Correspondence Analysis and Text Analysis

Methodology Institute, LSE, May 13 - 19, 2011

1. About the History of Correspondence Analysis

1901 - 1980

Ludovic Lebart
Télécom-ParisTech
ludovic@lebart.org

Reminder

Special issue of the Electronic Journal for the History of Probability and Statistics :

www.jehps.net (2008)

Nine contributions to the History of Data Analysis before 1980

- **John C. Gower** [*The biological stimulus to multidimensional data analysis*]
- **Fionn Murtagh** [*Origins of Modern Data Analysis Linked to the Beginnings and Early Development of Computer Science and Information Engineering*]
- **Michel Armatte** [*Histoire et Préhistoire de l'Analyse des données par J.P. Benzécri: un cas de généalogie rétrospective*]
- **Alain Desrosières** [*Analyse des données et sciences humaines : comment cartographier le monde social*?]
- **Willem Heiser** [*Psychometric Roots of Multidimensional Data Analysis in the Netherlands: From Gerard Heymans to John van de Geer*]
- **Antoine de Falguerolles** [*L'analyse des données ; before and around*]
- **Alfredo Rizzi** [*Italian Contributions to Data Analysis*]
- **Hans Hermann Bock** [*Origins and extensions of the k-means algorithm in cluster analysis*]
- **Boris Mirkin** and **Ilya Muchnik** [*Some topics of current interest in clustering: Russian approaches 1960-1985*]

1. About the History of Correspondence Analysis (MCA)

(1901 – 1980)

Content

1. Prehistory of CA (FA, PCA, SVD, CA) (1901 – 1940).

2. The discoverers of MCA: L. Guttman and C. Burt: (1941- 1953)

3. CA as a technology for Data Science (C. Hayashi, J.-B. Benzécri, and others) (1954 – 1980)

Part 1: Prehistory of CA

Karl Pearson,

1857-1936

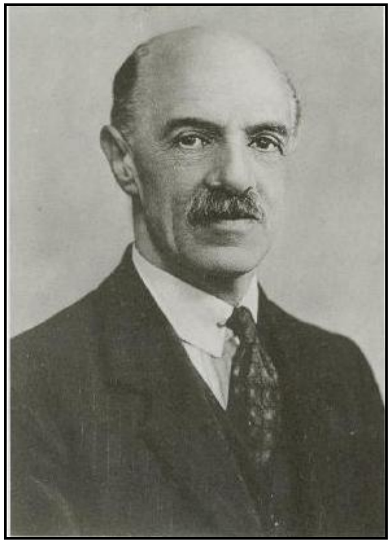
►. **Pearson K.** (1901) - On lines and planes of closest fit to systems of points in space. *Phil. Mag.* 2, n°ll, p 559-572.

Karl Pearson has been on the verge of discovering Correspondence Analysis, according to:

de Leeuw J. (1983) – On the prehistory of correspondence analysis. *Statistica Neerlandica*, vol 37, n°4, p 161-164.

Charles Spearman,

1863 - 1945



One factor:
Spearman C. (1904) – “General intelligence, objectively determined and measured”. *American Journal of Psychology, 15, p 201-293.*

Several factors:
Garnett J.-C. (1919) - General ability, cleverness and purpose. *British J. of Psych.*, 9, p 345-366.
Thurstone L. L. (1947) - *Multiple Factor Analysis*. The University of Chicago Press, Chicago.

One factor: (Spearman)

General factor for individual i

Known

$$x_i^j = a_j f^i + \varepsilon_i^j$$

Unknown

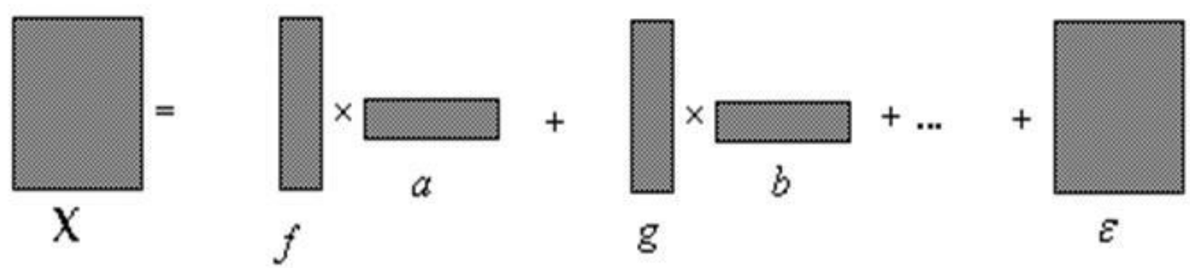
Value of variable j for individual i

Residual (hopefully small)

Coefficient of variable j

Several factors: (Garnett, Thurston)

$$x_i^j = a_j f^i + b_j g^i + \ldots + \varepsilon_i^j$$



Harold Hotelling, 1895-1973

Develops PCA as a technique of mathematical statistics. Recommends the use of the iterated power algorithm for computing eigenvalues. Proposes Canonical Analysis (1936).

▸ **Hotelling H.** (1933) - Analysis of a complex of statistical variables into principal components. *J. Educ. Psy.* 24, p 417-441, p 498-520.

With Hotelling and Eckart & Young, principal axes techniques are connected to both *multivariate analysis* and *modern linear algebra*.

$$\mathbf{X} = \sqrt{\lambda_1}\,\mathbf{v}_1 \times \mathbf{u}'_1 + \dots + \sqrt{\lambda_\alpha}\,\mathbf{v}_\alpha \times \mathbf{u}'_\alpha + \dots + \sqrt{\lambda_p}\,\mathbf{v}_p \times \mathbf{u}'_p$$

▸ **Eckart C., Young G.** (1936) - The approximation of one matrix by another of lower rank. *Psychometrika*, l, p 211-218.

CA: **1933, 1935** Two pioneering papers

► **Richardson M., Kuder G. F.** (1933) - Making a rating scale that measures. Procter and Gamble, *Personnel Journal*, 12, p 71-75.
[Reciprocal averaging]

► **Hirschfeld H.O.** (1935) - A Connection between correlation and contingency. *Proc. Camb. Phil. Soc.* 31, p 520-524.
[First manifestation of Correspondence Analysis]
[Paper long ignored, rediscovered by John Gower]

H.O. Hartley,
(Hirschfeld)
1912 – 1980

CA: 1940, 1941 Two other (independent) pioneering papers

► **Fisher R. A.** (1940) – The precision of discriminant functions. *Ann. Eugen. Lond.*, 10, 422-429.

► **Maung K.** (1941) – Measurement of association in a contingency table with special reference to the pigmentation of hair and eye colours of Scottish schoolchildren. *Ann. Eugen. Lond*. 11, 189-223. [Application following the previous Fisher's paper]

Ronald Aylmer **Fisher,**
1890 –1962

R.A. Fisher, 1955,
(p. 6, *Experiments in plant hybridisation* / G. Mendel. Edinburgh : Oliver & Boyd, 1965)

About the history of Multiple Correspondence Analysis before 1980

Multiple correspondence analysis (MCA) can be viewed as a simple extension of the area of applicability of Correspondence analysis (CA) from the case of a *contingency table* to the case of a *complete disjunctive binary table*. The properties of such a table are interesting, the computational procedures and the rules of interpretation of the obtained representations are simple, albeit specific.

MCA being both a particular case and a generalization of CA, it is not easy to disentangle its history from that of CA.

The basic formulas underlying MCA can be traced back to Guttman (1941) who devised it as a method of scaling, but also to Burt (1950), in a wider scope . The first applications of MCA as an exploratory tool probably dates back to Hayashi (1956). The availability of computing facilities entailed a wealth of new developments and applications in the early seventies, notably around Benzécri (1964,1973). The term *Multiple Correspondence Analysis* was coined at that time.

Multiple correspondence analysis has been developed in another theoretical framework (closer to the first approach of Guttman) under the name of *Homogeneity Analysis* by the research team of de Leeuw since 1973 (cf. Gifi, 1981/1990) and under the name of *Dual Scaling* by Nishisato (1980) more inspired by Hayashi.

Other types of extensions of correspondence analysis based on generalized canonical analysis have their foundation particularly in the works of Carroll (1968), Horst (1961) et Kettenring (1971).

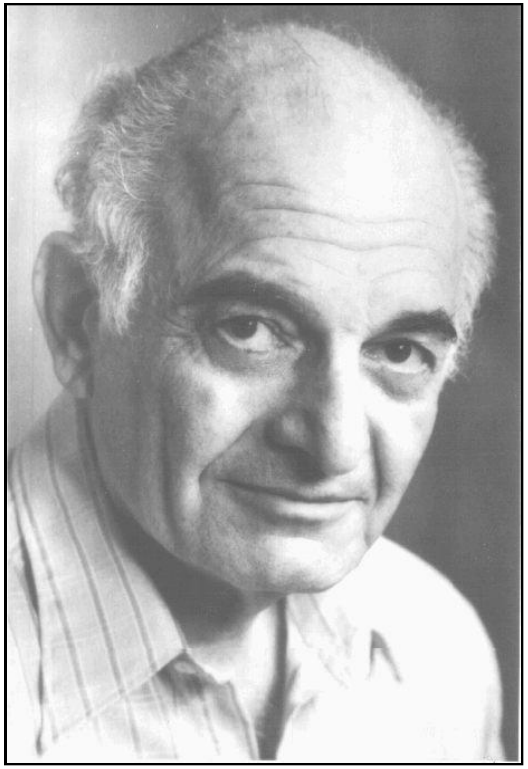
A first synthetic exposition of various approaches to MCA has been proposed by Tenenhaus and Young (1985).

Concerning a technique which is rather *specific* whose boundaries are so *fuzzy*, the term “history” may seem pretentious, almost provocative. In fact, the two important words in the title are “About” (we are dealing here with a point of view and a testimony) and “1980” (a distance of thirty years should, normally, provide us with a certain perspective).

Part 2: MCA , the discoverers

Louis Guttman,

1916-1987



► **Guttman L.** (1941) - The quantification of a class of attributes: a theory and method of a scale construction. In : *The prediction of personal adjustment* (Horst P., ed.) p 251 -264, SSCR New York.

Guttman Seminal 1941 paper

The Quantification of a Class of Attributes: A Theory and Method of Scale Construction

1 THE PROBLEM

In the social sciences we are often confronted with a set of acts of a population of individuals that we would like to consider as a single class of behavior. Examples would be that totality of acts which is considered to constitute attitude toward war, or that totality of acts which is considered to constitute competence in a job, or that totality of acts which is considered to constitute marital adjustment.

We are interested in the case where the acts are attributes recorded in the form of items with mutually exclusive subcategories.

Thus, we are given the responses of a population of U individuals to a set of m items which have a common content that is desired to be thought as a single class of behavior. These responses can be represented by check marks as in the following table (with hypothetical entries):

A complete disjunctive table

Thus, we are given the responses of a population of U individuals to a set of m items which have a common content that is desired to be thought as a single class of behavior. These responses can be represented by check marks as in the following table (with hypothetical entries):

SUBCATEGORY	INDIVIDUAL					
	1	2	3	4	...	U
A_1	√		√		...	
A_2		√			...	√
A_3				√	...	
B_1			√		...	√
B_2	√	√		√	...	
...	...	...	...	...	...	...
Z_1		√			...	√
Z_2	√		√		...	
Z_3				√	...	
Z_4					...	

A modern and thorough development

It will be convenient to express the right member of (3) in matric form.

We recognize immediately that

$$\frac{1}{m}\,\mathbf{x}M = (a_1\,a_2\;\cdots\;a_U),$$

so that

$$m\sum_{i=1}^{U} a_i^2 = \frac{1}{m}\,\mathbf{x}MM'\mathbf{x}'.$$

If we form the diagonal matrix

$$D = \left\|\begin{matrix} N_1 & & & \\ & N_2 & & \\ & & \cdots & \\ & & & N_n \end{matrix}\right\|,$$

we can write

$$\sum_{j=1}^{n} N_j x_j^2 = \mathbf{x}D\mathbf{x}'.$$

Then we can write (3) as

$$\eta_x^2 = \frac{\mathbf{x}MM'\mathbf{x}'}{m\mathbf{x}D\mathbf{x}'}. \qquad (4)$$

4 MAXIMIZING THE CORRELATION RATIO

Maximizing η_x^2 is seen to be equivalent to maximizing the quadratic form $\mathbf{x}MM'\mathbf{x}'$ under the restriction that $m\mathbf{x}D\mathbf{x}'$ be some finite constant. This we can do by maximizing the expression

$$\mathbf{x}MM'\mathbf{x}' - m\phi\mathbf{x}D\mathbf{x}', \qquad (5)$$

where ϕ is a Lagrange multiplier. Differentiating (5) with respect to $\mathbf{x}'$ and equating the result to zero for a maximum, we obtain the condition

$$\mathbf{x}(MM' - m\phi D) = 0. \qquad (6)$$

Let us suppose (6) is satisfied by a particular $\mathbf{x}_0$ and ϕ_0. Then, postmultiplying both members of (6) by $\mathbf{x}_0'$ and solving for ϕ_0,

$$\phi_0 = \frac{\mathbf{x}_0MM'\mathbf{x}_0'}{m\mathbf{x}_0D\mathbf{x}_0'}; \qquad (7)$$

or, comparing (7) with (4),

$$\phi_0 = \eta_{x_0}^2.$$

It will be shown that the number of independent solutions of (6) is equal to the rank of M, which we shall see to be $n-m+1$. One solution will be extraneous, yielding $\phi_0=\eta_{x_0}^2=1$. Apart from this extraneous solution, the solution of (6) that we want is the particular $\mathbf{x}_0$ which corresponds to the largest of the $n-m$ stationary values of $\phi_0=\eta_{x_0}^2$.

Let us first examine the extraneous solution. It is the row vector of n elements, each of which is unity,

$$I_n = (1\;1\;1\cdots 1).$$

We first see that

$$I_nM = mI_U, \qquad (8)$$

where I_U is the same type of vector as I_n except that it has U elements of unity.

Substituting I_n for $\mathbf{x}_0$ in the right member of (7), the numerator becomes

$$I_nMM'I_n' = mI_UI_U'm = m^2U;$$

and the denominator becomes

$$mI_nDI_n' = m\sum_{j=1}^{n} N_j = m^2U;$$

so that

$$\phi_0 = \eta_{x_0}^2 = \frac{m^2U}{m^2U} = 1.$$

Using this value for ϕ in (6), we shall now see that I_n satisfies (6). We have to verify that

$$I_n MM' = mI_n D. \tag{9}$$

Using (8), we derive from (9) that

$$I_U M' = I_n D. \tag{10}$$

Both members of (10) are recognized to be the row vector

$$(N_1\, N_2\, \cdots\, N_m),$$

which completes the verification.

However, the solution I_n is extraneous since it does not satisfy the restriction (2). Its appearance as a solution is an artifact since the value of $\phi_0 = \eta^2_{x_0} = 1$ that corresponds to it is not actually the value of a correlation ratio.

We shall now show that all other solutions of (6) will in general satisfy (2). Postmultiply both members of (6) by I_n'. Then

$$xMM'I_n' = m\phi x D I_n'.$$

This is verified to be in scalar notation

$$m\sum_{j=1}^{n} N_j x_j = m\phi \sum_{j=1}^{n} N_j x_j.$$

Therefore, if $\phi \neq 1$, it must be that (2) is satisfied.

5 THE "CHI-SQUARE" METRIC

The multiplicity of solutions in Section 4 shows that we are dealing with a problem akin to that of factor analysis in psychology. We can, as a matter of fact, throw our solution into the form of a principal axis solution, as we shall do in this section.

There is an essential difference, however, between the present problem of quantifying a class of attributes and the problem of "factoring" a set of quantitative variates. The principal axis solution for a set of quantitative variates depends on the preliminary units of measurement of those variates. In the present problem, the question of preliminary units does not arise since

we limit ourselves to considering the presence or absence of behavior. But we shall now see that in a sense a metric has arisen out of our analysis, a metric that we shall call the "chi-square" metric.

To obtain the form of a principal axis solution, let

$$\bar{x} = xD^{1/2}, \qquad \overline{M} = D^{-1/2}M. \tag{11}$$

Using (11) in (6), we obtain the matric equation for the characteristic vectors of $\overline{MM}'$:

$$\bar{x}(\overline{MM}' - m\phi I) = 0. \tag{12}$$

Hence, the $n-m+1$ solutions for x can be obtained from the principal axes of $\overline{MM}'$, and the corresponding stationary values $\phi_0 = \eta^2_{x_0}$ are proportional to the latent roots of $\overline{MM}'$.

The major principal axis corresponds to the largest stationary ϕ_0 which is unity, for the extraneous solution I_n satisfies (6) with $\phi_0 = 1$. Hence,

$$I_n D^{1/2}$$

is proportional to the major characteristic vector of $\overline{MM}'$. The contribution of this vector, normalized to m, to $\overline{MM}'$ is the matrix of rank one

$$\frac{1}{U} D^{1/2} I_n' I_n D^{1/2}.$$

The maximum, nonextraneous value of $\eta^2_{x_0}$ then corresponds to the largest latent root of

$$G = \overline{MM}' - \frac{1}{U} D^{1/2} I_n' I_n D^{1/2},$$

and the subcategory weights are obtainable from the components of the major characteristic vector of G.

To obtain a scalar formula for the general element of G, denote the general element of MM' by N_{jk}. It is the number of individuals who checked both subcategories j and k. Clearly $N_{jj} = N_j$; and N_{jk} is zero if j and k are distinct but pertain to

Mention of the « Chi-Square metric »

the same item, since responses within an item are mutually exclusive. Then the general element of $\overline{MM'}$ is

$$\frac{N_{jk}}{\sqrt{N_j N_k}}$$

and the general element of $\mathbf{G}$ is

$$\frac{N_{jk}}{\sqrt{N_j N_k}} - \frac{\sqrt{N_j N_k}}{U} = \frac{N_{jk} - \dfrac{N_j N_k}{U}}{\sqrt{N_j N_k}} \cdot$$

This element is recognized to be precisely that used in the chi-square test of significance of association between two attributes. (Actually, the extraneous axis that we have subtracted out is what is called "chance expectation" in the sampling theory.) Whereas in the case of quantitative variates the question must be answered beforehand as to what product-moments to "factor," the chi-square product-moment emerges automatically as a result of the analysis of the problem of attributes.

It must be remembered, however, that this is but an oblique way of looking at the matter. We obtained the chi-square metric because we were looking for a *single axis*. If the largest $\eta^2_{x_0}$ is not large enough to account for enough of the variability in $\mathbf{M}_x$, then trying to reproduce $\mathbf{M}_x$ by frequency functions obtained from a single axis will not be very effective; and we cannot usefully think of the class of attributes as comprising approximately a single variate. Then we should be tempted to try a "multiple factor" analysis. But the present rationale was devised specifically for a "single factor" analysis and does not necessarily carry over to the other case. It may be quite a different task to devise a rationale for "multiple factor" analysis of attributes, and the chi-square metric may not hold at all.

Furthermore, converting equation (6) into (12) may be considered merely an artifice. In (12) we solve for the major principal axis in the form of $\bar{\mathbf{x}}_0$, but it is $\mathbf{x}_0$ we are really interested in; and $\mathbf{x}_0$ is *not* the principal axis in general.

6 THE NUMBER OF INDEPENDENT SOLUTIONS

We now wish to fill in the proof for the statement that there are in general $n-m+1$ independent solutions to (6). We can show this by showing the rank of $\mathbf{G}$ to be $n-m$, for $\mathbf{G}$ has the extraneous solution subtracted out. Now the rank of $\mathbf{G}$ is the rank of

$$\overline{\mathbf{G}} = \mathbf{D}^{1/2}\mathbf{G}\mathbf{D}^{1/2},$$

for which the general element is

$$N_{jk} - \frac{N_j N_k}{U} \cdot$$

Let us consider the columns in $\overline{\mathbf{G}}$ pertaining to a single item which has, say, s subcategories. To show that we are restricting our attention to a single item, let us change notation to let $U_1, U_2, \cdots, U_s$ denote the number of persons checking the respective subcategories. Then

$$U_1 + U_2 + \cdots + U_s = U.$$

It is well known in the theory of attributes that the sum of the s columns (or rows) in $\overline{\mathbf{G}}$ pertaining to a single item is zero. This we can express in matric notation. Let

$$I_{)s(}$$

be a row vector of n elements which has s entries of unity corresponding to the s columns of $\overline{\mathbf{G}}$ in which we are interested, and which has all other entries zero. Then

$$\overline{\mathbf{G}}I'_{)s(} = 0 \tag{13}$$

for all items.

Hence, any particular item contributes only $s-1$ linearly independent columns to $\overline{\mathbf{G}}$, so that the total number of linearly independent columns cannot exceed

$$\sum (s-1) = n - m,$$

the summation extending over the m items. Therefore, the rank

But an unexpected limited scope…

Cyril Lodowic
Burt
(1883-1971)

C. Burt has rediscovered the formulas of L. Guttman. However, the eigth following slides will show that his scope and his point of view about both the use and the usefulness of the method (MCA) are much wider (and more modern in some respect) than that of L. Guttman.

C. Burt, an experienced practitioner, saw immediately the interest of using (interpreting) more than one axis.

► **Burt C. (1950) - The factorial analysis of qualitative data. *British J. of Statist. psychol*. 3, 3, p 166-185.**

About the polemics concerning the alleged fraud about some data used by Sir Cyril Burt, let us quote the Encyclopedia Britannica :

From the late 1970s it was generally accepted that "he had fabricated some of the data, though some of his earlier work remained unaffected by this revelation".

A sample of references:

Gould S.J. (1982) . The real error of Cyril Burt. In: *The Mismeasure of Man*. W.W. Norton and Company, New York. Chapter 6, p 234-320.
Hearnshaw, L. (1979). *Cyril Burt: Psychologist*. Ithaca, NY: Cornell University Press. Also published London: Hodder and Stoughton, (1979).
Joynson, R.B. (1989). *The Burt Affair*. New York: Routledge. *(supporting C.B.)*
Fletcher, R. (1991). *Science, Ideology and the Media: The Cyril Burt Scandal*. New Brunswick, USA: Transaction Publishers. *(supporting C.B.)*

Burt
1950

THE FACTORIAL ANALYSIS OF QUALITATIVE DATA

By CYRIL BURT
Psychological Department, University College, London

I. *The Importance of Qualitative Data in Psychology*. II. *Alternative Statistical Techniques*. III. *The Treatment of Multiple Determinates*. IV. *A Factorial Analysis of Physical Attributes*. V. *Summary and Conclusions*.

I. THE IMPORTANCE OF QUALITATIVE DATA IN PSYCHOLOGY

The Form of the Data. In many investigations within the field of individual differences, the available data are expressed, not as quantitative measurements stating magnitude or degree, but in terms of classes or attributes which are essentially qualitative. Statistical psychologists, and particularly those who have worked with standardized tests and employed factorial methods, are often accused of ignoring the qualitative aspects of their problems. As a rule, their critics seem to assume that, because such observations are not recorded in the form of quantitative assessments, they are no longer amenable to quantitative treatment, and can therefore have no use or interest for the statistical investigator. That, however, is a patent fallacy. If they are to be tabulated, such observations, it is true, must be set down in what the schoolboy calls the 'nought-and-one' style, where each 'one' will signify that yet another individual belongs to the class named or possesses the attribute specified, and 'nought' will signify that he does not. But, when that has been done, it is a simple matter to count the ones and compare their sum with that of the ones and noughts added together: in this way we can readily summarize our observations in the form of a table of frequencies or probabilities; and these can manifestly be subjected to statistical treatment: (cf. 3, 4, 5).

In psychology the oldest and most familiar instance of this procedure is furnished by the marks given for the Binet-Simon tests. Here each examinee is in effect awarded a measurement of 'one' for every test he passes and 'nought' for every failure. A similar device has long been adopted by teachers in marking the simpler type of examination paper: each child's total score is obtained by counting every correct answer as 'one,' and adding up the total. So-called personality-tests—the Rorschach, the personal inventory, the biographical questionnaire, and enquiries about interests and attitudes, for example—are frequently scored in this way.

Again a complete disjunctive table…

TABLE I. OBSERVED TABLE SHOWING PRESENCE OR ABSENCE OF ATTRIBUTES FOR A GIVEN SAMPLE OF PERSONS

	Persons (N)					Total for Row
	Tom	Dick	Harry	...	George	
Determinables (m)						
Hair-colour						
Fair	1	0	0	...	0	N_1
Red	0	1	0	...	0	N_2
Dark	0	0	1	...	1	N_3
Subtotal	1	1	1	...	1	N
Eye-colour						
Light	1	0	1	...	0	N_4
Brown	0	1	0	...	1	N_5
Subtotal	1	1	1	...	1	N
Total for Column ..	m	m	m	...	m	Nm

It will be convenient to adopt the following notation :[2]

Number of determinables .. .. m,
Number of determinate values in the 1st, 2nd, . . . jth, . . . mth determinable .. .. $n_1, n_2 \ldots n_j, \ldots, n_m$,
Total number of determinates .. .. $n = n_1 + n_2 + \ldots + n_m$
Total number of persons for each determinate .. $N_1, N_2, \ldots N_{n_j}$,
Total number of persons in sample .. .. .. $N = N_1 + N_2 + \ldots + N_{n_j}$,
Number of persons possessing both the jth and the kth determinates .. .. N_{jk}.

The BURT Contingency table

The data were collected at Liverpool, a district where representatives of different nationalities—Welsh, Irish, Scots, as well as foreigners—were easily accessible. In most cases temperamental assessments were obtained at the same time; but these will also be omitted from the present tables, as they raise somewhat different issues. In all, 217 individuals were examined, about two-thirds of them males. But, partly to simplify the calculations and partly because the later observations were rather more trustworthy, I shall here restrict my analysis to the data obtained from the last hundred males in the series.

The Crude and Standardized Contingency-Tables. The number of persons characterized by the several attributes specified, taken in pairs, are set out in Table II (this may be taken as representing the initial type of contingency-table which was designated C_t on p. 172 above).

TABLE II. OBSERVED FREQUENCIES

Number of Persons exhibiting Characteristics Specified

Trait	F	R	D	Tot.	L	M	B	Tot.	N	W	Tot.	T	S	Tot.
HAIR														
Fair ..	22	0	0	22	14	6	2	22	14	8	22	13	9	22
Red ..	0	15	0	15	8	5	2	15	11	4	15	10	5	15
Dark ..	0	0	63	63	11	25	27	63	44	19	63	20	43	63
Total ..	22	15	63	100	33	36	31	100	69	31	100	43	57	100
EYES														
Light ..	14	8	11	33	33	0	0	33	27	6	33	29	4	33
Mixed ..	6	5	25	36	0	36	0	36	20	16	36	10	26	36
Brown ..	2	2	27	31	0	0	31	31	22	9	31	4	27	31
Total ..	22	15	63	100	33	36	31	100	69	31	100	43	57	100
HEAD														
Narrow..	14	11	44	69	27	20	22	69	69	0	69	30	39	69
Wide ..	8	4	19	31	6	16	9	31	0	31	31	13	18	31
Total ..	22	15	63	100	33	36	31	100	69	31	100	43	57	100
STATURE														
Tall ..	13	10	20	43	29	10	4	43	30	13	43	43	0	43
Short ..	9	5	43	57	4	26	27	57	39	18	57	0	57	57
Total ..	22	15	63	100	33	36	31	100	69	31	100	43	57	100

Louis Guttman's comments…

Vol. VI Part I The British Journal of Statistical Psychology May, 1953

A NOTE ON SIR CYRIL BURT'S 'FACTORIAL ANALYSIS OF QUALITATIVE DATA'

By LOUIS GUTTMAN

Scientific Director, Israel Institute of Applied Social Research, Jerusalem

I. *Procedures for Dealing with Qualitative Data.* II. *Analysis into Principal Components.* III. *Applications to Scalable and Non-Scalable Data.* IV. *The Study of Deviations : Image Analysis and Nodal Analysis.*

I. PROCEDURES FOR DEALING WITH QUALITATIVE DATA

World War II interrupted communication between scientists in different countries, and only gradually is the exchange of information being made up. A case in point is the monograph *The Prediction of Personal Adjustment* by Paul Horst and others, which appeared just before the United States actively entered the holocaust. This was published by the Social Science Research Council in New York as *Bulletin 48* in 1941. During the war, only a handful of copies was able to reach Europe. In a recent article in this *Journal* (1), Sir Cyril Burt has called attention to the importance of developing a theory for qualitative data, as distinct from quantitative data. He develops a particular algebraic formulation which leads to the resolution of the data into principal components (latent vectors). This happens to be a topic treated also in the above-mentioned monograph, by the present writer. It is gratifying to see how Professor Burt has independently arrived at much the same formulation. This convergence of thinking lends credence to the suitability of the approach. The purpose of the present note is to call attention to the points of similarity and to describe further developments which have occurred in the United States and in Israel in the theory of qualitative data.

In a chapter of the above-mentioned Social Science Research Council monograph entitled 'The Quantification of a Class of Attributes' (2), it was proposed that qualitative data could be recorded in a manner amenable to treatment by matrix algebra. In form, the matrix *M* on page 326 of the monograph is identical with Table I of Professor Burt's article ((1), p. 171). My own paper proposes three different kinds of problems of quantification :

C. Burt's comments… about L. Guttman's comments

In the same issue of the BJSP

Vol. VI Part I — The British Journal of Statistical Psychology — May, 1953

SCALE ANALYSIS AND FACTOR ANALYSIS

Comments on Dr. Guttman's Paper [1]

By CYRIL BURT
University College, London

I. *The Analysis of Answer Patterns.* II. *Criticisms of the Factorial Approach.* III. *Factor Analysis of a Typical Scale Pattern.* IV. *Factor Analysis of the Corresponding Answer Pattern.* V. *Summary.*

I. THE ANALYSIS OF ANSWER PATTERNS

Quantitative and Qualitative Data. It is encouraging to find that Dr. Guttman has discerned points of resemblance between the methods we have independently reached for analysing qualitative data; and I willingly agree that this convergence in our lines of approach lends additional plausibility to the results. If, as I gather, he cannot wholly accept my own interpretations, that perhaps is attributable to the fact that our starting-points were rather different. My aim was to factorize such data; his to construct a scale. In the paper [2] to which he has referred, I sought to develop a practicable procedure for discovering the factors underlying qualitative characteristics, and then to apply that procedure to certain recurrent problems of my own. Dr. Guttman's purpose, as the title of his previous contribution indicates, was to present 'a theory and method of scale construction' by means of 'quantifying a class of attributes.' I have myself commented [3] on the similarity between my equations for factor-measurements and the equations for component scores given by Dr. Guttman, and ascribed it to the fact that his mode of deriving components seemed identical with the method proposed by Pearson for calculating 'index characters'—a method which had formed the original basis for my own factorial procedures.

Dr. Guttman's technique of scale analysis has been developed primarily in connexion with "the study of the internal structure of a universe of attitude and opinion items"; and the data he has used to illustrate his procedure consist, as a rule, of responses to questionnaires. I myself, however, first encountered the type of problem which he has discussed in a rather different field, namely, in the attempt to refine the scaling of the original Binet-Simon intelligence-tests. Before the publication of Binet's 1911 version, British psychologists relied chiefly on tests of an 'internally graded' type—opposites, dotting, card sorting, and the like. Binet introduced a composite *échelle* comprising a number of 'externally graded items,' i.e., questions or tasks for which the responses differ not in quantity but only in quality, and which are simply marked 'right' or 'wrong.' [4] Thus,

[1] I am much indebted to Dr. Guttman for his kindness in reading my manuscript and to Dr. M. L. Fraser for his help in checking my calculations.

[2] ... As there explained, the theoretical part of the paper was based

C. Burt comments…
In the same issue of the BJSP (continuation)

II. CRITICISMS OF THE FACTORIAL APPROACH

The Relations between Scale Analysis and Factor Analysis. In his very instructive paper (pp. 1–4) Dr. Guttman has indicated what he takes to be the differences, as well as the similarities, between his approach and my own. He observes that, while the 'principal components' employed in his method of scale analysis may be 'formally similar' to those employed in factor analysis, "nevertheless their interpretation may be quite different." The differences are examined more fully in his earlier contribution on 'The Relation of Scalogram Analysis to Other Techniques.'[1]

If I understand Dr. Guttman rightly, his objections turn on five main points.

1. First, he holds that factor analysis is "designed only for quantitative variables," and is consequently unsuited for qualitative data ((18), p. 191). The method, he says, "originated as a single factor theory by Spearman, and was developed into a multifactor theory by Thurstone and others." Were his own procedure to be described in factorial terms, then, he adds, we should have to treat it as "a single factor theory for qualitative data."

However, as other writers have pointed out (cf. this *Journal*, V, p. 206), such statements limit the term 'factor analysis' to very specific forms. In point of fact, the technique now generally known as factor analysis is much older than Spearman's procedures. It originated with Pearson's proposal to reduce a given multivariate distribution to terms of the 'principal axes of the frequency ellipsoid' and take these axes as representing 'index characters.' It was Pearson's investigation of the general problem that really supplied the earliest "algebraic formulation, leading" (if I may borrow Dr. Guttman's phrase) "to the resolution of the data into principal components (latent vectors)" ((2), pp. 559f.; cf. (15), p. 309). Factor analysis was thus a multifactor method from the start. Spearman's single factor method was developed several years later as a substitute, because he held that Pearson's approach was unsuited to the data obtained in psychology. Nevertheless, in spite of his criticisms, numerous researches were carried out in which Pearson's method was applied, not only to quantitative measurements, such as those furnished by graded tests, but also to qualitative data, such as were supplied by dichotomous test-problems and by questionnaires. In a footnote, Dr. Guttman ((18), p. 193) refers to the treatment worked out by Yule (Karl Pearson's assistant) for dealing with frequency-tables for qualitative variables (3), and considers it "strange that statistical text-books in the social sciences have not followed suit, but fail to discuss material of this kind at all." But as a matter of fact in this country psychologists have made free use of Yule's procedures, especially in relation to social data[2]; and numerous theses could be cited where factorial methods have been applied to tables of frequencies, contingencies, or Yule's coefficients of colligation or point-correlation. In particular, 'answer patterns' have regularly been subjected to a factorial analysis by various devices (cf. (6), p. 326, (11), p. 52, and refs.).

C. Burt comments…

In the same issue of the BJSP (continuation)

2. Secondly, Dr. Guttman argues that, in order to apply factor analysis, we must begin by calculating correlation coefficients, and that in the case of qualitative data such coefficients are bound to be misleading. With his criticisms of the uses made of the tetrachoric coefficient and the point correlation I very largely agree. Yet his arguments seem only to prove that these coefficients are not suitable for *all* occasions or for *every* purpose. There is one coefficient which he does not explicitly discuss—the ordinary product-moment coefficient applied to the data after they have been transformed to standard measure; and this, which has been frequently used for such problems, yields (as I shall show in a moment) results that are virtually the same as his. Nevertheless, nothing in the theory of factor analysis confines its application solely to coefficients of correlation. Indeed, for the factorist there is often a special advantage in working with frequencies, since (as I pointed out in the paper to which Dr. Guttman refers) the higher order frequencies may prove especially serviceable in the calculation of group factors.[1]

3. His third argument runs as follows. The principal criterion for scalability is reproducibility. But factor analysis does not allow us to reproduce the original data from the so-called factor-measurements. Hence factor analysis can never show whether a scale is perfect or not. This objection, however, applies only to the special procedures advocated by Spearman and Thurstone, which seek to analyse, not the total variance, but merely the common factor-variance. The method of principal axes, on the other hand, requires the full test-variances to be retained in the covariance matrix; and with Pearson's procedure the factor-measurements are obtained by pre-multiplying the initial measurements by the matrix of direction cosines (the latent vectors). Now such a matrix is necessarily orthogonal. Hence its transpose can be used as a second pre-multiplier to reproduce the initial measurements from the factor-measurements (cf. (10), Appendix II). An exact reproduction is therefore possible. If, then, we can also show that, with a perfect scale, one of the factors so obtained is in perfect correlation with the rank of the persons, it would seem that the method can after all provide an entirely satisfactory criterion.

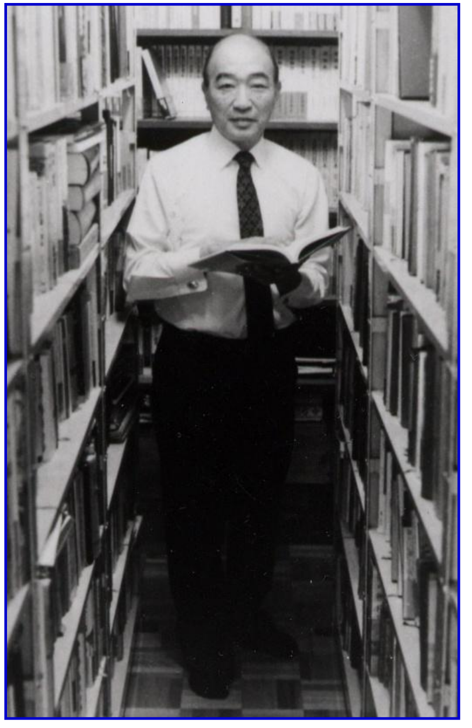
4. "The Spearman-Thurstone approach to factor analysis," as Dr. Guttman says, "is completely linear, and is therefore not adequate for analysing the curvilinearities inherent in the scale pattern." Certainly in that mode of approach the factor-measurements are always estimated by the method of linear regression, as developed in Pearson's earlier papers. But Pearson himself also elaborated a method for dealing with curvilinear regressions.[2] His treatment was intended primarily for problems involving an external criterion; but it is equally applicable to the case of an internal criterion or factor. Elsewhere I have argued that it is quite unnecessary to restrict the theory of factor analysis to linear relations only ((10), p. 258); and, with the aid of the orthogonal polynomials used in the theory of curvilinear regression,[3] it is simple to estimate factor-measurements from test data or test data from factor-measurements on the assumption of non-linear relations.

5. Finally, Dr. Guttman concludes that "from a scale analysis it can be known what a factor analysis will show; from a factor analysis it will usually be difficult, if not impossible, to know what a scale analysis will show." To determine this point I propose to apply a factorial procedure to his own table, and see how far the results achieved are similar to those reached by his own scale analysis. A concrete numerical example will probably help best to explain how far our methods are similar and in what ways they seem to differ.

Part 3: CA , a technology for Data Science

Chikio Hayashi,
(1918 - 2002)

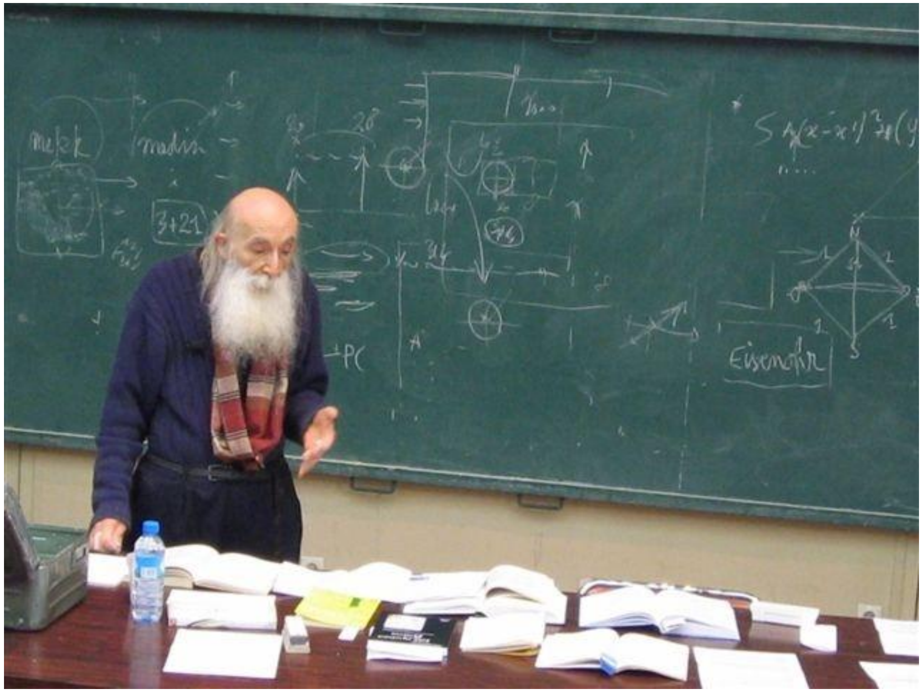
First applications of MCA



Hayashi C.(1952) - On the quantification of qualitative data from the mathematico-statistical point of view. *Annals of the Institute of Statist. Math.* (2), p 69-98.
(The 1941 Guttman paper is quoted in this article)

Hayashi C.(1956) - Theory and examples of quantification. (II), *Proc. of the Institute of Statist. Math.* 4 (2), p 19-30.

Jean-Paul Benzécri – (born 1932)



A mathematician of the highest level according to French selective procedures, and also a linguist, Benzécri considers with suspicion the diversification of techniques (diversification stimulated by the publish or perish system).
A few versatile and robust techniques mastered by the user, together with a deep knowledge of the data (in collaboration with the scientist) are more productive than a weak grasp of many seemingly more adapted methods.

Benzécri J.-P. (1964) – Cours de Linguistique Mathématique. *Faculté des Sciences de Rennes.*

Benzécri J.-P. (1969) - Statistical analysis as a tool to make patterns emerge from clouds. In : *Methodology of Pattern Recognition* (S.Watanabe, Ed.) Academic Press, p 35-74.

(J.P. Benzécri, L'avenir de l'analyse des données, Behaviormetrika, Tokyo, 1983, n° 14, 1-11.)
(this paper, published in French in the Japanese journal Behaviormetrika, is posterior to our limit of 1980, but its content was published about fifteen years earlier. It exemplifies in general terms the similarities between the « Data Science » of Hayashi and the « Analyse des données » of Benzécri: In both cases, an interdisciplinary project of experimental statistics.

Translation of the conclusion: **The future of Data Analysis** ***(in fact: of multidimensional exploratory data analysis).***
…This vision is philosophical. It does not directly translate into mathematical terms the system of concepts of a particular discipline to bind them in the equations of a model, or to accept data as they are collected; but to elaborate them into a profound synthesis that discovers new entities, and, between these, simple relationships.

Thanks to calculus, experimental situations, admirably dissected into simple components, have been translated into as many fundamental laws. We believe that Data Analysis should adequately express the laws of those phenomenons, complex by nature (living being, social body, ecosystem), that cannot be dissected without losing their character.

Benzécri and Hayashi, 1979

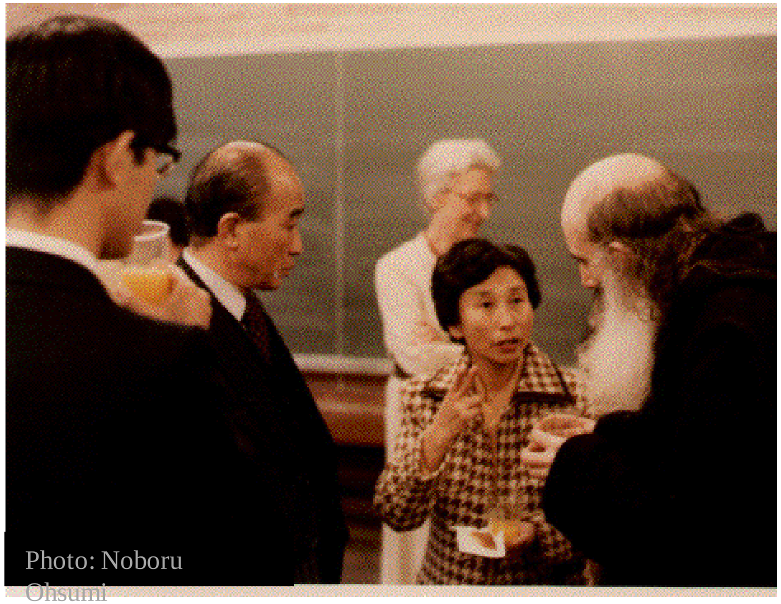


Photo: Noboru Ohsumi

Brigitte Escofier, 1941 - 1994

► **Escofier B. [Cordier B.]** (1965) - *l'Analyse des correspondances.* Thèse, Faculté des Sciences de Rennes ; [Supervisor: Jean-Paul Benzecri]

Photo: Noboru Ohsumi

1960 - 1970

Selection of noteworthy papers and books having real or potential links with MCA

Horst P. (1961) - Relation among m sets of measures. *Psychometrika*, 26, p 129-149.

Lancaster H. O. (1963) - Canonical correlation and partition of Chi-Square. *Quart. J. Math.*, 14, p 220-224.

Horst P. (1965) - *Factor Analysis of Data Matrices*. Holt, Rinehart, Winston, New York.

Carroll J. D. (1968) - Generalization of canonical correlation to three or more sets of variables. *Proc. Amer. Psychological Assoc.* p 227-228.

Lancaster H. O. (1969) - *The Chi-squared Distribution*. J. Wiley, New York.

CA re-discovered and applied to linguistics. Simultaneous displays of rows and columns of data matrices…

Benzecri J.-P. (1964) – *Cours de Linguistique mathématique*, Faculté des Sciences de Rennes.

Escofier B. [Cordier B.] (1965) - *l'Analyse des correspondances.* Thèse, Faculté des Sciences de Rennes ; published in 1969 in: *Cahiers du Bureau Universitaire de Recherche Opérationnelle*, n°13.

Benzecri J.-P. (1968) – *Sur l'analyse des tableaux de correspondances. Working paper of the L.S.M.*

The geometry of data analysis:

Gower J. C. (1966) - Some distance properties of latent and vector methods used in multivariate analysis. *Biometrika*, 53, p 325-328.

Gower J. C. (1968) - Adding a point to vector diagram in multivariate analysis. *Biometrika*, 55, 582-585.

Dissemination of CA:

Benzécri J.-P. (1969) - Statistical analysis as a tool to make patterns emerge from clouds. In : *Methodology of Pattern Recognition* (S.Watanabe, Ed.) Academic Press, p 35-74.

First anxieties about validity of results in CA:

Lebart L. (1969) - Introduction à l'analyse des données : Analyse des correspondances et validité des résultats. *Consommation*, Dunod. 4, p 65-87.

1971 – 1975

Seminal paper about the biplot (neologism) in PCA

Gabriel K.R. (1971) - The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58, 3, p 453-467.

Series of papers or technical reports about MCA (under various names)

Benzécri J.-P. (1972) - Sur l'analyse des tableaux binaires associés à une correspondance multiple. *Note multigraphiée du Laboratoire de Statistique Mathématique*, Université Pierre et Marie Curie.

Nakache J.P. (1973) - Influence du codage des données en analyse factorielle des correspondances. Etude d'un exemple pratique médical. *Revue Statist. Appl*., 21, (2).

de Leeuw J. (1973) – *Canonical Analysis of Categorical Data*. Unpublished Dissertation. University of Leiden, Leiden.

Lebart L., Tabard N. (1973) - *Recherches sur la description automatique des données socio-economiques*. Rapport CORDES-CREDOC, Convention de Recherche n°13/1971.

Treatise of Benzécri (36 contributors) . Useful, timely (and patronizing) historical paper of Hill.

Benzécri J.-P. (1973) - *L'Analyse des Données*. Tome 1: *La Taxinomie*. Tome 2: *L'Analyse des Correspondances* (2de. éd. 1976). Dunod, Paris.

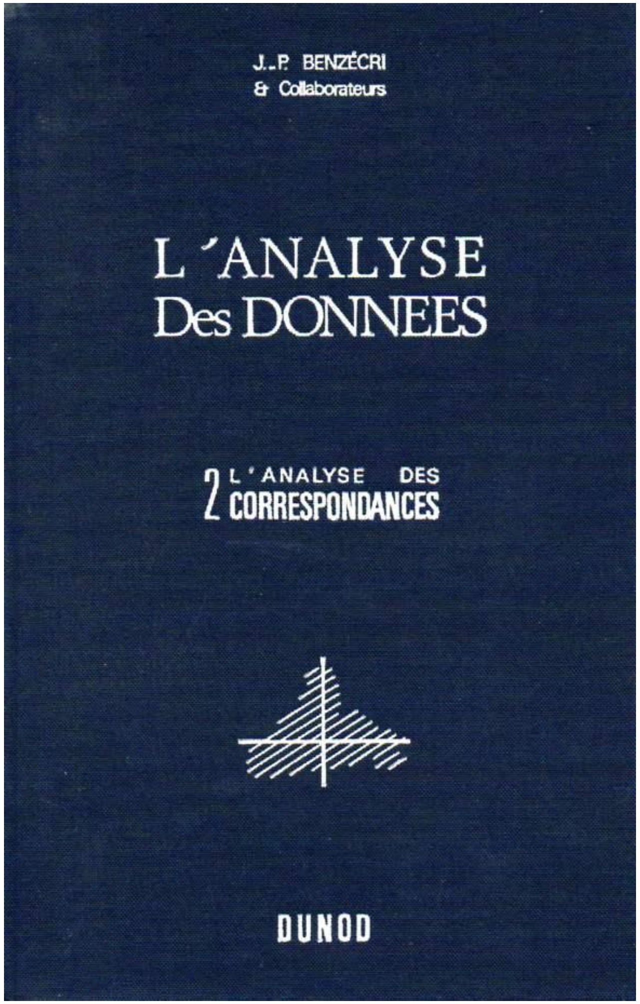
Hill M.O. (1974) - Correspondence analysis: a neglected multivariate method. *Appl. Statist.* 3, p 340-354.

More about MCA and related techniques

Lebart L. (1974) - On the Benzécri's method for finding eigenvectors by stochastic approximation. *Proceedings in Comp. Statist., In: COMPSTAT*, Physica verlag, Vienna, p 202-211.

Lebart L. (1975) - L'orientation du dépouillement de certaines enquêtes par l'analyse des correspondances multiples. *Consommation*, 2, p 73-96. Dunod.

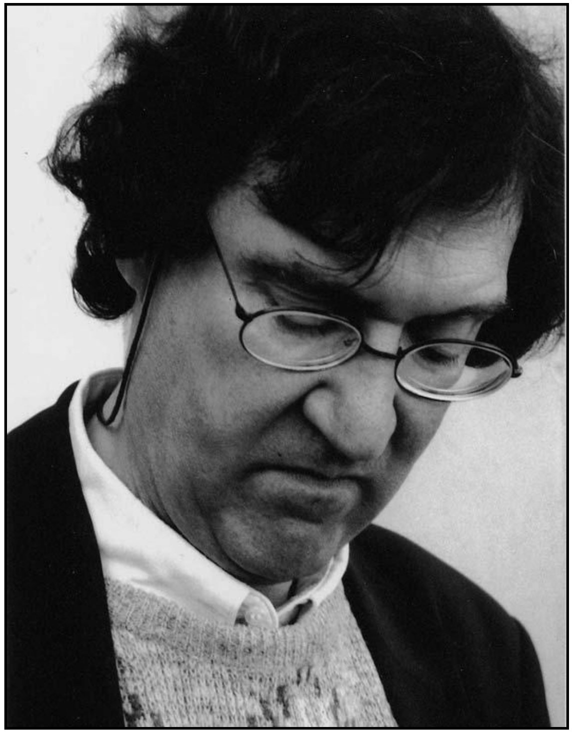
Saporta G. (1975) - Dépendance et codage de deux variables aléatoires. *Revue Statist. Appl.* 23, p 43-63.



1973

Benzécri J.-P. (1973) - *L'Analyse des Données*. Tome 1: *La Taxinomie*. Tome 2: *L'Analyse des Correspondances* (2de. éd. 1976). Dunod, Paris.

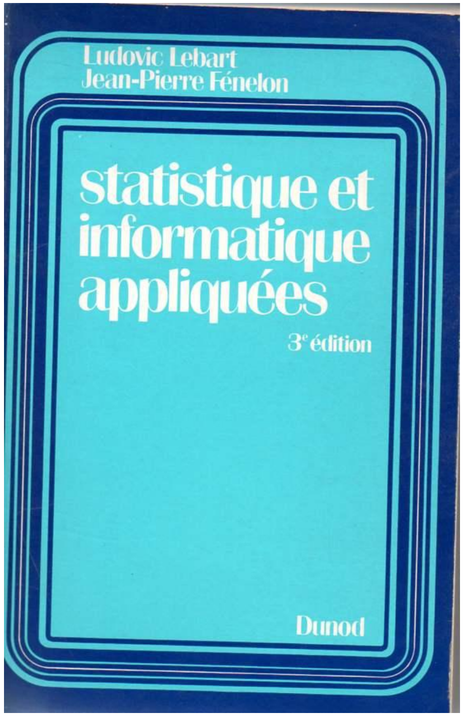
Benzécri J.-P. (1972) - Sur l'analyse des tableaux binaires associés à une correspondance multiple. *Note multigraphiée du Laboratoire de Statistique Mathématique*, Université Pierre et Marie Curie.



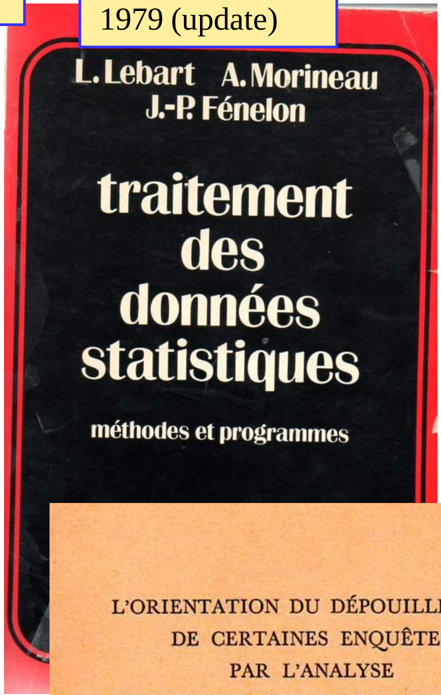
Jean-Pierre Fénelon, 1940-2002

(Photo : Marie-Odile Lebeaux, archives personnelles)

1971 – CA- PCA - + Fortran codes



1979 (update)



L'ORIENTATION DU DÉPOUILLEMENT
DE CERTAINES ENQUÊTES
PAR L'ANALYSE
DES CORRESPONDANCES MULTIPLES

par

Extrait de CONSOMMATION - ANNALES DU C.R.E.D.O.C
N° 2 (1975)
24-26, Bd de l'Hôpital **DUNOD** 75005 PARIS

1975 : MCA and the methodology of sample survey data processing

INFLUENCE DU CODAGE DES DONNÉES
EN ANALYSE FACTORIELLE DES CORRESPONDANCES
ÉTUDE D'UN EXEMPLE PRATIQUE MÉDICAL

Jean-Pierre NAKACHE
Groupe de recherche U 88
C.H.U. Pitié Salpétrière (Service du Professeur Grémy)

I – INTRODUCTION

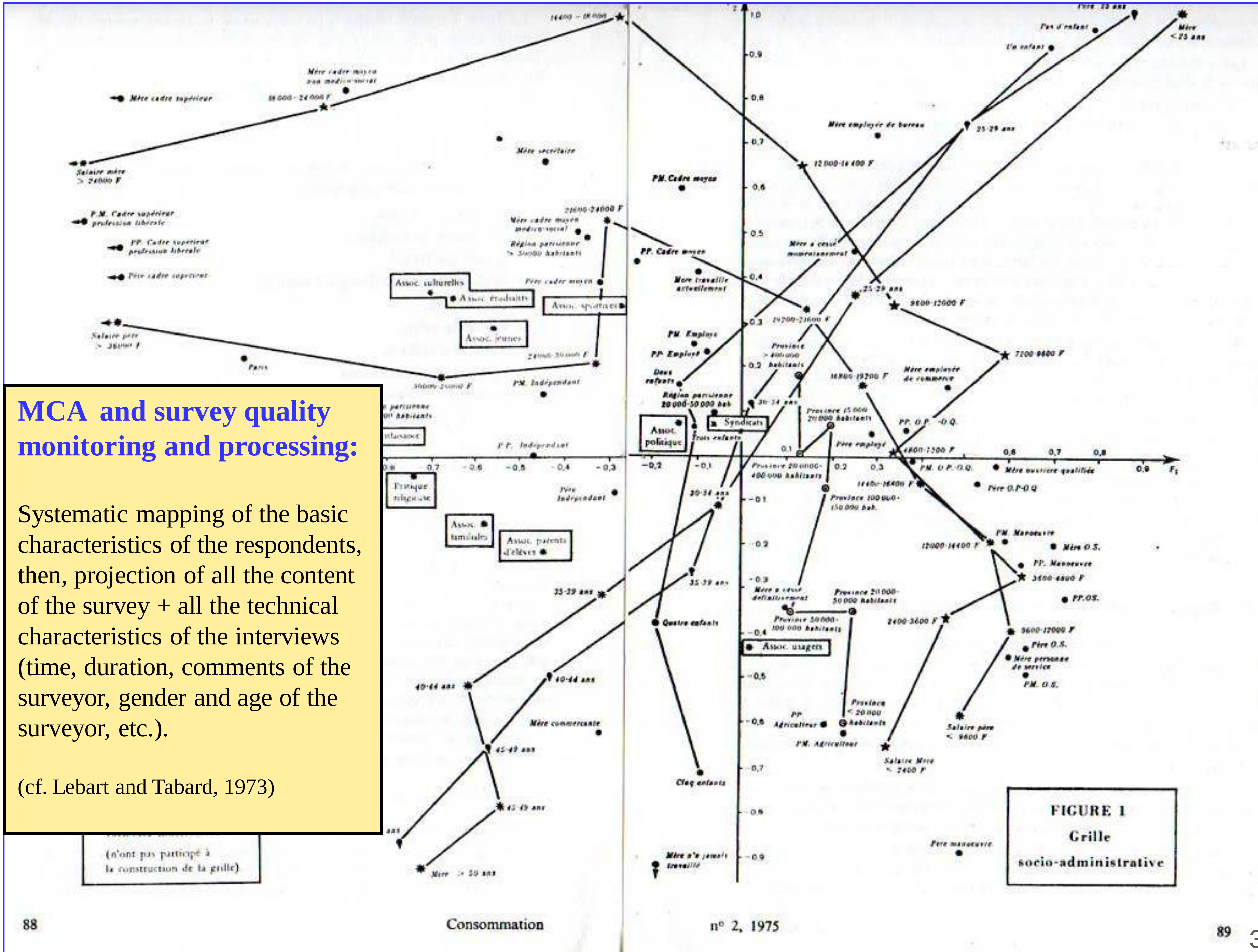
L'analyse factorielle des correspondances (Ref. ...), dont le principe est exposé succinctement en annexe, est une méthode d'analyse descriptive multidimentionnelle qui s'applique rigoureusement à des tableaux de contingence à n lignes et p colonnes. Les lignes représentent en général les "indi-

Nakache (1973) An application of MCA … under the name of CA…

MCA and survey quality monitoring and processing:

Systematic mapping of the basic characteristics of the respondents, then, projection of all the content of the survey + all the technical characteristics of the interviews (time, duration, comments of the surveyor, gender and age of the surveyor, etc.).

(cf. Lebart and Tabard, 1973)



About the overuse of MCA in the seventies

Samples of articles in weekly French magazines popularizing MCA (1974-1975)

24/09/04 15:05

Le prix d'un Français

Chaque semaine, nouvelobs.com publie le *Nouvel Observateur*... 30 ans avant. Au menu, l'article de couverture du numéro en question.


SPÉCIAL

LE PRIX D'UN FRANÇAIS

L'article de cover du Nouvel Observateur n° 514 du 16 septembre 1974

Cet article a paru dans *Le Nouvel Observateur* n°514 du 16 septembre 1974

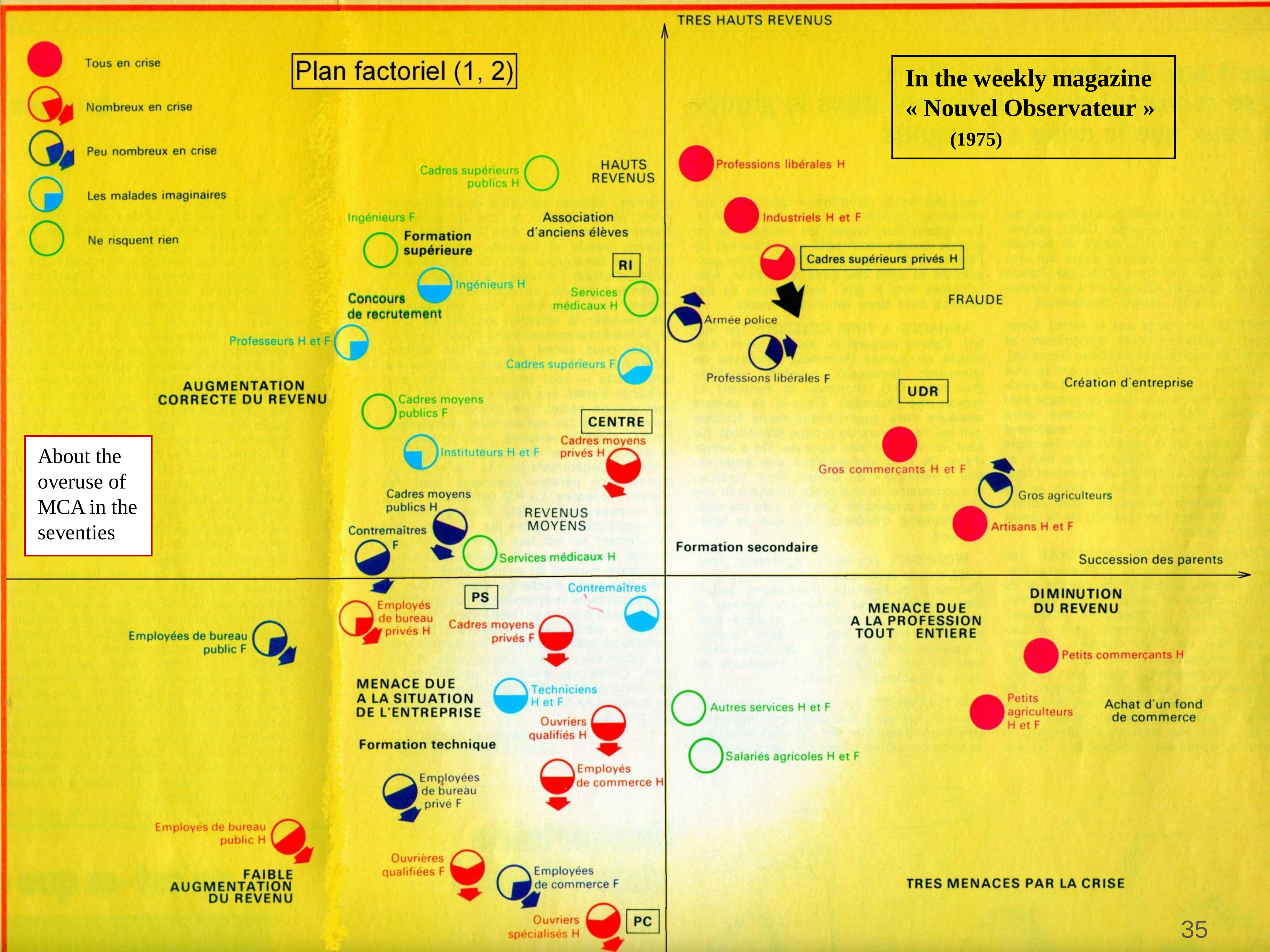
+T -T
Imprimer
Envoyer
Partager
Traduire
J'aime

Une grande enquête dirigée par François-Henri de Virieu

LE GATEAU EST PLUS GROS ? Bien sûr. Cela s'appelle l'expansion. Et la France a été, trente années durant, un pays en expansion. Mais les parts de ce gâteau ? Sont-elles moins arbitrairement découpées ? Sommes-nous plus près de cette « égalité » promise

In the weekly magazine « Nouvel Observateur » (1975)

About the overuse of MCA in the seventies



About the overuse of MCA in the seventies

In the same issue, trying to calm down the frenzy of journalists about CA and MCA (1974-1975) (Nouvel Observateur)

Les contrats faits par les sociétés d'i
sont de quarante heures par semaine mais l
correspond à trente-cinq h

Discours sur la "méthode"

Les résultats contenus dans la série d'articles publiés cette année encore par « le Nouvel Observateur » reposent sur des éléments d'importance variable. Un travail considérable sur le terrain, diversifié dans sa forme — questionnaires fermés et interviews —, un travail de documentation, enfin une étape d'analyse et de rédaction. Je crois qu'il est difficile de ne pas rendre hommage à la hardiesse, à la compétence et au talent des auteurs de cette entreprise unique en son genre.

La mise en avant d' « outils mathématiques sophistiqués » appelle cependant, de la part du chercheur et enseignant que je suis, quelques remarques ou questions.

Une question que beaucoup de lecteurs se poseront peut-être aussi : quelle est la part, dans les résultats finaux, de ces outils mathématiques ? Qu'on le veuille ou non, leur utilisation peut exercer une pression, engendrer un sentiment de dépossession chez les lecteurs non prévenus. Il est sans doute utile de dire ce qu'il serait advenu de cette étude en l'absence de ces traitements « sophistiqués ».

— Au départ, les auteurs auraient probablement hésité à recueillir une information aussi volumineuse auprès de chaque enquêté, de peur d'être submergés par des liasses de tableaux difficiles à lire, puis à synthétiser.

— Par la suite, en l'absence de traitements globaux, la cohérence interne des données n'aurait pu être éprouvée.

— Enfin, en l'absence de visualisation globale, la recherche des associations significatives aurait été laborieuse ; certains regroupements pertinents auraient pu échapper aux analystes.

En somme, les « outils mathématiques » ont sûrement permis un gain de productivité et de qualité.

Mais la partie la plus « massive » et indiscutable des résultats aurait pu être obtenue (avec des délais et des coûts plus importants) par des traitements classiques. Cela devrait réduire une frustration non prise en compte dans l'étude : celle du... lecteur non mathématicien lisant « le Nouvel Observateur ».

N'est-il pas alors dangereux de publier de telles « cartes », qui ne sont que des intermédiaires techniques au même titre que les radiographies ou les microphotographies utilisées lors de travaux médicaux ou biologiques ? Ces cartes, de prime abord très vivantes, car parsemées d'intitulés évocateurs ou familiers, ne peuvent, en fait, être lues qu'à la suite d'un véritable apprentissage.

Il faut féliciter les auteurs d'avoir insisté sur les précautions d'emploi des graphiques. Cependant, certaines règles de lecture auraient pu être mentionnées : la proximité entre deux points a d'autant plus de sens que ceux-ci sont éloignés de l'origine des axes. Autour de chaque point peut être tracée une zone de confiance qui permet d'apprécier l'incertitude sur la position du point due aux fluctuations d'échantillonnage, etc.

Ce n'est pas chercher à « confisquer le savoir » que de proposer que la publication de telles cartes soit réservée à des revues spécialisées : c'est simplement avoir le souci de limiter une certaine forme de vulgarisation et les malentendus, mythologies et contresens qui lui sont inhérents.

Les auteurs de l'étude sur « le Prix d'un Français », pionniers en la matière, ne méritent pas ces reproches. Mais on peut craindre que des émules moins attentifs ne suivent leurs traces et imputent à l'ordinateur et à la « science » des résultats qui ne refléteraient en fait que des a priori politiques ou autres, non explicitement formulés.

LUDOVIC LEBART
professeur de statistiques
à l'Institut de Statistiques
des Universités de Paris

mais il n'en trouve pas : *« Dans les annonces des journaux spécialisés, on propose des salaires vraiment dérisoires. »* Et il ne s'agit pas d'un

définitive, un blocage déguisé des salaires.
Il est évidemment dérisoire de parler ici de sécurité de l'emploi : *« On signe des contrats*

tandis que quatre-vingt-dix ont
porte. Manpower, la plus célèb
« Nous arrivons demain chez vou
tons quand le travail est fait »), a
son bulletin confidentiel que ses
premier trimestre de 1975 ont ét
à ceux de la période corespondar
En l'espace de quelques mois,
vient d'ouvrir des agences à Ma
Dunkerque, Saint-Nazaire, Vénissi
banne. Elle s'apprête à en ouvrir
Paris, dans le quartier Denfer
Selon la C.F.D.T., Manpower «
de prendre la tête des boîtes q
paient le moins ». Ceci explique
Certaines sociétés d'intérim
même à manquer de bras. *« Envo*
amis actuellement libres de tout
quel que soit leur âge. Et vous
francs par personne recommand
Contact Office, qui distribue des d
de petites silhouettes que les c
bras n'ont plus qu'à *« détacher s*
tillé » et à remplir. La Fédératio
nelle du Travail temporaire estime
vingt mille personnes sont emplo
manence dans les mille entrepris
ce qui représente un « passage »
personnes dans l'année !
Le travail intérimaire s'infiltre
faveur de la crise, assurant au
une main-d'œuvre docile et peu
phénomène a pris aujourd'hui u
pleur en France que c'est tout l'
rapports sociaux qui est en train
fier, par contagion. Le travail ter
met aux patrons non seulement c
sion sur les salaires pour sauve
profits, mais aussi de grignoter de
de la législation du travail, jugée
gnante. Les dispositions de la
travail sont bafouées. La duré
redevient arbitraire. Les conver
tives sont tournées. Les avantages
réduits. Et les syndicats, pris à c
une évolution qu'ils n'avaient pa
peuvent rien. Le temps de la g
quête patronale s'ouvre.
Qui sont-ils, ces nouveaux
« jaunes » malgré eux, dont la s

1976 – 1977

Series of research papers or books related to MCA

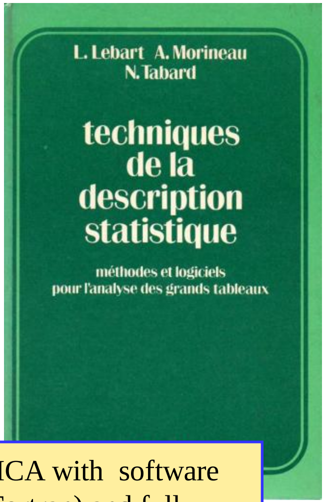
Pagès J.-P., Escoufier Y., Cazes P. (1976) - Opérateurs et analyse de tableaux à plus de deux dimensions. *Cahiers du BURO*, ISUP, Paris, p 61-89.

Escofier B., Le Roux B. (1976) – Influence d'un élément sur les facteurs en analyse des correspondances. *Les Cahiers de l'Analyse des Données*, 3, p 297-318.

Rouanet H., Lépine D. (1976) – A propos de l' *Analyse des Données* selon Benzécri. *L'Année Psychologique*, 76, p 133-144.

Caillez F., Pagès J.P. (1976) - *Introduction à l'Analyse des Données*. S.M.A.S.H., Paris.

Lebart L., Morineau A., Tabard N. (1977) - *Techniques de la description statistique.* Dunod, Paris. English translation in 1984: Lebart L., Morineau A., Warwick K.- *Multivariate Descriptive Statistical Analysis.* J. Wiley, NY. (translated without the (excellent) sociological examples from N. Tabard, by E. Berry, under the supervision of K. Warwick).



MCA with software (Fortran) and full-sized examples of application.

Discriminant analysis using MCA

Saporta G. (1977) - Une méthode et un programme d'analyse discriminante sur variables qualitatives. In : *Premières Journées Int. Analyse des Données et informatiques*, INRIA, Rocquencourt.

Contributions to MCA methodology

Cazes P. (1977) - Etude des propriétés extrêmales des sous-facteurs issus d'un sous-tableau d'un tableau de Burt. *Les Cahiers de l'Analyse des Données*, 2, p 143-160.

Cazes P. (1977 – collection of notes published in 1982) - Note sur les éléments supplémentaires en analyse des correspondances. *Les Cahiers de l'Analyse des Données*, 1, p 9-23; 2, p 133-154.

1979 – 1980

Improvements of « MCA technology »

Escofier B. (1979 a) - Traitement simultané de variables qualitatives et quantitatives. *Les Cahiers de l'Analyse des Données*, 4, (2), p 137-146.

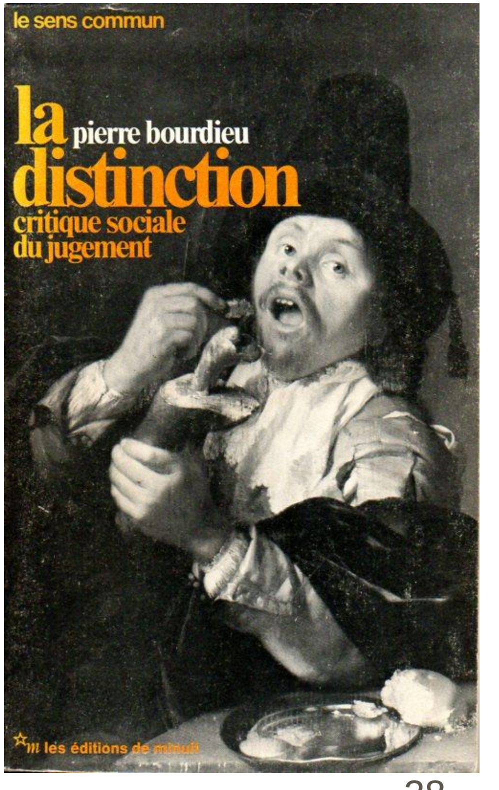
Escofier B. (1979 b) - Une représentation des variables dans l'analyse des correspondances multiples. *Revue de Statist. Appl.* , 27, p 37-47.

Benzécri J.-P. (1979) - Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire. *Cahiers de l'Analyse des Données*, 3, p 377-378 .

Cazes P. (1980) - Analyse de certains tableaux rectangulaires décomposés en blocs. *Les Cahiers de l'Analyse des Données*, 5, p 145-161, et p 387-403.

Emblematic application of MCA in sociology

Bourdieu P. (1979) – *La Distinction. Critique sociale du jugement.* Les Editions de Minuit. Paris.



Dual Scaling

Nishisato S.(1980) - *Analysis of Categorical Data. Dual Scaling and its Application*. Univ. of Toronto Press.

Homogeneity Analysis

Gifi A. (1981) - *Non Linear Multivariate Analysis*, Department of Data theory, University of Leiden. [and: Gifi A. (1990) - *Non Linear Multivariate Analysis*, J. Wiley, Chichester.] **(book recapitulating works done before 1980).**

References in connection with the presentation (before 1980, except for historical papers)

Pearson K. (1901) - On lines and planes of closest fit to systems of points in space. *Phil. Mag.* 2, n°ll, p 559-572.

Spearman C. (1904) - General intelligence, objectively determined and measured. *Amer. Journal of Psychology*, 15, p 201-293.

Richardson M., Kuder G. F. (1933) - Making a rating scale that measures. Procter and Gamble, *Personnel Journal*, 12, p 71-75.

Hirschfeld H. D. (1935) - A Connection between correlation and contingency. *Proc. Camb. Phil. Soc.* 31, p 520-524.

Eckart C., Young G. (1936) - The approximation of one matrix by another of lower rank. *Psychometrika*, l, p 211-218.

Hotelling H. (1933) - Analysis of a complex of statistical variables into principal components. *J. Educ. Psy.* 24, p 417-441, p 498-520.

Hotelling H. (1936) - Relation between two sets of variables. *Biometrika*, 28, p 129-149.

Fisher R. A. (1940) – The precision of discriminant functions. *Ann. Eugen. Lond.*, 10, 422-429.

Guttman L. (1941) - The quantification of a class of attributes: a theory and method of a scale construction. In : *The prediction of personal adjustment* (Horst P., ed.) p 251 -264, SSCR New York.

Maung K. (1941) – Measurement of association in a contingency table with special reference to the pigmentatin of hair and eye colours of Scottish schoolchildren. *Ann. Eugen. Lond*. 11, 189-223.

Thurstone L. L. (1947) - *Multiple Factor Analysis.* The Univ. of Chicago Press, Chicago.

Guttman, L. (1950) - The principal components of scale analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, &J . A. Clausen. *Measuremenat nd prediction.* Princeton: Princeton University Press.

Burt C. (1950) - The factorial analysis of qualitative data. *British J. of Statist. psychol*. 3, 3, p 166-185.

Williams E., J. (1952) – Use of scores for the analysis of association in contingency tables. *Biometrika*, 44, 274-289.

Hayashi, C. (1952) - On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view. *Annals of the Institute of Statistical Mathematics*, *3* (No. 2), 69-98.

Guttman L. (1953) – A note on Sir Cyril Burt's Factorial Analysis of Qualitative Data, *British J. of Statist. psychol*. 6, p 1-4.

Burt C. (1953) – Scale Analysis and factor analysis. Comments on Dr Guttman paper. *British J. of Statist. psychol*. 6, p 5-20.

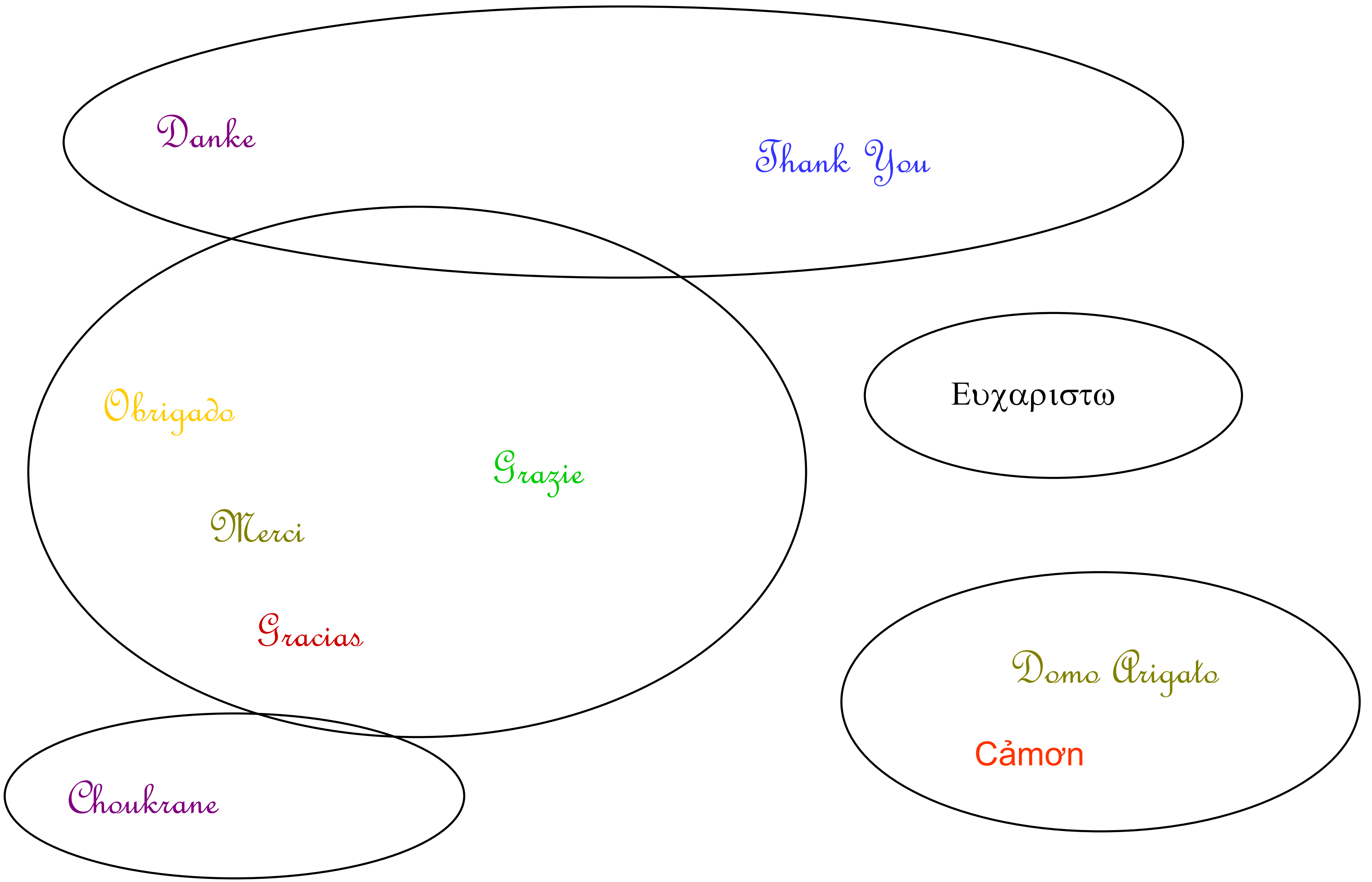
Hayashi C.(1956) - Theory and examples of quantification. (II) *Proc. of the Institute of Statist. Math.* 4 (2), p 19-30.
Horst P. (1961) - Relation among m sets of measures. *Psychometrika*, 26, p 129-149.
Lancaster H. O. (1963) - Canonical correlation and partition of Chi-Square. *Quart. J. Math.*, 14, p 220-224.
Benzecri J.-P. (1964) – *Cours de Linguistique mathématique*, Faculté des Sciences de Rennes.
Horst P. (1965) - *Factor Analysis of Data Matrices*. Holt, Rinehart, Winston, New York.
Escofier B. [Cordier B.] (1965) - *l'Analyse des correspondances.* Thèse, Faculté des Sciences de Rennes ; published in 1969 in: *Cahiers du Bureau Universitaire de Recherche Opérationnelle*, n°13.
Gower J. C. (1966) - Some distance properties of latent and vector methods used in multivariate analysis. *Biometrika*, 53, p 325-328.
Gower J. C. (1968) - Adding a point to vector diagram in multivariate analysis. *Biometrika*, 55, 582-585.
Carroll J. D. (1968) - Generalization of canonical correlation to three or more sets of variables. *Proc. Amer. Psychological Assoc.* p 227-228.
Lancaster H. O. (1969) - *The Chi-squared Distribution*. J. Wiley, New York.
Benzécri J.-P. (1969) - Statistical analysis as a tool to make patterns emerge from clouds. In : *Methodology of Pattern Recognition* (S.Watanabe, Ed.) Academic Press, p 35-74.
Lebart L. (1969) - Introduction à l'analyse des données : Analyse des correspondances et validité des résultats. *Consommation*, Dunod. 4, p 65-87.
Gabriel K.R. (1971) - The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58, 3, p 453-467.
Kettenring R. J. (1971) - Canonical analysis of several sets of variables. *Biometrika*, 58, (3), p 433-450.
Benzécri J.-P. (1972) - Sur l’analyse des tableaux binaires associés à une correspondance multiple. *Note multigraphiée du Laboratoire de Statistique Mathématique*, Université Pierre et Marie Curie.
Nakache J.P. (1973) - Influence du codage des données en analyse factorielle des correspondances. Etude d'un exemple pratique médical. *Revue Statist. Appl.*, 21, (2).
de Leeuw J. (1973) – Canonical Analysis of Categorical Data. Unpublished Dissertation. University of Leiden, Leiden.
Benzécri J.-P. (1973) - *L'Analyse des Données*. Tome 1: *La Taxinomie*. Tome 2: *L'Analyse des Correspondances* (2de. éd. 1976). Dunod, Paris.
Lebart L., Tabard N. (1973) - *Recherches sur la description automatique des données socio-economiques*. Rapport CORDES-CREDOC, Convention de Recherche n°13/1971.

Lebart L. (1974) - On the Benzécri's method for finding eigenvectors by stochastic approximation. *Proceedings in Comp. Statist., In: COMPSTAT*, Physica verlag, Vienna, p 202-211.
Hill M.O. (1974) - Correspondence analysis: a neglected multivariate method. *Appl. Statist.* 3, p 340-354.
Lebart L. (1975) - L'orientation du dépouillement de certaines enquêtes par l'analyse des correspondances multiples. *Consommation*, 2, p 73-96. Dunod.
Saporta G. (1975) - Dépendance et codage de deux variables aléatoires. *Revue Statist. Appl.* 23, p 43-63.
Saporta G. (1975) – Données supplémentaires sur l'analyse des données. *L'Echo des Messaches*, 4, 4-5.
Pagès J.-P., Escoufier Y., Cazes P. (1976) - Opérateurs et analyse de tableaux à plus de deux dimensions. *Cahiers du BURO,* ISUP, Paris, p 61-89.
Rouanet H., Lépine D. (1976) – A propos de l' *Analyse des Données* selon Benzécri. *L'Année Psychologique*, 76, p 133-144.
Escofier B., Le Roux B. (1976) – Influence d'un élément sur les facteurs en analyse des correspondances. *Les Cahiers de l'Analyse des Données,* 3, p 297-318.
Caillez F., Pagès J.P. (1976) - *Introduction à l'Analyse des Données*. S.M.A.S.H., Paris.
Lebart L., Morineau A., Tabard N. (1977) - *Techniques de la description statistique.* Dunod, Paris. English translation in 1984: Lebart L., Morineau A., Warwick K.- *Multivariate Descriptive Statistical Analysis.* J. Wiley, NY. (translated without the (excellent) sociological examples from N. Tabard, by E. Berry, under the supervision of K. Warwick).
Saporta G. (1977) - Une méthode et un programme d'analyse discriminante sur variables qualitatives. In : *Premières Journées Int. Analyse des Données et informatiques*, INRIA, Rocquencourt.
Cazes P. (1977) - Etude des propriétés extrêmales des sous-facteurs issus d'un sous-tableau d'un tableau de Burt. *Les Cahiers de l'Analyse des Données,* 2, p 143-160.
Cazes P. (1977 – collection of notes published in 1982) - Note sur les éléments supplémentaires en analyse des correspondances. *Les Cahiers de l'Analyse des Données,* 1, p 9-23; 2, p 133-154.
Escofier B. (1979 a) - Traitement simultané de variables qualitatives et quantitatives. *Les Cahiers de l'Analyse des Données*, 4, (2), p 137-146.
Escofier B. (1979 b) - Une représentation des variables dans l'analyse des correspondances multiples. *Revue de Statist. Appl.* , 27, p 37-47.
Benzécri J.-P. (1979) - Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire. *Cahiers de l'Analyse des Données,* 3, p 377-378 .
Bourdieu P. (1979) – *La Distinction. Critique sociale du jugement.* Les Editions de Minuit. Paris.

Cazes P. (1980) - Analyse de certains tableaux rectangulaires décomposés en blocs. *Les Cahiers de l'Analyse des Données*, 5, p 145-161, et p 387-403.
Bastin C., Benzécri J.-P., Bourgarit C., Cazes P. (1980) – *Pratique de l'analyse des données*. Dunod, Paris.
Nishisato S.(1980) - *Analysis of Categorical Data. Dual Scaling and its Application*. Univ. of Toronto Press.
Gifi A. (1981) - *Non Linear Multivariate Analysis*, Department of Data theory, University of Leiden. [and: Gifi A. (1990) - *Non Linear Multivariate Analysis*, J. Wiley, Chichester.] (book recapitulating works done before 1980).

Historical papers or books

Benzecri J.-P. (1976) – Histoire et Préhistoire de l'Analyse des Données. *Les Cahiers de l'Analyse des Données*, 1976, 1, p 9-32 ; 1976, 2, p 101-120, 1976, 3, p 221-241 ; 1977, 1, p 9-40.
Guttman L. (1978) – Cyril Burt and the careless star worshippers. *Technical Report, The Hebrew University*, Department of Sociology, p 1- 7.
Hearnshaw L. S. (1979) – *Cyril Burt, Psychologist*. Hodder and Stoughton, London.
de Leeuw J. (1983) – On the prehistory of correspondence analysis. *Statistica Neerlandica*, vol 37, n°4, p 161-164.
Gould S.J. (1983) – The real error of Cyril Burt. In: *The Mismeasure of Man*. W.W. Norton and Company, New York. Chapter 6, p 234-320.
Tenenhaus M., Young F. W. (1985) - An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika*, p 91-119.
Benzécri J.-P. (1992) – *Correspondence Analysis Handbook*. Marcel Dekker, New York. (This book is not the translation of the 1973 book, but recapitulates series of lectures aiming at a larger readership).
van der Heijden P.G.M., Sijtsma K. (1996) – Fifty years of measurement and scaling in the Dutch social science. *Statistica Neerlandica*, vol 50, n°1, p 111-135.
Gower J.C. (2008) - The biological stimulus to multidimensional data analysis. *JEHPS*. December 2008, vol 4, n°2. (www.jehps.net).
Murtagh F. (2008) - Origins of modern data analysis linked to the beginnings and early development of computer science and information engineering. . *JEHPS*. December 2008, vol 4, n°2. (www.jehps.net).
Heiser W (2008) - Psychometric roots of multidimensional data analysis in the Netherlands: From Gerard Heymans to John van de Geer. *JEHPS*. December 2008, vol 4, n°2. (www.jehps.net).


Danke
Thank You
Obrigado
Grazie
Merci
Gracias
Ευχαριστω
Domo Arigato
Cảmơn
Choukrane