JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

P. THIONET

Analyse factorielle d'une activité administrative

Journal de la société statistique de Paris, tome 107 (1966), p. 114-129

<http://www.numdam.org/item?id=JSFS_1966__107__114_0>

III

ANALYSE FACTORIELLE D'UNE ACTIVITÉ ADMINISTRATIVE

L'étude dont nous confions la publication à la Société de Statistique de Paris nous semble présenter un certain intérêt pour un large public. Portant sur des données vieilles de bientôt dix ans, elle remonte à une époque où nous appartenions à l'I. N. S. E. E. et étions chargé d'études au ministère des Finances (Service d'Études Économiques et Financières). Il nous fut demandé un jour de 1958, de faire calculer par l'atelier mécanographique de la Direction de la Comptabilité publique un ensemble de coefficients de corrélations concernant l'activité des services locaux du Trésor. Ces corrélations se révélèrent beaucoup moins fortes qu'on ne le pensait; des corrélations partielles négatives tout à fait insoupçonnées apparurent même; bref il fallut tourner bride.

En pareil cas la bonne solution consiste à déplacer le problème : l'Administration centrale peut toujours demander à ses représentants des informations chiffrées plus raffinées sur leur activité; cet aspect du problème nous échappe totalement. En revanche nous pouvions chercher à interpréter par les techniques statistiques les corrélations calculées déjà. N'ayant jamais pratiqué ni étudié ni enseigné l'analyse factorielle, nous trouvâmes extrêmement plaisant de travailler avec le manuel de Faverge et d'expérimenter pas à pas les diverses méthodes de calcul décrites par cet excellent ouvrage qui, par-dessus le marché, évite de noyer le novice dans un océan de termes du jargon psychométrique.

C'est le résultat de ce travail *expérimental* que nous donnons ici, sans prendre parti sur le fond, et même sans faire usage des *notations matricielles* qui (nous nous en sommes aperçu ensuite) sont actuellement indispensables pour exposer et discuter les méthodes en concurrence (plutôt que les notations *vectorielles*). Nous voulons simplement montrer les résultats, tantôt concordants et tantôt discordants, de techniques reposant sur des hypothèses structurelles différentes, contradictoires même, et surtout gratuites.

Nous nous réservons de publier plus tard, peut-être, sur ces modèles, des observations d'ordre théorique. Ici au contraire, ce sont des explications pratiques qui paraissent nécessaires.

Et d'abord, qu'entend-on ici par *Trésor?* Il s'agit des perceptions, placées dans chaque département sous l'autorité du trésorier-payeur général (T. P. G.). A l'échelon central, c'est le directeur de la Comptabilité publique qui correspond (pour le ministre) avec les T. P. G. et non pas (comme on le croit parfois) le directeur du Trésor ou encore l'agence judiciaire du Trésor. Précisons qu'on ne dispose pas ainsi à Paris de ce que le statisticien appelle les données individuelles, mais seulement d'informations beaucoup moins « riches » : les récapitulations par département. Les calculs de corrélations pouvaient se faire alors sur machine de bureau; mais nous avons préféré employer le procédé en usage à l'I. N. S. E. E. : carrés et produits perforés sur cartes, totaux obtenus à la tabulatrice; la formule finale de la variance et de la covariance, les extractions de racine, le calcul final de la corrélation sont *manuels*. Chaque fois qu'une corrélation paraît un peu faible, on remonte la chaîne des calculs à la recherche de l'erreur possible. De nos jours on disposerait vraisemblablement de moyens électroniques.

Il était aberrant de calculer des corrélations entre des « pourcentages », variables astreintes à rester entre 0 et 100, contrairement à l'hypothèse gaussienne sous-jacente en analyse factorielle, mais nous n'avons songé qu'après coup à cette analyse. D'ailleurs c'est surtout le concept de pression fiscale qui, à la réflexion, manque de réalisme. Il était à première vue naturel de juger que le pourcentage de pénalités dont on ne faisait pas remise finalement au contribuable était un indicateur de cette pression au même titre que le pourcentage de commandements et de saisies. Une certaine expérience de la province nous fait soupçonner après coup qu'il est remis d'autant plus de pénalités qu'on les a mises à la légère : chèques de paiements égarés dans les perceptions, changements de domicile ignorés, bref imperfection des services locaux, tempérée par le coût excessif du contentieux fiscal.

Ainsi s'expliquerait que la variable P (pénalités) dont il est question plus loin ait un facteur spécifique considérable.

Achevons ce préambule en mettant le lecteur en garde contre deux erreurs fréquentes chez les non-spécialistes : analyse factorielle et plans factoriels d'expérience désignent deux techniques statistiques qui n'ont à peu près rien de commun; l'analyse factorielle est entrevue comme un programme pour cerveau électronique, fournissant un résultat incontestable, décisif : si l'étude qui suit prouve quelque chose, c'est bien le contraire.

I. LES DONNÉES DU PROBLÈME

Les trésoriers-payeurs généraux communiquent à la Comptabilité publique chaque année certaines données statistiques (en pour cent) relatives à leurs 89 département (Belfort est rattaché à la Haute-Saône ou au Haut-Rhin, peu importe). Ces données sont strictement comparables. Il est évidemment assez paradoxal de mettre sur le même plan les départements de la Seine et des Basses-Alpes, de ne pas « *pondérer* » les données; mais le but de ces données

est de permettre la comparaison des départements entre eux; et nous ferons semblantd'ignorer que ceux-ci sont de tailles inégales.

Données centrées et normées

Considérons par exemple le pourcentage de recouvrements des rôles effectués dans les délais; c'est l'une des données à analyser. Nous disposons des 89 données R s'échelonnant (disons) de 50 % à 90 % (notamment pour 1955, 1956).

Faisons la moyenne (non pondérée) de ces 89 nombres; soit 73 %.

Retranchons 73 de chacun des R, ce qui donne les *données centrées* R', les unes négatives les autres positives, leur total algébrique étant zéro.

Changeons alors l'échelle des R' en les remplaçant par des r qui leur soient proportionnels mais tels que la somme des r^2 soit égale à 88; c'est ce que nous appellerons les *données normées.*

Il est facile de comprendre leur signification. Les recouvrements R dans les délais s'échelonnaient entre 50 et 90 %, les proportions C de commandements lancés contre les contribuables retardataires allaient disons de 5 à 20 % et les proportions S de saisies de 0,2 à 1,5 % (nous donnons là des chiffres hypothétiques et non pas les chiffres réels). Il était très mal commode dans ces conditions de comparer entre eux les R, les C et les S, et il n'était pas tellement plus facile de comparer les R', C' et S'. Au contraire les r, c et s ont en commun leurs moyennes égales à 0 et leur somme de carrés égales à 88.

Il est inutile de chercher à justifier ici le nombre 88 lui-même; pourquoi pas 100? pourquoi pas plutôt 89, c'est-à-dire le nombre de départements? En réalité cela n'a aucune importance; c'est la conséquence de théories beaucoup plus générales et qui resteront à l'arrière-plan de nos préoccupations.

Dans ce qui suit il sera exclusivement question des données *normées.*

Notations

La direction de la Comptabilité publique disposait pour chaque département les données suivantes :

D : pourcentage de recouvrement dans les délais,

C : pourcentage de commandements,

S : pourcentage de saisies,

P : pourcentage de pénalités dont la remise a été refusée,

R : pourcentage de rentrées après expiration de délais et jusqu'au 31 décembre de l'année suivante,

D + R : Pourcentage de rentrées au total (toujours légèrement inférieur à 100 %).

On avait ainsi une physionomie de l'action et des résultats des services du Trésor dans chaque département, tant pour les rôles émis en 1955 que pour ceux émis en 1956. Cette documentation de base était bien entendu *confidentielle* mais ceci ne nous empêchera pas de faire état des corrélations calculées entre ces variables.

	Matrice des corrélations 1955					Matrice des corrélations 1956				
	d	*c*	*s*	*p*	*r*	*d*	*c*	*s*	*p*	*r*
d	1					1				
c	— 0,614	1				— 0,730	1			
s	— 0,581	0,920	1			— 0,705	0,922	1		
p	— 0,324	0,443	0,379	1		— 0,269	0,394	0,435		
r	— 0,970	0,495	0,472	0,253	1	— 0,939	0,556	0,547	0,195	1

Initialement l'inspecteur des Finances avait seulement demandé qu'on calculât ces corrélations; mais celles-ci sont apparues en fait assez faibles. A l'action répressive du Trésor, représentée par les variables C, S, P, déjà peu corrélées (car dans tel département on préfère agir avec C et dans tel autre avec P) s'ajoutent des facteurs variés; essentiellement :

facteurs locaux (esprit frondeur de certains départements, prospérité ou calamités dans d'autres, action persuasive variable des services du Trésor); et facteurs propres à l'année considérée.

Seule l'analyse de données portant sur de nombreuses années permettrait de tenir compte des seconds; or on ne disposait que de deux années; et l'interprétation des données va reposer pour une bonne part sur des comparaisons subjectives (peu scientifiques) entre ces 2 années.

En revanche il semblait possible d'assimiler à des variations aléatoires la masse des facteurs locaux qui créent les différences observées entre les 89 départements.

Ainsi nous avons pensé qu'il convenait, non pas de se borner au calcul des corrélations, mais d'essayer de les interpréter à l'aide d'un modèle *annuel.*

Une autre idée aurait pu être retenue : établir *un modèle séquentiel* 1956-1955 qui expliquerait les résultats *Rôles de 1956* non seulement par les données *Rôles de 1956* mais également en partie par celles de 1955. Toutefois il a semblé difficile de faire sérieusement cas de ces seuls résultats; il eut fallu au minimum disposer des données relatives aux rôles de 1957; on aurait pu alors construire un modèle 1957-1956, le comparer à celui de 1956-1955 et chercher à interpréter les changements constatés.

II. UNE PREMIÈRE SORTE D'INTERPRÉTATION : L'ÉQUATION DE RÉGRESSION

L'analyse statistique nous propose divers moyens. Tout d'abord on a pensé à la théorie des *corrélations partielles.*

Il est à noter qu'historiquement, cette théorie est la première qu'on ait appliquée (avec Galton et Pearson) sur les données de nature psychologique, vers 1900 (les résultats en furent décevants).

Travaillant avec les données *normées* et employant la méthode des moindres carrés (comme ils le firent) nous avons trouvé ainsi que les *résultats* des services du Trésor suivaient le modèle suivant :

Équation dite de régression :

$$1955 \qquad r = [0{,}593\, c + 0{,}174\, s + 0{,}079\, p] - 0{,}479\, d + 0{,}28\, \zeta$$
$$1956 \qquad r = [0{,}516\, c + 0{,}340\, s - 0{,}065\, p] - 0{,}349\, d + 0{,}48\, \zeta$$

La variable ζ est une variable aléatoire normée; et le terme qui la renferme représente la partie des rentrées d'impôt r que la connaissance des variables $c\, s\, p\, d$, ne permet pas d'expliquer. Cette part est très petite pour 1955 et encore assez faible pour 1956 comme le montre *l'indice de corrélation totale* qu'on définit comme

$$\sqrt{1 - (0{,}28)^2} = 0{,}96 \text{ en } 1955$$
$$\sqrt{1 - (0{,}48)^2} = 0{,}88 \text{ en } 1956$$

Ainsi l'équation de régression amputée du terme en ζ représenterait convenablement les données (surtout celles de 1955). L'ennui est que cette équation ait pour 1956 des coefficients en s et p aussi profondément différents de ceux de 1955; cette instabilité du modèle lui ôte tout son intérêt prévisionnel.

Essayons maintenant d'interpréter ce type d'équation du point de vue fiscal.

La quantité entre crochets représente « l'action répressive » du Trésor. Cette action semble rencontrer une « résistance » représentée par le terme soustractif; mais c'est là une interprétation incorrecte.

Si l'on considère les rentrées totales d'impôts R + D = T (avant et après échéance bloquées) les mêmes équations de régression prennent l'expression suivante :

$$1955 \begin{cases} t = 3{,}250\, r + 2{,}453\, d \\ \quad = 1{,}927\, c + 0{,}565\, s + 0{,}257\, p + \mathit{0{,}986}\, d + 0{,}91\, \zeta \end{cases}$$

$$1956 \begin{cases} t = 2{,}486\, r + 1{,}823\, d \\ \quad = 1{,}283\, c + 0{,}845\, s - 0{,}162\, p + \mathit{0{,}978}\, d + 1{,}19\, \zeta. \end{cases}$$

Cette fois on voit que les départements où d est grand sont aussi ceux qui payent le mieux (t grand), à supposer que l'action répressive soit partout la même. Il s'agit des départements où l'état d'esprit est « bon » et où les difficultés économiques sont moindres qu'ailleurs [1]; ces départements ont payé presque tous leurs impôts avant échéance (valeurs négatives de r). Inversement, les départements « mauvais payeurs » (d négatifs) font des rentrées après échéances plus grandes que les autres, à supposer que l'action répressive (c, s, p) soit la même partout.

On notera le changement de signe du coefficient de p entre 1955 et 1956. Par suite pour agir dans le sens le plus favorable au Trésor, il aurait fallu (semble-t-il) maintenir le plus possible de pénalités en 1955, mais en remettre le plus possible en 1956. D'ailleurs les conséquences de ces pénalités restent très médiocres, vu le faible coefficient dont est affectée (dans les 2 cas) la variable p. Le modèle n'a donc guère d'intérêt et nous en chercherons un autre.

A part quelques vétérans des luttes du début du siècle (tel Burt) [2], personne ne tient la représentation des données par une formule de régression pour de l'analyse factorielle. Nous allons voir un second mode d'interprétation des mêmes données, *par une analyse des composantes*. Beaucoup pensent qu'il s'agit bien cette fois d'analyse factorielle, encore que telle ne soit pas l'opinion du professeur Kendall [3] ni la nôtre.

III. UNE NOUVELLE INTERPRÉTATION : L'ANALYSE DES COMPOSANTES [4]

Il s'agit simplement de remplacer les n variables *normées* représentant les données brutes (ici $c\ s\ p\ d\ r$, ou $d + r = t$) par n variables, *normées* elles aussi mais *indépendantes* entre elles ($: f\ g\ h\ k\ j$).

Ceci veut dire que deux *quelconques* d'entre elles, soit f et g, vérifient la relation

$$f \cdot g = f_1 \cdot g_1 + f_2 \cdot g_2 + \ldots + f_{89} \cdot g_{89} = 0$$

1. Le « poujadisme » battait son plein en 1955.
2. BURT C., *The ai ms and methods of factorial analysis* (Dans hommage à Henri Pieron, l'Année Psychologique, 1951, p. 61).
3. KENDALL et LAWLEY, *The principles of factor analysis*, J. Royal Stat. Soc., A 1, 1956.
4. Ce n'est pas un néologisme. Le terme est employé par : Hotelling, Kendall, Faverge, pourquoi en chercher un autre?

Il existe en fait diverses méthodes qui aboutissent à ce mode de présentation, pour la simple raison que le problème mathématique ainsi posé admet une infinité de solutions. Et le critère qu'on doit ajouter aux conditions précédentes pour fixer son choix et calculer *une certaine solution* du problème, n'est pas de ceux qui s'imposent sans hésitation à l'esprit.

Personnellement nous avons employé la méthode « centroïde » (Burt, Thurstone) qui a au moins le mérite de permettre des calculs (*manuels*) simples et précis. Il existe également une belle méthode due à Hotelling qui (pour nos calculs *manuels*) est moins précise et plus longue, nécessitant le recours à des approximations successives; les calculateurs électroniques la suivent couramment.

Dans les divers cas, on détermine *successivement* les facteurs [1] $f, g, h \ldots j$ (1^{er}, 2^e, 3^e ... n^e facteurs), de telle façon que *f rende compte* le mieux possible des données, — puis que g rende compte le mieux possible des données qui *ne sont pas expliquées* par f, — puis que h, etc., — et ainsi de suite jusqu'à épuisement. Il y a en général autant de facteurs que de variables ($n = 5$ ici). Alors variables et facteurs sont *fonctions linéaires les uns des autres.* Il peut pourtant arriver qu'il y ait moins de facteurs que de variables.

Dans le cas présent, voici ce que la méthode centroïde a donné. (Nous avons remplacé D par 100 — D, c'est-à-dire le déficit à l'instant de l'échéance; d devient ainsi — d et il n'y a plus de corrélations négatives) :

1955

$$
\begin{aligned}
c &= 0{,}871\,f - 0{,}328\,g - 0{,}310\,h - 0{,}133\,k + 0{,}141\,j \\
s &= 0{,}841\,f - 0{,}333\,g - 0{,}372\,h + 0{,}198\,k - 0{,}071\,j \\
p &= 0{,}602\,f - 0{,}404\,g + 0{,}682\,h - 0{,}063\,k - 0{,}071\,j \\
-d &= 0{,}875\,f + 0{,}474\,g - 0{,}029\,h - 0{,}065\,k - 0{,}071\,j \\
r &= 0{,}800\,f + 0{,}592\,g + 0{,}030\,h + 0{,}065\,k + 0{,}071\,j
\end{aligned}
$$

1956

$$
\begin{aligned}
c &= 0{,}890\,f - 0{,}228\,g - 0{,}343\,h + 0{,}028\,k + 0{,}194\,j \\
s &= 0{,}891\,f - 0{,}265\,g - 0{,}299\,h + 0{,}091\,k - 0{,}196\,j \\
p &= 0{,}566\,f - 0{,}501\,g + 0{,}648\,h - 0{,}094\,k - 0{,}000\,j \\
-d &= 0{,}900\,f + 0{,}411\,g - 0{,}076\,h - 0{,}123\,k + 0{,}013\,j \\
r &= 0{,}799\,f + 0{,}583\,g + 0{,}075\,h + 0{,}125\,k - 0{,}021\,j
\end{aligned}
$$

Indiquons quelques propriétés des coefficients figurant dans ces formules.

On remarquera d'abord (et c'est essentiel) que la somme des carrés des coefficients de chaque ligne est *égale à 1;* par exemple

$$(0{,}871)^2 + (0{,}328)^2 + \ldots + (0{,}141)^2 = 0{,}999\,895$$

En outre on s'apercevra que pour tout couple de lignes, la somme des produits de coefficients homologués est égale au coefficient de corrélation correspondant; par exemple pour les 2 premières lignes (1955), on a :

$$0{,}871 \times 0{,}841 + 0{,}328 \times 0{,}333 + \ldots = 0{,}920710$$

c'est-à-dire pratiquement 0,920 (corrélation entre c et s en 1955).

1. On note bien entendu que ces « facteurs » s'ajoutent; ils n'ont rien à voir avec les facteurs de l'algèbre qui se multiplient entre eux.

Bien entendu les n^2 coefficients (de *fgh* ... dans *csp* ...) vérifient ainsi $n + n\ (n-1)/2 = n\ (n+1)/2$ conditions, ce qui ne suffit pas à les déterminer. On y ajoute d'autres conditions.

La méthode « centroïde » employée ici est conçue pour donner des coefficients dont les sommes verticales des valeurs absolues sont aussi grandes que possible (une autre méthode consiste à maximiser les sommes verticales des carrés de coefficients).

On remarque que les gros coefficients sont très stables, lorsqu'on passe de 1955 à 1956; tandis que les petits coefficients varient relativement beaucoup. C'est un signe que, s'il existe quelque structure analogue stable, elle doit correspondre à des *valeurs nulles* pour ces petits coefficients. Bref on aurait envie de remplacer la matrice des 5×5 coefficients ci-dessus par une matrice où il y aurait beaucoup de zéros dans les colonnes de droite; mais en arrondissant ainsi certains coefficients à 0, on ne retrouverait plus ni 1 comme somme de leurs carrés, ni les corrélations comme sommes de leurs produits.

Par exemple (1re ligne 4e facteur) $ni - 0{,}133$ (1955) $ni + 0{,}028$ (1956) ne signifient rien; il faudrait pouvoir mettre zéro. Quant aux coefficients du 5e facteur (1956) pour p, $-d$ et r, leur signe lui-même est incertain; seuls ceux de c et s ont une signification.

Pourtant nous avons fait tous les calculs avec un nombre effectif de décimales très supérieur à 3. Si l'on songe à la précision réelle de certaines données (pourcentage de saisies surtout) on doit bien admettre que les équations ci-dessus sont un peu illusoires.

Analyse des composantes de 3 ou 4 variables seulement.

Avant de calculer les systèmes de 5 équations ci-dessus, nous nous étions avisé de faire l'analyse des composantes des 3 variables *c s p* (action répressive du Trésor) puis des 4 variables *c s p r*. Chaque fois on constate l'existence du 1er facteur à coefficients tous grands et positifs; on l'appellera *f*. Ensuite le 2e facteur est affecté de signes opposés pour le couple (*c s*) (poursuites du Trésor) d'une part et le couple (*p r*) ou encore *p* seul; ce caractère nous le fait appeler *m* (et non pas *g* comme s'il s'agissait du *2e facteur* de (*c s p d r*).

Pour (*c s p*) il existe un 3e facteur (faible) commun à *c* et *s* (non à *p*) : soit *j*.

Dans l'analyse de (*c s p r*) apparaît un *3e facteur* m' (*fort*) *commun à* p *et* r (la part de *c* et *s* est négligeable) et un 4e facteur (faible) commun à tous.

$$
\begin{aligned}
1955 &\begin{cases} c = 0{,}928\,f - 0{,}316\,m + 0{,}196\,j \\ s = 0{,}903\,f - 0{,}382\,m - 0{,}196\,j \\ p = 0{,}716\,f - 0{,}699\,m + 0\,j \end{cases} \\
1956 &\begin{cases} c = 0{,}908\,f - 0{,}370\,m + 0{,}149\,j \\ s = 0{,}924\,f - 0{,}327\,m - 0{,}149\,j \\ p = 0{,}717\,f + 0{,}696\,m \quad\ 0\,j \end{cases} \\
1955 &\begin{cases} c = 0{,}907\,f - 0{,}371\,m - 0{,}074\,m' + 0{,}183\,j \\ s = 0{,}880\,f - 0{,}433\,m + 0{,}074\,m' - 0{,}183\,j \\ p = 0{,}659\,f + 0{,}441\,m - 0{,}581\,m' + 0{,}183\,j \\ r = 0{,}705\,f + 0{,}363\,m + 0{,}581\,m' - 0{,}183\,j \end{cases} \\
1956 &\begin{cases} c = 0{,}904\,f - 0{,}380\,m + 0{,}075\,m' + 0{,}182\,j \\ s = 0{,}914\,f - 0{,}355\,m - 0{,}075\,m' - 0{,}182\,j \\ p = 0{,}637\,f + 0{,}447\,m - 0{,}601\,m' - 0{,}182\,j \\ r = 0{,}723\,f + 0{,}288\,m + 0{,}601\,m' + 0{,}182\,j \end{cases}
\end{aligned}
$$

Rappel 1955 —

$$\begin{cases} c = 0{,}871\, f - 0{,}322\, g - 0{,}310\, h - 0{,}133\, k + 0{,}141\, j \\ s = 0{,}841\, f - 0{,}333\, g - 0{,}372\, h + 0{,}198\, k - 0{,}071\, j \\ p = 0{,}602\, f - 0{,}404\, g + 0{,}682\, h - 0{,}063\, k - 0{,}071\, j \leftarrow \\ d = 0{,}875\, f + 0{,}474\, g - 0{,}029\, h - 0{,}065\, k - 0{,}071\, j \\ r = 0{,}800\, f + 0{,}592\, g - 0{,}030\, h + \text{p},065\, k + 0{,}071\, j \end{cases}$$

Rappel 1956 —

$$\begin{cases} c = 0{,}890\, f - 0{,}228\, g - 0{,}343\, h + 0{,}028\, k + 0{,}194\, j \\ s = 0{,}891\, f - 0{,}265\, g - 0{,}299\, h + 0{,}091\, k - 0{,}186\, j \\ p = 0{,}566\, f - 0{,}501\, g + 0{,}648\, h - 0{,}094\, k - 0{,}000\, j \\ d = 0{,}900\, f + 0{,}411\, g - 0{,}076\, h - 0{,}123\, k + 0{,}013\, j \\ r = 0{,}799\, f + 0{,}583\, g + 0{,}075\, h + 0{,}125\, k - 0{,}021\, j \end{cases}$$

On est frappé par la permanence des coefficients de *f*, le « *facteur général* », et par contraste on s'étonne de voir surgir un facteur *m'*, puis le couple *m m'* disparaître aux dépens d'un couple *g h* dont les coefficients n'ont plus les mêmes signes. On est fondé à y voir la preuve que ces 2e et 3e facteurs ne signifient rien de concret.

Lorsqu'on analyse une vingtaine de variables au lieu de 4 ou 5, il y a de grandes chances pour que le nombre de facteurs communs *présentant* une *réelle consistance* ne dépasse pas quelques unités. En psychométrie, le grand mérite de Spearman (1904) est d'avoir mis en lumière la présence dans la plupart des tests *d'un certain facteur commun* (qu'il appelle *g*); et il a même semblé qu'on n'en pouvait trouver qu'un seul; à présent, on sait trouver un 2e, un 3e facteur commun, mais par des méthodes dont les résultats ne sont pas toujours très concordants.

Nous allons donc abandonner cette *Analyse des Composantes* pour chercher autre chose :

donnant des équations *plus simples*,
moyennant une théorie *plus délicate*.

Mais avant de changer de chapitre, notons ceci :

Nous avons là un système de n (= 5) équations à n inconnues. On peut résoudre ce système en $f\ g\ k\ j$; et écrire (*inversion de matrices*)

$$f = A_1\, c + A_2\, s + A_3\, p + A_4\, d + A_5\, r$$
$$g =$$
etc.

(Nous ne calculerons pas numériquement ici les coefficients A_1 A_2, etc.).

On peut donc *calculer* pour chaque département, connaissant les valeurs de *c s p d r*, *les valeurs* correspondantes des facteurs $f\ g \ldots j$.

Nous ne pourrons plus en faire autant avec les analyses factorielles au sens strict. Chaque département n'aura plus alors *ses facteurs à lui*.

IV. PRINCIPE DE L'ANALYSE FACTORIELLE PROPREMENT DITE

On pourrait croire qu'il est plus facile d'interpréter les données quand, au lieu d'un système de n équations à n facteurs dont les n^2 coefficients se calculent pas à pas, on écrit

un système de n équations à $n + 1$, ou $n + 2$, ou $n + 3$, etc., facteurs tout en décidant à l'avance que certains coefficients (au nombre de N) seraient nuls.

C'est ce qu'on fait depuis Spearman (analyse factorielle). Or ceci revient à écrire en fait

N équations de plus,

avec n, $2\,n$, $3\,n$ ... inconnues de plus (les cœfficients).

Au lieu d'avoir un problème en partie *indéterminé*, on arrive donc très vite à un problème en général *impossible*, qu'on ne peut résoudre exactement qu'avec des données numériques très particulières et qu'on résoud d'une façon plus ou moins approchée avec des données numériques réalistes.

V. LE SCHÉMA DE SPEARMAN

Le schéma le plus simple (celui de Spearman) est bien entendu à $n + 1$ *facteurs*, c'est-à-dire :

un facteur *commun* à toutes les variables,
plus un facteur *spécifique* pour chaque variable.

Mais nous allons voir que c'est un schéma simpliste et peu réaliste : car il ne s'adapte plus aux données dès qu'on suppose les facteurs spécifiques indépendants entre eux comme le fait Spearman.

Alors la matrice de $n\,(n + 1)$ coefficients renferme $n\,(n + 1)$ zéros et seulement $2\,n$ coefficients inconnus (non nuls), pour $n + n\,(n - 1)/2$ conditions. Pour $n = 3$ il y a autant de conditions que d'inconnues mais ce ne sont pas des conditions linéaires. Pour $n > 3$, il y a trop de conditions pour qu'on puisse les satisfaire toutes.

Voici le schéma que nous avons calculé (par un procédé de Spearman) avec les données de 1955 :

$$
\begin{aligned}
c &= 0{,}863\,f + 0{,}505\,u \\
s &= 0{,}802\,f + 0{,}597\,v \\
p &= 0{,}483\,f + 0{,}663\,w \\
-d &= 0{,}865\,f + 0{,}502\,x \\
r &= 0{,}713\,f + 0{,}701\,y
\end{aligned}
$$

f est le facteur *commun*.
$u\ v\ w\ x\ y$ sont les facteurs *spécifiques*.

Les facteurs sont des variables *normées*. Les coefficients de $u\ v \ldots y$ (ou *spécifités*) sont calculés de façon que f soit strictement *indépendant* de u, de v, ... de y; c'est-à-dire qu'on ait *formellement* [1] :

$$c^2 = (0{,}863)^2 + (0{,}550)^2 = 1 \qquad \text{et analogues}$$

1. Donnons ici le calcul formel *sans autre justification* :

$$
\begin{aligned}
c^2 &= (0{,}863\,f + 0{,}505\,u)^2 \\
&= 0{,}745\,f^2 + 0{,}255\,u^2 \,(+\,0{,}871\,fu) = 1 \\
cs &= (0{,}863\,f + 0{,}505\,u)\,(0{,}802\,f + 0{,}597\,v) \\
&= 0{,}692\,f^2 + (0{,}405\,uf + 0{,}515\,fv + 0{,}302\,uv) = 0{,}692
\end{aligned}
$$

L'indépendance de f et u s'écrit ... $f\,u = 0$
f et v s'écrit ... $f\,v = 0$
u et v s'écrit ... $u\,v = 0$
} les facteurs étant normés on a $\left\{ \begin{array}{l} f^2 = 1 \\ u^2 = 1 \\ v^2 = 1 \end{array} \right.$

Les lettres représentent des vecteurs de l'espace à 89 dimensions (axes orthonormés).
Le détail du calcul est expliqué en *Annexe*.

Les coefficients de f (ou *communautés*) sont calculés de façon que u et v, u et w, etc., soient autant que possible *indépendants*. S'ils l'étaient strictement, on devrait avoir formellement (1) :

$$\text{CS} = \text{Corrélation de } c \text{ et } s = 0{,}863 \times 0{,}802 = 0{,}692$$
$$\text{et non pas } 0{,}920$$

Mettons côte à côte le tableau des corrélations données et celui des corrélations ainsi recalculées qui correspondraient à l'hypothèse de Spearman (données de 1955).

	Corrélations données					Corrélations recalculées				
	c	s	p	$-d$	r	c	s	p	$-d$	r
c	—					—				
s	0,920					0,692	—			
p	0,443	0,379	—			0,365	0,339	—		
$-d$	0,614	0,581	0,324	—		0,748	0,692	0,366	—	
r	0,495	0,472	0,253	0,970	—	0,616	0,571	0,302	0,616	—

Il est clair que ces deux tableaux sont profondément différents et que nous devrons renoncer à l'analyse factorielle de nos données suivant le schéma de Spearman, c'est-à-dire

$$\left.\begin{array}{c}\text{en 1 facteur commun}\\ \text{et } n \text{ facteurs spécifiques}\end{array}\right\} \text{ indépendants}$$

Remarque 1. — Lorsqu'on analyse les résultats de tests psychométriques, les données sont affectées d'une très grande incertitude; et il est possible alors qu'on hésite à rejeter catégoriquement le schéma de Spearman dans des cas analogues, en mettant sur le compte des erreurs de mesure les différences qui existent entre les 2 tableaux de corrélations.

Remarque 2. — Les 3 données (c s p) relatives à l'action répressive des services du Trésor sont dans ce qu'on appelle le « *cas de Heywood* ». On veut dire par là que, bien qu'il y ait alors autant d'équations que d'inconnues, il est impossible de calculer un facteur f commun à c s p, parce que le carré du coefficient de f dans c (ou communauté)

$$\frac{0{,}920 \times 0{,}443}{0{,}379}$$

est plus grand que 1, de sorte que le calcul de la *spécifité* (coefficient de g) a pour résultat un nombre *imaginaire*.

Autrement dit : $0{,}379 - 0{,}920 \times 0{,}443$ est négatif, ce qui signifie que la corrélation *partielle* de p et s (c donné) est négative.

Remarque 3. — Les données (c s p r) sans tenir compte de d) ne s'opposent pas à l'hypothèse de l'existence du facteur commun unique. En effet les 2 tableaux ci-dessous se ressemblent fort :

	Corrélations données				Corrélations recalculées		
	c	s	p		c	s	p
s	0,920	—	—	s	0,890	—	
p	0,443	0,379	—	p	0,442	0,400	—
r	0,495	0,472	0,253	r	0,519	0,471	0,234

Ainsi lorsqu'on fait abstraction de la variable d (ou $-d$), l'analyse factorielle de Spearman fournit un schéma qui s'adapte assez bien aux données.

1. Voir note page 42.

Exemple : Données 1955.

$$
\begin{aligned}
c &\sim 0{,}990\, f + 0{,}138\, u \\
s &\sim 0{,}899\, f + 0{,}438\, v \\
p &\sim 0{,}446\, f + 0{,}895\, w \\
r &\sim 0{,}524\, f + 0{,}852\, y
\end{aligned}
$$

On a mis $\sim$ au lieu de $=$ parce que l'égalité n'est qu'approximative.

Il est à noter que le facteur général f apparaît essentiel dans c et s (commandements, saisies) mais ne joue qu'un rôle secondaire dans p et r (pénalités, paiements tardifs).

VI. L'ANALYSE FACTORIELLE AVEC 2 FACTEURS COMMUNS INDÉPENDANTS

Expliquer les mêmes données par un *couple* de facteurs communs (f, g) indépendants (et n facteurs spécifiques indépendants des précédents) est alors une solution particulièrement tentante.

1. Écrivons notre système d'analyse factorielle :

$$
\begin{aligned}
c &= \mathrm{A}\, f + \mathrm{A}'\, g + \mathrm{A}''\, u \\
s &= \mathrm{B}\, f + \mathrm{B}'\, g + \mathrm{B}''\, v \\
p &= \mathrm{C}\, f + \mathrm{C}'\, g + \mathrm{C}''\, w \\
-d &= \mathrm{D}\, f + \mathrm{D}'\, g + \mathrm{D}''\, x \\
r &= \mathrm{E}\, f + \mathrm{E}'\, g + \mathrm{E}''\, y.
\end{aligned}
$$

Les lettres $u\ v\ w\ x\ y$ désignent encore les facteurs *spécifiques* (indépendants entre eux et indépendants de f et g).

Les inconnues sont : A A′ A″ ... E″. Il y en a 15 en tout.

On leur impose 15 conditions, à savoir :

— 10 équations à 10 inconnues AA′, BB′, ... EE′ du type $\mathrm{AB} + \mathrm{A'B'} = (cs)$ (corrélation entre c et s),

— et 5 équations donnant les 5 inconnues A″ B″ ... E″ (lorsqu'on connaît les 10 autres du type $\mathrm{A''^2} = 1 - \mathrm{A^2} - \mathrm{A'^2}$.

Sous réserve qu'on ne rencontre pas de cas d'Heywood, c'est-à-dire que jamais $\mathrm{A^2} + \mathrm{A'^2}$ (par exemple) ne dépasse l'unité, on pourrait espérer que ce système d'équation ait une solution valable. Mais en étudiant le problème, on s'aperçoit *qu'en général* il n'a pas de solution; et que lorsqu'il en a une, il en a une *infinité*. Ceci tient au fait que l'équation $\mathrm{AB} + \mathrm{A'B'} = cs$ équivaut à $a.\ b.\ \cos\,(a, b) = cs$

avec $$a^2 + \mathrm{A'^2} = a^2, \qquad \mathrm{B^2} + \mathrm{B'^2} = b^2$$

AA′ étant les composantes d'un vecteur a, et BB′ celles d'un vecteur b. Ainsi nos 10 équations ne déterminent les 5 vecteurs $a\ b\ c\ d\ e$ *qu'à une rotation près* (si toutefois elles sont compatibles, ce qui n'est pas le cas en général).

Ainsi le calcul *de composantes* AA′ A″, etc., avec des équations de la forme ci-dessus est *impossible* ou *indéterminé* — sans juste milieu. Ceci reste vrai si l'on augmente le nombre de facteurs indépendants communs.

2. On est conduit à essayer de résoudre le même problème de façon *approximative;* c'est-à-dire à calculer des coefficients AA′ A″, etc., tels que l'on ne retrouve peut-être pas *exactement* la matrice des corrélations initiales mais une matrice qui en diffère le moins possible.

On a vu que, avec un seul facteur, cela ne donnerait pas des résultats très brillants· Comment va-t-on s'y prendre avec 2 facteurs?

Il y a une infinité de façon de faire.

Mais il est assez *naturel* et en tous cas *économique* de conserver comme coefficients du premier facteur ceux qu'on a calculés par la méthode de Spearman, et de calculer les coefficients du second facteur par la méthode même qui a servi au premier mais en opérant sur un tableau des résidus des corrélations. Ce n'est d'ailleurs pas aussi *empirique* qu'on pourrait le croire à première vue :

Quand on fait des calculs sur des *variables réduites convenables*, il arrive (et c'est le cas ici) qu'en retranchant une corrélation d'une autre corrélation, on trouve encore une corrélation (qu'on n'aille pas croire que cette affirmation soit valable en tout état de cause !).

Nous formons donc la matrice des corrélations résiduelles, c'est-à-dire (pour les données 1955) :

	c	*s*	*p*	$-d$	*r*
c	—				
s	0,228	—			
p	0,078	0,040	—		
$-d$	— 0,134	— 0,111	— 0,042	—	
r	— 0,221	— 0,099	— 0,049	0,354	—

Et nous lui appliquons toujours la même règle de Spearman (*cf. Annexe*).

Il vient cette fois (toujours pour 1955) :

$$
\begin{aligned}
c &\sim 0{,}863\,f - 0{,}474\,g + 0{,}175\,u \\
s &\sim 0{,}802\,f - 0{,}295\,g + 0{,}519\,v \\
p &\sim 0{,}483\,f - 0{,}373\,g + 0{,}792\,w \\
-d &\sim 0{,}865\,f + 0{,}421\,g + 0{,}273\,x \\
r &\sim 0{,}713\,f + 0{,}515\,g + 0{,}475\,y
\end{aligned}
$$

(L'ajustement n'étant pas parfait le symbole $\sim$ remplace le signe $=$).

L'imperfection du modèle se manifeste à la fois par des coefficients encore massifs affectés aux facteurs *spécifiques* et par des divergences encore notables entre matrice recalculée et matrice de corrélations donnée.

	Corrélations données				Corrélations recalculées			
	c	*s*	*p*	$-d$	*c*	*s*	*p*	$-d$
s	0,920	—			0,832	—		
p	0,443	0,379	—		0,541	0,449	—	
$-d$	0,614	0,581	0,324	—	0,548	0,568	0,209	—
r	0,495	0,472	0,253	0,970	0,372	0,419	0,110	0,833

Rien ne nous empêcherait de nous attaquer alors à la 2e matrice résiduelle.

	c	*s*	*p*	— *d*
s	0,088	—		
p	— 0,098	— 0,070	—	
— *d*	+ 0,066	+ 0,013	+ 0,115	—
r	+ 0,023	— 0,053	+ 0,143	+ 137

et d'en rendre compte (dans la mesure du possible) par un 3e facteur commun. Mais nous en resterons là pour cette méthode.

3. *Autre procédé.*

Nous avons expérimenté également l'une des méthodes d'approximations successives, celle où l'on estime les « spécifités » inconnues par la plus grande des corrélations de la même colonne; c'est-à-dire qu'avec les données 1955, par exemple :

$$m - IK = \begin{bmatrix} (1 - A'')^2 & 0{,}920 & 0{,}443 & 0{,}614 & 0{,}495 \\ 0{,}920 & (1 - B''^2) & 0{,}379 & 0{,}581 & 0{,}472 \\ 0{,}443 & 0{,}379 & (1 - C''^2) & 0{,}324 & 0{,}253 \\ 0{,}614 & 0{,}581 & 0{,}324 & (1 - D''^2) & 0{,}970 \\ 0{,}495 & 0{,}472 & 0{,}253 & 0{,}970 & (1 - E''^2) \end{bmatrix}$$

on remplace $1 - A''^2$, $1 - B''^2$, $1 - C''^2$, $1 - D''^2$, $1 - E''^2$ respectivement par 0,920; 0,920; 0,443; 0,970; 0,970.

moyennant quoi on suivait le même processus que pour l'analyse des composantes (ci-dessus, p. 115) pour déterminer les composantes du 1er facteur, lesquelles « expliquent » une partie des termes de la matrice précédente; on en déduit une matrice résiduelle, dont les termes de la 1re diagonale sont douteux et qu'on remplace encore par les plus grands de la colonne, etc.

Nous trouvons ainsi (comparer au résultat précédent)

$$\begin{aligned} c &\sim 0{,}872\, f - 0{,}399\, g + 0{,}284\, u \\ s &\sim 0{,}841\, f - 0{,}376\, g + 0{,}389\, v \\ p &\sim 0{,}474\, f - 0{,}228\, g + 0{,}851\, w \\ -d &\sim 0{,}889\, f + 0{,}458\, g + 0\, x \\ r &\sim 0{,}812\, f + 0{,}553\, g + 0{,}184\, y \end{aligned}$$

Il est bien entendu que ce ne sont que des valeurs grossièrement approchées des coefficients à trouver. Quelle est la matrice de corrélations de ce schéma?

Corrélations recalculées				Corrélations données			
—				—			
0,883	—			0,920	—		
0,504	0,484	—		0,443	0,379	—	
0,592	0,575	0,317	—	0,614	0,581	0,324	—
0,487	0,475	0,258	0,975	0,495	0,472	0,253	0,970

Sans crier au miracle, il est impossible de nier que la matrice des corrélations recalculée est cette fois *bien plus proche* de celle proposée que la matrice obtenue au n° 2. Il y a donc certainement des raisons pratiques très sérieuses pour abandonner les méthodes faciles à comprendre et leur substituer des procédés délicats (nous voulons dire que la raison pour laquelle pareilles approximations doivent converger vers les vraies valeurs et notamment pour les spécifités n'a rien d'évident en soi).

VII. LE PROCÉDÉ DES FACTEURS OBLIQUES

Bien que nous ne cherchions pas à être complet, il nous faut encore signaler le procédé des facteurs obliques.

Thurstone qui fut le principal champion de la polyfactorisation a imaginé des schémas à *plusieurs facteurs communs* qui ne sont plus nécessairement *indépendants* l'un de l'autre.

Par exemple on peut avoir intérêt, quand on analyse des données anthropométriques, à adopter la Taille et le Poids comme facteurs des autres mensurations, tout en sachant pertinemment que ces 2 données ne sont pas indépendantes; rien n'empêche de leur substituer ultérieurement un couple de facteurs indépendants dépourvus de toute signification concrète. Ainsi l'intérêt que présentent les facteurs non indépendants ou comme on dit *obliques*, c'est au fond qu'on peut les *choisir* et même leur donner une signification physique.

La variable $(-d)$ aurait de bonnes raisons de figurer comme *facteur* f dans notre analyse, car c'est compte tenu des rentrées aux échéances que les services du Trésor entreprennent leur action, laquelle pourrait fournir un 2e facteur g, d'où les résultats r se déduiraient logiquement.

On doit ajouter au modèle 4 facteurs spécifiques indépendants de f et g.

On a 12 inconnues (les 3 coefficients dans $c\ s\ p\ r$) à déterminer par

4 équation du type $1 = A^2 + A'^2 + A''^2$,

10 équations du type $(cs) = AB + A'B' + (AB' + BA')\ (fg)$.

On n'est pas obligé de supposer g indépendant de f (l'action du Trésor dépend du retard des versements spontanés); ceci nous fournit une 13e inconnue $x = (fg)$.

Il est possible que, dans le cas précis, ce système soit satisfait bien qu'il ait une ou deux équations de trop. Ce serait le signe qu'on aurait enfin trouver un modèle adéquat.

Résultat :

a) Si l'on suppose g *indépendant de* f, donc $x = (fg) = 0$, nous trouvons

$$
\begin{aligned}
-d &= f \\
c &= 0{,}614\,f - 0{,}763\,g + 0{,}202\,u \\
s &= 0{,}581\,f - 0{,}676\,g + 0{,}453\,v \\
p &= 0{,}324\,f - 0{,}300\,g + 0{,}903\,w \\
r &= 0{,}970\,f + 0{,}146\,g + 0{,}194\,z
\end{aligned}
$$

Les corrélations entre $-d$ et $c\ s\ p$ sont *exactement* celles proposées. La matrice recalculée des autres corrélations n'est pas mauvaise (on a employé la règle de Spearman pour déterminer les coefficients de g).

	Corrélations données (1955)			Corrélations recalculées		
	c	*s*	*p*	*c*	*s*	*p*
s	0,920	—		0,873	—	
p	0,443	0,379	—	0,448	0,391	—
r	0,495	0,472	0,253	0,484	0,465	0,270

Nous pouvons donc penser que les données sont assez bien expliquées par les 2 facteurs :

f : ampleur des retards hors des échéances

g : action des Services du Trésor indépendante des retards.

On notera que : { l'indulgence de ces Services profite au Trésor
les *pénalités* ont un facteur *spécifique excessivement élevé.*

b) Si on laisse g dépendre de f, la corrélation x entre f et g est astreinte à rester entre 2 limites + 0,453 et — 0,202 (sans quoi le modèle présente quelque impossibilité). Moyennant quoi on peut modifier les équations précédentes; mais on n'a pas vu qu'on puisse se servir de x pour *améliorer le modèle :* car les corrélations recalculées sont en fait *indépendantes de x.* La présentation des équations sera simplifiée si l'on s'arrange ainsi pour annuler l'un des 8 coefficients (de f et g), mais ceci ne présente aucun intérêt majeur.

Cas des données 1956 : On trouve de même pour 1956 (avec $(fg) = 0$).

$$d = -f$$
$$c = 0{,}730\, f - 0{,}616\, g + 0{,}296\, u$$
$$s = 0{,}705\, f - 0{,}674\, g + 0{,}222\, v$$
$$p = 0{,}269\, f - 0{,}338\, g + 0{,}902\, w$$
$$r = 0{,}939\, f + 0{,}185\, g + 0{,}290\, z$$

	Corrélations données			Corrélations recalculées		
	c	s	p	c	s	p
s	0,922	—		0,930	—	
p	0,394	0,435	—	0,404	0,417	—
r	0,556	0,547	0,195	0,571	0,538	0,191

On notera la grande ressemblance existant entre les modèles ajustés pour 1955 et 1956. Quant à l'ajustement des données, il s'avère excellent.

Tel est, en définitive, le moins mauvais usage que nous ayons pu faire de l'analyse factorielle pour le problème posé par la Comptabilité publique. On notera que l'emploi de l'une des variables (— d) comme le facteur f paraît en principe réprouvé par les spécialistes de l'analyse factorielle. Mais si l'on tient compte des *liens de causalité,* c'est-à-dire de l'ordre chronologique (la valeur de d est pratiquement connue quand se précisent les valeurs de c, s et p, d'où suivra ultérieurement r) notre choix de f *a priori* ne paraît pas absurde :

Chaque problème mérite examen sans préjugés excessifs.

VIII. QUESTIONS PENDANTES

Problème 1. Quelle interprétation donner aux facteurs communs (parties III, IV, V, VI).

Le facteur f est le facteur *général,* qui dirige dans le même sens les retards à payer, l'action répressive des services du Trésor et les rentrées tardives.

Le facteur g a au contraire ceci de singulier qu'il entre avec des signes opposés d'une part dans r et — d, d'autre part dans $c\ s\ p$; l'indulgence et la compréhension des services du Trésor (compte tenu d'un supplément des retards à l'échéance) se traduit apparemment par des rentrées supplémentaires récupérées après échéance.

On notera que les coefficients de f et g du VI ressemblent fort à ceux obtenus déjà par la méthode centroïde (du III). Mais on n'est plus en mesure de calculer pour *chaque département* les valeurs de f et g. C'est ce qui distingue l'analyse factorielle de l'analyse des composantes.

Problème 2. Convient-il de s'arrêter quand on a trouvé 1 facteur, 2 facteurs, 3 facteurs... communs? Ceci peut s'étudier en recourant à un modèle mettant en jeu le calcul des probabilités.

Il ne semble pas que (jusqu'à ce point précis, bien entendu) les questions qu'on vient d'évoquer soient du domaine des hautes mathématiques.

Mais il est aujourd'hui reconnu que les problèmes *d'identification* d'une certaine *structure* voilée par des données affectées *d'erreurs*, sont parmi les plus difficiles qui se posent en économétrie comme en psychométrie, et il est caractéristique que les recherches les plus récentes comme celles d'Anderson et Rubin [1] (éditées par Neyman, 1956) ne s'adressent plus qu'aux spécialistes de la pure mathématique statistique.

CONCLUSION : Nous avons vu finalement qu'il semblait possible d'expliquer la matrice des corrélations calculée pour la Comptabilité publique en adoptant la variable D comme l'un des facteurs et en lui adjoignant un second facteur commun (indépendant). Nous avons effleuré au passage divers problèmes délicats :

Les modèles proposés sont variés : lequel est le moins mauvais?

Comment faudrait-il s'y prendre pour calculer les meilleurs coefficients (en admettant qu'on ait choisi le modèle)?

Quelle est l'interprétation concrète du modèle?

Quelle est la stabilité du modèle à travers le temps?

Ajoutons qu'il serait concevable d'utiliser le modèle pour prévoir l'effet sur les rentrées d'impôts de petits changements apportés aux variables stratégiques, bien que l'Administration centrale ne puisse pas agir directement sur elles. Mais toute modification notable de ces variables pourrait conduire à des « pourcentages » inférieurs à 0 % ou supérieurs à 100 %, ou (sans en arriver là) sortirait du domaine étroit où le modèle est réaliste.

P. THIONET

ANNEXE

Formule de Spearman pour calculer les « communautés » :

Le coefficient 0,863, par exemple, est la racine carrée du rapport obtenu en totalisant les numérateurs entre eux et les dénominateurs entre eux, pour toutes les fractions suivantes qui, d'après la théorie de Spearman, sont égales (aux erreurs près) :

$$\frac{0{,}920 \times 0{,}443}{0{,}379} = \frac{0{,}920 \times 0{,}614}{0{,}581} = \frac{0{,}920 \times 0{,}495}{0{,}472} = \frac{0{,}443 \times 0{,}614}{0{,}324}$$

$$= \frac{0{,}443 \times 0{,}495}{0{,}253} = \frac{0{,}614 \times 0{,}495}{0{,}970}$$

1. ANDERSON et RUBIN, Statistical inference in factor analysis (3e symposium, Berkeley, t. V).