Statistique et analyse des données

I. Anastassakos
E. Gaussens

Approche systémique de l'analyse factorielle, les problèmes posés par les deux univers

Statistique et analyse des données, tome 8, nº 3 (1983), p. 1-15.
<http://www.numdam.org/item?id=SAD_1983__8_3_1_0>

Statistiques et Analyse de données
1983 - Vol. 8 n° 7 pp. 1-15

APPROCHE SYSTEMIQUE DE L'ANALYSE FACTORIELLE
LES PROBLEMES POSES PAR LES DEUX UNIVERS

I. ANASTASSAKOS *
E. GAUSSENS *

* Laboratoire de Statistique et d'Etudes Economiques et Sociales
Centre d'Etudes Nucléaires de Fontenay-aux-Roses

Résumé : *L'introduction de l'analyse factorielle à partir de l'analyse des systèmes apporte un nouvel éclairage sur les rôles qui peuvent être attribués aux deux univers (individus et variables) en analyse des données. Le choix de la méthode (analyse en composantes principales ou analyse factorielle au sens de Spearman) dépend alors directement de la stratégie retenue pour la collecte des données. De ces considérations découlent des propositions concrètes pour la construction des questionnaires.*

Abstract : *When factor analysis or principal component analysis is considered within the framework of systems analysis the part of the two universes (variables and individuals) is emphasized. The specific method employed (factor analysis or principal component analysis) is then directly linked to the data collection's strategy. From this approach derive proposals of practical interest to build questionnaries.*

Mots clefs : *analyse factorielle, univers des individus, univers des variables, analyse des systèmes, construction des questionnaires.*

INTRODUCTION

L'introduction du modèle factoriel en utilisant le langage de l'automatique permet de préciser les problèmes que posent les deux univers, celui des individus et celui des variables, en analyse factorielle. Si le problème de stabilité des axes factoriels avec le premier univers est du ressort de la statistique classique, il n'en est pas de même avec l'univers des variables dont la définition renvoie au phénomène particulier qu'il s'agit d'analyser, sur lequel on a des idées a priori traduites sous la forme d'un modèle supposé linéaire. Dans cet article l'analyse factorielle est abordée non comme un ensemble de techniques permettant la description de données, mais comme un moyen pour produire les paramètres d'un modèle qui est satisfaisant dans bien des cas malgré sa simplicité, particulièrement dans les sciences sociales. Ceux qui connaissent l'histoire de l'analyse factorielle retrouveront donc ici, exposées différemment, certaines préoccupations des factorialistes de l'époque qui a suivi les travaux de Spearman puis de Thurstone.

Ayant introduit un formalisme, nous précisons les liens qui existent entre le critère choisi pour faire émerger les paramètres du modèle (les composantes factorielles) et la stratégie à utiliser au niveau de la sélection des variables. Nous arrivons à une conclusion qui a son importance pour comprendre l'histoire de l'analyse factorielle : en suivant Spearman et Thurstone on a moins de précautions à prendre au niveau de la sélection des variables qu'en emboîtant le pas à Hotelling.

Que dire des interprétations qui sont données dans les enquêtes psycho-sociales quand rien n'est dit sur le mode de production des questionnaires ? Les axes factoriels extraits pourraient-ils être, si des précautions sont prises, indépendants d'une imagination qu'il est difficile de maîtriser ? L'approche de l'analyse factorielle proposée ici a l'avantage d'offrir un fil conducteur à celui qui veut analyser un phénomène à travers des opinions : s'il respecte certaines règles de construction des questions, il peut effectivement espérer obtenir des résultats que l'on saura reproduire en ne conservant que l'univers auquel les questions font référence et la procédure qui permet de les extraire.

1. *DE L'ANALYSE DES SYSTEMES A L'ANALYSE FACTORIELLE : UN CHEMINEMENT POSSIBLE*

Dans beaucoup de domaines, comme la biologie et la physique théorique ou encore la gestion des entreprises, l'analyse des systèmes permet d'énoncer des hypothèses, qui conduisent à des modèles décrivant simplement les phénomènes étudiés lorsqu'elles sont vérifiées. De la même façon, dans les Sciences Sociales, l'individu est fréquemment assimilé à un système (ou un sous-système) ; c'est le cas en ergonomie, où l'on insiste sur les interfaces entre les trois sous-systèmes : l'environnement, l'homme et la machine.

Dans les enquêtes psycho-sociales, pour analyser le comportement des individus face à un problème donné, on soumet à un échantillon représentatif de la population étudiée, un ensemble de questions qui doivent prendre en compte les différents aspects de ce problème. Pour analyser l'ensemble des réponses ainsi obtenues, plusieurs méthodes sont envisageables ; le choix d'une méthode particulière est toujours conditionné, implicitement ou explicitement, par l'existence d'un modèle de référence. Nous montrons que l'analyse factorielle, comme méthode de traitement, peut être justifiée par l'introduction de diverses hypothèses issues d'une description du phénomène étudié en termes d'analyse des systèmes.

En effet, on peut retenir, à propos de l'individu, la représentation formelle décrite dans le schéma 1 .

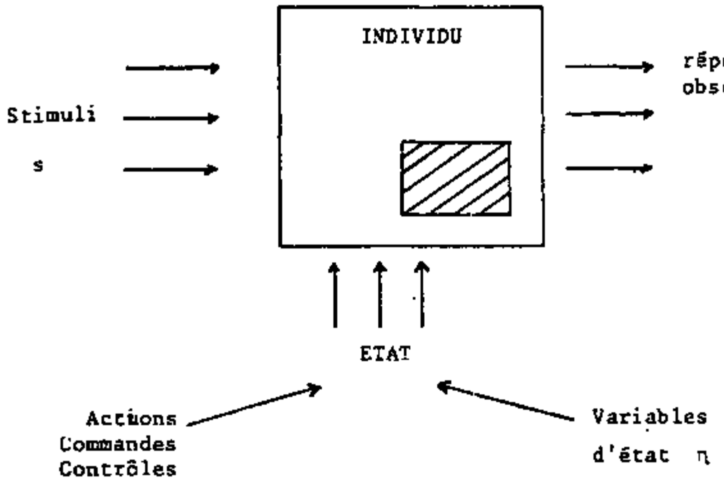


Schéma 1

Le système individu

Dans cette représentation le "système individu" est caractérisé par :

. un état : il est en général saisi à travers des "variables d'états" η^j

. des commandes : elles constituent le moyen dont disposerait un pilote pour modifier l'état.

Par l'intermédiaire des stimuli s, l'état η et les commandes étant fixés, l'individu est confronté à une situation particulière qui engendre sa réponse y .

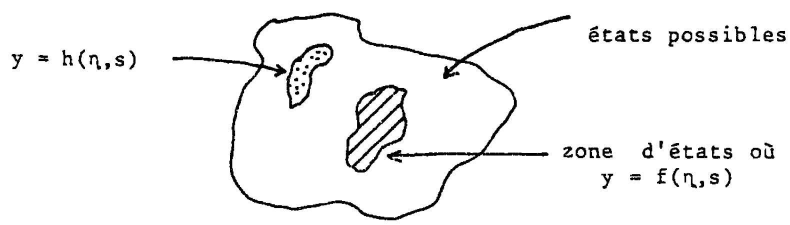
Dans le cas d'une enquête, l'ensemble des stimuli correspond au questionnaire et la variable y aux réponses de l'individu aux questions. En général les enquêtes ne cherchent pas à appréhender l'individu dans son intégralité, ce qui est encore l'apanage des psychologues, mais plutôt à solliciter chez lui, un ensemble de mécanismes, caractéristiques du phénomène que l'on étudie : par exemple l'attitude individuelle face a une catégorie de risques. On suppose donc que ces mécanismes forment un sous-système (partie hachurée du schéma

1) ; c'est lui que l'on cherche à étudier ; c'est donc sur lui que repose la modélisation. Avant d'aborder ce problème, on doit cependant faire deux remarques ; puisque seule nous intéresse l'étude du sous-système, on doit faire en sorte d'une part que l'observation soit caractéristique des réactions du sous-système et d'autre part que les intéractions entre ce dernier et le reste du système soient identiques, ou tout au moins analogues, quels que soient les stimuli envisagés. Considérer ce sous-système revient donc à retenir une classe particulière de stimuli ; nous l'appellerons : <u>l'ensemble des situations permises</u>.

Une fois cet ensemble défini il faut préciser plus avant la nature des liaisons qui régissent les stimuli (situations), les états du système et les réponses. Par analogie avec les méthodes de l'automatique, on peut alors plonger les ensembles de stimuli, d'états et de réponses, dans des espaces topologiques, et utiliser un modèle qui sera supposé ajuster les comportements du sous-système dans une certaine <u>zone d'états</u>. Nous nous limitons au cadre des espaces vectoriels réels (typiquement des espaces R^n), et nous supposons qu'une relation fonctionnelle particulière f relie la réponse y du système au stimulus auquel il est soumis ; cette fonction dépend de l'état du système :

$$y = f(\eta , s)$$

On donne le nom de barrière à la frontière qui sépare la zone d'états (domaine de définition de f) de l'ensemble des états possibles ; au-delà de cette barrière d'autres relations fonctionnelles sont envisageables.



En automatique on connait en général la forme de la fonction f et les états sont saisis à travers des variables d'états clairement définies. On cherche alors à préciser les réponses y du système dans différentes situations ; on se heurte alors souvent à une difficulté : la fonctionnelle f est formalisée comme solution d'un système d'équations différentielles en général non intégrable et lourd à manipuler. Pour trouver la réponse y correspondant à un état du système on éprouve donc déjà des difficultés ; pour définir la "commande" qui, fixant l'état, conduit à une réponse y optimale au sens d'un critère, la difficulté est encore plus grande... si les valeurs prises par certains paramètres caractérisant la fonction f ne sont pas connues, le problème d'estimation qui se pose complique encore la tâche...

Quand on aborde l'individu dans les enquêtes psycho-sociales on ne sait pas grand chose a priori sur le mode de fonctionnement du sous-système qui, confronté à la situation à laquelle on s'intéresse, produit la réponse... on s'appuie sur les théories existantes pour énoncer certaines hypothèses qui permettent de considérer cette situation comme un élément de ce que nous avons appelé la famille des "situations permises". Dans la zone d'états la réponse peut être alors reproduite par l'équation caractéristique du modèle :

$$(0) \qquad y = f(\eta \ , \ s) \ .$$

Ainsi, quand on traite du comportement individuel face au risque, on fait l'hypothèse que toutes les situations qui provoquent un "sentiment d'insécurité" peuvent être considérées comme équivalentes relativement aux mécanismes mis en jeu.

On ne connait pas a priori la fonctionnelle f :

Quelles sont les conditions qui doivent être remplies pour qu'à partir de réponses on puisse trouver une formule du type (0), qui permette, connaissant le stimulus et l'état, de reconstruire la réponse ?

En automatique pour ausculter le modèle au voisinage d'un état η_0 , on utilise des méthodes de perturbations : on admet que le modèle peut être approché au voisinage de η_0 par l'équation :

$$(1) \qquad y - y_0 = f'_\eta(\eta_0, s)(\eta - \eta_0) + \text{résidu} \quad .$$

Quand on traite de l'individu comme d'un système, passer d'un individu à l'autre c'est modifier l'état. Tirer au hasard dans une population peut être considéré comme une stratégie de perturbation des états dans une zone dont la définition coïncide avec celle de la population. Dans les enquêtes psycho-sociales que nous analysons, il s'agit de la population des français d'âge adulte vivant en métropole. On admet ici l'hypothèse selon laquelle tout se passe comme si on ne perturbait qu'une fois autour d'un "individu standard" et que cette perturbation affectait toute la zone d'états...

Si "l'individu standard", dont la réponse aux p stimuli considérés est $\bar{y}$, est assimilé au centre de gravité $\bar{\eta}$ des vecteurs η_i décrivant dans un espace vectoriel H de dimension k les différents états (ce sont nos individus échantillonnés), les équations (1) s'écrivent :

$$y_i - \bar{y} = f'_\eta(\bar{\eta}, s)(\eta_i - \bar{\eta}) + \text{résidu} \; ; \qquad i = 1, \ldots, n \quad .$$

Si on suppose que pour les stimuli permis la dérivée par rapport à η est constante dans toute la zone d'états (hypothèse de linéarité), on aboutit, pour les p stimuli considérés, au modèle linéaire :

$$(2) \qquad x_i = Uc_i + e_i \quad ; \qquad i = 1, \ldots, n \quad .$$

Dans cette équation le vecteur centré $x_i = y_i - \bar{y}$ des réponses $x_i^j = y_i^j - \bar{y}^j$ de l'individu i , décrit par les variables x^j , est situé dans un espace de dimension p ; le vecteur centré $c_i = \eta_i - \bar{\eta}$ est assimilé à un point d'un espace H de dimension k ; la matrice (p x k) U n'est autre que le jacobien $\frac{\partial f}{\partial \eta}$: ses lignes sont les gradients correspondant aux différents stimuli.

Nous allons maintenant montrer comment l'analyse factorielle, à travers un problème d'optimisation, permet d'étudier le modèle (2). Rappelons que pour utiliser cette représentation, on doit, de façon suffisamment précise, définir l'ensemble des situations permises, c'est-à-dire le domaine dans lequel les stimuli doivent être choisis.

2. ANALYSE FACTORIELLE : CRITERES D'OPTIMISATION ET SELECTION DE VARIABLES

Reprenons le modèle linéaire introduit à la fin du paragraphe 1 :

$$(2) \qquad x_i = Uc_i + e_i \quad ; \qquad i = 1,\ldots,n \quad .$$

Si les états sont effectivement paramétrés à travers les k variables $c^1, c^2, \ldots, c^k$ le système des équations (2) permet, connaissant les réponses x_i , de trouver les éléments de la matrice U. On effectuera autant de régressions qu'il y a de stimuli , cela peut se faire de différentes manières. La matrice U étant connue, pour trouver les vecteurs d'état c_i à partir des n réponses x_i (problème classique de "calibration"), on a recours encore à la régression.

Mais la situation qui nous intéresse est plus délicate : nous ne savons pas paramétrer les états car nous ne connaissons pas les variables d'état c^j et la matrice des coefficients U est inconnue ! La seule chose que l'on sache est qu'une relation, du type (2) parait raisonnable pour relier les états aux réponses.

L'énoncé de certaines hypothèses complémentaires concernant les propriétés que devraient observer les résidus e_i conduit à poser le problème de la recherche de U et des variables c^j sous la forme d'un problème d'optimisation. La façon de poser ce problème, c'est-à-dire de choisir le critère à optimiser, est intimement liée à la stratégie employée pour recueillir l'information : on ne peut discuter de l'un sans aborder l'autre.

En effet, jusqu'à maintenant, on a indiqué qu'il fallait s'assurer d'une certaine diversité des états du système ; elle peut être obtenue par une manière de perturber les états : l'échantillonnage. Il reste à définir la méthode de sélection des stimuli : nous savons déjà qu'ils doivent appartenir à l'ensemble que l'on a appelé précédemment "ensemble des situations permises" ; mais comment sélectionner dans cet ensemble ? Il faut évidemment, et c'est tout ce qui peut être dit pour l'instant, que l'exigence (A) suivante se trouve respectée au mieux :

(A) : "La solution du problème d'optimisation est indépendante du sous-ensemble de stimuli choisi par la méthode de sélection envisagée".

C'est à travers cette exigence que le critère d'optimisation est indissociablement lié au mode de collecte des données. Nous examinerons dans les deux paragraphes suivants les familles de critères qui sont couramment utilisés.

2.1 *En suivant C. Spearman et T. Thurstone*

Pour les disciples de Spearman et de Thurstone l'analyse factorielle est la recherche d'un ensemble de dimensions "cachées" $c^1, c^2, \ldots, c^k$ permettant d'expliquer parfaitement les corrélations constatées entre les réponses du système aux différents stimuli (Cf. Mulaik 1972). On pose alors deux conditions :

. les réponses sont fonctions des dimensions cachées à travers une formule du type (2) : $x_i = Uc_i + e_i$,

. les résidus e^j associés à chacun des stimuli sont orthogonaux deux à deux et orthogonaux aux variables d'état.

Ces deux conditions impliquent que les éléments de chacune des lignes de la matrice U sont les coefficients de régression des réponses x^j par rapport aux variables d'état c^ℓ ; la covariance entre les réponses à deux stimuli, les variables d'état c^ℓ étant fixées, est donc nulle :

$$(3) \qquad \operatorname{cov}(x^j, x^{j'} / c^1, c^2, \ldots, c^k) = 0 \quad ;$$

et ceci est vérifié pour tout couple (j,j') de stimuli pris parmi les p situations permises sélectionnées, mais aussi, compte tenu de l'exigence (A), quels que soient les stimuli j et j' choisis dans l'ensemble permis pris dans sa totalité.

Le modèle évoqué n'a pas d'existence concrète, en tant que tel. Il a pour principal intérêt de suggérer un problème d'optimisation : ayant sélectionné p stimuli, on recherche dans l'espace des variables F à n dimensions le sous-espace W de dimension k tel que les projections des variables x^j sur l'orthogonal $W^\perp$ de W soient les plus orthogonales possibles ; toute base de W définit alors un ensemble optimal de variables d'état. Bien des indices ont été proposés pour traduire les corrélations entre les résidus ; il leur correspond différentes façons de mener à bien une analyse factorielle au sens de Spearman.

Qu'implique le problème d'optimisation précédent au niveau de la sélection des stimuli ? Peut-on garantir la stabilité de la solution W ?

Quand le modèle (2) est bien adapté au phénomène étudié et en particulier quand la condition (3) peut être considérée comme vérifiée, on montre qu'il suffit que la méthode de sélection assure une certaine diversité des stimuli, ceux-ci étant tirés en assez grand nombre par rapport à k , pour que l'on retombe bien, dans la plupart des cas, sur la solution cherchée (le sous-espace W engendré par les dimensions cachées). Ce résultat bien connu des factorialistes de l'école classique, résulte des travaux de L. Guttman et explique pour une bonne part le succès du modèle de Spearman et Thurstone.

En théorie on n'a donc guère de problèmes pour sélectionner les stimuli avec ce modèle ; il suffit d'en choisir un assez grand nombre et de s'assurer de leur diversité.

En faisant intervenir des stimuli supplémentaires, la formule (3) donne un moyen simple, pour contrôler la stabilité de la solution trouvée.

2.2 *L'approche de H. Hotelling*

Une généralisation du critère décrit par Hotelling (1933) conduira à rechercher, en procédant par étapes, le système orthonormé des k variables c^{ℓ} , rendant maximum la quantité (Cf. Cailliez-Pagès 1976)

$$(4) \qquad J = \sum_{\ell} \sum_{(j,j')} m_{jj'} \, \mathrm{cov}(x^{j}, c^{\ell}) \, \mathrm{cov}(x^{j'}, c^{\ell}) \ .$$

Lorsque les termes $m_{jj'}$ définissent une matrice M symétrique définie positive, une solution de ce problème d'optimisation est fournie par l'analyse en composantes principales du triplet (X, M, D_p) ; où :

. X est le tableau (p x n) des réponses x_i^j des individus i aux stimuli j

. M est la matrice associée à la métrique choisie dans $E = R^p$ pour mesurer les proximités entre les réponses x_i

. D_p est la métrique des poids permettant de mesurer en termes de covariance les angles entre variables dans $F = R^n$.

Nous avons vu que la solution au problème d'optimisation posé, n'était acceptable que si elle respectait l'exigence (A). La solution fournie par l'analyse en composantes principales du triplet (X, M, D_p) est-elle indépendante du sous-ensemble de stimuli sélectionné ? Quelles conditions doit-on imposer à la procédure de sélection des stimuli pour qu'il en soit ainsi ?

Remarquons d'abord que l'on doit disposer d'une règle pour définir de façon systématique les nombres $m_{jj'}$; plusieurs règles sont classiques :

. $M = I_p$

On a alors : $J = \sum_{\ell} \sum_{j} \mathrm{cov}^2(x^j, c^{\ell})$

. $M = (\mathrm{diag}\ V)^{-1} = D_{1/\sigma^2}$

C'est cette règle qui avait été retenue par H. Hotelling

. $M = \mathrm{diag}\ V^{-1}$.

On a alors : $J = \sum_{\ell} \sum_{j} \frac{1}{1-r_j^2} \mathrm{cor}^2(x^j, c^\ell)$, où r_j est le coefficient de corrélation multiple entre la variable x^j et les autres variables. Avec cette règle l'introduction d'une nouvelle variable modifie le jeu des pondérations. Cette règle permet de relier l'analyse en composantes principales aux méthodes de Spearman et de Thurstone (Cailliez et Pagès, 1976).

Les composantes fournies par l'analyse en composantes principales sont des combinaisons linéaires des variables x^j ; chacune de ces variables, supposées centrées et réduites, est représentée dans F par un point de l'hypersphère de rayon unité. Assimilant variables x^j et stimuli (deux stimuli conduisant aux mêmes réponses sont considérés comme équivalents), l'univers des situations permises peut donc être représenté sur une hypershère de F de rayon unité.

Si, allant plus loin, on précise que cette représentation coïncide avec un continuum de points sur la sphère (espace probabilisé des stimuli), on dispose alors d'une méthode de tirage de stimuli respectant l'exigence (A) : les composantes principales calculées à partir des réponses aux stimuli associés à p points de l'hypersphère sélectionnés au hasard peuvent être considérées comme de bonnes approximations des composantes principales associées à la population des stimuli, si le nombre p de stimuli est choisi assez grand et si le nombre n d'individus est suffisamment grand par rapport à p (ici n peut être infini). Si donc on considère les variables comme éléments d'un espace probabilisé, on sait définir une stratégie de sélection de stimuli : le tirage au hasard, qui nous permet d'aboutir à des résultats qui peuvent être considérés comme sufisamment indépendants des stimuli effectivement sélectionnés. Pour les problèmes que pose l'univers des variables en analyse factorielle on pourra consulter S. Mulaik (1966, 1972) ou A. Pousse et J. Dauxois (1976).

Mais nous n'avons aucune connaissance mathématique précise sur l'ensemble des situations permises ; nous ne savons que les produire suivant certaines règles... et faire en sorte, en particulier, que le respect de ces règles permette d'assimiler la production des stimuli à un tirage au hasard dans une population, dont on ne pourrait se faire une idée qu'en faisant croître le nombre p de stimuli produits vers l'infini.

Voici les règles que nous nous imposons au niveau de la sélection des questions (stimuli) dans nos enquêtes psycho-sociales.

. La méthode de sélection est la même au cours des expériences (les différentes enquêtes),

. elle opère sur un vaste ensemble de situations permises très diversifiées défini suivant une procédure fixe (l'ensemble évolue d'une expérience à l'autre),

. elle est "erratique" et cela toujours de la même manière.

Ces conditions, peu précises pour le statisticien et draconiennes pour le psycho-sociologue, permettent de respecter l'exigence (A), fondamentale quand on prétend produire par l'analyse en composantes principales des dimensions ayant une signification universelle. Le laxisme au niveau de la sélection des variables qu'autorisait le modèle de Spearman n'est donc plus permis quand on retient le critère de Hotelling.

3. CONCLUSION

Les enquêtes nationales sur les conflits (système AESOP) ont permis de mettre à l'épreuve les idées exposées dans cet article ; effectuées chaque année elles s'appuient sur un questionnaire remodelé à chaque fois en respectant les règles décrites :

. l'ensemble des conflits (situations permises) est obtenu par une analyse systématique de la presse nationale effectuée un mois avant le déroulement de l'enquête. Les plus de cinq cents conflits obtenus sont rangés suivant une nomenclature fixée une fois pour toutes.

. Un groupe d'une dizaine d'experts bien au fait des enquêtes et aux pôles d'intérêt bien diversifiés sélectionne la cinquantaine de conflits, à partir desquels sont bâties les propositions sur lesquelles les interviewés se prononcent suivant une échelle d'accord en cinq paliers.

La stabilité des six dimensions factorielles significatives extraites des données recueillies, à l'aide de questionnaires qui n'ont qu'un certain nombre de questions en commun, est tout à fait remarquable. Sans être une preuve, ce résultat conforte très largement la démarche qui a été décrite puisque six enquêtes ont été réalisées depuis 1977.

BIBLIOGRAPHIE

CAILLIEZ F. et PAGES J.P. (1976). Introduction à l'analyse des données. SMASH, Paris.

J. DAUXOIS, A. POUSSE (1976). Les analyses factorielles en calcul des probabilités et en Statistique : Essai d'étude synthétique. Thèse, Université de Toulouse.

EUGENE J. (1981). Aspects de la théorie générale des systèmes. Maloine, Paris.

MULAIK S.A. (1966). Inferring the communality of a variable in a universe of variables. Psychological Bulletin Vol. 66, n° 2, 119-124.

MULAIK S.A. (1972). The foundations of Factor Analysis. McGraw-Hill, New-York.

Les structures de l'opinion publique en 1981 (3 tomes, édités par AESOP), Paris.