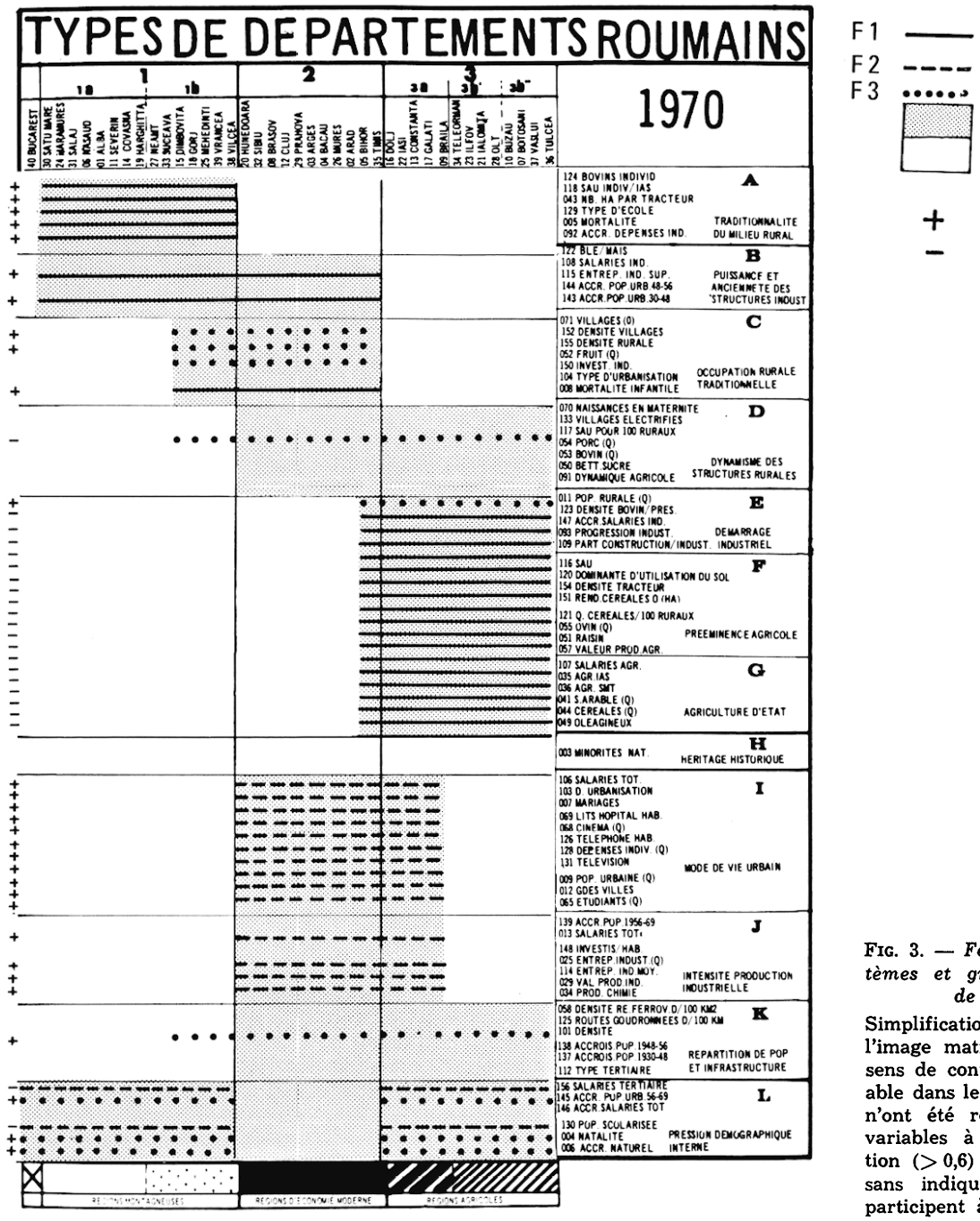


FIG. 3. — *Facteurs d'ACP : systèmes et groupes de variables de la matrice.*

Simplification en gris-blanc de l'image matricielle. — (+, −) : sens de contribution de la variable dans le facteur. Remarque : n'ont été représentées que les variables à très forte corrélation (> 0,6) avec un facteur et sans indiquer le cas où elles participent à plusieurs facteurs.

fait presque un sous-système. Leur situation médiane sur l'image globale, reflet des distributions moins différenciées des variables en question, exprime le rôle second de ces variables dans la différenciation générale, tout comme le 10 % de contribution à la variance totale l'exprime pour F3.

Dans l'ACP incluant Bucarest, F1 devient le système industriel urbain et F2 le système rural; les valeurs extrêmes, prises par Bucarest pour les variables industrielles et urbaines, renversent l'ordre des deux facteurs. Il est donc vain de mettre une hiérarchie entre ces deux facteurs de structuration de l'espace roumain au début des années 1970.

Une AFC avec le tableau de ces données brutes confirme les réserves théoriques déjà émises quant à l'emploi d'une telle analyse sur des données hétérogènes.

Ici, les données sont non seulement hétérogènes dans leur nature de mesure (têtes de bétail, habitants, etc.), mais aussi dans leurs unités (milliers, effectifs totaux, etc.), par commodité d'écriture. Or qu'en est-il advenu sur les résultats ? ô surprise, dans un premier passage, la variable étudiants devenait premier facteur à elle seule (F1 : 72 % de la variance avec une très forte contribution), par double effet de taille et de dispersion; ses valeurs, non réduites en

milliers, à la différence de la population, sont nulles sur 24 départements et atteignent le chiffre maximum de 65 000 : le premier axe était complètement « attiré » par cette mesure. Dans un second passage sans la variable étudiants, trois facteurs se dégagent avec 35, 25, et 12 % de la variance (soit 72 % au total, nettement plus qu'en ACP). Leur composition repose sur 20 variables seulement, dont 9 réapparaissent dans chacun des facteurs. Il ne faudrait pas en conclure que les 75 variables ont leur « résumé » le plus approché avec ces 9 variables, car elles correspondent encore aux variables qui ont à la fois les plus fortes valeurs absolues exprimées sur le tableau initial et la plus grande dispersion dans leur profil; mais rien de ce qui relève des variables de structure, exprimées en pourcentage, n'apparaît.

La signification des axes factoriels, est, de ce fait, sans grand rapport avec les résultats précédents. Pour F1, il s'agit de l'importance de la population rurale et à forte activité agricole traditionnelle, ne recevant pas d'investissements (variables à coordonnées positives éparpillées dans le système graphique I, variables de coordonnées négatives dans le système II). Pour F2, c'est la grosse agriculture céréalière moderne (variables de coordonnées positives, groupes G, F du système graphique I), qui s'oppose au niveau de vie élevé et en progression (variables de coordonnées négatives, groupes A et I). Avec F3, apparaissent des activités urbaines et industrielles à bon niveau de vie (coordonnées positives dans les groupes I et J), mais sans progression des dépenses individuelles (coordonnées négatives groupe A). Or, s'il y a une réalité incontestable dans ces oppositions fortes d'activités et de niveaux de vie, cet exemple montre surtout combien une manipulation inconsidérée de chiffres peut aboutir à une interprétation abusive; dans sa structuration, l'espace roumain ne relève pas d'oppositions aussi schématiques.

Afin de mieux utiliser l'AFC et d'en obtenir une analyse structurale plus pertinente, nous avons « adapté » l'information à l'outil et transformé le tableau des données brutes en un tableau des données qualitatives ordonnées et homogènes selon trois classes (valeur forte, moyenne ou faible de l'individu dans chaque variable). Cependant, cette réduction en trois classes rend l'analyse ultérieure très dépendante de la distribution choisie pour les départements, soit en classes d'égale étendue, soit en classes d'effectifs égaux, comme le montre le tableau suivant :

Variable		Nombre de départements en classe		
		forte	moyenne	faible
à classe d'amplitude égale	Nombre de villages	9	17	14
	% salariés	10	21	9
à classe d'effectifs égaux		13 (± 1)	13 (± 1)	13 (± 1)

La distribution en classes d'étendue égale individualise trop les départements aux valeurs extrêmes aux dépens d'une masse de « non classés » dont la diversité réelle compromet toute interprétation d'ensemble. C'est pourquoi nous n'en retiendrons pas ici le commentaire; nous nous limiterons aux résultats de *l'AFC faite avec des classes d'effectifs égaux*. Chaque classe ayant été codée en valeur binaire, toutes les variables ont le même poids dans l'analyse, ce qui la rend particulièrement comparable aux traitements par l'ACP et la matrice ordonnable.

Avec ces données transformées, la répartition des trois premiers facteurs redevient proche de celle obtenue par l'ACP, aussi bien par les pourcentages de variance absorbée que par leur signification. F1 (12 %) oppose une structure industrielle et urbaine évoluée avec agriculture individuelle (coordonnées positives, groupes B, A, H) à une forte agriculture d'Etat avec démarrage industriel (coordonnées négatives, groupes F, G). F2 (11 %) caractérise un mode de vie urbain et ses équipements (coordonnées positives, groupes I, J). F3 (7 %) oppose les campagnes peu peuplées, mais à forte agriculture d'Etat (coordonnées positives, groupes F, G) à celles très peuplées, mais à résultats agricoles moyens en coopératives (variables négatives, groupes C, D). Ainsi, malgré la simplification des données en trois classes, les effets de profil des variables l'emportent sur leurs effets de taille et mettent bien en valeur les faits de structure, d'héritage comme de dynamisme; en F1, c'est « la puissance et l'ancienneté industrielle » (groupe B) qui l'emporte sur « l'intensité de la production » (groupe J).

3. *Le niveau des regroupements d'individus : typologies spatiales.*

a. *Matrice graphique et classification hiérarchique automatique.*

Dans le cas des villes françaises, les résultats sont sensiblement différents (fig. 4) : classes hiérarchiques et groupes de villes ne se correspondent guère d'une analyse à l'autre, sauf pour le groupe I de la classification (villes ouvrières) et les groupes 1 - 2 de la matrice; il s'agit à la fois des catégories socio-professionnelles les plus discriminantes, ayant des corrélations négatives avec pratiquement toutes les autres, et des villes dont le profil est le plus éloigné du profil moyen. Hormis cet ensemble, il n'y a plus guère correspondance entre groupes dans les deux analyses. Une aussi faible stabilité des typologies vis-à-vis des techniques qui prétendent toutes deux regrouper des profils d'après leurs ressemblances poserait la question de la validité de telles typologies ou du moins de certaines d'entre elles, si l'exemple sur la recherche scientifique n'apportait pas un résultat plus conforme. Il est possible que, dans la typologie comparée des villes, on retrouve trace des incertitudes des corrélations visuelles constatées au tout début.

Classification hiérarchique / Matrice ordonnable	Groupe I			Groupe II			Groupe III				Groupe IV
	a	b	c	d	e	f	g	h	i	j	k
1		DOUAI BRUAY	LENS FORBACH								
2 a	MONTLUÇON ROANNE St ETIENNE ROUEN	BOULOGNE TROYES St QUENTIN LILLE	BETHUNE ARMENTIERES CALAIS								
2 b	VALENCIENNES DENAIN St CHAMOND MAUBEUGE DUNKERQUE LE HAVRE	MANTES CREIL BELFORT	LONGWY HAGONDANGE MONTBELIARD THIONVILLE	St NAZAIRE							
2 c		MULHOUSE ANNECY		CHERBOURG CHALON/S BOURGES					BREST LORIENT		
3 a		REIMS		VALENCE NANTES NEVERS CHARTRES LA ROCHELLE ANGERS LIMOGES	BESANÇON MELUN ARRAS AMIENS CHARLEVILLE COLMAR CHALON/MARNE CLERMONT FD	LYON		TARBES St BRIEUC			
3 b				ANGOULEME CHATEAUROUX TOURS ORLEANS LE MANS	METZ CAEN			CHAMBERY BORDEAUX PERIGUEUX ALBI MARSEILLE AVIGNON SETE PAU	TOULON		
4 a						GRENOBLE	AIX EN PROV. MONTPELLIER TOULOUSE				PARIS
4 b					NANCY STRASBOURG DIJON		RENNES POITIERS	NIMES			
5 a				NIORT				QUIMPER AGEN PERPIGNAN			
5 b								BEZIERS BAYONNE		VIVHY NICE MENTON CANNES	

FIG. 4. — *Comparaison des groupes de villes issus de la matrice ordonnable et d'une classification automatique.*

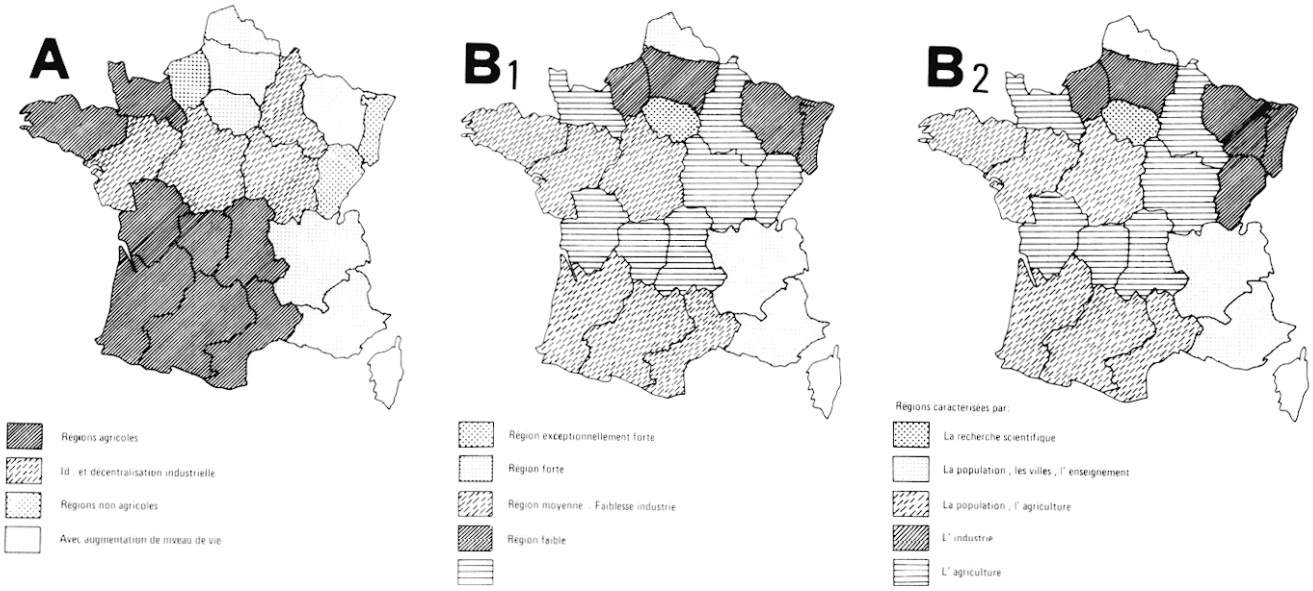
Dans le cas de la recherche scientifique, la typologie régionale obtenue par la classification hiérarchique arborescente est presque semblable à celle qui ressort du traitement visuel (fig. 5) : les groupes A, B, C sont identiques. Par contre, le groupe D est scindé en deux par la matrice, et l'ordre des régions diffère partiellement.

Cette différence tient à la nature de chaque type de traitement (visuel ou mathématique). Il s'agit de régions qui se ressemblent deux à deux pour une partie seulement de leur profil : dans le traitement visuel, le chercheur est obligé de choisir entre différents regroupements possibles. C'est ainsi que la Franche-Comté, en tant que région industrielle, ressemble à la Picardie. Mais, sur le reste du profil, donc la majorité des indicateurs, elle apparaît aussi « faible » que l'Auvergne ou le Poitou, régions avec lesquelles elle a été finalement classée. Cela est parfaitement visible pour le lecteur.

Il est plus difficile de trouver les raisons des regroupements dans la classification ascendante hiérarchique, qui donne un résultat sans qu'on sache ce qui y a contribué dans l'information initiale.

b. *Matrice graphique et analyse factorielle.*

Dans l'exemple de *la Roumanie,* compte tenu des similitudes de résultats de corrélation, les typologies spatiales sont voisines d'une analyse à l'autre. Nous utiliserons la carte typologique obtenue par la matrice graphique comme base de comparaison; seules, les différences spatiales introduites par les autres analyses sont figurées ici (fig. 6). La matrice graphique a fourni, Bucarest mis à part, trois types distincts de départements : le type rural montagnard, avec le sous-type de piémont externe; le type industriel urbain, dans lequel trois sous-types mineurs auraient pu être distingués; le type agricole des grandes plaines avec le sous-type à pôle de développement urbain; soit cinq catégories spatiales différentes. Aucune classification hiérarchique n'ayant été faite sur l'exemple roumain, nous avons classé les départements sur des tableaux croisés en fonction des trois premiers axes factoriels : sur 9 types (décomposables en 18 sous-types) théoriquement possibles, 6 ou 7 seulement se sont différenciés.



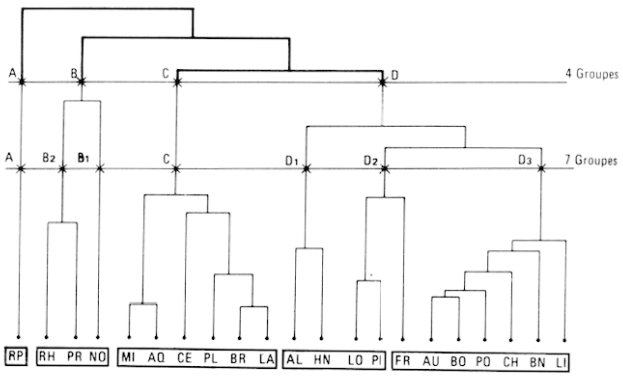


FIG. 5. — *Comparaison des typologies régionales.*
A. Classification ascendante hiérarchique. — B. Matrice visuelle (B_1 : système I, B_2 : système II). — C. Classification ascendante hiérarchique des régions. En bas, encadrés, les groupes de la typologie de la matrice visuelle.

A partir de l'ACP, quelle est la signification des deux divergences spatiales observées ?

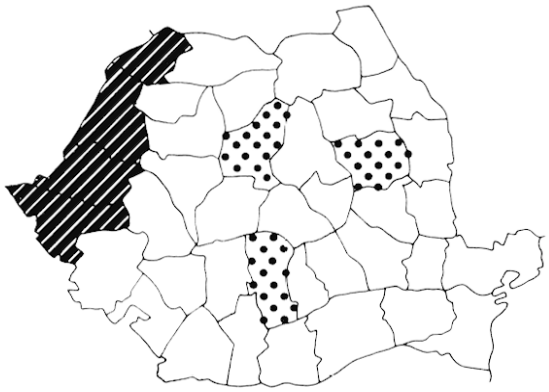
Les départements Argeş, Bacau et Mureş se détachent en groupe autonome, comme départements à faible importance agricole (valeurs négatives sur F1), industrialisation moyenne (F2) et occupation spécifique de l'espace rural (« podgoria », valeurs positives sur F3). Sur le classement graphique, tous trois, intégrés dans le type industriel urbain, auraient pu constituer un sous-groupe de cette catégorie.

Les quatre départements occidentaux sont, dans l'ACP, regroupés avec le type « plaine agricole à pôle de développement ». Dans la matrice, trois d'entre eux, Timiş, Arad, Bihar, constituent un sous-groupe que nous avions jugé inutile d'individualiser dans la typologie synthétique, et ont comme plus proche voisin le type auquel l'ACP les a associés. La position charnière joue là encore et, à l'interprétation, la rupture typologique enregistrée par la matrice exprime une approche de la réalité aussi pertinente que l'association de l'ACP. Quant au quatrième département, Satu Mare, il a le profil le plus moyen (valeur centrale sur les trois facteurs) comme dans la matrice; visuellement, sa place fut difficile à trouver. Dans la division régionale qui fit suite à cette typologie, nous avons identifié une région de « marges parcimonieuses », qui correspond à ces quatre départements occidentaux.

A partir de l'AFC à classes d'effectifs égaux, la combinaison des facteurs met en valeur une plus grande diversité typologique dont, en particulier, le « département à structure intermédiaire ». Ce type, à activités industrielles, urbaines et agricoles moyennes (c'est-à-dire mal classé sur F1 et F2 pour des origines diverses) est subdivisé par les notions de densité et de forme urbaine contenues dans F3. La carte générale met fortement en retrait les départements de Iaşi et Dolj dont les villes principales n'assurent pas l'entraînement général de leur circonscription; elle fait aussi mieux ressortir les individualités régionales traditionnelles de Moldavie et d'Olténie.

En définitive, quelle que soit l'analyse, certains départements se définissent identiquement compte tenu de la vigueur de leur structure, renforcée par un effet de taille. D'une part, il s'agit de l'axe de fort développement Sibiu — Brasov — Brahora; des cinq départements, Timiş et Cluj côté intracarpatique, à solide tradition urbaine et industrielle, Galaţi, Braila, Constanţa côté bas-danubien, à profil agricole et industrialo-portuaire. D'autre part, il s'agit de départements ruraux et peu peuplés de montagne et piémont. Pour les autres, qui combinent diversement agriculture plus ou moins modernisée, tradition urbaine et expansion économique modeste ou récente, donc qui ont un statut de transition, chaque analyse met en évidence ce à quoi elle est le plus sensible, l'écart par rapport à une structure moyenne en AFC, l'ordre des corrélations en ACP.

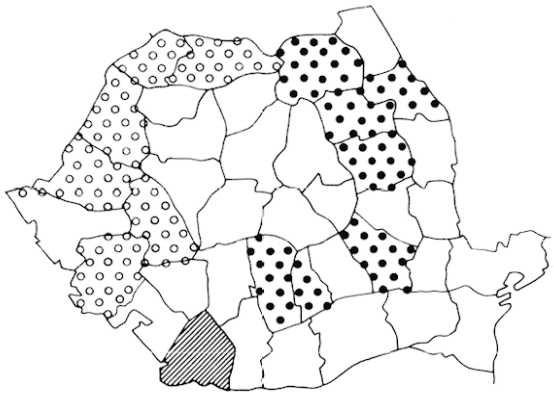
d' Après L' ANALYSE EN COMPOSANTES PRINCIPALES



Départements agricoles à pôle de développement

Départements mixtes

d' après L' ANALYSE FACTORIELLE DES CORRESPONDANCES



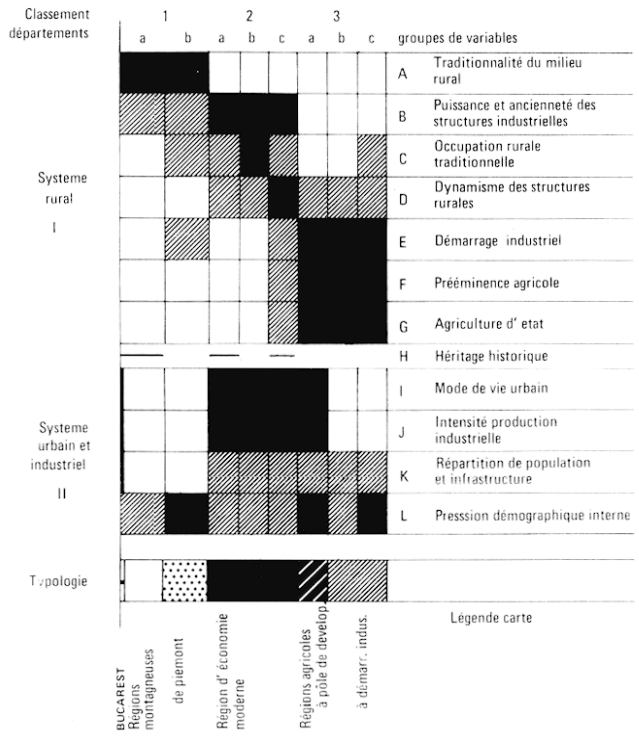
Départements de structure intermédiaire

à faible densité

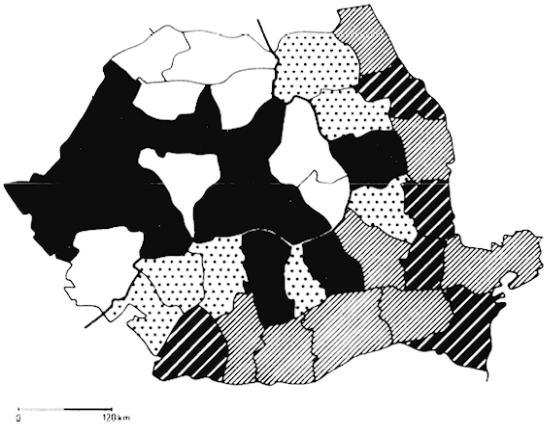
à forte densité

Département agricole à démarrage industriel

MATRICE D' INTERPRETATION



d' après la MATRICE GRAPHIQUE



Ligne de crête des Carpates

FIG. 6. — *Types de départements roumains en* 1970. Les cartes, à gauche, représentent les divergences de classement apportées par les analyses factorielles. Pour les départements laissés en blanc, se reporter à la carte ci-dessus, leur représentation est identique.

CONCLUSIONS.

Convergences...

Une fois encore est apparue, comme un préalable fondamental, la prise en compte de la nature statistique des données et des caractéristiques limites de chaque outil, afin d'éviter des résultats aberrants dus à un mauvais usage.

A cet égard, l'exercice de comparaison de plusieurs méthodes de traitement sur un même tableau de données est un peu artificiel, dans la mesure où l'élaboration de ce tableau doit être conçue en fonction du type de traitement statistique auquel il sera soumis.

Ceci étant posé, c'est la relative *stabilité des résultats globaux,* quelle que soit l'analyse, qui nous a frappés dans ces triturations multiples de données identiques; stabilité telle qu'un chercheur attentif peut avoir l'assurance de cerner des structures fortes sur les variables et les individus au terme d'une série d'analyses de même type.

Deux questions importantes restent de toute façon mal résolues par les deux types de méthodes. Dans quelle mesure la structure obtenue reflète-t-elle une structure contenue dans les données ou n'est-elle qu'un artefact ? L'incertitude quant à la validité des pondérations introduites, le caractère parfois indécis de la hiérarchie d'ensemble, les limites floues de la plupart des groupes de variables et d'individus obligent à relativiser la portée d'un seul résultat. De même en est-il de l'évaluation et de l'interprétation des effets de taille, diversement pris en compte par chaque analyse sans que l'on puisse identifier les effets de seuils qui les sous-tendent.

... et spécificités.

Compte tenu des limites de notre expérience à travers ces exemples, sont apparues des propriétés communes et des propriétés spécifiques de quatre types de traitement des données. Pouvons-nous éclairer le choix d'une méthode mathématique ou graphique en résumant les principaux avantages et inconvénients de chacune d'elles ?

L'intérêt de l'analyse mathématique tient en sa précision, sa reproductibilité, sa comparabilité, et à la possibilité d'utiliser ses résultats pour la modélisation et la prévision. Toutefois, l'indépendance (l'orthogonalité) des facteurs et la présentation analytique et chiffrée entraînent souvent un discours relativement étroit, énumérant le contenu des facteurs successifs. En outre, il est vraisemblable que, dans la pratique, la réduction mathématique en ce petit nombre de facteurs, où certaines variables apparaissent plusieurs fois tandis que d'autres « disparaissent », vu leur faible contribution, risque de limiter le chercheur à un discours d'interprétation rapide sur « l'essentiel » (souvent un peu « trivial »), en lui faisant négliger des variables auxquelles au départ il accordait de l'importance.

En revanche, la matrice ordonnable oblige à un discours global plus structuré et sans doute plus complet dans la description de l'image. Son atout le plus apparent tient dans la présentation simultanée de la totalité de l'information initiale et de son classement final (3). Toutefois, il est difficile d'aller au-delà de ce résultat, de le comparer à d'autres et de le soumettre à un traitement d'une autre nature. Au moins serait-il souhaitable que le tableau chiffré des données initiales accompagne toujours la présentation des résultats graphiques.

En définitive, la comparaison des traitements visuel et mathématique montre qu'il s'agit de *deux langages différents,* c'est-à-dire de deux modes d'expression sur la réalité. Le choix de l'un ou de l'autre peut dépendre des moyens disponibles, des préférences personnelles du chercheur, et de la finalité assignée à la communication des résultats.

Si ces techniques sont dans une certaine mesure interchangeables, c'est qu'elles s'insèrent dans une démarche générale, un processus de recherche identique. Elles assurent un double gain d'information, en organisant les données d'une part, en permettant d'aller plus loin d'autre part : la recherche et l'insertion d'informations nouvelles prennent un tout autre sens avant et après le traitement. Bien comprise, l'analyse des données n'est donc pas seulement un instrument parmi beaucoup d'autres, mais elle constitue une étape essentielle de toute recherche.

(3) A condition de publier la matrice des données. Dès lors, c'est aussi le cas des analyses factorielles ! (N.D.L.R.).

BIBLIOGRAPHIE

M. Brocard et V. Rey, Le traitement graphique de l'information, deux exemples d'utilisation de la matrice. *Analyse de l'Espace,* 1975, n° 3, p. 1-28.

M. Brocard et P. Rodocanachi, La recherche dans les régions françaises : apport d'une analyse multivariée des données, à paraître.

D. Noin et D. Pumain, Typologie des agglomérations urbaines françaises d'après les catégories socio-professionnelles. Communication au Colloque National de Démographie, Nice, avril 1976.

D. Pumain, Utilisation de l'analyse des correspondances et de la classification hiérarchique pour la constitution d'une typologie, à paraître.

V. Rey, Les structures de l'espace roumain, typologie par matrice graphique ordonnable. *L'Espace Géographique,* 1973, n° 1, p. 37-49.

V. Rey, *La Roumanie, essai d'analyse régionale*. Paris, S.E.D.E.S., 1975.

La bibliographie de base est trop volumineuse pour être rappelée ici. Voir l'article de Y. Chauviré dans ce fascicule pour la matrice ordonnable; pour les analyses mathématiques, voir le n° 4, 1975, de *l'Espace géographique*. Pour la comparaison des techniques mathématiques, voir : J. Dreyfus, *Implication ou neutralité des méthodes statistiques appliquées aux sciences humaines. L'analyse des correspondances*. Paris, CREDOC, 1975. — H. Lebras, Vingt analyses multivariées d'une structure connue. *Mathématique et Sciences Humaines*, 1974, vol. 47, p. 37-57.

Point de vue

PERCEPTION ET CALCUL DANS L'ANALYSE TYPOLOGIQUE

Roger BRUNET

L'ensemble de réflexions et d'études de cas présentant ici une forme de traitement « visuel » de l'information soulève des questions intéressantes. Voici quelques remarques critiques, qui sont bien entendu ouvertes à la contradiction.

1. La qualité de certains résultats n'est pas douteuse. Comme pour certaines analyses factorielles, ce n'est pas un argument, dans la mesure où la connaissance préalable du sujet par l'auteur ne peut être écartée. Laissons également de côté les problèmes du choix des variables : ils sont les mêmes dans les deux cas.

2. La technique utilisée, en revanche, autorise bien plus de « manipulations », dans tous les sens du mot, que ne le permet le calcul. L'abondance des choix stratégiques successifs, décrits par les auteurs, y contribue; et l'on voit bien que chacun peut modifier à son gré la plupart des typologies secondaires — et parfois primaires — obtenues.

3. L'ensemble manque de rigueur, au moins dans le vocabulaire : trop de termes flous sont employés pour désigner des catégories ailleurs très strictes, comme variables, classe, étendue, effectif, et à plus forte raison facteurs ou système.

4. Les techniques de calcul (analyse factorielle par exemple) ont sans doute leurs mystères; du moins leur logique est-elle explicitée; les mathématiciens et les utilisateurs qualifiés savent exactement quels sont les partis pris, et à peu près exactement les biais qui en résultent. Dans le type de technique « visuelle » employée ici, deux graves questions sont soulevées : a) celle de la perception visuelle : on trie des taches noires, grises ou blanches, avec toutes les distorsions connues et inconnues que la perception peut introduire (cf. les articles de S. Rimbert et J.C. Muller dans *l'Espace géographique*, 1973-4 et 1976-1); b) celle de la logique interne de ces rangements successifs. Ces « manipulations » ressortissent à un bricolage qui, pour être « bien de chez nous », n'en paraît pas moins laborieux, voire aléatoire. La grande idée de base est la diagonalisation du tableau (initiale ou par sous-parties), seule façon de mettre de l'ordre dans le « bruit » : mais elle pré-suppose que le principal facteur de classement permet de ranger tous les individus sur un continuum, du « plus » au « moins »; si l'hypothèse se vérifie parfois, elle reste gratuite — et laisse en outre trop de part aux redondances, aux paquets de variables répétitives, dont le groupement influe sur le classement, et ossifie la diagonale.

5. Les résultats sont très sensibles à la façon dont les données sont regroupées en classes; l'article de M. Brocard et al. (§2, a, 1er alinéa) le dit clairement; ce regroupement entraîne aussi une perte d'information non négligeable; et il écrase les extrêmes lorsqu'on choisit des classes à effectifs égaux.

6. La technique ne procure aucune économie réelle, ne se fragmente pas, et va moins loin que le calcul. Les travaux publiés dans le fascicule précédent de *l'Espace géographique* le montrent (articles de G. Lazarev et C. Cretin). Un essai réalisé par nous sur les données de S. Bonin (région Centre) a donné des résultats plus complets pour une dépense de vingt minutes de perforation et 9,5 F de frais de calcul !... Ajoutons que le calcul procède par étapes et qu'il suffit souvent d'en rester à un bon graphe de corrélation, comme je l'ai plusieurs fois montré. Il ne serait pas juste de croire qu'il y aurait une technique du pauvre (sans calcul) et une technique de luxe (avec calcul).

7. Dans l'ensemble, il y a un danger incontestable à se fonder sur des techniques visuelles à manipulations successives — les superpositions de cartes n'échappent pas à cette critique — : c'est celui de la fausse facilité, de la recherche purement empirique par tâtonnements, sans effort pour poser les problèmes, pour chercher les logiques internes, pour choisir les procédures adaptées à la nature des problèmes; et de l'emploi de techniques qui favorisent les distorsions dues à la perception, à l'autocorrélation, aux enchaînements a priori.

8. Mais la matrice ordonnable a un très grand mérite. Instrument d'initiation, à vertus pédagogiques, elle me semble devoir très vite donner l'idée de rechercher ces éléments de logique, et de passer au calcul. C'est le cheminement suivi par certains chercheurs. C'était peut-être une étape utile pour certains d'entre eux — un autre danger étant de faire faire des analyses factorielles sans avoir idée de leur contenu. En ce sens, elle a peut-être encore un certain avenir.