L'Actualité économique

Un exemple de l'utilisation des méthodes de l'analyse des données dans la démarche scientifique en économie
The use of multivariate analysis procedures in economic methodology

Henri Leredde et J.-François Outreville

Volume 57, numéro 4, octobre–décembre 1981

URI : https://id.erudit.org/iderudit/601004ar
DOI : https://doi.org/10.7202/601004ar

Aller au sommaire du numéro

Éditeur(s)

HEC Montréal

ISSN

0001-771X (imprimé)
1710-3991 (numérique)

Découvrir la revue

Citer cet article

Leredde, H. & Outreville, J.-F. (1981). Un exemple de l'utilisation des méthodes de l'analyse des données dans la démarche scientifique en économie. *L'Actualité économique*, *57*(4), 507–524. https://doi.org/10.7202/601004ar

Résumé de l'article

The purpose of this paper is both to report on a new methodology in the field of automatic classification, and to discuss the need of these many techniques which are referred to as factor analysis.

The paper is designed to explore one group of procedures and to illustrate with a specific example that such a methodology can lead to an awareness of the problem and to a better understanding and forecasting of economic phenomena.

UN EXEMPLE DE L'UTILISATION DES MÉTHODES DE L'ANALYSE DES DONNÉES DANS LA DÉMARCHE SCIENTIFIQUE EN ÉCONOMIE*

I. AVANT-PROPOS

L'origine de cet article est une thèse de H. Leredde (1979) utilisant une nouvelle famille d'algorithmes de représentation géométrique et de classification automatique à caractère non hiérarchique. Le point de départ de cette recherche fut l'étude d'un procédé particulier : « La Sériation » (Lerman, 1972). Il s'agissait de tester cette méthodologie dans divers domaines d'application (archéologie, démographie, économie, médecine, systèmes informatiques, ...).

L'aboutissement de la recherche de Henri Leredde démontre que ces algorithmes peuvent aussi fournir une représentation graphique très proche de l'analyse factorielle en composantes principales normées. Il est d'ailleurs souvent tentant, à partir des résultats d'une analyse factorielle, d'opérer une classification en reprenant les principaux agrégats. Le fait que ces méthodes aboutissent à des résultats semblables n'apporte pas grand-chose de nouveau à l'économiste dans l'interprétation de ses résultats, puisque l'analyse factorielle est un outil statistique désormais très répandu. Par contre, l'aspect classification est encore très rarement employé en économie.

Les données suggérées pour le domaine de l'économie sont les variables d'un modèle économétrique du secteur financier français développé à l'Institut orléanais de Finance (1975).

L'intérêt de l'utilisation de ces séries économiques était double :

1 — pour l'analyse des données, elles fournissaient un ensemble cohérent de variables;

* Les auteurs remercient A. Dionne et J. Pottier pour leurs commentaires et les dissocient bien sûr de toutes erreurs possibles.

2 — pour l'économiste, il s'agissait de vérifier si cette méthodologie pouvait s'insérer dans la démarche scientifique et permettre de comprendre les problèmes dus à la période d'étude et à la nature des séries statistiques.

II. INTRODUCTION

> « L'analayse des données est actuellement, chez les statisticiens, l'objet d'un véritable phénomène de mode, caractérisé à la fois par l'engouement et le rejet. » (M. Volle, 1978)

Il est bien connu pour les économistes que la simple recherche de liaisons statistiques ne peut donner de résultats fiables sans une réflexion théorique préalable. Toute vérification empirique ne peut être que l'aboutissement d'une démarche scientifique, l'élaboration d'un modèle retenant de la réalité ce qu'elle compte d'essentiel. Comme l'a écrit P. Valéry, « Tout ce qui est simple est faux, tout ce qui ne l'est pas est inutilisable. »

Or, de nombreux auteurs ne font pas mention de la collecte des données comme élément fondamental du processus de recherche contrairement à la démarche proposée par Russel Ackoff (1962) :

1 — formulation du problème;
2 — construction du modèle;
3 — solution du modèle;
4 — collecte des données;
5 — test du modèle;
6 — validation du modèle.

Déterminer les caractéristiques d'un ensemble de données n'est évidemment pas une fin en soi[1], mais cet oubli est malencontreux car le type de données disponibles influence toujours la direction de la recherche.

Pourquoi ne pas élargir le terme « collecte » au terme « analyse »? Cet essai se propose de montrer comment l'analyse des données peut s'avérer un auxiliaire précieux à l'économiste, non pas dans la formulation du modèle, mais bien comme un des éléments de la démarche scientifique.

Dans ce contexte seulement, l'économiste peut situer et délimiter la place de l'anlyse des données — statistique descriptive car elle s'applique à des résultats statistiques bruts — dont elle vise à facili-

1. Remarque de Malinvaud (1964).

ter le maniement; elle se situe immédiatement en aval de la production de ces résultats, et immédiatement en amont des raisonnements empiriques (probabilistes et/ou économétriques).

III. MÉTHODOLOGIE

1. *L'analyse des données*[2]

L'appellation d'« Analyse des données » recouvre une collection d'instruments de statistique descriptive. Elle comporte deux grands groupes de méthodes qui sont les méthodes d'analyse factorielle[3] et les méthodes de classification automatique[4]. Elle regroupe aussi l'analyse discriminante et l'analyse canonique.

Ces diverses méthodes emploient toutes le même procédé : étant donné un nuage de points munis de masses, et situés dans un espace métrique dont le grand nombre de dimensions interdit la visualisation du nuage, il s'agit de trouver les axes d'inertie du nuage et d'obtenir des visualisations sur les plans formés par des couples d'axes. En d'autres termes, nous sommes en présence de données de nature multidimentionnelle. Pour visualiser géométriquement les données, l'analyse va fournir sur une succession de plans un certain nombre de projections[5].

L'aspect des résultats de l'analyse, s'il apporte des indications, conduit aussi à se poser des questions imprévues au départ, et ces questions sont souvent l'apport le plus intéressant de l'analyse. Ainsi, un nuage peut contenir un ou plusieurs « points aberrants » très éloignés du centre de gravité. Ces points correspondent, soit à des erreurs (et dans ce cas l'analyse aura au moins permis de les repérer), soit à des « individus » ou événements » très « originaux », et dans ce cas, une étude détaillée est nécessaire.

2. *Les algorithmes utilisés dans cette recherche*

Les algorithmes utilisés dans cette recherche appartiennent à une méthode relativement nouvelle d'analyse des données. Ils sont issus d'une famille d'algorithmes de représentation géométrique et de classification automatique à caractère « non hiérarchique » [Lerman (1972), Lerman-Leredde (1977) et Leredde (1979)].

---

2. Pour une présentation complète des méthodes de l'analyse des données, voir l'article de M. Volle (1978).

3. Voir par exemple, Benzecri (1973) ou Bertier-Bouroche (1975).

4. On distingue les méthodes de classification hiérarchique des méthodes de classification non hiérarchique. Voir à cet effet, Lerman (1970), Diday (1970) et Eisenbeis-Avery (1972).

5. Le traitement graphique de l'information est une discipline fondée par J. Bertin. Voir son dernier livre publié en 1977.

L'origine même de cette recherche était de tester sur des variables du domaine de l'économie cette nouvelle approche, de comparer les résultats à ceux de l'analyse factorielle traditionnelle et de vérifier l'apport d'information d'une telle approche dans la conception de modèles économétriques.

a) *La méthode des pôles d'attraction*

Le principe général de cette méthode est de détecter par une analyse de la variance, des proximités parmi les éléments de l'ensemble à classifier, des « pôles d'attraction » à partir desquels il est possible, soit de former des classes, soit de fournir une représentation géométrique du nuage des points.

Nous avons utilisé deux algorithmes de la famille des pôles d'attraction, l'une utilisant les distances (la méthode des pôles d'attraction sur les distances MPATD), l'autre utilisant les similarités (la méthode des pôles d'attraction sur les similarités MPATS).

Ces méthodes dont nous tirerons l'aspect classification automatique, nous permettent d'aborder une méthodologie encore très rarement employée en économie[6].

Le principe général de ces méthodes est de constituer de manière itérative des agrégats parmi les éléments à classer. Pour former un agrégat, il suffit de déterminer un élément pertinent appelé « pôle » auquel on agrège les éléments qui lui sont proches, en fonction d'un certain rayon (ou seuil) d'agrégation. Cet agrégat constitué, le processus itératif permet de classer, tant qu'il en reste, les éléments non encore agrégés. À la fin du processus, on obtient une partition de l'ensemble des éléments (algorithme MPATD en annexe).

b) *L'analyse factorielle*

L'analyse factorielle en composantes principales normées (ACP), est particulièrement adaptée à la réduction de tableaux de mensurations, où l'on travaille le plus souvent sur une matrice de corrélation entre caractères. L'analyse factorielle substitue aux mesures primitives de nouvelles variables ou facteurs (qui sont des combinaisons linéaires des variables primitives) à l'aide desquels il est donc possible d'optimiser la visualisation des données en un nombre restreint de figures.

---

6. En économie, voir le livre de W.D. Fisher (1969), en finance, deux articles dans le *Journal of Financial and Quantitative Analysis* démontrent l'intérêt de cette méthodologie : Elton-Gruber (1970) et Joy-Tollefson (1975).

IV. LES DONNÉES

En ce qui concerne l'analyse des données, nous avons étudié les variations de vingt paramètres économiques relevés par trimestre sur une période allant de 1967 à 1974 (tableau 1).

Quant au modèle que nous présentons, il s'agissait d'un modèle d'inspiration keynésienne, dans le cadre d'une économie fermée, mais intégrant le système bancaire dans une approche voisine aux modèles du genre développés aux U.S.A. (FRB-MIT par exemple).

Il s'agissait d'un modèle trimestriel de dimension réduite (9 équations) couvrant la même période 1967-1974 dont l'objectif était double :

1. mettre en évidence les influences déterminantes pour l'évolution des variables financières caractéristiques que sont le crédit, les dépôts à terme et à vue, les taux d'intérêts ;
2. l'examen détaillé des voies de propagation de la politique monétaire et ses répercussions sur le secteur réel de l'économie.

Le bilan simplifié du système bancaire consolidé est représenté comme suit :

$$(RR + RL) + CB + EP = DV + DT$$

à savoir :

réserves totales des banques (libres et obligatoires),
plus crédit bancaire,
plus portefeuille d'effets publics, d'une part,
dépôts à vue plus dépôts à terme, d'autre part.

Le portefeuille d'effets publics essentiellement formé de bons du trésor est, de nos jours, d'un volume négligeable. Pour cette raison, nous n'en tiendrons pas compte.

L'évolution du portefeuille de réescompte de la Banque de France (MCC) a été marquée au cours de la période d'étude par une transformation structurelle de grande ampleur tenant à la réforme des mécanismes d'intervention de la politique d'open-market plus affirmée. La variable *DUMMY* traduit la volonté des autorités monétaires de passer désormais en priorité par le marché monétaire en laissant progressivement le taux sur celui-ci *(TAjj)* descendre en-dessous du taux de réescompte *(TRESC)*.

Les autres variables instrumentales de la politique monétaire sont les coefficients des réserves obligatoires des banques sur les dépôts à vue (*TRDV*) et les dépôts à terme (*TRDT*).

TABLEAU 1
LES VARIABLES DU MODÈLE

CURR	Billets et monnaie divisionnaire
DT	Dépôts à terme dans les banques
DV	Dépôts à vue dans les banques
RR + RL	Réserves totales des banques (libres et obligatoires)
Y	Production intérieure brute
FBCF	Formation brute de capital fixe
CP	Consommation des ménages
DY/Y	Variation relative de la production
TRESC	Taux de réescompte de la Banque de France
TCC	Taux de base bancaire
TDT	Taux d'intérêt sur les dépôts à terme
TOBL	Taux de rendement des obligations
TAjj	Taux de l'argent au jour le jour
TACT	Taux de rendement des actions
TRDT	Coefficient des réserves obligatoires sur les dépôts à terme
TRDV	Coefficient des réserves obligatoires sur les dépôts à vue
DUMMY	Modifications de la politique monétaire de la Banque de France
MCC	Portefeuille de réescompte de la Banque de France
MCG	Monnaie centrale gratuite (or et devises)

Deux autres variables intéressantes sont prises en compte, à savoir la somme des billets et monnaie divisionnaire (*CURR*) et la monnaie centrale gratuite (or et devises) (*MCG*).

Nous avons introduit un ensemble de taux d'intérêts représentant les coûts d'opportunité des agents sur les différents marchés financiers :

— le taux d'escompte de la Banque de France : *TRESC*
— le taux de l'argent au jour le jour : *TAjj*
— le taux de base bancaire sur les emprunts : *TCC*
— le taux bancaire sur les dépôts à terme : *TDT*
— le taux de rendement des obligations : *TOBL*
— le taux de rendement des actions : *TACT*

Enfin, la sphère réelle du modèle inclut :

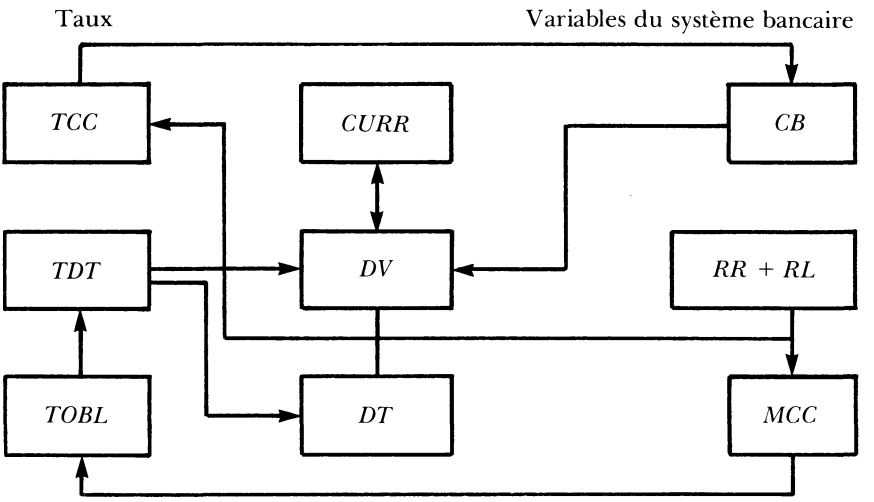
— la production intérieure brute : *Y*
— la consommation des ménages : *CP*
— la formation brute de capital fixe : *FBCF*

Les séries trimestrielles sont tirées du Bulletin trimestriel de la Banque de France et du Bulletin mensuel de statistiques de l'INSEE.

Les relations d'interdépendance entre les variables endogènes apparaissent dans la figure 1.

FIGURE 1

RELATIONS STRUCTURELLES DU MODÈLE



L'estimation des équations du modèle nous avait rapidement conduit à des spécifications mettant en évidence certaines particularités du secteur financier français : rôle très faible des bons du Trésor, importance rapidement décroissante du portefeuille d'escompte de la Banque centrale (*MCC*), absence de relation de structure à terme des taux d'intérêts, pas de différence entre les réserves obligatoires et les réserves totales des banques auprès de la Banque centrale *(RR + RL)*. Enfin, les résultats obtenus avec le taux bancaire sur les dépôts à terme *(TDT)* sont peu satisfaisants et nous avaient incité à tester d'autres variables alternatives[7].

V. LES RÉSULTATS

1. *Classification par pôles d'attraction*

L'application de l'algorithme de classification automatique nous donne une partition en cinq classes qui varient peu quand l'on

7. Sur la période d'étude, l'évolution de cette variable correspond à des seuils (marche en escaliers).

augmente le seuil d'agrégation (et par le fait même, la part d'inertie expliquée). Si l'on considère la seconde partition qui semble excellente, le rapport inertie inter-classe sur inertie totale du nuage est de 56,94% (tableau 2).

TABLEAU 2
CLASSIFICATION AUTOMATIQUE

	Partition 1	Partition 2	Partition 3
Nombre de classes de la partition	5	5	5
Part d'inertie expliquée	52,67%	56,94%	65,36%
Algorithme	*MPATD*	*MPATD*	*MPATS*
Classe 1	*CB* *DT* *DV*	*DT* *CB* *DV* *CURR* *RR + RL*	*DT* *CB* *DV* *CURR* *RR + RL*
Classe 2	*CP* *Y* *FBCF*	*Y* *FBCF* *CP* *DY/Y*	*CP* *Y* *FBCF* *TDT*
Classe 3	*TRDT* *TRDV* *DUMMY* *RR + RL*	*TRDV* *TRDT* *DUMMY*	*DY/Y*
Classe 4	*TRESC* *TCC* *TOBL* *TDT* *TAjj* *MCC* *DY/Y*	*TRESC* *TCC* *TOBL* *TDT* *TAjj* *MCC*	*TAjj* *TRESC* *TCC* *TOBL* *MCC*
Classe 5	*MCG* *CURR* *TACT*	*MCG* *TACT*	*TRDV* *TRDT* *DUMMY* *MCG* *TACT*

La première classe correspond bien au système bancaire auquel s'est ajoutée la variable représentant les billets et monnaie divisionnaire (*CURR*). La seconde classe regroupe l'ensemble de la sphère réelle y compris la variation relative de la production (*DY/Y*). La troisième classe est celle de la politique monétaire des autorités. Dans la quatrième classe, on retrouve l'ensemble des taux d'intérêts que nous avons introduits dans le modèle. Seule la dernière classe n'a pas grande signification, si ce n'est qu'elle permet de regrouper les deux variables qui se sont révélées non explicatives et non significatives en tant que variables exogènes dans l'analyse économétrique du modèle.

La même analyse s'applique dans le cas de la troisième partition avec à nouveau un élément « original » agrégé à la classe 2, à savoir le taux bancaire sur les dépôts à terme (*TDT*) qui est justement la variable que nous avions signalée dans la partie précédente.

L'utilisation de méthodes de classification automatique en économie semble donc possible tout au moins pour vérifier l'existence d'ensembles cohérents et révéler éventuellement des problèmes inhérents à certaines variables.

2. *Analyse factorielle en composantes principales*

Nous présentons dans quatre graphiques (figures 2 à 5), les résultats de la projection des variables et des observations sur les axes factoriels 1 et 2, et 1 et 3.

À titre de comparaison avec la méthode de classification, les deux premiers axes factoriels pour les variables expliquent 84,67% d'inertie. Nous retrouvons bien l'ensemble des taux qui s'opposent au taux de rendement des actions, et les variables du système bancaire et de la sphère réelle de l'économie qui s'associent pour entraîner le premier axe.

Les deux premiers axes nous permettent aussi d'observer le rôle particulier joué par le portefeuille de réescompte de la Banque centrale *(MCC)* et par la monnaie centrale gratuite (or et devises) *(MCG)*.

Nous pouvons interpréter les résultats de la projection des observations sur les axes factoriels 1 et 2 (fig. 4) et 1 et 3 (fig. 5), pour la période 1967-1973, car ils sont très significatifs de la validité d'une telle méthodologie.

Afin d'interpréter ces projections des observations sur le premier plan factoriel, il nous faut revenir sur les projections des variables sur ce même plan (fig. 2). Le premier axe factoriel est

entraîné par l'ensemble des grands agrégats économiques et en particulier par la production qui est croissante sur la période observée. Cet axe sous-tend donc également l'évolution dans le temps. C'est ce que nous retrouvons de manière très frappante le long du premier axe pour la projection des trimestres (fig. 4). Le second axe lui (fig. 4) oppose l'ensemble des taux d'intérêts au taux des actions. Or, toujours sur la figure 4, nous voyons apparaître une courbe continue dans le temps, entraînée précisément par le second axe factoriel qui correspond parfaitement à l'évolution des principaux taux d'intérêts sur cette même période (fig. 6). Cette ressemblance mérite d'être soulignée.

Pour le plan factoriel 1-3 (fig. 5), reportons-nous d'abord à la projection des variables (fig. 3) sur laquelle nous pouvons voir le troisième axe entraîné par les variations relatives de production (*DY/Y*), le taux de rendement des actions (*TACT*) ayant une certaine tendance à « suivre » cette variation de production. Or, une fois encore, en étudiant la projection des observations sur ce même axe et sur le premier axe qui reflète le temps (fig. 5), nous retrouvons la tendance de la courbe des variations de production au cours du temps sur la période considérée. Deux points remarquables sont à souligner, la chute de la production du deuxième trimestre 1968 et la vigoureuse reprise du troisième trimestre. Cette cassure du printemps 1968 apparaît très en évidence sur notre graphique.

Plus que la projection des variables qui n'apporte pas d'informations nouvelles par rapport à l'analyse de classification, c'est surtout la projection des observations qui fournit de précieux renseignements sur l'évolution des variables dans le temps et surtout les modifications structurelles ou les perturbations (le printemps 1968 en France) qui risquent de biaiser les résultats de l'analyse économétrique.

FIGURE 2

ANALYSE FACTORIELLE (*ACP*)

Projection des variables sur les axes 1 et 2

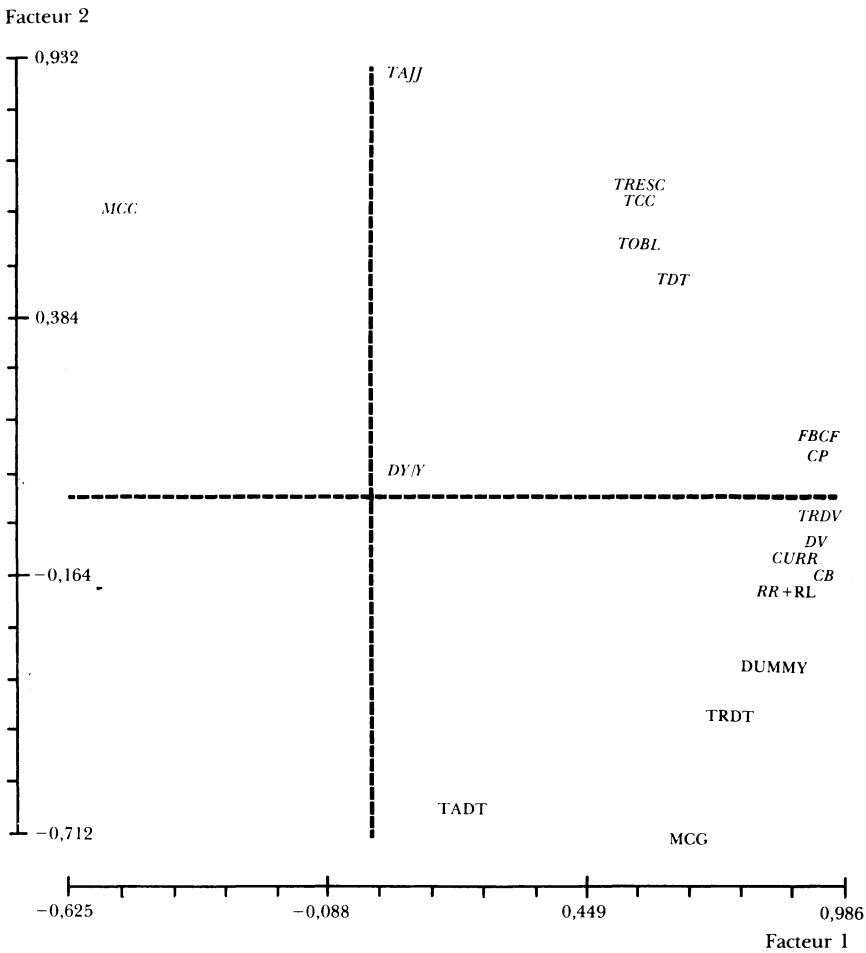


FIGURE 3
ANALYSE FACTORIELLE (*ACP*)
Projection des variables sur les axes 1 et 3

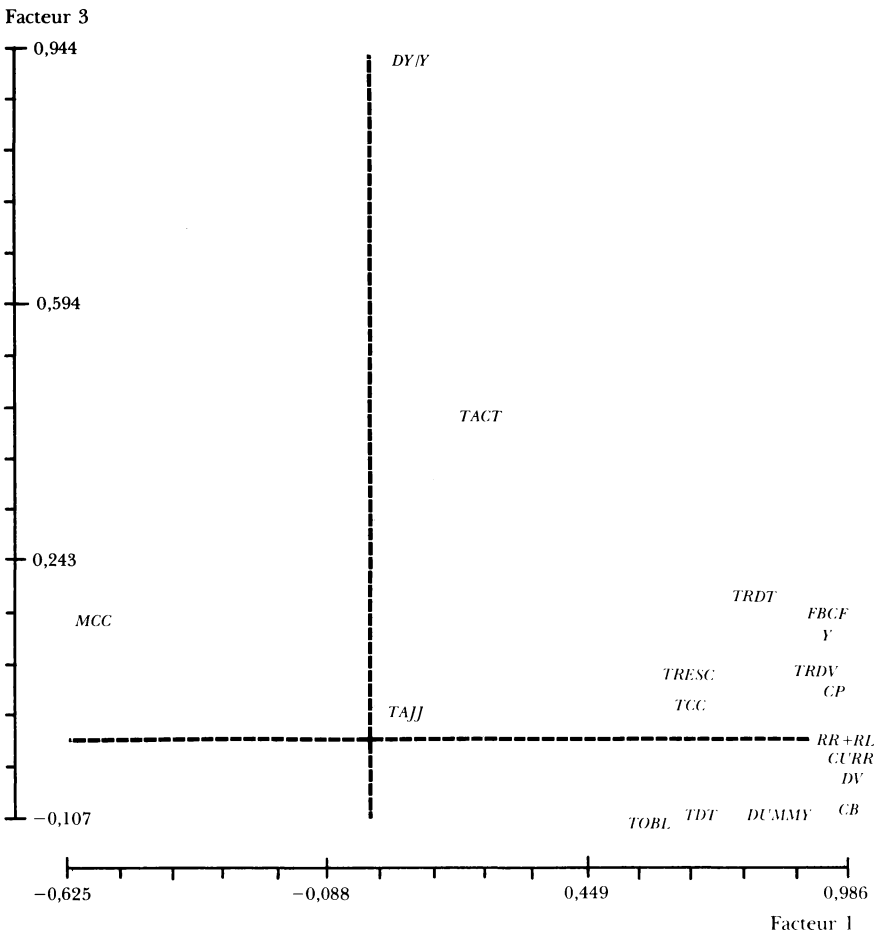


FIGURE 4
ANALYSE FACTORIELLE (*ACP*)
Projection des observations sur les axes 1 et 2

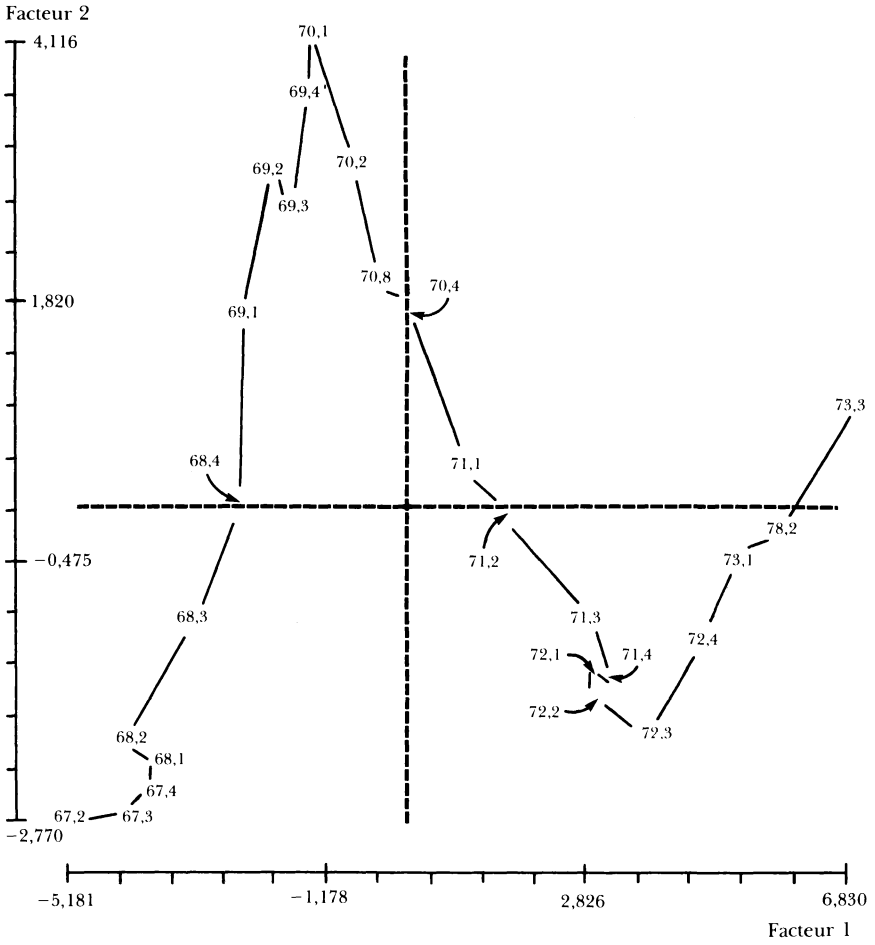


FIGURE 5

ANALYSE FACTORIELLE (*ACP*)

Projection des observations sur les axes 1 et 3

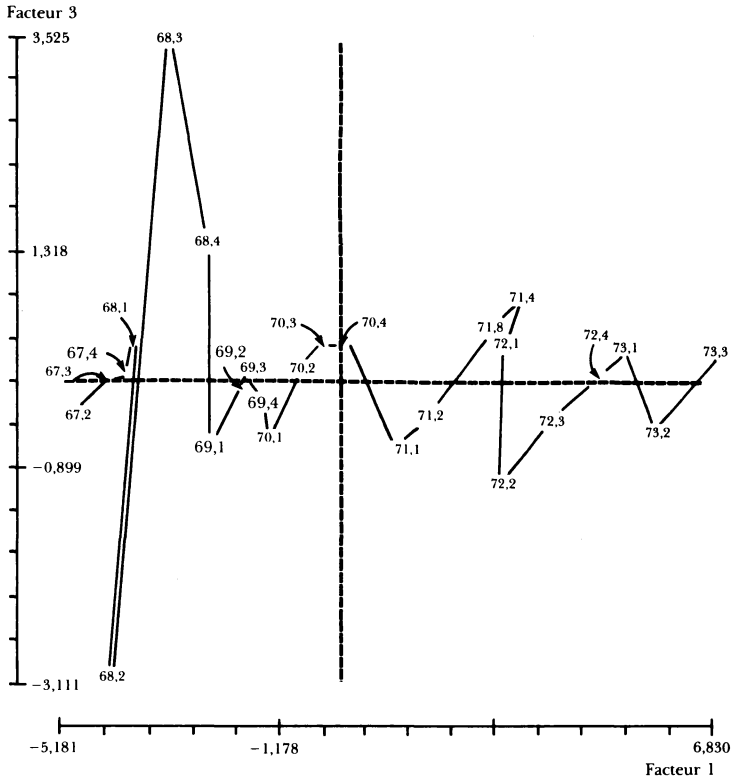
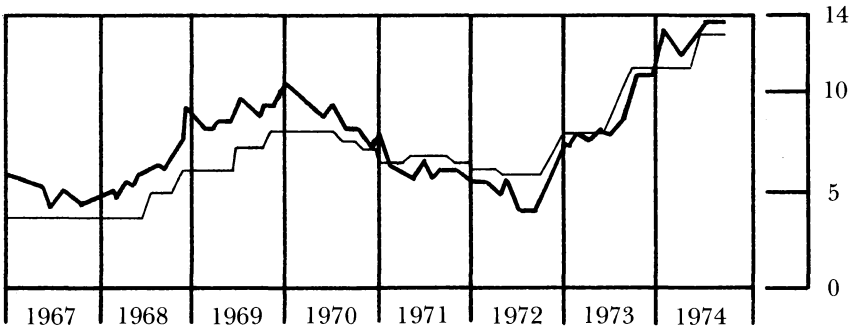


FIGURE 6

LE MARCHÉ MONÉTAIRE



Taux de l'argent au jour le jour (effets privés)

Taux d'escompte de la Banque de France

SOURCE: INSEE

VI. CONCLUSION

Les résultats fournis par la méthode de classification automatique sont très encourageants. Ce type d'outil statistique a été jusqu'à ce jour très peu utilisé en économie. C'est en voulant tester la validité de nouveaux algorithmes en analyse des données que nous avons pu juger de l'intérêt et de l'apport de ces méthodes dans ce domaine. Il existe déjà en analyse de données plusieurs autres méthodes de classification qu'il serait certainement très intéressant d'utiliser [Benzecri (1973), Diday (1970), Lerman (1970)].

Les méthodes de classification automatique obligent à une précision rigoureuse dans le choix des variables prises en compte et des critères d'agrégation. S'il est vrai que l'analyse des données en général ne révèle que des évidences, cela est finalement heureux et permet d'autant plus d'intégrer cet outil dans la démarche de recherche. Comme M. Volle (1978) le fait remarquer : « Il est trop facile de dire, une fois un résultat établi, qu'il était évident. On oublie trop facilement que le résultat contraire était bien souvent tout aussi évident à priori... ».

« *In principle, tolls have a servant's status. The best choice of tools depends on the problem area selected and on the extent to which at least partial answers have been found... The availability of certain tools may lead to an awareness of problems, important or not, that can be solved with their help.* » T. C. Koopmans, *Three Essays on the State of Economic Science,* 1957, McGraw-Hill, N.Y..

Henri LEREDDE
Université Paris-Nord

et

J.-François OUTREVILLE
Université Laval, Québec

BIBLIOGRAPHIE

ACKOFF, Russel L., *Scientific Method, Optimizing Applied Research Decision,* John Wiley and Sons Inc., N.Y., 1962.

BENZECRI, J. P., *L'analyse des données,* (vol. I et II), Dunod, Paris, 1973.

BERTIER, P. et BOUROCHE, , J. M., *Analyse des données multidimensionnelles,* P.U.F., Paris, 1975.

BERTIN, J., *Le graphique et le traitement graphique de l'information,* Paris, 1977, Flammarion.

DIDAY, Edwin, « Une nouvelle méthode de classification automatique et reconnaissance des formes : la méthode des nuées dynamiques », *Revue de Statistique appliquée,* vol. XIX, no 2, 1970.

EISENBEIS, Robert A. and AVERY, Robert B., *Discriminant Analysis and Classification Procedures : Theory and Applications,* D.C. Health and Col., Lexington Mass., 1972.

ELTON, Edwin J. and GRUBER, Martin J., « Homogeneous Groups and the Testing of Economic Hypothesis », *Journal of Financial and Quantitative Analysis,* January 1970, pp. 581-602.

FISHER, W. D., *Clustering and Aggregation in Economics,* Johns Hopkins Press, Baltimore, 1969.

Institut orléanais de Finance, *Un modèle trimestriel du secteur financier français,* Université d'Orléans, Document interne non publié, 1975.

JOY, Maurice O. and TOLLEFSON, John O., « On the Financial Applications of Discriminant Analysis », *Journal of Financial and Quantitative Analysis,* décembre 1975.

LEREDDE, Henri, *La méthode des pôles d'attraction, la méthode des pôles d'agrégation,* Thèse de 3e cycle en mathématiques appliquées, Paris, 1979.

LERMAN, I. César, *Les bases de la classification automatique,* Gauthier-Villars, Paris, 1970.

LERMAN, I. C., Analyse du phénomène de la sériation à partir d'un tableau d'incidence, *Mathématiques et Sciences humaines,* 1972, no 38, pp. 39-57.

LERMAN, I. C. et LEREDDE, H. « La méthode des pôles d'attraction », *Analyse des données et informatique,* (éd. IRIA), 1977, pp. 37-49.

MALINVAUD, E., *Méthodes statistiques de l'Économétrie,* Dunod, Paris, 1964.

VOLLE, Michel, « L'analyse des données », *Économie et Statistiques,* no 96, janvier 1978, pp. 2-23.

ANNEXE

L'ALGORITHME MPTD:
(Extrait de H. Leredde (1979))

1. Centrer et réduire les variables, puis former une matrice des similarités entre variables ou entre objets selon le nuage de points à visualiser.

Soit: p nombre de variables,
n nombre d'observations,
x_{ij} valeur de la j^e variable pour la i^e observation,
μ_j moyenne de la j^e variable.

$$\mu_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

σ_j écart-type de la j^e variable :

$$\sigma^2{}_j \quad \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \mu_j)^2$$

Les données sont centrées et réduites :

$$X_{ij} = \frac{x_{ij} - \mu_j}{\sqrt{\sigma^2{}_j}\sqrt{n}}$$

Indice de similarités entre variables :

$$S(k,1) = \sum_{i=1}^{n} X_{i,\,k} X_{i,\,1} = \sqrt{n} p(k,1) \quad (P(k,1) = \text{coefficient de corrélation})$$

incide de similarités entre observations :

$$P(r,s) = \sum_{j=1}^{p} X_{r,j} X_{s,j}$$

Nous désignerons désormais indifféremment par $S(*,*)$ la matrice des similarités aussi bien entre variables qu'entre observations, et par N, le nombre d'éléments de l'ensemble à organiser ($N = p$ si l'on étudie les variables, $N = n$ si l'on étudie les observations).

2. Dans cet algorithme, former ensuite une matrice des distances entre éléments à classifier. La métrique utilisée est celle sous-jacente aux similarités.

$$D^2(i,j) = S(i,i) + S(j,j) - 2S(i,j)$$

À partir de cette matrice de distances, déterminer des *pôles d'attraction* autour desquels on constitue des agrégats avec les éléments les plus

proches. Pour cela, on calcule la variance des distances de chaque élément aux autres éléments :

$$\mathrm{VAR}(i) = \frac{1}{N-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} D^2(i,j)$$

le premier pôle est l'élément de variance maximum :

$$P_1 \text{ tel que } \mathrm{VAR}(P_1) = \max_{1 \leq q \leq N} \left[\mathrm{VAR}(q)\right]$$

Ce premier pôle étant choisi, agréger autour de lui tous les éléments qui en sont le plus proche, fonction d'un certain rayon ou seuil d'agrégation que l'on fixe au départ.

Puis, retirer de l'ensemble des éléments à classifier, cette classe d'éléments qui viennent d'être agrégés.

$$P_2 \text{ tel que } \mathrm{VAR}(P_2) \cdot D^2(P_1, P_2) = \max_{\substack{1 \leq q \leq N \\ q \neq P_1}} - \left[\mathrm{VAR}\, q \cdot D^2(P_1, q)\right]$$

On recommence alors l'algorithme en déterminant un nouveau pôle d'agrégation, toujours selon le principe de la variance maximum, mais choisi cette fois-ci parmi les éléments restant à classer. L'algorithme se termine quand il ne reste plus d'éléments à classer.